Atividade Aula 04
Encontrando Pistas e Prevendo Resultados com Datasets Brasileiros

Objetivo:

- Trabalhar com datasets reais e multifacetados do Kaggle.
- Aprender a calcular e interpretar uma **matriz de correlação** para identificar as relações mais fortes em um conjunto de dados.
- Praticar a **seleção de características (features)** com base em evidências numéricas e visuais.
- Construir, treinar e avaliar um modelo de Regressão Linear Simples para resolver um problema específico.

Instruções Gerais:

Olá, detetives de dados! Cada dupla receberá dois casos para investigar. Para cada caso, vocês seguirão estas etapas:

1. **Carregar e Inspecionar o Caso:** Usem o link do Kaggle para encontrar e carregar o dataset no Google Colab. Façam a inspeção inicial com .head(), .info() e .describe() para entender as variáveis disponíveis.
2. **Encontrar a Pista Principal (Análise de Correlação):** O coração da atividade! Para descobrir qual característica (X) é a melhor para prever seu alvo (y), vocês usarão o método .corr().
    - Calculem a matriz de correlação do DataFrame: df.corr(numeric_only=True).
    - Identifiquem na matriz qual variável numérica tem a **maior correlação** (o valor mais próximo de 1.0 ou -1.0) com a sua variável alvo. Essa será a sua "pista principal".
3. **Confirmar a Pista (Visualização):** Criem um **gráfico de dispersão (scatterplot)** entre a "pista principal" (X que vocês escolheram) e a variável alvo (y) para confirmar visualmente se a relação linear realmente existe.
4. **Construir o Modelo:**
    - Preparem os dados X (sua pista principal) e y (o alvo).
    - Treinem um modelo de LinearRegression.
    - Usem o modelo treinado para responder à **pergunta de predição** específica do seu tema.
5. **Relatório Final:** Em uma célula de texto, escrevam uma breve conclusão: Qual foi a variável que vocês escolheram como melhor preditora e por quê? Qual foi o resultado da predição?

Temas Específicos para Duplas

Dupla 1: Mercado Imobiliário do Distrito Federal

- **Dataset:** Preço do Aluguel de Imóveis no Distrito Federal no Kaggle
- **Instrução para Carregar:** A melhor rota para este dataset é fazer o download do arquivo `aluguel_df.csv` diretamente da página do Kaggle e, em seguida, fazer o **upload manual** para o ambiente do Google Colab, como já praticamos.
- **Código para carregar (após o upload):** `df = pd.read_csv('aluguel_df.csv')`
- **Tema 1.1:**
    - **Alvo da Predição (y):** `preco` (Preço do Aluguel)

    - **Sua Missão:** O que mais influencia o preço de um aluguel no DF? Investiguem as correlações e descubram se a ***area***, o número de quartos ou de banheiros é a melhor "pista" para prever o ***preco***. Usem a característica com a correlação mais forte para construir o modelo.
    - **Pergunta de Predição:** Qual o aluguel previsto para um imóvel com uma área de **75 m²**?
- **Tema 1.2:**
    - **Alvo da Predição (y):** ***preco*** (Preço do Aluguel)
    - **Sua Missão:** O valor do ***condominio*** é um bom previsor para o ***preco*** do aluguel? Geralmente, imóveis mais caros têm condomínios mais altos. Investiguem a correlação entre essas duas variáveis e construam um modelo para testar essa hipótese.
    - **Pergunta de Predição:** Para um imóvel com taxa de condomínio de **R$ 500,00**, qual seria o valor do aluguel previsto pelo modelo?

Dupla 2: E-commerce Brasileiro (Olist)

- **Dataset:** Brazilian E-Commerce Public Dataset by Olist no Kaggle (Focar no arquivo ***olist_orders_dataset.csv*** e ***olist_order_payments_dataset.csv***).
- **Tema 2.1:**
    - **Alvo da Predição (y):** ***payment_value*** (Valor do Pagamento)
    - **Sua Missão:** Carreguem o arquivo ***olist_order_payments_dataset.csv***. A quantidade de parcelas (***payment_installments***) tem uma forte relação com o valor total da compra? Investiguem e criem um modelo.
    - **Pergunta de Predição:** Qual o valor de compra previsto para um pagamento feito em 10 parcelas?
- **Tema 2.2:**
    - **Alvo da Predição (y):** ***tempo_de_entrega*** (Tempo de Entrega)
    - **Sua Missão:** Carreguem o arquivo ***olist_orders_dataset.csv***. Vocês precisarão criar a coluna tempo_de_entrega calculando a diferença entre ***order_delivered_customer_date*** e ***order_purchase_timestamp***. (Dica: convertam as colunas para ***datetime*** primeiro). Depois, usem o dataset de geolocalização (***olist_geolocation_dataset.csv***) para tentar prever o tempo de entrega com base na latitude ou longitude do cliente. (Este é mais avançado!).

Dupla 3: Preços de Carros Usados no Brasil

- **Dataset:** Used Cars Prices in Brazil no Kaggle
- **Tema 3.1:**
    - **Alvo da Predição (y):** ***price*** (Preço)
    - **Sua Missão:** Qual a relação entre a quilometragem (***mileage***) de um carro e seu preço? Investiguem a correlação (espera-se que seja negativa!) e construam o modelo.
    - **Pergunta de Predição:** Qual o preço esperado de um carro com 80.000 km rodados?
- **Tema 3.2:**
    - **Alvo da Predição (y):** ***price*** (Preço)
    - **Sua Missão:** O ano de fabricação (***year***) é um bom previsor para o preço? Investiguem a correlação e construam o modelo.
    - **Pergunta de Predição:** Qual o preço esperado de um carro fabricado em 2021?

Dupla 4: Incêndios Florestais no Brasil

- **Dataset:** Amazon Forest Fires in Brazil no Kaggle
- **Tema 4.1:**
  - **Alvo da Predição (y):** *number* (Número de incêndios)
  - **Sua Missão:** Filtrem os dados para exibir apenas o estado de 'Amazonas'. O *year* (ano) tem alguma correlação com o número de incêndios? Modelem essa relação.
  - **Pergunta de Predição:** Com base na tendência histórica, qual seria o número de incêndios previsto para o Amazonas em um determinado mês de 2020?
- **Tema 4.2:**
  - **Alvo da Predição (y):** *number* (para o estado do 'Mato Grosso')
  - **Sua Missão:** Crie um novo DataFrame que tenha o número de incêndios do 'Mato Grosso' e de 'Rondonia' lado a lado para cada data. Existe uma correlação entre os incêndios nos dois estados? Modele essa relação.
  - **Pergunta de Predição:** Se em um mês Rondonia registrar 500 incêndios, quantos seriam esperados no Mato Grosso?

Dupla 5: Renda e Evasão Escolar no Brasil

- **Dataset:** Taxa de abandono escolar por Renda Média Brasil no Kaggle
- **Instrução para Carregar:** Para este dataset, a melhor rota é fazer o download do arquivo *evasao-renda.csv* do Kaggle e fazer o **upload manual** para o Google Colab, como aprendemos anteriormente.
- **Código para carregar (após o upload):** *df = pd.read_csv('evasao-renda.csv')*
- **Tema 5.1:**
  - **Alvo da Predição (y):** *Taxa_Abandono*
  - **Sua Missão:** Existe uma relação entre a renda média (*Renda_Media*) de uma localidade e a taxa de abandono escolar? Investiguem a correlação (espera-se que seja negativa, ou seja, quanto maior a renda, menor o abandono) e construam um modelo para quantificar essa relação.
  - **Pergunta de Predição:** Para uma localidade com renda média de **R$ 1.800,00**, qual seria a taxa de abandono escolar prevista pelo modelo?
- **Tema 5.2:**
  - **Alvo da Predição (y):** *Taxa_Abandono*
  - **Sua Missão:** A taxa de abandono escolar no Brasil mudou ao longo do tempo? Investiguem a correlação entre o Ano e a *Taxa_Abandono* para descobrir se há uma tendência de queda ou de aumento. Construam um modelo baseado nessa tendência.
  - **Pergunta de Predição:** Com base na tendência histórica, qual seria a taxa de abandono escolar prevista para o ano de **2023**?

Dupla 6: Vendas de Videogames no Brasil

- **Dataset:** Video Game Sales no Kaggle (Filtrar para vendas no Brasil, que estão na coluna *Other_Sales*, conforme descrição de alguns notebooks do Kaggle).
- **Tema 6.1:**
  - **Alvo da Predição (y):** *Global_Sales* (Vendas Globais)
  - **Sua Missão:** As vendas na América do Norte (*NA_Sales*) são um bom previsor para as vendas globais? Investiguem a forte correlação e construam o modelo.
  - **Pergunta de Predição:** Se um jogo vende 5 milhões de cópias na América do Norte,

qual seria sua venda global esperada?

- **Tema 6.2:**
  - **Alvo da Predição (y):** *EU_Sales* (Vendas na Europa)
  - **Sua Missão:** As vendas no Japão (*JP_Sales*) se correlacionam com as vendas na Europa? Crie um modelo para investigar essa relação de mercados.
  - **Pergunta de Predição:** Para um jogo que vende 2 milhões de cópias no Japão, quantas cópias seriam esperadas vender na Europa?