Universidade Presbiteriana Mackenzie

ANDREI SOUZA DE OLIVEIRA – TIA: 22520600

DANIELE DOS SANTOS ROSA – TIA: 22510631

GABRIELA OHASHI DE SOUZA – TIA: 22521097

MARINA OHASHI DE SOUZA – TIA: 22520971

MIGUEL MAURÍCIO TADEU PITALI DA SILVA – TIA: 22507310

PROJETO APLICADO II: BOA VIAGEM

São Paulo

2023

SUMÁRIO

RESUMO

Neste projeto de análise de dados para a agência de viagens "Boa Viagem" iremos aplicar uma variedade de técnicas de análise de dados para extrair informações valiosas dos registros de reservas de passagens aéreas. Essas técnicas incluem análise descritiva para compreender a distribuição de características-chave, segmentação de clientes para personalização de ofertas, análise de tendências temporais para previsão de demanda sazonal, previsão de demanda por destino para alocação eficaz de recursos e análise de desempenho de voos para identificar áreas de melhoria nas operações.

Essas análises nos permitirão tomar decisões mais informadas, melhorar a eficiência operacional e, em última instância, aprimorar a experiência do cliente da "Boa Viagem" no setor de viagens e turismo.

INTRODUÇÃO

A agência de viagens "Boa Viagem" está empenhada em aprimorar seus serviços e aumentar a satisfação do cliente por meio da análise de dados relacionados às reservas de passagens aéreas. Neste projeto, aplicaremos uma variedade de técnicas de análise de dados para atingir esses objetivos.

A "Boa Viagem" atua no setor de turismo, oferecendo serviços de viagens personalizadas e reserva de passagens aéreas. Possui um conjunto de dados detalhado que abrange informações sobre passageiros, voos, destinos e comportamento de reserva. O principal objetivo deste projeto é aproveitar esses dados para melhorar a tomada de decisões estratégicas. Isso inclui segmentar os clientes com base em características específicas, prever a demanda por destinos, analisar tendências temporais e avaliar o desempenho dos voos.

A análise de dados desempenha um papel fundamental na compreensão do comportamento dos clientes, na identificação de oportunidades de negócios e no aprimoramento da eficiência operacional. Por meio dela, a "Boa Viagem" poderá tomar decisões mais informadas e oferecer experiências de viagem personalizadas aos clientes.

Neste projeto, abordaremos técnicas como análise descritiva, segmentação de clientes, análise de tendências temporais, previsão de demanda por destino e análise de desempenho de voos. Acreditamos que essas análises fornecerão insights valiosos para ajudar a "Boa Viagem" a aprimorar seus serviços, aumentar a satisfação do cliente e prosperar no competitivo mercado de viagens e turismo.

Empresa Escolhida: A empresa escolhida é a "Boa Viagem", uma agência de viagens e turismo fictícia que atua globalmente, oferecendo pacotes de viagens personalizados e serviços de reserva de passagens aéreas.

Área de Atuação: A "Boa Viagem" atua no setor de turismo, fornecendo serviços de viagens e turismo, incluindo a organização de viagens de lazer e negócios, bem como a reserva de passagens aéreas.

Dados Disponíveis: Os dados disponíveis consistem em informações detalhadas sobre passagens aéreas, incluindo o seguinte conjunto de atributos:

- ID do Passageiro
- Primeiro Nome
- Último Nome
- Gênero
- Idade
- Nacionalidade
- Nome do Aeroporto de Partida
- País do Aeroporto de Partida
- Código do Aeroporto de Partida

- Nome do País
- Continentes Envolvidos na Rota de Voo
- Data de Partida
- Aeroporto de Destino
- Nome do Piloto
- Status do Voo

Conjunto de dados escolhido

Airline dataset, disponível em:

https://www.kaggle.com/datasets/iamsouravbanerjee/airline-dataset

Apresentação disponível em:

https://youtu.be/y10MDfolarI

OBJETIVOS E METAS

Os objetivos do projeto são os seguintes:

Análise de Demanda por Destino: Analisar a demanda por destinos específicos com base nas reservas de passagens aéreas, identificando os destinos mais populares e os segmentos de mercado mais relevantes.

Segmentação de Clientes: Segmentar os clientes com base em critérios como idade, gênero, nacionalidade e preferências de viagem para oferecer pacotes de viagens mais personalizados.

Análise de Desempenho de Voos: Avaliar o desempenho de voos com base no status do voo e identificar áreas de melhoria na eficiência das operações de voo.

Previsão de Tendências de Viagem: Prever tendências futuras de viagem com base em análises históricas, ajudando a empresa a planejar e adaptar seus serviços.

METODOLOGIA

Foi utilizada a base de dados "*Airline Dataset"*, disponível no *Kaggle*, na qual é apresentado um conjunto de dados que relacionam operações aéreas em escala global. No contexto deste trabalho, entende-se por dados as principais características dos passageiros (gênero, idade, nacionalidade) e dos voos (país, continente e status).

Para realizar a análise de dados, criar modelos de aprendizado de máquina e avaliar seu desempenho, foram utilizadas as seguintes bibliotecas Python:

- *pandas*: para realização de operações como leitura, filtragem, agregação e transformação de dados;
- *numpy*: para operações numéricas, incluindo manipulação de arrays multidimensionais;
- *matplotlib*: para criação gráficos e visualizações de dados, ajudando na interpretação e comunicação de resultados;
- *scikit-learn* (também conhecida como *sklearn*): para implementações eficientes de uma ampla gama de algoritmos de aprendizado de máquina, incluindo regressão linear, o *K-means* dentre outros.

Para o desenvolvimento de um modelo de aprendizado de máquina supervisionado, foi utilizada a regressão linear, sendo necessária a utilização de técnica de pré-processamento "*LabelEncoder*", que transforma variáveis categóricas em variáveis numéricas.

FUNDAMENTAÇÃO TEÓRICA

ANÁLISE EXPLORATÓRIA DOS DADOS

A Análise Exploratória de Dados (EDA, Exploratory Data Analysis) é usada para analisar e investigar conjuntos de dados e resumir suas características principais, empregando métodos quantitativos e de visualização dos dados. Dentro do modelo CRISP-DM, a EDA compreende as fases de Entendimento e Preparação dos Dados, incluindo também a fase de Modelagem.

O modelo CRISP-DM (Cross-Industry Standard Process for Data Mining) é uma metodologia abrangente de mineração de dados e um modelo de processo que oferece modelo para a execução de um projeto de mineração de dados. Ele é estruturado em seis etapas distintas: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação (Shearer, 2000).

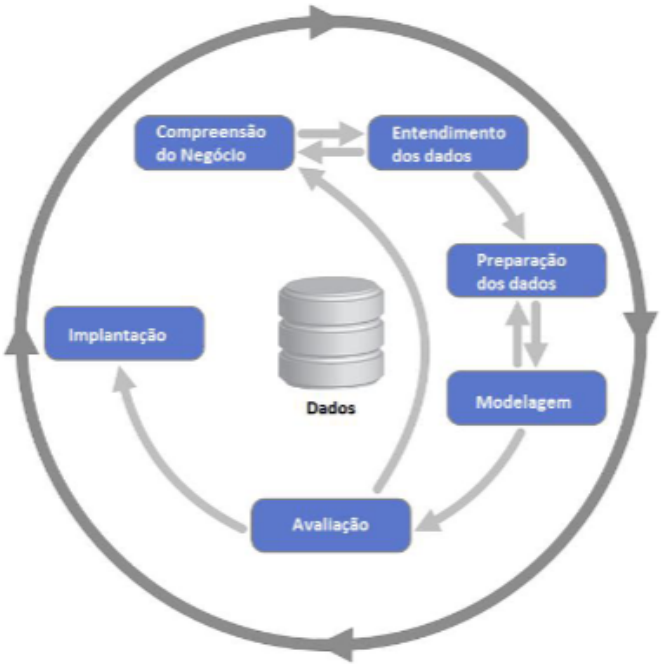


Figura 1 - Fases do CRISP-DM (Shearer, 2000).

A sequência de fases no CRISP-DM não segue uma ordem rígida. Na maioria dos projetos, há um movimento flexível entre as etapas, permitindo retornos quando necessário. Esta metodologia abrange descrições das fases típicas de um projeto, as tarefas necessárias em cada fase e explicações das relações entre essas tarefas. Como modelo de processo, o CRISP-DM oferece uma visão geral do ciclo de vida da mineração de dados (Chapman, 2000).

A fase inicial envolve a aquisição do conhecimento do domínio de negócios, ou seja, compreender os objetivos do projeto de mineração de dados a partir da perspectiva do negócio. Esse entendimento se transforma em um problema de mineração de dados. É amplamente reconhecido que essa etapa é uma das mais cruciais do processo, pois é a base para o desenvolvimento de um plano preliminar do projeto de mineração, direcionado para a realização dos objetivos (Shearer, 2000).

A segunda fase, que envolve a compreensão dos dados, inicia-se com a coleta inicial de dados, com o propósito de desenvolver uma familiaridade com os dados, identificar possíveis problemas de qualidade dos dados, obter *insights* iniciais e identificar subconjuntos de interesse que possam levar à formulação de hipóteses sobre informações ocultas. Esse estágio é de fundamental importância para prevenir surpresas indesejadas durante a fase subsequente, a preparação de dados, que frequentemente é a etapa mais extensa de um projeto (Shearer, 2000).

Na fase subsequente, ocorre a preparação dos dados, que engloba todas as atividades necessárias para construir o conjunto de dados final a partir dos dados brutos iniciais. Esses dados preparados servirão como entrada para a ferramenta de modelagem na etapa seguinte. As tarefas de preparação de dados são flexíveis e podem ser realizadas em várias iterações, sem uma ordem estritamente definida. Essas atividades envolvem a seleção de tabelas, registros e atributos, bem como a realização de transformações e a limpeza dos dados (Chapman et al., 2000).

A etapa de preparação de dados é a mais crítica do processo e frequentemente a que demanda maior tempo em projetos de mineração de dados. Estima-se que, em geral, essa fase absorva entre 50-70% do tempo e dos recursos de um projeto. Alocar recursos adequados para as fases iniciais de compreensão do negócio e esforços de tratamento de dados pode ajudar a minimizar a carga relacionada a essa etapa, mas, ainda assim, será necessário um esforço substancial para a preparação e formatação dos dados para fins de mineração (IBM, 2016).

A etapa de modelagem ocorre na quarta fase do processo. Dependendo da natureza do problema de mineração, diversas técnicas podem ser aplicadas. Tipicamente, a modelagem envolve várias iterações, nas quais o analista de dados executa múltiplos modelos, inicialmente com as configurações padrão e, em seguida, ajustam os parâmetros para obter valores otimizados. Além disso, é comum retornar à fase de preparação de dados, se necessário, para realizar manipulações específicas exigidas pelos modelos (Shearer, 2010; IBM, 2016).

A sexta e última fase é a etapa de implantação, na qual os novos *insights* e conhecimentos descobertos são aplicados para promover melhorias na organização. Durante essa etapa, é fundamental que todo o conhecimento adquirido seja organizado e apresentado de maneira que o cliente possa utilizá-lo eficazmente no processo de tomada de decisão. (Shearer, 2000; Chapman et al., 2000).

APRENDIZADO DE MÁQUINA

Nos últimos anos, houve um notável crescimento na pesquisa em aprendizado de máquina. O aprendizado de máquina é uma disciplina da inteligência artificial que se concentra no desenvolvimento de técnicas computacionais para a aquisição automática de conhecimento e na construção de sistemas capazes de aprender com base em experiências adquiridas por meio da resolução bem-sucedida de problemas anteriores (Rezende, 2005). Essa área representa a interseção entre estatística, inteligência artificial e ciência da computação, e tem aplicação significativa no reconhecimento de padrões (Guido, 2016).

O Aprendizado de Máquina (*Machine Learning*) é uma disciplina que emprega uma ampla gama de procedimentos e algoritmos para a identificação automatizada de padrões, agrupamentos e tendências nos dados, com o propósito de extrair informações valiosas para análise. Em termos simples, pode ser descrito como o uso de métodos matemáticos para treinar algoritmos a fim de reconhecer padrões (Nelli, 2015).

APRENDIZADO DE MÁQUINA SUPERVISIONADO

O aprendizado supervisionado é um processo que envolve a extração de um modelo de conhecimento a partir de dados apresentados na forma de pares ordenados, consistindo de uma entrada e uma saída desejada. A entrada representa o conjunto de atributos ou características que são fornecidos ao algoritmo para um caso específico, enquanto a saída desejada corresponde ao valor de uma característica-alvo que se espera que o algoritmo possa prever sempre que receber determinados valores de entrada (Goldschmidt, 2015). Alguns exemplos de algoritmos que se encaixam nesse modelo incluem K-NN, Modelos Lineares, Classificador Naive Bayes, Support Vector Machines e Redes Neurais Artificiais.

No presente trabalho, foi utilizado o modelo de Regressão Linear para Análise de demanda por destino e para Previsão de Tendências de viagem que serão descritos posteriormente.

REGRESSÃO LINEAR

A regressão linear é uma ferramenta matemática fundamental usada para criar modelos que descrevem e explicam o relacionamento entre variáveis. Existem dois modelos comuns: a regressão linear simples, que estabelece uma relação linear entre a variável dependente e uma variável independente, e a regressão linear múltipla, que estabelece uma relação linear entre a variável dependente e várias variáveis independentes.

O principal objetivo ao empregar a regressão linear é compreender o comportamento de uma variável específica, denominada variável dependente, em relação a um conjunto de variáveis independentes. Ao estabelecer uma relação estatística entre elas, é possível criar um modelo que represente essa relação e utilizá-la para fazer previsões (Seber, 2003).

REGRESSÃO POLINOMIAL

A regressão polinomial é um modelo de regressão no qual a relação entre as variáveis independentes e a variável dependente pode ser não linear e tem a forma de um polinômio de grau n.

Embora o polinômio de aproximação seja não linear, o problema de estimação dos parâmetros do modelo é linear e o método também é considerado uma forma de regressão linear (Mueller, 2020).

APRENDIZADO DE MÁQUINA NÃO SUPERVISIONADO

A aprendizagem não supervisionada é caracterizada pela ausência de saídas desejadas para as entradas, em que o conjunto de treinamento consiste exclusivamente de vetores de entrada (Rezende, 2005). Nesse contexto, não há saídas conhecidas que permitam ensinar à máquina quais resultados são esperados. Em vez disso, espera-se que a máquina agrupe os dados com base em suas características. O principal desafio dos modelos não supervisionados reside na avaliação da adequação da clusterização de dados com características semelhantes.

Os modelos não supervisionados são aplicados a conjuntos de dados desprovidos de rótulos, o que torna a saída desconhecida. Devido a essa característica, é complexo determinar se o modelo fez previsões corretas das informações. Uma aplicação comum desse modelo é usá-lo como um passo preliminar ao modelo supervisionado, pois a combinação do não supervisionado com o supervisionado pode resultar em um aumento na acurácia (Karen, 2012).

A tarefa de *clustering*, por sua vez, envolve a subdivisão do conjunto de dados em grupos, denominados *clusters*. O objetivo principal é dividir os dados de forma que os pontos dentro de um mesmo cluster sejam altamente semelhantes, enquanto os pontos pertencentes a diferentes clusters sejam distintos (Guido, 2016).

Neste trabalho, utilizamos o *K-means* para a Segmentação dos Clientes a ser descrito posteriormente.

K-MEANS

O *K-Means Clustering* é uma técnica que realiza a clusterização por meio do método de particionamento (Guido, 2016). Esse método envolve a criação de várias partições dos dados e a avaliação delas com base em critérios específicos. O algoritmo K-Means tem como objetivo identificar os centros das regiões que representam diferentes tipos de dados. Ele opera alternando entre a atribuição de cada ponto ao centro mais próximo e a seleção do centro do cluster como a média dos pontos atribuídos a ele. O processo continua até que as atribuições aos *clusters* não sofram mais alterações, o que marca o ponto de convergência do algoritmo (Tan, 2014).

MÉTRICAS DE AVALIAÇÃO DE EFICIÊNCIA

A avaliação das métricas de eficiência em modelos de regressão é uma parte essencial do processo de desenvolvimento de modelos de aprendizado de máquina. Essas métricas fornecem uma visão abrangente da qualidade do modelo e permitem tomar decisões informadas com base em previsões precisas. No Projeto utilizamos as seguintes métricas:

Coeficiente de Determinação (R^2): O R^2 mede a proporção da variabilidade nos dados de resposta que é explicada pelo modelo. Um valor de R^2 próximo de 1 indica um ajuste excelente, enquanto um valor próximo de 0 indica um ajuste ruim. O R^2 é uma métrica crucial para entender o poder explicativo do modelo. Um R^2 mais alto é preferível; e

Erro Quadrático Médio (MSE): O MSE calcula a média das diferenças ao quadrado entre as previsões do modelo e os valores reais. O MSE atribui mais peso a erros maiores, sendo sensível a *outliers* e indicando a magnitude geral dos erros do modelo. Um MSE menor indica um melhor ajuste do modelo aos dados.

METODOLOGIA - PROJETO BOA VIAGEM

Para o Projeto, realizou-se uma análise exploratória minuciosa dos dados usando diversas técnicas e ferramentas em Python utilizando a base de dados escolhida.

Entender os dados é o primeiro passo crucial em qualquer análise de dados. Envolve explorar as características dos dados como sua distribuição, tendências, padrões e relações entre variáveis.

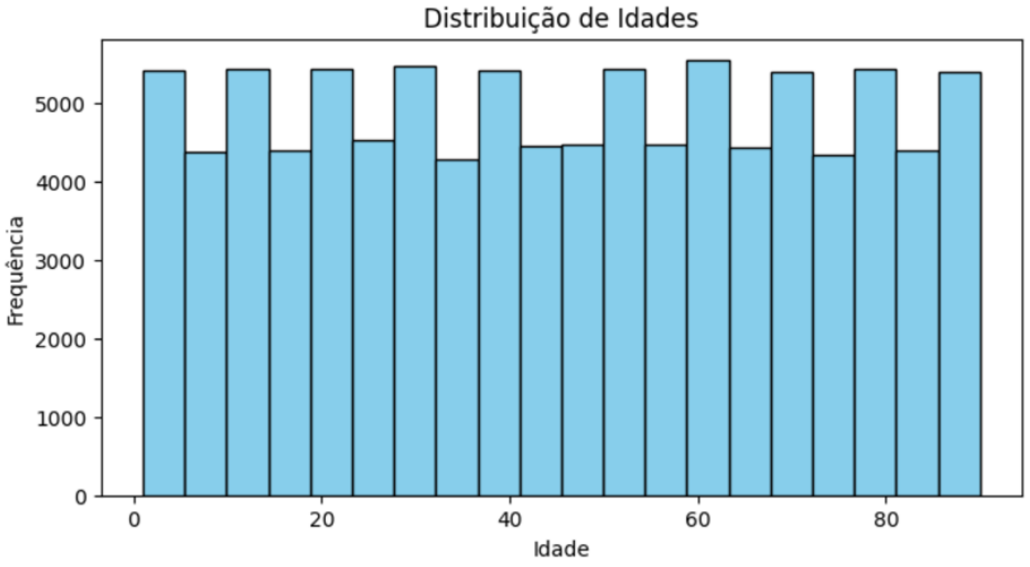
Segundo Ferreira (2021), na classificação dos dados, as “variáveis podem representar diferentes valores, como numéricos e não numéricos”. Foi verificada na nossa base que apenas a variável “idade” é numérica, sendo todas as demais não numéricas.

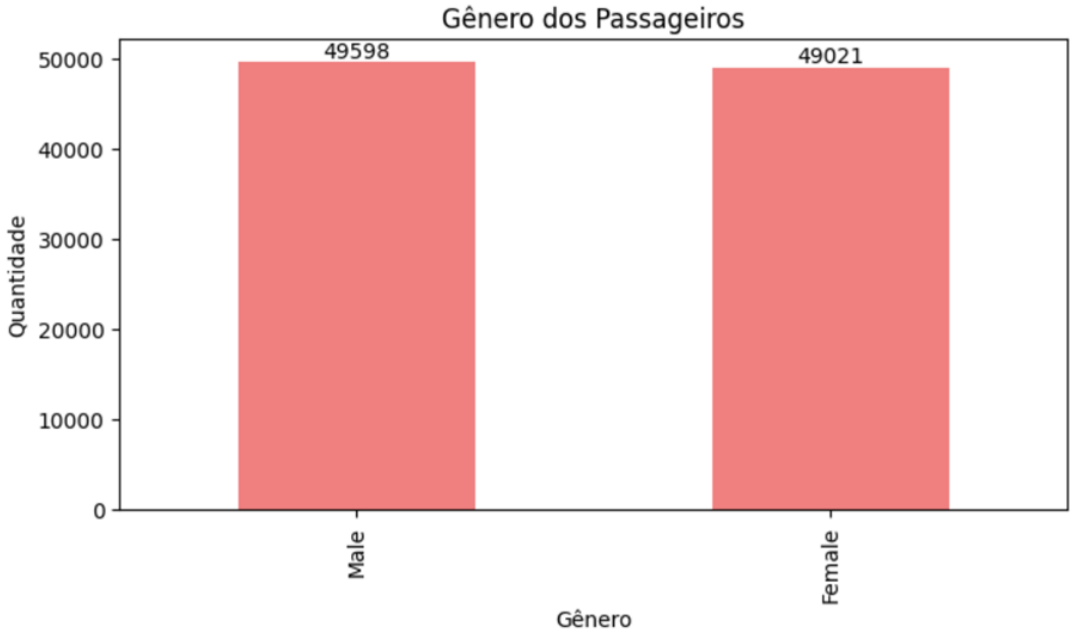
“Conhecer e preparar de forma adequada os dados para análise é uma etapa [...] que pode tornar todo o processo de mineração muito mais eficiente e eficaz. Por outro lado, dados mal ou não pré-processados podem inviabilizar uma análise ou invalidar um resultado” (CASTRO, 2016).

A integridade dos dados foi verificada, através da constatação de inexistência de valores ausentes no *DataFrame*. A ausência de valores faltantes é crucial para uma análise precisa.

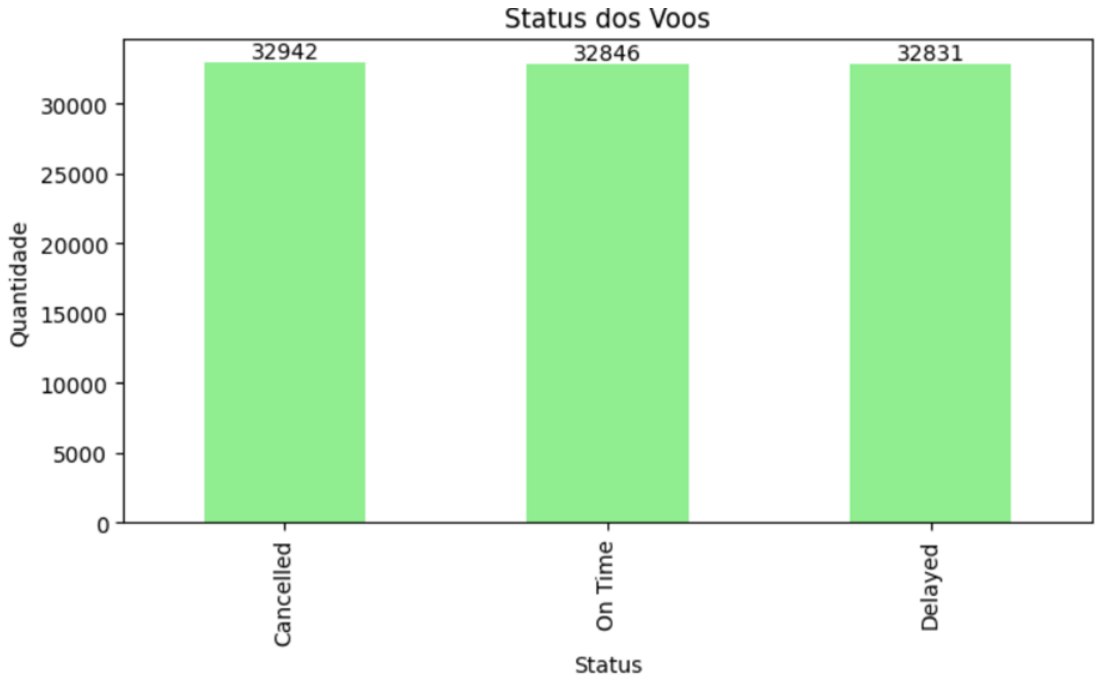
O atributo “Gênero” foi convertido para o tipo de dados categoria, com o objetivo de economizar espaço e facilitar análises categóricas. Já o atributo “idade” foi definido como inteiro para facilitar operações numéricas e análises estatísticas.

Para visualização da distribuição dos dados relacionados aos passageiros, foi utilizado o histograma para o atributo “Idade” e o gráfico de barras para o atributo “Gênero” para identificação de discrepâncias ou tendências do perfil dos viajantes.

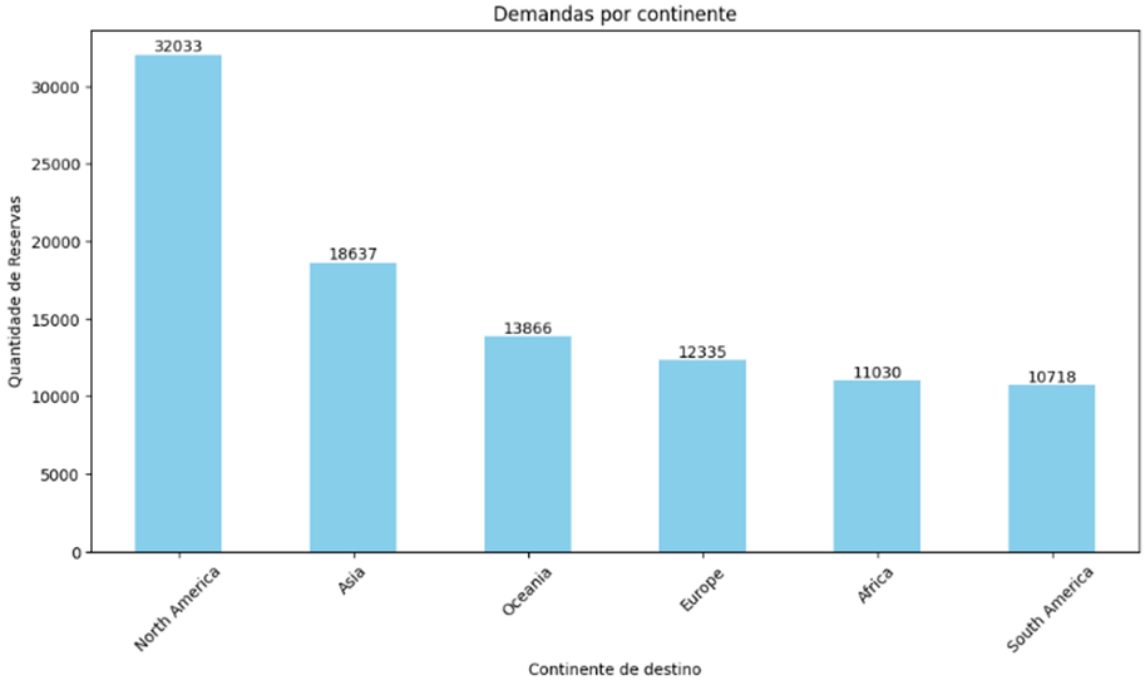


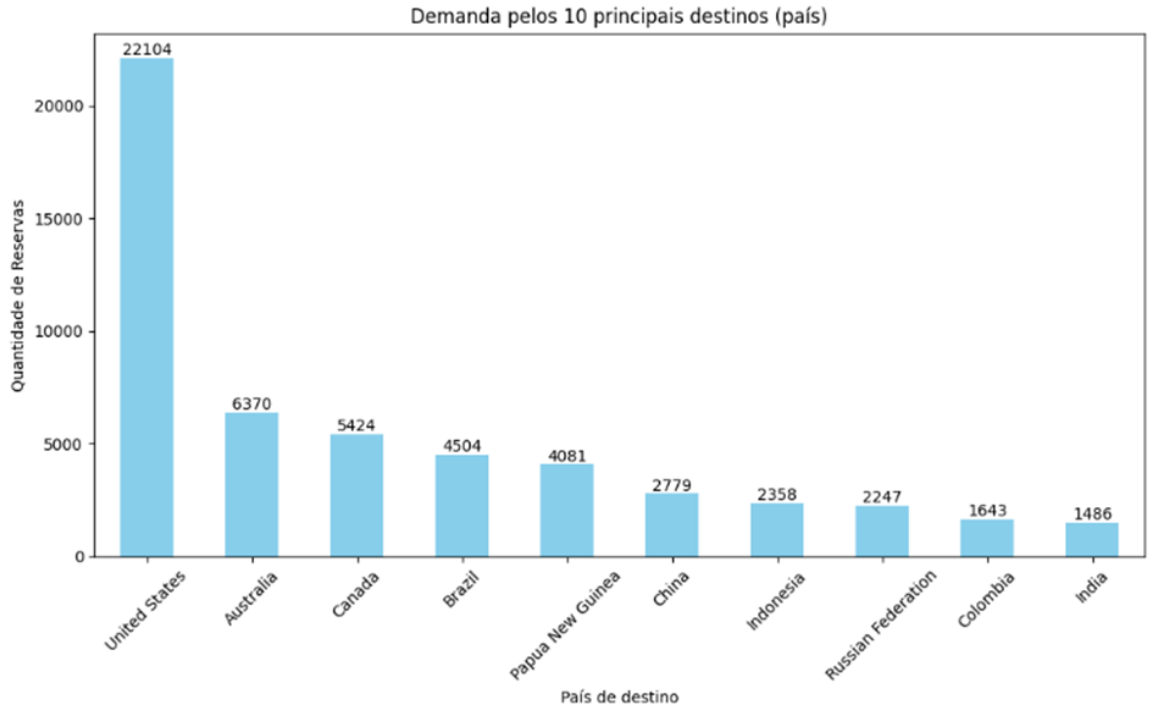


Foram explorados os diferentes “Status dos voos” utilizando gráfico de barras. Isso permitiu identificar padrões como a frequência de voos atrasados, cancelados ou pontuais.



As demandas por “Continentes” e pelos principais “Países” de destino foram investigadas, com a utilização de gráficos de barras, proporcionando insights sobre as preferências dos passageiros e auxiliando nas estratégias de marketing e expansão.





ANÁLISE DE DEMANDA POR DESTINO

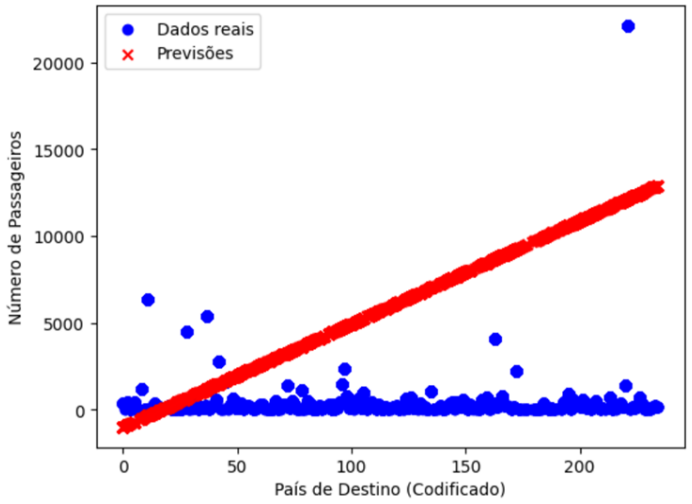
Para a divisão dos dados em conjunto de treinamento e teste foram realizadas as etapas: Codificação da variável categórica “Nome do País” em valores numéricos, usando a técnica *Label Encoding*. Em seguida, foi calculado o número de registros para cada país de destino, organizando esses valores em um *DataFrame* com duas colunas “Nome do País” e “Contagem de Passageiros”. Esses resultados são, então, integrados ao *DataFrame* original, proporcionando uma visão mais completa e contextualizada dos dados.

Posteriormente, os dados foram preparados para o treinamento do modelo. A variável independente (x) foi definida como “Nome do País”, enquanto a variável dependente (y) é definida como “Contagem de Passageiros”.

Para avaliar a eficácia do modelo, os dados são divididos em conjuntos de treinamento e teste, com 80% dos dados usados para treinamento e 20% para teste.

Com os dados preparados, foi aplicado o modelo de regressão linear. Utilizando a biblioteca *scikit-learn*, o modelo é treinado com os dados de treinamento, permitindo que ele aprenda padrões nos dados. Em seguida, o modelo é utilizado para fazer previsões com base nos dados de teste.

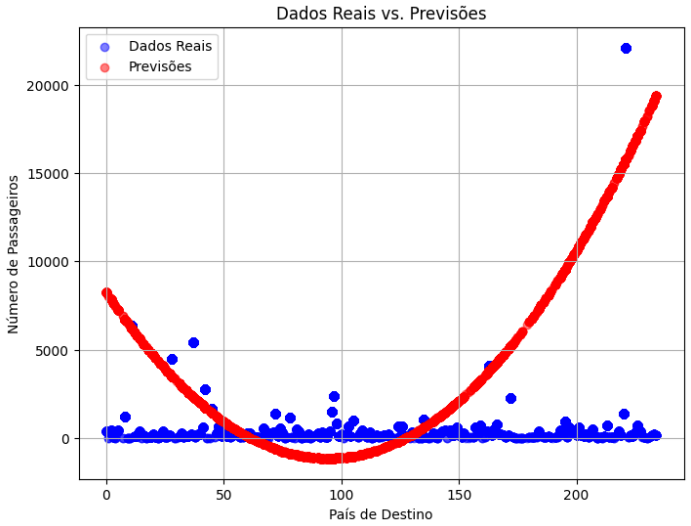
Para proporcionar uma compreensão visual das previsões, foi gerado um gráfico de dispersão que mostra tanto os dados reais quanto as previsões do modelo. Os dados reais são marcados em azul, enquanto as previsões do modelo são representadas em vermelho, marcadas com 'x'.



Finalmente, foi realizada a avaliação da qualidade do modelo. Foram apurados o Erro Quadrático Médio (MSE) e o Coeficiente de Determinação (R^2) para o modelo. Como resultado, obtivemos os seguintes resultados:

```
Erro Quadrático Médio (MSE): 51689893.02
Coeficiente de Determinação (R²): 0.30
```

Considerando que o Coeficiente de Determinação do Modelo ficou baixo, foi implementado o modelo de Regressão Polinomial que apresentou os seguintes resultados:




```
Erro Quadrático Médio (MSE): 32125604.94
Coeficiente de Determinação (R²): 0.56
```

Comparando-se a análise realizada pela Regressão Linear e a Polinomial quanto ao Erro Quadrático Médio (MSE) e o Coeficiente de Determinação (R^2), foi apurado que a Regressão Polinomial melhor se adequa aos dados apresentados, considerando que apresentou menor MSE e maior R^2.

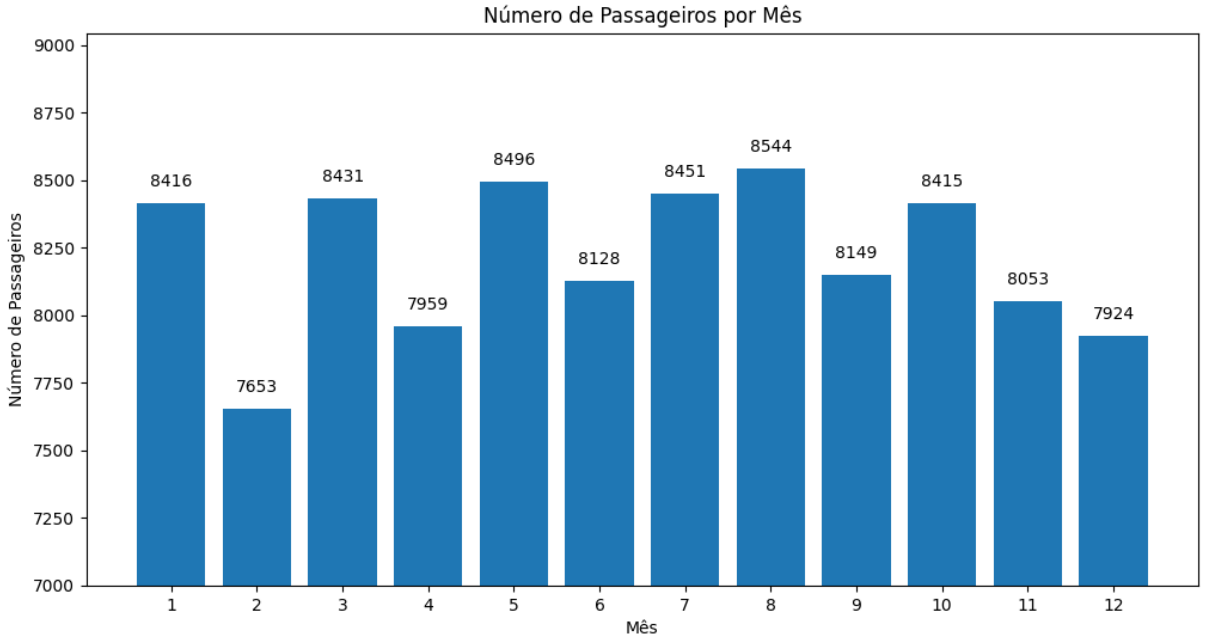
ANÁLISE DE DEMANDA POR DESTINO - NÚMERO DE PASSAGEIROS AO LONGO DO ANO

Inicialmente, a coluna *"*Data do Voo*"* foi convertida para o formato *datetime* utilizando a biblioteca *pandas*. Isso foi realizado para facilitar a manipulação e análise de dados de data e hora. Em seguida, foram extraídos o ano e o mês a partir da coluna "Data do Voo" com o armazenamento dessas informações em colunas separadas "Ano" e "Mês*"*. Essa extração de dados permitiu a análise dos voos em níveis mensais.

Foi calculado o número de passageiros para cada mês específico usando a coluna "Identificação Passageiro". Os resultados são armazenados em um novo *DataFrame* chamado "Data Mensal", apresentando o número de passageiros por mês.

Em seguida, foi feito um gráfico de barras usando a biblioteca matplotlib, onde o eixo "x" representa o mês e o eixo "y" representa o número de passageiros. Para aprimorar a visualização, o intervalo do eixo "y" é ajustado para começar em 7000 e

ir até o valor máximo do número de passageiros, garantindo que as barras sejam claramente visíveis.



SEGMENTAÇÃO DE CLIENTES

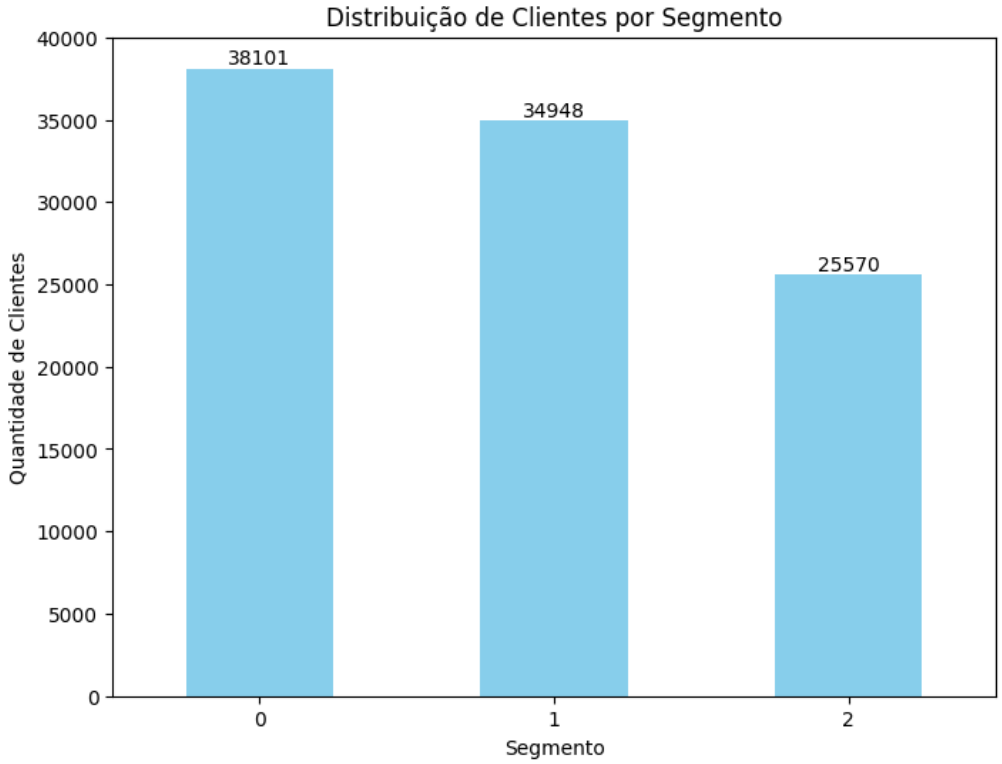
Para segmentação, foi utilizado o algoritmo de *K-means*.

Inicialmente, foi empregada a técnica de codificação de variáveis categóricas. Para isso, utilizou-se o *LabelEncoder* para transformar todas as variáveis inicialmente não numéricas em numéricas.

Foram selecionadas apenas as variáveis relevantes “Idade”, “Gênero” e “Nacionalidade”. Em seguida, determinou-se o número de clusters desejado, neste caso, definido como 3.

A seguir, o algoritmo *K-means* foi aplicado aos dados utilizando a biblioteca *scikit-learn*. O *K-means* é um algoritmo de aprendizado não supervisionado usado para dividir dados em grupos ou *clusters*, onde cada ponto de dados é atribuído ao cluster mais próximo do centróide.

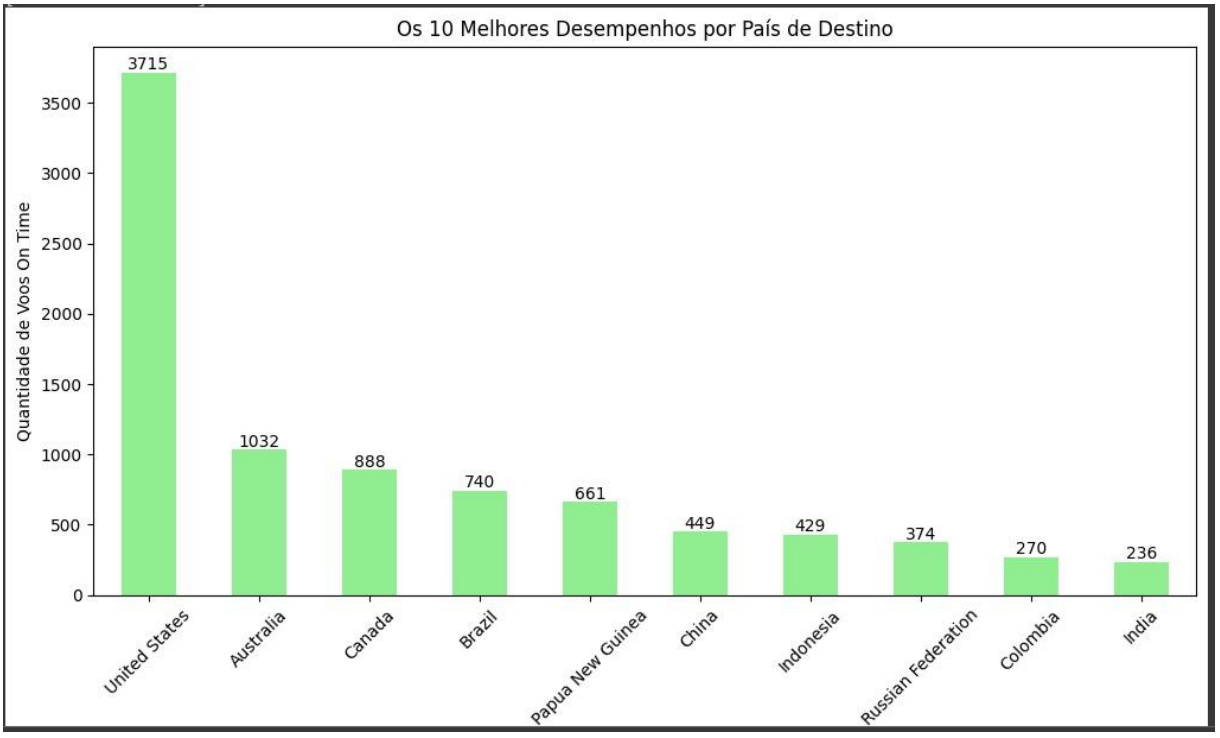
O resultado da aplicação do *K-means* adicionou uma nova coluna chamada "Segmento” ao *DataFrame* original, indicando a que cluster cada entrada pertence. Além disso, foi realizada a contagem do número de pontos de dados em cada segmento, fornecendo uma visão da distribuição dos dados nos *clusters*.

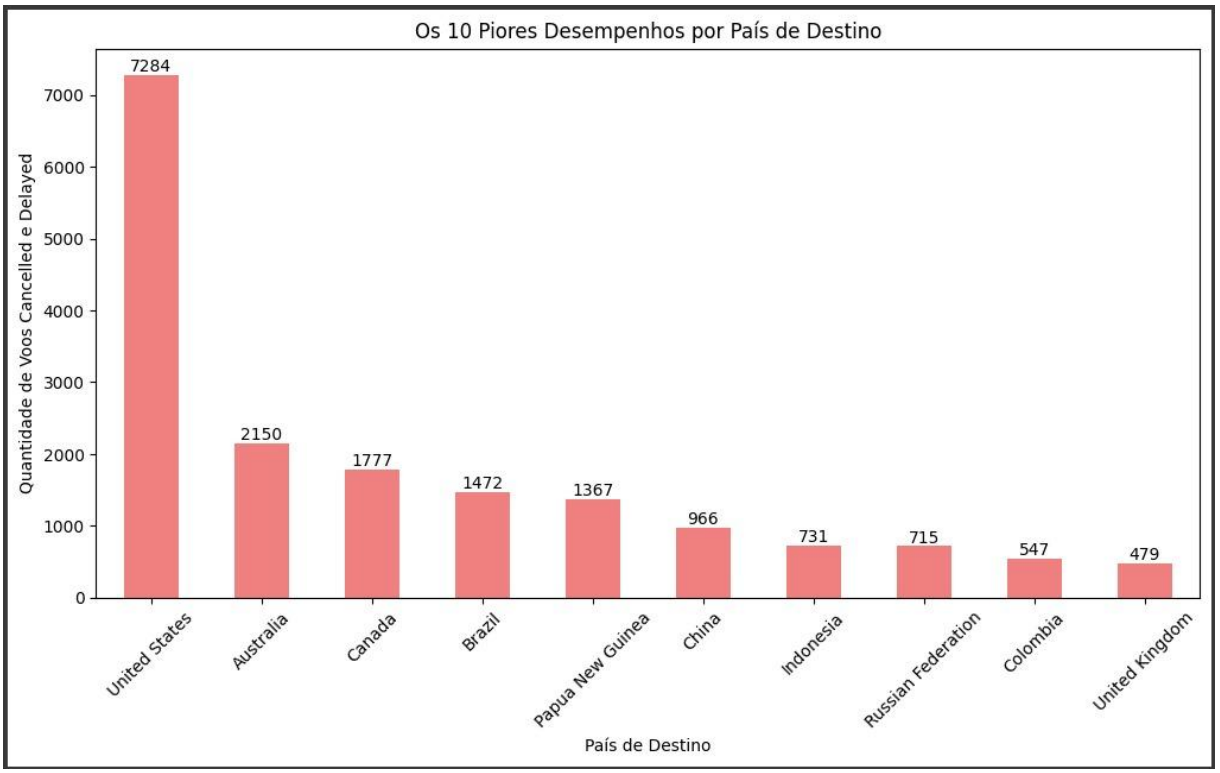


ANÁLISE DE DESEMPENHO DE VOOS

Foi realizada uma análise do desempenho de diferentes países em termos de status de voos. Inicialmente, foram agrupados os dados pelo nome do país (Nome do País) e pelo status do voo (Status do voo). Em seguida, foi realizada a contagem de ocorrência de cada status de voo em cada país.

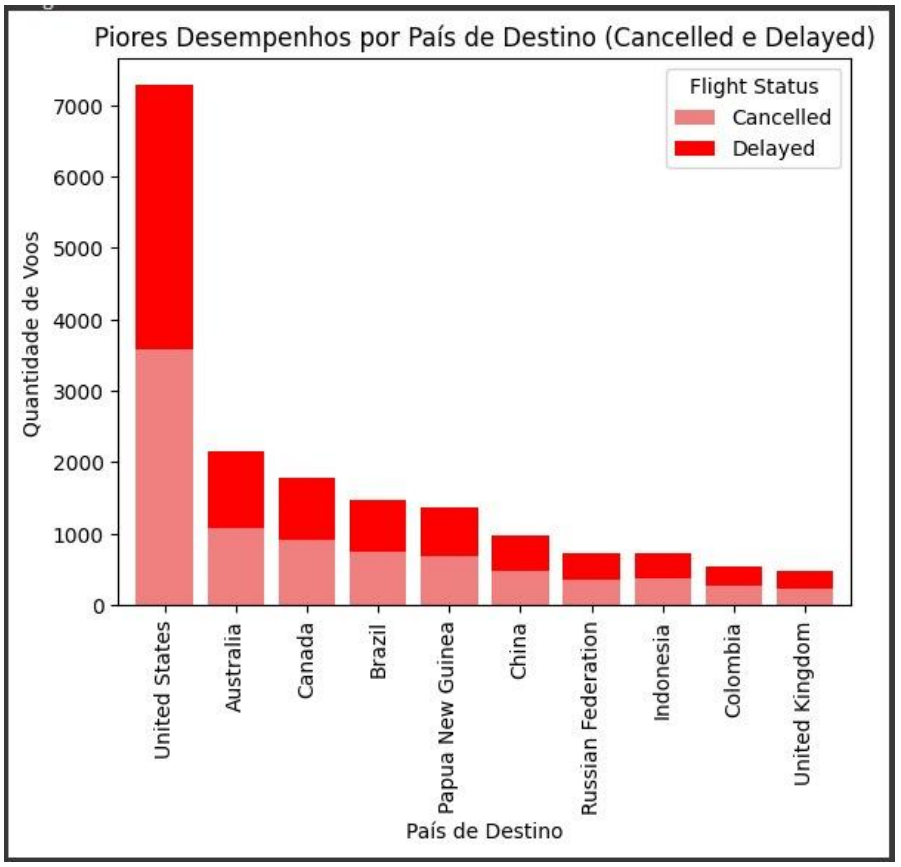
Esta análise oferece uma visão organizada do desempenho dos voos em diferentes países, nesta etapa, foram segmentados os 10 melhores e 10 piores desempenhos por país de destino, permitindo uma análise comparativa e identificação de padrões ou tendências nos dados.



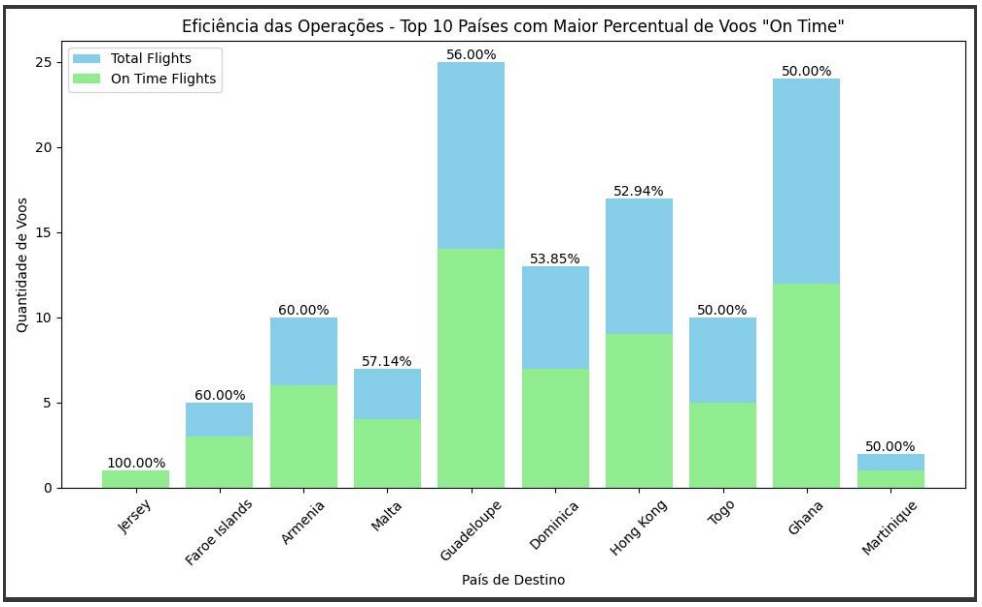


Pode-se notar que os países que possuem maior quantidade de voos são os que estão relacionados tanto dentre os que possuem maior quantidade de voos “on time” quanto os que apresentam mais voos atrasados e/ou cancelados. Possivelmente por terem maior quantidade de voos possuem maiores quantidades de intercorrências que acarretam atrasos ou cancelamentos de voos.

Segue abaixo, gráfico para visualização dos voos com intercorrências. É possível perceber que as quantidades de voos cancelados são próximas dos atrasados.



O gráfico abaixo traz os 10 países mais eficientes quando comparamos o total de voos e voos *on time*. Nota-se que países com menor quantidade de voos são os mais eficientes.



PREVISÃO DE TENDÊNCIA DE VIAGEM

Os dados foram agrupados pelo nome do país de destino (Nome do País) e, em seguida, a contagem de passageiros é calculada para cada país, resultando em um *DataFrame* chamado “Contagem de passageiros”, que possui duas colunas: “Nome do País” e “Contagem de Passageiros”, onde o primeiro representa o país de destino e o segundo o número de passageiros correspondente.

Após essa etapa inicial, os dados foram preparados para o treinamento do modelo de regressão linear. A variável independente (x) é definida com os países de destino (codificados através de variáveis *dummy* para representação categórica), enquanto a variável dependente (y) é definida com o número de passageiros.

Em seguida, os dados são divididos em conjuntos de treinamento (80%) e teste (20%) usando a função *train_test_split* da biblioteca *scikit-learn*. Posteriormente, um modelo de regressão linear é criado e treinado utilizando os dados de treinamento.

Após o treinamento, o modelo é utilizado para fazer previsões sobre o número de passageiros para cada país de destino. O modelo de regressão linear aprende padrões nos dados de treinamento e tenta estender esses padrões para fazer previsões precisas para os dados de teste.

As análises realizadas estão disponíveis no google colab, através do link:

Projeto Aplicado 2.ipynb

RESULTADO

No âmbito deste projeto, almejamos alcançar resultados significativos que impulsionarão o sucesso da "Boa Viagem" no dinâmico mercado de viagens e turismo.

Nossos objetivos abrangem diversas áreas-chave, cada uma contribuindo de forma crucial para o crescimento e aprimoramento da empresa.

Análise de Demanda por Destino: O primeiro objetivo concentrou-se em analisar a demanda por destinos específicos com base em reservas de passagens aéreas. Foram identificados os destinos (países / continentes) mais populares, os meses de maior e menor concentração de voos. Isso permitirá que a "Boa Viagem" direcione melhor seus recursos e esforços promocionais para destinos de alto potencial e ajuste suas estratégias de *marketing* de acordo com as tendências de voo.

Segmentação de Clientes: A segmentação de clientes é um importante ponto do nosso projeto. Ao segmentar os clientes com base nos critérios idade, gênero e nacionalidade, nosso objetivo foi personalizar as ofertas de pacotes de viagem de acordo com as necessidades e interesses específicos de cada grupo. Esperamos com isso melhorar significativamente a experiência do cliente, tornando-a mais relevante e atraente.

Análise de Desempenho de Voos: A avaliação do desempenho de voos é vital para garantir operações eficientes e confiáveis. Ao analisar o status dos voos e identificar áreas de melhoria na eficiência das operações, pudemos perceber que países com maior quantidade de voos são passíveis de maior quantidade de intercorrências. Foram listados também os top 10 países com maior percentual de voos “*on time*” e pudemos observar que são países com menor quantidade de voos.

Previsão de Tendências de Viagem: A capacidade de prever tendências futuras de viagem com base em análises históricas foi um dos pilares do nosso projeto. Com isso, a empresa poderá planejar estrategicamente seus serviços e ofertas. Isso não apenas aumentará a competitividade, mas também garantirá que a "Boa Viagem" permaneça na vanguarda do setor de viagens e turismo.

Estamos confiantes de que as informações e recomendações resultantes deste projeto servirão como base sólida para decisões estratégicas informadas e para a construção de relacionamentos duradouros com seus clientes.

BIBLIOGRAFIA

CASTRO, Daniel Gomes Ferrari Leandro Nunes de. Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações. [Digite o Local da Editora]: Editora Saraiva, 2016. E-book. ISBN 978-85-472-0100-5. Disponível em: https://app.minhabiblioteca.com.br/#/books/978-85-472-0100-5/. Acesso em: 17 set. 2023.

CHAPMAN, P.et al.CRISP-DM 1.0: Step-by-step data mining guide.SPSS inc, v. 9, p. 13, 2000. Disponível em: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf. Acesso em: 29 out. 2023.

FERREIRA, Rafael G C.; MIRANDA, Leandro B. A de; PINTO, Rafael A.; et al. Preparação e Análise Exploratória de Dados. [Digite o Local da Editora]: Grupo A, 2021. E-book. ISBN 9786556902890. Disponível em: https://app.minhabiblioteca.com.br/#/books/9786556902890/. Acesso em: 19 set. 2023.

GOLDSCHMIDT, Ronaldo. Data Mining. [Digite o Local da Editora]: Grupo GEN, 2015. E-book. ISBN 9788595156395. Disponível em: https://app.minhabiblioteca.com.br/#/books/9788595156395/. Acesso em: 29 out. 2023.

GUIDO, A. C. S. Introduction to Machine Learning with Python. [S.l.]: O'Reilly Media, 2016.

IBM SPSS. IBM SPSS modeler text analytics 16 user guide. 2016. Disponível em: https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/16.0/en/ta_guide_book.pdf. Acesso em: 29 out 2023.

KAREM, F.; DHIBI, M.; MARTIN, A. Combination of supervised and unsupervised classification using the theory of belief functions. Belief Functions: Theory and Applications, 2012.

LAROSE, C. D.; LAROSE, D. T. Data science using Python and R. Hoboken: Wiley, 2019 (Series on Methods and Applications in Data Mining).

MUELLER, John P.; MASSARON, Luca. Python Para Data Science Para Leigos. Editora Alta Books, 2020. E-book. Disponível em: https://app.minhabiblioteca.com.br/#/books/9786555201512/. Acesso em: 06 nov. 2023.

NELLI, F. Python Data Analytics. Nova York: Apress. Springer Science+Business, 2015.

REZENDE, S. O. Sistemas inteligentes: fundamentos e aplicações. [S.l.]: Manole, 2005.

SEBER, G. A. F.; LEE, A. J. Linear Regression Analysis. [S.l.]: Wiler, 2003.

SHEARER, C. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, v. 5, n. 4, 2000. Disponível em: https://www.academia.edu/42079490/CRISP_DM_The_New_Blueprint_for_Data_Mining_Colin_Shearer_Fall_2000. Acesso em: 29 out. 2023.

TAN, M. S. P.-N.; KUMAR, V. Introduction to Data Mining. [S.l.]: PEARSON, 2014.