

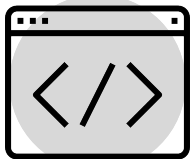
Profissão: Cientista de Dados



BOAS PRÁTICAS



Descritiva II – Análise descritiva para resposta contínua



- Explore covariância e correlação
- Compreenda matriz de correlação
- Identifique outliers
- Analise Variáveis Bivariadas
- Compreenda a estabilidade



Explore covariância e correlação

- Lembre-se de que a correlação e a covariância são medidas de associação linear. Se a relação entre as variáveis não for linear, essas medidas podem não ser adequadas.
- Não assuma que uma correlação de zero significa que as variáveis são independentes. Uma correlação de zero apenas indica uma associação linear fraca.
- Ao analisar a correlação, observe a intensidade da correlação. Uma correlação mais forte resultará em pontos mais próximos à linha de tendência.
- Lembre-se de que uma correlação perfeita ocorre quando os pontos estão exatamente na linha de tendência, independentemente da inclinação da linha (desde que não seja horizontal).



Compreenda matriz de correlação

- Utilize gráficos de dispersão para explorar a estrutura de associação entre as variáveis. No entanto, esteja ciente de que esses gráficos podem se tornar poluídos quando há muitas variáveis.
- Considere o uso de um cluster map para visualizar a matriz de correlação quando há muitas variáveis. Este gráfico agrupa variáveis que têm alta correlação entre si, facilitando a interpretação.
- Ao criar um cluster map, centralize a cor no zero. Isso pode ser feito usando o parâmetro 'center'.
- Ao realizar análises de correlação, utilize ferramentas como o Jupyter Notebook, pandas e seaborn. O método 'corr' do pandas pode ser usado para calcular a matriz de correlação, enquanto a biblioteca seaborn pode ser usada para criar um cluster map.



Identifique outliers

- Outliers podem distorcer significativamente os resultados de suas análises, deslocando a média e inflacionando a variância. Portanto, é crucial identificar e lidar com eles adequadamente.
- Em alguns casos, uma transformação de dados, como a aplicação de um logaritmo, pode ser útil para transformar uma escala de incrementos multiplicativos em uma escala de incrementos aditivos, tornando os dados mais gerenciáveis.
- Existem várias maneiras de lidar com outliers, incluindo removê-los, usar técnicas robustas ou buscar um modelo que se ajuste melhor à natureza dos outliers. A escolha da solução depende do objetivo da análise e do contexto dos dados.
- O ranking pode ser uma boa alternativa para identificar associações entre variáveis sem ter que ajustar a escala das variáveis.



Analise Variáveis Bivariadas

- Conceitos como erro padrão e intervalo de confiança são fundamentais para a análise de dados. O erro padrão ajuda a determinar se uma diferença observada é fortuita ou não, enquanto o intervalo de confiança ajuda a entender a variabilidade em torno da média da amostra.
- Não tire conclusões precipitadas: Mesmo que a média de um grupo caia dentro do intervalo de confiança de 95%, isso não significa necessariamente que a diferença observada na amostra também se verificaria na população total.
- Explore diferentes perspectivas: Ao analisar os dados, é importante explorar diferentes perspectivas. Isso pode revelar insights que não seriam aparentes a partir de uma única análise.



Compreenda a estabilidade

- A distribuição de uma variável pode mudar ao longo do tempo devido a vários fatores. Portanto, é importante monitorar a distribuição de uma variável ao longo do tempo para identificar quaisquer alterações significativas.
- Assim como a distribuição, a média de uma variável também pode mudar ao longo do tempo. Portanto, é importante avaliar a média de uma variável ao longo do tempo para identificar quaisquer alterações significativas.
- Alterações na distribuição ou na média de uma variável ao longo do tempo podem indicar que a população está mudando. Isso pode exigir uma revisão do modelo.



Bons estudos!

