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 3](#_Toc147747271)

[INTRODUÇÃO 4](#_Toc147747272)

[OBJETIVOS E METAS 6](#_Toc147747273)

[METODOLOGIA 7](#_Toc147747274)

[ANÁLISE EXPLORATÓRIA DOS DADOS 8](#_Toc147747275)

[RESULTADO (ESPERADO) 14](#_Toc147747276)

[CRONOGRAMA DO PROJETO 15](#_Toc147747277)

[BIBLIOGRAFIA 16](#_Toc147747278)

# 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>

# 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*, em que 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 usadas as seguintes bibliotecas do 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;
* sklearn – para implementações eficientes de uma ampla gama de algoritmos de aprendizado de máquina, incluindo regressão linear, árvores de decisão, máquinas de vetores de suporte (SVM), k-vizinhos mais próximos (K-NN), entre 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.

# ANÁLISE EXPLORATÓRIA DOS DADOS

No contexto da base de dados escolhida, realizamos uma análise exploratória minuciosa dos dados usando diversas técnicas e ferramentas em Python.

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, p.15), 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 numérica.

“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, p.47).

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.

Gráfico, Gráfico de barras, Histograma

Descrição gerada automaticamente

Gráfico, Gráfico de barras

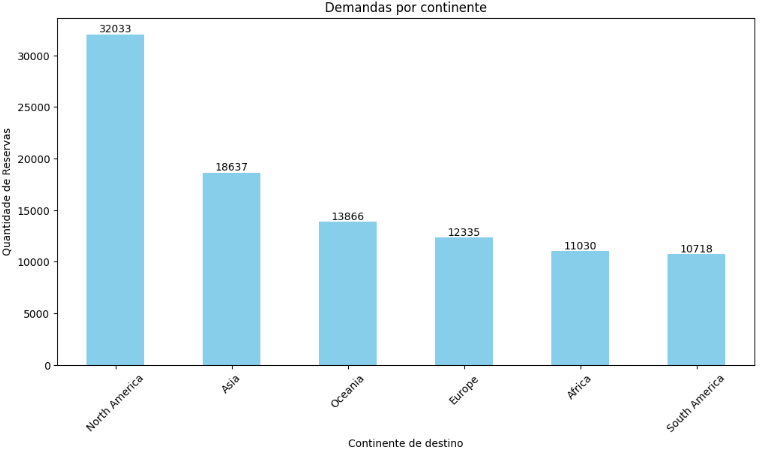
Descrição gerada automaticamente

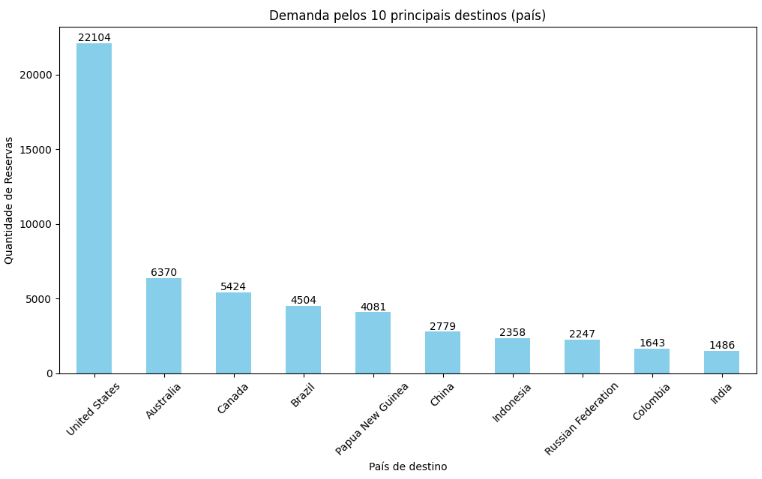
Foram explorados os diferentes “Status do voo” usando gráfico de barras. Isso permitiu identificar padrões, como a frequência de voos atrasados, cancelados ou pontuais.

Gráfico, Gráfico de cascata

Descrição gerada automaticamente

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 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, o código gera 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'.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Finalmente, foi realizada a avaliação opcional da qualidade do modelo. Ele calcula o Erro Quadrático Médio (MSE), uma métrica que mede a diferença entre as previsões do modelo e os dados reais. Essa métrica fornece insights sobre o quão bem o modelo está performando, sendo uma ferramenta valiosa para avaliação e refinamento do modelo em aplicações do mundo real.

**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 toas as variáveis inicialmente não numéricas em numéricas.

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

A seguir, o algoritmo K-means foi aplicado aos dados usando 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 centroide.

O resultado da aplicação do K-means adionou uma nova coluna chamada " Segmento” ao DataFrame original, indicando a que cluster cada entrada pertence. Além disso, o código conta e imprime o 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, permitindo uma análise comparativa e identificação de padrões ou tendências nos dados.

**Previsão de Tendências 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”. Este DataFrame possui duas colunas: “Nome do País” e “Contagem de Passageiros”, em que 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 como os países de destino (codificados através de variáveis dummy para representação categórica), enquanto a variável dependente (y) é 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.

# RESULTADO (ESPERADO)

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 concentra-se em analisar a demanda por destinos específicos com base em reservas de passagens aéreas. Esperamos identificar os destinos mais populares, compreender os padrões sazonais e os comportamentos de reserva dos clientes. 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 sazonais.

Segmentação de Clientes: A segmentação de clientes é um importante ponto do nosso projeto. Ao segmentar os clientes com base em critérios como idade, gênero, nacionalidade e preferências de viagem, nosso objetivo é personalizar as ofertas de pacotes de viagem de acordo com as necessidades e interesses específicos de cada grupo. Esperamos 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, buscamos aumentar a pontualidade, reduzir atrasos e minimizar cancelamentos. Isso resultará em uma experiência de viagem mais tranquila e agradável para os clientes da "Boa Viagem".

Previsão de Tendências de Viagem: A capacidade de prever tendências futuras de viagem com base em análises históricas é um dos pilares do nosso projeto. Com isso, a empresa poderá se antecipar às mudanças nas preferências dos clientes e 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.

# CRONOGRAMA DO PROJETO

Tabela

Descrição gerada automaticamente

Arquivo Excel disponível no GitHub.

Observação: Dias restantes contados da data da elaboração desse documento, 09/10/2023.

# 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.

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: 17 set. 2023.