SUSTAINABLE BUSINESS INTERNATIONAL JOURNAL
Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponível em: https://periodicos.uff.br/sbijournal
PPGAd uff

Estimação da quantidade de combustível consumido em aeronaves utilizando regressão linear

Rafael Vernizzi Oliveira rafael.vernizzi@gmail.com
Julia Hosken juliahosken@hotmail.com

Resumo

O transporte aéreo é o principal meio de transporte utilizado pelos turistas internacionais e um dos principais pelos turistas nacionais no Brasil, sendo importante para um país de dimensão continental, integralizando os mais diferentes destinos. Logo, a eficiência dos voos precisa ser continuamente revisada e melhorada. Esse fato levantou a questão que conduziu o tema deste trabalho: "Estimação da quantidade de combustível consumido em aeronaves utilizando regressão linear". A ciência de dados é uma poderosa ferramenta para análise de dados podendo ser aplicada com êxito na área da aviação civil. Este projeto foi realizado com auxílio da linguagem de programação Python. Com o Python foram realizados a importação, limpeza, visualização e por fim a análise de regressão linear dos dados. Conclui-se que a previsão da demanda da utilização de combustível em relação a cada variável independente pode trazer uma série de benefícios, como melhorar a eficiência, a competitividade e a margem de lucro de empresas através da aplicação das informações aqui disponibilizadas em novas estratégias das companhias aéreas.

Palavras-Chave: Transporte Aéreo, Regressão Linear, Python.

Abstract

Air transport is the main means of transport used by international tourists and one of the main means by national tourists in Brazil, being important for a continental-sized country, covering the most different destinations. Therefore, flight efficiency needs to be continually reviewed and improved. This fact raised the question that corresponds to the theme of this work: "Estimating the amount of fuel consumed in aircraft using linear regression". Data science is a powerful tool for data analysis and can be successfully applied in the field of civil aviation. This project was carried out with the help of the Python programming language. Using Python, data import, cleaning, visualization and, finally, linear regression analysis were carried out. It is concluded that forecasting the demand for fuel use in relation to each independent variable can bring a series of benefits, such as improving the efficiency, competitiveness and profit margin of companies through the application of the information made available here in new company strategies. airline companies.

Keywords: Air Transport, Linear Regression, Python.

1. Introdução

O mundo no qual vivemos está cada vez mais empenhado na busca por decisões acertadas em um curto período de tempo, sendo está, um dos grandes desafios pelos quais as empresas se deparam. Até pouco tempo atrás as grandes decisões eram baseadas principalmente na intuição humana e na experiência profissional (softiskils). Com certeza tais primícias continuam sendo importantes, porém sozinhas, sem uma base concreta, podem enviesar decisões, baixar o nível de lucratividade, manchar a reputação de uma organização, lançar empresas em dívidas bancárias e além de tudo isso aumentar o número de demissões. Estes são apenas alguns exemplos das consequências negativas quando desconsideramos aplicar metodologias modernas de análise de forma antecipada e apenas depois disso fazer as escolhas, no entanto, são inúmeros benefícios quando as decisões são robustas e assertivas. A ciência de dados chegou com esse propósito, independentemente do tamanho da base de dados disponível, ela se tornou uma ferramenta poderosa para se entregar valor aos tomadores de decisão.

As estratégias adotadas devem sempre estar alinhas ao crescimento econômico, porém o crescimento econômico é um grande influenciador dos transportes terrestres, marítimos e aéreos. Em destaque, pode-se citar o crescimento do transporte aéreo em todo o mundo. Novas rotas surgiram como nichos a serem exploradas por companhias que se esforçam a proporcionar facilidades oferecendo serviços de qualidade e buscando uma maior eficiência para lidar com a concorrência existente nesse ramo. (FERIYANTO; SALEH; FAUZI; DZAKIYULLAH; IWAPUTRA, 2015)

Para atender a grande demanda, empresas importantes como Boing, Airbus e Embraer precisam analisar a tendência de mercado e buscar tecnologias inovadoras, utilizando novas fontes de energias sustentáveis para suas aeronaves. Aviões movidos a combustíveis menos agressivos e até mesmo versões elétricas ou híbridas despontam como opções interessantes no transporte aéreo moderno.

A grande pergunta que surge é “como otimizar ainda mais os voos comerciais visto que grande parte das novas tecnologias já estão sendo aplicadas pela aviação?”. Sim a busca por novas tecnologias, novos conhecimentos, novos materiais estruturais, rotas inteligentes, combustíveis sustentáveis e mais eficientes é o alvo daquelas instituições que atuam na produção e operação dessas aeronaves.

Segundo Caetano e Alves (s.d.), as áreas com maior número de estudos incluem a indústria aeronáutica (eficiência energética, processo industrial e redução de ruído e emissão de poluentes), companhias aéreas (modelo de negócios, TI e planejamento e gestão), políticas (transporte sustentável, mecanismos de incentivo e aspirações sociais) e aeroporto (serviços, segurança, autofinanciamento e controle e projetos de tráfego aéreo). Ainda acrescentam que partir dessa revisão, determinou-se que, além dos estudos limitados sobre o assunto, também há carência de pesquisas voltadas a inovação no que tange estruturas aeroportuárias, como por exemplo: pavimentação de pistas e otimização de terrenos aeroportuários, bem como em novas formas de destinação de resíduos gerados nos voos, treinamento oferecidos aos tripulantes e planejamento integrado de inovação no setor. Estes podem direcionar estudos futuros sobre essa temática nas quatro áreas de aplicação identificadas e promover o desenvolvimento de um sistema integrado de inovação na gestão do transporte aéreo.

No mundo moderno, existe um grande desafio rumo ao crescimento econômico de forma sustentável devido a degradação das condições ambientais necessárias à vida. Partindo

desta preocupação, acordos e regras internacionais são criadas para disciplinar até certo ponto as atividades das empresas, como por exemplo o sistema de “crédito de carbono”.

Existe uma grande dificuldade de descarbonizar a indústria devido aos custos envolvidos e limitações dos combustíveis alternativos. Segundo especialistas do setor, ela é responsável por 2% das emissões globais de CO2. Os executivos alertam para importância das vendas de novas aeronaves com maior eficiência energética que ajudariam a financiar o investimento em carbono zero, citando como exemplo, aviões elétricos ou movidos por hidrogênio. (HANCOCK; GEORGIADIS; PFEIFER, 2023).

Sobre as características influenciadoras do transporte aéreo, Oliveira diz que se destacam a seguintes características do transporte aéreo: importância na economia, alavancagens da cadeia produtiva, inserção internacional do País e vulnerabilidade e choques externos, impacto nas contas externas, posição efeito de integração e desenvolvimento ao longo do território nacional. (2009, p.26)

Para Oliveira (2009) é necessário existir acesso para novas empresas aéreas com a utilização de critérios equilibrados de aprovação, sendo que, potenciais interessados em operar no transporte aéreo devem passar por uma inspeção rigorosa do ponto de vista técnico, no entanto, deve-se tomar cuidado com requisitos econômicos à entrada para que os mesmos não se transformem em verdadeiros obstáculos com prejuízo às alternativas de consumo dos passageiros. Destacando-se que, uma vez cuidada da fiscalização das operações e das condições de segurança de vôo, deve ser garantido o livre acesso, livre mobilidade e a liberdade estratégica.

É de conhecimento que a aviação é contemplada com uma vasta quantidade de tecnologia visando oferecer um serviço de confiança e qualidade às pessoas de todo mundo. Essas tecnologias geram uma gigantesca massa de dados, rica em informações, que precisam ser tratadas e exploradas da forma correta por aqueles que fornecem e fiscalizam as atividades desenvolvidas de transporte aéreo.

No Brasil, os dados do transporte aéreo são regulamentados pela Resolução ANAC nº 191 de 2011 e pelas Portarias ANAC nº 1.189 e 1.190/SER/2011. (ANAC, 2022)

Os dados, conforme citado na regulamentação mencionada, são mensalmente fornecidos à ANAC, até o dia 10 do mês subsequente ao de referência, pelas empresas brasileiras e estrangeiras que exploram os serviços de transporte aéreo público regular e não regular no Brasil. (ANAC, 2022)

Sobre os procedimentos aos quais tais dados são submetidos pela ANAC, pode-se dizer que na busca pela melhoria continua da qualidade da informação e com o propósito de atingir o maior nível de consistência possível, os dados são submetidos a críticas, validações e procedimentos de auditoria realizados pela Agência. Desta forma, os dados estão sujeitos a revisões, correções e alterações, podendo apresentar diferenças em relação àqueles divulgados anteriormente ou mesmo discrepâncias e observações, conhecidas como outliers, que devem ser consideradas em sua análise.” (ANAC, 2022)

Os dados recebidos pela ANAC estão abertos a todos interessados que desejem baixa-los. A quantidade de dados é grande o suficiente para que precisemos de ferramentas de ciência de dados para tirarmos informações de valor dos dados. Uma ferramenta muito útil em situações como está, é o Machine Learning que é uma das facetas da Inteligência Artificial, sendo considerada também uma parte importante da ciência de dados.

O crescimento acelerado de fontes de dados e posteriormente dos próprios dados fez com que a ciência de dados fosse um dos campos de crescimento mais rápido em todos os

setores. As organizações dependem cada vez mais deles para interpretar dados e fornecer recomendações em atendimento às suas demandas com o objetivo de melhorar os resultados de negócios. (IBM, *s.d*)

Segundo o instituto Brasileiro de Pesquisa e Análise de Dados (IBPAD), a Ciência de Dados é uma atividade interdisciplinar que concilia principalmente duas grandes áreas: Ciência da Computação e Estatística, além de ser aplicada como apoio em diferentes áreas do conhecimento, tais como: Medicina, Biologia, Economia, Comunicação, Ciências Políticas, etc.

No mundo moderno de automação, computação em nuvem, algoritmos, inteligência artificial e big data, dois temas de grande destaque são a ciência de dados e aprendizado de máquina. Essa importância vai muito além de sua aplicação às questões reais do nosso dia a dia, mas também devido sua própria natureza, sendo uma área multidisciplinar, incluindo matemática, estatística, ciência da computação, engenharia, ciência e finanças. (CROESE; BOTEV; TAIMRE; VAISMAN, 2022, tradução nossa).

Este estudo tem como objetivo usar regressão linear para prever a quantidade de litros de combustível que a aeronave precisa de acordo com características intrínsecas dos voos. As ferramentas escolhidas para criar os algoritmos e fazer as verificações necessária foram o Jupyter notebook e Visual Studio Code, sendo sua linguagem padrão “Python”.

2. Revisão bibliográfica

Empesas como a Embraer, fabricantes já consolidadas de aviões, estão focadas em soluções mais econômicas. Segundo a Embraer, devido aos atuais valores de venda de petróleo praticados, existe uma tendência de aumento dos custos operacionais das empresas do ramo da aviação. (EMBRAER, 2006)

Segundo divulgado pela BBC, cerca de 2,4% das emissões globais de CO2 vêm do setor da aviação, ainda acrescenta, as emissões dos aviões estão aumentando de forma muito rápida - elas cresceram 32% entre os anos de 2013 e 2018. (BBC, 2023)

É de extrema importância destacar que a economia de combustível não ocorre apenas por desenvolvimento de novos aviões, mas também por uso mais inteligente dos produtos já disponíveis. Para uma gestão mais eficiente dos recursos energéticos, é imprescindível conhecer muito bem cada vetor atuante no uso de combustível das aeronaves.

Conforme mencionado a aviação comercial gera uma enorme massa de dados, seu controle é rigoroso e por isso tecnologias modernas são utilizadas. Existe uma busca incansável por novas tecnologias, sendo algumas dessas, consideradas disruptivas, estas por sua vez, são diferenciais estratégicos que podem levar empresas aéreas a um outro patamar. A ciência de dados entra como uma ferramenta poderosa ao se determinar causas, padrões e tendências.

De acordo com Neto, a fase de conhecimento dos dados é de muita importância:

> Se torna essencial conhecer, ter experiência e domínio do negócio sobre o problema que está sendo resolvido. Tudo começa por uma pergunta sobre dados, vontade de identificar algo fora da curva, algo que remeta a nova descoberta, um “insight”. (NETO, 2019)

Neto classifica as etapas percorridas ao se praticar a ciência de dados da seguinte forma: (1) Formulação das perguntas corretas; (2) Aquisição dos dados; (3) dados; (5) Análise dos

dados; (6) Demonstração dos dados encontrados; (7) Transformação dos insights em ações. (NETO, 2019)

A respeito da etapa 1, onde são elaboradas as perguntas, neto diz “Procure identificar o problema e descreva os ingredientes que o compõem para uma posterior análise de dados”. (NETO, 2019)

Discorrendo sobre a aquisição dos dados (etapa 2), neto destaca “Dados provêm de vários locais (Armazém de Dados, Redes Sociais, Documentos) estruturados ou não, e devem ser capturados para análise”. (NETO, 2019)

Ao mencionar a etapa 3, onde ocorre a exploração dos dados, Neto acrescenta “Explique a importância e descreva os dados. Processe dados, limpe e os transforme. Identifique métodos para realizar uma análise preliminar com correlações, anomalias, visualização”. (NETO, 2019)

Neto fala a respeito da análise dos dados (etapa 4) “Aplique técnicas de análise de dados, como classificação, agrupamentos, regressão, associação, para identificar as possibilidades. Escolha a melhor delas e construa um modelo para tentar responder as perguntas iniciais”. (NETO, 2019)

Ao relatar o que foi encontrado (etapa5), neto admoesta “Forneça relatórios dos insights, identifique as melhores técnicas de apresentação e de convencimento para comunicar os resultados. Utilize os melhores softwares de visualização e apresentação”. (NETO, 2019)

Ao explicar a etapa de transformação de insight sem ações (etapa 6), neto conclui “Conecte resultados em ações práticas de negócios, em resultados empresariais. Crie produto de dados para a empresa”. (NETO, 2019)

Segundo Rautenberg e Carmo (2019), alguns conhecimentos são imprescindíveis na etapa de visualização dos dados:

> O conhecimento sobre Matemática e Estatística também é necessário para a realização de atividades de Análise de Dados. Ou seja, os profissionais da Ciência de Dados devem entender o funcionamento dos algoritmos de Aprendizado de Máquina, bem como, saber interpretar os resultados, estatisticamente. Interdisciplinarmente, a atividade de interpretação é facilitada pela visualização da informação, a qual privilegia a utilização de elementos de representação gráfica da informação.

Após a visualização inicial é necessário verificar o total de valores faltantes para cada atributo e o quanto isso representa do total de dados. Valores ausentes ou faltantes precisam ser tratados antes de se realizar as análises dos dados pois podem causar distorções e erros ao rodar os códigos de análise no Python.

2.1. A presença de valores vazios (null)na análise dos dados

Em Python a palavra null é usada para representar valores vazios. Esses valores ausentes são esperados ao manipular grande volume de dados, devendo receber um tratamento adequado:

> Dados ausentes estão presentes em muitos problemas da vida real. Geralmente, ao trabalhar com dados incompletos, vetores de recursos, onde um ou mais valores estão ausentes, é típico excluir completamente tal vetor dos dados (que pode distorcer os dados) ou imputar (adivinhar) seus valores ausentes a partir dos dados disponíveis; (CROESE; BOTEV; TAIMRE; VAISMAN, 2022, tradução nossa).

Dados ausentes, conforme já destacado, precisam ser identificados, sendo assim, ao lidar com um conjunto de dados novos, é necessário saber qual é a quantidade e proporção dos dados missingvalues.

A função describe retorna estatísticas descritivas, como por exemplo, desvio padrão, máximo, mínimo e outras tendências centrais, além da forma da distribuição. Isso exclui os valores Nando resumo. Além disso, permite se ter uma ideia sobre a distribuição dos campos de dados e outiliers, se houver. O percentil da saída pode ser personalizado mencionando a faixa de percentis no respectivo parâmetro da função. (LIMA; PERES, 2021).

A aplicação da função info() permite constatar o tipo de cada variável, além dos valores faltantes. (ANSELMO, 2022)

O código df.isnull().sum informa se há algum valor ausente presente em um objeto do tipo array. Esta função retorna valores booleanos após verificar a existência de valores ausentes. Quando estamos criando uma lista com um valor nulo e quando ela é passada pela função isnull(), ela fornece como saída com uma lista booleana.Isso também pode ser bem útil quando é necessário conferir se existem valores ausentes em um grande dataframe. Podemos calcular a soma total de valores ausentes de uma coluna adicionando a função sum() ao final da função isnull(). Desta forma, a função isnull() poderá ser aplicada em todo o dataset, verificando para cada coluna se há algum valor ausente e mostrando o mesmo.(LIMA; PERES, 2021).

2.2. A presença de valores indefinidos (nan) na análise dos dados

Em python um valor indefinido ou irrepresentável é identificado como NaN.

De acordo com Chistian Hill (2020), nan refere-se ao valor que não pode ser definido matematicamente ou não finito.

A escolha de utilizar NaN internamente para representar dados ausentes em um dataframe foi em grande parte por motivos de simplicidade e desempenho. A partir do pandas 1.0, alguns tipos de dados opcionais começam a experimentar um NA escalar nativo usando uma abordagem baseada em máscara.(PANDAS DOCUMENTATION, 2023)

Como os dados vêm em muitas formas, o pandas pretende ser flexível no que diz respeito ao tratamento de dados ausentes. Embora NaN seja o marcador de valor ausente padrão por motivos de velocidade e conveniência computacional, precisamos ser capazes de detectar facilmente esse valor com dados de diferentes tipos: ponto flutuante, inteiro, booleano e objeto geral. Em muitos casos, no entanto, no Python, surgirá None que pode ser também considerado “ausente” ou “não disponível” ou “NA”.(PANDAS DOCUMENTATION, 2023)

É importante lidar de maneira adequada com os valores ausentes. Muitos algoritmos de aprendizado de máquina falham se o conjunto de dados apresenta valores ausentes. No entanto, algoritmos como K-nearest e NaiveBayes oferecem suporte a dados com valores ausentes. Um modelo de aprendizado de máquina tendencioso pode levar a resultados incorretos se os valores ausentes não forem tratados adequadamente.A falta de dados pode levar a uma falta de precisão na análise estatística. (TAMBOLI, 2021)

2.3. A presença de valores negativos na análise dos dados

Valores negativos podem ser encontrados em uma base de dados, porém é necessário conferir se estes são números reais ou são apenas erros da fase de registro e armazenamento dos mesmos.

Números negativos detectados precisam fazer sentido em relação a sua variável. Apenas como exemplo, suponha que estamos registrando em uma coluna de um dataframe o quanto cada voo de avião de uma rota gasta de combustível. Durante a fase de visualização dos dados descobre-se que 80% dos dados são números positivos e 20% são negativos. Em situações como é esta é necessário questionar se fazem sentido estes registros negativos. No

exemplo citado, não faria nenhum sentido manter tais dados, pois não é possível um avião decolar e não ter gasto de combustível e muito menos ter um acréscimo deste. Este foi um exemplo no qual os dados negativos são inconsistentes e podem atrapalhar, caso permaneçam, as análises dos dados.

Blanco, Geb e Pitner(2021, tradução nossa) falam o seguinte sobre a origem de dados inconsistentes. “Isso pode ser causado possivelmente pelo mau funcionamento de determinado dispositivo da Internet das Coisas ou por erros humanos”.

Valores negativos podem ser mantidos se fazerem sentido. Caso esses valores não sejam coerentes poderão ser substituídos por zero ou excluídos. É fundamental, no caso da última opção, o uso de cautela pois a exclusão de um registro afeta todas as outras colunas do data-frame.

2.4. Classificação

Variáveis e atributos são sinônimos, ou seja, é o título de cada coluna. Cada um desses atributos tem um tipo determinado, por exemplo, float aceita números com pontos decimais, int, numéricos inteiros, string, caracteres, além disso Python trabalha com um tipo especial denominado Scategory. Corresponde a uma determinada faixa de valores. (ANSELMO, 2022)

Segundo a documentação oficial do python:

> “Todo objeto tem uma identidade, um tipo e um valor. A identidade de um objeto nunca muda depois de criado; você pode pensar nisso como endereço de objetos em memória. O operador ‘is’ compara as identidades de dois objetos; a função id() retorna um inteiro representando sua identidade”. (PYTHON DOCUMENTATION, 2001)

Anselmo indica um método para converter o tipo do item em atributo categórico da seguinte, sendo este:

```
df['Item_Type'] = df.Item_Type.astype('category')
```

Figura 1- Conversão do tipo do item em atributo categórico.
Fonte: ANSELMO, 2020, p. 73.

Sobre a comando astype, a documentação oficial do Pandas acrescenta entendimento a sua finalidade “Lança um objeto pandas para um dtype especificado”. (PANDAS DOCUMENTATION, 2023)

2.5. Seleção de atributos e variáveis

Um dataframe é formado por colunas e linhas, cada linha representa um registro. As colunas são as varáveis desses registros. Em muitos casos, depois de uma análise inicial constatamos que não é de nosso interesse utilizar certa parte do dataframe, como por exemplo, algumas colunas não estão relacionadas com nossa variável principal ou existem linhas contém dados discrepantes ou vazios. Neste caso talvez consideremos a possibilidade de exclusão de tais itens. Uma possibilidade de realizar tal alteração do dataframe é a aplicação do método drop.

O comando drop pode ser aplicado nesse tipo de situação, de acordo com MCKINNEY (2018), “drop - calcula um novo Index apagando os valores recebidos”.

Apesar de ser um ótimo recurso o comando drop deve ser utilizado com bom critério, pois ao ser acionado sobre um registro, não será apagada apenas a célula com conteúdo

indesejado ou ausente, a linha inteira será apagada do dataframe incluindo todas as outras células existentes na mesma.

Podemos dizer que é comum encontrarmos relações entre varáveis que foram obtidas em um mesmo levantamento. Podemos citar como exemplo o aumento de velocidade com aumento de consumo de combustível, ou seja, para se atingir uma velocidade maior é necessário um acréscimo também na aceleração e isso demanda maior utilização de combustível.

De acordo com Araujo, Santos e Gomes (2019) “O coeficiente de correlação de Pearson mede a correlação linear entre duas variáveis, devendo esta correlação estar compreendida no intervalo de -1 a 1, sendo -1fortemente correlacionadas negativamente e 1 fortemente correlacionadas positivamente”.

Em muitos casos, na ciência de dados, é necessário a criação ou exclusão de certas variáveis. Esse tratamento pode ser feito pela somatória direta entre duas colunas do dataframe. Antes de executar tal comando, é preciso verificar se faz sentido tal junção. Por exemplo, podemos citar um dataframe que possui duas colunas com características de automóveis, sendo uma do o modelo A e a outra do o modelo B. Supomos que queremos saber as informações sobre os carros em geral sem discriminar o modelo do veículo, neste caso seria muito mais viável criar uma coluna nova com a soma das duas colunas. As colunas que deram origem poderão ser descartadas deste dataframe apenas depois de se ter certeza que não serão necessárias em algum momento da respectiva análise. Uma forma prática para juntar as duas variáveis em uma única variável é pelo uso do comando loc de forma direta, adicionando o sinal de + entre as mesmas. O resultado será uma nova coluna no dataframe.

```
df.loc[:, "NOVA_COLUNA"] = df.MODELO_1 + df.modelo_2
```

Figura 2: Criando uma nova coluna com o comando loc.
Fonte: Autor, 2023.

Segundo a documentação oficial da biblioteca pandas, o comando loc pode ser utilizado para acessar um grupo de linhas e colunas por rótulo(s) ou uma matriz booleana, ainda acrescenta que.loc[]é baseado principalmente em rótulos, mas também pode ser utilizado com uma matriz booleana. (PANDAS DOCUMENTATION, 2023)

O comando drop, já mencionado, pode também ser utilizado neste caso para exclusão das variáveis originais.

2.6. Aplicação da regressão linear

Muitas vezes o analista de dados se depara com a necessidade de prever valores de dados desconhecidos. A regressão linear tem a função ajudar em tal situação.

Filho, (*s.d.*) “Quando analisamos dados que sugerem a existência de uma relação funcional entre duas variáveis, surge então o problema de se determinar uma função matemática que exprima esse relacionamento, ou seja, uma equação de regressão”.

O objetivo da regressão linear é estabelecer uma relação linear, quando possível, entre dois conjuntos de variáveis, independente e dependente, respectivamente. (FARIA; OLIVEIRA; PINTO; SZWARCFITER, 2021).

A AWS menciona que os modelos de regressão linear são considerados relativamente simples e fornecem uma fórmula matemática de fácil interpretação possibilitando por sua vez a geração previsões. Pode-se dizer que a regressão linear é uma técnica estatística consolidada e se aplica facilmente pelo uso de softwares e na computação. Muitas empresas aplicam a

regressão linear para conversão de dados brutos de forma confiável e previsível em business intelligence e insights práticos. (AWS, 2023)

Segundo Rong e Bao-wen (2018) o Python tem recebido um destaque especial na área do machinelearning “Python, como a linguagem de programação mais popular no campo de aprendizagem de máquina, tem sido usado cada vez mais amplamente”.

Ainda a AWS acrescenta sobre a necessidade do uso de duas variáveis para plotagem de um gráfico:

> Em sua essência, uma técnica de regressão linear simples tenta traçar um gráfico de linhas entre duas variáveis de dados, x e y. Como variável independente, x é plotada ao longo do eixo horizontal. Variáveis independentes também são chamadas de variáveis explicativas ou variáveis preditoras. A variável dependente, y, é plotada no eixo vertical. Você também pode fazer referência aos valores de y como variáveis de resposta ou variáveis previstas. (AWS, 2023)

Menotti (*s.d*) considera que a relação existente entre a resposta e as variáveis é uma função linear de alguns parâmetros.

O modelo de regressão linear simples e o modelo de regressão linear múltipla podem ser definidos da seguinte forma:

Regressão Linear Simples: Quando existe uma relação casual entre duas variáveis, e pode ser traçada uma reta. Neste tipo de regressão temos uma variável denominada dependente, e uma outra denominada independente. Ela é utilizada para determinar a equação da reta ajustada (modelo matemático linear). (FARIA, 2016)

Regressão Linear Múltipla: Quando existe uma relação casual com mais de duas variáveis. Isto é, quando o comportamento de Y é explicado por duas ou mais variáveis independentes X1, X2,....Xn. Essa técnica é uma solução adequada para se utilizar quando se deseja investigar simultaneamente os efeitos, sobre Y, de duas ou mais variáveis preditoras. (FARIA, 2016)

Antes de executar uma regressão linear, é importante conferir se os dados estão adequados a este procedimento de análise. Os dados devem estar atendendo a certas premissas, conforme o site oficial da IBM menciona: (IBM, 2015)

- “As variáveis devem ser medidas a nível contínuo. Exemplos de variáveis contínuas são tempo, vendas, peso e pontuações de teste”. (IBM, 2015)

- “Use um gráfico de dispersão para descobrir rapidamente se há um relacionamento linear entre essas duas variáveis”. (IBM, 2015)

- “As observações devem ser independentes umas das outras (isto é, não deve haver dependência)”. (IBM, 2015)

- “Seus dados não devem possuir valores discrepantes significativos”. (IBM, 2015)

- “Verifique a homoscedasticidade, que é um conceito estatístico no qual as variações ao longo da linha de regressão linear de melhor ajuste permaneçam semelhantes por toda a linha”. (IBM, 2015)

- “Os resíduos (erros) da linha de regressão de melhor ajuste seguem a distribuição normal”. (IBM, 2015)

Conforme mencionado por Filho (*s.d.*), o comportamento conjunto de duas variáveis quantitativas pode ser analisado por meio do gráfico de dispersão.

2.6.1. Homocedasticidade e heterocedasticidade

A hipótese de homoscedasticidade é uma condicional às variáveis explicativas, a variância do erro, u, é constante.Se isso não for verdade, ou seja, se a variância é diferente para diferentes valores de x's, então os erros são considerados heterocedásticos. (USP, *s.d.*)

É necessário se preocupar com heterocedasticidade. Primeiro precisamos lembrar que o Método dos Mínimos Quadrados (MMQ)continua não tendencioso e consistente, mesmo sem a hipótese de homocedasticidade. Outro ponto é que os erros-padrão dos coeficientes estimados serão viesados se há heterocedasticidade, nesta situação, se os erros-padrão são viesados, não podemos utilizar as estatísticas t, F e LM usuais. (USP, *s.d.*)

A aplicação do método dos mínimos quadrados, conforme Filho, poderá ser utilizado para estimar os parâmetros do modelo (α e β) e consiste em fazer com que a soma dos erros quadráticos seja minimizado, ou seja, este método consiste em obter os valores d e α e β que reduzem a expressão: (FILHO, sd)

$$S = \sum \varepsilon_i = \sum (Y_i - \alpha - \beta x_i)^2$$

Figura 3- valores d e α e β que minimizam a expressão.
Fonte: FILHO, s.d.

O teste de Breusch-Paganpermite detectar formas de heterocedasticidade lineares, já o teste de White permite encontrar os não-linearidades por utilizar quadrados e produtos cruzados de todos os x's. Basta computar a estatística F ou LM para testar se todos os xj, xj2 e xjxh são conjuntamente significativos. (USP, *s.d.*)

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error.$$

Figura 4 - Teste de significância.
Fonte: USP, *s.d.*

Existe uma forma alternativa de aplicação do teste deWhite. Suponha que o valores ajustado por MQO, ŷ, é função de todos os x's. Logo, ŷ2 será função dos quadrados e produtos cruzados e, portanto, ŷ e ŷ2 serão proxies para todos osxj, xj2 e xjxh. Em tal situação é necessário fazer a regressão dos resíduos ao quadrado em ŷ e ŷ2 e use o R2 para obter a estatística F ou LM, sendo este teste para apenas duas restrições. (USP, *s.d.*)

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error.$$

Figura 5- Forma alternativa do teste de White
Fonte: USP, *s.d.*

O Método dos Mínimos Quadrados Ponderados pode ser usado para construir funções que modelam fenômenos de diferentes naturezas a partir de dados observacionais, sendo particularmente útil para fenômenos nos quais algumas medidas são mais precisas do que outras. (VAZ; BECK, 2012)

Embora seja possível estimar os erros-padrão robustos para os estimadores de MQO, considerando o Método dos Mínimos Quadrados Ponderados, se soubermos alguma coisa sobre a forma especificada heterocedasticidade, poderemos obter estimadores mais eficientes que os de MQO.Como devemos especificar a natureza da heterocedasticidade, o processo de

estimação é mais trabalhoso. A ideia básica é transformar o modelo em outro cujos erros sejam homocedásticos. (USP, *s.d.*)

É importante ressaltar que se utiliza MQP apenas por eficiência, pois MQO continua não tendencioso e consistente. As estimativas serão diferentes devido a erros amostrais, mas se forem muito diferentes, podemos considerar que alguma outra hipótese de Gauss-Markov também deve estar sendo violada.(USP, *s.d.*)

2.6.2. Calculo do P valor

O valor-p é definido, segundo Ferreira e Patino (2015), como:

> O valor-p é definido como a probabilidade de se observar um valor da estatística de teste maior ou igual ao encontrado. Tradicionalmente, o valor de corte para rejeitar a hipótese nula é de 0,05, o que significa que, quando não há nenhuma diferença, um valor tão extremo para a estatística de teste é esperado em menos de 5% das vezes.

Muitos pesquisadores acham que o valor-p é o número mais importante a ser relatado. No entanto, é fundamental se concentrar no tamanho do efeito. É necessário relatar o valor-p de forma isolada e, preferencialmente, relatar os valores médios para cada grupo, a diferença, o intervalo de confiança de 95% e, então, o valor-p. (FERREIRA; PATINO, 2015)

2.6.3.Overfitting

O Overfitting é algo que deve ser levado à sério pelo cientista de dados, pois pode induzir ao erro:

> O overfitting do modelo é um problema sério e pode fazer com que o modelo produza informações enganosas. Uma das técnicas para superar o overfitting é a Regularização. A regularização, em geral, penaliza os coeficientes que causam o overfitting do modelo (SHAH, 2021).

O overfitting do modelo ocorre quando o modelo aprende "bem demais" sobre os dados. Isso pode parecer uma vantagem, mas não é. Quando um modelo é submetido a um treinamento excessivo dos dados, ele apresenta pior desempenho nos dados de teste ou em quaisquer novos dados fornecidos. Podemos dizer que o modelo aprende os detalhes, bem como o ruído dos dados. Isso seria prejudicial ao desempenho de quaisquer novos dados fornecidos ao modelo, pois os detalhes aprendidos e o ruído não podem ser aplicados aos novos dados. Este é o caso quando dizemos que o desempenho do modelo não é adequado. Existem várias formas de evitar o overfitting de um modelo, como validação cruzada K-fold, reamostragem, redução do número de recursos, etc. Uma das técnicas é aplicar a regularização ao modelo. (SHAH, 2021).

2.7. Livros utilizados neste projeto

Autor	Livro	Contribuição para este artigo
Dirk P. Kroese, Radislav Vaisman, Thomas Taimre, Zdravko I. Botev	Data Science and Machine Learning - Mathematical and Statistical Methods	Fundamentação da ciência de dados e tratamento de dados ausentes
Flávia Chein	Introdução aos Modelos de Regressão Linear	Analise de regressão linear
Fabiano de Souza Oliveira Jayme Luiz Szwarcfiter Luerbio Faria Paulo Eustáquio Duarte Pinto	Ciência de Dados: Algoritmos e Aplicações	Entendimento sobre o objetivo da regressão linear
Christian Hill	Learning Scientific Programming with Python	Tratamento de valores indefinidos (nan)
Wes McKinney	Python para Análise de Dados: Tratamento de Dados com Pandas, Numpy e Ipython	Execussão da limpeza dos dados
Jose Antonio Ribeiro Neto	Big Data para Executivos e Profissionais: Tecnologias, Aplicações e Carreira	Tratamento dos dados
Alessandro Oliveira	Transporte Aéreo: Economia e Políticas Públicas	Conhecimentos sobre o transporte aéreo

Figura 6 –Bibliografias
Fonte: Autor, 2023.

3. Estudo de caso

3.1. A base de dados

Conforme já mencionado, foi utilizada neste estudo uma base de dados que se encontra disponível para download, sendo esta, uma fonte rica de informações sobre voos que ocorrem dentro do território brasileiro, possuindo um total de 38 colunas e 945.830 linhas e abrangendo um período que inicia no ano de 2000 e vai até 2022. Os registros dos voos estão relacionados a locais de decolagens e pousos, distâncias de deslocamento, peso das cargas e número passageiros.

Segundo informado no site da ANAC, a intenção da disponibilização dos dados é possibilitar a ampliação do conhecimento e do entendimento sobre o tema:

> Com o intuito de ampliar o conhecimento da sociedade brasileira e de subsidiar a realização de pesquisas, estudos e análises mais abrangentes sobre o setor, a ANAC tem disponibilizado, na seção "Dados e Estatísticas" do seu portal na internet, relatórios, estudos e informações sobre as condições de mercado. (ANAC, 2022)

Reforçando o livre acesso e a qualidade de sua base de dados, a Agência menciona: “para a livre consulta de qualquer interessado, a série histórica dos dados estatísticos do transporte aéreo do Brasil, com elevado grau de detalhamento” (ANAC, 2022)

Segundo a ANAC, as etapas básicas são aquelas realizadas pela aeronave desde seu ponto de partida (decolagem) até o próximo pouso, independentemente de onde tenha sido realizado o embarque ou o desembarque do objeto de transporte. Em cada etapa, são gerados dados estatísticos do voo, demonstrando a movimentação de cargas e de passageiros entre os respectivos aeroportos. Os dados referentes a operação são aqueles que são gerados entre a

decolagem até o próximo pouso, ou seja, a ligação entre dois aeroportos. As variáveis que se referem diretamente aos aeroportos são: AEROPORTO DE ORIGEM (SIGLA); AEROPORTO DE ORIGEM (NOME); AEROPORTO DE ORIGEM (UF); AEROPORTO DE ORIGEM (REGIÃO); AEROPORTO DE ORIGEM (PAÍS); AEROPORTO DE ORIGEM (CONTINENTE); AEROPORTO DE DESTINO (SIGLA); AEROPORTO DE DESTINO (NOME); AEROPORTO DE DESTINO (UF); AEROPORTO DE DESTINO (REGIÃO); AEROPORTO DE DESTINO (PAÍS) e AEROPORTO DE DESTINO (CONTINENTE). (ANAC, 2022)

A base de dados ainda possui outras variáveis, as numéricas, ou seja, aquelas que podem ser quantificadas, sendo estas: ANO; MÊS; PASSAGEIROS_PAGOS; PASSAGEIROS_GRATIS; CARGA_PAGA_KG; CARGA_GRATIS_KG; CORREIO_KG; ASK; RPK; ATK; RTK; COMBUSTIVEL_LITROS; DISTANCIA_VOADA_KM; DECOLAGENS; CARGA_PAGA_KM; CARGA_GRATIS_KM; CORREIO_KM; ASSENTOS; PAYLOAD, HORAS_VOADAS; BAGAGEM_KG.

Para facilitar o entendimento da base de dados da ANAC, considere sobre as siglas/nomes de algumas das variáveis:

- ASK: “número de assentos disponíveis multiplicado pelos quilômetros voados”. (AZUL, 2019)

- RPK: “passageiros pagantes transportados em um quilômetro. O RPK é calculado ao multiplicar-se o número de passageiros pagantes pelos quilômetros voados. (AZUL, 2019)

- ATK: “soma do produto entre o payload, que é a capacidade total de peso disponível na aeronave, expressa em quilogramas, disponível para efetuar o transporte de passageiros, carga e correio, e a distância das etapas, dividido por 1.000”. (ANAC, 2022)

- RTK: “soma do produto entre os quilogramas carregados pagos, onde cada passageiro possui o peso estimado de 75 Kg, e a distância das etapas, dividido por 1.000”. (ANAC, 2022)

- Payload: “capacidade total de peso na aeronave, expressa em quilogramas, disponível para efetuar o transporte de passageiros, carga e correio”. (ANAC, 2022)

O tratamento dos dados envolve toda manipulação dos dados, a primeira parte desse trabalho de pesquisa foi a realização de download dos dados abertos disponibilizados pelo site da ANAC denominados “METADADOS DO CONJUNTO DE DADOS: DADOS ESTATÍSTICOS DO TRANSPORTE AÉREO”. Mas antes de continuar este estudo foi necessário realizar um pré-tratamento dos dados.

3.2 Tratamento e Análise de Dados

Na primeira etapa os dados foram lidos e armazenados em um dataframe. Após a importação foi realizada a primeira visualização dos dados com a função head() do método Pandas através da IDE Jupyter Notebook para conhecimento e familiarização do mesmo. Por padrão o Python coloca apenas as 5 primeiras linhas, mas em nosso projeto colocamos as 100 primeiras linhas dessa base de dados, para ter uma noção inicial da forma como os registros estão organizados e verificar se os dados e sua estrutura estão corretos.

3.2.1. Identificando e excluindo valores negativos da variável COMBUSTIVEL_LITROS

A verificação da existência de valores negativos em relação a variável dependente denominada 'COMBUSTIVEL_LITROS' foi vital para as próximas etapas, partindo do pressuposto que todo voos tiveram um gasto de combustível representado por um valor

positivo. Um valor negativo seria equivalente a um aumento de combustível no tanque do avião, por isso tais valores, no caso de sua existência, precisavam ser identificados, avaliados e submetidos a algum tratamento adequado.

A identificação desses valores foi feita com o código (df['COMBUSTIVEL_LITROS'] < 0).sum().sum(). A utilização de <0 realizou a filtragem de valores menores que zero, ou seja, os valores negativos referentes ao gasto de combustível. O resultado do output foi um total de 27 ocorrências de valores negativos. Essa visualização tornou possível inferir que a quantidade de valores negativos seria irrisória em relação ao total de registros. Por se tratar de quantidade bem pequena de valores negativos, foi decido pela exclusão destes dados, pois não iria causar impacto significativo em nosso projeto de pesquisa. Depois de excluir todos os negativos desta coluna, foi feita uma conferência e constatou-se 0 registros de negativo.

3.2.2. Visualizando informações das colunas

O método info() tornou possível obter de forma rápida uma descrição dos dados, como total de linhas, número de colunas, tipo de cada série e se há valores nulos.

O output deste comando destacou a presença de tipos de colunas identificadas de forma equivocada. As variáveis HORAS_VOADAS, DISTANCIA_VOADA_KM, ASK, RPK que deveriam ser do tipo float estavam rotuladas como strings.

Como o dataframe foi criado a partir de um arquivo csv, as colunas foram importadas e o tipo de dados foi definido automaticamente, e alguns tipos de colunas não foram atribuídas corretamente quanto ao seu tipo.

Como o método de entrada em python aceita o objeto string para entrada do usuário, foi necessário convertê-los explicitamente em float para que fosse possível realizar as operações necessárias sobre eles, como adição, multiplicação, etc.

A transformação das strings em floats com o uso da função replace() juntamente com a função astype().A função replace() possibilitou substitui tipos textuais (strings) para tipos decimais (floats). A palavra replace(), do inglês, significa substituir. A função astype( ) foi usada para invocar os tipo das colunas e o tipo ao qual seriam transformadas.

Após a transformação, as variáveis HORAS_VOADAS, DISTANCIA_VOADA_KM, ASK, RPK apareceram corretamente como floats.

3.2.3. Realizando a seleção das variáveis para etapa de correlação

A seleção das variáveis foi feita para adequar os dados para as etapas seguintes. Como critério foram escolhidas varáveis numéricas pois essas poderiam ser utilizadas na etapa de regressão.As variáveis numéricas selecionadas foram: PASSAGEIROS_PAGOS; PASSAGEIROS_GRATIS; PASSAGEIROS; CARGA_PAGA_KG; CARGA_GRATIS_KG; CARGA_KG; CORREIO_KG; ATK; RTK; COMBUSTIVEL_LITROS; DECOLAGENS; CARGA_PAGA_KM; CARGA_GRATIS_KM; CARGA_KM; CORREIO_KM; ASSENTOS; PAYLOAD; BAGAGEM_KG e LOTACAO.Após a escolha das variáveis que seriam úteis para continuidade deste projeto, atribuiu-se as mesmas a um novo dataframe, para facilitar sua manipulação e visualização.

3.2.4. Criação de novas variáveis

Para melhorar a análise dos dados foram criadas cinco novas variáveis a partir de outras variáveis já existentes. A variável CARGA_KM surgiu da soma das variáveis CARGA_PAGA_KM com a variável CARGA_GRATIS_KM. Essa somatória foi realizada porque era preciso descobrir a correlação da CARGA_KM (carga total) com

COMBUSTIVEL_LITROS. O mesmo raciocínio levou a criação das variáveis CARGA_KG e PASSAGEIROS.

A variável LOTACAO foi criada pela divisão da variável PASSAGEIROS pela variável ASSENTOS multiplicado por 100 que apontou a taxa de aproveitamentos dos assentos ocupados por voo.

Outra variável que foi introduzida foi a VOO_DE_CARGA. Essa variável serviu para selecionar os registros exclusivos de voos de carga. Para isso foi desconsiderando os voo de passageiros atribuindo df.ASSENTOS == 0. O raciocínio usado neste caso foi considerar que os voos com zero assentos se referiam aos dos aviões de cargas.

3.2.5. Analisando a correlação entre as variáveis

Antes de começar a análise de regressão linear, precisou ser detectada a correlação entre as variáveis. A correlação mediu a força ou o grau de relacionamento entre as variáveis. Nosso objetivo nesta etapa do tratamento foi descobrir qual o nível de correlação das variáveis independentes com a variável dependente “COMBUSTIVEL_LITROS”.

Para a realização da análise de correlação utilizou-se o método heatmap da biblioteca seaborn. O heatmap é um mapa de calor que indica no nosso caso: (1) quanto mais forte a cor azul marinho, maior é a correlação existente entre as variáveis e (2) quanto mais forte for a cor vermelha, menoré correlação entre elas. O output do heatmap gerou o seguinte gráfico:

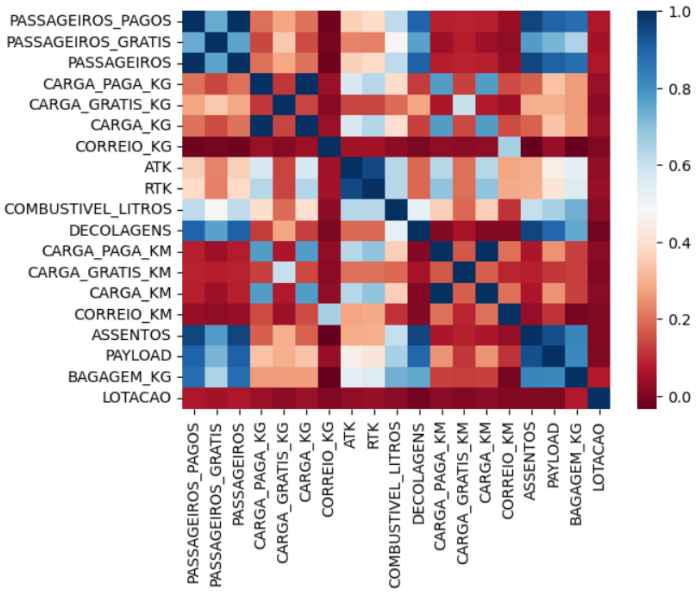

Figura 7 - Visualização da correlação usando o heatmap.
Fonte: Autor, 2023.

A inserção dos dados no código é o argumento mais nobre na função, pois o objetivo final é traçar uma correlação. O método corr() adicionou os dados e se passou como primeiro argumento possibilitando a correlação.

A visualização proporcionada pelo heatmap permitiu selecionaras variáveis que apresentaram alguma correlação com a variável dependente (COMBUTIVEL_LITROS), porém,foi levado também em consideração variáveis independentes que possuem alta correlação entre si. Essa seleção obteve como resultado as seguintes variáveis: BAGAGEM_KG; PAYLOAD; ASSENTOS; DECOLAGENS; ATK; RTK; PASSAGEIROS; desta vez incluindo a variável VOO_DE_CARGA.

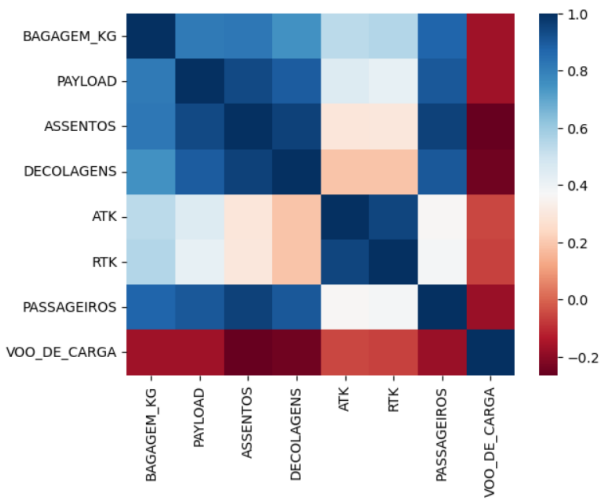

Figura 8 - Visualização da correlação existentes entre as variáveis independentes.
Fonte: Autor, 2023.

Após a visualizações proporcionadas pelo heatmap, chegou-se a um conjunto de variáveis apropriadas, um dataframe final (df_final), que será usado na realização da regressão linear: BAGAGEM_KG; PAYLOAD; ASSENTOS; DECOLAGENS; ATK; RTK; PASSAGEIROS; VOO_DE_CARGA e COMBUTIVEL_LITROS.

3.2.6. Identificando e excluindo valores igual a zero da variável COMBUSTIVEL_LITROS

Assim como no caso da limpeza dos registros negativos da variável COMBUSTIVEL_LITROS, um voo com zero gasto de combustível é irreal. Para exclusão de tais valores do dataframe final foi utilizado o código df_final = df_final[df_final.COMBUSTIVEL_LITROS> 0]. Desta forma apenas valores maiores que zero seriam mantidos em df_final.

3.2.7. Criando variáveis dummies para os meses

No início deste projeto havia a suspeita de influência temporal exercida pelos meses sobre o consumo de combustível. Por esse motivo decidiu-se pela criação de dummies para cada mêsa partir da variável MES. Este processo foi feito pelo código dummies_mes = pd.get_dummies(df_final.MES, prefix='MES') que possui a função de criar as dummies e juntar ao df final. Para etapa de regressão, foi excluído o mês de janeiro, pois os 11 meses já seriam suficientes para a análise do aspecto temporal.

3.2.8. Realizando a regressão linear

Na etapa de regressão, o eixo x foi atribuído às variáveis independentes e o eixo y a variável dependente. Após definir-se o x e y foi feito separação na baseem treino e teste para criação do modelo.

O treinamento envolveu a apresentação dos dados ao algoritmo de machinelearning para criação do modelo. Determinou-se que seria usado 70% da totalidade dos dados para essa finalidade. Para o teste, o valor determinado foi de 30%. Esses dados foram apresentados ao modelo, simulando previsões reais, permitindo assim a verificação de seu desempenho real.

Para que o algoritmo pudesse realizar os cálculos de forma eficiente e precisa foi necessário adicionar uma constante a matriz X. No primeiro treinamento, selecionamos 19 variáveis que acreditamos ter uma relação com a variável COMBUTIVEL_LITROS.

Para criação do modelo, foram utilizados os seguintes métodos: OLS, fit, summary e predict.

OLS foi implementado para melhorar o ajuste do modelo através da minimização dos quadrados do erro de regressão, já método fit() foi empregado para ajustar o modelo aos dados fornecidos. Após o modelo de regressão estar ajustado e armazenado em results aplicou-se o método summary() para visualizar os resultados da regressão OLS. O método predict() permitiu prever os rótulos dos valores de dados com base no modelo treinado.

4. Análise dos resultados

A regressão é uma equação que descreve um relacionamento entre variáveis em linguagem matemática. Ela é o estabelecimento de uma reta que representa a correlação entre as variáveis, sendo, portanto, uma média. A primeira rodagem do modelo já apresentou um coeficiente de determinação de 95% que foi considerado bom a análise.

A regressão linear múltipla é dada por $Y_i=\beta_0 + \beta_1 X_i + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$, onde ϵ é um erro aleatório. Evidentemente quando é acrescentado mais variáveis na equação, resulta-se em mais problemas vistos que a variável dependente acaba condicionada a valores de n fatores, no entanto o cientista de dados deve se atentar com o R-squared ou R^2 (em português R-quadrado), pois ele indica a variabilidade total do modelo de regressão.

A criação das dummies dos meses possibilitou verificar a relevância do aspecto temporal no consumo de combustível, ou seja, ver se existia algum padrão temporal. O mês de abril apresentou, no início da etapa de treinamento e teste, um p-valor bem alto de 0.938, ou seja, sem nenhuma relevância para o propósito deste projeto, sendo este excluído. A cada vez que o código do treinamento e teste era executado retirava-se o mês com p-valor mais alto. Assim foram extraídas sucessivamente as variáveis dos meses com exceção de maio e agosto que ficaram dentro do intervalo de confiança. A cada execução do código um modelo diferente é treinado. Devido o valor de R-quadrado já estar alto e ter restado apenas dois meses, optou-se pela exclusão completa dos meses, pois não seria possível considerar o aspecto temporal, além do que, a não retirada dos meses não apresentaram mudança no R-quadrado. Como haviam várias outras variáveis ainda disponíveis, a exclusão dos meses não resultaria em precarização.

Os meses das variáveis descartadas apresentaram os seguintes p-valor:

MESES	P-VALOR	MESES	P-VALOR
abril	0.938	outubro	0,062
março	0,859	setembro	0,527
novembro	0,508	janeiro	0,685
fevereiro	0,639	junho	0,184
julho	0,094	maio	0,000
outubro	0,062	agosto	0,014

Figura 9 – Variáveis descartadas (meses).
Fonte: Autor, 2023.

Após a exclusão dos meses, restaram apenas 6 variáveis mais a constante que é o beta zero, ou seja, aquelas cujos coeficientes são menores que o valor estabelecido de 0,05, sendo assim, considerados relevantes para o modelo. Veja o modelo final da regressão linear na figura 9.

OLS Regression Results

Dep. Variable:	COMBUSTIVEL_LITROS	**R-squared:**	0.947
Model:	OLS	**Adj. R-squared:**	0.947
Method:	Least Squares	**F-statistic:**	1.187e+06
Date:	Wed, 26 Apr 2023	**Prob (F-statistic):**	0.00
Time:	19:45:01	**Log-Likelihood:**	-5.1074e+06
No. Observations:	400716	**AIC:**	1.021e+07
Df Residuals:	400709	**BIC:**	1.021e+07
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-703.4156	178.048	-3.951	0.000	-1052.385	-354.447
PAYLOAD	-0.0869	0.001	-108.397	0.000	-0.089	-0.085
ASSENTOS	12.5922	0.097	129.876	0.000	12.402	12.782
DECOLAGENS	1062.0613	10.406	102.058	0.000	1041.665	1082.458
ATK	0.1847	0.000	558.104	0.000	0.184	0.185
RTK	0.0155	0.000	31.543	0.000	0.015	0.016
VOO_DE_CARGA	8779.8069	384.936	22.808	0.000	8025.344	9534.270

Omnibus:	177969.694	**Durbin-Watson:**	1.999
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	720099762.657
Skew:	-0.102	**Prob(JB):**	0.00
Kurtosis:	210.675	**Cond. No.**	6.48e+06

Figura 10 - Modelo final.
Fonte: Autor, 2023.

Foi utilizado o valor do R-quadrado (coeficiente de determinação falar) para comparar os modelos e poder no final chegar no modelo que agrada mais.

No teste de hipótese avaliou-se se os betas, coeficientes de cada variável é zero ou não.

A equação aplicada ao modelo final de regressão utilizado neste projeto ficou da seguinte forma:

COMBUSTIVEL_LITROS = -703 + -0,08.PAYLOAD + 12.ASSENTOS + 1062.DECOLAGENS + 0,18.ATK + 0,01.RTK + 8779.VOO_DE_CARGA

A princípio pareceu um tanto estranho a variável PASSAGEIROS apresentar coeficiente negativo. Considerou-se, por essa razão, sua remoção da análise, porém com uma reflexão mais a fundo, foi considerado que poderia se tratar de uma relação entre peso de passageiro e peso de carga. Partindo desta primícia um metro quadrado ocupado por carga até o limite máximo permito do avião representaria um peso bem maior do que o peso de um passageiro, por sua vez um peso maior exigiria um maior consumo de combustível.

A variável BAGAGEM_KG também apresentou valor negativo de 0,07 litros para consumo de combustível, o que causou, como no caso de passageiros, certa estranheza. Mas empregando o mesmo raciocínio usado para variável PASSAGEIROS, cada metro quadrado de carga como aproveitamento até o limite máximo da capacidade representaria um peso maior do que se estivesse com bagagens de passageiros, levando em consideração que nela predomina peças de vestuário, que em geral são compostos de materiais leves.

O decréscimo no uso de combustível tem outro ponto a se considerar sobre a variável BAGAGEM_KG, que o aumento de bagagem no vôo é um indicativo de um número maior de passageiro, ou seja, existe uma relação direta entre as variáveis PASSAGEIROS e BAGAGEM_KG.

A variável PAYLOAD exibiu um coeficiente negativo de 0,08 litros por unidade payload. Considerou-se a hipótese de tal valor negativo estar sendo influenciado pela variável PASSAGEIROS, pois esta é um dos elementos que compõe o PAYLOAD.

A variável ASSENTOS teve um coeficiente de 13 litros por unidade, isso significa um avião equivalente ao E195-E2 da Embraer que possui 124 assentos tem um acréscimo de consumo de 1.612 litros de combustível. O consumo de combustível a cada decolagem, obtido da variável DECOLAGENS, sofre um aumento de 1.031 litros de combustível. Isso significa que a cada cem voos ocorre um acréscimo de 103.100 litros consumidos. ATK e RTC tiveram coeficientes de 0,18 litros e 0,12 litros de combustível consumidos respectivamente, já a variável VOO_DE_CARGA representa um acréscimo de 8711 litros.

Na análise dos resíduos era esperado média zero e variância constante. Conforme o gráfico da figura 8 os valores estão em torno de zero com uma variância moderada.

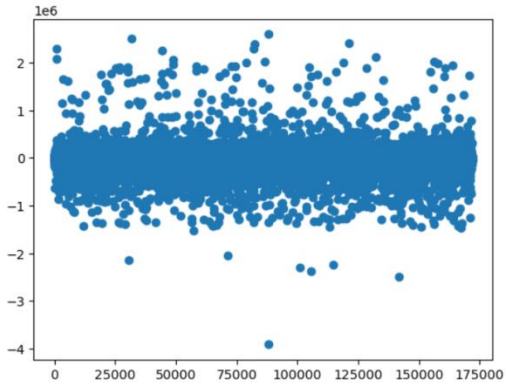

Figura 11 -Análise dos resíduos
Fonte: Autor, 2023.

Para confirmar a qualidade do modelo, realizou-se o t-test, que é um teste de hipótese.

O valor do p-valor obtido ao executar o t-test foi de 0.90. Neste teste o p-valor maior que 0.05 indicou que as medias são iguais entre os grupos, ou seja, o resíduo é zero.

Toda vez escolhemos um modelo devemos levar em consideração o coeficiente de determinação que é R-squared e análise dos resíduos que são critérios para estabelecer se nosso modelo é bom ou não. Desta forma este modelo foi validado pois atendeu aos requisitos mínimos estabelecidos por esses critérios.

5. Conclusões e trabalhos futuros

O presente estudo, através de uma análise de regressão linear, buscou estimar a influência de variáveis independentes sobre o consumo de combustível. A regressão linear foi realizada de forma prática com uso da linguagem Python.

Na etapa de limpeza dos dados, algumas variáveis foram excluídas pois não eram de interesse em relação ao tema deste projeto, porém é necessário dar atenção a uma delas, a variável CORREIO_KG. A variável CORREIO_KG foi retirada, pois apresentou um valor quase igual a zero de correlação com a variável dependente. Isso significa que não foi impactante no consumo de combustível. No entanto, foi considerado a possibilidade do tipo de desse material ser bem mais leve que outros tipos de carga. Apesar de a variável CORREIO_KG não ter entrado na etapa de regressão, seria interessante para as companhias aéreas verificar a viabilidade de aumentar o espaço destinado para esse tipo de transporte, pois poderia resultar em um aumento na receita que não traria gastos adicionais significativos com combustível.

Conforme os objetivos deste projeto, os procedimentos de regressão linear, os de treinos e testes possibilitaram chegar a um modelo ajustado com 95% de confiança.Pode-se dizer que foi escolhido um modelo que evita um overfitting dos dados. Isso é um bom indicador porque um modelo com overfitting seria incapaz de realizar uma estimativa boa para dados novos.

Não foi utilizado a biblioteca scikit-learn mas sim a biblioteca statsmodels pois esta retorna o p-valor. O p-valor permitiu selecionar as variáveis do modelo final, ou seja, as que possuíam maior influência sobre a variável combustível, sendo estas: BAGAGEM_KG; PAYLOAD; ASSENTOS; DECOLAGENS; ATK; RTK e VOO_DE_CARGA.

Este projeto poderá contribuir em estudos futuros de consumo de combustível de aviões de passageiros, de cargas e mistos, além de trabalhos de pesquisa voltados ao aumento da margem de lucro devido a maior eficiência relacionada a redução de consumo do mesmo.

No início deste estudo, pensava-se haver uma relação temporal dos dados entre os meses e a varial dependente, ou seja, uma variação da quantidade de combustível de acordo com a demanda existente em cada mês dos voos. É de conhecimento comum que existe sazonalidade no transporte aéreo.

Os meses não relevantes não captaram a variabilidade intrínseca dos dados para explicar a quantidade de combustível consumida. Porém na etapa do treinamento e teste do modelo foi observado a existência de outras variáveis que podiam explicar muito melhor o consumo de combustível.

Foi observado que estas informações foram suficientes para obter uma estimativa de combustível gasto. A quantidade de variáveis do modelo final se mostrou capaz de atender as necessidades deste projeto. As correlações foram consideradas lineares.

7. Referências

ALVES, C.J.P.; CAETANO, M; Innovation system in airtransport management. Scielo, Goiás, [*s.d.*]. Disponível em: <https://www.scielo.br/j/jistm/a/x77MYrG7Kj74nfMMqLfMQbB/?lang=en>. Acesso em 06 de abr. de 2023.

ANAC. Metadados do conjunto de dados: Registro Aeronáutico Brasileiro- Agência Nacional de Aviação Civil - ANAC. Disponível em: <https://www.anac.gov.br/acesso-a-informacao/dados-abertos>. Acesso em: 05de maio de 2022.

ANSELMO, F.; Machine Learning na Prática - modelos em Python, 1.ed. [S.I.:s.n.], 2020. 47p. ibid., p. 73.

ARAUJO, J. V. G. A.; GOMES, C. F. S.; SANTOS, M. Desenvolvimento de um código em Python para geração de matrizes de correlação de Pearson com laços a partir de "n" variáveis tomadas duas a duas. In: SIMPOSIO DE PESQUISA OPERACIONAL E LOGISTICA DA MARINHA, 19., 2019, Rio de Janeiro, RJ. Anais […]. Rio de Janeiro:Centro de Analises de Sistemas Navais, p. 1, 2019. ibid., p. 3.

AWS. O que é regressão linear? Disponível em: <https://aws.amazon.com/pt/what-is/linear-regression/#:~:text=A%20regress%C3%A3o%20linear%20%C3%A9%20uma,independente%20como%20uma%20equa%C3%A7%C3%A3o%20linear>. Acesso em: 24 de abr. de 2023.

AZUL.Glossário. Disponível em:<https://ri.voeazul.com.br/servicos/glossario/#:~:text=Assentos%2Dquil%C3%B4metro%20oferecidos%20(ASK)%3A,total%20de%20assentos%2Dquil%C3%B4metro%20oferecidos.>Acesso em: 30 de abr. de 2023.

BAO-WEN, Z.; RONG, S. The research of regression model in machine learning field. Disponível em:<https://www.matec-conferences.org/articles/matecconf/pdf/2018/35/matecconf_ifid2018_01033.pdf>. Acesso em: 26 de abr. de 2023.

BBC. A aviação pode se tornar sustentável um dia? Disponível em: < https://www.bbc.com/portuguese/articles/crgn8ny2e8xo>. Acesso em: 19 de jan. de 2024.

BLANCO, J.M.; GEB, M.; PITNER, T. Modeling Inconsistent Data for Reasoners in Web of Things. Disponível em: <https://reader.elsevier.com/reader/sd/pii/S1877050921016197?token=F61C0F12197BD9082E437FB1F2F5D34BB7D421E7D3BDF4F268FCA5249CDCA6264353A2900BD6C62085C89C5CECB2C58E&originRegion=us-east-1&originCreation=20230430112431>. Acesso em 30 de abr. 2023.

BOTEV, Z.I.; CROESE, D.P; TAIMRE, T; VAISMAN, R. Data Science and Machine Learning - mathematical and statistical methods, 1.ed. Queensland e New South Wales:[*s.n.*],2022. 13p.

ibid., p. 301p.

CARMO, P.R.V.; RAUTERNBERG, S. Big Data e Ciência de Dados: complementariedade conceitual no processo de tomada de decisão. Brazilian Journal of Information Studies, 22 de mar. de 2019. Disponível em: <https://brapci.inf.br/index.php/res/download/112105>. Acesso em 13 de abr. de 2023.

CHEIN, F. Introdução aos Modelos de Regressão Linear, 1.ed. Brasília: Enap - Escola Nacional de Administração Pública, 2019. 9p.

EMBRAER. Informação à imprensa: Embraer Anuncia Melhora no Consumo de Combustível dos Embraer 190/195. https://www.rad.cvm.gov.br/ENET/frmDownloadDocumento.aspx?Tela=ext&numProtocolo=94642&descTipo=IPE&CodigoInstit>. Acesso em 19 de jan. de 2024.

FARIA, J.C. Regressão Linear Simples e Múltipla. Disponível em: <https://lec.pro.br//download/faria/seminarios/rl.pdf>. Acesso em: 25 de abr. de 2023.

FARIA, L.; OLIVEIRA., F.S; PINTO, P.E.D.; SZWARCFITER, J.L.; Ciência de Dados: algoritmos e aplicações, 1.ed. Rio de Janeiro: Impa, 2021. 165p.

FAUZI, A.; FERIYANTO, N.; IWAPUTRA, K.R ; DZAKIYULLAH, N.R; SALEH, C. The Route Analysis Based On Flight Plan. Disponível em: <https://iopscience.iop.org/article/10.1088/1757-899X/114/1/012147/pdf>. Acesso em: 05 de abr. de 2023.

FERREIRA, J. C.; PATINO, C.M. O que Realmente Significa o Valor-p? Disponível em: <https://www.scielo.br/j/jbpneu/a/SWk5XsCsXTW7GBZq8n7mVMJ/?format=pdf&lang=pt>. Acesso em: 28 de abr. de 2023.

GEORGIADIS, P.; HANCOCK, A; PFEIFER, S. Aviation sector pushes EU for green investment status. Financial Times, Bruxelas e Londres, 07 de fev. de 2023. Disponível em: <https://www.ft.com/content/96d6d00d-ab3f-45dc-a65d-845f74987561>. Acesso em 10 de abr. de 2023.

HILL, C.; Learning Scientific Programming with Python, 2.ed. New York: Cambridge University Press, 2020. 204p.

IBM. O que é Regressão Linear? Disponível em: < https://www.ibm.com/br-pt/analytics/learn/linear-regression>. Acesso em: 25 de abr. de 2023.

IBM. What is Data Science? Disponível em: <https://www.ibm.com/topics/data-science>. Acesso em: 10 de abr. de 2023.

FILHO, L.M.A.L. Correlação e Regressão. Disponível em: <http://www.de.ufpb.br/~luiz/AED/Aula9.pdf>. Acesso em: 26 de abr. de 2023.

KUMAR, S. Tutorial de Regressão Linear. Disponível em:<https://www.kaggle.com/code/sudhirnl7/linear-regression-tutorial>. Acesso em: 26 de abr. de 2023.

LIMA, M.; PERES; L. 10 Funções Mais Usadas para Manipular Dataframes no Pandas. Laboratório de pesquisa em Ciência de Dados e Inteligência Artificial da Universidade Federal do Ceará, Ceará, 30 de jun. de 2021. Disponível em: <https://www.insightlab.ufc.br/10-funcoes-mais-usadas-para-manipular-dataframes-no-pandas/>. Acesso em 17 de abr. de 2023.

McKinney, W. Python para Análise de Dados: tratamento de dados com pandas, numpy e ipython, 2.ed. United States: Novatec, 2018. 26p.

MENOTTI, D. Boas Maneiras Aprendizado Não Supervisionado - Regressão. Disponível em: <https://www.inf.ufpr.br/menotti/am-18a/slides/ML-1112cluster-regression.pdf>. Acesso em: 26 de abr. de 2023.

NETO, J.H.R.; Big Data para Executivos e Profissionais: tecnologias, aplicações e carreiras, 2.ed.[*S.I.*]: Kindle, 2019. 44p-45p.

OLIVEIRA, A. Transporte Aéreo: Economia e Políticas Públicas, 1.ed. São Paulo: Pezco, 2009. 26p.

2. ibid., p. 175.

PANDAS DOCUMENTATION: Pandas. DataFrame.astype. Disponível em: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.astype.html>. Acesso em 18 de abr. 2023.

PYTHON DOCUMENTATION: 3. Data Model. Disponível em: <https://docs.python.org/3/reference/datamodel.html>. Acesso em 18 de abr. 2023.

SHAH, R. Prevent Overfitting Using Regularization Techniques. Disponível em: <https://www.analyticsvidhya.com/blog/2021/07/prevent-overfitting-using-regularization-techniques/>. Acesso em: 28 de abr. de 2023.

TAMBOLI, N. Effective Strategies for Handling Missing Values in Data Analysis. Disponível em: <https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/>. Acesso em 28 de abr. 2023.

USP. Heterocedasticidade. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/176812/mod_resource/content/1/Slides%20-%20Heterocedasticidade.pdf>. Acesso em 28 de abr. 2023.

VAZ, R. G.; BECK, V. C. Introdução ao Método dos Mínimos Quadrados Ponderados. Disponível em:<http://www2.ufpel.edu.br/cic/2012/anais/pdf/CE/CE_00146.pdf>. Acesso em: 28 de abr. de 2023.

Submetido pelos autores em: 25/09/2023.
1a rodada de avaliação concluída em: 10/12/2023.
2ª rodada de avaliação - aprovação em: 02/02/2024.