Avaliação de Estatística para Ciência de Dados

Base de Dados: Análise de Vendas de Videogames
Esta base de dados, contém informações sobre vendas de videogames.

Link para o dataset:
https://www.kaggle.com/datasets/gregorut/videogamesales

Dicionário de Dados:

- ***Rank:*** *Ranking de vendas gerais*
- ***Name:*** *Nome do jogo*
- ***Platform:*** *Plataforma do jogo (ex: PS2, X360, etc.)*
- ***Year:*** *Ano de lançamento do jogo*
- ***Genre:*** *Gênero do jogo*
- ***Publisher:*** *Empresa que publicou o jogo*
- ***NA_Sales:*** *Vendas na América do Norte (em milhões)*
- ***EU_Sales:*** *Vendas na Europa (em milhões)*
- ***JP_Sales:*** *Vendas no Japão (em milhões)*
- ***Other_Sales:*** *Vendas em outros países/regiões (em milhões)*
- ***Global_Sales:*** *Total de vendas mundial (em milhões)*

Objetivo: *Realizar uma análise completa do ciclo de vida de um projeto de ciência de dados, desde a obtenção e preparação dos dados, passando pela análise estatística até a apresentação dos resultados, utilizando as ferramentas Python, Power BI e KNIME.*

1: Extração, Transformação e Limpeza com Python e KNIME (3,0 pts)

Com Python:

1. **Carregamento dos Dados:** Carregar o arquivo vgsales.csv em um DataFrame do Pandas.
2. **Tratamento de Dados Ausentes:** Identificar e tratar valores ausentes nas colunas 'Year' e 'Publisher' com o método fillna() e/ou dropna(). Sugestão: remover as linhas com dados faltantes ou preenchê-las com um valor apropriado (ex: a moda para 'Publisher' e a mediana ou média para 'Year'), **justificando a escolha**.
3. **Limpeza de Dados:** Verificar e corrigir eventuais inconsistências nos nomes dos jogos ou empresa, se houver.
4. **Exportação:** Salvar o DataFrame limpo em um novo arquivo CSV chamado vgsales_limpo.csv.
5. **IMPORTANTE:** ***Em seu notebook, exiba sempre a contagem de linhas antes e depois das operações.***

Com KNIME:

1. **Leitura dos Dados:** Utilizar o nó CSV Reader para carregar o arquivo vgsales.csv.
2. **Manipulação de Dados:**
    - Utilizar o nó Missing Value para tratar os dados ausentes nas colunas 'Year' e 'Publisher', de forma similar ao que foi feito com Python.

   - Utilizar o nó Column Filter para remover colunas que não serão utilizadas em um primeiro momento (ex: 'Rank').
   - Empregar o nó Row Filter para selecionar um subconjunto dos dados, como por exemplo, jogos de uma plataforma específica ou de um determinado gênero.
3. **Documentação do Fluxo:** Organizar o workflow utilizando anotações e metanodes para explicar cada etapa do processo de limpeza e transformação.
4. **Exportação:** Utilizar o nó CSV Writer para salvar os dados transformados.

Parte 2: Análise e Modelagem com Estatística Descritiva (4,0 pontos)

Com Python (utilizando Pandas e Matplotlib/Seaborn):

1. **Análise Descritiva Geral:** Calcular as principais medidas de tendência central (média, mediana, moda) e de dispersão (desvio padrão, variância, amplitude) para as colunas de vendas (NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales).
2. **Correlação:** Calcular e visualizar a matriz de correlação entre as variáveis de vendas para entender a relação entre os mercados.
3. **IMPORTANTE: EXIBA OS VALORES ENCOTRANDOS NO NOTEBOOK.**

Com Power BI:

1. **Criação de Medidas DAX:**
   - Criar medidas para calcular a média, mediana e moda das vendas globais (Global_Sales).
   - Criar medidas para calcular a variância e o desvio padrão das vendas globais.
2. **Tabela de Frequência:** Criar uma tabela de frequência para a variável 'Platform', mostrando a contagem absoluta e relativa de jogos por plataforma.

Com KNIME:

1. **Agregação de Dados:**
   - Utilizar o nó GroupBy para agrupar os dados por 'Genre' e 'Platform', calculando a soma das Global_Sales.
   - Utilizar o nó Pivoting para criar uma tabela que mostre as vendas por gênero em diferentes regiões (colunas NA_Sales, EU_Sales, JP_Sales).
2. **Estatísticas:** Utilizar o nó Statistics para gerar um resumo estatístico das colunas de vendas.

Avaliação de Estatística para Ciência de Dados

Parte 3: Visualização de Dados com Power BI e KNIME (Valor: 3,0 pontos)

Com Power BI:

1. **Dashboard Interativo:** Criar um dashboard contendo:
    - Um gráfico de barras mostrando o total de Global_Sales por 'Genre'.
    - Um gráfico de pizza ou de rosca mostrando a distribuição de vendas por região (NA_Sales, EU_Sales, JP_Sales, Other_Sales) para um gênero selecionado.
    - Um gráfico de linhas mostrando a evolução do total de Global_Sales ao longo dos anos ('Year').
    - Cartões para exibir as principais métricas descritivas (média, mediana, desvio padrão das vendas globais).
    - Filtros (slicers) para 'Platform' e 'Year'.
    - Crie um visual que permita agrupar os dados por 'Genre' e calcular a soma das vendas globais para identificar os gêneros mais populares.
    - Crie um visual para agrupar os dados por 'Publisher' e calcular a média de vendas globais para avaliar o desempenho médio das publicadoras.

Com KNIME:

1. **Visualizações no Workflow:**
    - Utilizar o nó Bar Chart para visualizar o número de jogos por 'Platform'.
    - Utilizar o nó Scatter Plot para visualizar a relação entre NA_Sales e EU_Sales.
    - Criar uma "Composite View" combinando diferentes gráficos (ex: um Bar Chart e uma Table View) para apresentar uma análise consolidada.
2. **Exportação de Relatório:** Configurar o componente para gerar um relatório em PDF com as visualizações criadas.

Entregáveis

Ao final da avaliação, entregar:

- O script Python (.ipynb).
- O arquivo do Power BI (.pbix).
- O workflow do KNIME (print da tela no relatório abaixo).
- Um relatório em formato de texto (.docx ou .pdf) explicando as principais conclusões da análise, além do passo a passo para o desenvolvimento do projeto.