Bacharelado em Engenharia de Software

Disciplina: Ciência de Dados

Para esta atividade use a seguinte base de dados: Board Games Dataset Complete Features

A. Estatística Descritiva e Exploração Inicial

1. Carregue o arquivo `board_games.csv` em um DataFrame do Pandas. Fique atento ao separador de colunas, que pode não ser uma vírgula. Use o parâmetro `sep` do `pd.read_csv()`.
2. Exiba as 5 primeiras linhas do DataFrame para entender a estrutura dos dados.
3. Use o método `info()` para verificar os tipos de dados de cada coluna e a existência de valores ausentes.
4. Use o método `describe()` para gerar as estatísticas descritivas das colunas numéricas como `'average_rating'` e `'playing_time'`.
5. Responda, em forma de comentário no código:
    - Qual é a nota média de um jogo de tabuleiro neste dataset?
    - Qual é o tempo de jogo médio e o desvio padrão?
    - Existe alguma coluna com valores nulos que possa impactar as análises futuras?

B. Tratamento de Dados Ausentes e Outliers

1. Verifique a quantidade de valores nulos na coluna `'average_rating'`.
2. Se houver valores ausentes, preencha-os com a **mediana** da coluna. Justifique por que a mediana foi uma escolha melhor que a média nesse caso.
3. Utilize o **boxplot** para visualizar a distribuição da coluna `'playing_time'` e identificar visualmente os outliers.
4. Calcule o **IQR** (Intervalo Interquartil) para a coluna `'playing_time'` e identifique a quantidade de jogos que podem ser considerados outliers de tempo de jogo.

C. Visualização e Transformação de Dados

1. Crie um **histograma** da coluna `'average_rating'` para visualizar a sua distribuição.
2. Com base no histograma, comente se a distribuição é simétrica ou assimétrica (positiva ou negativamente).
3. Aplique a **transformação logarítmica** na coluna `'playing_time'`.
4. Crie um **novo histograma** da coluna `'playing_time'` após a transformação. Compare visualmente com a distribuição original e comente os resultados.

D. Análise da Relação entre Variáveis

1. Crie um **gráfico de dispersão (scatter plot)** com `'min_players'` no eixo x e `'max_players'` no eixo y.
2. A partir do gráfico, comente se você consegue identificar visualmente alguma tendência ou correlação entre o número mínimo e máximo de jogadores.
3. Calcule a **matriz de correlação** de todas as colunas numéricas do seu DataFrame.
4. Crie um **heatmap** (mapa de calor) para visualizar a matriz de correlação, tornando mais fácil a identificação das relações.
5. Em forma de comentário no código, responda:
    - Qual par de colunas numéricas possui a maior correlação positiva?
    - Qual par possui a correlação mais próxima de zero?

E. Análise Temporal

1. Identifique a coluna que representa a data de publicação do jogo (`'yearpublished'`).
2. Crie uma nova coluna chamada `'Decada'` que categorize os jogos por década de publicação (ex: 1990, 2000, 2010).
3. Faça a contagem de jogos lançados por década (agrupando por `'Decada'`) para entender a tendência de produção de jogos de tabuleiro ao longo do tempo.
4. Em forma de comentário no código, responda qual foi a década com o maior número de lançamentos e quantos jogos foram lançados nesse período.