Profa. Solange Kanso

Disciplina: Ciência de Dados

Lista 2

Com base no que trabalhamos em sala de aula, escreva um código em Python seguindo as solicitações.

1. Utilize o arquivo “igm_modificado.csv” (disponível no Classroom dentro de tabela de dados) e o código “Ciência de Dados - Aula 03 04 - Churn - Análise exploratória.ipynb” como norte para desenvolver seu próprio código.
2. Essa lista pode ser resolvida em grupo de no mínimo 5 alunos e no máximo 10 alunos (**sem exceção**)
3. Todas as análises e resultados devem estar no código em formato de comentário ou texto
4. **Não esqueça: no início do código colocar o RA e o NOME COMPLETO de cada integrante do grupo**
5. Para iniciar, faça o upload da tabela “csv” no seu drive para fazer a leitura da tabela em Python

Exercícios

1ª parte – utilize todos os municípios para as análises

1. Faça uma checagem do arquivo (vimos diversos comandos, apresente pelo menos 3 deles)
2. Nesse arquivo, quais são os tipos de dados que encontramos?
3. Escolha 2 variáveis e analise sua frequência absoluta e percentual
4. Calcule o 1º quartil, o 3º quartil de alguma variável numérica e analise seus resultados
5. Escolha uma variável numérica e por meio das estatísticas de tendência central (média, mediana e moda), identifique o tipo de assimetria da variável. Analise o seu resultado e justifique.
   Sabendo que:
   - Distribuição assimétrica positiva (ou assimétrica à direita): moda < mediana < média
   - Distribuição simétrica ou ausente de assimetria: moda = mediana = média
   - Distribuição assimétrica negativa (ou assimétrica à esquerda): média < mediana < moda
6. Há dados faltantes em algumas variáveis? Quais? Sugira formas de tratá-las. E em uma delas substitua o NaN pela mediana
7. Faça um gráfico de setores de alguma variável categórica e analise os resultados
8. Faça o boxplot de alguma variável numérica contínua por outra variável categórica e analise os resultados comparativos

2ª parte – crie outro dataframe com apenas duas regiões

9. Faça uma análise comparativa das estatísticas de 4 variáveis e analise as diferenças entre as regiões. Seria interessante levantar uma hipótese para essa comparação