UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
CURSO ENGENHARIA DE PRODUÇÃO CIVIL

Joao Marcos Cipriani Bertelli Correa

Uma análise exploratória de indicadores do sistema aéreo brasileiro a partir de Modelos *Machine Learning*

Florianópolis
2023

Joao Marcos Cipriani Bertelli Correa

Uma análise exploratória de indicadores do sistema aéreo brasileiro a partir de Modelos Machine Learning

Trabalho de Conclusão de Curso submetido ao curso de Engenharia de produção civil do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de bacharel em Engenharia de Produção Civil .

Orientador(a): Prof. Dr. Ricardo Villarroel Dávalos

Florianópolis

2023

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

CORREA, JOAO MARCOS CIPRIANI BERTELLI
Uma análise exploratória de indicadores do sistema aéreo brasileiro a partir de Modelos Machine Learning / JOAO MARCOS CIPRIANI BERTELLI CORREA ; orientador, RICARDO VILLAROEL DÁVALOS, 2023.
146 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Civil, Florianópolis, 2023.

Inclui referências.

1. Engenharia de Produção Civil. 2. Ciência de dados,. 3. Sistema Aéreo. 4. Machine Learning. I. DÁVALOS, RICARDO VILLAROEL . II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Civil. III. Título.

Joao Marcos Cipriani Bertelli Correa

Título: Uma análise exploratória de indicadores do sistema aéreo brasileiro a partir de Modelos *Machine Learning*

Local Florianópolis, 28 de junho de 2023.

Este Trabalho de Conclusão de Curso foi avaliado e aprovado pela banca examinadora composta dos seguintes membros

Prof.(a) Ricardo Villaroel Dávalos, Dr.(a)
Orientador(a)

Prof.(a) Maurício Uriona Maldonado, Dr.(a)

Prof.(a) Lucas Vieira Werner, Mestre

Certifico que esta é a versão final do Trabalho de Conclusão de Curso apresentado pelo autor e julgado adequado por mim e pelos demais membros da banca para obtenção do título de Bacharel em Engenharia de produção civil

ICP-Edu
Documento assinado digitalmente
Ricardo Villarroel Davalos
Data: 10/07/2023 15:31:07-0300
CPF: ***.328.619-**
Verifique as assinaturas em https://v.ufsc.br

Prof.(a) Ricardo Villaroel Dávalos, Dr.(a)
Orientador(a)

AGRADECIMENTOS

Agradeço aos meus pais, Lenita Cipriani e Antônio Marcos Bertelli Corrêa e toda minha família que esteve comigo em todos os momentos da graduação.

Aos meus amigos e colegas que fizeram parte dessa jornada ao longo desses anos todos.

Ao professor Maurício Uriona Maldonado com quem foi possível consolidar diversos conhecimentos necessários para a realização deste trabalho e para meu crescimento profissional.

E principalmente, ao orientador Ricardo Villarroel Dávalos, por me auxiliar e incentivar em todos os momentos do desenvolvimento deste trabalho.

RESUMO

A utilização da ciência de dados se torna cada vez mais crucial nos dias atuais, impulsionada pelo crescente volume de dados disponíveis e pela necessidade de tomar decisões informadas em diversas áreas. A ciência de dados combina habilidades estatísticas, conhecimento em programação e expertise em domínio para extrair insights valiosos e criar valor a partir dos dados. Da mesma forma o sistema aéreo desempenha um papel de extrema importância em nossa sociedade globalizada e conectada. Esse sistema abrange uma série de elementos, como aeroportos, companhias aéreas, rotas de voo, controle de tráfego aéreo e infraestrutura relacionada, que trabalham em conjunto para facilitar o transporte aéreo de pessoas e mercadorias em todo o mundo. Dentro desse cenário este trabalho pretende realizar a modelagem de dados a fim de gerar visualizações para os principais indicadores de desempenho aeroportuário e de reclamações dos clientes para os três estados de Santa Catarina, Paraná e Rio Grande do Sul. Para realizar tal levantamento necessita-se realizar a extração, carregamento e limpeza dos dados para que não se gere visualizações com distorções. Além disso o trabalho se propõe a avaliar e definir o melhor modelo de *machine learning* entre regressão linear, árvore de decisão e floresta aleatória para prever o tempo de resposta das chamadas abertas pelos consumidores. A partir desses modelos é possível medir o sucesso de cada um através de métrica estabelecidas pelo trabalho. Através dessas métricas foi possível definir que a floresta aleatória gerou resultados melhores o que permitiu modelar os dados finais para que se fosse possível prever o tempo de resposta de uma reclamação, além de quais são os fatores mais preponderantes para o tempo de resposta.

Palavras-chave: Ciência de dados, *Machine Learnig*, sistema aéreo brasileiro, análise exploratória dos dados.

ABSTRACT

The utilization of data science has become increasingly crucial in today's world, driven by the growing volume of available data and the need to make informed decisions in various fields. Data science combines statistical skills, programming knowledge, and domain expertise to extract valuable insights and create value from data. Similarly, the aviation system plays an extremely important role in our globalized and interconnected society. This system encompasses various elements such as airports, airlines, flight routes, air traffic control, and related infrastructure, all working together to facilitate air transportation of people and goods worldwide. Within this context, this study aims to perform data modeling in order to generate visualizations for key airport performance indicators and customer complaints for the states of Santa Catarina, Paraná, and Rio Grande do Sul. To accomplish this, it is necessary to extract, load, and clean the data to ensure that visualizations are not distorted. Additionally, the study aims to evaluate and determine the best machine learning model among logistic regression, decision trees, and random forests to predict the response time of customer inquiries. Using these models, the success of each one can be measured through established metrics in the study. Based on these metrics, it was determined that the random forest model produced better results, enabling the modeling of the final data to predict the response time of a complaint and identify the most influential factors in the response time. By combining the power of data science with insights from the aviation system, this study aims to provide valuable information and actionable insights that can contribute to enhancing airport performance and improving customer satisfaction. Through the analysis of data and the application of machine learning models, the study seeks to uncover patterns, trends, and factors that impact response times and enable more efficient and effective management of customer complaints in the aviation industry. The findings of this research have the potential to drive improvements in service delivery, operational efficiency, and overall customer experience in the aviation system.

Keywords: Data Science, *Machine Learning*, Brazilian aviation system, Exploratory data analysis

SUMÁRIO

1 INTRODUÇÃO

O sistema aéreo brasileiro movimenta anualmente milhões de passageiros e toneladas de carga. No entanto, a complexidade e a grande quantidade de informações envolvidas na operação e gestão do sistema podem dificultar a tomada de decisões efetivas (JAKOB, 2018). Nesse contexto, a ciência de dados surge como uma ferramenta capaz de extrair insights e conhecimentos a partir dos dados gerados pelo sistema aéreo.

Desde o surgimento dos sistemas computacionais, um dos principais objetivos das organizações tem sido o de armazenar um grande volume de dados. Essa tendência ficou ainda mais visível com a grande queda nos custos de aquisição de hardware, devido a implantação de novas tecnologias, tornando possível armazenar quantidades cada vez maiores de informações. Novas e mais complexas estruturas de armazenamento foram desenvolvidas, tais como: depósitos de dados virtuais, bibliotecas virtuais e armazenamento em nuvem, pois além de armazenar estes dados é necessário buscá-los com facilidade (CAMILO; SILVA, 2009).

Grande parte das atividades realizadas em um aeroporto geram metadados que são armazenados em bancos de dados virtuais, que isolados podem não transmitir uma informação útil diretamente, porém ao se relacionar o conteúdo, o conceito e o histórico daquele dado, pode-se obter informações úteis do processo. Estas informações reorganizadas podem gerar insights relacionados à situações-chave podendo contribuir na otimização deste, seja na diminuição de recursos no processo, redução do número de atividades ou custos, tempo de espera, entre outros. O armazenamento de grande volume de dados não é uma proposta recente e remonta do início da maior utilização de sistemas de informação pelas empresas, na década de 1950 (CHEN; CHIANG; STOREY, 2012).

A utilização de ciência de dados evolui diariamente e diversas soluções já foram apresentadas. Empresas aéreas utilizam técnicas de análise de dados para melhorar a gestão de suas operações. Por exemplo, a Delta Airlines desenvolveu um modelo de previsão de atrasos em voos, utilizando dados históricos e informações em tempo real para minimizar atrasos e melhorar a satisfação dos passageiros (GORMAN, 2019).

Este trabalho tem como proposta desenvolver um dashboard na linguagem R visando centralizar e avaliar um conjunto de indicadores do tráfego aéreo e de reclamação dos consumidores. O estudo de caso considera técnicas de *machine learning* aplicadas às bases de dados dos três estados da região sul.

1.1 PROBLEMA

Dados são uma parte fundamental da tomada de decisões e análise de negócios. Eles fornecem informações valiosas que podem ser usadas para entender tendências, identificar oportunidades e tomar decisões informadas. No entanto, os dados podem enfrentar vários problemas que podem afetar sua qualidade, precisão e utilidade. Conjuntos de dados extremamente grandes e complexos podem não ser

manuseados facilmente utilizando ferramentas convencionais. Dada a dificuldade de se visualizar informações contidas nesses dados, propõe-se o desenvolvimento de uma ferramenta capaz de trazer indicadores a respeito da área de estudo, além de fornecer previsões sobre determinado ponto e se poder definir quais são os critérios mais importantes, ou seja, que tem maior relação com a variável que se deseja prever.

A aviação é uma indústria altamente regulamentada, com normas e procedimentos estabelecidos para garantir a integridade das aeronaves e a proteção dos passageiros. Um sistema de controle eficiente permite a monitorização constante dos indicadores. Isso possibilita a identificação precoce de possíveis problemas e a adoção de medidas preventivas para evitar a ocorrência de incidentes e acidentes aéreos (Deng et al., 2017).

Dentro deste contexto o desenvolvimento de ferramentas para apoio na tomada de decisão se torna muito importante, pois aumentam as chances das decisões serem assertivas. A ferramenta apresenta diversos indicadores como horários e datas com mais voos, cancelamentos, utilização por companhias, reclamações por companhia entre outras. Esses dados permitem a quem os analisa uma série de possibilidades para ajudar na hora de decidir alguma questão.

Um dos principais problemas enfrentados por aqueles que não utilizam a análise exploratória dos dados é a falta de conhecimento sobre a estrutura e as características dos dados em questão. Sem uma exploração adequada, é difícil identificar possíveis padrões, tendências ou relações existentes nos dados. Isso pode levar a interpretações equivocadas ou conclusões precipitadas, comprometendo a validade das análises realizadas.

Outro desafio é a possibilidade de não identificar a presença de outliers, valores discrepantes que podem distorcer as análises estatísticas. Sem uma análise exploratória, esses valores atípicos podem passar despercebidos, levando a conclusões imprecisas e decisões inadequadas baseadas nos dados incorretos. A identificação e o tratamento adequado dos outliers são essenciais para garantir a confiabilidade e a robustez dos resultados.

Machine Learning é uma tecnologia que está sendo cada vez mais utilizada em diferentes setores para tomadas de decisão mais precisas e informadas. Por meio da análise de grandes volumes de dados, os modelos de *Machine Learning* são capazes de identificar padrões e fazer previsões com alto grau de acurácia.

O tempo de resposta dos chamados abertos pelos consumidores é um importante dado gerado da utilização do sistema aéreo, dado que a partir do tempo de resposta é possível entender o grau de dificuldade para resolução do problema que gerou a abertura do chamado. Além disso o tempo de resposta permite traçar um paralelo com a avaliações dos chamados abertos.

1.2 OBJETIVOS

1.2.1 Objetivo geral

O objetivo geral deste trabalho é apresentar uma ferramenta para avaliar os tempos de resposta para reclamações dos clientes do sistema aéreo brasileiro a partir de *machine learning*.

1.2.2 Objetivos específicos

Os objetivos específicos deste trabalho encontram-se definidos a seguir.

- Avaliar e definir o melhor modelo de *machine learning* entre regressão linear, árvore de decisão e floresta aleatória para prever o tempo de resposta das chamadas abertas pelos consumidores.
- Carregamento e modelagem dos dados a partir de tecnologias livres;
- Pesquisar e definir por estado os principais indicadores de desempenho aeroportuário e de reclamações dos clientes;
- Validar os resultados através de técnicas de erros (R^2, MAE, RMSE, MAPE).

1.3 JUSTIFICATIVA

Em um cenário altamente competitivo, as empresas aéreas buscam constantemente maneiras de aumentar sua eficiência e reduzir seus custos, sem comprometer a segurança e a qualidade do serviço prestado aos clientes. Nesse sentido, a utilização de técnicas de ciência de dados pode ser um diferencial para as empresas, permitindo que elas tomem decisões baseadas em informações precisas e em tempo real.

A implementação de um sistema de controle eficiente dos indicadores de tráfego no sistema aéreo é de extrema relevância para garantir a segurança, a eficiência operacional e a qualidade dos serviços (Blom et al., 2016). Com o crescimento do tráfego aéreo em escala global, é fundamental desenvolver estratégias e metodologias de controle que possam lidar com os desafios e demandas crescentes do setor.

O controle dos indicadores de tráfego é essencial para manter a qualidade das operações aéreas. A qualidade do sistema aéreo é um aspecto fundamental para garantir a segurança e a satisfação dos passageiros. Um sistema aéreo de alta qualidade envolve a eficiência operacional, a pontualidade dos voos, a confiabilidade dos serviços e a excelência no atendimento ao cliente" (SMITH, 2019).

Além da segurança, o controle eficiente dos indicadores de tráfego também contribui para a eficiência operacional do sistema aéreo. Ao monitorar e gerenciar esses indicadores de forma adequada, é possível otimizar o fluxo de tráfego, minimizar atrasos, evitar congestionamentos e melhorar a pontualidade dos voos (Mendes et al., 2018). Isso resulta em uma utilização mais eficiente do espaço aéreo

e dos recursos disponíveis, maximizando a capacidade dos aeroportos e proporcionando uma experiência mais fluida e satisfatória para os passageiros.

Combinando as informações de tráfego aos dados de reclamação dos consumidores contribui ainda mais para uma análise completa da situação. A análise dos consumidores de tráfego aéreo possibilita a identificação de tendências de demanda, auxiliando as companhias aéreas na otimização de suas rotas e na oferta de voos mais adequados às necessidades dos passageiros" (CHOI et al., 2017). Permitindo que as companhias aéreas identifiquem padrões de reclamação e resolvam problemas recorrentes, aprimorando assim a qualidade do serviço oferecido (GONZÁLEZ-PRIETO et al., 2020). Essas informações juntas permitem uma tomada de decisão mais informada e embasada em dados reais.

Análogo a isso, o uso de dados nos mercados produtivos vem crescendo cada vez mais, com intuito de tornar decisões mais assertivas e lucrativas. Com esse nível imenso de informações, as empresas detêm a oportunidade de trabalhar com informações cada vez mais precisas e em uma granularidade cada vez maior, podendo identificar tendências ou características mais detalhadas. Com base nessa gigantesca quantidade de dados, se gerenciados de forma apropriada, pode-se fornecer às empresas vantagens competitivas sustentáveis e, ainda, a geração de valor para os stakeholders (JANSSEN; VOORT; WAHYUDI, 2016).

Para uma previsão mais assertiva o trabalho se propõe a treinar os dados em três modelos de *machine learning*, regressão linear, árvore de decisão e *random forest*, a fim de definir baseado em seus resultados estatísticos, R-Quadrado, MAPE, MAE e RMSE o melhor modelo para previsão do tempo de resposta.

Prever o tempo de resposta para a resolução de um problema aberto pela central de consumidores permite o controle da satisfação do cliente, visto que notas da avaliação do serviço se tornam cada vez mais maiores a medida que o tempo de resposta decresce.

Ao usar mais de um modelo de treinamento, é possível comparar seus resultados para escolher o mais adequado para um determinado conjunto de dados. Essa estratégia aumenta a chance de selecionar o modelo mais adequado para a tarefa em questão, dado que os modelos não são inerentemente bons ou ruins, apenas se adequam mais ou menos para tarefas específicas

Utilizar mais de um modelo de treinamento também pode ajudar a reduzir possíveis erros e vieses em um único modelo, aumentando a confiabilidade e a precisão das previsões. Além disso, é importante ter em mente que um único modelo pode apresentar limitações ou suposições incorretas, tornando a utilização de várias abordagens importantes (CORTHELL, 2016).

1.4 ESTRUTURA DO TCC

Este trabalho conta com cinco capítulos referentes à utilização de técnicas de análise exploratória dos dados e aprendizado de máquina sobre dados de tráfego aéreo e de consumidores do setor aéreo da região sul. O primeiro capítulo contém a introdução, onde se é situado o contexto do presente trabalho. No segundo capítulo,

apresenta-se o referencial teórico que contempla as técnicas e indicadores presentes nos capítulos posteriores. O terceiro capítulo apresenta informações do trabalho como as etapas metodológicas seguidas e delimitações necessárias do estudo. O quarto capítulo contempla o estudo de caso e seus resultados. E no último capítulo, apresenta-se o fechamento e suas devidas conclusões.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta um enquadramento teórico das três áreas do conhecimento necessárias para o desenvolvimento de um Dashboard, desde sua extração, tratamento e visualização dos dados.

A primeira aborda o sistema aéreo e seus indicadores. Na segunda parte apresenta-se um enquadramento teórico de ciência de dados, abordando os principais conceitos ligados à elaboração deste trabalho. Na terceira apresenta-se o aprendizado de máquina e os modelos abordados que corroboram este estudo.

2.1 SISTEMA AÉREO

2.1.1 Tráfego aéreo e utilização

O controle do tráfego aéreo desempenha um papel crucial na gestão e segurança do sistema aéreo. Ele envolve a coordenação e monitoramento das aeronaves em voo, garantindo a separação adequada entre elas e permitindo um fluxo eficiente de tráfego. A implementação de um sistema eficaz de controle do tráfego aéreo é essencial para prevenir colisões, minimizar atrasos e congestionamentos, e garantir a segurança das operações aéreas.

Além disso, o controle do tráfego aéreo também envolve o gerenciamento do espaço aéreo e a coordenação de atividades em diferentes regiões e aeroportos. Isso inclui a designação de rotas de voo, o planejamento de capacidade dos aeroportos e a gestão do fluxo de tráfego em momentos de alta demanda. Para realizar essas tarefas de forma eficiente, são utilizados sistemas de gerenciamento de fluxo de tráfego aéreo, que utilizam algoritmos e modelos de previsão para otimizar a utilização do espaço aéreo e evitar congestionamentos (D'Auria et al., 2020).

Os sistemas de gerenciamento do tráfego aéreo também desempenham um papel essencial no controle do sistema aéreo. Eles coletam e processam dados em tempo real sobre o tráfego aéreo, permitindo uma visualização clara e abrangente da situação. Esses sistemas utilizam algoritmos e modelos matemáticos avançados para prever a demanda de tráfego, otimizar rotas e garantir a utilização eficiente do espaço aéreo (Rao et al., 2017).

Os indicadores de utilização de voos no sistema aéreo brasileiro são elementos essenciais para a compreensão e o monitoramento do fluxo de tráfego aéreo no país. Esses indicadores fornecem informações valiosas sobre a demanda, a ocupação das aeronaves, a eficiência operacional das companhias aéreas e a capacidade dos aeroportos. O estudo desses indicadores é de grande importância para a tomada de decisões estratégicas e o planejamento adequado do setor aéreo.

Um indicador importante é a taxa de utilização dos aeroportos, que mede a capacidade de atendimento desses locais em relação ao número de voos programados. A análise desse indicador permite identificar gargalos e possíveis congestionamentos nos aeroportos, além de auxiliar no planejamento de

investimentos em infraestrutura aeroportuária (ANAC, 2019). O estudo da taxa de utilização dos aeroportos é fundamental para garantir a eficiência operacional do sistema aéreo e proporcionar uma experiência satisfatória para os passageiros.

Indicador relevante é a pontualidade dos voos. Esse indicador mede o percentual de voos que decolam e aterrissam dentro do horário programado. A pontualidade é um fator crucial para garantir a confiabilidade e a satisfação dos passageiros, além de otimizar o uso dos recursos aeroportuários (Aguiar et al., 2020). Além disso, a densidade de tráfego aéreo é um indicador importante para avaliar a utilização do espaço aéreo. Esse indicador mede a quantidade de aeronaves em um determinado espaço e tempo. Uma alta densidade de tráfego pode indicar a necessidade de adotar medidas para garantir a segurança e a fluidez das operações (Medeiros et al., 2019).

2.1.2 Dados dos consumidores

A coleta e análise de dados dos consumidores que utilizaram o sistema aéreo brasileiro desempenham um papel crucial na compreensão das necessidades, preferências e experiências dos passageiros. Esses dados oferecem aspectos fundamentais dos dados dos consumidores é a avaliação da satisfação do cliente para as companhias aéreas, agências reguladoras e demais envolvidos no setor, permitindo a tomada de decisões embasadas e a melhoria contínua dos serviços prestados.

Por meio de pesquisas de satisfação e feedback direto dos passageiros, é possível obter informações sobre diversos aspectos, como atendimento ao cliente, conforto das aeronaves, pontualidade dos voos, qualidade dos serviços de bordo e eficiência dos procedimentos de embarque e desembarque. Esses dados ajudam as companhias aéreas a identificar áreas de melhoria e a implementar ações corretivas para proporcionar uma experiência mais satisfatória aos passageiros (ANAC, 2020). A segurança é outra área em que os dados dos consumidores desempenham um papel crucial. Através do registro e análise de incidentes de segurança reportados pelos passageiros, as agências reguladoras podem identificar padrões e tendências, tomar medidas preventivas e garantir a contínua melhoria dos padrões de segurança no sistema aéreo brasileiro. (ANAC, 2020).

A análise dos indicadores de dados dos consumidores desempenha um papel fundamental na compreensão da experiência dos usuários. Esses indicadores permitem levantar informações, permitindo medir a satisfação dos passageiros, qualidade dos serviços prestados e as necessidades dos consumidores. Por meio da análise desses indicadores, é possível identificar áreas de melhoria, tomar decisões embasadas e aprimorar a experiência de voo para os passageiros.

Um indicador importante é o número de reclamações recebidas pelos órgãos reguladores e pelas companhias aéreas. Essas reclamações podem abranger uma variedade de questões, desde atrasos e cancelamentos de voos até problemas com bagagem e atendimento ao cliente. A análise dessas reclamações ajuda a identificar os principais problemas enfrentados pelos passageiros e a implementar medidas corretivas. (ANAC, 2020).

Além disso, a análise dos indicadores de dados dos consumidores pode incluir a avaliação da qualidade dos serviços de atendimento ao cliente, como a rapidez e eficácia no tratamento de reclamações e solicitações. Essa avaliação pode ser feita por meio de pesquisas de satisfação e feedback direto dos passageiros. (ANAC, 2020).

Santos et al. (2019) destaca a importância da análise dos indicadores de dados dos consumidores para a personalização dos serviços. Ao compreender as preferências individuais dos passageiros, como destinos frequentes, horários de voo preferidos e serviços adicionais desejados, as companhias aéreas podem customizar sua oferta, proporcionando uma experiência mais personalizada e adequada às necessidades de cada passageiro. Outro aspecto relevante é a análise dos indicadores de dados dos consumidores para a gestão da reputação das companhias aéreas. A coleta e análise de avaliações e feedbacks dos passageiros permitem monitorar a percepção do público em relação aos serviços prestados. Isso permite que os envolvidos identifiquem eventuais problemas e tomem medidas corretivas.

2.1.3 Indicadores-Chave de Desempenho (KPI)

Os Indicadores-Chave de Desempenho são ferramentas para mensurar e avaliar o desempenho de uma organização, equipe ou processo em relação a objetivos e metas estabelecidos. Os KPIs podem ser definidos com base em diversos critérios, tais como desempenho financeiro, satisfação do cliente, eficiência operacional, qualidade do produto ou serviço, entre outros. Os KPIs são ferramentas essenciais para a gestão e tomada de decisões nas organizações. Eles fornecem informações relevantes e atualizadas sobre o desempenho dos processos e atividades em relação aos objetivos e metas estabelecidos, permitindo identificar tendências, identificar oportunidades de melhoria, e tomar decisões embasadas em dados concretos (Parmenter, 2015)

A definição e o acompanhamento adequado dos KPIs podem trazer diversos benefícios para as organizações. Eles permitem identificar áreas de melhoria, detectar problemas e oportunidades, direcionar esforços para os aspectos mais relevantes e mensurar os resultados alcançados (Ittner et al., 2003). Além disso, a utilização de KPIs permite uma visualização rápida, alinhando facilmente pontos entre diferentes níveis e departamentos da organização.

Os KPIs (Key Performance Indicators), ou Indicadores-Chave de Desempenho, desempenham um papel crucial na medição e avaliação do desempenho de uma organização, fornecendo informações valiosas para a tomada de decisões estratégicas (Cokins, 2013). São métricas quantificáveis e objetivas que permitem acompanhar o progresso em direção aos objetivos e metas estabelecidos.

A aplicação de KPIs é amplamente utilizada em diferentes setores e contextos, incluindo o sistema aéreo, onde desempenham um papel fundamental na avaliação da eficiência operacional, segurança e qualidade dos serviços. Esses indicadores podem abranger uma ampla gama de áreas, como pontualidade dos voos, taxa de ocupação de assentos, número de passageiros transportados, tempo médio de espera em filas, entre outros (Lee et al., 2014).

A utilização de KPIs é amplamente adotada em diferentes setores e áreas de negócio, pois permitem uma avaliação objetiva e quantificável do desempenho. Essas métricas podem ser aplicadas em diversas dimensões, como financeira, operacional, de vendas, de atendimento ao cliente, entre outras (Marr, 2016). A escolha dos KPIs adequados depende dos objetivos e das necessidades da organização.

Um exemplo de KPI comumente utilizado na aviação é o OTP (On-Time Performance), que mede a pontualidade dos voos, ou seja, o percentual de voos que decolam e chegam dentro do horário programado. Outro exemplo é o LF (Load Factor), que indica a taxa de ocupação dos assentos em uma aeronave, medindo a eficiência no aproveitamento da capacidade disponível (Singh et al., 2020).

Dentro do sistema aéreo, os KPIs podem abranger diversas áreas-chave, como segurança, pontualidade, eficiência operacional e satisfação do cliente. Esses indicadores fornecem uma visão objetiva e mensurável do desempenho da companhia aérea, aeroporto ou do sistema de controle de tráfego aéreo.

2.2 CIÊNCIA DE DADOS

2.2.1 Ferramentas

A ciência de dados é o estudo dos dados para extrair insights significativos para os negócios. Ela é uma abordagem multidisciplinar que combina princípios e práticas das áreas de matemática, estatística, inteligência artificial e engenharia da computação para analisar grandes quantidades de dados. Uma linguagem muito utilizada para trabalhar com ciência de dados é R, visto que traz bibliotecas e ferramentas complementares para resolução de problemas no ambiente de dados, bem como no contexto da inteligência artificial. Além disso, a linguagem possui facilidade na obtenção das bibliotecas, como também uma documentação robusta que busca simplificar todas as ferramentas contidas na linguagem, tendo também um código de fácil leitura e aprendizado, sendo por esses motivos a escolha dessa linguagem para o presente trabalho.

A biblioteca Tidyverse é uma coleção de pacotes de software criados para trabalhar com dados em R de forma mais intuitiva e eficiente. É uma ferramenta de manipulação de dados de alto nível e sua estrutura de dados chave é chamada de *DataFrame*.

Os *DataFrames* possibilitam armazenar e manipular dados tabulares em linhas e colunas. A estrutura do *DataFrame* pode ser vista como uma planilha oferecem formas versáteis de trabalhar, podendo manipular facilmente qualquer conjunto de dados da maneira desejada, adicionando ou removendo colunas ou linhas. Ele também fornece funções de alto desempenho para tratamento de dados ausentes, agregar, mesclar e unir conjuntos de dados, bem como importar e exportar dados de diferentes formatos. Essa biblioteca foi usada para manusear o conjunto de dados utilizados neste trabalho.

O objetivo de se utilizar esse pacote é tornar a análise de dados em R mais coerente, padronizada e fácil de entender, com ênfase na legibilidade do código e na facilidade de uso. Dentro do Tidyverse existem diversos outros pacotes muito utilizados, como o Ggplot2 para gerar as visualizações, Dplyr para manipulação de dados e Tdyr para organização de dados. O Tidyverse é detalhado no manifesto tidy Trata-se de um documento que formaliza uma série de princípios que norteiam o desenvolvimento do Tidyverse. Como os pacotes do Tidyverse compartilham os mesmos princípios, podem ser utilizados naturalmente em conjunto.

Flexdashboard é um pacote que permite criar dashboards interativos com os dados e visualizações em um único documento. Desenvolvido pela equipe do RStudio, o pacote flexdashboard permite criar dashboards com uma interface de usuário responsiva que se adapta automaticamente ao tamanho da tela do usuário, tornando-a ideal para uso em desktops, tablets e smartphones.

O pacote flexdashboard inclui suporte para uma ampla gama de gráficos e visualizações, incluindo Ggplot2 utilizado para criar os gráficos, bem como suporte para tabelas dinâmicas e outras visualizações. A biblioteca também nos possibilita demonstrar os resultados obtidos através dos modelos de previsão e compará-los.

Para abordagem do aprendizado de máquina, utilizou-se os seguintes pacotes: Vip, Broom, Forcats, Ranger, além do Tidymodels importado junto com o Tidyverse. O pacote Vip (Plotagem de variáveis importantes) é uma biblioteca do R que oferece uma maneira fácil e intuitiva de visualizar a importância das variáveis em modelos de aprendizado de máquina.

O pacote vip é compatível com vários modelos de aprendizado de máquina, incluindo modelos baseados em árvore, modelos lineares e modelos baseados em redes neurais. Com o pacote, o usuário pode identificar as variáveis mais importantes em um modelo e entender como essas variáveis contribuem para a sua precisão.

O Broom é uma biblioteca do R que fornece ferramentas para manipular e visualizar modelos estatísticos de uma maneira mais padronizada e coerente. Permite transformar os resultados de modelos em data frames de uma maneira estruturada e coerente.

Já o Forcats é uma biblioteca que fornece mecanismos para manusear com variáveis categóricas. O pacote inclui funções para ordenar fatores de acordo com valores, níveis ou ordens personalizadas, renomear níveis de fatores, combinar níveis comuns, entre outras funcionalidades. É possível transformar variáveis categóricas em fatores e trabalhar com eles de uma maneira padronizada e que faça sentido para o projeto. Por último, o pacote ranger fornece uma implantação segura de árvores de decisão e florestas aleatórias para problemas de classificação e regressão. O pacote é capaz de lidar com grandes conjuntos de dados e com um grande número de variáveis preditoras. Além disso, ele é desenvolvido para otimizar e aproveitar ao máximo os recursos do processador, com paralelização *multi-thread* e *multi-core*. O pacote ranger oferece uma série de recursos para ajuste e ajuste fino dos modelos, incluindo ajuste de hiper parâmetros, seleção de variáveis, avaliação da importância das variáveis, entre outros.

2.2.2 Análise exploratória dos dados

A análise exploratória de dados é uma abordagem inicial e crucial em todo processo de análise de dados. Ela tem como objetivo identificar padrões, tendências e insights que possam ser úteis na tomada de decisões. Segundo Chatterjee e Hadi (2015), a análise exploratória de dados é uma etapa importante para entender as características dos dados e definir a melhor estratégia para a análise. As etapas mais comuns para a preparação de dados envolvem, segundo Igual et al (2017), as operações de obtenção, verificação, limpeza e construção dos dados

A AED é uma área relativamente nova da estatística, visto que a estatística clássica se concentrava em inferência (BRUCE; BRUCE, 2019). Com a disponibilidade de poder computacional e expressivos softwares de análise de dados, Bruce e Bruce (2019) afirma que:

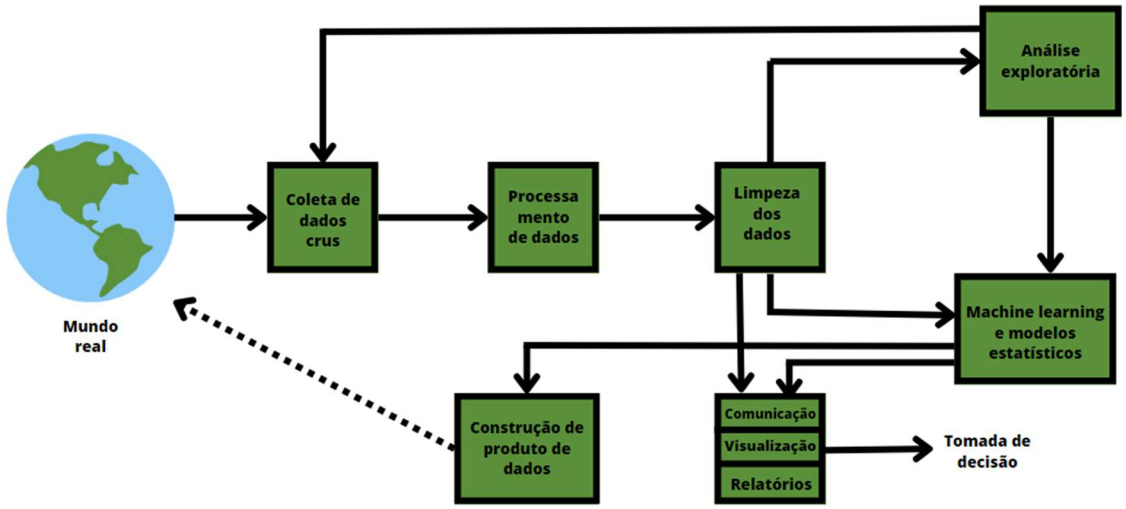
[...] a análise exploratória de dados evoluiu muito além de seu escopo original. As principais características dessa modalidade tem sido o rápido desenvolvimento de novas tecnologias, o acesso a dados maiores e em maior quantidade e o maior uso de análises quantitativas em diversas modalidades. (BRUCE; BRUCE, 2019, p. 2).

Preparar uma base de dados para se adequar a uma tarefa de ciência de dados é a parte mais demorada do processo. É extremamente incomum que os conjuntos de dados sejam disponibilizados na forma exigida pelos algoritmos de ciência de dados. A maioria dos algoritmos de ciência de dados requer que os dados sejam estruturados em um formato tabular com registros nas linhas e atributos nas colunas. Se os dados estiverem em qualquer outro formato, eles precisarão ser transformados aplicando técnicas para condicionar os dados na estrutura necessária.

Após a etapa de coleta, pode-se dizer que a fase de análise é para incluir a sumarização dos dados, usando porções de dados (amostras) para fazer inferências sobre o contexto mais amplo e a visualização dos dados apresentando-os em tabelas, gráficos e até animações (SALTZ; STANTON, 2017).

Uma das principais ferramentas utilizadas na análise exploratória de dados é a visualização gráfica. Segundo Tufte (2001), a visualização gráfica é uma forma eficaz de representar dados complexos e possibilitar uma compreensão rápida dos padrões e tendências presentes nos dados. Segundo Wickham e Grolemund (2017), a visualização gráfica é essencial para a identificação de padrões em grandes conjuntos de dados, pois permite que sejam detectadas relações que não seriam identificadas por meio de técnicas estatísticas tradicionais.

Figura 1 - Processo de análise de dados segundo a ciência de dados



Fonte: Adaptado de Rachel Schutt and Cathy O'Neil, 2014, pelo autor.

A análise de correlação é uma técnica importante na análise exploratória de dados, que permite identificar a relação entre variáveis e entender como elas se relacionam. Segundo Myers et al. (2010), a análise de correlação pode ajudar a identificar padrões e tendências nos dados, além de fornecer informações úteis para a modelagem estatística. No entanto, a correlação não implica necessariamente causalidade, e outras técnicas devem ser utilizadas para confirmar as hipóteses da causalidade.

De acordo com Aggarwal (2013), a análise exploratória de dados deve ser realizada de forma iterativa e interativa, para que sejam identificados novos insights e problemas ao longo do processo de análise de dados. É importante também que a análise exploratória de dados seja realizada por profissionais capacitados e experientes, que tenham conhecimento das técnicas e ferramentas disponíveis e saibam interpretar corretamente os resultados obtidos.

Os dados brutos são um dos maiores desafios da ciência de dados (BRUCE; BRUCE, 2019). Segundo eles, os dados brutos e não estruturados devem ser processados e manipulados, para tomarem uma forma estruturada, sendo dividido em dois tipos.

[...] numérico e categórico. Os dados numéricos aparecem de duas formas: contínua, como velocidade do vento ou tempo de duração, e discreta, como a contagem de ocorrências de um evento. Dados categóricos assumem apenas um conjunto fixo de valores, como um tipo de TV (plasma, LCD, LED, etc.) ou nome de um estado (Alabama, Alasca, etc.). Dados binários são um importante caso especial de dados categóricos que assumem apenas um de dois valores, como 0/1, sim/não ou verdadeiro/falso. Outro tipo de dados categóricos é o dado ordinal, no qual as características são ordenadas. (BRUCE; BRUCE, 2019, p. 3).

2.2.3 Transformação dos dados

Na transformação de dados, de acordo com Mahalle (2021), os dados são convertidos na forma em que o modelo de aprendizado de máquina pode aprender. Há possibilidades de que os dados tenham intervalos diferentes ou que os dados possam estar na forma categórica, então a análise torna-se difícil visto que a maioria dos algoritmos de análise funcionam apenas com dados numéricos, portanto, a transformação de dados torna-se uma etapa necessária.

Sendo assim, espera-se que os atributos de entrada sejam numéricos e normalizados, pois de acordo com Kotu e Deshpande (2018) o algoritmo compara os valores de diferentes atributos e calcula a distância entre os pontos de dados. A normalização evita que um atributo domine os resultados de distância por causa de valores grandes.

2.2.4 Limpeza dos dados

A limpeza de dados é uma etapa crucial em qualquer projeto de análise de dados, uma vez que dados sujos ou inconsistentes podem levar a resultados incorretos e decisões equivocadas. Segundo Redman (2016), dados de baixa qualidade podem levar a erros em processos de tomada de decisão, além de aumentar o tempo necessário para realizar a análise e diminuir a confiança nos resultados obtidos.

Uma das principais fontes de dados sujos são os dados faltantes ou ausentes. De acordo com Graham (2015), dados faltantes podem ocorrer por diversos motivos, como falhas na coleta de dados, problemas técnicos ou erros humanos. Existem diversas abordagens para lidar com dados faltantes, como a exclusão dos registros com dados faltantes ou a imputação de valores para esses dados. No entanto, a escolha da abordagem depende do contexto e dos objetivos do projeto de análise de dados.

2.2.5 Análises gráficas

São as análises em forma de plotagem de gráfico. Essas auxiliam a visualização de informações que, na forma bruta, seriam dificilmente analisadas. Com a análise gráfica, os dados são tratados em forma de gráficos, resumindo o comportamento dos dados a uma forma mais amigável ao ser humano, facilitando o entendimento e a detecção de tendências ou correlações (WICKHAM, GROLEMUND e GARRETT, 2017).

2.2.5.1 Diagrama de barras

Os gráficos de barras são uma das formas mais comuns de visualização de dados e são amplamente utilizados em diversas áreas, como negócios, ciências sociais e saúde. Um dos principais tipos de gráficos de barras é o gráfico de barras simples, que consiste em barras verticais ou horizontais que representam as frequências ou proporções de cada categoria. Segundo Cleveland e McGill (1984), os

gráficos de barras simples são mais efetivos do que os gráficos de pizza ou de áreas, pois permitem uma comparação mais precisa entre as categorias.

Outro tipo comum de gráfico de barras é o gráfico de barras agrupadas ou empilhadas, que é utilizado para comparar as frequências ou proporções de diversas categorias em diferentes grupos. De acordo com Few (2012), os gráficos de barras agrupadas ou empilhadas são úteis para mostrar como as proporções ou frequências mudam em relação a diferentes categorias e grupos.

Por fim, é importante mencionar que os gráficos de barras podem ser utilizados em conjunto com outras formas de visualização de dados, como gráficos de dispersão e histogramas, para fornecer uma visão mais completa e detalhada dos dados. Como destaca Cairo (2016), a escolha do tipo de gráfico mais adequado deve levar em consideração as características dos dados e os objetivos da análise.

2.2.5.2 Gráfico de linhas

Gráficos de linhas são uma forma comum de visualização de dados que permitem a representação da relação entre duas variáveis contínuas. Os gráficos de linhas são ideais para apresentar dados em séries cronológicas ou dados que mudam continuamente no tempo. (CLEVELAND e MCGIL,1984)

Um dos principais benefícios dos gráficos de linhas é a sua capacidade de destacar tendências e padrões nos dados. Como mencionado por Few (2012), o uso de linhas para conectar pontos em um gráfico de linhas permite que as mudanças e flutuações nos dados sejam visualizadas de forma mais clara e intuitiva.

Os gráficos de linhas também são úteis para comparar diferentes séries de dados ao longo do tempo. De acordo com Cairo (2016), o uso de linhas com diferentes cores ou estilos permite que múltiplas séries de dados sejam representadas em um mesmo gráfico, facilitando a comparação entre elas.

Outro aspecto importante a ser considerado na criação de gráficos de linhas é a escolha dos eixos e escalas. Segundo Tufte (2001), a escala dos eixos deve ser escolhida de forma apropriada para o tipo de dados e os objetivos da análise, de modo a maximizar a clareza e a precisão do gráfico.

Por fim, é importante ressaltar que os gráficos de linhas podem ser combinados com outras formas de visualização de dados, como gráficos de barras e áreas, para fornecer uma visão mais completa e detalhada dos dados. Como mencionado por Few (2012), a escolha do tipo de gráfico mais adequado deve levar em consideração as características dos dados e os objetivos da análise.

2.2.5.3 Gráfico de pizza

Os gráficos de pizza são uma das formas mais comuns de representação visual de dados, principalmente quando se trata de comparar a distribuição de categorias ou proporções. Os gráficos de pizza são especialmente úteis para apresentar a

composição de um conjunto de dados em termos de percentuais ou proporções. (CAIRO, 2016)

No entanto, é importante destacar que os gráficos de pizza possuem algumas limitações e críticas em relação à sua eficácia na comunicação de informações. Segundo Tufte (2001), os gráficos de pizza são menos precisos do que outras formas de visualização de dados, como os gráficos de barras, pois a percepção das diferenças entre as áreas das fatias pode ser afetada por diversos fatores, como a inclinação ou a disposição das fatias no gráfico.

Além disso, os gráficos de pizza podem ser menos eficazes na comparação de quantidades, especialmente quando o número de categorias é grande ou as diferenças entre as proporções são sutis. Em muitos casos, os gráficos de barras ou de colunas são mais adequados para a comparação entre categorias ou para representar dados em que as diferenças entre as proporções são pequenas. (FEW, 2012)

No entanto, quando utilizado corretamente e em situações adequadas, os gráficos de pizza podem ser uma forma eficaz de representação visual de dados. Como destacado por Cairo (2016), a escolha do tipo de gráfico mais adequado deve levar em consideração as características dos dados, os objetivos da análise e o público-alvo, para que a comunicação das informações seja clara e efetiva.

2.2.5.4 Gráfico bigote

Gráficos de bigode, também conhecidos como *boxplots*, são uma representação visual de dados estatísticos que fornecem informações importantes sobre a distribuição e a variação dos valores de um conjunto de dados. Esses gráficos consistem em uma caixa retangular que representa o intervalo interquartil (25% a 75%) dos dados, com uma linha vertical que indica a mediana. Além disso, linhas chamadas de "bigodes" se estendem a partir da caixa para representar a amplitude dos dados, excluindo os valores atípicos. (MCGILL, TUKEY e LARSEN, 1978)

Uma das principais vantagens dos gráficos de bigode é sua capacidade de fornecer uma visão resumida da distribuição dos dados, permitindo identificar rapidamente características como a assimetria e a presença de valores atípicos. Ao comparar múltiplos gráficos de bigode, é possível observar diferenças significativas nas medianas e na amplitude dos dados entre os grupos, facilitando a identificação de padrões e tendências. (WICKHAM, 2011) E ainda esses gráficos permitem a detecção de possíveis discrepâncias ou valores extremos que podem influenciar a interpretação dos resultados.

Os gráficos de bigode têm sido amplamente utilizados em diversas áreas, como ciências sociais, economia, medicina e engenharia, devido à sua eficácia na análise de dados. Esses gráficos proporcionam uma representação visual clara e concisa dos principais elementos estatísticos de um conjunto de dados, ajudando na tomada de decisões e no suporte a argumentos baseados em evidências. (SPIEGELHALTER, PEARSON e SHORT, 2011). Além disso, esses gráficos podem ser combinados com outras técnicas estatísticas para uma análise mais aprofundada, como testes de

hipóteses ou regressão. Portanto, os gráficos de bigode são uma ferramenta valiosa para explorar e comunicar informações importantes contidas nos dados estatísticos.

2.3 APRENDIZADO DE MÁQUINA

Aprendizado de máquina (AM), também conhecido como *Machine Learning* (ML), é um ramo da inteligência artificial (IA) que envolve a criação de algoritmos capazes de aprender a partir de dados. A essência do aprendizado de máquina é criar modelos que possam aprender a reconhecer padrões e fazer previsões ou classificações a partir de dados.(ALPAYDIN, 2010)

Os modelos de aprendizado de máquina são desenvolvidos com base em um conjunto de dados de treinamento, que é usado para ajustar os parâmetros do modelo e permitir que ele aprenda a generalizar a partir dos exemplos apresentados. Em seguida, o modelo pode ser aplicado a novos dados para fazer previsões ou classificações. Existem diversas técnicas de aprendizado de máquina, como as redes neurais, árvores de decisão, regressão linear, entre outras. (GOODFELLOW, et al. 2016)

Uma das principais vantagens do uso de aprendizado de máquina é a capacidade de lidar com dados complexos e grandes quantidades de informações. De acordo com Jordan e Mitchell (2015), a disponibilidade de grandes quantidades de dados e o poder de processamento computacional têm impulsionado o avanço da área de aprendizado de máquina nas últimas décadas, permitindo que sejam desenvolvidos modelos mais sofisticados e precisos.

No entanto, é importante destacar que o sucesso do aprendizado de máquina depende não apenas da escolha da técnica correta, mas também da qualidade e representatividade dos dados de treinamento. Como mencionado por Alpaydin (2010), o processo de preparação dos dados é crucial para o sucesso do modelo de aprendizado de máquina, e pode envolver técnicas de limpeza, transformação e seleção de dados.

Além disso, a interpretabilidade dos modelos de aprendizado de máquina é uma questão importante em muitas aplicações práticas. Como destacado por Domingos (2015), os modelos de aprendizado de máquina podem ser vistos como caixas-pretas, em que é difícil compreender como as decisões são tomadas. Por isso, o desenvolvimento de técnicas para explicar os modelos de aprendizado de máquina e torná-los mais transparentes é uma área de pesquisa em expansão.

2.3.1 Métodos de Aprendizagem de Máquina

Em um nível fundamental, a maioria dos problemas de ciência de dados pode ser categorizada em classes ou problemas de predição numérica. Na classificação deve-se tentar usar as informações dos preditores para classificar amostras de dados em duas ou mais classes ou segmentos distintos, tentando prever o valor de uma

variável dependente usando os valores assumidos pelas variáveis independentes (KOTU; DESHPANDE, 2018).

Existem diferentes métodos de *Machine Learning*, que podem ser classificados em três categorias: aprendizado supervisionado, não supervisionado e por reforço.

O aprendizado supervisionado é um dos métodos mais comuns em *Machine Learning*. Ele é usado para criar modelos que aprendem a partir de dados rotulados, ou seja, dados que já possuem uma resposta conhecida. "O aprendizado supervisionado é baseado na ideia de que um modelo pode ser construído para prever um resultado desconhecido, fornecendo um conjunto de entradas e saídas previamente rotuladas".(ALPAYDIN, 2010).

Um modelo de *machine learning* aprende os padrões dos dados e cria matematicamente uma função para gerar previsões, portanto algoritmos a função para algoritmo supervisionados gerarem os valores de resposta é:

$$\hat{y} = f(X) = f(x_1, x_2, \ldots, x_n) \quad (1)$$

onde:

f(x): é a função que o algoritmo irá criar.
x: é a variável independente, ou seja, os atributos.
yˆ: a saída estimada, com base na função

Já o aprendizado não supervisionado é utilizado quando os dados não possuem rótulos ou respostas conhecidas. Ele é usado para descobrir padrões e estruturas ocultas nos dados, sem a necessidade de um modelo prévio. "O aprendizado não supervisionado é uma abordagem exploratória, na qual o objetivo é encontrar uma estrutura útil nos dados sem a necessidade de um conhecimento prévio".(BISHOP, 2006)

O aprendizado por reforço é usado para criar modelos que aprendem a tomar decisões a partir de interações com um ambiente. "o aprendizado por reforço é baseado na ideia de que um agente deve aprender a tomar ações em um ambiente para maximizar uma recompensa numérica, ao longo do tempo". (SUTTON e BARTO, 2018)

Segundo Goodfellow et al. (2016), "a escolha do método de *Machine Learning* apropriado depende da natureza dos dados e do problema em questão, e o sucesso do modelo depende da qualidade dos dados, da escolha do algoritmo e da configuração de seus parâmetros".

2.3.1.1 Regressão Linear

A regressão linear é uma técnica de aprendizado de máquina que busca estabelecer uma relação linear entre uma variável de saída e uma ou mais variáveis

de entrada, com o objetivo de prever ou estimar valores futuros da variável de saída. Regressão linear é um tipo de algoritmo supervisionado.

Segundo Draper e Smith (1981), a regressão linear é uma técnica que envolve a análise de um conjunto de dados para determinar a relação linear entre duas ou mais variáveis. Essa relação é expressa por meio de uma equação linear que pode ser usada para prever o valor da variável de saída com base nos valores das variáveis de entrada.

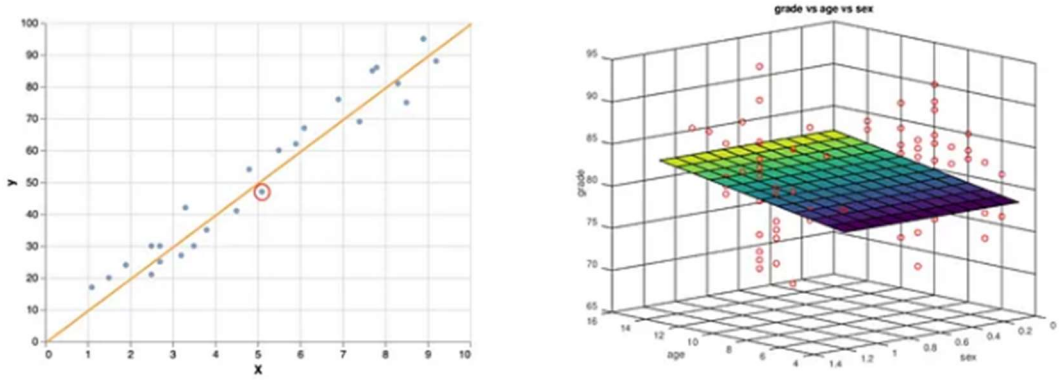
A regressão linear pode ser dividida em duas categorias principais: a regressão linear simples, que envolve apenas uma variável de entrada, e a regressão linear múltipla, que envolve duas ou mais variáveis de entrada.

- Regressão linear simples: refere-se quando temos somente uma variável independente (X) para fazermos a predição.

- Regressão linear múltipla: refere-se a várias variáveis independentes (X)usadas para fazer a predição.

Desta maneira, a forma de representação de cada regressão varia conforme o tipo. A figura 2 mostra a representação gráfica em uma regressão linear simples, onde há uma reta com um plano de 2 dimensões. Já em uma regressão linear múltipla, sua representação é feita em um plano que pode ser nD.

Figura 2 - Representação gráfica da regressão



Fonte: (Raghunath D,2019)

O objetivo da regressão linear é encontrar uma reta que consiga definir bem os dados e minimizar a diferença entre o valor real e a saída calculada pelo modelo. A função que representa bem a regressão linear é dado pela equação 2.

$$f(X) = w_0 + w_1 * x_1 \quad (2)$$

Onde w0 (representa o ponto inicial da reta) e w1 (representa a inclinação da reta, ou seja, o quanto que essa variável cresce conforme o tempo passa) são variáveis que o algoritmo calcula para poder definir a reta, e x1 seria o atributo de entrada que foi dada ao modelo. E com esses valores ele consegue fazer as previsões.

A regressão linear é uma técnica bastante flexível e pode ser aplicada em diversas áreas, como ciências sociais, medicina, economia e engenharia, entre outras. Além disso, a regressão linear é capaz de identificar a importância relativa de cada variável de entrada para a variável de saída, permitindo uma melhor compreensão do problema em questão. (NETER, KUTNER, NACHTSHEIM e WASSERMAN, 1996)

2.3.1.2 Árvore de decisão

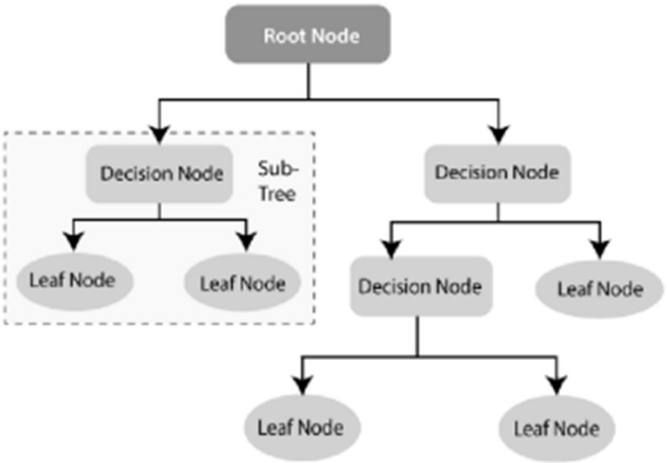
Uma árvore de decisão é uma das técnicas mais simples e intuitivas em aprendizado de máquina, baseada no paradigma dividir para conquistar (IGUAL et al., 2017). A árvore de decisão é um dos métodos mais antigos de aprendizado de máquina e, embora seja simples tanto no treinamento quanto na previsão, é precisa em muitos domínios (ALPAYDIN, 2016).

Essa técnica é usadas para separar um conjunto de dados em classes pertencentes à variável de resposta. Normalmente, a variável de resposta possui duas classes: Sim ou Não (1 ou 0), sendo usadas quando a resposta ou variável de destino é de natureza categórica (KOTU; DESHPANDE, 2018).

Um modelo de árvore de decisão, como mostra a Figura 3, assume a forma de fluxograma de decisão em que um atributo é testado em cada nó (KOTU; DESHPANDE, 2018). A árvore é composta de nós de decisão e folhas, sendo iniciado na raiz, onde cada nó de decisão aplica um teste de divisão à entrada e, dependendo do resultado, pega-se um dos ramos. Quando chega-se a uma folha, a pesquisa para e encontra-se as instâncias de treinamento mais semelhantes (ALPAYDIN, 2016).

No final do caminho da árvore de decisão está um nó folha onde uma previsão é feita sobre a variável de destino com base nas condições estabelecidas pelo caminho de decisão (KOTU; DESHPANDE, 2018).

Figura 3 - Exemplo da estrutura de uma árvore de decisão.



Fonte: (SARKER, 2021)

O principal benefício da Árvore de Decisão é sua capacidade de gerar um modelo preditivo facilmente interpretável e explicável. Isso ocorre porque a estrutura da árvore reflete o processo de tomada de decisão e os critérios de divisão utilizados para classificar os dados. (QUINLAN, 1986)

Além disso, a Árvore de Decisão pode ser utilizada para identificar os atributos mais relevantes para a classificação ou previsão do problema em questão. Segundo Breiman et al. (1984), "as árvores de decisão são capazes de identificar padrões complexos e não lineares nos dados, e podem ser usadas para identificar os atributos mais importantes para a classificação".

Ressalta-se a importância que a construção da Árvore de Decisão pode ser sensível a ruídos e outliers nos dados, e que a escolha dos critérios de divisão pode afetar significativamente o desempenho do modelo.

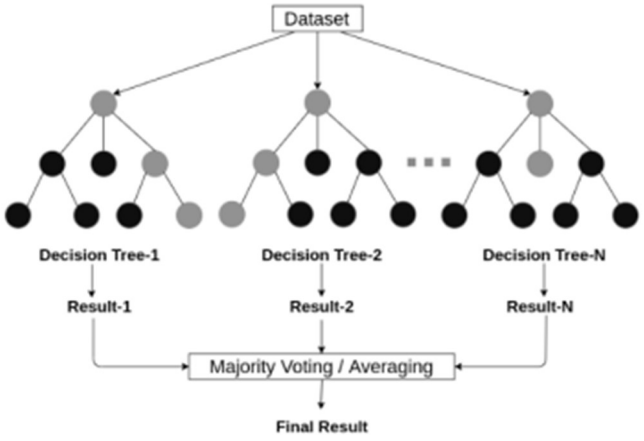
2.3.1.3 Random Forest

Random Forest ou Floresta Aleatória é, segundo Géron (2019), um conjunto de previsores de Árvores de Decisões, ou seja, um conjunto de Árvores de Decisão, Figura 4, fazendo com que, na maioria das vezes, obtenha uma melhor previsão do que com um previsor individual.

Ao decidir sobre a divisão de cada nó em uma árvore de decisão, a floresta aleatória considera apenas um subconjunto aleatório dos atributos no conjunto de treinamento (KOTU; DESHPANDE, 2018). Para realizar a previsão, deve-se, segundo Géron (2019)), obter todas as árvores individuais do ensemble e, então, prever a classe que tem a maioria dos votos.

Para reduzir o erro de generalização, o algoritmo é randomizado em dois níveis, seleção de registros de treinamento e seleção de atributos, no funcionamento interno de cada classificador base (KOTU; DESHPANDE, 2018).

Figura 4 - Exemplo da estrutura de uma Floresta Aleatória.



Fonte: (SARKER, 2021)

Diferentemente do que acontece na criação de uma árvore de decisão simples, ao utilizar o *RandomForest*, o primeiro passo executado pelo algoritmo será selecionar aleatoriamente algumas amostras dos dados de treino, e não a sua totalidade. Nesta etapa é utilizado o *bootstrap*, que é um método de reamostragem onde as amostras selecionadas podem ser repetidas na seleção. Com esta primeira seleção de amostras será construída a primeira árvore de decisão.

No RandomForest a definição desta variável não acontece com base em todas as variáveis disponíveis. O algoritmo irá escolher de maneira aleatória (*random*) duas ou mais variáveis, e então realizar os cálculos com base nas amostras selecionadas, para definir qual dessas variáveis será utilizada no primeiro nó. ara escolha da variável do próximo nó, novamente serão escolhidas duas (ou mais) variáveis, excluindo as já selecionadas anteriormente, e o processo de escolha se repetirá. Desta forma a árvore será construída até o último nó. A quantidade de variáveis a serem escolhidas pode ser definida na criação do modelo.

De acordo com Liaw e Wiener (2002), *Random Forest* é uma técnica que apresenta várias vantagens em relação a outros métodos de aprendizado de máquina, como a capacidade de lidar com dados categóricos e numéricos, a robustez contra ruídos e outliers nos dados, e a possibilidade de identificar os atributos mais importantes para a classificação ou previsão do problema em questão.

2.3.2 Métricas

Uma tarefa de classificação pode ser avaliada de muitas maneiras diferentes para atingir objetivos específicos. Avaliar um classificador, de acordo com Géron (2019), é muitas vezes significativamente mais complicado do que avaliar um regressor. Existem algumas ferramentas principais que estão disponíveis para testar a qualidade de um modelo de classificação, como R-Quadrado, MSE, MAE e MAPE.

2.3.2.1 R-Quadrado

O R-Quadrado é uma medida estatística de quão próximos os dados estão da linha de regressão ajustada. O coeficiente de determinação, é uma métrica comumente utilizada em modelos de regressão para avaliar o quão bem o modelo se ajusta aos dados observados.

De acordo com Hair Jr. et al. (2010), R-Quadrado mede a proporção da variação total na variável de saída que é explicada pelo modelo de regressão. Em outras palavras, R-Quadrado indica o quanto das variações na variável de saída pode ser explicado pelas variáveis de entrada incluídas no modelo.

R-Quadrado pode ser interpretado como uma medida de ajuste do modelo aos dados observados, onde valores próximos a 1 indicam um ajuste muito bom do modelo aos dados, enquanto valores próximos a 0 indicam que o modelo não explica bem a variação na variável de saída. (KVALSETH, 1985)

2.3.2.2 MAE (Erro Médio absoluto)

O erro médio absoluto (MAE — do inglês *Mean Absolut Error*), como demonstrado na equação 3, mede a média da diferença entre o valor real com o predito. Pelo fato de haver valores positivos e negativos, considera-se o valor entre as diferenças em módulo. Segundo Hastie et al. (2009), o MAE é menos sensível a outliers nos dados do que o MSE e pode ser mais adequado para problemas em que erros maiores não são tão críticos quanto erros menores.

$$MAE\,(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (3)$$

O valor de saída da equação tem a mesma escala dos dados utilizados para previsão, facilitando a interpretação. Caso o valor de MAE resultante for igual a 10,01 m, por exemplo, este resultado significa que o modelo pode estar errando em média 10,01 m para mais quanto para menos em relação ao valor correto.

2.3.2.3 MSE (Erro Médio absoluto)

O erro quadrático médio (MSE — do inglês *Mean Squared Error*) é uma métrica que calcula a média de diferença entre o valor predito com o real, como a métrica MAE. Entretanto, ao invés de usar o módulo do resultado entre o valor de y e ŷ, nesta métrica a diferença é elevada ao quadrado. Desta maneira penalizando valores que sejam muito diferentes entre o previsto e o real. Portanto, quanto maior é o valor de MSE, significa que o modelo não performou bem em relação às previsões.

$$MSE\,(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (4)$$

Apesar de sua ideia poderosa, a métrica MSE apresenta um problema de interpretabilidade. Por haver a elevação ao quadrado, a unidade fica distorcida, em outras palavras, se a unidade medida for metros (m), o resultado será em m². Por isso que uma adaptação da MSE é a RMSE que será apresentada abaixo.

2.3.2.4 MSE (Raiz do Erro Quadrático Médio

A raiz do erro quadrático médio (RMSE — do inglês, *Root Mean Squared Error*) é basicamente o mesmo cálculo de MSE, contendo ainda a mesma ideia de penalização entre diferenças grandes do valor previsto e o real. Porém, para lidar com o problema da diferença entre unidades, é aplicada a raiz quadrática como demonstrado na equação 5. Assim a unidade fica na mesma escala que o dado original, resultando em uma melhor interpretabilidade do resultado da métrica.

$$RMSE\,(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

Apesar do valor ter a mesma unidade, ele não costuma se assemelhar ao resultado encontrado de MAE, demonstrando como os outliers podem estar impactando nas previsões do modelo. Mas a sua interpretabilidade pode seguir a mesma lógica, onde o resultado da métrica sendo igual a 80,0 m, significa que o modelo pode estar errando em 80,0 m para mais ou para menos. Por essa razão, esta métrica pode ser uma boa opção quando é preciso ter uma avaliação mais criteriosa sobre as previsões do modelo. O MSE é a medida de desempenho mais comum para modelos de regressão e é amplamente utilizado em áreas como finanças, economia, ciências sociais e engenharia. (HASTIE et al, 2009).

2.3.2.5 MAPE (Erro Percentual Absoluto Médio)

O MAPE (Mean Absolute Percentage Error) calcula a porcentagem média do erro absoluto em relação aos valores reais, fornecendo uma medida de precisão em termos percentuais. Essa métrica é especialmente útil quando se deseja avaliar a acurácia das previsões e comparar o desempenho de diferentes modelos. No entanto, é importante ter em mente que o MAPE possui algumas limitações, como a sensibilidade a valores próximos de zero e a influência de outliers nos dados. Portanto, é recomendado utilizar o MAPE em conjunto com outras métricas e considerar o contexto específico do problema para uma análise mais completa e precisa.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i} \quad (6)$$

2.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foi apresentado o referencial teórico necessário para compreensão do trabalho. No primeiro subcapítulo apresenta-se o tema geral de cada base de dados tal qual a importância de se levantar um estudo baseado em dados para uma análise mais acertada dentro do contexto dos agentes envolvidos. Foi referenciado indicadores importantes tanto para utilização de voos, quanto para reclamações dos consumidores. No segundo subcapítulo, aborda-se sobre a análise exploratória de dados e como ela é uma abordagem crucial que acompanha todo o processo de análise de dados. Apresenta-se os métodos de visualização dos indicadores apresentados no subcapítulo anterior, como gráfico de barras, linhas, indicadores chave de desempenho, entre outros. No subcapítulo são apresentados os modelos de *machine learning* utilizados no presente estudo (regressão linear, árvore de decisão e *random forest),* assim como as métricas para determinar o sucesso ou não do modelo (R-Quadrado, Erro Médio absoluto, Raiz do Erro Quadrático Médio, Erro Percentual Absoluto Médio).

Os próximos capítulos serão demonstrados as aplicações dos conceitos abordados na prática, sendo possível visualizar os indicadores propostos, tal qual os resultados dos modelos de aprendizado de máquina descritos.

3 METODOLOGIA

Neste capítulo foi apresentada a caracterização científica deste trabalho, suas etapas metodológicas e delimitações.

3.1 TIPO DE PESQUISA

A pesquisa pode ser classificada como aplicada, visto que objetiva gerar conhecimentos para a aplicação prática de problemas.

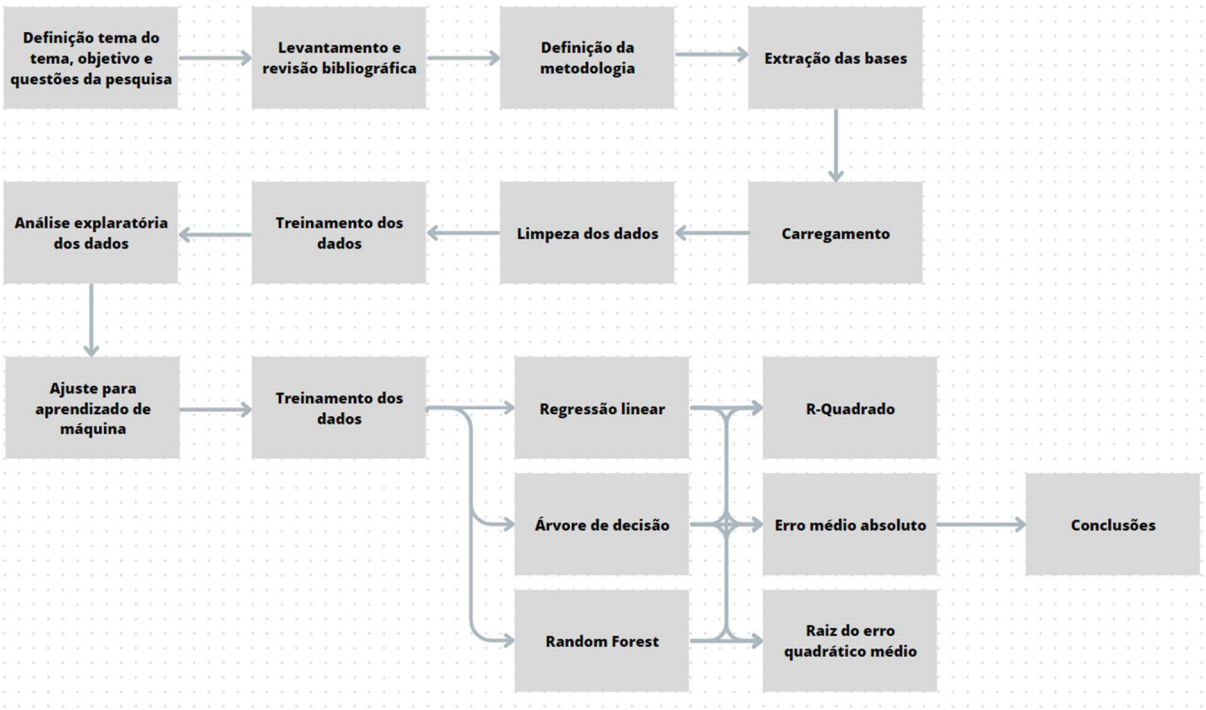
A abordagem pode ser classificada como quantitativa, visto que os indicadores analisados são puramente quantitativos.

Os procedimentos técnicos são classificados como estudo de caso, pois o trabalho foca no estudo de bases de dados referentes ao sistema aéreo da região sul.

3.2 ETAPAS METODOLÓGICAS

As etapas metodológicas do trabalho foram desenvolvidas de acordo com o fluxograma da figura 5. Primeiramente, define-se o tema, a questão da pesquisa e os objetivos, apresentados na introdução. Em seguida, realiza-se uma pesquisa bibliográfica relacionada à análise exploratória e modelos de *machine learning* abordados neste estudo. Com os conhecimentos consolidados da pesquisa bibliográfica é possível extrair, carregar e limpar os dados. A partir dos dados limpos, é possível gerar visualizações como os gráficos e kpi 's. Em seguida na etapa de *machine learning* novos ajustes são feitos nos dados para que se treine e execute os modelos. Com os dados treinados se obtém as métricas que são definidas para o sucesso ou não dos modelos. Por fim, foram extraídas conclusões do estudo de caso, onde as partes envolvidas podem tirar suas conclusões baseadas em dados, para uma decisão com mais acurácia.

Figura 5 – Etapas metodológicas do trabalho



Fonte: elaborado pelo autor.

3.3 DELIMITAÇÕES

O trabalho aborda todos os dados fornecidos no ano de 2021, alguns dados não serão contemplados dado que foram disponibilizados nas bases de 2020 ou 2022.

Não é apresentada uma análise exploratória dos dados nacionais, apenas da região delimitada pelo estudo.

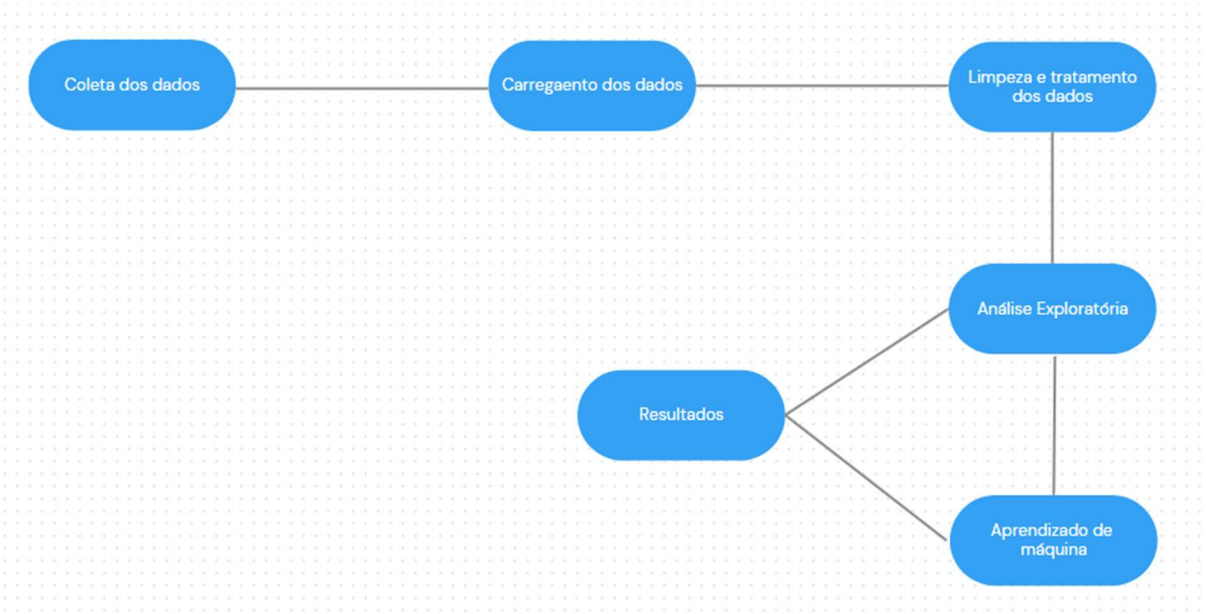
O trabalho consiste em fornecer informações, resultados e métricas para ser possível uma tomada de decisão mais acertada, porém, não tem como objetivo tomar alguma decisão para mudanças na forma de gerir os processos.

Nenhum dado foi aferido, dados nulos são desconsiderados pelos modelos.

4 ESTUDO DE CASO

Esse capítulo aborda as etapas seguidas para o desenvolvimento deste trabalho, apresentando todos os passos para análises e aplicações do processo de desenvolvimento do aprendizado de máquina e criação do modelo, desde a obtenção dos dados até o treinamento do modelo. A metodologia abordada neste trabalho é mostrada na Figura 6.

Figura 6 - Fluxo de projeto



Fonte: Autoria própria

4.1 COLETA DE DADOS

Os conjuntos de dados deste trabalho foram extraídos do site oficial da Agência Nacional de Aviação Civil (ANAC), que disponibiliza o Plano de Dados Aberto. O Plano de Dados Abertos é o documento que informa a priorização dos dados que serão disponibilizados. Um dos critérios para construção do plano é a necessidade manifestada pelo cidadão. Dados de diversas áreas são disponibilizados como:

- Informação de Aeródromos
- Informação de Aeronaves
- Certificação e Outorgas
- Fiscalização
- Gestão Interna
- Operador Aéreo

- Operador Aeroportuário
- Organizações de Formação
- Organizações de Manutenção
- Pessoal da Aviação Civil
- Regulamentação
- Segurança Operacional
- Voos e Operações Aéreas

Para este estudo foram utilizados dois conjuntos de arquivos, referentes ao ano de 2021, disponibilizados na categoria Vôos e Operações Aéreas. As duas áreas temáticas foram Vôo Regular Ativo(VRA) e Dados do consumidor.

O Voo Regular Ativo é uma base de dados composta por informações de voos de empresas de transporte aéreo regular que apresenta alterações de voos (atrasos, antecipações e cancelamentos), bem como horários em que os voos ocorreram. O VRA é formado pela junção das informações, fornecidas pelas empresas de transporte aéreo, relativas aos voos planejados e aos voos realizados.

Como os dados de VRA são disponibilizados mensalmente, foi necessário juntar os meses em um único *Dataframe*. O conjunto de dados final contém 607582 registros com 13 colunas, conforme tabela 1.

Tabela 1 - Detalhamento das colunas Voo Regular Ativo - VRA

Variável	Descrição
Sigla ICAO Empresa Aérea	Sigla/Designador ICAO Empresa Aérea
Número Voo	Numeração do voo
Código	Caractere usado para identificar o Dígito Identificador (DI) para cada etapa de voo
Código Tipo Linha	Caractere usado para identificar o Tipo de Linha realizada para cada etapa de voo
Sigla ICAO Aeroporto Origem	Sigla/Designador ICAO Aeroporto de Origem
Sigla ICAO Aeroporto Destino	Data e horário da partida prevista informada pela empresa aérea, em horário de Brasília
Partida Prevista	Data e horário da partida realizada informada pela empresa aérea, em horário de Brasília
Partida Real	Data e horário da chegada prevista informada pela empresa aérea, em horário de Brasília
Chegada Prevista	Data e horário da chegada realizada, informada pela empresa aérea, em horário de Brasília

Chegada Real	Campo informando a situação do voo: realizado, cancelado ou não informado.

Fonte: Autoria própria

Desde 2019 o site do consumidor tornou-se o sistema eletrônico de atendimento adotado pela ANAC para reclamações de consumidores contra empresas aéreas que operam no Brasil serviços de transporte aéreo regular de passageiros, doméstico e internacional. As empresas aéreas, como previsto no art. 39 da Resolução ANAC no 400, de 13 de dezembro de 2016, devem aderir e estar ativas no Consumidor.gov.br, com cumprimento de prazo, as reclamações registradas na plataforma. A Agência monitora as reclamações registradas na plataforma e avalia, em âmbito coletivo, os serviços prestados pelas empresas do transporte aéreo.

O arquivo conta com 322827 registros brutos, com 38 colunas. Filtrou-se apenas as colunas que continham registros não nulos. Restando assim 29 colunas, conforme tabela 2.

Tabela 2 - Detalhamento das colunas Dados do Consumidor

Variável	**Descrição**
Gestor	Nome da entidade de defesa do consumidor, responsável pela gestão da reclamação do consumidor.
Canal de Origem	Origem do registro no Consumidor.gov.br.
Região	Sigla da região geográfica do consumidor.
UF	Sigla do estado do consumidor.
Cidade	Município do consumidor.
Sexo	Sigla do sexo do consumidor.
Faixa Etária	Faixa etária do consumidor.
Ano Abertura	Ano de abertura da reclamação pelo consumidor.
Mês Abertura	Número do mês de abertura da reclamação pelo consumidor.

Data Abertura	Data de abertura da reclamação pelo consumidor.
Hora Abertura	Hora de abertura da reclamação pelo consumidor.
Data Resposta	Data de resposta da reclamação pela empresa.
Hora Resposta	Hora de resposta da reclamação pela empresa.
Data Finalização	Data de finalização da reclamação.
Hora Finalização	Hora de finalização da reclamação.
Prazo Resposta	Data limite para resposta da empresa. Caso a reclamação tenha sido recusada pela empresa e encaminhada para análise do Gestor, o prazo se altera, considerando o tempo que a reclamação tenha ficado em análise pelo gestor.
Tempo Resposta	Número de dias para a resposta da reclamação, entre a Data de Resposta e a Data de Abertura, desconsiderado o tempo que a reclamação tenha ficado em análise pelo Gestor (se for o caso).
Nome Fantasia	Nome pelo qual a empresa reclamada é conhecida no mercado.
Segmento de Mercado	Principal segmento de mercado da empresa participante.
Área	Área à qual pertence o assunto objeto da reclamação.
Assunto	Assunto objeto da reclamação.
Grupo Problema	Agrupamento do qual faz parte o problema classificado na reclamação.
Problema	Descrição do problema objeto da reclamação.
Como Comprou Contratou	Descrição do meio utilizado para contratação/aquisição do produto ou serviço reclamado.
Procurou Empresa	Sigla da resposta do consumidor à pergunta: “Procurou a empresa para solucionar o problema?".
Respondida	Sigla que indica se a empresa respondeu à reclamação ou não.

Situação	Situação atual da reclamação no sistema.
Avaliação Reclamação	Classificação atribuída pelo consumidor sobre o desfecho da reclamação.
Nota do Consumidor	Número da nota de 1 a 5 atribuída pelo consumidor ao atendimento da empresa.

Fonte: Autoria própria

4.2 CARREGAMENTO E TRATAMENTO DE DADOS

Os dados coletados do mundo real estão, na maior parte dos casos, em um estado bruto e não-processados, de tal forma a ter informações inconsistentes e consequentemente mais difíceis de trabalhar. Valores inseridos de forma errada, valores ausentes, entradas inválidas, colunas com nomes não apropriados para manuseio entre outros. Os dados podem ser confusos se não forem organizados adequadamente, podendo levar a ocasionar resultados com erros, o que implicaria em uma análise equivocada. Sendo assim, os dados precisam passar por um processamento para serem finalizados apropriadamente a fim de serem usados em uma análise exploratória e em um modelo de aprendizado de máquina, fatores esses que o presente trabalho abordará.

4.2.1 Renomeação de colunas

Ao extrair os arquivos da fonte e carregá-los no ambiente de trabalho, é possível que os nomes das colunas contenham acentos, espaços ou outros caracteres indesejados. O fato destes caracteres estarem presentes nos nomes de colunas dificultam trabalhar com o conjunto de arquivos, por isso se faz o processo de renomear as colunas. Em ambas as bases de dados trabalhadas foi necessário realizar este processo, conforme figuras 7.

Figura 7 - Renomeação do nome das colunas.

```
```{r, include=FALSE}
#VRA
#Alterando nome das colunas

voos <- rename(voos, "Empresa"= "ICAOEmpresaAérea")
voos <- rename(voos, "NumeroVoo"= "NúmeroVoo")
voos <- rename(voos, "CodAutorizacao"= "CódigoAutorizaçãoDI")
voos <- rename(voos, "CodTipoLinha"= "CódigoTipoLinha")
voos <- rename(voos, "Arpt_Origem"= "ICAOAeródromooOrigem")
voos <- rename(voos, "SituacaoVoo"= "SituaçãoVoo")
voos <- rename(voos, "CodJustificativa"= "CódigoJustificativa")
voos <- rename(voos, "Arpt_Destino"= "ICAOAeródromooDestino")


```

```{r, include=FALSE}
#Consumidores

data <- rename(data, "Area" = "Área")
data <- rename(data, "Regiao" = "Região")
data <- rename(data, "FaixaEtaria" = "FaixaEtária")
data <- rename(data, "MesAbertura" = "MêsAbertura")
data <- rename(data, "DataFinalizacao" = "DataFinalização")
data <- rename(data, "HoraFinalizacao" = "HoraFinalização")
data <- rename(data, "Situacao" = "Situação")
data <- rename(data, "AvaliacaoReclamacao" = "AvaliaçãoReclamação")

```

Fonte: Autoria própria

Os nomes das colunas devem ser descritivos e refletir claramente o conteúdo ou a natureza dos dados armazenados naquela coluna. Nomes genéricos ou abreviações devem ser evitados, a fim de evitar ambiguidades e garantir a compreensão dos dados por parte dos usuários (Wickham, 2015).

O uso de *underscores* (também conhecido como *snake_case*) para separar palavras nos nomes das colunas é uma prática comum e recomendada. Isso ajuda a tornar os nomes das colunas mais legíveis e evita o uso de espaços em branco, que podem causar problemas em alguns sistemas de gerenciamento de bancos de dados e linguagens de programação (Markham, 2020).

4.2.2 Conversão para datas

A conversão de tipos de colunas em bases de dados é fundamental para garantir a correta interpretação e manipulação dos dados. Os tipos de dados em cada coluna determinam como os dados serão armazenados, processados e apresentados. A escolha adequada dos tipos de dados é crucial para garantir a precisão das análises e consultas realizadas sobre os dados (Müller, 2017). Além disso, a conversão de tipos de colunas é importante para garantir a integridade dos dados, evitando a ocorrência de erros e inconsistências resultantes de conversões automáticas realizadas pelos sistemas de gerenciamento de bancos de dados (Hoffman, 2017).
```

Nas duas bases de dados apresentadas a únicas colunas que foi necessário realizar uma conversão manual foi as colunas referentes a datas. A figura 8 mostra as funções chamadas para realizar a conversão.

Figura 8 - Conversão de colunas para tipo data

```
```{r, include=FALSE}
#VRA
#Convertendo colunas para datas

voos$PartidaPrevista <- as.POSIXct(voos$PartidaPrevista,format = "%d/%m/%Y %H:%M")
voos$PartidaReal <- as.POSIXct(voos$PartidaReal,format = "%d/%m/%Y %H:%M")


voos$ChegadaPrevista <- as.POSIXct(voos$ChegadaPrevista,format = "%d/%m/%Y %H:%M")
voos$ChegadaReal <- as.POSIXct(voos$ChegadaReal,format = "%d/%m/%Y %H:%M")


```

```{r}
#Consumidores

data$DataAbertura <- as.Date( data$DataAbertura, format="%d/%m/%Y" )
data$DataResposta <- as.Date( data$DataResposta, format="%d/%m/%Y" )
data$DataFinalização <- as.Date( data$DataFinalização, format="%d/%m/%Y" )
data$PrazoResposta <- as.Date( data$PrazoResposta, format="%d/%m/%Y" )
```
```

Fonte: Autoria própria

4.2.3 Separando áreas de estudo

Após o tratamento e limpeza dos dados é necessário separar a base de dados em *Dataframes*. Ambas as bases são divididas nos três estados de origem e o aeroporto de sua capital. Para garantir que os dados utilizados contemplam apenas o ano de 2021, mais uma condição é adicionada ao filtro. Conforme figura 9.

Figura 9 - Separando áreas de estudo

```
#VRA
#Separando as regioes e ano de estudo

cur <- voos %>% filter(Arpt_Origem=="SBCT"  | Arpt_Destino == "SBCT")

flo <- voos %>% filter(Arpt_Origem=="SBFL"  | Arpt_Destino == "SBFL")

poa <- voos %>% filter(Arpt_Origem=="SBPA"  | Arpt_Destino == "SBPA")

cur <- cur %>% filter(ano == 2021)

flo <- flo %>% filter(ano == 2021)

poa <- poa %>% filter(ano == 2021)


```
```{r, include=FALSE}
##Consumidores
#Separando as regioes e ano de estudo
cur_cons <- data %>% filter(UF=="PR")

flo_cons <- data %>% filter(UF=="SC")

poa_cons <- data %>% filter(UF=="RS")


flo_cons <- flo_cons %>% filter(ano == 2021)

cur_cons <- cur_cons %>% filter(ano == 2021)

poa_cons <- poa_cons %>% filter(ano == 2021)

```

Fonte: Autoria própria

4.3 ANÁLISE EXPLORATÓRIA

4.3.1 Indicadores-Chave de Desempenho (KPI)

Neste trabalho seis indicadores foram mensurados. Número de voos anuais, média de voos mensais e taxa de cancelamento são indicadores referentes à base de voos. E outros três indicadores abordando a base de consumidores. Taxa de resposta, média de avaliação e média do tempo de resposta. Disponíveis no apêndice A.

4.3.2 Cancelamentos mensais

Mensurar os cancelamentos de voos pode fornecer informações valiosas para compreender as causas e impactos dessas ocorrências na indústria da aviação.Ao realizar a análise exploratória de dados de cancelamentos de voos, é possível utilizar diversas técnicas estatísticas e de visualização de dados para examinar as variáveis relevantes (McKinney, 2017). Entre as principais métricas a serem consideradas estão: a quantidade de voos cancelados por período, como diário, mensal e anual; as razões para o cancelamento, como causas climáticas, mecânicas, operacionais, entre outras (Han et al., 2011). Neste trabalho foram analisados cancelamentos por mês. Disponíveis no apêndice B.

4.3.3 Dados temporais

A partir do *dataset* de voos regulares, é possível fazer análise exploratórias de acordo com dados temporais. A análise exploratória de dados temporais é uma abordagem essencial na análise de conjuntos de dados que variam ao longo do tempo. É uma disciplina especializada que envolve a compreensão de padrões e tendências em séries temporais, que são sequências de observações registradas em intervalos regulares de tempo. Utiliza-se gráficos de séries temporais para identificar padrões sazonais, tendências e possíveis anomalias nos dados ao longo do tempo. Essas técnicas podem ser aplicadas em conjunto para obter insights valiosos sobre a dinâmica temporal dos dados (Brockwell & Davis, 2016; Chatfield, 2004).

Os gráficos obtidos trazem informações da distribuição de voos ao longo do ano, tanto para partidas quanto para chegadas. A linha Azul representa a linha azul representa a função *geom_smooth* que tem o objetivo de auxiliar o olho humano a identificar padrões nos dados, mesmo quando há sobreposição de pontos no gráfico. A sobreposição de pontos pode dificultar a identificação de padrões em um gráfico, e essas funções ajudam a suavizar os dados para tornar os padrões mais evidentes. O outro gráfico representa os horários com mais voos durante o ano, dividido em chegadas e partidas. Disponíveis no apêndice C.

4.3.4 Atividades por companhia aérea

Fator importante também que pode trazer *insights* para algum tomador de decisão são as companhias aéreas que mais fazem trajetos partindo ou chegando no aeroporto de estudo. Através dessa análise, é possível obter informações valiosas sobre a participação de mercado das companhias aéreas em um aeroporto específico, identificar tendências e padrões de fluxo de tráfego ao longo do tempo, monitorar e gerenciar a performance operacional, e embasar decisões estratégicas.

A análise dos dados de chegadas e partidas das companhias aéreas pode ser útil para compreender a concorrência entre as empresas no mercado da aviação. Estudos como o de Neumark et al. (2017) têm explorado a relação entre a concorrência no setor aéreo e os preços das passagens, evidenciando como a participação de mercado das companhias aéreas em um aeroporto pode influenciar na formação de preços e nas estratégias de negócio. Disponíveis no apêndice D.

4.3.5 Aeroportos com mais movimentações

A análise da movimentação de partidas e chegadas de voos em um aeroporto é fundamental para a compreensão do desempenho operacional, demanda de passageiros e carga, planejamento operacional e análise de concorrência. No planejamento operacional, a movimentação de partidas e chegadas de voos é crucial para a alocação eficiente de recursos. Autoridades aeroportuárias e companhias aéreas utilizam esses dados para planejar a quantidade de funcionários necessários em áreas como atendimento ao cliente, segurança e manutenção. Além disso, a movimentação de voos também é utilizada para a alocação estratégica de portões de

embarque, posições de estacionamento de aeronaves e esteiras de bagagem, otimizando assim o fluxo de passageiros e aeronaves no aeroporto. Disponíveis no apêndice E.

4.3.6 Visualizações sociodemográficas

Para realizar uma análise de dados eficiente e embasar decisões informadas, é crucial considerar os dados socioeconômicos e demográficos, que fornecem informações valiosas sobre a população em estudo. De acordo com Babbie (2016) e Hair et al. (2014), esses dados são essenciais para contextualizar a análise, identificar tendências e padrões, segmentar o mercado, avaliar o impacto de políticas e programas, embasar a tomada de decisões e auxiliar no planejamento e desenvolvimento de políticas públicas e projetos.

Todas as visualizações com dados de sexo, idade e cidade foram geradas a partir do *dataset* de abertura de chamados do consumidor. A análise de dados desagregados por sexo é uma abordagem fundamental para entender as disparidades de gênero em diferentes áreas da sociedade. Essa abordagem permite identificar diferenças significativas entre homens e mulheres, e é crucial para promover a equidade de gênero e tomar decisões informadas. Por isso uma série de visualizações foram geradas segmentando os sexos. Uma visualização simples e muito importante obtida é a porcentagem de reclamações feitas por cada sexo. Foi possível segmentar também o tempo de espera, a fim de entender se há alguma diferença notável. Além das notas dadas por ambos os sexos.

Observa-se também as cidades com maior número de avaliações das chamadas segmentando a parcela de cada nota atribuída por aquela cidade. Além do grupo etário subdividido em quais problemas cada reclamação apresentou. Disponíveis no apêndice F.

4.3.7 Visualizações de aberturas de reclamações

Os gráficos de aberturas de reclamações trazem aspectos interessantes que podem ser abordados e analisados. No primeiro gráfico é possível acompanhar o número total de reclamações em um ano, separado por mês. Além disso, o gráfico permite identificar quais os grupos de problemas mais aparecem e suas variações.

Na segunda visualização é possível ver quais companhias aéreas têm mais reclamações. Este gráfico é interessante ao ser comparado com o gráfico de vôos por companhias. O dashboard tem como intuito gerar visualizações rápidas e comparativas e já supor e descartar possíveis hipóteses para analisar os dados. Disponíveis no apêndice G.

4.3.8 Matriz de resultados

A matriz de resultado, tabela 3, para este estudo foi organizada em linhas e colunas, onde as linhas representam os diferentes valores levantados neste estudo. As colunas se referem aos três estados abordados. Os resultados são comparados por estados atribuindo a cor verde para o melhor/maior, amarelo para segundo e vermelho para o último.

Tabela 3 - Matriz de Resultados

Resultados	Santa Catarina	Paraná	Rio Grande do Sul
Número de vôos anuais	17580	27610	37808
Média de vôos mensais	1465	2301	3151
Taxa de Cancelamento	2,23%	3,34%	3,81%
Taxa de Resposta	99,41%	99,67%	99,03%
Média de avaliação	3,11	3,19	3,1
Média do tempo de resposta	5,61	5,52	5,63
Mês com mais cancelamentos/ Quantidade	Dezembro / 57	Novembro / 138	Outubro / 271
Horário mais movimentado - chegadas/Quantidade	07h / 1807	07h / 2151	11h / 2781
Horário mais movimentado - partidas/Quantidade	05h / 1531	19h / 2112	17h / 2968
Companhia com mais partidas/ Quantidade	Gol / 2860	Azul / 7004	Azul / 8936
Companhia com mais chegadas/ Quantidade	Azul / 2868	Azul / 7000	Azul / 8956
Aeroporto com mais partidas/ Quantidade	Guarulhos / 2850	Guarulhos / 3532	Guarulhos / 5294
Aeroporto com mais chegadas/ Quantidade	Guarulhos / 2786	Guarulhos / 3598	Guarulhos / 5238
% de homens	58,07%	55,90%	54,00%
% de mulheres	41,93%	44,10%	46,00%
Tempo de resposta mais comum homens	1 minuto / 1554	1 minuto / 2280	1 minuto / 1155
Tempo de resposta mais comum mulheres	1 minuto / 1290	1 minuto / 1710	1 minuto / 1149
Cidade com mais avaliações/ Quantidade	Florianópolis / 2679	Curitiba / 6840	Porto Alegre / 3288
Empresa com mais aberturas / Quantidade	Gol / 6357	Gol / 8973	Gol / 5367

Mês com mais aberturas de chamado/ Grupo problema /Quantidade Total	Novembro / Cobrança Contestação / 1932	Dezembro / Cobrança Contestação / 27899	Dezembro / Cobrança Contestação / 1746

Fonte: Autoria própria

A matriz possibilita identificar os atributos relevantes do estudo e que, portanto, podem ser abordados por um tomador de decisão gerando maior competitividade na prestação de serviços.

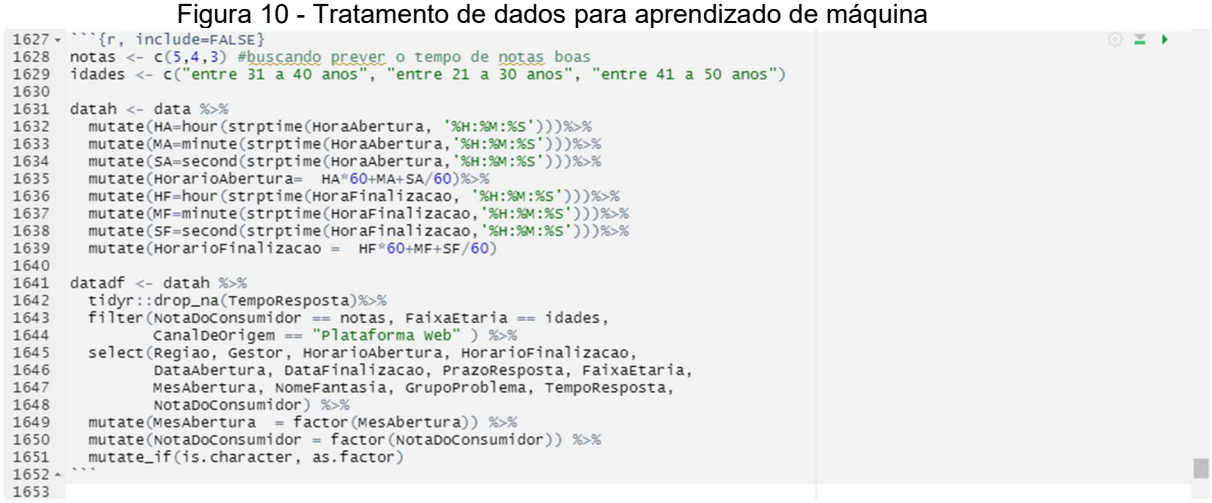
4.4 PREPARAÇÃO DOS DADOS PARA APRENDIZADO DE MÁQUINA

4.4.1 Filtros e transformação em fatores

A escolha adequada de dados é um fator crítico para o sucesso de um modelo de aprendizado de máquina (Witten, Frank, & Hall, 2016). Dados de treinamento de qualidade e representativos são fundamentais para garantir que o modelo possa aprender padrões relevantes e generalizá-los para novos dados durante a fase de predição. Além disso, a qualidade dos dados de treinamento influencia diretamente na capacidade do modelo de lidar com questões de enviesamento e discriminação, garantindo a equidade e justiça na tomada de decisões automatizada (Mitchell et al., 2019). Portanto, é crucial selecionar dados de treinamento adequados que sejam representativos do domínio de interesse e estejam de acordo com os objetivos do modelo.

Conforme a figura 10, as notas desejadas são alocadas a variável “notas”, enquanto na variável “idades” foi atribuído três grupos de idade pois são os intervalos de idade que mais aparecem no *dataset*. Uma das razões para desconsiderar resultados menos frequentes é a redução de ruído nos dados de treinamento. Dados não frequentes podem introduzir instabilidades e estimativas estatísticas menos confiáveis, uma vez que sua presença limitada pode resultar em amostras com variabilidade alta. Essa variabilidade pode interferir na capacidade do modelo de identificar padrões relevantes e generalizar para novos dados. Portanto, desconsiderar dados raros pode ser uma estratégia para reduzir o ruído e melhorar a qualidade dos dados utilizados no treinamento do modelo (Witten, Frank & Hall, 2016).

Figura 10 - Tratamento de dados para aprendizado de máquina



```r
```{r, include=FALSE}
notas <- c(5,4,3) #buscando prever o tempo de notas boas
idades <- c("entre 31 a 40 anos", "entre 21 a 30 anos", "entre 41 a 50 anos")

datah <- data %>%
  mutate(HA=hour(strptime(HoraAbertura, '%H:%M:%S')))%>%
  mutate(MA=minute(strptime(HoraAbertura,'%H:%M:%S')))%>%
  mutate(SA=second(strptime(HoraAbertura,'%H:%M:%S')))%>%
  mutate(HorarioAbertura=  HA*60+MA+SA/60)%>%
  mutate(HF=hour(strptime(HoraFinalizacao, '%H:%M:%S')))%>%
  mutate(MF=minute(strptime(HoraFinalizacao,'%H:%M:%S')))%>%
  mutate(SF=second(strptime(HoraFinalizacao,'%H:%M:%S')))%>%
  mutate(HorarioFinalizacao =  HF*60+MF+SF/60)

datadf <- datah %>%
  tidyr::drop_na(TempoResposta)%>%
  filter(NotaDoConsumidor == notas, FaixaEtaria == idades,
         CanalDeOrigem == "Plataforma Web" ) %>%
  select(Regiao, Gestor, HorarioAbertura, HorarioFinalizacao,
         DataAbertura, DataFinalizacao, PrazoResposta, FaixaEtaria,
         MesAbertura, NomeFantasia, GrupoProblema, TempoResposta,
         NotaDoConsumidor) %>%
  mutate(MesAbertura  = factor(MesAbertura)) %>%
  mutate(NotaDoConsumidor = factor(NotaDoConsumidor)) %>%
  mutate_if(is.character, as.factor)
```
```

Fonte: Autoria própria

Separa-se o tempo em horas, minutos e segundos para melhor alimentação de dados, seleciona-se as colunas desejadas que farão parte do modelo, além da transformação em fatores de colunas tipo caracter em fatores.

A transformação de dados em fatores é uma etapa fundamental na preparação de dados para modelos de *machine learning*. Essa prática envolve a conversão de variáveis em formatos que podem ser mais facilmente compreendidos e processados pelos algoritmos de aprendizado de máquina. A transformação de dados em fatores pode trazer benefícios significativos para o desenvolvimento de modelos precisos e eficazes.

Uma das principais razões para transformar os dados em fatores é a padronização e normalização dos dados. Dados em formatos diferentes, com unidades diferentes ou escalas diferentes, podem prejudicar a capacidade do modelo de aprender padrões relevantes e fazer comparações adequadas entre as variáveis. A transformação dos dados em fatores permite que as variáveis sejam representadas em uma mesma escala ou formato, tornando mais fácil a comparação e o processamento pelos algoritmos de *machine learning*. Isso pode melhorar a estabilidade e a confiabilidade do modelo, evitando distorções causadas por diferenças de escala ou unidade (Hastie, Tibshirani & Friedman, 2009).

4.4.2 Divisão de dados

A divisão do *dataset* em treinamento e teste é uma prática essencial quando se trabalha com aprendizado de máquina, serve para avaliar a performance dos modelos e garantir sua capacidade de generalização, conforme figura 11. O processo envolve separar os dados disponíveis em duas partes distintas: o conjunto de treinamento, que é usado para treinar o modelo, e o conjunto de teste, que é reservado para avaliar o desempenho do modelo em dados não vistos. Essa divisão é crucial por várias razões. A divisão ocorre através da coluna TempoResposta, pois é a variável de predição desejada.

Um modelo pode se ajustar muito bem aos dados de treinamento, mas pode não ser capaz de generalizar corretamente para novos dados. Ao usar um conjunto de teste independente, é possível obter uma estimativa imparcial da capacidade de generalização do modelo (Hastie, Tibshirani & Friedman, 2009). A divisão do dataset em treinamento e teste ajuda a prevenir o *overfitting*, que ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, resultando em uma performance ruim em dados de teste. A avaliação do modelo em um conjunto de teste separado permite identificar se o modelo está sofrendo de *overfitting*, indicando a necessidade de ajustes ou regularização para melhorar sua capacidade de generalização (Bishop, 2006).

A divisão do *dataset* em treinamento e teste deve ser feita de forma aleatória e estratificada, para garantir que os conjuntos de treinamento e teste sejam representativos dos dados originais. Além disso, é comum usar técnicas como a validação cruzada, que envolve a divisão do dataset em múltiplos conjuntos de treinamento e teste, para obter uma avaliação mais robusta do modelo (Raschka & Mirjalili, 2020; Géron, 2019).

Figura 11 - Divisão em treinamento e teste

```
1658 ```{r, include=FALSE}
1659 #Dividindo os dados em treinamento e teste
1660 set.seed(1234)
1661 datadf_split <- initial_split(datadf, strata=TempoResposta)
1662
1663 datadf_train <- training(datadf_split)
1664 datadf_test <- testing(datadf_split)
1665
1666
1667
1668 ```
1669
```

Fonte: Autoria própria

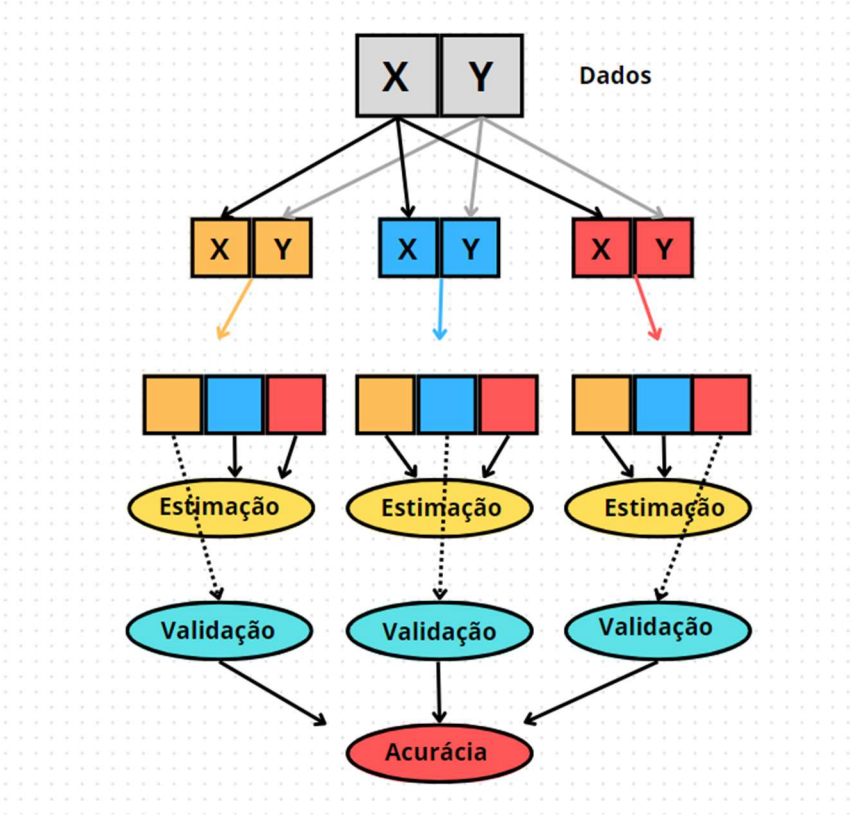
4.4.3 Validação cruzada

A validação cruzada é uma técnica utilizada em aprendizado de máquina para avaliar o desempenho de modelos de forma mais robusta e confiável, conseguindo superar limitações de apenas realizar a divisão tradicional do *dataset* em treinamento e teste. Essa abordagem envolve a divisão dos dados em múltiplos conjuntos de treinamento e teste, permitindo que o modelo seja treinado e avaliado em diferentes combinações de dados, proporcionando uma estimativa mais precisa de sua performance em dados não vistos (Hastie, Tibshirani & Friedman, 2009; Géron, 2019). Com a validação cruzada, todos os dados são utilizados tanto para treinamento quanto para teste em diferentes combinações, maximizando o uso das informações disponíveis e fornecendo uma avaliação mais confiável do desempenho do modelo (Raschka & Mirjalili, 2020).

Um ganho da utilização de validação cruzada é a redução da sensibilidade à aleatoriedade da divisão dos dados. A divisão tradicional em treinamento e teste pode ser influenciada pela aleatoriedade dos dados, resultando em diferentes desempenhos do modelo dependendo da divisão específica. Uma das abordagens

mais comuns de validação cruzada é a *k-fold cross-validation*, em que o dataset é dividido em k conjuntos de tamanho aproximadamente igual. O modelo é treinado k vezes, cada vez usando k-1 conjuntos como treinamento e o conjunto restante como teste. A performance do modelo é avaliada pela média dos resultados obtidos nas k iterações. Essa abordagem é amplamente utilizada e aceita na comunidade de machine learning devido à sua simplicidade e eficácia na avaliação de modelos (Hastie, Tibshirani & Friedman, 2009).

Figura 12 - Exemplo do esquema de particionamento e execução do método *k-fold* com k = 3



Fonte: Autoria própria

Em R para realizar validação cruzada k-fold utiliza-se a função "*vfold_cv*", conforme a figura 13, onde v é número de validações cruzadas realizadas. Neste estudo foram feitas um total de dez validações.

Figura 13 - Validação cruzada

```
1670 ```{r}
1671 #validação cruzada
1672 datadf_fold <- vfold_cv(datadf_train, v=10)
1673 ```
```

Fonte: Autoria própria

4.4.4 Treinamento

Assim que todo o *dataset* está apropriado para se manusear, é possível realizar o treinamento dos modelos. O processo de treinamento de um modelo de regressão envolve a busca pela melhor função matemática que se ajusta aos dados de treinamento disponíveis (Bishop, C. M., 2006) Os três modelos que foram utilizados para treinar as previsões, conforme referencial teórico são: Regressão linear, árvore de decisão e *random forest*. Conforme figura 14 os modelos foram treinados.

Figura 14 - Treinando modelos

```
1681 ```{r, include=FALSE}
1682 lm_spec <- linear_reg() %>%
1683   set_engine("lm")
1684
1685 tree_spec <- decision_tree() %>%
1686   set_engine("rpart") %>%
1687   set_mode("regression")
1688
1689 rf_spec <- rand_forest(trees=1000) %>%
1690   set_engine("ranger") %>%
1691   set_mode("regression")
1692 ```
1693
```

Fonte: Autoria própria

Na primeira função é atribuído a variável “lm_spec” o treinamento do modelo de regressão linear com “lm” na função *set_engine* que usa o método dos mínimos quadrados ordinários para ajustar modelos com resultados numéricos.

Na segunda função atribui-se “tree_spec” o treinamento da árvore de decisão. *Set_mode* indica que se trate de uma regressão e *set_engine* junto de “rpart” indica um ajuste do modelo como um conjunto de declarações de 'se/então' que cria uma estrutura baseada em árvore.

A última variável atribuída é “rf_spec” que se trata do treinamento de *random forest*. Onde primeiro é definido o número de árvores, em seguida, utiliza-se a biblioteca ranger para gerar e manipular dados de *random forest*. A função ranger ajusta um modelo que cria um grande número de árvores de decisão, cada uma independente das outras. A predição final utiliza todas as predições das árvores individuais e as combina.

4.5 APRENDIZADO DE MÁQUINA

Conforme a preparação e o treinamento feito anteriormente é possível coletar e mostrar os resultados das métricas delimitadas pelo trabalho no capítulo 2. Conforme figura 15 é possível verificar a adição do modelo de regressão linear junto da coleta de métricas.

Figura 15 - Coleta de resultados

```
```{r, include=FALSE}
#coletando resultados

lm_rs <- datadf_wf %>%
  add_model(lm_spec) %>%
  fit_resamples(resamples=datadf_fold,metrics=metric_set(rmse, rsq, mae),
                control=control_resamples(save_pred=TRUE))


collect_metrics(lm_rs)


```
```

Fonte: Autoria própria

Na figura 16, vê-se o resultado apresentado pelo modelo de regressão linear. Métricas foram calculadas utilizando a média das dez validações.

Figura 16 - Resultados de regressão linear

```
```{r}
print(collect_metrics(lm_rs))
```
```

A tibble: 3 × 6

.metric	.estimator	mean	n	std_err	.config
<chr>	<chr>	<dbl>	<int>	<dbl>	<chr>
mae	standard	0.6135097	10	0.005244105	Preprocessor1_Model1
rmse	standard	0.7741223	10	0.005494377	Preprocessor1_Model1
rsq	standard	0.4019200	10	0.007265244	Preprocessor1_Model1

3 rows

Fonte: Autoria própria

Conforme figura 17 é possível verificar a coleta das métricas para o modelo de árvore de decisão.

Figura 17 - Coleta de métrica de árvore de decisão

```
```{r, include=FALSE}
#coletando resultados

tree_rs <- datadf_wf %>%
  add_model(tree_spec) %>%
  fit_resamples(resamples=datadf_fold,
                metrics=metric_set(rmse, rsq, mae),
                control=control_resamples(save_pred=TRUE))

collect_metrics(tree_rs)
```
```

Fonte: Autoria própria

Na figura 18, mostra-se o resultado coletado anteriormente pelo modelo de árvore de decisão.

Figura 18 - Resultado de árvore de decisão

```
1725
1726 ```{r}
1727 print(collect_metrics(tree_rs))
1728 ```
```

A tibble: 3 × 6

.metric <chr>	.estimator <chr>	mean <dbl>	n <int>	std_err <dbl>	.config <chr>
mae	standard	0.6009704	10	0.004367125	Preprocessor1_Model1
rmse	standard	0.7734378	10	0.005113049	Preprocessor1_Model1
rsq	standard	0.4018639	10	0.007024661	Preprocessor1_Model1

3 rows

Fonte: Autoria própria

O processo análogo foi realizado no modelo de *random forest*. A coleta e os resultados estão sendo mostrados nas figuras 19 e 20.

Figura 19 - Coletando resultados de *random forest*

```
1731 ```{r, include=FALSE}
1732 #coletando resultados
1733
1734 rf_rs <- datadf_wf %>%
1735   add_model(rf_spec) %>%
1736   fit_resamples(resamples=datadf_fold,
1737                 metrics=metric_set(rmse, rsq, mae),
1738                 control=control_resamples(save_pred=TRUE))
1739
1740 collect_metrics(rf_rs)
1741 ```
1742
```

Fonte: Autoria própria

Figura 20 - Resultados de *random forest*

```
1743 ```{r}
1744 print(collect_metrics(rf_rs))
1745 ```
```

A tibble: 3 × 6

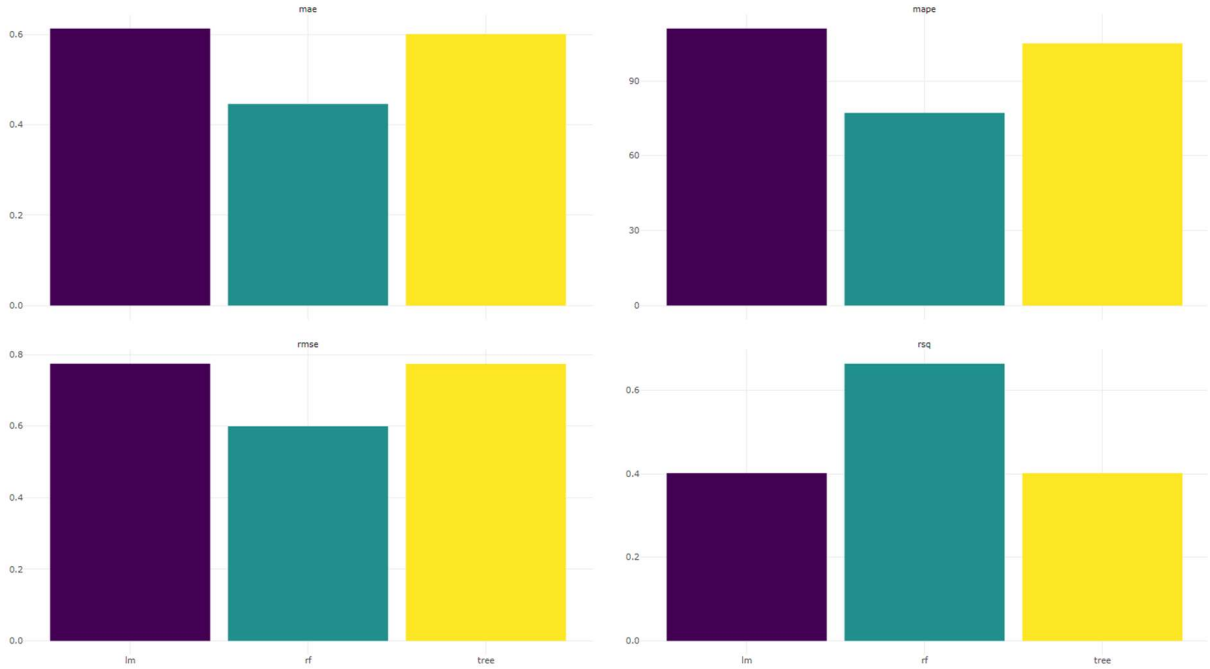
.metric <chr>	.estimator <chr>	mean <dbl>	n <int>	std_err <dbl>	.config <chr>
mae	standard	0.4459491	10	0.004599323	Preprocessor1_Model1
rmse	standard	0.5987752	10	0.006639510	Preprocessor1_Model1
rsq	standard	0.6648287	10	0.008797727	Preprocessor1_Model1

3 rows

Fonte: Autoria própria

Com os resultados dos modelos utilizados para treinamento é necessário fazer uma comparação dos resultados a fim de determinar qual modelo obteve os melhores resultados, segundo metodologia apresentada no capítulo 2.

Figura 21 - Comparação dos resultados



Fonte: Autoria própria

Conforme capítulo 2 quanto mais próximo de 1 o R-Quadrado e quanto mais baixo a raiz do erro quadrático médio e erro médio absoluto melhor o modelo. Por isso o método de *random forest* foi escolhido para gerar as previsões com os dados que não foram treinados. Para isso, a função *last_fit* realiza um último ajuste usando os dados de treino (dentro da especificação *rf_spec*) e automaticamente uma última estimação, usando os dados de teste que a função reconhece os dados de treino e teste no objeto dividido anteriormente. Conforme figura 22.

Figura 22 - Coletando métricas finais

```
```{r}
#Resultado Final

library(kableExtra)



modelo_final <- datadf_wf %>%
  add_model(rf_spec) %>%
  last_fit(datadf_split)

cm <- collect_metrics(modelo_final,
                      metrics = metric_set(mae, rsq, rmse))
cm %>%
  kbl() %>%
  kable_material_dark()

```
```

Fonte: Autoria própria

Com este último ajuste gerou-se os resultados finais do modelo de aprendizado de máquina. Na figura 23 é mostrado os resultados.

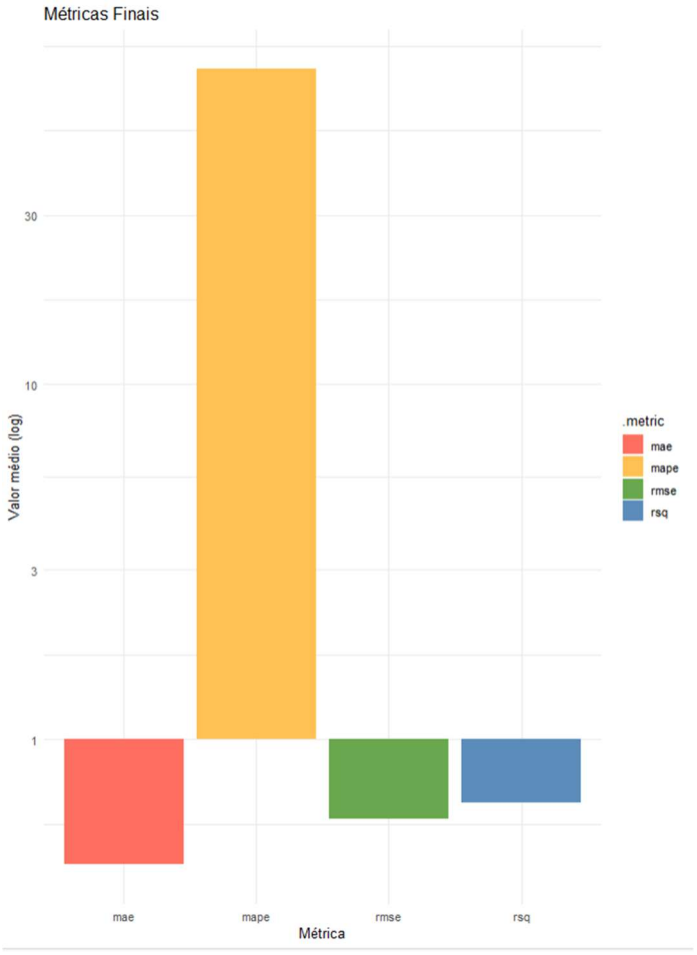
Figura 23 - Resultados finais

.metric	.estimator	.estimate	.config
rmse	standard	0.6043263	Preprocessor1_Model1
rsq	standard	0.6565988	Preprocessor1_Model1

Fonte: Autoria própria

É possível observar os resultados gráficos finais dos modelos na figura 24. Nota-se que pelo fato de o MAPE ser um número absoluto, ou seja, não está entre 0 e 1 a escala logarítmica transforma o R^2. MAE, e RMSE em valores negativos.

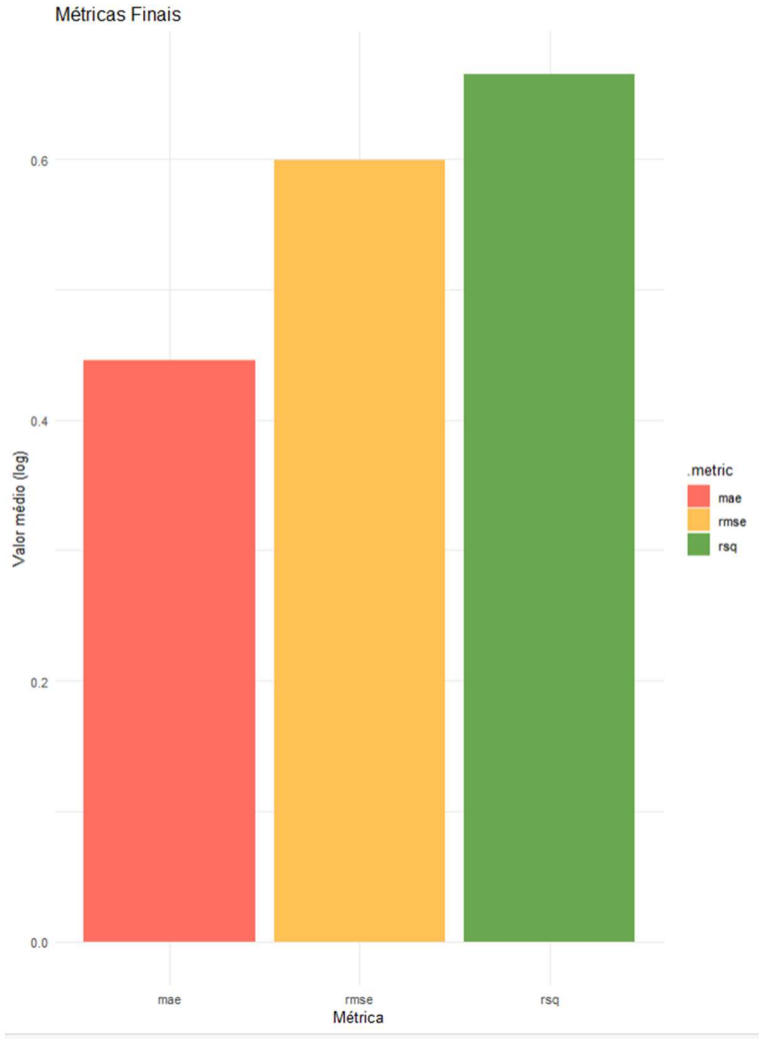
Figura 24 - Resultados finais



Fonte: Autoria própria

Na figura 25 é possível observar os resultados sem o MAPE para melhor visualização.

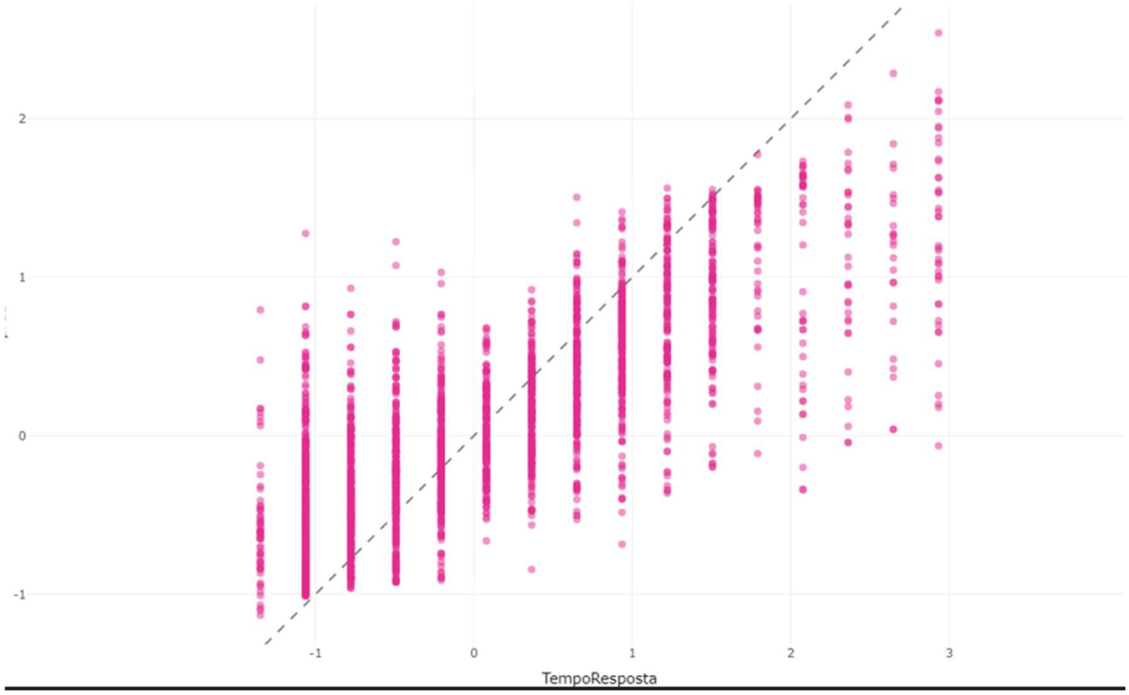
Figura 25 - Resultados finais sem MAPE



Fonte: Autoria própria

A partir dos resultados obtidos pelo modelo de regressão *random forest*, foi possível levantar dados das predições. No gráfico da figura 26, mostra-se o confronto entre as previsões e os tempos de resposta obtidos.

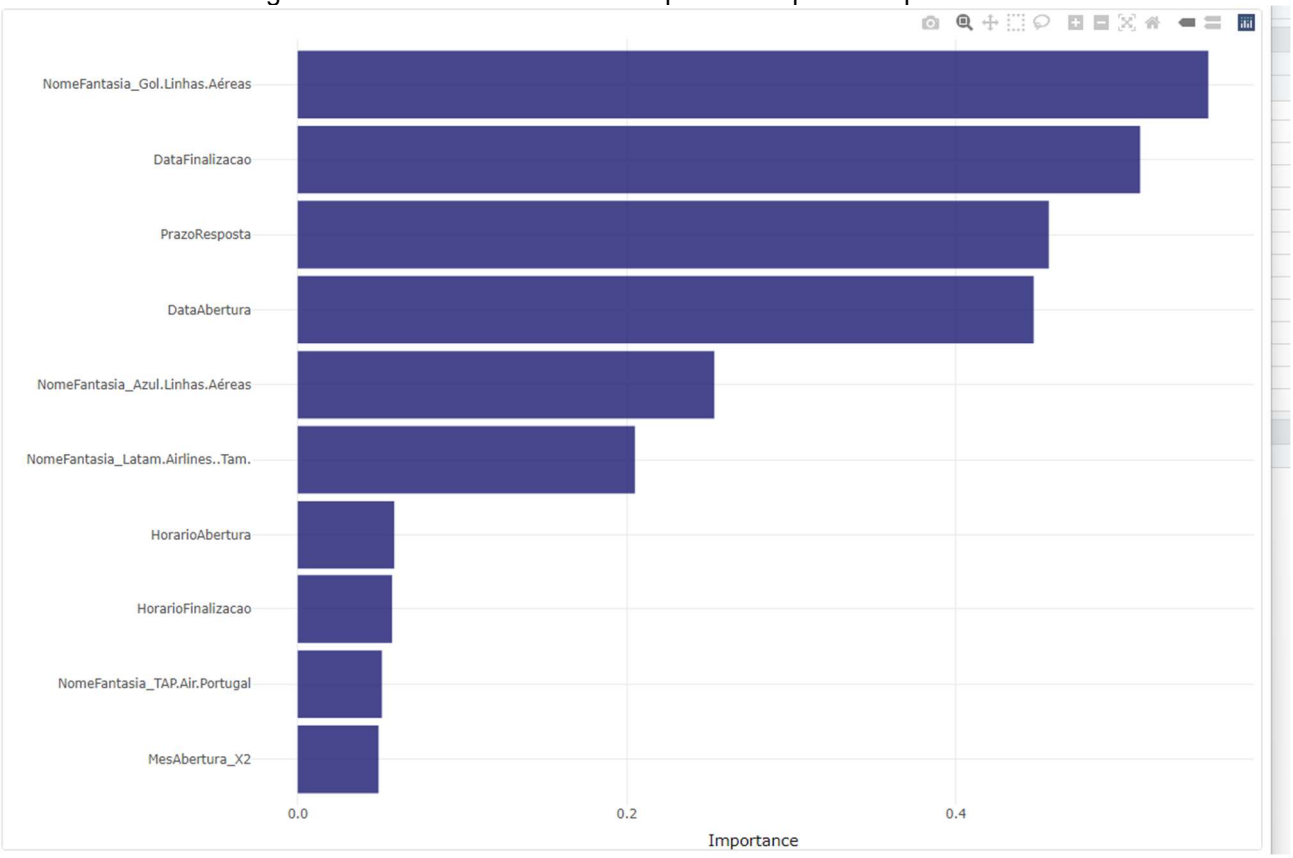
Figura 26 - confronto previsão X tempo de resposta real



Fonte: Autoria própria

Outra visualização possível de extrair dos resultados coletados é definir por grau de importância quais atributos são mais responsáveis por uma variação no tempo de resposta de cada chamada. Conforme figura 27, é possível verificar os fatores mais importantes.

Figura 27 - Fatores determinísticos para o tempo de resposta



Fonte: Autoria própria

4.6 DISCUSSÃO E ANÁLISE DOS RESULTADOS

Os resultados obtidos no capítulo anterior na etapa de *machine learning* se dão através do modelo final gerado pelo *machine learnig*. Através dos resultados dos dados de teste, foi possível concluir que o modelo com maior acurácia, ou seja, deveria ser escolhida para o modelo final, foi o *random forest*. Essa escolha se deu devido a comparação das métricas estabelecidas. Isso se deu por seu R-Quadrado maior e MSE, MSAE e MAPE menor em relação aos outros modelos.

A partir do modelo final foi possível verificar quais fatores são mais preponderantes para definir se o tempo de resposta de uma reclamação. Percebe-se que o maior fator é se a companhia aérea é Gol Aerolineas, isso acontece pelo fato de a companhia ser a com maior número de reclamações, o que gera um comportamento particular se for esta companhia aérea ou outra.

Os resultados do modelo final gerado foram comparados com estudos realizados em bibliografia. Para uma melhor comparação dos resultados, pesquisou-se testes realizados em variáveis correlacionáveis. Segundo a tabela 4, mede-se o sucesso de cada modelo, segundo R^2, prevendo-se a taxa de atrasos em voos realizados.

Tabela 4 – Resultados obtidos por bibliografia

N°	Autor	Método	R²
1	Wei e Vazea (2018)	Simulação	0,8805
2	Chen e Wang (2019)	Regressão usando dados de séries temporais	0,718
3	Chen, Whang e Zhou (2021)	Regressão linear Árvore de decisão Método ensemble Support vector machine (SVM).	0,69 0,65 0,74 0,71
4	Oliveira et al. (2021)	Modelo de regressão logit.	0,4714
5	Arora e Mathur (2020)	Regressão logística multinomial.	0,3
6	Liu, Yin e Hansen (2019)	Modelo de regressão hedônica log-linear	0,716

Fonte: Autoria própria.

Segundo resultados encontrados em bibliografia, nota-se que o valor do R² pode variar dependendo do modelo utilizado, variações dos resultados ocorrem devido ao isolamento de determinadas variáveis. A partir dos resultados obtidos pelo estudo se conclui que os resultados encontrados são satisfatórios e dentro do que se foi encontrado em estudos similares.

5 CONCLUSÕES E TRABALHOS FUTUROS

5.1 CONCLUSÕES

As Tecnologias de Informação (TI) desempenham um papel crucial no atual padrão mundial, no qual o conhecimento e a informação ocupam posições centrais no mercado, impulsionando o progresso e o desenvolvimento. Dentro dessas tecnologias a ciência de dados surge como ferramenta de apoio nas tomadas de decisões, sendo possível transformar dados em informação a fim de melhorar a acurácia das escolhas cotidianas. Dentro da ciência de dados a área de *machine learning* tem sido utilizada em situações reais e tem se tornado cada vez mais prevalente e impactante.

Ao mesmo tempo que o sistema aéreo desempenha um papel fundamental na conectividade global, no comércio internacional, no turismo e no desenvolvimento econômico dos países. Sua importância vai além do simples transporte de passageiros e carga, influenciando positivamente diversos setores da sociedade.

No estudo realizado foi realizado um levantamento de dados relacionados ao sistema aéreo de estados do Brasil, desde sua extração até as visualizações gráficas. Indicadores foram levantados segundo bibliografia e os resultados foram expostos em uma matriz de resultados. Destes resultados pode-se realizar diversas conclusões.

Para o estado do Rio Grande do Sul vê-se que contém mais voos durante o ano, 37808, consequentemente com a maior média de voos por mês 3151. O estado ainda conta com a maior taxa de cancelamento. Uma particularidade do Rio Grande do Sul é seu pico de horário de chegada, diferente dos outros estados o horário mais movimentado é às 11h da manhã. Ainda se destaca que o estado obteve a pior taxa de resposta, média de avaliação e média do tempo de resposta do estudo, ao mesmo tempo que conteve o menor pico de reclamações em um único mês.

No estado do Paraná nota-se que é o segundo estado com mais voos, destaca-se que o estado teve a melhor taxa de resposta, média de avaliação e média no tempo de resposta do estudo. Além de ter uma maior participação no número de avaliações por cidade. Curitiba a cidade que mais avalia no estado tem mais que o dobro de Porto Alegre, e cerca de 2,5 vezes Florianópolis, cidades que mais participam nos outros estados do estudo de caso.

Santa Catarina é o estado que com menor número de voos anuais, 1780 voos anuais, o que impacta na taxa de cancelamento, a menor do estudo com 2,23%. É possível perceber que o mês com mais cancelamento foi dezembro com 57 contra 138 e 271 de Paraná e Rio Grande do Sul, respectivamente. Destaca-se os horários de partidas mais comuns no estado, enquanto os outros estados horários no final do dia são mais comuns, como 17h ou 19h, em Santa Catarina o horário mais comum é as 5h da manhã. Outro ponto levantado é que mesmo o estado tendo uma quantidade menor de voos em um ano, as reclamações tais qual os tempos de respostas ficaram atrás do Paraná.

É possível perceber alguns pontos que se repetem para os três estados, tanto para chegadas quanto para partidas os aeroportos que mais fazem conexões com os aeroportos do estudo são, aeroporto de Guarulhos, aeroporto de Congonhas e

aeroporto de Viracopos, todos presentes no estado de São Paulo. Assim como as empresas que mais utilizam os aeroportos, Azul, Latam e Gol. Interessante observar que a Azul é a empresa com mais maiores utilizações tanto em chegadas, quanto em partidas em todos os estados exceto chegadas em Santa Catarina. Ressalta-se que mesmo a companhia Azul ser líder em quase todos os trajetos a empresa com maior número de reclamações é a companhia Gol para todos os estados.

Na parte onde se abordou o modelo de *machine learning* a fim de prever o tempo de resposta, vê-se que os resultados das métricas estabelecidas pelo trabalho. Os resultados da regressão linear foram muito parecidos com os resultados da árvore de decisão o com R-Quadrado 0,4, enquanto, *random forest* obteve resultado de 0,66. O RMAE e o RSE do *random forest* obteve os valores mais baixos, cerca de 0,54 e 0,49 respectivamente. Enquanto regressão linear obteve aproximadamente 0,77 e 0,61. E árvore de decisão cerca de 0,77 e 0,6, valores próximos da regressão linear. Com os resultados obtidos se conclui que o modelo foi *random forest* obteve maior acurácia, por este motivo o modelo final foi gerado com utilizando este modelo obtendo um R-Quadrado 0,65, muito próximo do treinamento dos dados. Concluindo que os fatores mais determinísticos para um bom tempo de resposta é se a companhia é Gol ou não, a data de finalização, tempo de resposta e data de abertura.

5.2 TRABALHOS FUTUROS

Para trabalhos futuros, sugere-se que se aumenta a áreas de estudo, podendo fazer realizar a elaboração de visualizações para todos os estados do Brasil, além de fornecer gráficos para todo o território brasileiro, abrindo possibilidades para comparar determinado estado em relação cenário nacional. Indica-se que para uma melhor avaliação nas tomadas de decisão se desenvolva uma quantidade mais diversa de tipos de visualizações.

No que se refere a parte de aprendizado de máquina, recomenda-se prever outras variáveis como, notas dos consumidores, cancelamento de voos ou aberturas de chamados por mês. Para uma maior robustez dos resultados, sugere-se definir maior número de métricas para medir o sucesso dos modelos de *machine learning*.

Dado que agência reguladora fornece grandes volumes de dados, em diferentes arquivos, é possível em trabalhos futuros incluir mais dados oriundos do mesmo período a fim de que se aumente a quantidade de visualizações e as possibilidades para os tomam as decisões.

REFERÊNCIAS

AGÊNCIA NACIONAL DE AVIAÇÃO CIVIL (ANAC). **Relatório de Desempenho Operacional**. Disponível em: https://www.anac.gov.br/assuntos/dados-e-estatisticas/desempenho-operacional. Acesso em: 11 abr. 2023.

AGÊNCIA NACIONAL DE AVIAÇÃO CIVIL (ANAC). **Relatório de Indicadores de Qualidade do Serviço de Transporte Aéreo**. Disponível em: https://www.anac.gov.br/assuntos/dados-e-estatisticas/relatorios-de-indicadores-de-qualidade-do-servico-de-transporte-aereo. Acesso em: 11 abr. 2023.

AGGARWAL, C. C. **Data Mining: The Textbook**. Springer, 2013.

AGUIAR, L.; BORGES, R.; VILELA, C. **Análise da pontualidade dos voos comerciais no Brasil**. Revista do BNDES, v. 53, n. 1, p. 257-278, 2020.

ALPAYDIN, E. **Introduction to Machine Learning (2nd ed.)**. MIT Press, 2010.

BABBIE, E. **The Practice of Social Research**. Cengage Learning, 2016.

BISHOP, C. M. **Pattern recognition and machine learning**. New York: Springer, 2006.

BLOM, H. A.; VISSER, H. G.; WESTERVELD, T. H. **Indicators for strategic planning of air traffic control: a case study**. Procedia Computer Science, v. 80, p. 2062-2071, 2016.

BRUCE, A.; BRUCE, P. **Estatística Prática para Cientistas de Dados.** [S.l.]: Alta Books, 2019.

CAIRO, A. **The Truthful Art: Data, Charts, and Maps for Communication**. New Riders, 2016.

CAMILO, C. O.; SILVA, J. C. d. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas**. Universidade Federal de Goiás (UFC), 2009.

CHEN, H.; CHIANG, R. H.; STOREY, V. C. **Business intelligence and analytics: From big data to big impact**. MIS quarterly, v. 36, n. 4, 2012.

CHOI et al. **Customer satisfaction factors of low-cost carriers and full-service airlines: A comparative analysis**. Journal of Travel & Tourism Marketing, v. 34, n. 1, p. 50-65, 2017.

CLEVELAND, W. S.; MCGILL, M. E. **Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods.** Journal of the American Statistical **Association**, v. 79, n. 387, p. 531-554, 1984.

D'ANDREA, A. et al. **Air Traffic Flow Management: Review and Critical Analysis.** Aerospace, v. 4, n. 4, p. 58, 2017.

D'AURIA, R. et al. **Data-Driven Air Traffic Flow Management: Models and Algorithms for Traffic Prediction and Delay Optimization**. Springer, 2020.

DAYALA, R. **Linear Regression**. Disponível em: <www.medium.com/@rndayala/linear-regression-a00514bc45b0>. Acesso em: junho de 2023.

DENG, M. et al. **Study on the Traffic Flow Control and Operation Mode of Large Airports in China**. In: Advances in Engineering Research, v. 92, p. 151-155, 2017.

DINO. **US$ 965 milhões em 2018: o promissor mercado dos dados requer integração.** [S. l.], 8 fev. 2018. Disponível em: https://www.Mundodomarketing.com.br/noticiascorporativas/conteudo/125458/us$-965-milhoes-em-2018-o-promissor-mercado-dos-dados-requer-integracao. Acesso em: 10 Abr. 2023.

DOMINGOS, P. **The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World**. Basic Books, 2015.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 2nd Edition. Wiley, 1981.

EXAME. **O segredo das empresas que sabem usar os dados a seu favor.** [S. l.], 1 dez. 2021. Disponível em: https://exame.com/inovacao/osegredodasempresasque-sabemusarosdadosaseufavor/. Acesso em: 3 agosto. 2022.

FEW, S. **Show Me the Numbers: Designing Tables and Graphs to Enlighten**. Analytics Press, 2012.

GONZÁLEZ-PRIETO et al. **Airline passengers' perceptions of service quality and their complaint behavior: A segmentation approach**. Journal of Travel Research, v. 59, n. 2, p. 337-351, 2020.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016.

GORMAN, B. **The Delta TechOps Predictive Maintenance Story. InformationWeek**, 2019. Disponível em: https://www.informationweek.com/big-data/ai-machine-learning/the-delta-techops-predictive-maintenance-story/a/d-id/1335245. Acesso em: 11 abr. 2023.

GRAHAM, J. W. **Missing data: Analysis and design**. Springer, 2015.

HAIR JR., J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E. **Multivariate Data Analysis**. Pearson, 2014.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. Morgan Kaufmann, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction (2ª ed.).** Springer, 2009.

IGUAL, L. et al. **Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications.** [S.l.]: Springer International Publishing, 2017.

ITTNER, C. D.; LARCKER, D. F.; MEYER, M. W. **Subjectivity and the weighting of performance measures: evidence from a balanced scorecard**. The Accounting Review, v. 78, n. 3, p. 725-758, 2003.

JAKOB, M. **Air Traffic Management and Systems**. Springer, 2018.

JANSSEN, M.; VOORT, H, V der.; WAHYUDI, A. **Factors influencing big data decisionmaking quality**. Journal of Business Research, v. 70, p. 338345, 2016.

JORDAN, M. I.; MITCHELL, T. M. **Machine learning: Trends, perspectives, and prospects**. Science, v. 349, n. 6245, p. 255-260, 2015.

KAPLAN, R. S.; NORTON, D. P. **The balanced scorecard: translating strategy into action**. Harvard Business Press, 1996.

KIM, Y. J.; MIN, H.; PARK, D. **Do service quality, onboard service, and flight operations affect airlines' financial performance?**. Journal of Air Transport Management, v. 60, p. 30-37, 2017.

KOZYRKOV, C. **Building trustworthy AI**. O'Reilly Media, Inc., 2019.

KUABARA, M. T. **Avaliação do desempenho operacional de aeroportos brasileiros utilizando dados ADS-B**. Revista Transportes, v. 25, n. 1, p. 102-117, 2017.

KVALSETH, T. O. **Cautionary note about R^2**. The American Statistician, v. 39, n. 4, p. 279-285, 1985.

MARKHAM, K. **Data Wrangling with Python**. O'Reilly Media, 2020.

MCGILL, R.; TUKEY, J. W.; LARSEN, W. A. **Variations of box plots. The American Statistician**, v. 32, n. 1, p. 12-16, 1978.

MCKINNEY, W. **Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython**. O'Reilly Media, 2017.

MYERS, R. H.; MONTGOMERY, D. C.; ANDERSON-COOK, C. M. **Response Surface Methodology: Process and Product Optimization Using Designed Experiments**. John Wiley & Sons, 2010.

NETER, J. et al. **Applied Linear Statistical Models**. 4th Edition. New York: WCB McGraw-Hill, 1996.

NEUMARK, D.; ZHANG, J.; CICCARELLA, S. **Airline competition and domestic US airfares: A comprehensive reappraisal**. Journal of Transport Economics and Policy, v. 51, n. 3, 2017.

PARMENTER, D. **Key Performance Indicators: Developing, Implementing, and Using Winning KPIs**. John Wiley & Sons, 2015.

RAO, B. L.; REDDY, S. C.; VENKATESH, B. **Air Traffic Management System**. International Journal of Engineering Research and Technology, v. 6, n. 12.

RASCHKA, S.; MIRJALILI, V. **Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2**. Packt Publishing, 2020.

SALTZ, J.; STANTON, J. **An Introduction to Data Science**. [S.l.]: SAGE Publications, 2017.

SANTOS, L. C. et al. **Personalização de serviços em empresas aéreas: um estudo sobre as preferências dos passageiros brasileiros**. Revista Brasileira de Pesquisa em Turismo, v. 13, n. 1, p. 108-126, 2019.

SMITH, J. **Quality Management in the Aviation Industry.** International Journal of Aviation Management, v. 6, n. 2, p. 42-55, 2019.

SPIEGELHALTER, D.; PEARSON, M.; SHORT, I. **Visualizing uncertainty about the future. Science**, v. 333, n. 6048, p. 1393-1400, 2011.

SUTTON, R. S.; BARTO, A. G. **Reinforcement learning: An introduction (2nd ed.)**. Cambridge, MA: MIT Press, 2018.

TUFTE, E. R. **The Visual Display of Quantitative Information**. Graphics Press, 2001.

TURBAN, E.; SHARDA, R.; ARONSON, J. E.; KING, D. **Business intelligence: um enfoque gerencial para a inteligência do negócio**. Bookman Editora, 2009.

TURNER, R. W. **The many models fallacy: don't make every model a deep learning model**. Medium, 2021. Disponível em: https://towardsdatascience.com/the-many-models-fallacy-dont-make-every-model-a-deep-learning-model-a3c46fbcf6c2. Acesso em: 04 maio 2023.

WICKHAM, H. **R Packages: Organize, Test, Document, and Share Your Code**. O'Reilly Media, 2015.

WICKHAM, H. **The split-apply-combine strategy for data analysis**. Journal of Statistical Software, v. 40, n. 1, p. 1-29, 2011.

WICKHAM, H.; GROLEMUND, G. **R for Data Science: Import, Tidy, Transform, Visualize, and Model Data**. O'Reilly Media, Inc., 2017.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. Morgan Kaufmann, 2016.

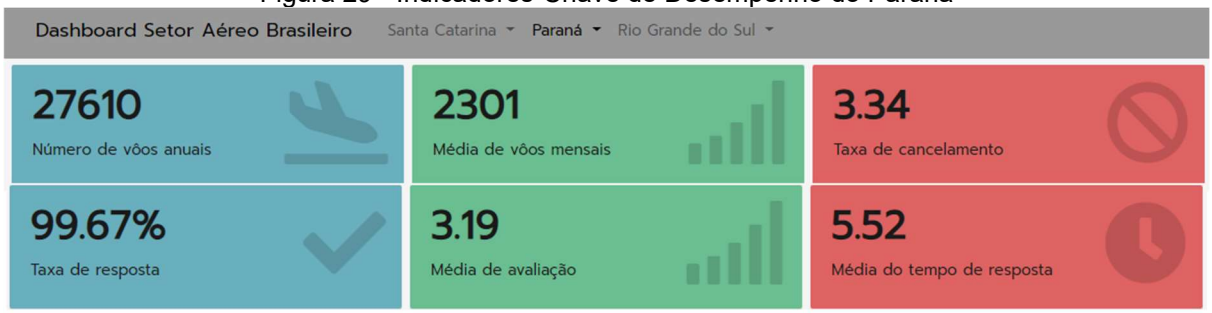
WIXOM, B.; ARIYACHANDRA, T.; GOUL, M.; GRAY, P.; KULKARNI, U.; PHILLIPS-WREN, G. **The current state of business intelligence in academia**. Communications of the Association for Information Systems, v. 29, n. 1, p. 16, 2011.

APÊNDICE A - INDICADORES-CHAVE DE DESEMPENHO

Figura 28 - Indicadores-Chave de Desempenho de Santa Catarina

Fonte: Autoria própria

Figura 29 - Indicadores-Chave de Desempenho do Paraná



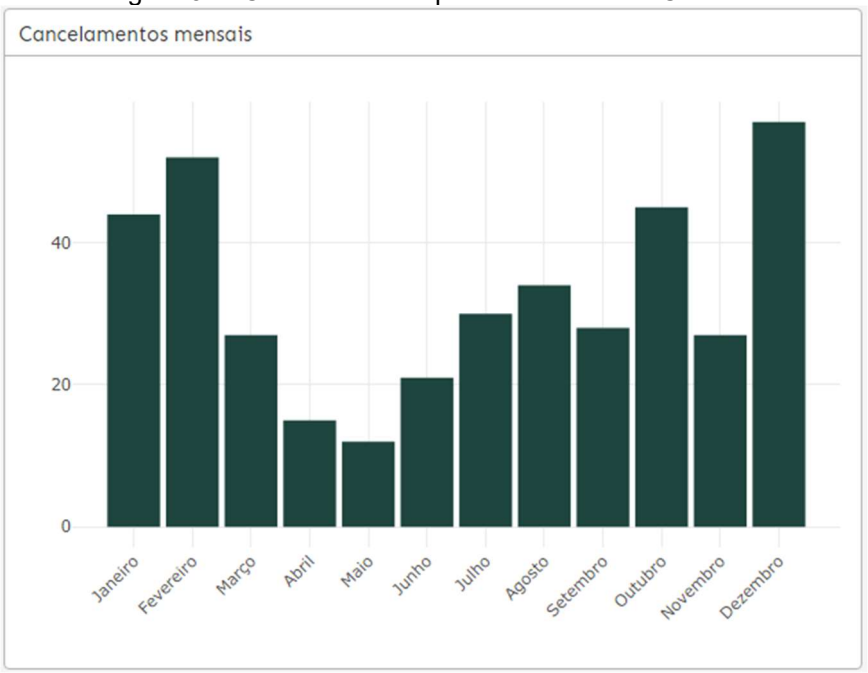
Fonte: Autoria própria

Figura 30 - Indicadores-Chave de Desempenho do Rio Grande do Sul

Fonte: Autoria própria

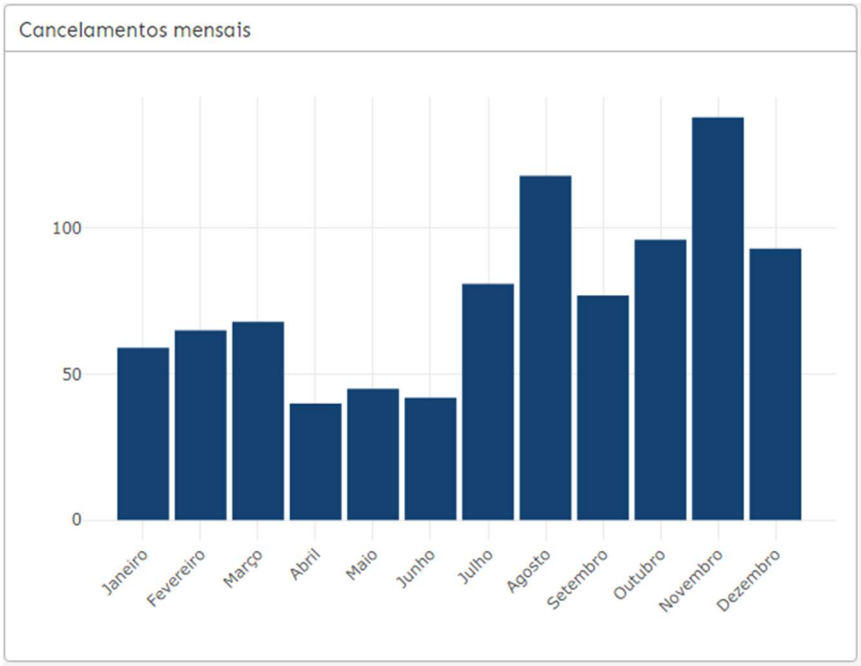
APÊNDICE B - CANCELAMENTOS POR MÊS

Figura 31 - Cancelamentos por mês em Santa Catarina



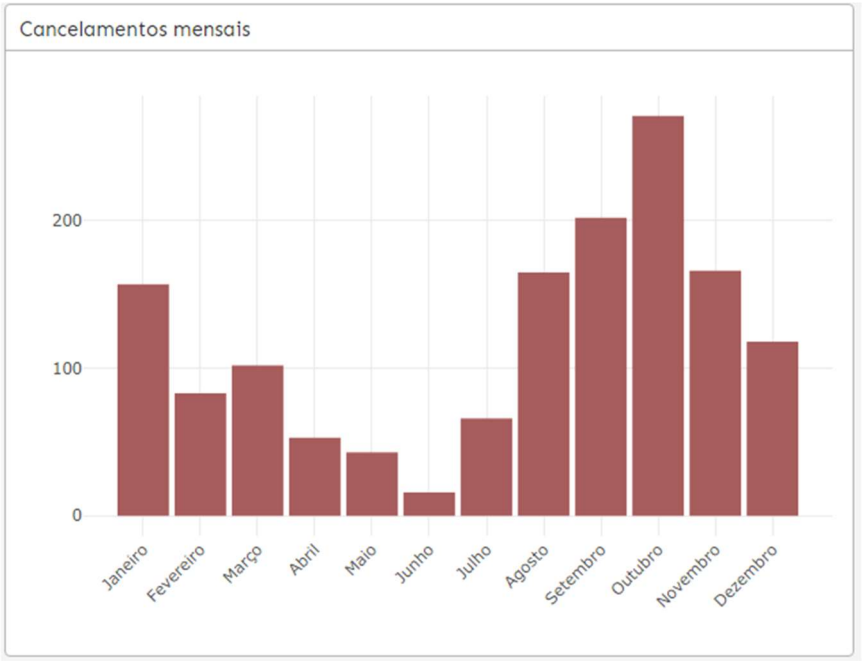
Fonte: Autoria própria

Figura 32 - Cancelamentos por mês no Rio Grande do Sul



Fonte: Autoria própria

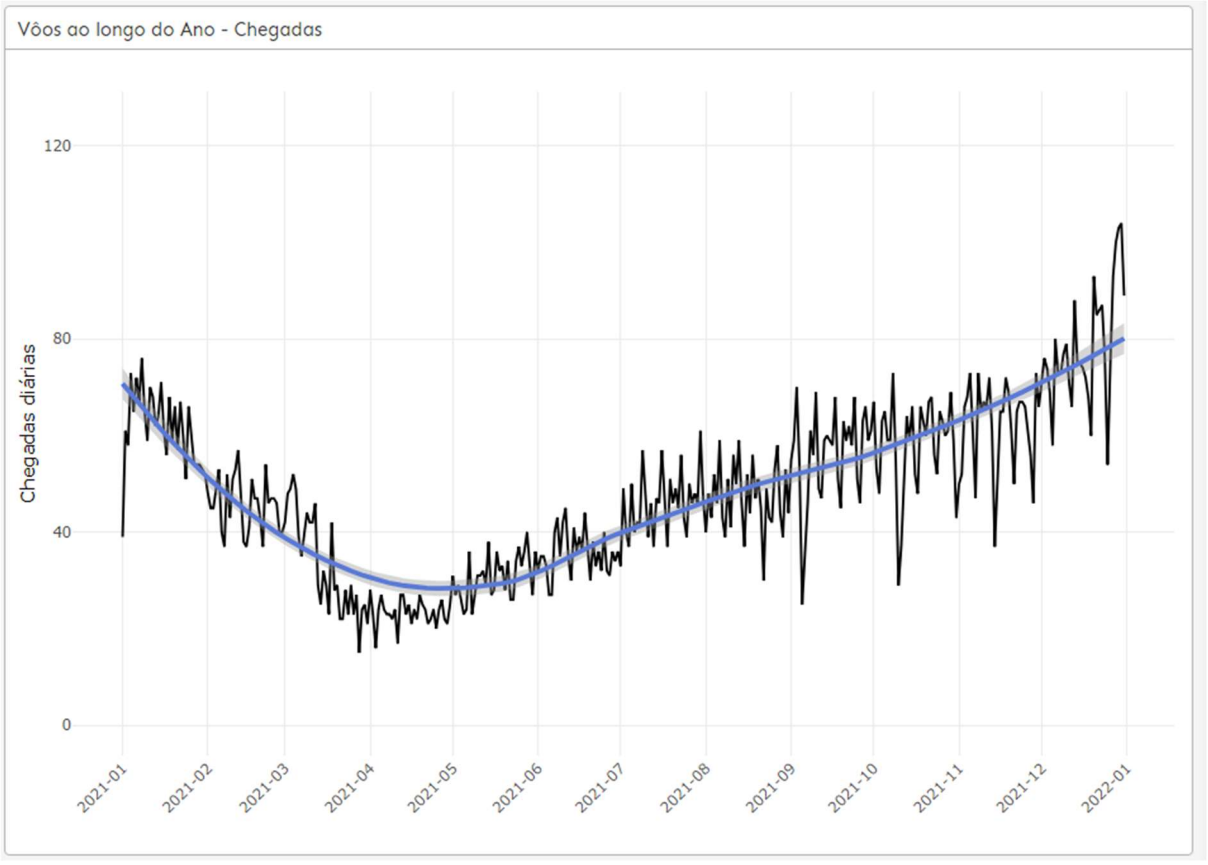
Figura 33 - Cancelamentos por mês no Rio Grande do Sul



Fonte: Autoria própria

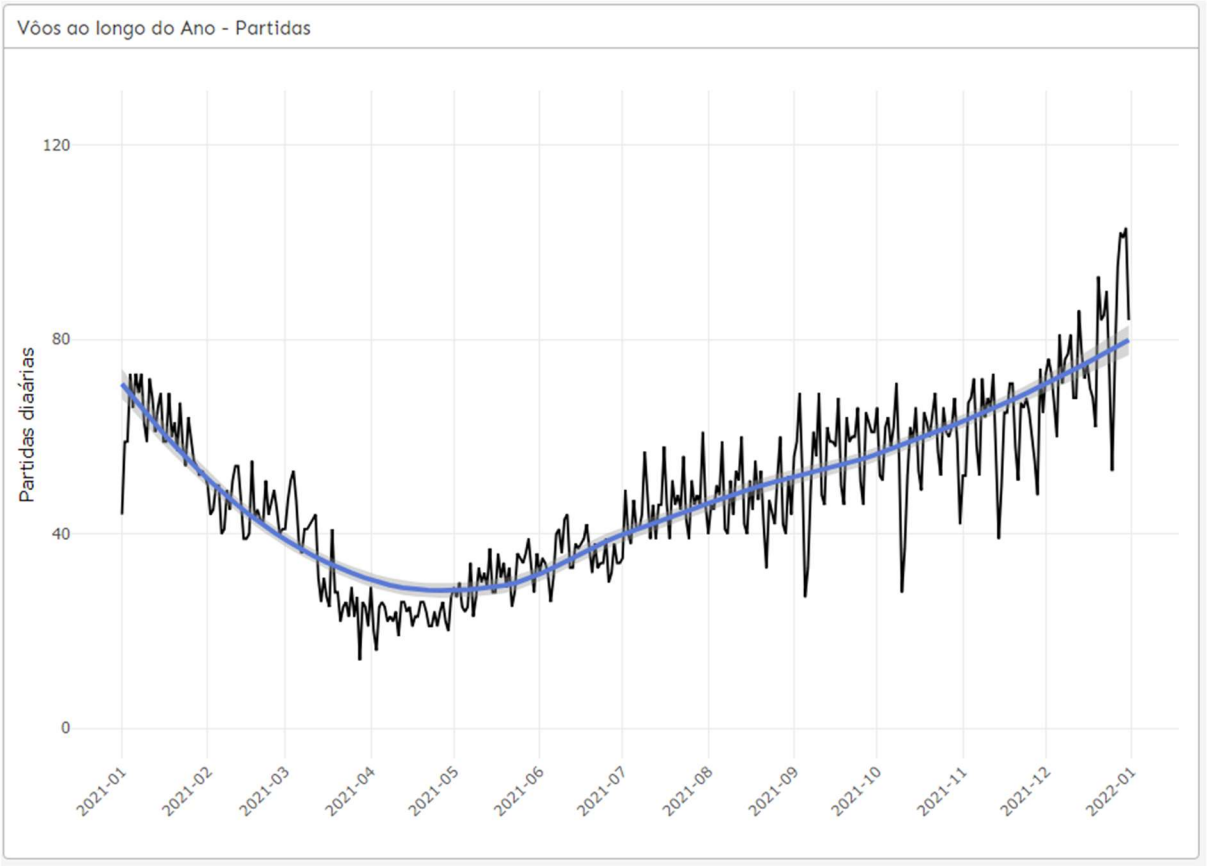
APÊNDICE C - DADOS TEMPORAIS

Figura 34 - Vôos ao longo do ano em Santa Catarina - Chegadas



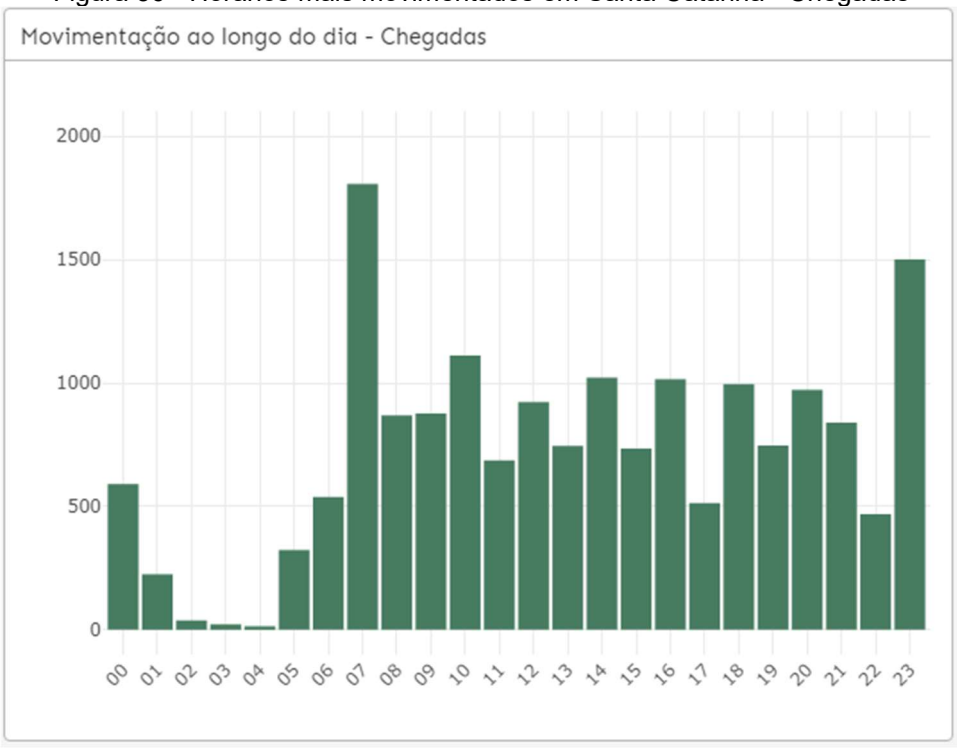
Fonte: Autoria própria

Figura 35 - Vôos ao longo do ano em Santa Catarina - Partidas



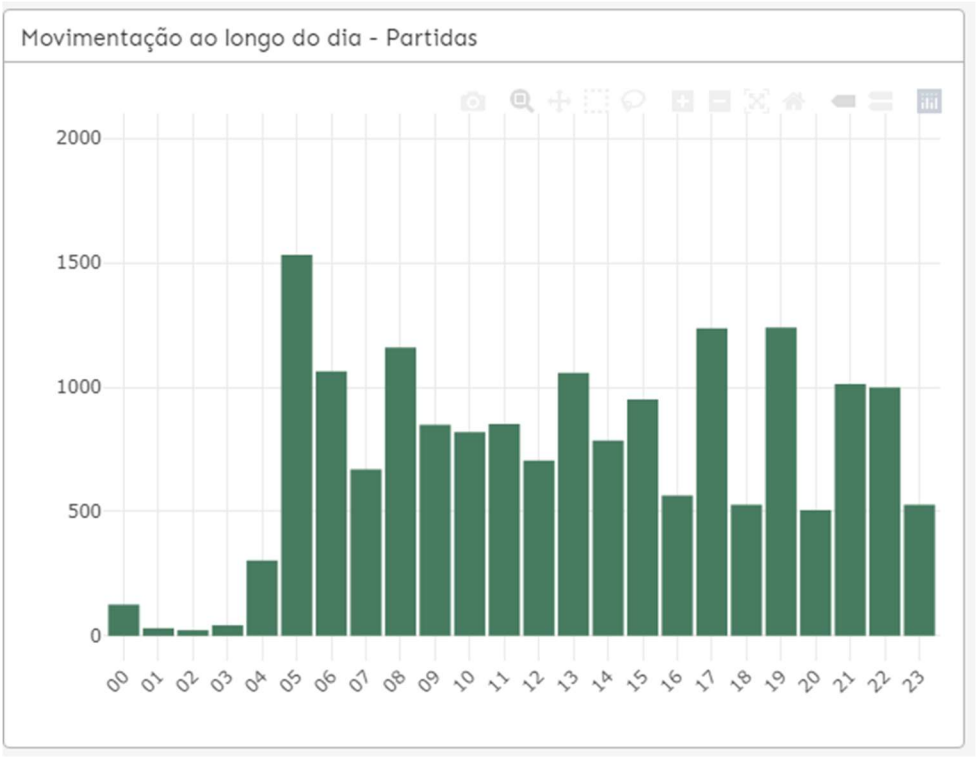
Fonte: Autoria própria

Figura 36 - Horários mais movimentados em Santa Catarina - Chegadas



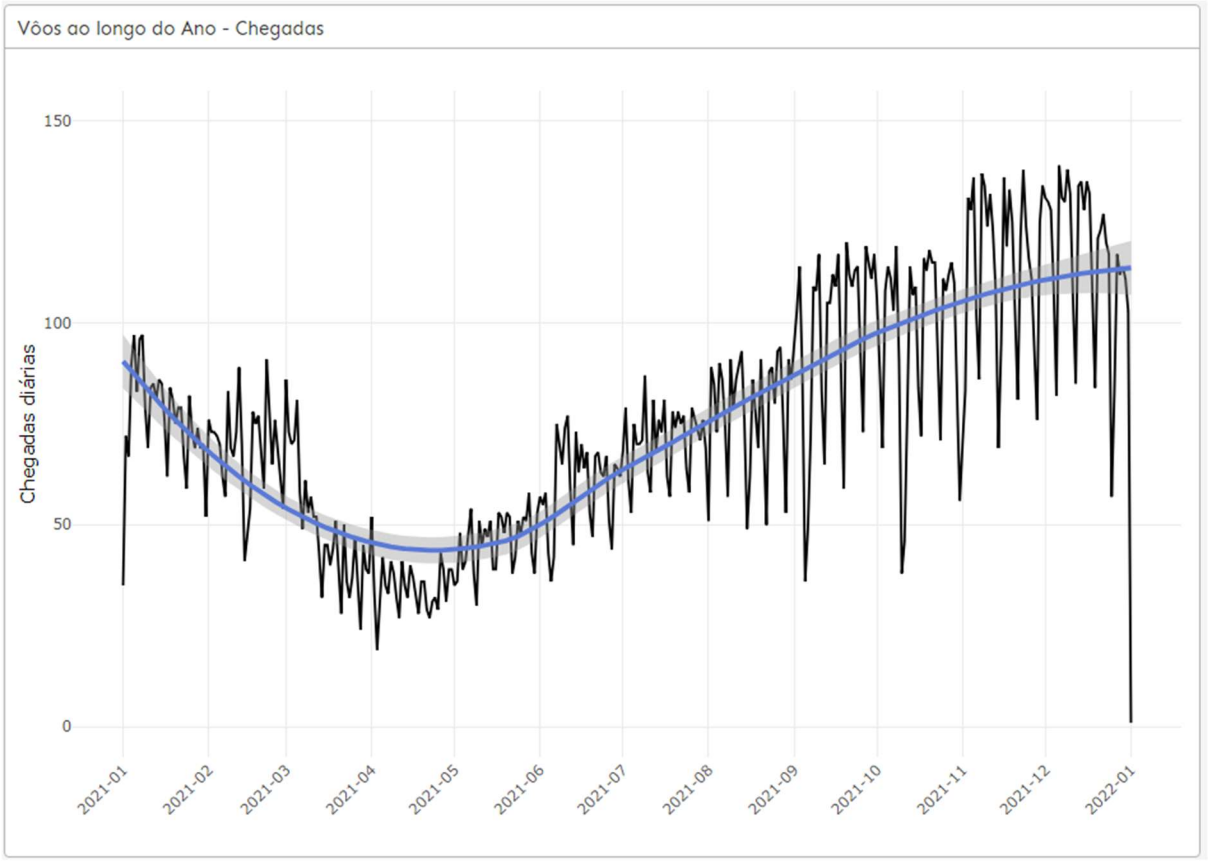
Fonte: Autoria própria

Figura 37 - Horários mais movimentados em Santa Catarina - Partidas



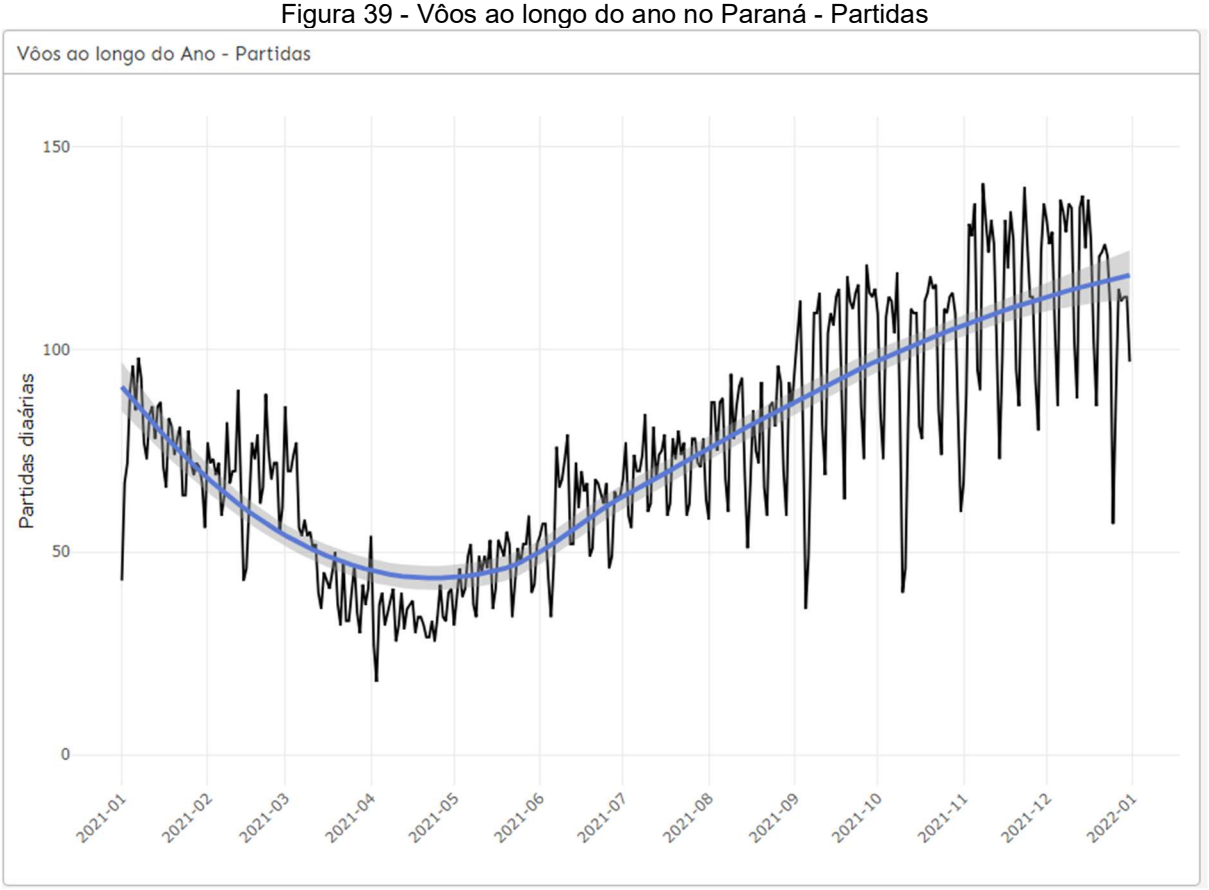
Fonte: Autoria própria

Figura 38 -Vôos ao longo do ano no Paraná - Chegadas



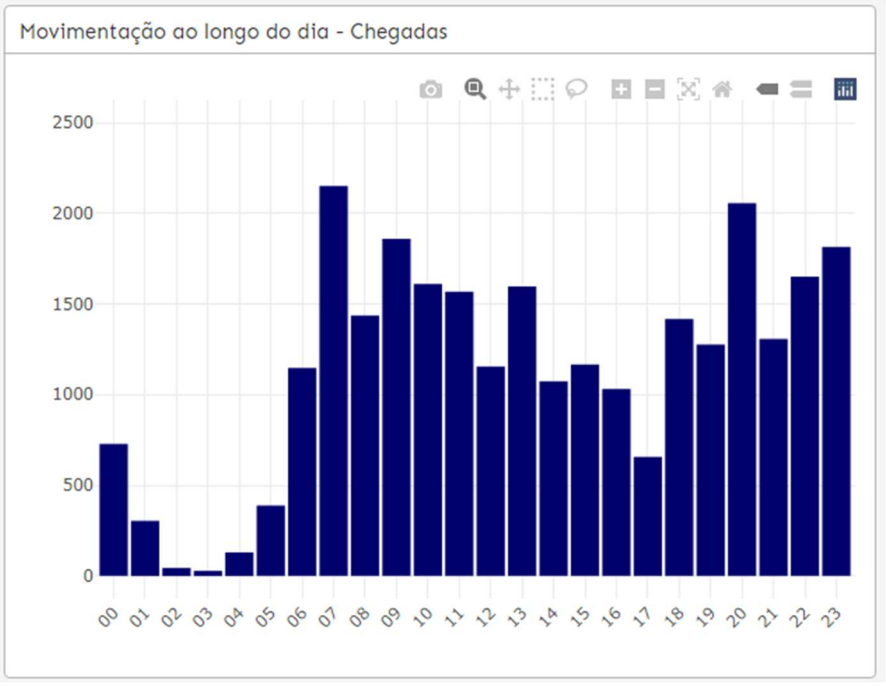
Fonte: Autoria própria

Figura 39 - Vôos ao longo do ano no Paraná - Partidas



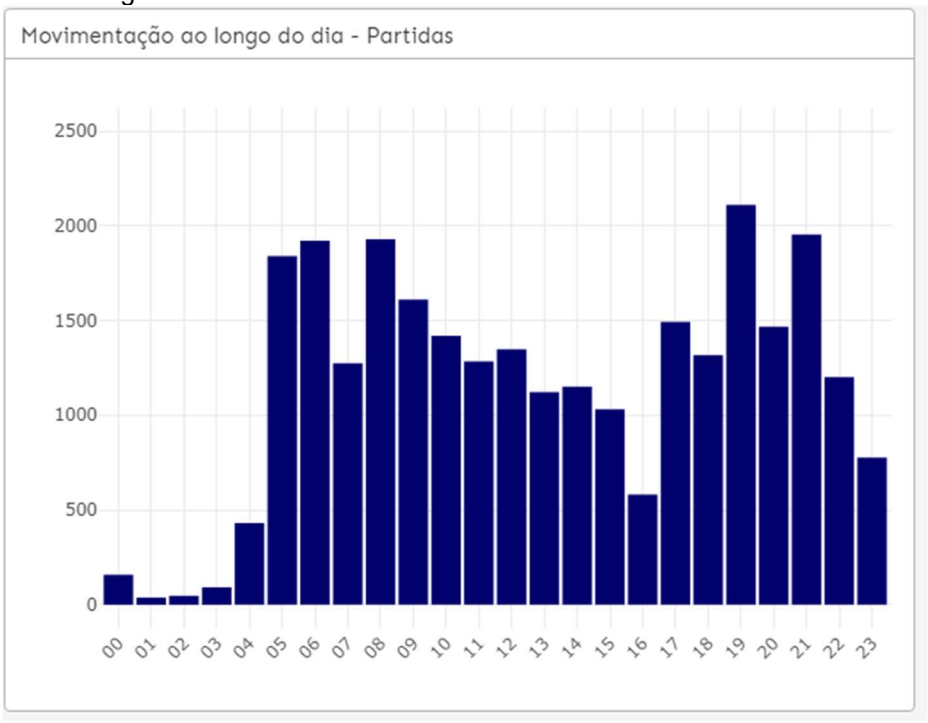
Fonte: Autoria própria

Figura 40 - Horários mais movimentados no Paraná - Chegadas



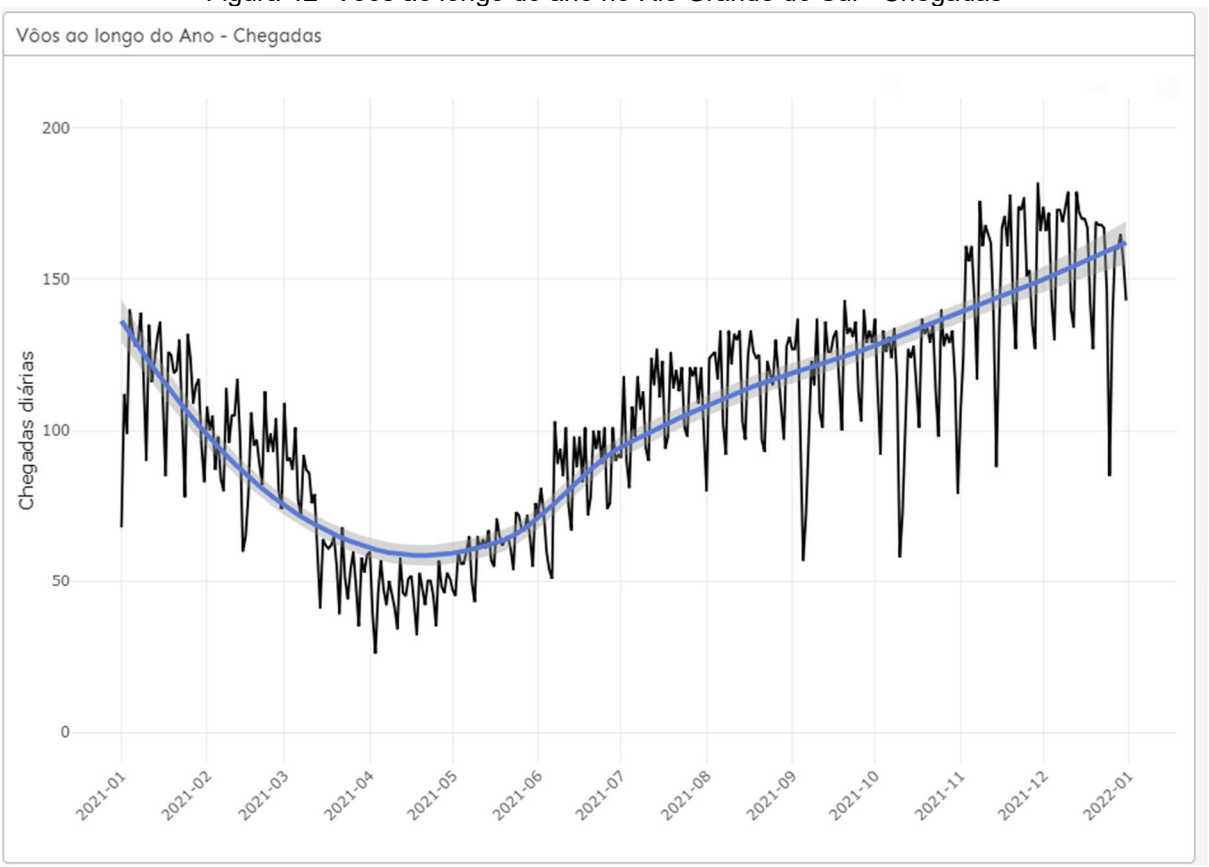
Fonte: Autoria própria

Figura 41 - Horários mais movimentados no Paraná - Partidas



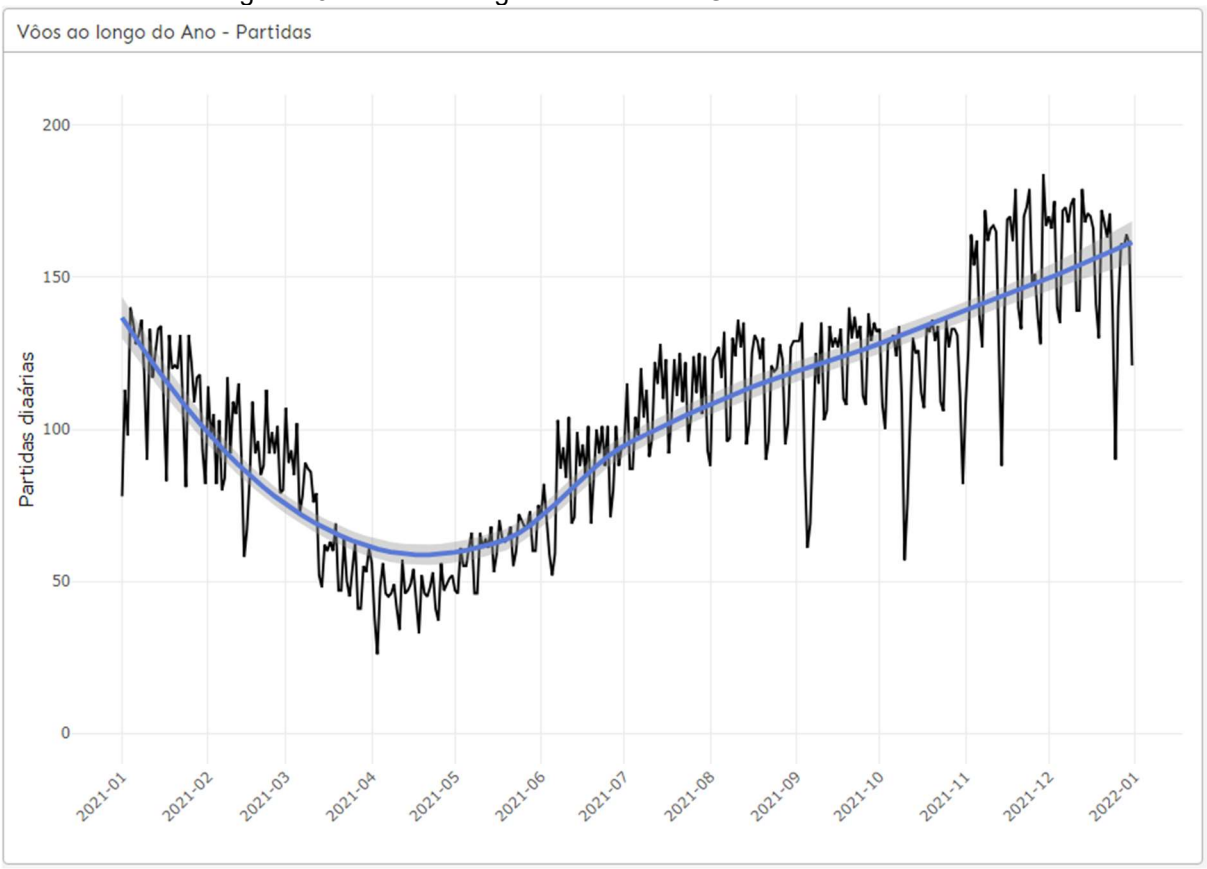
Fonte: Autoria própria

Figura 42 -Vôos ao longo do ano no Rio Grande do Sul - Chegadas



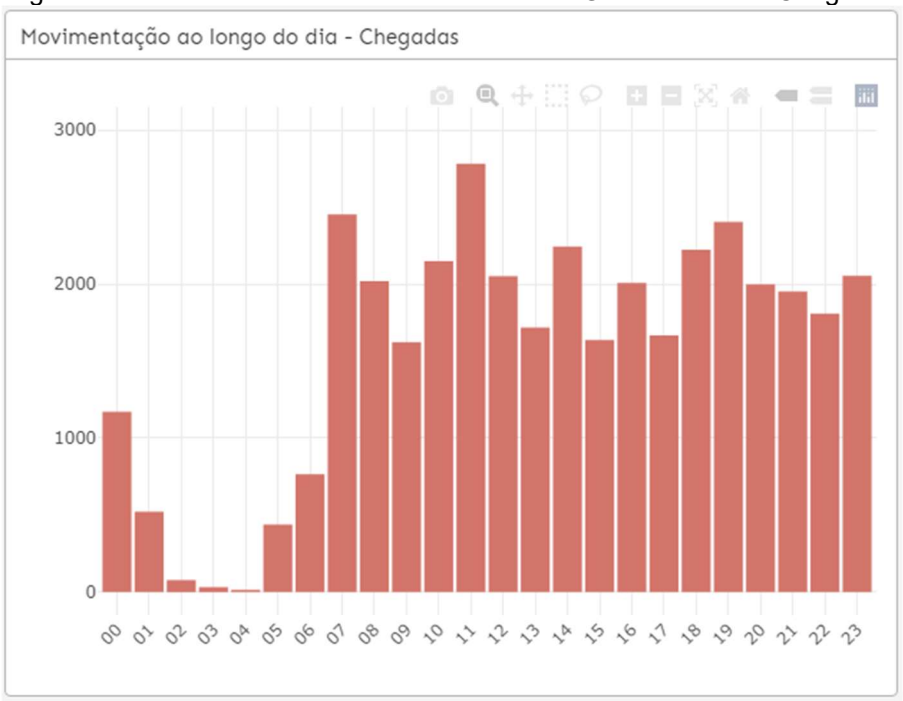
Fonte: Autoria própria

Figura 43 - Vôos ao longo do ano no Rio Grande do Sul- Partidas



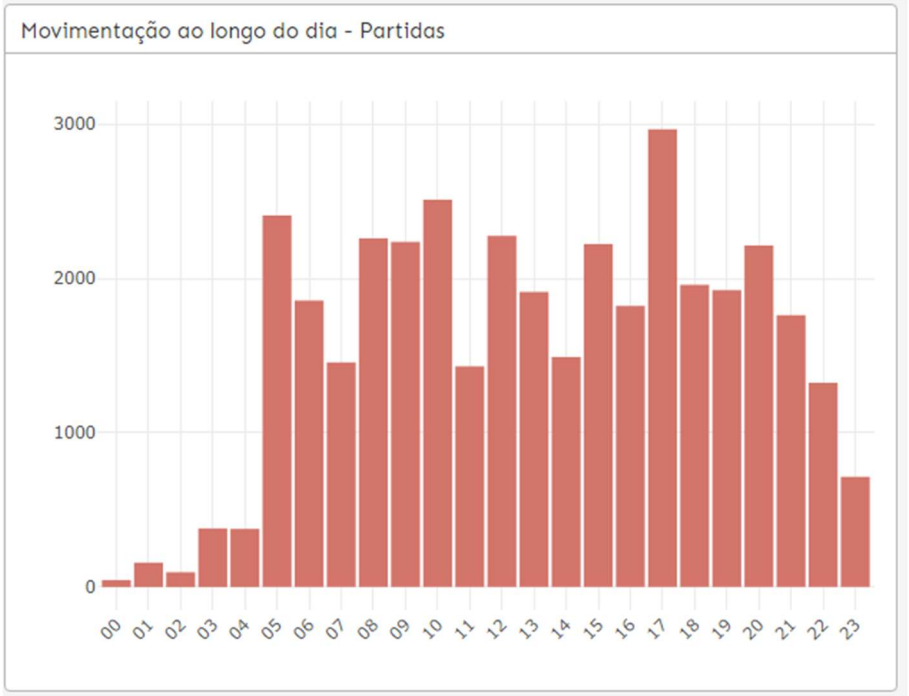
Fonte: Autoria própria

Figura 44 - Horários mais movimentados no Rio Grande do Su l- Chegadas



Fonte: Autoria própria

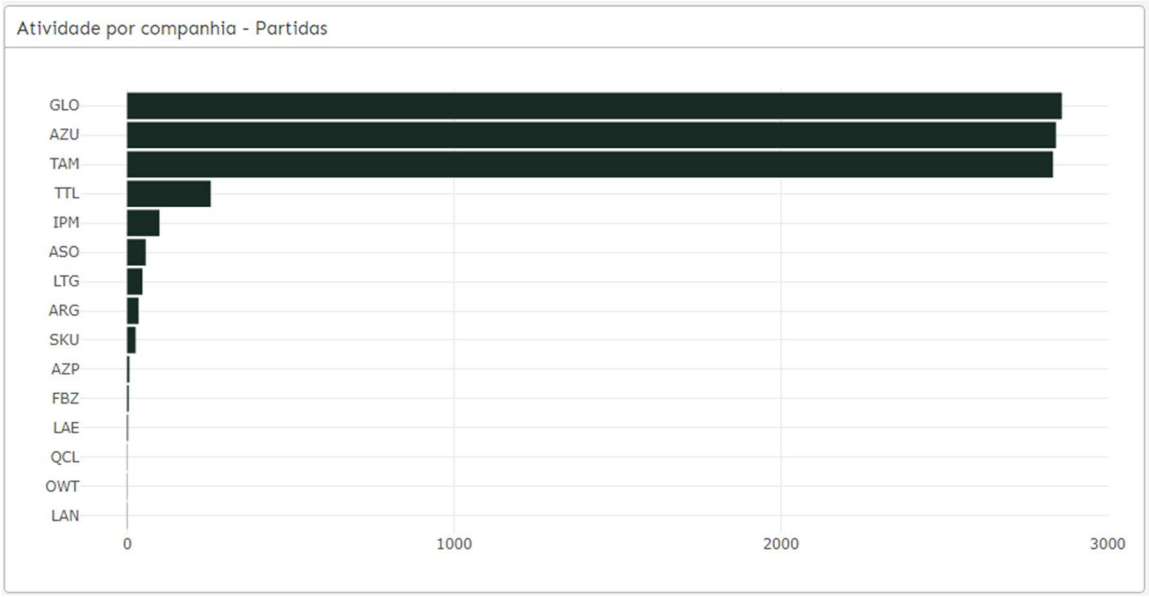
Figura 45 - Horários mais movimentados no Rio Grande do Sul - Partidas



Fonte: Autoria própria

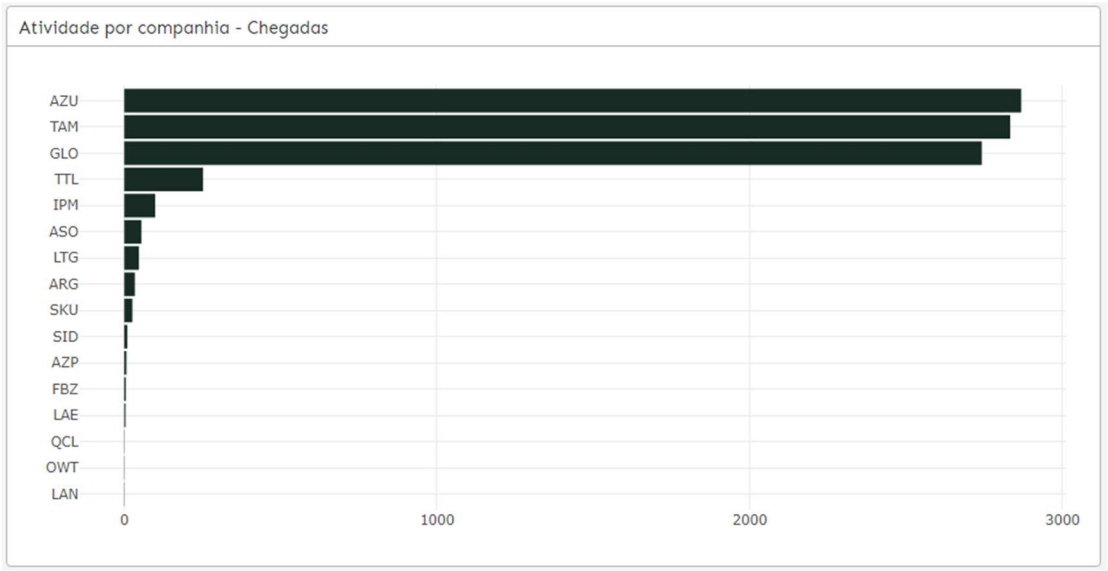
APÊNDICE D - ATIVIDADES POR COMPANHIA AÉREA

Figura 46 - Companhias com mais vôos em Santa Catarina - Partidas



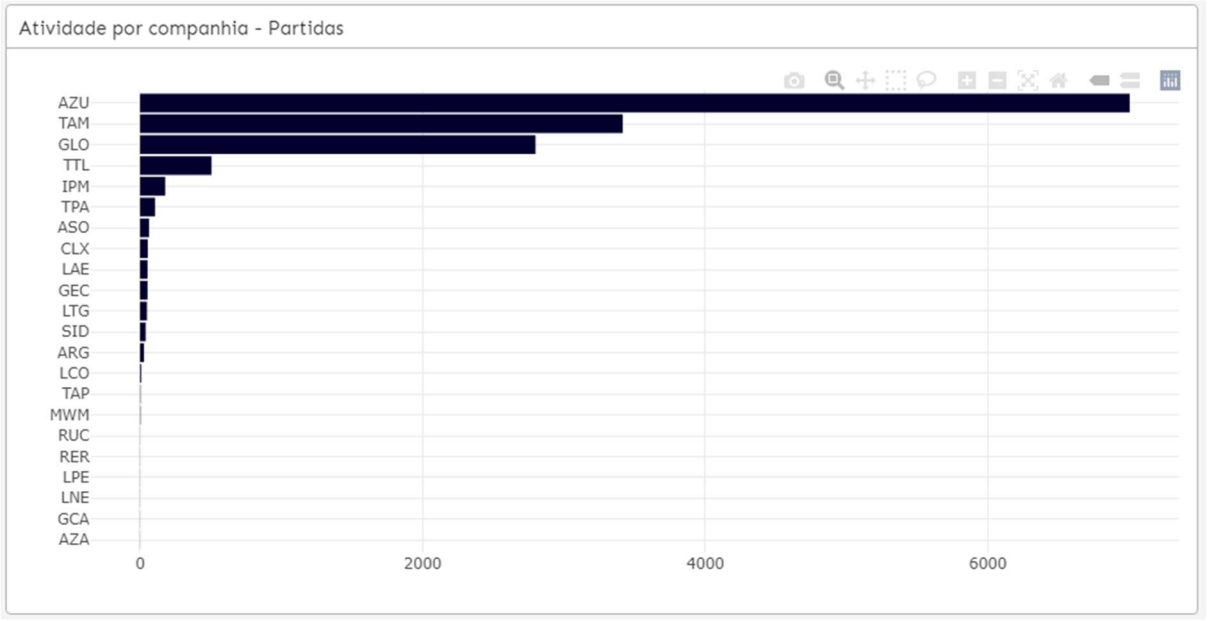
Fonte: Autoria própria

Figura 47 - Companhias com mais vôos em Santa Catarina - Chegadas



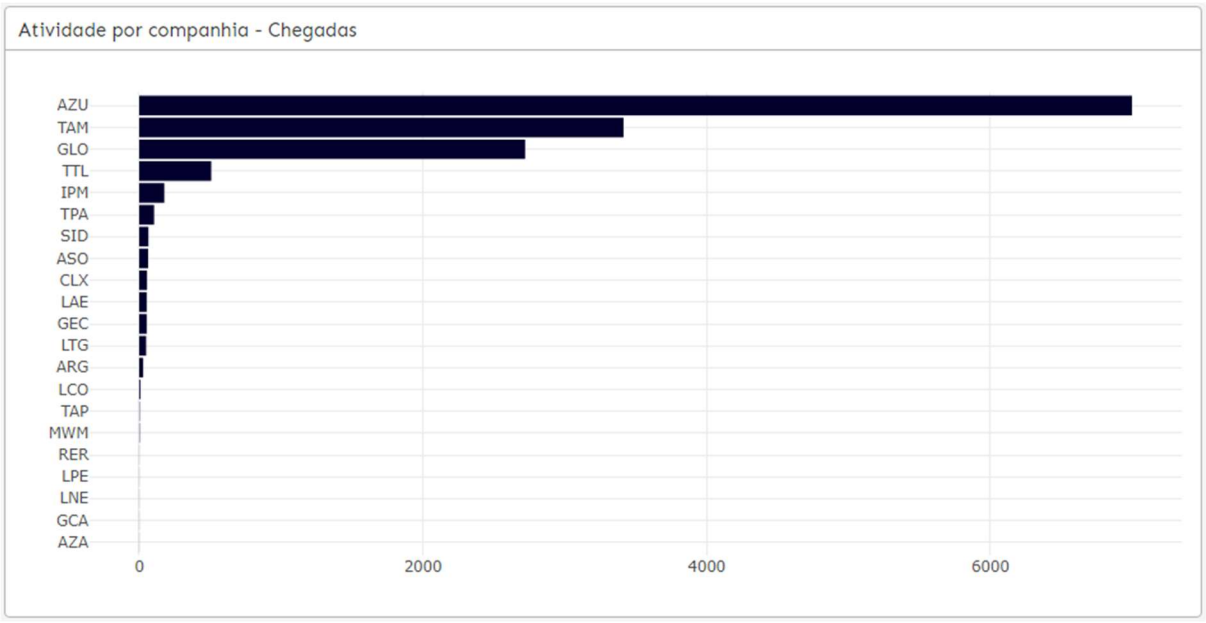
Fonte: Autoria própria

Figura 48 - Companhias com mais vôos no Paraná - Partidas



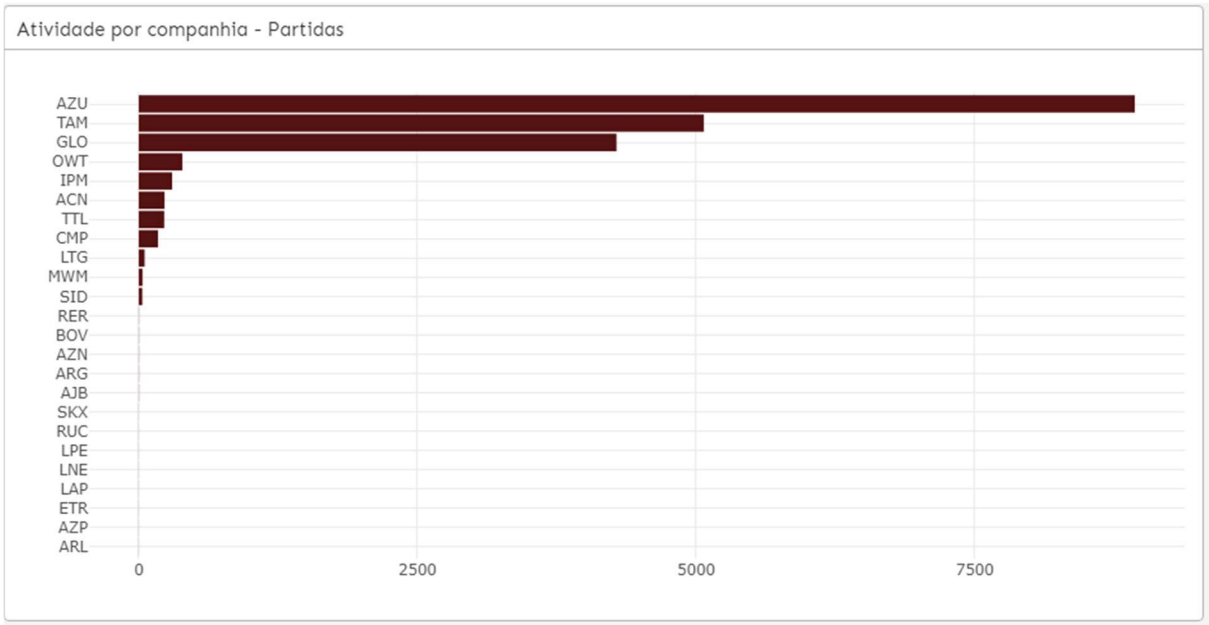
Fonte: Autoria própria

Figura 49 - Companhias com mais vôos no Paraná - Chegadas



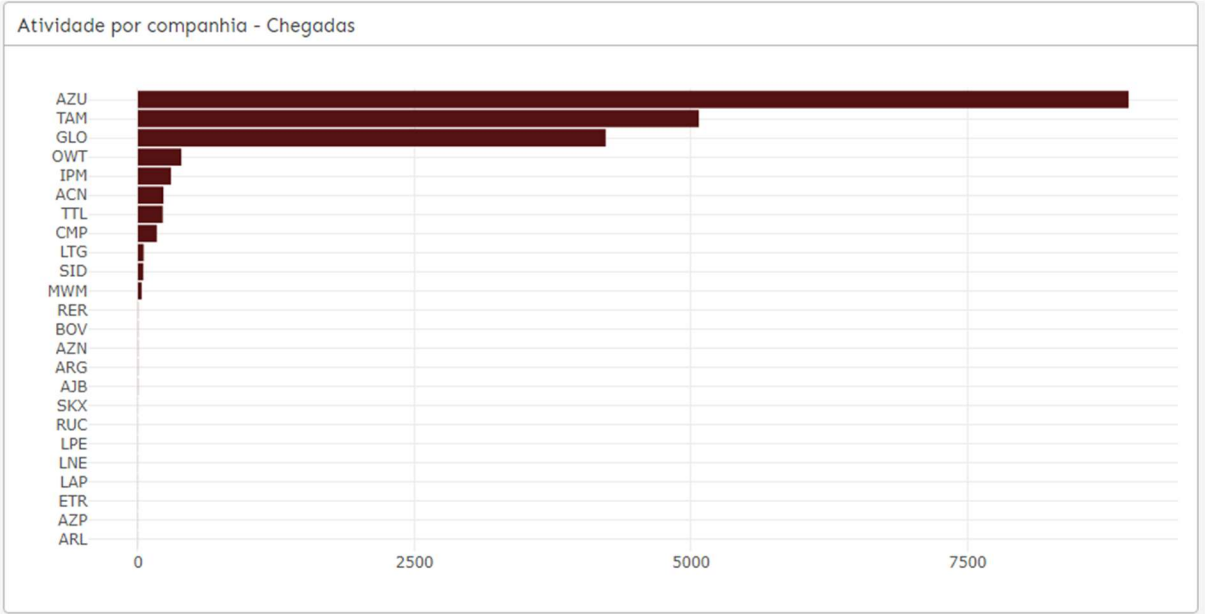
Fonte: Autoria própria

.Figura 50 - Companhias com mais vôos no Rio Grande do Sul- Partidas



Fonte: Autoria própria

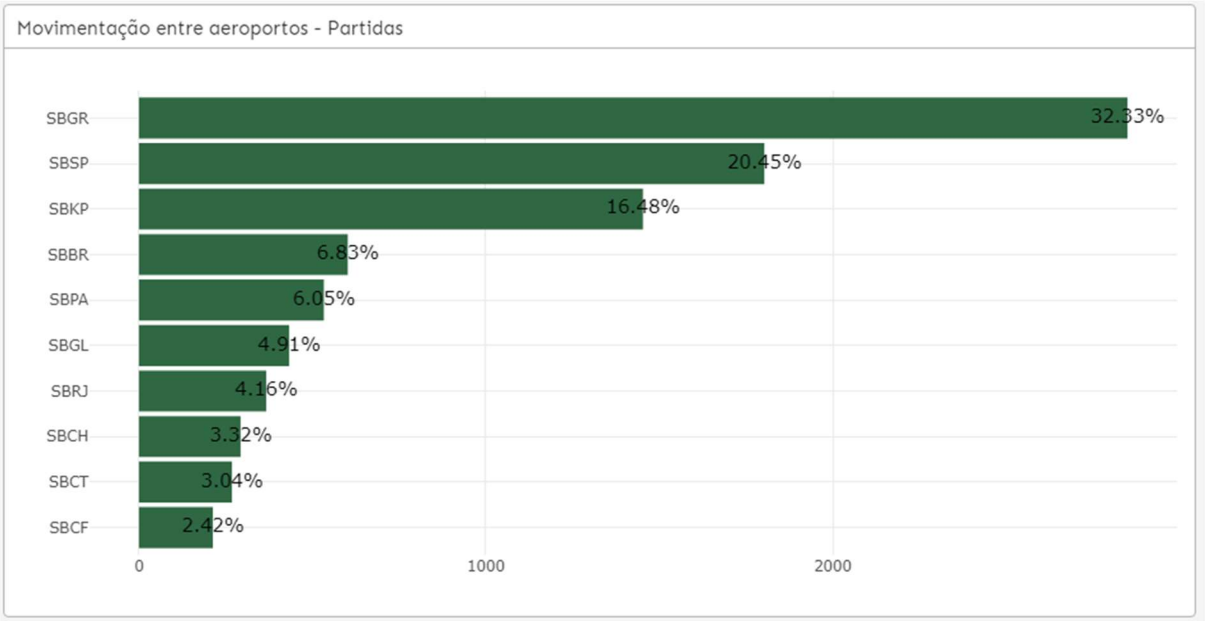
Figura 51 - Companhias com mais vôos no Rio Grande do Sul- Chegadas



Fonte: Autoria própria

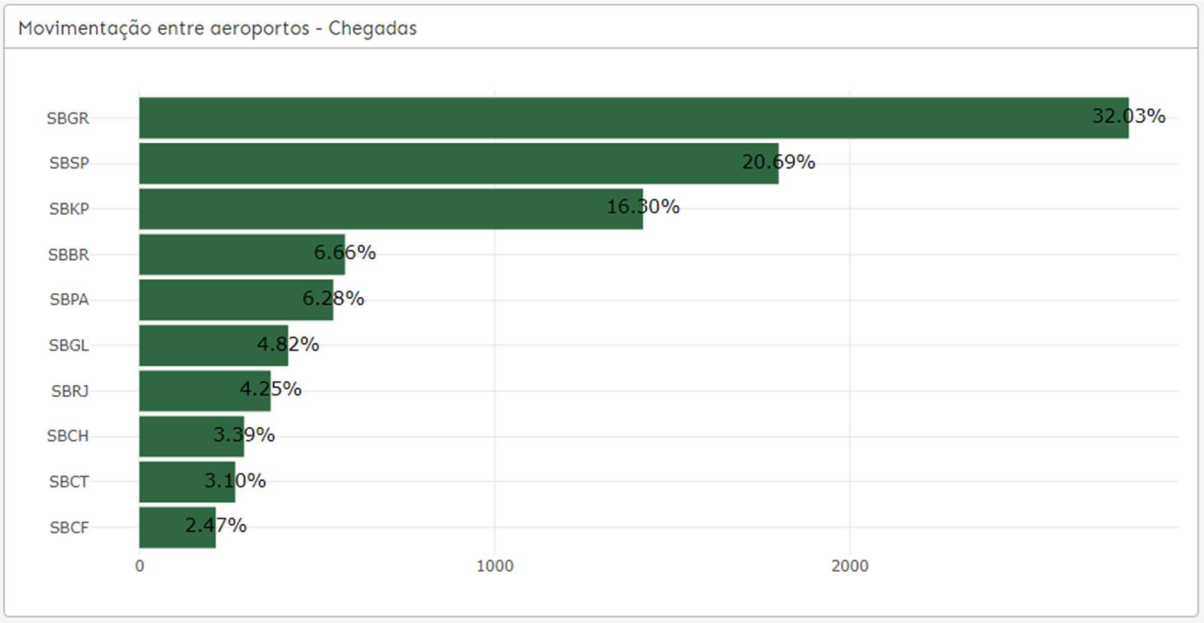
APÊNDICE E - AEROPORTOS COM MAIS MOVIMENTAÇÕES

Figura 52 - Aeroporto com mais vôos em Santa Catarina - Partidas



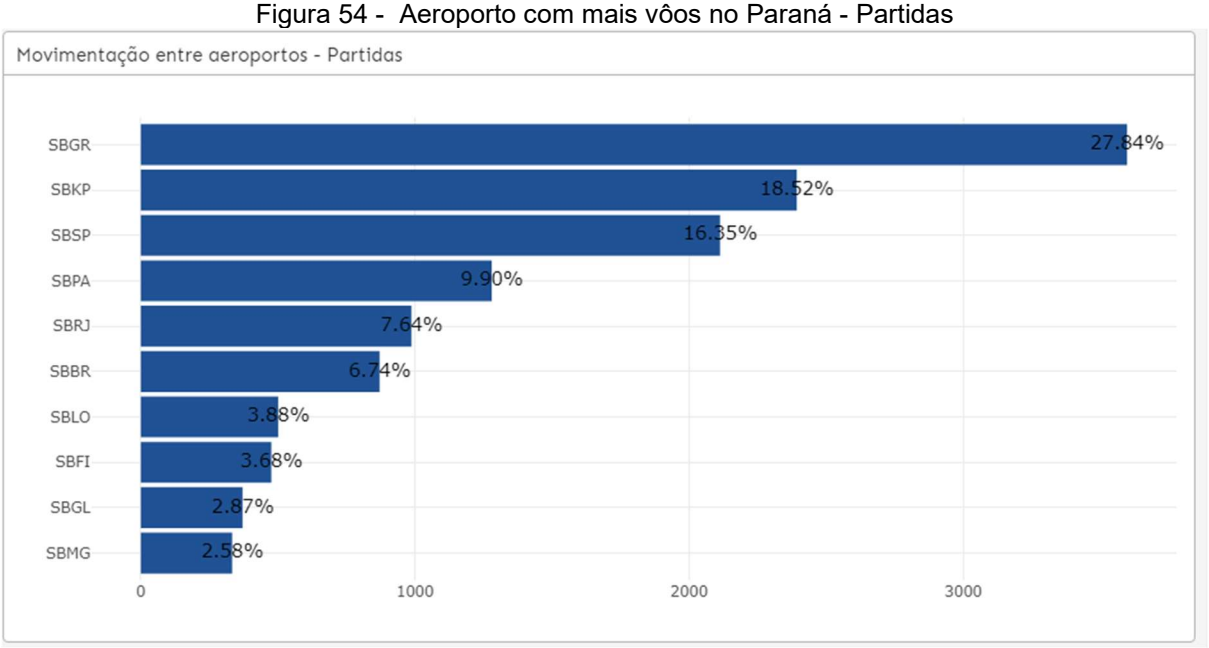
Fonte: Autoria própria

Figura 53 - Aeroporto com mais vôos em Santa Catarina - Chegadas



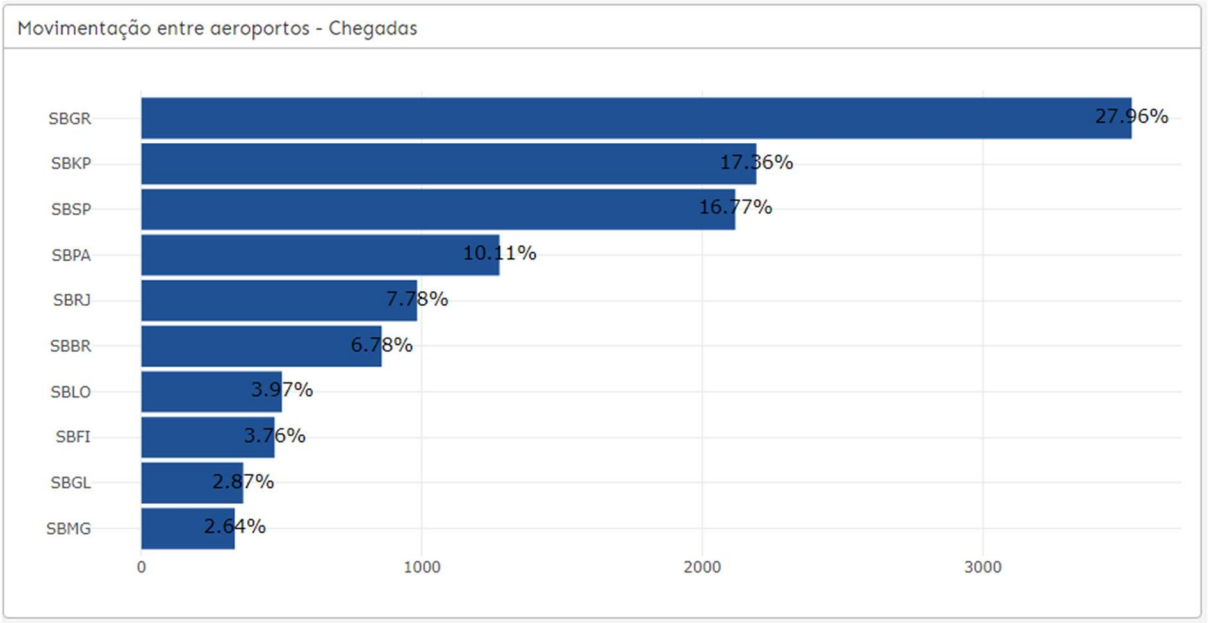
Fonte: Autoria própria

Figura 54 - Aeroporto com mais vôos no Paraná - Partidas



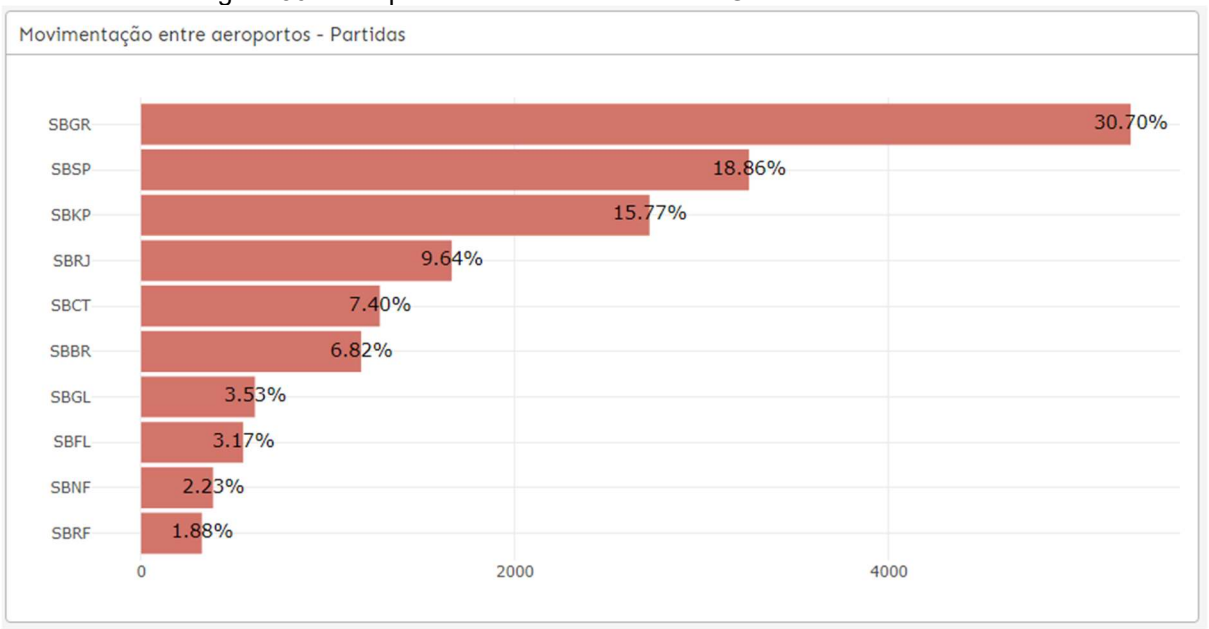
Fonte: Autoria própria

Figura 55 - Aeroporto com mais vôos no Paraná - Chegadas



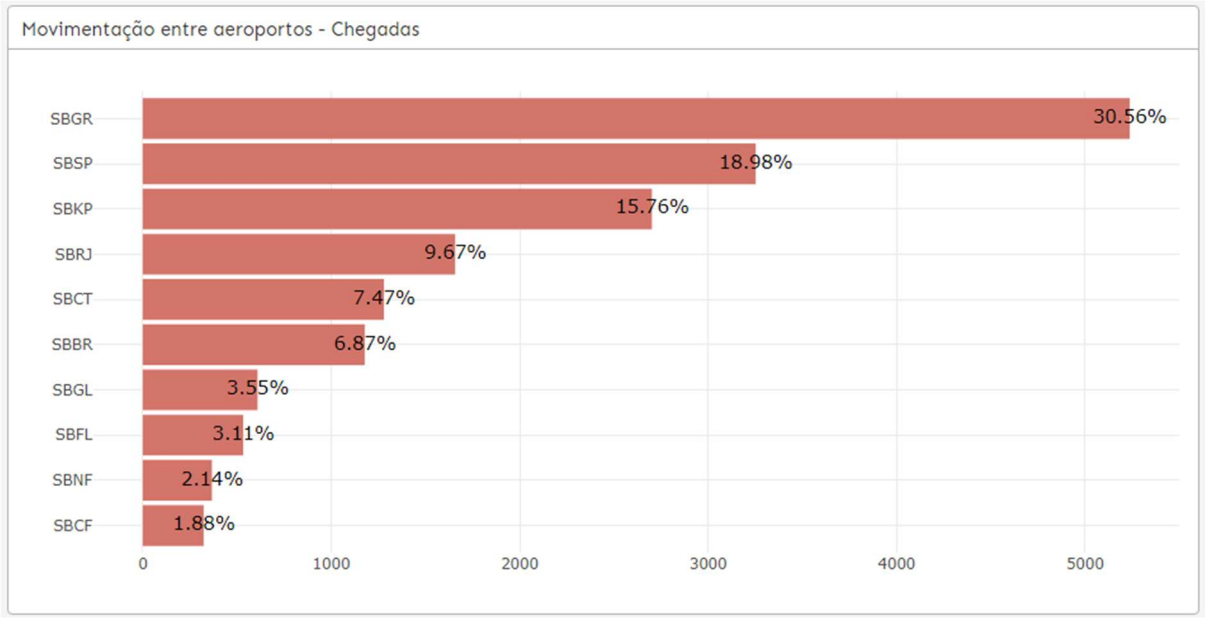
Fonte: Autoria própria

Figura 56 - Aeroporto com mais vôos no Rio Grande do Sul- Partidas



Fonte: Autoria própria

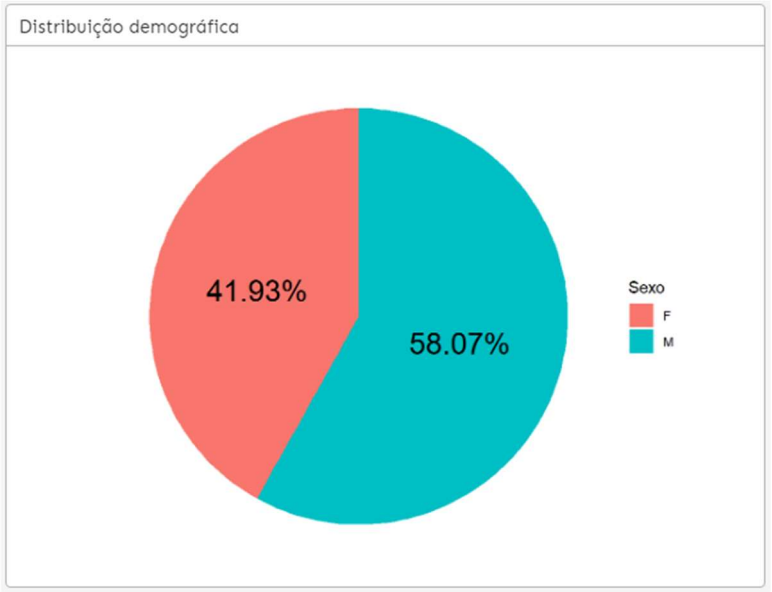
Figura 57 - Aeroporto com mais vôos no Rio Grande do Sul- Chegadas



Fonte: Autoria própria

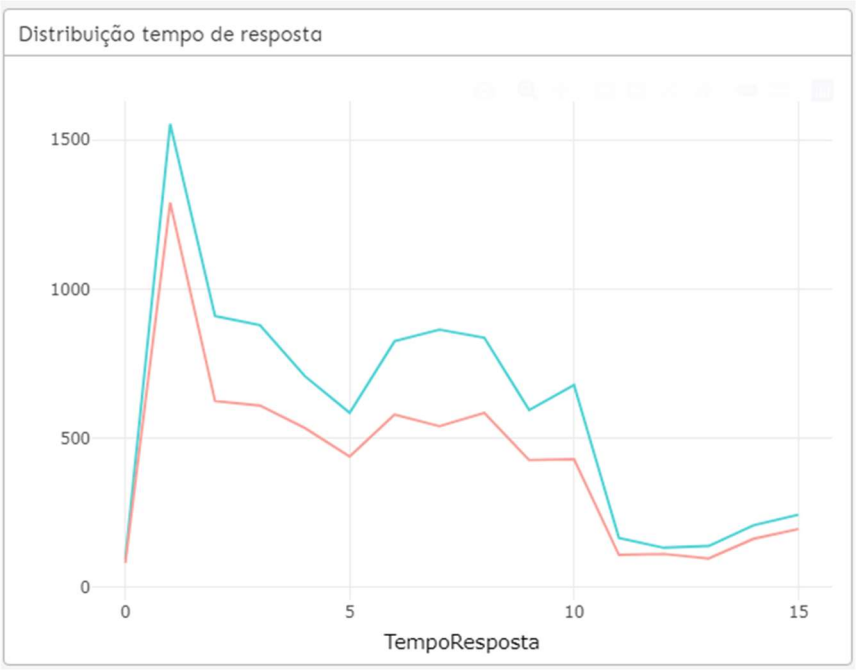
APÊNDICE F - VISUALIZAÇÕES SOCIODEMOGRÁFICAS

Figura 58 - Porcentagem relativa por sexo em Santa Catarina



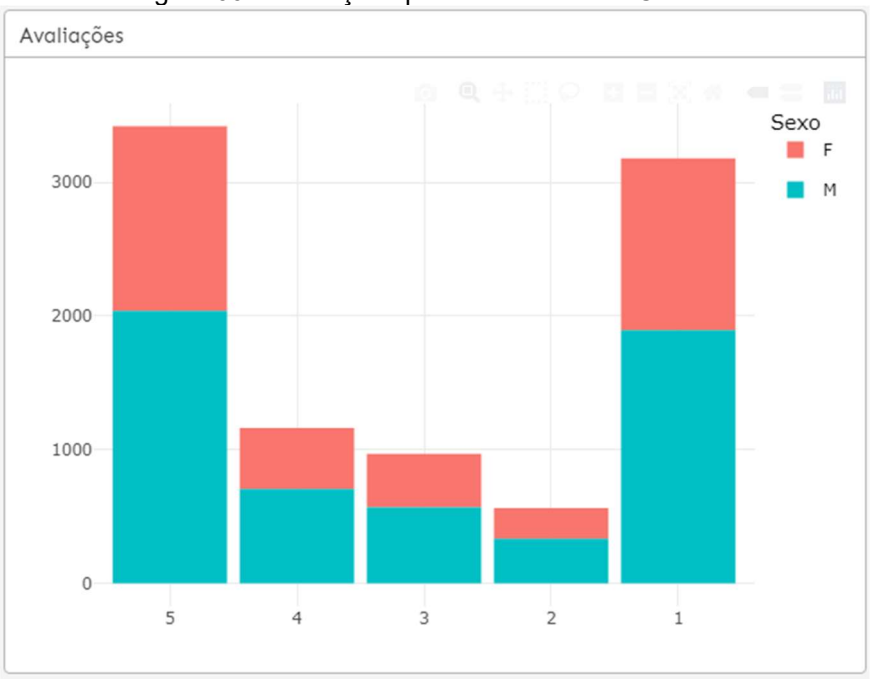
Fonte: Autoria própria

Figura 59 - Tempo de resposta por sexo em Santa Catarina



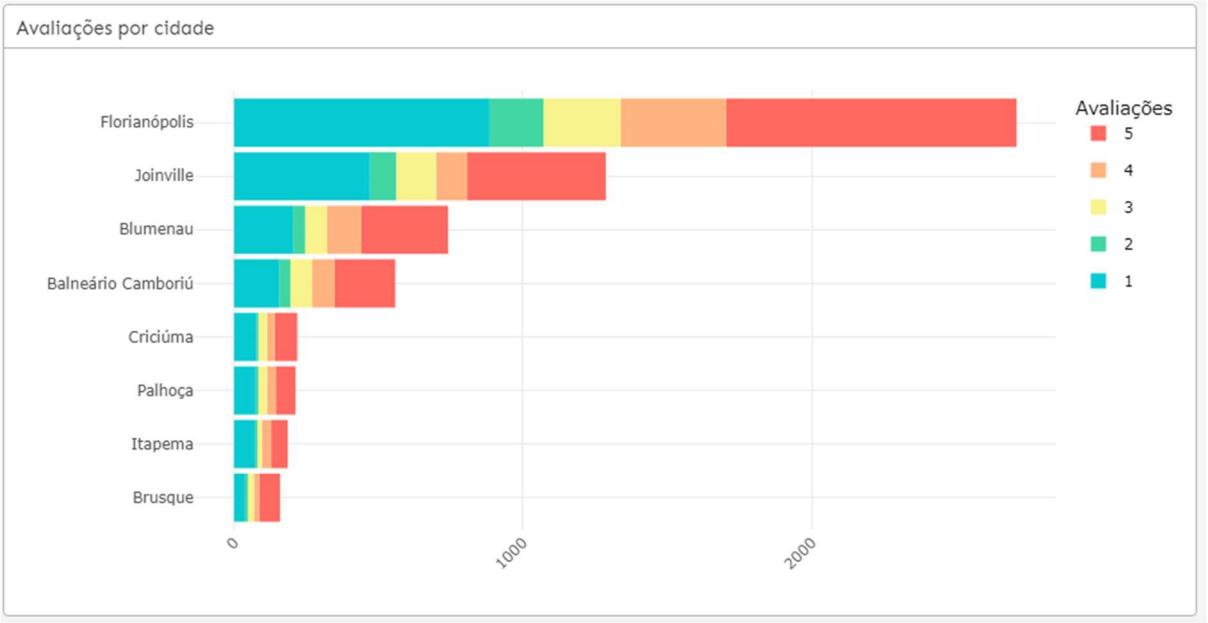
Fonte: Autoria própria

Figura 60 - Avaliações por sexo em Santa Catarina



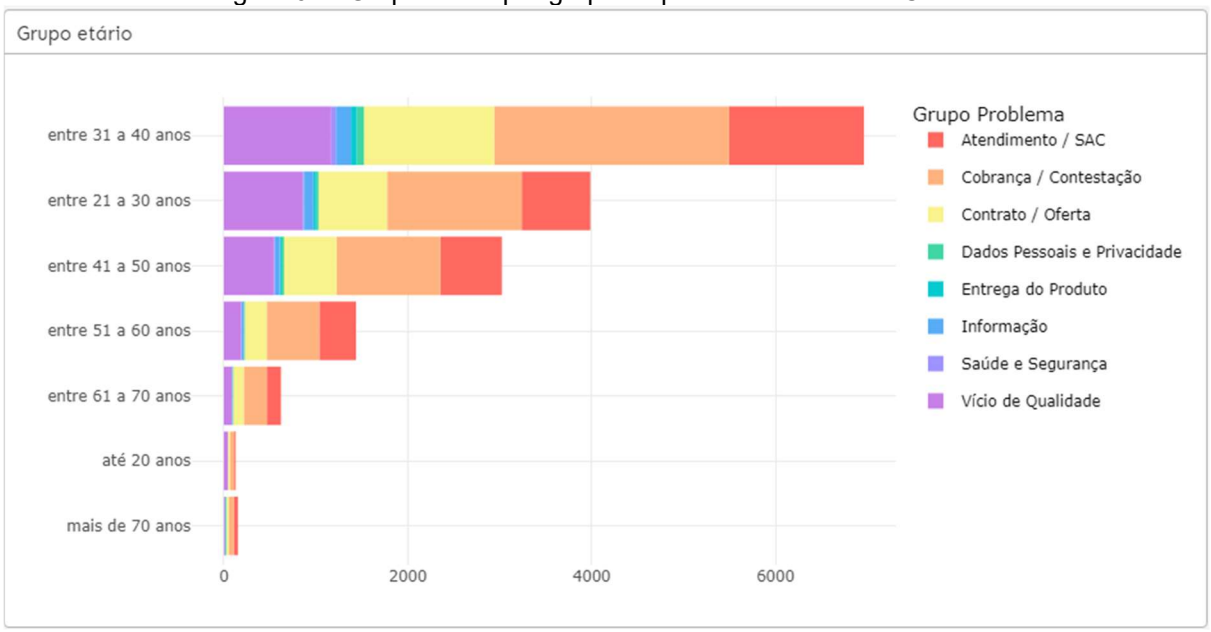
Fonte: Autoria própria

Figura 61 - Avaliações por cidades mais comuns em Santa Catarina



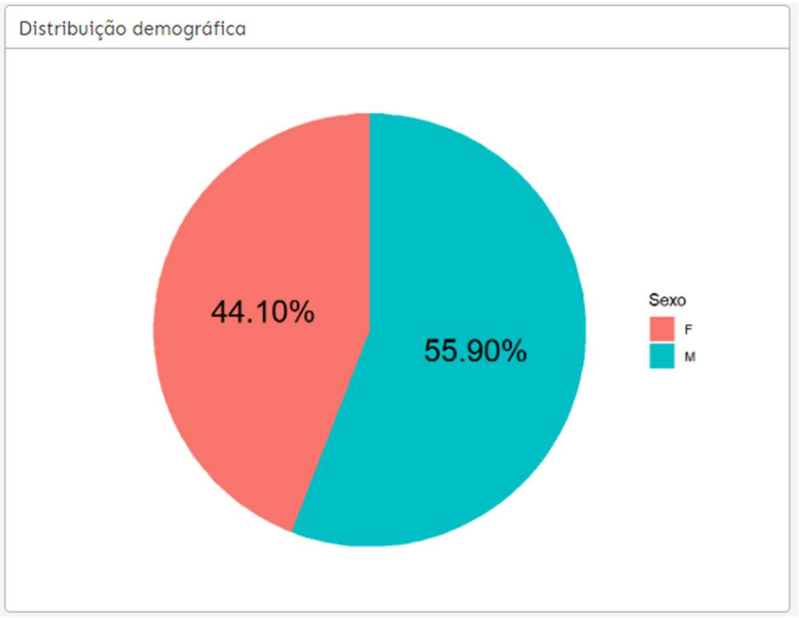
Fonte: Autoria própria

Figura 62 - Grupo etário por grupo de problema em Santa Catarina



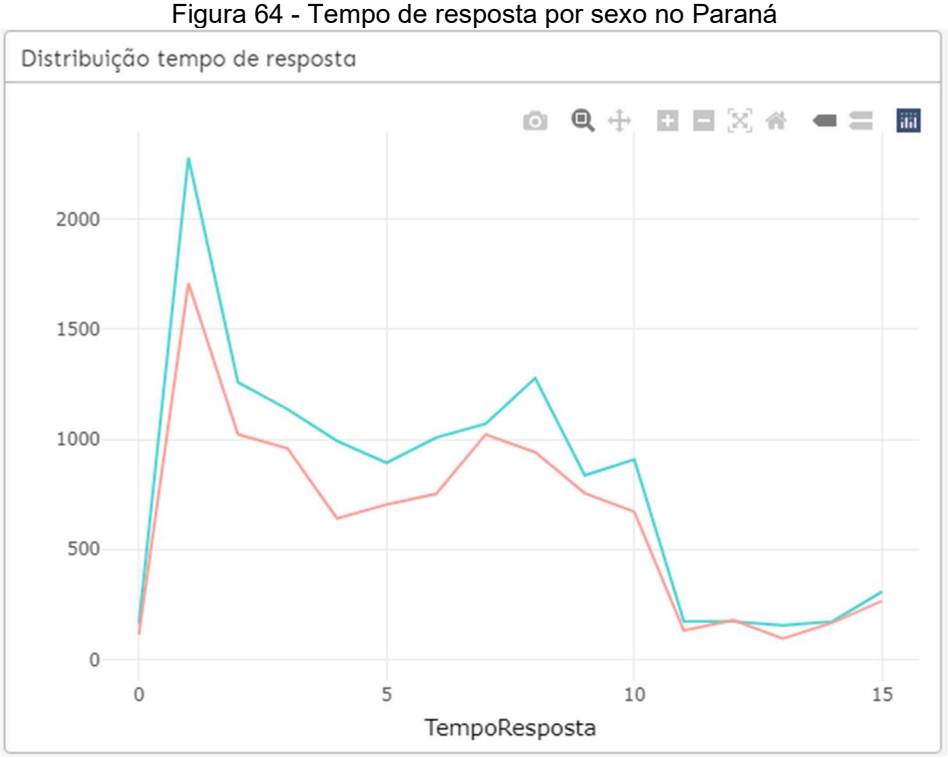
Fonte: Autoria própria

Figura 63 - Porcentagem relativa por sexo no Paraná



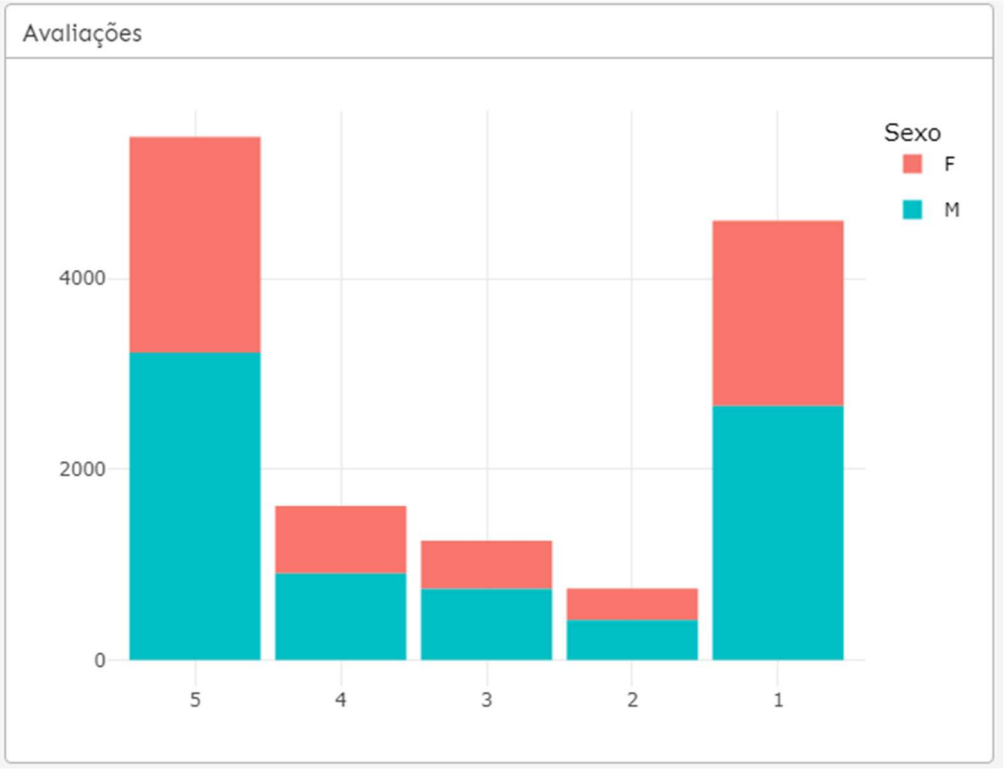
Fonte: Autoria própria

Figura 64 - Tempo de resposta por sexo no Paraná



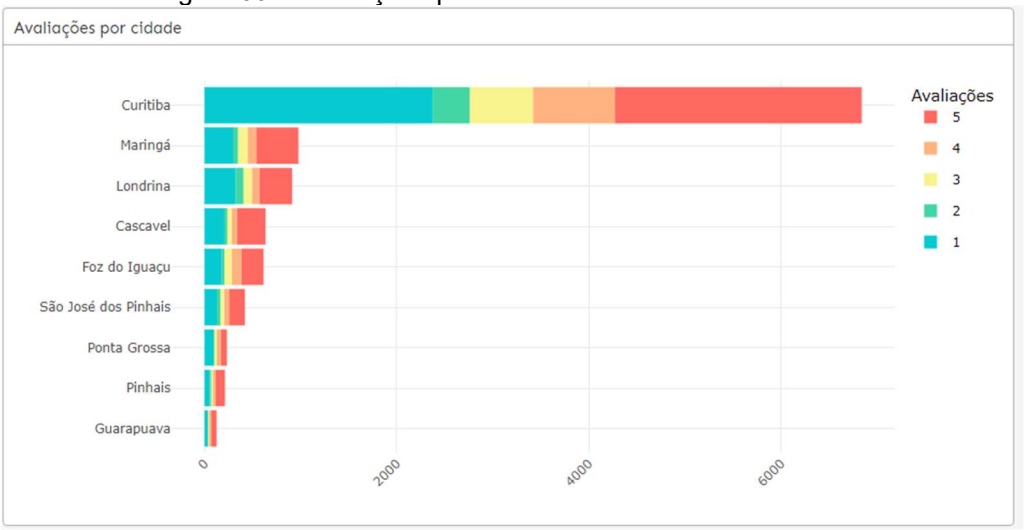
Fonte: Autoria própria

Figura 65 - Avaliações por sexo no Paraná



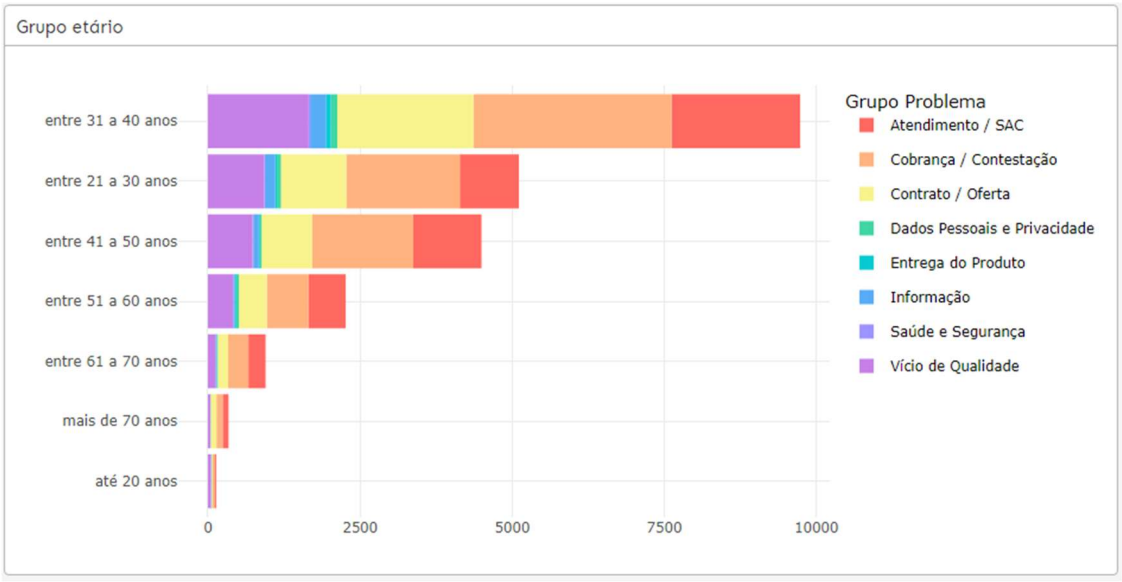
Fonte: Autoria própria

Figura 66 - Avaliações por cidades mais comuns no Paraná



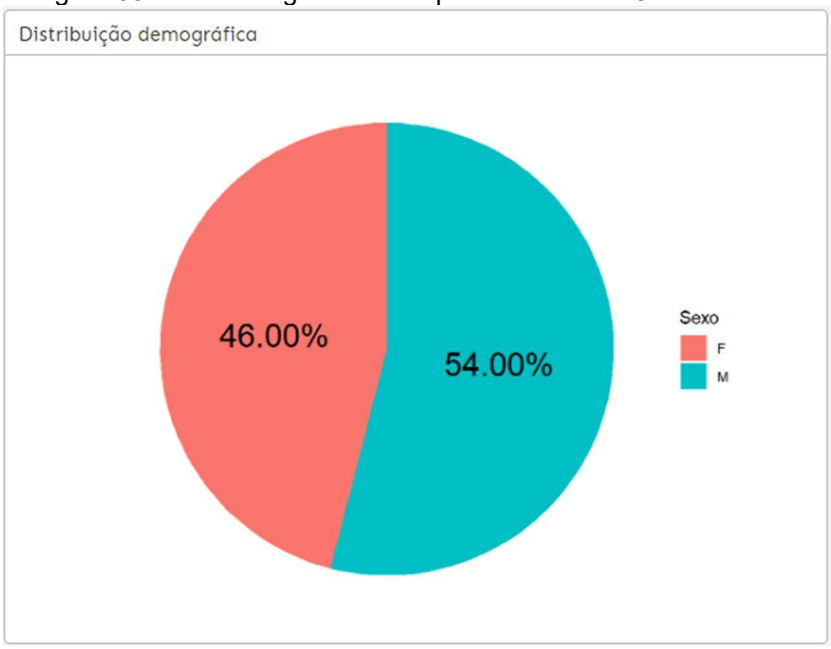
Fonte: Autoria própria

Figura 67 - Grupo etário por grupo de problema no Paraná



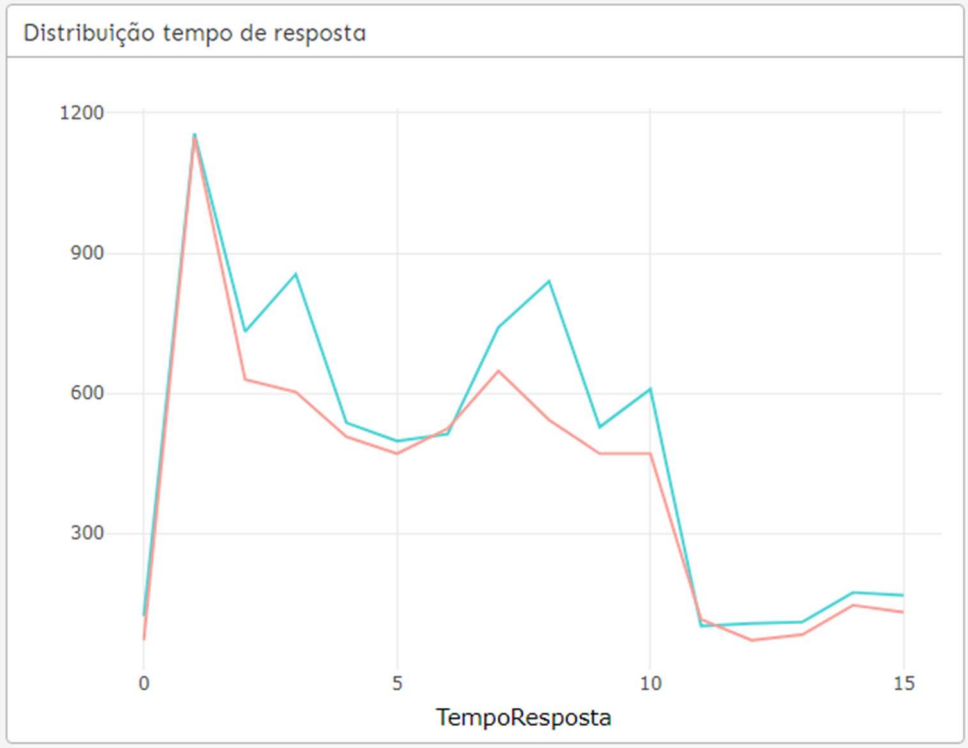
Fonte: Autoria própria

Figura 68 - Porcentagem relativa por sexo do Rio Grande no Sul



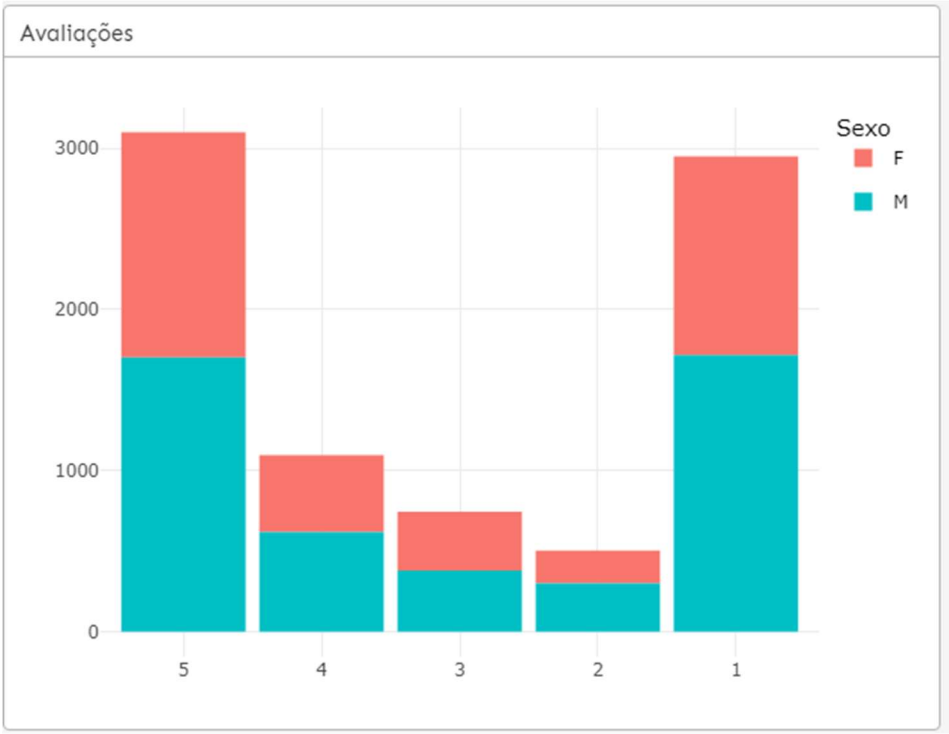
Fonte: Autoria própria

Figura 69 - Tempo de resposta por sexo do Rio Grande no Sul



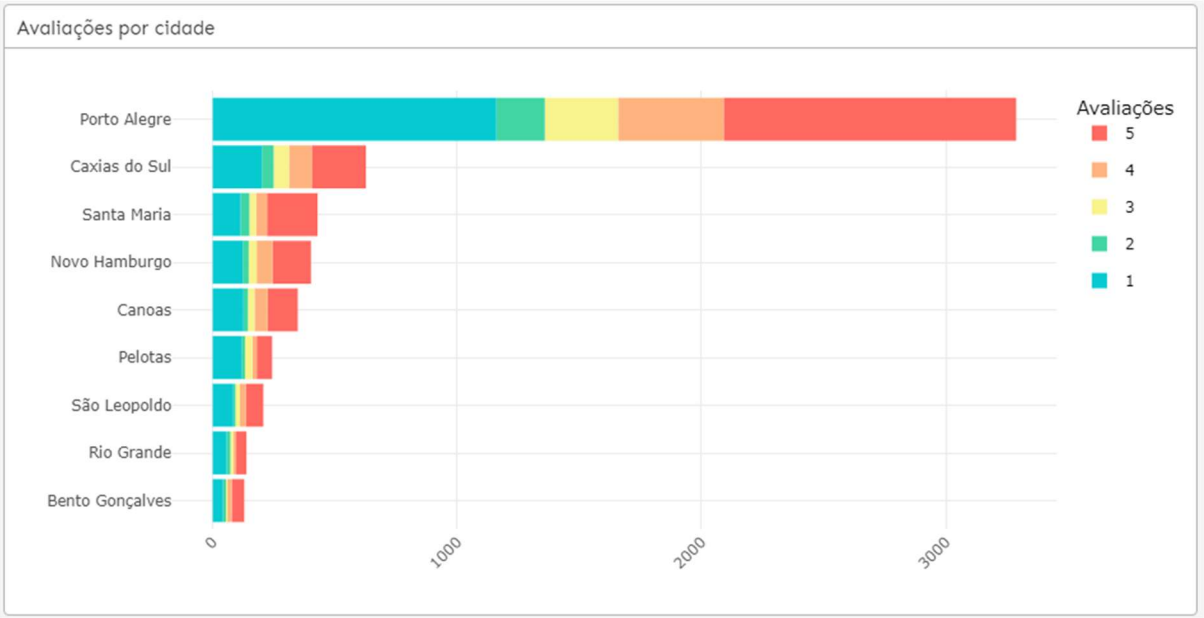
Fonte: Autoria própria

Figura 70 - Avaliações por sexo em no Rio Grande do Sul



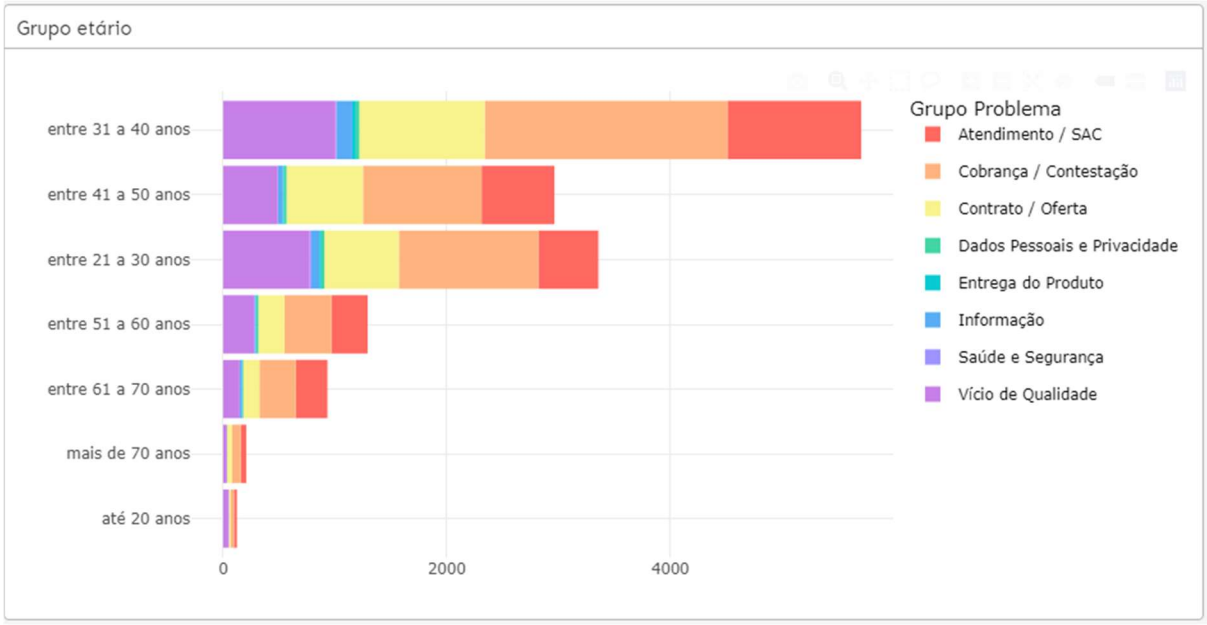
Fonte: Autoria própria

Figura 71 - Avaliações por cidades mais comuns no Rio Grande do Sul



Fonte: Autoria própria

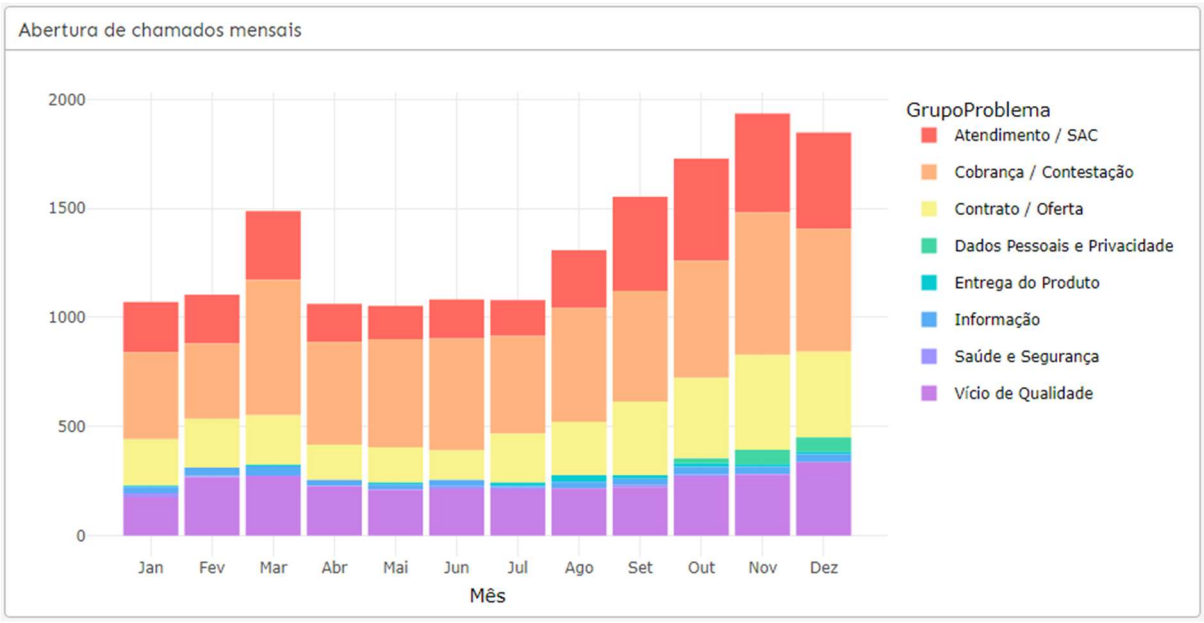
Figura 72 - Grupo etário por grupo de problema no Rio Grande do Sul



Fonte: Autoria própria

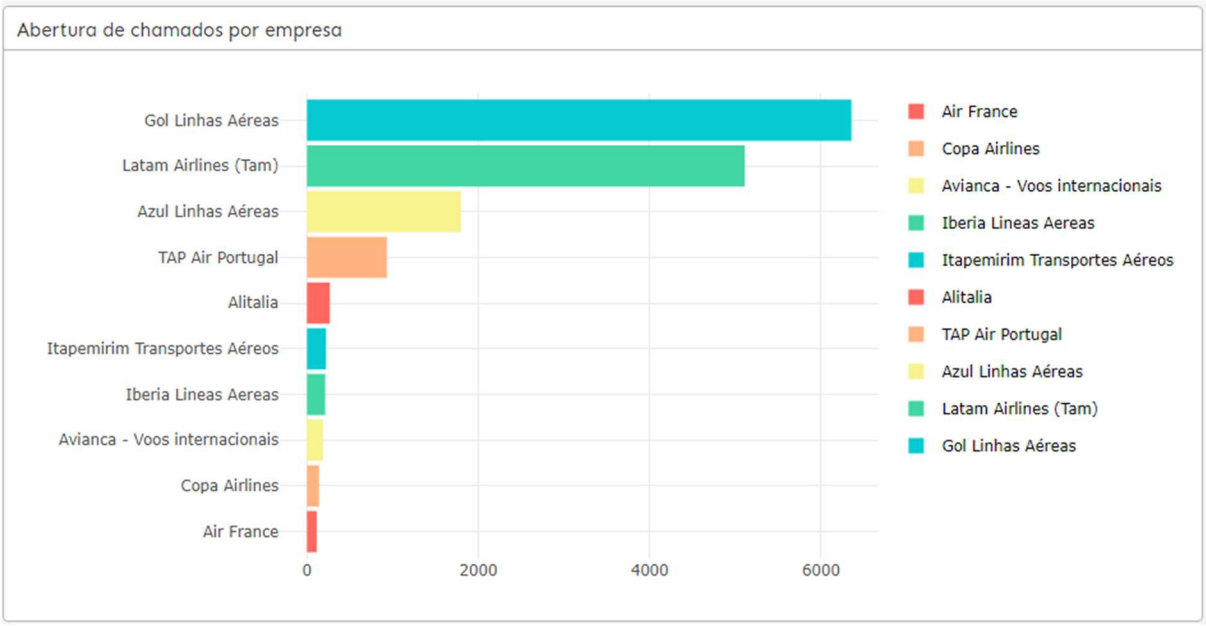
APÊNDICE G - VISUALIZAÇÕES DE ABERTURAS DE RECLAMAÇÕES

Figura 73 - Número de aberturas por problema em Santa Catarina



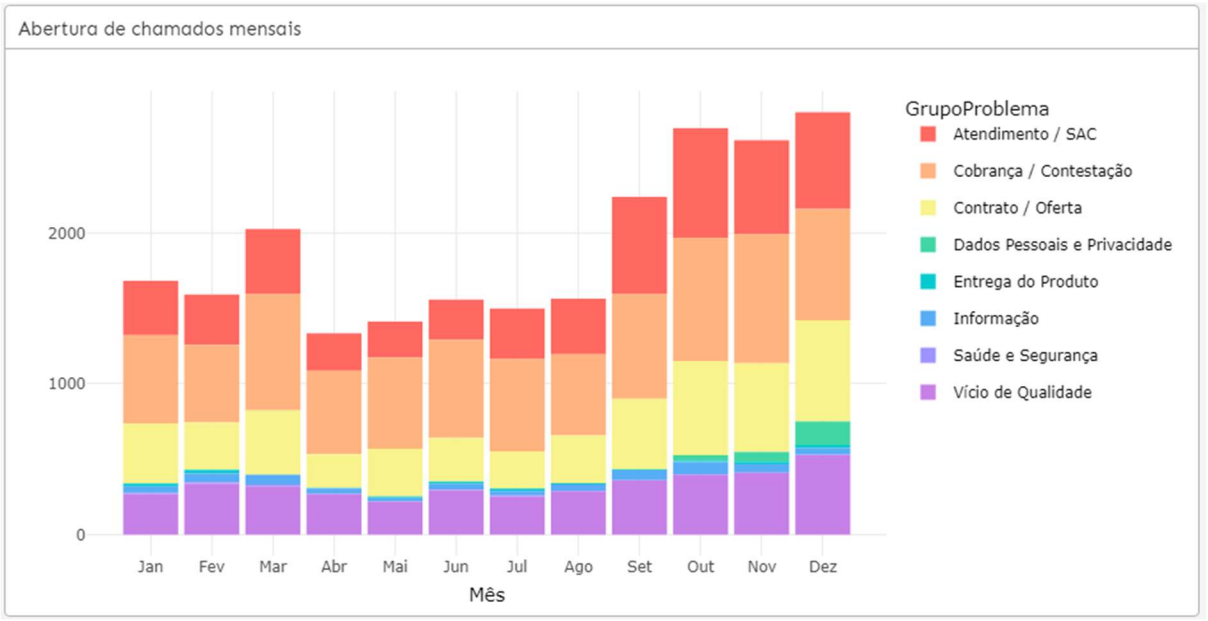
Fonte: Autoria própria

Figura 74 - Aberturas de chamado por empresa em Santa Catarina



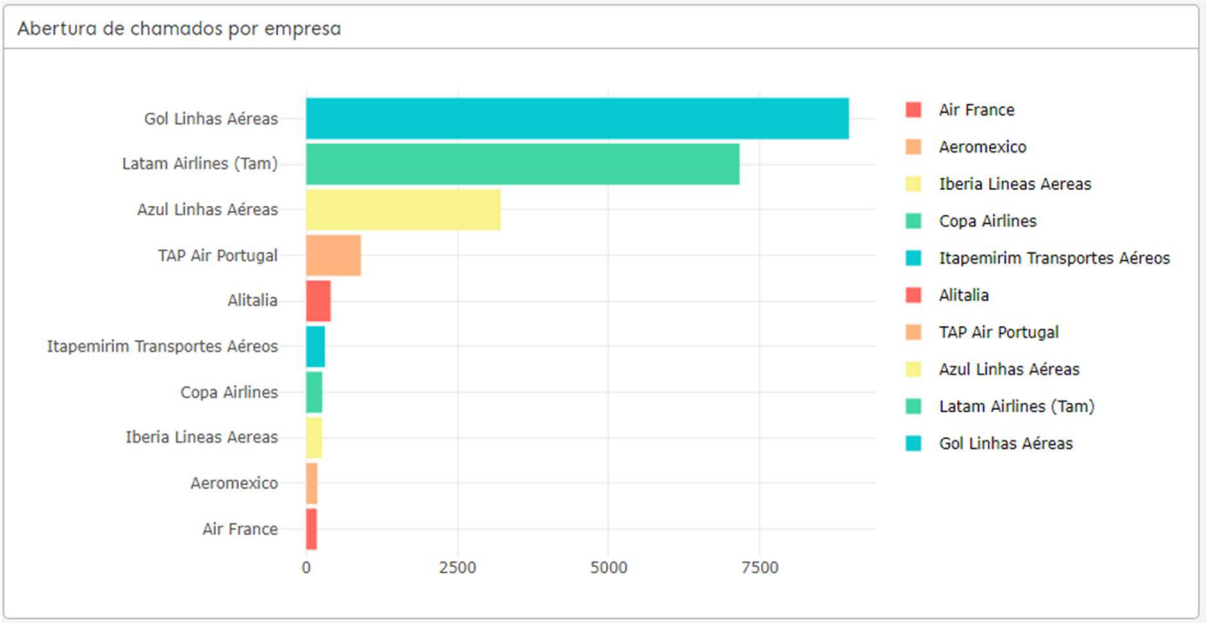
Fonte: Autoria própria

Figura 75 - Número de aberturas por problema no Paraná



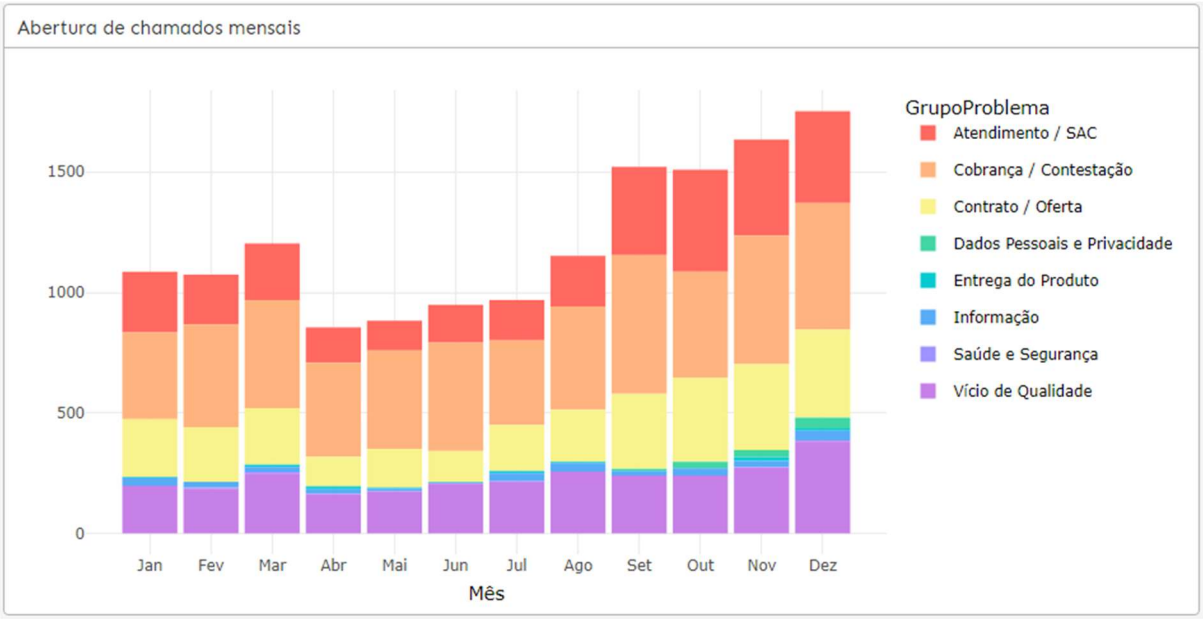
Fonte: Autoria própria

Figura 76 - Aberturas de chamado por empresa no Paraná



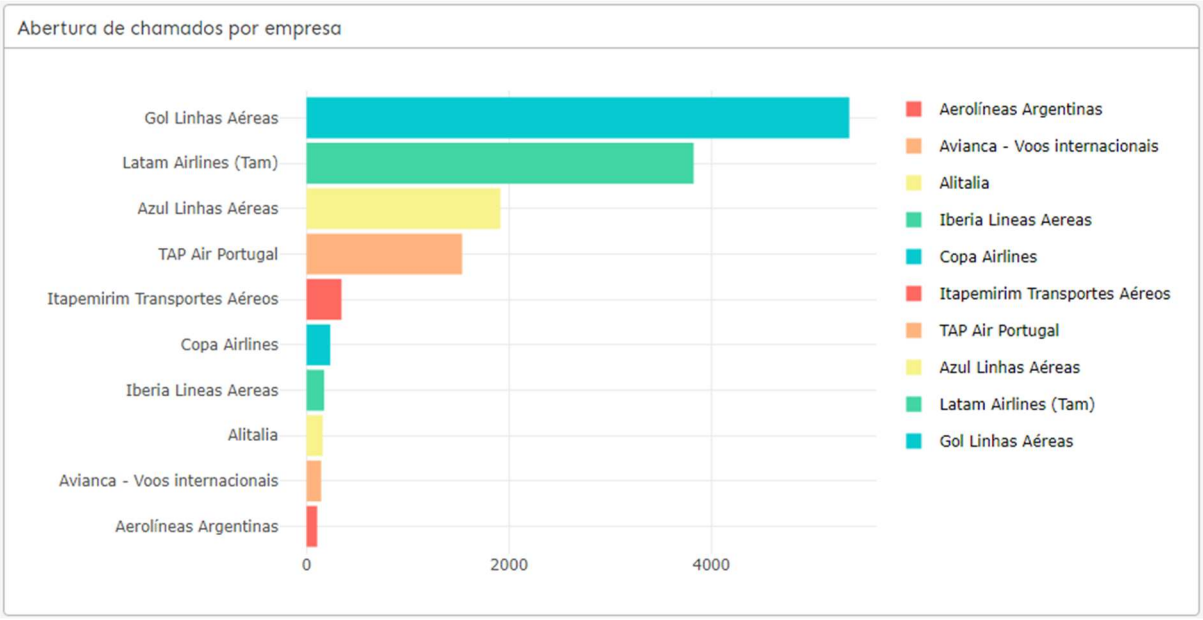
Fonte: Autoria própria

Figura 77 - Número de aberturas por problema no Rio Grande do Sul



Fonte: Autoria própria

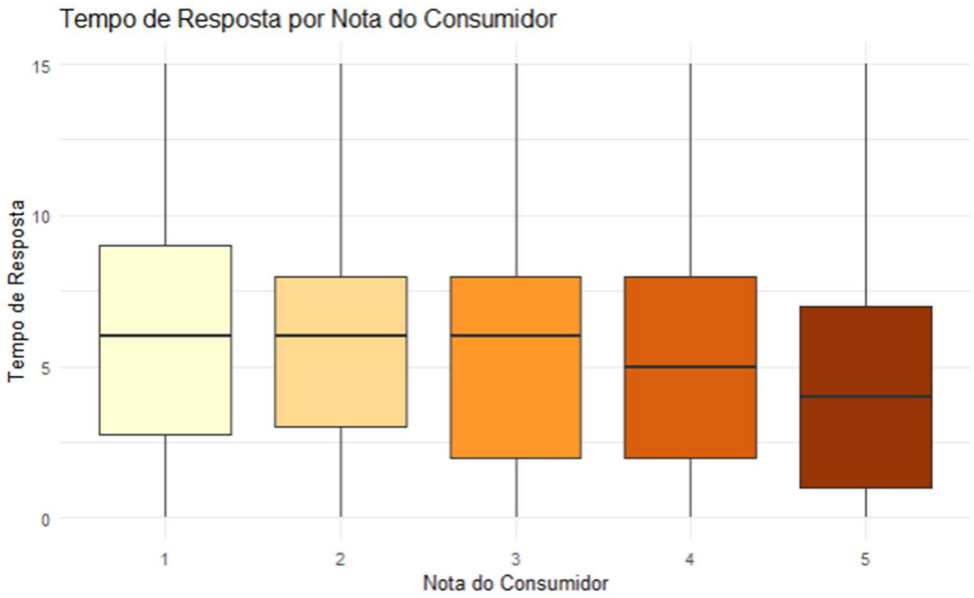
Figura 78 - Aberturas de chamado por empresa no Rio Grande do Sul



Fonte: Autoria própria

APÊNDICE H - RELAÇÃO NOTA DO CONSUMIDOR POR TEMPO DE RESPOSTA

Figura 79 - Relação nota do consumidor por tempo de resposta em Santa Catarina



Fonte: Autoria própria

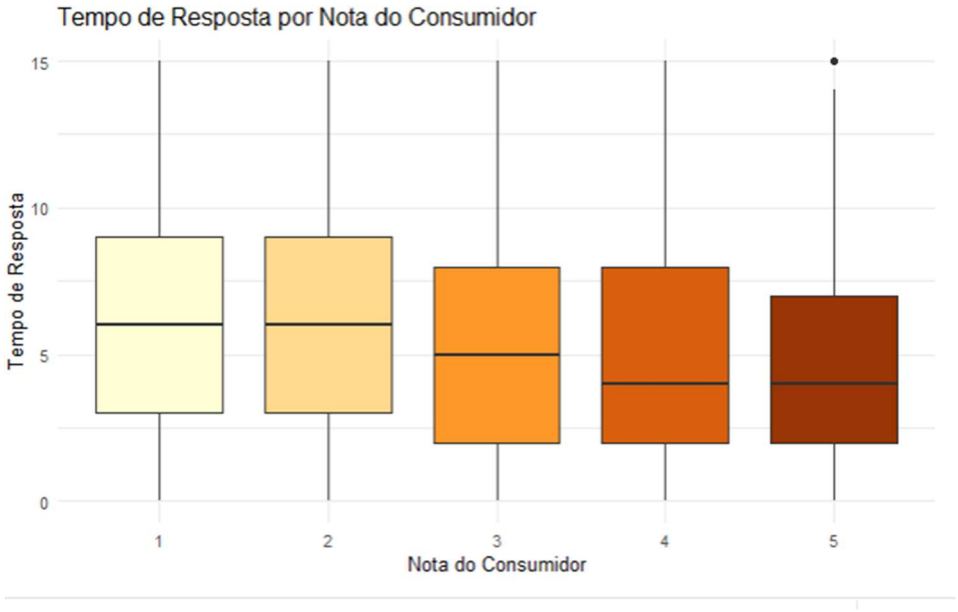
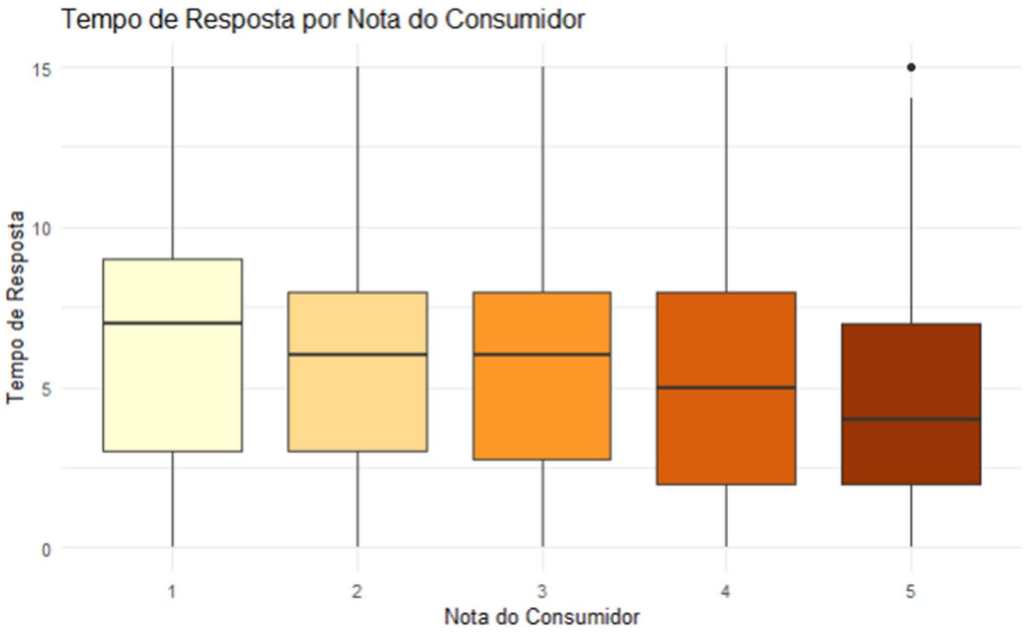


Figura 80 - Relação nota do consumidor por tempo de resposta no Paraná

Fonte: Autoria própria

Figura 81 - Relação nota do consumidor por tempo de resposta no Rio Grande do Sul



Fonte: Autoria própria

APÊNDICE I - CÓDIGO DESENVOLVIDO

```
---
title: "Dashboard Setor Aéreo Brasileiro"
output:
  flexdashboard::flex_dashboard:
    theme:
      version: 4
      bg: "#FFFFFF"
      fg: "#000"
      primary: "#999999"
      navbar-bg: "#3ADAC6"
      base_font:
        google: Prompt
      heading_font:
        google: Sen
      code_font:
        google:
          # arguments to sass::font_google()
          family: JetBrains Mono
          local: false
    orientation: rows
    vertical_layout: scroll
runtime: shiny
---

```{r setup, include=FALSE}

#importando bibliotecas

library(tidyverse)
library(flexdashboard)
library(readxl)
library(dplyr)
library(lubridate)
library(scales)
library(plotly)
library(shiny)
```
```

```
library (glue)
library(here)
library(vip)
library(kableExtra)
library(knitr)
library(forcats)
library(broom) #para manipular objetos de regressão
library(ranger)
library(tidymodels)
```

```{r, include=FALSE}
#Parte 01 - Arquivo de vôos

#Lendo arquivos

jan <-read.csv("VRA_202101.csv", sep = ";", encoding = "UTF-8")
fev <-read.csv("VRA_202102.csv", sep = ";", encoding = "UTF-8")
mar <-read.csv("VRA_202103.csv", sep = ";", encoding = "UTF-8")
abr <-read.csv("VRA_202104.csv", sep = ";", encoding = "UTF-8")
mai <-read.csv("VRA_202105.csv", sep = ";", encoding = "UTF-8")
jun <-read.csv("VRA_202106.csv", sep = ";", encoding = "UTF-8")
jul <-read.csv("VRA_202107.csv", sep = ";", encoding = "UTF-8")
ago <-read.csv("VRA_202108.csv", sep = ";", encoding = "UTF-8")
set <-read.csv("VRA_202109.csv", sep = ";", encoding = "UTF-8")
out <-read.csv("VRA_202110.csv", sep = ";", encoding = "UTF-8")
nov <-read.csv("VRA_202111.csv", sep = ";", encoding = "UTF-8")
dez <-read.csv("VRA_202112.csv", sep = ";", encoding = "UTF-8")

voos <- rbind(jan,fev,mar,abr,mai,jun,jul,ago,set,out,nov,dez)

```

```{r, include=FALSE}
#Parte 02 - Arquivo de consumidores

data <-read.csv("DadosDoConsumidor2021.csv", sep = ";")

data <- data %>% select(-DataAnálise,-HoraAnálise,-DataRecusa,-HoraRecusa,-PrazoAnaliseGestor,-InteraçãoDoGestor, -EdiçãoDeConteúdo, -AnáliseDaRecusa, -CódigoClassificadorANAC )
```

```

```{r, include=FALSE}
#VRA
#Alterando nome das colunas

voos <- rename(voos, "Empresa"= "ICAOEmpresaAérea")
voos <- rename(voos, "NumeroVoo"= "NúmeroVoo")
voos <- rename(voos, "CodAutorizacao"= "CódigoAutorizaçãoDI")
voos <- rename(voos, "CodTipoLinha"= "CódigoTipoLinha")
voos <- rename(voos, "Arpt_Origem"= "ICAOAeródromoOrigem")
voos <- rename(voos, "SituacaoVoo"= "SituaçãoVoo")
voos <- rename(voos, "CodJustificativa"= "CódigoJustificativa")
voos <- rename(voos, "Arpt_Destino"= "ICAOAeródromooDestino")

voos <- select(voos, -idvra,-CodJustificativa, -CodAutorizacao )

```

```{r, include=FALSE}
#Consumidores

data$DataAbertura <- as.Date( data$DataAbertura, format="%d/%m/%Y" )
data$DataResposta <- as.Date( data$DataResposta, format="%d/%m/%Y" )
data$DataFinalização <- as.Date( data$DataFinalização, format="%d/%m/%Y" )
data$PrazoResposta <- as.Date( data$PrazoResposta, format="%d/%m/%Y" )

data <- rename(data, "Area" = "Área")
data <- rename(data, "Regiao" = "Região")
data <- rename(data, "FaixaEtaria" = "FaixaEtária")
data <- rename(data, "MesAbertura" = "MêsAbertura")
data <- rename(data, "DataFinalizacao" = "DataFinalização")
data <- rename(data, "HoraFinalizacao" = "HoraFinalização")
data <- rename(data, "Situacao" = "Situação")
data <- rename(data, "AvaliacaoReclamacao" = "AvaliaçãoReclamação")
```

```{r, include=FALSE}
#P1
#Convertendo colunas para datas

voos$PartidaPrevista <- as.POSIXct(voos$PartidaPrevista,format = "%d/%m/%Y %H:%M")
voos$PartidaReal <- as.POSIXct(voos$PartidaReal,format = "%d/%m/%Y %H:%M")

voos$ChegadaPrevista <- as.POSIXct(voos$ChegadaPrevista,format = "%d/%m/%Y %H:%M")
voos$ChegadaReal <- as.POSIXct(voos$ChegadaReal,format = "%d/%m/%Y %H:%M")

```

```{r, include=FALSE}

#Separando data e hora

voos$PartidaPrevistaData <- as.Date(voos$PartidaPrevista)
voos$PartidaPrevistaHora <- format(voos$PartidaPrevista, "%H")

voos$PartidaRealData <- as.Date(voos$PartidaReal)
voos$PartidaRealHora <- format(voos$PartidaReal, "%H")

voos$ChegadaPrevistaData <- as.Date(voos$ChegadaPrevista)
voos$ChegadaPrevistaHora <- format(voos$ChegadaPrevista, "%H")

voos$ChegadaRealData <- as.Date(voos$ChegadaReal)
voos$ChegadaRealHora <- format(voos$ChegadaReal, "%H")

voos$ano <-
ifelse(is.na(voos$ChegadaPrevistaData),year(voos$PartidaRealData),year(voos$ChegadaPrevistaDat
a))
voos$mes <-
ifelse(is.na(voos$ChegadaPrevistaData),month(voos$PartidaRealData),month(voos$ChegadaPrevist
aData))

voos <- select(voos, -ChegadaReal,-ChegadaPrevista, -PartidaPrevista,-PartidaReal )

```

```{r, include=FALSE}

#Apresentação do Dataset P1

head(voos)

```

```{r, include=FALSE}

glimpse(voos)

```

```{r, include=FALSE}

colnames(voos)

```

```{r}
#P2

head(data)
#glimpse(data)
#colnames(data)
```

```{r, include=FALSE}
#P1
#Separando as regioes e ano de estudo

cur <- voos %>% filter(Arpt_Origem=="SBCT"  | Arpt_Destino == "SBCT")

flo <- voos %>% filter(Arpt_Origem=="SBFL"  | Arpt_Destino == "SBFL")

poa <- voos %>% filter(Arpt_Origem=="SBPA"  | Arpt_Destino == "SBPA")

cur <- cur %>% filter(ano == 2021)

flo <- flo %>% filter(ano == 2021)

```
poa <- poa %>% filter(ano == 2021)


```

```{r, include=FALSE}
#P2
cur_cons <- data %>% filter(UF=="PR")

flo_cons <- data %>% filter(UF=="SC")

poa_cons <- data %>% filter(UF=="RS")
```

```{r}
#TESTE BOXPLOT
library(RColorBrewer)

ex1 <-

dados_grafico <- subset(poa_cons, NotaDoConsumidor %in% c(1, 2, 3, 4, 5))
paleta_cores <- brewer.pal(5, "YlOrBr")



# Gerar o gráfico utilizando ggplot
ex1 <- ggplot(dados_grafico, aes(x = factor(NotaDoConsumidor), y = TempoResposta)) +
  geom_boxplot(fill = paleta_cores) +
  labs(title = "Tempo de Resposta por Nota do Consumidor",
       x = "Nota do Consumidor",
       y = "Tempo de Resposta")+ theme_minimal()


ex1
```


Vôos {data-icon="fa-plane" data-navmenu="Santa Catarina"}

Rows

Número de vôos anuais

```{r}

max_flo <- flo %>%
  filter(SituacaoVoo == "REALIZADO") %>%
  count()

valueBox(max_flo, icon = "fa-plane-arrival" , color='#6AAFBE')

```

Média de vôos mensais

```{r}
med_flo <- round(max_flo$n/12)

valueBox(med_flo, icon = 'fa-signal', color='#6ABE91')

```

Taxa de Cancelamento

```{r}

tax_flo  <- flo %>% filter(SituacaoVoo == "CANCELADO") %>%
            count()

tax_flo <- round((tax_flo$n/max_flo$n), digits=4)

tax_flo <- scales::percent(tax_flo, accuracy = 0.01)
```

```
valueBox(tax_flo, icon='fa-ban', color='#DE6363')
```

```
## Row {data-height="650"}

### Vôos ao longo do Ano - Partidas

```{r}

#voos ao longo do Ano - Partidas

f6 <- flo %>% group_by(PartidaRealData) %>%
  count() %>%
  ggplot(aes(PartidaRealData, n))+
  geom_line()+
  geom_smooth(method = 'loess', formula = 'y ~ x')+
  labs(x=NULL, y="Partidas diaárias")+
  scale_x_date(labels = date_format("%Y-%m"),
    breaks = "1 month")+
    theme_minimal()+
    theme(axis.text.x = element_text(angle = 45, hjust = 1))+
    ylim(0,125)


ggplotly(f6)


```

Vôos ao longo do Ano - Chegadas

```{r}


#voos ao longo do Ano - Chegadas

f7 <- flo %>% group_by(ChegadaRealData) %>%
  count() %>%
```
```

```
  ggplot(aes(ChegadaRealData, n))+
  geom_line()+
  geom_smooth(method = 'loess', formula = 'y ~ x')+
  theme_minimal()+
  labs(x=NULL, y="Chegadas diárias")+
  scale_x_date(labels = date_format("%Y-%m"),
    breaks = "1 month")+
    theme(axis.text.x = element_text(angle = 45, hjust = 1))+
    ylim(0,125)


ggplotly(f7)
```

Column {data-width="250"}

Cancelamentos mensais

```{r}

#Cancelamento por mes

f1<- flo %>%
  group_by(SituacaoVoo,mes) %>%
  filter(SituacaoVoo == "CANCELADO") %>%
  count() %>%
  arrange(-n) %>%
  ggplot(aes(mes, n))+
  geom_col(fill= "#1e453e")+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+
  scale_x_continuous(breaks=c(1,2,3,4,5,6,7,8,9,10,11,12),
                     label=  c("Janeiro",  "Fevereiro",  "Março",  "Abril",  "Maio",  "Junho",  "Julho",  "Agosto",
"Setembro", "Outubro", "Novembro", "Dezembro"))

ggplotly(f1)
```

Movimentação ao longo do dia - Chegadas

```{r}
#horarios comuns de chegadas

f90 <- flo %>% group_by(ChegadaRealHora) %>% count() %>% drop_na()

f9 <- f90 %>% ggplot(aes(ChegadaRealHora,n))+
  geom_col(fill = '#487c61')+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+
  ylim(0,2000)


ggplotly(f9)
```

Movimentação ao longo do dia - Partidas

```{r}
#horarios comuns de partidas
f80 <- flo %>% group_by(PartidaRealHora) %>% count() %>% drop_na()

f8 <- f80 %>% ggplot(aes(PartidaRealHora,n))+
  geom_col(fill='#487c61')+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+
  ylim(0,2000)


ggplotly(f8)
```

Column {data-table="column," data-width="250"}

Atividade por companhia - Chegadas

```{r}

#Empresa que mais chega

f2 <-   flo %>% filter(Arpt_Destino == 'SBFL') %>%
  group_by(Empresa) %>%
  count() %>%
  arrange(-n)%>%
  ggplot(aes(reorder(Empresa, n),n))+
  geom_col(fill='#182c25')+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())

ggplotly(f2)

```

Atividade por companhia - Partidas

```{r}

#Empresa que mais parte Vôos

f3 <-   flo %>% filter(Arpt_Origem == 'SBFL') %>%
  group_by(Empresa) %>%
  count() %>%
  arrange(-n)%>%
    ggplot(aes(reorder(Empresa, n),n))+
    geom_col(fill='#182c25')+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())

ggplotly(f3)

```

Column {data-width="350"}

Movimentação entre aeroportos - Chegadas

```{r}

#10 maiores chegadas

vv4 <- flo %>% filter(Arpt_Destino == 'SBFL') %>%
  group_by(Arpt_Origem) %>%
  count() %>%
  arrange(-n)


f5 <- vv4[1:10,] %>%
ggplot(aes(reorder(Arpt_Origem, n),n))+
  geom_col(fill= "#306844")+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())+
  geom_text(aes(label = scales::percent(n/sum(n))),
            position = "identity")

ggplotly(f5)

```

Movimentação entre aeroportos - Partidas

```{r}
#10 maiores destinos

v4 <- flo %>% filter(Arpt_Origem == 'SBFL') %>%
  group_by(Arpt_Destino) %>%
  count() %>%
  arrange(-n)
```

```
f4 <- v4[1:10,] %>%
ggplot(aes(reorder(Arpt_Destino, n),n))+
geom_col(fill= "#306844")+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())+
  geom_text(aes(label = scales::percent(n/sum(n))),
            position = "identity")

ggplotly(f4)
```

Consumidores {data-table="row," data-icon="fa-street-view" data-navmenu="Santa Catarina"}

Row

Taxa de resposta

```{r}

resposta_num <- flo_cons %>% group_by(Respondida) %>% filter(Respondida == 'S') %>% count()
resposta_den <- length(flo_cons$Respondida)

taxa_resposta <- round((resposta_num$n  / resposta_den), digits=4)

taxa_resposta <- scales::percent(taxa_resposta, accuracy = 0.01)

valueBox(taxa_resposta, icon='fa-check', color='#6AAFBE')

```

Média de avaliação

```{r}

nota_av <- flo_cons %>% select(NotaDoConsumidor) %>%  drop_na(NotaDoConsumidor) %>% group_by(NotaDoConsumidor)

res <- round(mean(x = nota_av$NotaDoConsumidor), digits=2)
```

```
valueBox(res, icon='fa-signal', color='#6ABE91')
```

Média do tempo de resposta

```{r}
tdr_av <- flo_cons %>% select(TempoResposta) %>% drop_na(TempoResposta) %>% group_by(TempoResposta)

res_tdr <- round(mean(x = tdr_av$TempoResposta), digits=2)

valueBox(res_tdr, icon='fa-clock', color='#DE6363')
```

Row {data-weight="500"}

Abertura de chamados mensais

```{r}
#Chamados por mês

calls <- flo_cons %>%
group_by(MesAbertura,GrupoProblema) %>%
count() %>%
ggplot(aes(MesAbertura, n, fill=GrupoProblema))+
geom_col()+
theme_minimal()+
theme (axis.title.y = element_blank())+
scale_fill_manual(values=c("#ff6961", "#ffb480", "#f8f38d", "#42d6a4", "#08cad1", "#59adf6", "#9d94ff", "#c780e8"))+
xlab("Mês")+
scale_x_continuous(breaks = 1:12, labels = c("Jan","Fev","Mar","Abr","Mai","Jun","Jul","Ago","Set","Out","Nov","Dez"))


ggplotly(calls)
```

Grupo etário

```{r}

#Faixa etaria

fx_et <- flo_cons%>%
                    group_by(FaixaEtaria, GrupoProblema) %>%
                    count()%>%
                    arrange(-n) %>%

ggplot((aes(reorder(FaixaEtaria,n),n, fill=GrupoProblema)))+
                      geom_col()+
    coord_flip()+
    theme_minimal()+
  scale_fill_manual(values=c("#ff6961",  "#ffb480",  "#f8f38d",  "#42d6a4",  "#08cad1",  "#59adf6",
"#9d94ff", "#c780e8"),  name="Grupo Problema")+
                      theme (axis.title.y = element_blank(),
            axis.title.x = element_blank())


ggplotly(fx_et)
```

Row {data-weight="500"}

Distribuição demográfica

```{r}


#Sexo #feito

genderf <- flo_cons%>% group_by(Sexo)%>% filter(Sexo== "F") %>% count()

genderm <- flo_cons%>% group_by(Sexo)%>% filter(Sexo== "M") %>% count()

gendero <- flo_cons%>% group_by(Sexo)%>% filter(Sexo== "O") %>% count()

gender = bind_rows(genderm, genderf,gendero)
```

```
chart<- gender %>% ggplot(aes(x="",y=n,fill=Sexo))+ geom_bar(width = 1, stat = "identity")+ theme_void()

pie <- chart+ coord_polar("y", start=0) + geom_text(aes(y = n / 2 + c(0, cumsum(n)[-length(n)]), label = percent(n / sum(n),accuracy=0.01)), size = 7)

pie
```

Distribuição tempo de resposta

```{r}
#tempo de resposta

time_resp_m <- flo_cons %>%
filter(Sexo == 'M') %>%
group_by(TempoResposta) %>%
count()

time_resp_f <- flo_cons %>%
filter(Sexo == 'F') %>%
group_by(TempoResposta) %>%
count()

times_resp <- ggplot()+
geom_line(data= time_resp_m, aes(x= TempoResposta, y=n), color = "#4CD2D5")+
geom_line(data= time_resp_f, aes(x= TempoResposta, y=n), color = "#FA9F98")+
theme_minimal()+
theme (axis.title.y = element_blank())


ggplotly(times_resp)
```

Avaliações

```{r}
```

```
#avaliações #feito

rating <- flo_cons %>% na.omit(flo_cons) %>%
group_by(NotaDoConsumidor, Sexo) %>%
count() %>%
arrange(-NotaDoConsumidor) %>%
ggplot((aes(reorder(NotaDoConsumidor, -NotaDoConsumidor),n, fill=Sexo)))+
                    geom_col()+
      theme_minimal()+
                    theme (axis.title.y = element_blank(),
            axis.title.x = element_blank())


ggplotly(rating)
```

Row {data-weight="500"}

Abertura de chamados por empresa

```{r}
#Empresa #feito


empr <- flo_cons%>%
count(NomeFantasia) %>%
 slice_max(order_by = n, n = 10) %>%
mutate(NomeFantasia = forcats::fct_reorder(NomeFantasia, n)) %>%
ggplot() +
  geom_col(
    aes(y = NomeFantasia, x = n, fill = NomeFantasia),
    show.legend = FALSE)+
  scale_fill_manual(values=c("#ff6961",  "#ffb480",  "#f8f38d",  "#42d6a4",  "#08cad1",  "#ff6961",
"#ffb480", "#f8f38d", "#42d6a4", "#08cad1"))+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
            axis.title.x = element_blank())
ggplotly(empr)
```

Avaliações por cidade

```{r}
cid_10 <- na.omit(flo_cons)

cid_10 <- cid_10 %>% group_by(Cidade, NotaDoConsumidor) %>% count() %>% arrange(-n)
cid_10 <- cid_10 %>%  filter(Cidade == "Blumenau" | Cidade == "Florianópolis" | Cidade == "Joinville" | Cidade == "Balneário Camboriú" | Cidade == "itajaí"| Cidade == "Brusque"| Cidade == "Itapema"| Cidade == "Palhoça"| Cidade == "Criciúma")

cid_10$NotaDoConsumidor <- factor(cid_10$NotaDoConsumidor, levels = c('5','4','3','2','1'))


cid_100 <- cid_10 %>% ggplot(aes(reorder(Cidade, n),n, fill= factor(NotaDoConsumidor)))+
  geom_col()+
  coord_flip()+
  theme_minimal()+
    theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+

    scale_fill_discrete(breaks=c('5','4','3','2','1'), labels=c('5','4','3','2','1'))+
    scale_fill_manual(values=c("#ff6961", "#ffb480", "#f8f38d", "#42d6a4", "#08cad1"),
                      name="Avaliações")

ggplotly(cid_100)



```

VÔOS {data-icon="fa-plane" data-navmenu="Paraná"}

Rows

Número de vôos anuais

```{r}
#Número total de voos
```

```
max_cur <- cur %>%
  filter(SituacaoVoo == "REALIZADO") %>%
  count()

valueBox(max_cur, icon = "fa-plane-arrival" , color='#6AAFBE')
```

Média de vôos mensais

```{r}
#Média de vôos

med_cur <- round(max_cur$n/12, digits = 0)

valueBox(med_cur,  icon = 'fa-signal', color='#6ABE91')
```

Taxa de cancelamento

```{r}
#Taxa de cancelamento



tax_cur  <- cur %>% filter(SituacaoVoo == "CANCELADO") %>%
            count()

tax_cur <- round((tax_cur$n/max_cur$n)*100, digits=2)

valueBox(tax_cur, icon='fa-ban', color='#DE6363')


```

Row {data-height="650"}

Vôos ao longo do Ano - Partidas

```{r}

#voos ao longo do Ano - Partidas

c6 <- cur %>% group_by(PartidaRealData) %>%
  count() %>%
  ggplot(aes(PartidaRealData, n))+
  geom_line()+
  geom_smooth(method = 'loess', formula = 'y ~ x')+
  labs(x=NULL, y="Partidas diaárias")+
  scale_x_date(labels = date_format("%Y-%m"),
    breaks = "1 month")+
    theme_minimal()+
    theme(axis.text.x = element_text(angle = 45, hjust = 1))+
    ylim(0,150)


ggplotly(c6)


```

Vôos ao longo do Ano - Chegadas

```{r}


#voos ao longo do Ano - Chegadas

c7 <- cur %>% group_by(ChegadaRealData) %>%
  count() %>%
  ggplot(aes(ChegadaRealData, n))+
  geom_line()+
  geom_smooth(method = 'loess', formula = 'y ~ x')+
  theme_minimal()+
  labs(x=NULL, y="Chegadas diárias")+
  scale_x_date(labels = date_format("%Y-%m"),
    breaks = "1 month")+
```

```
   theme(axis.text.x = element_text(angle = 45, hjust = 1))+
   ylim(0,150)


ggplotly(c7)
```

Column {data-width="250"}

Cancelamentos mensais

```{r}

#Cancelamento por mes

c1<- cur %>%
  group_by(SituacaoVoo,mes) %>%
  filter(SituacaoVoo == "CANCELADO") %>%
  count() %>%
  arrange(-n) %>%
  ggplot(aes(mes, n))+
  geom_col(fill= "#144272")+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+
  scale_x_continuous(breaks=c(1,2,3,4,5,6,7,8,9,10,11,12),
                     label=  c("Janeiro",  "Fevereiro",  "Março",  "Abril",  "Maio",  "Junho",  "Julho",  "Agosto",
"Setembro", "Outubro", "Novembro", "Dezembro"))

ggplotly(c1)
```

Movimentação ao longo do dia - Chegadas

```{r}
#horarios comuns de chegadas

c90 <- cur %>% group_by(ChegadaRealHora) %>% count() %>% drop_na()
```

```
c9 <- c90 %>% ggplot(aes(ChegadaRealHora,n))+
  geom_col(fill = '#02006c')+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+
  ylim(0,2500)


ggplotly(c9)
```

Movimentação ao longo do dia - Partidas

```{r}
#horarios comuns de partidas
c80 <- cur %>% group_by(PartidaRealHora) %>% count() %>% drop_na()

c8 <- c80 %>% ggplot(aes(PartidaRealHora,n))+
  geom_col(fill='#02006c')+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+
  ylim(0,2500)


ggplotly(c8)
```

Column {data-table="column," data-width="250"}

Atividade por companhia - Chegadas

```{r}

#Empresa que mais chega

c2 <-   cur %>% filter(Arpt_Destino == 'SBCT') %>%
  group_by(Empresa) %>%
```

```
  count() %>%
  arrange(-n)%>%
  ggplot(aes(reorder(Empresa, n),n))+
  geom_col(fill='#03002e')+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())

ggplotly(c2)

```

Atividade por companhia - Partidas

```{r}

#Empresa que mais parte Vôos

c3 <-   cur %>% filter(Arpt_Origem == 'SBCT') %>%
  group_by(Empresa) %>%
  count() %>%
  arrange(-n)%>%
    ggplot(aes(reorder(Empresa, n),n))+
    geom_col(fill='#03002e')+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())

ggplotly(c3)

```

Column {data-width="350"}

Movimentação entre aeroportos - Chegadas

```{r}
```

```
#10 maiores chegadas

cc4 <- cur %>% filter(Arpt_Destino == 'SBCT') %>%
  group_by(Arpt_Origem) %>%
  count() %>%
  arrange(-n)


c5 <- cc4[1:10,] %>%
ggplot(aes(reorder(Arpt_Origem, n),n))+
  geom_col(fill= "#205295")+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())+
  geom_text(aes(label = scales::percent(n/sum(n))),
            position = "identity")

ggplotly(c5)

```

Movimentação entre aeroportos - Partidas

```{r}
#10 maiores destinos

h4 <- cur %>% filter(Arpt_Origem == 'SBCT') %>%
  group_by(Arpt_Destino) %>%
  count() %>%
  arrange(-n)


c4 <- h4[1:10,] %>%
ggplot(aes(reorder(Arpt_Destino, n),n))+
geom_col(fill= "#205295")+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())+
```

```
  geom_text(aes(label = scales::percent(n/sum(n))),
            position = "identity")

ggplotly(c4)
```

CONSUMIDORES {data-table="row," data-icon="fa-street-view" data-navmenu="Paraná"}

Rows

Taxa de resposta

```{r}

resposta_num1 <- cur_cons %>% group_by(Respondida) %>% filter(Respondida == 'S') %>% count()
resposta_den1 <- length(cur_cons$Respondida)

taxa_resposta1 <- round((resposta_num1$n  / resposta_den1), digits=4)

taxa_resposta1 <- scales::percent(taxa_resposta1, accuracy = 0.01)

valueBox(taxa_resposta1, icon='fa-check', color='#6AAFBE')

```

Média de avaliação

```{r}

nota_av1 <- cur_cons %>% select(NotaDoConsumidor) %>%  drop_na(NotaDoConsumidor) %>% group_by(NotaDoConsumidor)

res1 <- round(mean(x = nota_av1$NotaDoConsumidor), digits=2)

valueBox(res1, icon='fa-signal', color='#6ABE91')

```

Média do tempo de resposta

```{r}
tdr_av1 <- cur_cons %>% select(TempoResposta) %>% drop_na(TempoResposta) %>% group_by(TempoResposta)

res_tdr1 <- round(mean(x = tdr_av1$TempoResposta), digits=2)

valueBox(res_tdr1, icon='fa-clock', color='#DE6363')
```

Row {data-weight="500"}

Abertura de chamados mensais

```{r}
#Chamados por mês

calls1 <- cur_cons %>%
group_by(MesAbertura,GrupoProblema) %>%
count() %>%
ggplot(aes(MesAbertura, n, fill=GrupoProblema))+
geom_col()+
theme_minimal()+
theme (axis.title.y = element_blank())+
scale_fill_manual(values=c("#ff6961", "#ffb480", "#f8f38d", "#42d6a4", "#08cad1", "#59adf6", "#9d94ff", "#c780e8"))+
xlab("Mês")+
scale_x_continuous(breaks = 1:12, labels = c("Jan","Fev","Mar","Abr","Mai","Jun","Jul","Ago","Set","Out","Nov","Dez"))


ggplotly(calls1)
```

Grupo etário

```{r}

#Faixa etaria

fx_et1 <- cur_cons%>%
```

```
                    group_by(FaixaEtaria, GrupoProblema) %>%
                    count()%>%
                    arrange(-n) %>%

ggplot((aes(reorder(FaixaEtaria,n),n, fill=GrupoProblema)))+
                      geom_col()+
    coord_flip()+
    theme_minimal()+
  scale_fill_manual(values=c("#ff6961",  "#ffb480",  "#f8f38d",  "#42d6a4",  "#08cad1",  "#59adf6",
"#9d94ff", "#c780e8"),  name="Grupo Problema")+
                      theme (axis.title.y = element_blank(),
            axis.title.x = element_blank())


ggplotly(fx_et1)
```

Row {data-weight="500"}

Distribuição demográfica

```{r}

#Sexo #feito

genderf1 <- cur_cons%>% group_by(Sexo)%>% filter(Sexo== "F") %>% count()

genderm1 <- cur_cons%>% group_by(Sexo)%>% filter(Sexo== "M") %>% count()

gendero1 <- cur_cons%>% group_by(Sexo)%>% filter(Sexo== "O") %>% count()

gender1 = bind_rows(genderm1, genderf1,gendero1)

chart1<- gender1 %>% ggplot(aes(x="",y=n,fill=Sexo))+ geom_bar(width = 1, stat = "identity")+
theme_void()

pie1 <- chart1+ coord_polar("y", start=0) + geom_text(aes(y = n / 2 + c(0, cumsum(n)[-length(n)]), label
= percent(n / sum(n),accuracy=0.01)), size = 7)

pie1
```

```
```

Distribuição tempo de resposta

```{r}
#tempo de resposta

time_resp_m1 <- cur_cons %>%
filter(Sexo == 'M') %>%
group_by(TempoResposta) %>%
count()

time_resp_f1 <- cur_cons %>%
filter(Sexo == 'F') %>%
group_by(TempoResposta) %>%
count()

times_resp1 <- ggplot()+
geom_line(data= time_resp_m1, aes(x= TempoResposta, y=n), color = "#4CD2D5")+
geom_line(data= time_resp_f1, aes(x= TempoResposta, y=n), color = "#FA9F98")+
theme_minimal()+
theme (axis.title.y = element_blank())


ggplotly(times_resp1)
```

Avaliações

```{r}


#avaliações #feito

rating1 <- cur_cons %>% na.omit(cur_cons) %>%
group_by(NotaDoConsumidor, Sexo) %>%
count() %>%
arrange(-NotaDoConsumidor) %>%
ggplot((aes(reorder(NotaDoConsumidor, -NotaDoConsumidor),n, fill=Sexo)))+
                        geom_col()+
```

```
      theme_minimal()+
                  theme (axis.title.y = element_blank(),
            axis.title.x = element_blank())


ggplotly(rating1)
```

Row {data-weight="500"}

Abertura de chamados por empresa

```{r}
#Empresa #feito


empr1 <- cur_cons%>%
count(NomeFantasia) %>%
 slice_max(order_by = n, n = 10) %>%
mutate(NomeFantasia = forcats::fct_reorder(NomeFantasia, n)) %>%
ggplot() +
  geom_col(
    aes(y = NomeFantasia, x = n, fill = NomeFantasia),
    show.legend = FALSE)+
  scale_fill_manual(values=c("#ff6961",    "#ffb480",    "#f8f38d",    "#42d6a4",    "#08cad1",    "#ff6961",
"#ffb480", "#f8f38d", "#42d6a4", "#08cad1"))+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
            axis.title.x = element_blank())
ggplotly(empr1)
```

Avaliações por cidade

```{r}
cid_101 <- na.omit(cur_cons)

cid_101 <- cid_101 %>% group_by(Cidade, NotaDoConsumidor) %>% count() %>% arrange(-n)
```

```
cid_101 <- cid_101 %>%  filter(Cidade == "Curitiba" | Cidade == "Maringá" | Cidade == "Londrina" | Cidade == "Cascavel" | Cidade == "Foz do Iguaçu"| Cidade == "São José dos Pinhais"| Cidade == "Ponta Grossa"| Cidade == "Pinhais"| Cidade == "Guarapuava")

cid_101$NotaDoConsumidor <- factor(cid_101$NotaDoConsumidor, levels = c('5','4','3','2','1'))


cid_1001 <- cid_101 %>% ggplot(aes(reorder(Cidade, n),n, fill= factor(NotaDoConsumidor)))+
  geom_col()+
  coord_flip()+
  theme_minimal()+
    theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+

    scale_fill_discrete(breaks=c('5','4','3','2','1'), labels=c('5','4','3','2','1'))+
    scale_fill_manual(values=c("#ff6961", "#ffb480", "#f8f38d", "#42d6a4", "#08cad1"),
                      name="Avaliações")

ggplotly(cid_1001)
```

```

# VÔOS {data-icon="fa-plane" data-navmenu="Rio Grande do Sul"}

## Rows

### Número de vôos anuais

```{r}

max_poa <- poa %>%
  filter(SituacaoVoo == "REALIZADO") %>%
  count()


valueBox(max_poa, icon = "fa-plane-arrival" , color='#6AAFBE')
```
```

```

### Média de vôos mensais

```{r}

#Média de voos totais por mes


med_poa <- round(max_poa$n/12)
valueBox(med_poa, icon = 'fa-signal', color='#6ABE91')


```

Taxa de cancelamento

```{r}


tax_poa  <- poa %>% filter(SituacaoVoo == "CANCELADO") %>%
          count()

tax_poa <- round((tax_poa$n/max_poa$n)*100, digits=2)

valueBox(tax_poa, icon='fa-ban', color='#DE6363')

```

Row {data-height="650"}

Vôos ao longo do Ano - Partidas

```{r}
```
```

```
#voos ao longo do Ano - Partidas

p6 <- poa %>% group_by(PartidaRealData) %>%
  count() %>%
  ggplot(aes(PartidaRealData, n))+
  geom_line()+
  geom_smooth(method = 'loess', formula = 'y ~ x')+
  labs(x=NULL, y="Partidas diaárias")+
  scale_x_date(labels = date_format("%Y-%m"),
    breaks = "1 month")+
    theme_minimal()+
    theme(axis.text.x = element_text(angle = 45, hjust = 1))+
    ylim(0,200)


ggplotly(p6)


```

Vôos ao longo do Ano - Chegadas

```{r}


#voos ao longo do Ano - Chegadas

p7 <- poa %>% group_by(ChegadaRealData) %>%
  count() %>%
  ggplot(aes(ChegadaRealData, n))+
  geom_line()+
  geom_smooth(method = 'loess', formula = 'y ~ x')+
  theme_minimal()+
  labs(x=NULL, y="Chegadas diárias")+
  scale_x_date(labels = date_format("%Y-%m"),
    breaks = "1 month")+
    theme(axis.text.x = element_text(angle = 45, hjust = 1))+
    ylim(0,200)
```

```
ggplotly(p7)
```

Column {data-width="250"}

Cancelamentos mensais

```{r}

#Cancelamento por mes

p1<- poa %>%
  group_by(SituacaoVoo,mes) %>%
  filter(SituacaoVoo == "CANCELADO") %>%
  count() %>%
  arrange(-n) %>%
  ggplot(aes(mes, n))+
  geom_col(fill= "#A75D5D")+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+
  scale_x_continuous(breaks=c(1,2,3,4,5,6,7,8,9,10,11,12),
                     label=  c("Janeiro",  "Fevereiro",  "Março",  "Abril",  "Maio",  "Junho",  "Julho",  "Agosto",
"Setembro", "Outubro", "Novembro", "Dezembro"))

ggplotly(p1)
```

Movimentação ao longo do dia - Chegadas

```{r}
#horarios comuns de chegadas

p90 <- poa %>% group_by(ChegadaRealHora) %>% count() %>% drop_na()

p9 <- p90 %>% ggplot(aes(ChegadaRealHora, n))+
  geom_col(fill = '#D3756B')+
  theme_minimal()+
```

```
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+
  ylim(0,3000)


ggplotly(p9)
```

Movimentação ao longo do dia - Partidas

```{r}
#horarios comuns de partidas
p80 <- poa %>% group_by(PartidaRealHora) %>% count() %>% drop_na()

p8 <- p80 %>% ggplot(aes(PartidaRealHora,n))+
  geom_col(fill='#D3756B')+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+
  ylim(0,3000)


ggplotly(p8)
```

Column {data-table="column," data-width="250"}

Atividade por companhia - Chegadas

```{r}

#Empresa que mais chega

p2 <-   poa %>% filter(Arpt_Destino == 'SBPA') %>%
  group_by(Empresa) %>%
  count() %>%
  arrange(-n)%>%
  ggplot(aes(reorder(Empresa, n),n))+
```

```
  geom_col(fill='#541212')+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())

ggplotly(p2)
```

Atividade por companhia - Partidas

```{r}

#Empresa que mais parte Vôos

p3 <-   poa %>% filter(Arpt_Origem == 'SBPA') %>%
  group_by(Empresa) %>%
  count() %>%
  arrange(-n)%>%
    ggplot(aes(reorder(Empresa, n),n))+
    geom_col(fill='#541212')+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())

ggplotly(p3)
```

Column {data-width="350"}

Movimentação entre aeroportos - Chegadas

```{r}

#10 maiores chegadas

pp4 <- poa %>% filter(Arpt_Destino == 'SBPA') %>%
```

```
  group_by(Arpt_Origem) %>%
  count() %>%
  arrange(-n)


p5 <- pp4[1:10,] %>%
ggplot(aes(reorder(Arpt_Origem, n),n))+
  geom_col(fill= "#D3756B")+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())+
  geom_text(aes(label = scales::percent(n/sum(n))),
            position = "identity")

ggplotly(p5)
```

Movimentação entre aeroportos - Partidas

```{r}
#10 maiores destinos

g4 <- poa %>% filter(Arpt_Origem == 'SBPA') %>%
  group_by(Arpt_Destino) %>%
  count() %>%
  arrange(-n)


p4 <- g4[1:10,] %>%
ggplot(aes(reorder(Arpt_Destino, n),n))+
geom_col(fill= "#D3756B")+
  coord_flip()+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
         axis.title.x = element_blank())+
  geom_text(aes(label = scales::percent(n/sum(n))),
            position = "identity")
```

```
ggplotly(p4)
```

CONSUMIDORES {data-table="row," data-icon="fa-street-view" data-navmenu="Rio Grande do Sul"}

Rows

Taxa de resposta

```{r}
resposta_num2 <- poa_cons %>% group_by(Respondida) %>% filter(Respondida == 'S') %>% count()
resposta_den2 <- length(poa_cons$Respondida)

taxa_resposta2 <- round((resposta_num2$n  / resposta_den2), digits=4)

taxa_resposta2 <- scales::percent(taxa_resposta2, accuracy = 0.01)

valueBox(taxa_resposta2, icon='fa-check', color='#6AAFBE')
```

Média de avaliação

```{r}
nota_av2 <- poa_cons %>% select(NotaDoConsumidor) %>%  drop_na(NotaDoConsumidor) %>% group_by(NotaDoConsumidor)

res2 <- round(mean(x = nota_av2$NotaDoConsumidor), digits=2)

valueBox(res2, icon='fa-signal', color='#6ABE91')
```

Média do tempo de resposta

```{r}
tdr_av2 <- poa_cons %>% select(TempoResposta) %>%  drop_na(TempoResposta) %>% group_by(TempoResposta)
```

```
res_tdr2 <- round(mean(x = tdr_av2$TempoResposta), digits=2)

valueBox(res_tdr2, icon='fa-clock', color='#DE6363')
```

Row {data-weight="500"}

Abertura de chamados mensais

```{r}
#Chamados por mês

calls2 <- poa_cons %>%
group_by(MesAbertura,GrupoProblema) %>%
count() %>%
ggplot(aes(MesAbertura, n, fill=GrupoProblema))+
geom_col()+
theme_minimal()+
theme (axis.title.y = element_blank())+
scale_fill_manual(values=c("#ff6961", "#ffb480", "#f8f38d", "#42d6a4", "#08cad1", "#59adf6", "#9d94ff",
"#c780e8"))+
xlab("Mês")+
scale_x_continuous(breaks = 1:12, labels =
c("Jan","Fev","Mar","Abr","Mai","Jun","Jul","Ago","Set","Out","Nov","Dez"))


ggplotly(calls2)
```

Grupo etário

```{r}

#Faixa etaria

fx_et2 <- poa_cons%>%
                group_by(FaixaEtaria, GrupoProblema) %>%
                count()%>%
                arrange(-n) %>%
```

```
ggplot((aes(reorder(FaixaEtaria,n),n, fill=GrupoProblema)))+
                        geom_col()+
    coord_flip()+
    theme_minimal()+
 scale_fill_manual(values=c("#ff6961",  "#ffb480",  "#f8f38d",  "#42d6a4",  "#08cad1",  "#59adf6",
"#9d94ff", "#c780e8"),  name="Grupo Problema")+
                        theme (axis.title.y = element_blank(),
            axis.title.x = element_blank())


ggplotly(fx_et2)
```

Row {data-weight="500"}

Distribuição demográfica

```{r}

#Sexo #feito

genderf2 <- poa_cons%>% group_by(Sexo)%>% filter(Sexo== "F") %>% count()

genderm2 <- poa_cons%>% group_by(Sexo)%>% filter(Sexo== "M") %>% count()

gendero2 <- poa_cons%>% group_by(Sexo)%>% filter(Sexo== "O") %>% count()

gender2 <- bind_rows(genderm2, genderf2)


chart2<- gender2 %>% ggplot(aes(x="",y=n,fill=Sexo))+ geom_bar(width = 1, stat = "identity")+
theme_void()

pie2 <- chart2+ coord_polar("y", start=0) + geom_text(aes(y = n / 2 + c(0, cumsum(n)[-length(n)]), label
= percent(n / sum(n),accuracy=0.01)), size = 7)

pie2
```

Distribuição tempo de resposta

```{r}
#tempo de resposta

time_resp_m2 <- poa_cons %>%
filter(Sexo == 'M') %>%
group_by(TempoResposta) %>%
count()

time_resp_f2 <- poa_cons %>%
filter(Sexo == 'F') %>%
group_by(TempoResposta) %>%
count()

times_resp2 <- ggplot()+
geom_line(data= time_resp_m2, aes(x= TempoResposta, y=n), color = "#4CD2D5")+
geom_line(data= time_resp_f2, aes(x= TempoResposta, y=n), color = "#FA9F98")+
theme_minimal()+
theme (axis.title.y = element_blank())


ggplotly(times_resp2)
```

Avaliações

```{r}


#avaliações #feito

limp <- poa_cons %>% filter(Sexo != 'O')

rating2 <- limp %>% na.omit(poa_cons) %>%
group_by(NotaDoConsumidor, Sexo) %>%
count() %>%
arrange(-NotaDoConsumidor) %>%
ggplot((aes(reorder(NotaDoConsumidor, -NotaDoConsumidor),n, fill=Sexo)))+
                        geom_col()+
```

```
        theme_minimal()+
                        theme (axis.title.y = element_blank(),
              axis.title.x = element_blank())


ggplotly(rating2)
```

Row {data-weight="500"}

Abertura de chamados por empresa

```{r}
#Empresa #feito


empr2 <- poa_cons%>%
count(NomeFantasia) %>%
 slice_max(order_by = n, n = 10) %>%
mutate(NomeFantasia = forcats::fct_reorder(NomeFantasia, n)) %>%
ggplot() +
  geom_col(
    aes(y = NomeFantasia, x = n, fill = NomeFantasia),
    show.legend = FALSE)+
  scale_fill_manual(values=c("#ff6961",    "#ffb480",    "#f8f38d",    "#42d6a4",    "#08cad1",    "#ff6961",
"#ffb480", "#f8f38d", "#42d6a4", "#08cad1"))+
  theme_minimal()+
  theme (axis.title.y = element_blank(),
              axis.title.x = element_blank())
ggplotly(empr2)
```

Avaliações por cidade

```{r}
cid_102 <- na.omit(poa_cons)

cid_102 <- cid_102 %>% group_by(Cidade, NotaDoConsumidor) %>% count() %>% arrange(-n)
```

```
cid_102 <- cid_102 %>%  filter(Cidade == "Porto Alegre" | Cidade == "Caxias do Sul" | Cidade == "Santa Maria" | Cidade == "Novo Hamburgo" | Cidade == "Canoas"| Cidade == "Pelotas"| Cidade == "São Leopoldo"| Cidade == "Rio Grande"| Cidade == "Bento Gonçalves")

cid_102$NotaDoConsumidor <- factor(cid_102$NotaDoConsumidor, levels = c('5','4','3','2','1'))


cid_1002 <- cid_102 %>% ggplot(aes(reorder(Cidade, n),n, fill= factor(NotaDoConsumidor)))+
  geom_col()+
  coord_flip()+
  theme_minimal()+
    theme (axis.title.y = element_blank(),
         axis.title.x = element_blank(),
         axis.text.x = element_text(angle = 45, hjust = 1))+

    scale_fill_discrete(breaks=c('5','4','3','2','1'), labels=c('5','4','3','2','1'))+
    scale_fill_manual(values=c("#ff6961", "#ffb480", "#f8f38d", "#42d6a4", "#08cad1"),
                      name="Avaliações")

ggplotly(cid_1002)


```

Modelos de previsão {data-table="row"}


```{r, include=FALSE}
#abrindo arquivo novamente e acertando os tipos, e colunas desejadas


data <- read.csv ("DadosDoConsumidor2021.csv", sep= ";")

data <- data %>%
  select(-DataAnálise, -HoraAnálise, -DataRecusa, -HoraRecusa,
         -PrazoAnaliseGestor, -InteraçãoDoGestor, -EdiçãoDeConteúdo,
         -AnáliseDaRecusa, -CódigoClassificadorANAC )
data$DataAbertura <- as.Date( data$DataAbertura, format="%d/%m/%Y" )
data$DataResposta <- as.Date( data$DataResposta, format="%d/%m/%Y" )
data$DataFinalização <- as.Date( data$DataFinalização, format="%d/%m/%Y" )
```

```
data$PrazoResposta <- as.Date( data$PrazoResposta, format="%d/%m/%Y" )
```

```{r, include=FALSE}
data <- rename(data, "Area" = "Área")
data <- rename(data, "Regiao" = "Região")
data <- rename(data, "FaixaEtaria" = "FaixaEtária")
data <- rename(data, "MesAbertura" = "MêsAbertura")
data <- rename(data, "DataFinalizacao" = "DataFinalização")
data <- rename(data, "HoraFinalizacao" = "HoraFinalização")
data <- rename(data, "Situacao" = "Situação")
data <- rename(data, "AvaliacaoReclamacao" = "AvaliaçãoReclamação")
```

```{r, include=FALSE}
notas <- c(5,4,3) #buscando prever o tempo de notas boas
idades <- c("entre 31 a 40 anos", "entre 21 a 30 anos", "entre 41 a 50 anos")

datah <- data %>%
  mutate(HA=hour(strptime(HoraAbertura, '%H:%M:%S')))%>%
  mutate(MA=minute(strptime(HoraAbertura,'%H:%M:%S')))%>%
  mutate(SA=second(strptime(HoraAbertura,'%H:%M:%S')))%>%
  mutate(HorarioAbertura=  HA*60+MA+SA/60)%>%
  mutate(HF=hour(strptime(HoraFinalizacao, '%H:%M:%S')))%>%
  mutate(MF=minute(strptime(HoraFinalizacao,'%H:%M:%S')))%>%
  mutate(SF=second(strptime(HoraFinalizacao,'%H:%M:%S')))%>%
  mutate(HorarioFinalizacao =  HF*60+MF+SF/60)

datadf <- datah %>%
  tidyr::drop_na(TempoResposta)%>%
  filter(NotaDoConsumidor == notas, FaixaEtaria == idades,
         CanalDeOrigem == "Plataforma Web" ) %>%
  select(Regiao, Gestor, HorarioAbertura, HorarioFinalizacao,
         DataAbertura, DataFinalizacao, PrazoResposta, FaixaEtaria,
         MesAbertura, NomeFantasia, GrupoProblema, TempoResposta,
         NotaDoConsumidor) %>%
  mutate(MesAbertura  = factor(MesAbertura)) %>%
  mutate(NotaDoConsumidor = factor(NotaDoConsumidor)) %>%
  mutate_if(is.character, as.factor)
```

```{r, include=FALSE}
#Dividindo os dados em treinamento e teste
set.seed(1234)
datadf_split <- initial_split(datadf, strata=TempoResposta)

datadf_train <- training(datadf_split)
datadf_test <- testing(datadf_split)
datadf_fold <- vfold_cv(datadf_train)

datadf_rec <- recipe(TempoResposta ~ ., data = datadf_train) %>%
  step_dummy(all_nominal()) %>%
  step_impute_knn(TempoResposta) %>%
  step_normalize(all_numeric())

datadf_wf <- workflow() %>%
  add_recipe(datadf_rec)
```

```{r, include=FALSE}
lm_spec <- linear_reg() %>%
  set_engine("lm")

tree_spec <- decision_tree() %>%
  set_engine("rpart") %>%
  set_mode("regression")

rf_spec <- rand_forest(trees=1000) %>%
  set_engine("ranger") %>%
  set_mode("regression")
```

```{r}

lm_rs <- datadf_wf %>%
  add_model(lm_spec) %>%
  fit_resamples(resamples=datadf_fold,metrics=metric_set(rmse, rsq, mae, mape),
```

```
            control=control_resamples(save_pred=TRUE))


collect_metrics(lm_rs)
```

```{r, include=FALSE}
tree_rs <- datadf_wf %>%
  add_model(tree_spec) %>%
  fit_resamples(resamples=datadf_fold,
            metrics=metric_set(rmse, rsq, mae, mape),
            control=control_resamples(save_pred=TRUE))

collect_metrics(tree_rs)
```

```{r}
rf_rs <- datadf_wf %>%
  add_model(rf_spec) %>%
  fit_resamples(resamples=datadf_fold,
            metrics=metric_set(rmse, rsq, mae, mape),
            control=control_resamples(save_pred=TRUE))

final <- collect_metrics(rf_rs)
```

```{r}
metricas_finais <- final %>%
  filter(.metric %in% c("mae", "rsq", "rmse"))


# Criar o gráfico de barras
grafico_metricas <- ggplot(metricas_finais, aes(x = .metric, y = mean, fill = .metric)) +
```

```
  geom_bar(stat = "identity") +
  labs(x = "Métrica", y = "Valor médio (log)", title = "Métricas Finais") +
  scale_fill_manual(values = c("#FF6F61", "#FFC154", "#6AA84F", "#5C8DBA")) +  # Definir as cores das barras
  theme_minimal()

# Exibir o gráfico
grafico_metricas
```

```
## Row

### Comparação estatística dos modelos

```{r}
graph <- collect_metrics(lm_rs) %>% mutate(modelo="lm") %>% rbind(collect_metrics(tree_rs) %>% mutate(modelo="tree")) %>% rbind(collect_metrics(rf_rs) %>% mutate(modelo="rf")) %>%
  ggplot(aes(modelo, mean, fill=modelo))+
  geom_col() +
  facet_wrap(vars(.metric
  ), scales = "free_y")+
  scale_fill_viridis_d()+
  theme_minimal()+
  theme(axis.title.y = element_blank(),
        axis.title.x = element_blank(),
        legend.position = "none")
ggplotly(graph)
```

Resultados Finais

```{r}
library(kableExtra)



modelo_final <- datadf_wf %>%
```
```

```
  add_model(rf_spec) %>%
  last_fit(datadf_split)

cm <- collect_metrics(modelo_final,
            metrics = metric_set(mae(truth, estimate), rsq(truth, estimate), rmse(truth, estimate),
mape(truth, estimate)))
#cm %>%
 # kbl() %>%
  #kable_material_dark()

cm
```

```
## Row

### Previsões
```

```{r}
graph2 <- collect_predictions(modelo_final) %>%
  ggplot(aes(TempoResposta, .pred)) +
  geom_abline(lty = 2, color = "gray50") +
  geom_point(alpha = 0.5, color = "#e32d91") +
  coord_fixed()+
  theme_minimal()

ggplotly(graph2)
```

```
### Critérios de importância
```

```{r}
library(vip)

imp_spec <- rf_spec %>%
  set_engine("ranger", importance = "permutation")

graph3 <- datadf_wf %>%
  add_model(imp_spec) %>%
```

```
  fit(datadf_train) %>%
  pull_workflow_fit() %>%
  vip(aesthetics = list(alpha = 0.8, fill = "midnightblue"))+
  theme_minimal()

ggplotly(graph3)
```

```{r}

metricas_finais <- rf_rs %>%
  filter(.metric %in% c("mae", "rsq", "rmse", "mape"))

# Criar o gráfico de barras
grafico_metricas <- ggplot(metricas_finais, aes(x = .metric, y = .estimate, fill = .metric)) +
  geom_bar(stat = "identity") +
  labs(x = "Métrica", y = "Valor médio", title = "Métricas Finais") +
  scale_fill_manual(values = c("#FF6F61", "#FFC154", "#6AA84F", "#5C8DBA")) +  # Definir as cores das barras
  theme_minimal()

# Exibir o gráfico
grafico_metricas
`
```