UNIVERSIDADE FEDERAL DE MATO GROSSO
CAMPUS UNIVERSITÁRIO DO ARAGUAIA
Instituto de Ciências Exatas e da Terra
Curso de Bacharelado em Ciência da Computação

Disciplina: Programação I
Professor: Ivairton M. Santos

Trabalho de análise e exploração de dados

Trabalho baseado no tutorial:
https://dadosaocubo.com/analise-exploratoria-de-dados-com-python-parte-i/

Neste trabalho vamos abordar uma área em tecnologia que está em expansão, a Ciência dos Dados (Data Science). São vários conceitos, tecnologias e métodos empregados nesta área, tenha consciência que este trabalho é uma introdução, da introdução!

Para qualquer trabalho nesta área o primeiro passo consiste de conhecer os dados, levantar informações a respeito de todos os detalhes do conjunto de dados, métricas, valores de referência, entre outros. Este trabalho trata justamente desta primeira etapa, geralmente denominada de "análise exploratória dos dados". Para isso vamos entender alguns conceitos iniciais e trabalhar em um exemplo para então manipular um conjunto de dados que será fornecido para realização do trabalho.

Configuração inicial:

Para a realização deste trabalho serão utilizadas as bibliotecas:

- Numpy (cálculo matemático) - https://numpy.org/
- Pandas (análise e manipulação de grandes volumes de dados) - https://pandas.pydata.org/
- Seaborn (geração de gráficos) - https://seaborn.pydata.org/

Conceitos fundamentais:

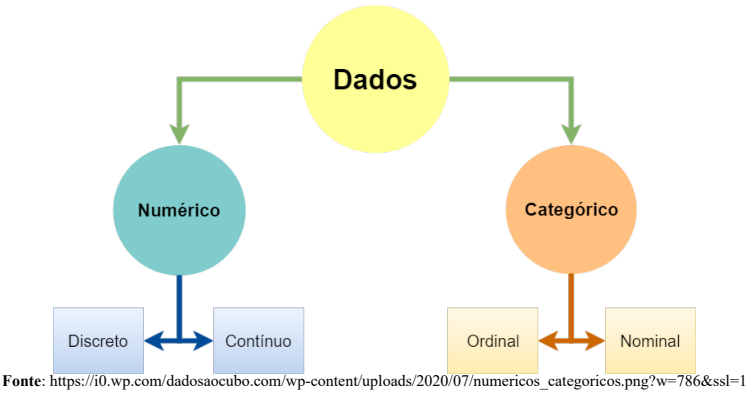
A análise exploratória de dados precede qualquer trabalho com dados, é fundamental conhecer as características e propriedades de uma base de dados. Neste contexto, um conhecimento fundamental é a estatística. A estatística é a base teórica da ciência de dados, sem ela não seria possível desenvolver com segurança e consistência nenhum dos modelos ou algoritmos que utilizamos.

É comum ouvir a expressão que cientistas de dados têm como missão contar histórias utilizando dados e é muito difícil contar uma boa história a partir dos dados sem usar estatística. Para ajudar com o entendimento elementar de estatística, vamos entender conceitos básicos como *variável* e *tipos de variáveis*.

Variáveis:

No processo de amostragem, ou seja, quando avaliamos parte dos dados, não devemos avaliar apenas a informação do que realmente temos interesse, mas todas as outras disponíveis na base, pois elas podem nos ajudar no entendimento dos dados, indicar correlações e até complementar o entendimento da informação que é objeto de estudo.

Cada uma das características da base de dados, também chamada de "população amostrada", como por exemplo peso, altura, sexo, ou idade, é denominada de variável. As variáveis podem assumir valores distintos, que basicamente podemos separá-los em Numéricos (variáveis quantitativa) ou Categóricos (variáveis qualitativas). A figura a baixo esquematiza essa classificação das variáveis:



Fonte: https://i0.wp.com/dadosaocubo.com/wp-content/uploads/2020/07/numericos_categoricos.png?w=786&ssl=1

Vamos entender cada uma dessas duas categorias:
Variáveis numéricas, podem ser de 2 tipos*:*

- **Discretas**: possuem apenas valores inteiros. Ex.: número de irmãos, número de passageiros.
- **Contínuas**: possuem qualquer valor, incluindo números reais (float). Ex.: peso, altura.

Variáveis *categóricas*, podem ser de 2 tipos:

- **Nominais**: quando as categorias não podem ser ordenadas de alguma maneira. Ex.: nomes, cores, sexo.
- **Ordinais**: nesse caso as categorias podem ser ordenadas. Ex.: tamanho (pequeno, médio, grande), classe social (baixa, média, alta), grau de instrução (básico, médio, graduação, pós-graduação).

Conhecendo melhor os conceitos inerentes às variáveis, vamos comentar algumas das principais técnicas estatísticas:

Média:

A média ou média aritmética, nada mais é do que a soma de todos os dados da amostra dividido pela quantidade de amostras.

Mediana:

Mediana é o valor que representa o valor central da amostra, por exemplo, num conjunto com os valores { 2, 2, 3, 8, 9 } a mediana é 3. Estando as amostras ordenadas em valores crescentes, caso o total de elementos for par, será necessário calcular a média dos dois valores centrais, por exemplo, no conjunto {1, 1, 4, 5} a mediana será (1+4)/2= 2,5.

Moda:

A Moda é o valor que aparece com mais frequência em um conjunto de dados, ou seja, o valor que se repete mais vezes. Para fazermos o cálculo da moda de um conjunto de dados, basta

encontrar os dados que mais aparecem no conjunto.

Variância:

A variância é uma medida de dispersão dos dados, mede o quão afastados os dados estão da média. Quanto maior a variância, mais afastados (dispersos) os dados encontram-se da média.

Desvio padrão:

O desvio padrão é uma medida que expressa o grau de dispersão de um conjunto de dados. Ou seja, o desvio padrão indica o quanto um conjunto de dados é uniforme. Quanto mais próximo de zero for o desvio padrão, mais homogêneo são os dados. O desvio padrão é calculado a partir da raiz quadrada da variância.

Quantis:

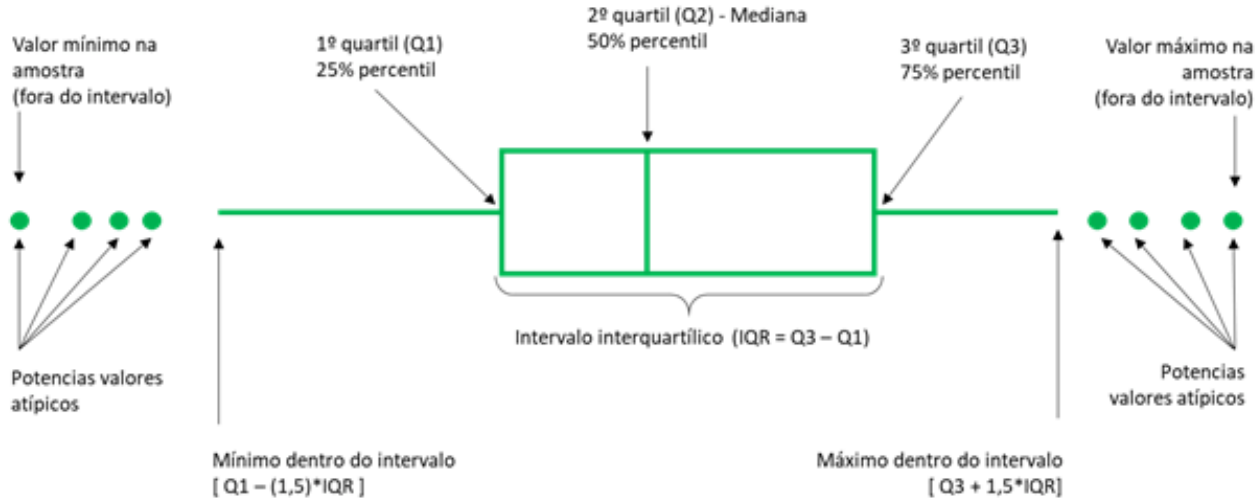
Quantis são pontos que dividem uma distribuição de probabilidade em partições de tamanhos iguais. Eles podem ser *quartis:* sendo o 1º quartil correspondente a 25% dos dados; o 2º quartil correspondente a 50% dos dados (ou seja, a mediana); e o 3º quartil correspondente a 75% dos dados). Ou ainda, o Quantis pode ser do tipo *percentis,* que divide a amostra em 100 partes.

Intervalo Interquartil (IQR):

O IQR (*Interquartile Range*), é a diferença entre o 3º e 1º *quartis* (IQR = Q3−Q1). É uma medida de dispersão robusta muito utilizada, por exemplo, quando os dados contêm muitos *outliers* (valores que estão muito fora do padrão observado na amostra) por ser menos sensível às variações nos extremos do conjunto.

Boxplots:

Boxplot é utilizado para avaliar e comparar o formato, tendência central e variabilidade de distribuições de uma amostra, especialmente é utilizado para identificar *outliers*. Por padrão, um boxplot demonstra a mediana, os quartis, o intervalo interquartil (IQR) e *outliers* para cada variável. A imagem a seguir descreve a estrutura e os componentes que o boxplot apresenta:



Histogramas:

Um histograma é uma visualização gráfica de dados usando barras de diferentes alturas. Em um histograma, cada barra agrupa números em intervalos. As barras mais altas mostram que mais dados estão nesse intervalo. Um histograma exibe a forma e distribuição de dados amostrais,

tanto para variáveis discretas ou contínuas.

Assimetria:

Apesar do termo, assimetria é na verdade uma medida de simetria (ou podemos dizer de não simetria). Ela nos diz o quão simétrica é a distribuição dos dados em torno da média. E junto com a *Curtose* (*Kurtosis*), veremos em seguida, é uma medida importante para informar a aparência ou forma da distribuição dos dados. Quando o valor for zero, indica que os dados são simétricos, ou seja, média, moda e mediana são iguais, enquanto que valores negativos, ou positivos, indicam uma assimetria e quanto mais longe de zero, mais "deformada" é a distribuição dos dados.

Curtose (Kurtosis):

A *Curtose*, ou achatamento, também é uma medida que nos ajuda a dar forma à distribuição dos dados. A *Curtose*, diferente da assimetria, tenta capturar em uma medida a forma das caudas da distribuição. Em outras palavras, ela indica o grau de achatamento da distribuição (gráfico mais "pontudo", ou "achatado"), isto é, quão espalhados os dados estão em torno da média.

Tutorial:

Acesse o link abaixo e estude o tutorial que demonstra como calcular cada uma dessas métricas

https://colab.research.google.com/drive/1dfyK02m879I9WRr_sTFS-bTBomD4Xu8A?usp=sharing

Desafios:

Agora é sua vez de realizar uma análise exploratória de dados:

Requisito 1:

Utilize a base de dados "California Housing Dataset", que consiste de informações de residências e seus respectivos valores em diferentes regiões da Califórnia. (https://scikit-learn.org/stable/datasets/real_world.html - Seção 7.2.7)

Esta base de dados contém mais de 20 mil registros, com 8 atributos:

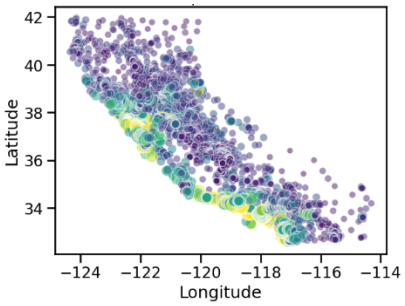
- **MedInc**: renda média (familiar) na região;
- **HouseAge:** idade média do imóvel na região;
- **AveRooms**: Número médio de cômodos no imóvel;
- **AveBedrms:** Número médio de quartos no imóvel;
- **Population:** População na região;
- **AveOccup**: Número médio de membros na família
- **Latitude:** latitude da região
- **Longitude**: longitude da região

A variável alvo consiste do valor do imóvel expresso em múltiplos de US$ 100.000 (campo com nome 'MedHouseVal').

Carregue a base de dados, crie um Pandas.DataFrame com os dados e execute a função que

descreve a base (`describe()`)

Requisito 2:
Utilizando a biblioteca SeaBorn e dos campos 'Latitude' (eixo y) e 'Longitude' (eixo x) da base de dados, gere um gráfico que apresente pontos representando os imóveis em sua localização espacial, o que corresponde ao mapa. O gráfico deverá ficar parecido com o gráfico abaixo:



Requisito 3:
Calcule as métricas:

- Média
- Mediana
- Moda
- Variância
- Desvio padrão
- Quantis
- IQR

Para as 6 primeiras variáveis da base de dados (medinc, houseage, averooms, avebedrms, population, aveoccup) e apresente os resultados de maneira adequada.

Requisito 4:
Gere o BoxPlot e o Histograma para as mesmas variáveis descritas no Requisito 3.

Requisito 5:
Identifique ao menos 2 pares de variáveis (dois conjuntos com 2 variáveis cada) que apresenta correlação entre si. Ou seja, identifique variáveis que seus valores estão correlacionados de uma tal maneira que juntos determinam uma variação no valor do imóvel. Por exemplo, será que quanto menor a idade do imóvel e menor o número de membros da (correlação positiva), maior o valor do imóvel? Ou, por exemplo, quanto maior a renda familiar e menor a idade do imóvel (correlação negativa), maior o valor do imóvel?

Requisito 6:
Ao final, apresente seu estudo a respeito da base de dados escrevendo um relatório, apresentando todos os itens solicitados nos requisitos anteriores. Entregue como resposta a este trabalho o relatório (em PDF) e o código fonte produzido.