

Atividade 1 – Probabilidade e Estatística – 2025.1

Atividade para ser feita e apresentada em dupla.

Utilizando uma ferramenta de IA Generativa escolher uma base de dados adequada para uma análise exploratória de dados, com no mínimo, 500 ocorrências (linhas) e 10 atributos (colunas), com variáveis numéricas e variáveis nominais, com um resumo dos dados e detalhes dos atributos de cada coluna.

Utilizando base de dados escolhida para a atividade:

1. Proceder uma Análise Exploratória de Dados, com o intuito entender a estrutura, qualidade e características básicas dos dados antes de aplicar técnicas estatísticas.
 - Importar os dados para o ambiente R e verificar: Número de linhas (observações) e colunas (variáveis). Tipos de variáveis (numéricas, categóricas, datas, etc.). Dicionário de dados para entender o significado de cada coluna ou variável. Modificar os nomes das colunas, caso ache necessário para facilitar o entendimento.
 - Identificar: Valores ausentes (NA/NaN): Decidir se remove ou ignora. Outliers: Verificar se são erros ou dados válidos. Dados inconsistentes: Ex.: Idade negativa, categorias mal codificadas (ex.: "Masc" e "Masculino"). Padronização: Unificar formatos (ex.: datas, categorias em maiúsculas/minúsculas).
 - Gravar a base de dados escolhida em formato Excel padrão .xlxs.
 - O que fizer com a base de dados registrar, identificar e explicar no trabalho.
2. Com uma coluna numérica contínua escolhida da base de dados, calcular: valor mínimo e o máximo valor, média, mediana, desvio padrão, os quartis, construir o histograma, boxplot e interpretar os resultados.
3. Fazer uma análise bivariada com duas colunas categóricas escolhidas da base de dados, fazer gráficos e interpretar os resultados.
4. Com uma coluna numérica contínua e uma coluna nominal escolhida da base de dados, gerar um gráfico boxplot considerando as duas colunas. Justifique as escolhas e interprete os resultados.
5. Gerar e mostrar a matriz de correlação com todas as variáveis numéricas da base de dados e explicar algumas correlações.
6. Fazer uma análise bivariada com duas colunas numéricas escolhidas da base de dados, fazer gráfico de dispersão, a linha de tendência dos dados, a correlação entre as duas variáveis e gerar gráfico de dispersão, mostrar a equação da reta ajustada e o coeficiente de determinação.
7. Considerando base de dados escolhida, formule três perguntas associadas a estatística bivariada e verificar se dados confirmam ou não as suas suspeitas ou dúvidas e mostrar através de gráficos ou análise estatística e interpretar os resultados.
 - Enviar para o classroom, até o horário da aula do dia 11/08/25, o código R, o arquivo .Rmd, o arquivo Excel gravado no item 1. acima, o arquivo html ou pdf gerado e se preparar para apresentação da atividade. Se organizar, para uma apresentação com tempo de duração em torno de 7 minutos, nos dias 11 e 13/08. As apresentações serão escolhidas através de sorteio.

- O trabalho vale 50 pontos para a nota AB1.
- Critérios de pontuação:
 - Escolha da base de dados → 0,5 ponto
 - Preparação e organização dos dados (item 1) → 0,5 ponto
 - Questões 2, 3, 4, 5 e 6 → 0,4 ponto, totalizando 2,0 pontos
 - Questão 7 → 1,0 ponto
 - Arquivos solicitados e a apresentação → 1,0 ponto.
- Quando a dupla escolher a base de dados, registrar uma descrição mínima da base, a quantidade de ocorrências e de atributos para que não seja mais escolhida por outra dupla.
 - Link para registrar a escolha do tema:
<https://docs.google.com/document/d/1k41iRTOeCqUINwSOPxJrYoeepYaGv9W7nukjktlAujg/edit?usp=sharing>