Análise Descritiva de Dados com Python
CAIC 2023

Profº Gustavo Miranda
gustavo.oliveira@penedo.ufal.br

UFAL
CAMPUS ARAPIRACA
U. E. PENEDO

Quem sou eu?

Professor Gustavo Miranda:

- **Doutor** em Inteligência Computacional - **UFPE**
- **Mestre** em Inteligência Computacional - **UFPE**
- **Graduado** em Licenciatura em Computação - **UPE**
- **Experiência** nos seguintes campos:
  - Inteligência Artificial
  - Ciência dos Dados
  - Desenvolvimento de Software

Informações Adicionais

- https://github.com/**GustavoHFMO**

Tópicos da Aula

1º Importação e Entendimento dos Dados

2º Adição de Novas Variáveis

3º Utilização de Estatísticas Univariadas

4º Apresentação de Gráficos Univariados

5º Estatísticas Bivariadas

Motivação

Análise Estatística de Dados

- A estatística tem como objetivo **fornecer informação** utilizando quantidades numéricas. A estatística divide o estudo da análise dos dados em três fases:
  - Obtenção dos dados.
  - Descrição, classificação e apresentação dos dados.
  - Conclusão a tirar dos dados.

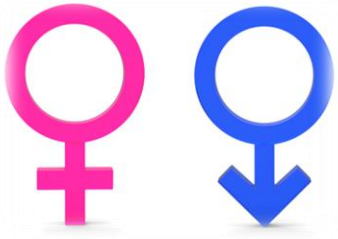
- É fundamental para que as **conclusões** decorrentes da análise de dados **utilizem métodos estatísticos para ter uma base sólida**, reduzindo os riscos de erro de avaliação.

- A conclusão sobre os dados podem partir de **variáveis qualitativas** e também **quantitativas**.

Variáveis Qualitativas

- Quando as variáveis representam **atributos**, códigos, categorias e características expressas preferencialmente por meio **textual**, os dados resultantes da observação dessas **variáveis são ditos qualitativos.**
  - Nominais, ordinais ou intervalares.

- **Qualitativas nominais:**
  - Sexo, religião, Estado civil, Cidade.



- **Qualitativas ordinais:**
  - Nível médio, Graduação, Mestrado.

- **Qualitativas Intervalares:**
  - Ótimo, bom, regular, ruim e péssimo.

Variáveis Quantitativas

- Variáveis que representam **contagem, medição** e outros dados cujos os valores devem ser expressos preferencialmente **por meios numéricos** tais variáveis são ditas **quantitativas.**
  - Discretas ou contínuas.

- **Quantitativas discretas:**
  - Assumem valores dentro dos números naturais.
  - Ex: quantidade de filhos de um casal.

- **Quantitativas contínuas:**
  - Assumem qualquer valor dentro dos números reais.
  - Ex: Altura de uma pessoa.

Estatística Descritiva

- **A estatística descritiva** tem como objetivo **compreender** em maior nível de detalhes **o conjunto de dados coletados, através da exploração** de suas características, tais como:
  - Média, moda, mediana, variância, desvio padrão e outros.

- **Além disso, ela também pode incluir:**
  - Verificação da representatividade ou da falta de dados;
  - Ordenação dos dados;
  - Compilação dos dados em tabelas;
  - Criação de gráficos com os dados;

Estatística Inferencial

- A partir da análise exploratória dos dados, obtém-se maior grau de conhecimento sobre os dados. **Após esse entendimento dos dados, podemos tirar conclusões**. Essas conclusões são chamadas de **estatísticas inferenciais**.

- A estatística inferencial pode **nortear a tomada de decisões** com segurança e assertividade, através da estimação de informações sobre uma população a partir de uma amostra.

- **População:** conjunto total de informações.
- **Amostra:** conjunto reduzido da entidade.

Importação e Entendimento dos Dados

Código!

Adição de Novas Variáveis

Por que criar novas variáveis?

Síntese de Informações:

- Você pode criar novas variáveis para resumir ou sintetizar informações presentes nas variáveis originais. Por exemplo, você pode calcular a média, mediana, moda, desvio padrão ou outras medidas resumo para criar uma variável que represente melhor o comportamento geral dos dados.

Transformação de Dados:

- Às vezes, é benéfico aplicar transformações aos dados para torná-los mais adequados para análise estatística. Isso pode incluir logaritmos, raízes quadradas, ou outras transformações que ajudem a estabilizar a variabilidade ou a normalizar a distribuição dos dados.

Criação de Categorias ou Grupos:

- Você pode criar novas variáveis categóricas para agrupar dados de acordo com critérios específicos. Isso pode facilitar a análise de subgrupos ou a comparação de diferentes categorias.

Análise Temporal:

- Se os dados incluírem informações temporais, você pode criar variáveis adicionais para analisar tendências ao longo do tempo, como calcular a variação percentual, médias móveis ou outras métricas temporais relevantes.

Normalização ou Padronização:

- Em algumas situações, normalizar ou padronizar variáveis pode ser útil para comparar diferentes conjuntos de dados. Isso envolve criar novas variáveis que representam os dados em uma escala comum, facilitando a comparação entre eles.

Facilitar a Visualização:

- Criar novas variáveis pode facilitar a visualização dos dados. Por exemplo, você pode criar variáveis que representam proporções, percentagens ou outras métricas que são mais interpretáveis e informativas em gráficos.

Utilização de Estatísticas Univariadas

Média

- A **média** é uma medida que tem como objetivo **determinar o valor central de um conjunto** de dados.
- A média possibilita encontrar o **comportamento padrão ou mais comum** de um grupo de dados.
- **Problemas:** sensível a pontos aberrantes, ou seja, valores que destoam muito do padrão.

$$\overline{X} = \frac{(X_1 + X_2 + X_3 + \ldots + X_n)}{n}$$

- n: representa a quantidade de elementos calculados.
- $x_1, \ldots, x_n$: representam as observações do conjunto.

Mediana

- A mediana representa o “**valor do meio**” dos dados analisados, que metade dos valores estão acima e a outra metade abaixo.

- A mediana é **eficiente contra pontos aberrantes**, pois considera o centro dos dados.

- **Para calcular a mediana:**

1. Ordenar os valores de forma crescente.
2. Escolher o número do meio.
3. **Obs:** quando a quantidade de elementos é par, soma-se os valores centrais e divide-se por dois.

Moda

- A moda é uma medida que **representa o valor mais frequente** no conjunto de dados.
- Possibilita entender **qual elemento** ou atividade que mais ocorre dentro do conjunto.
- Também é **eficiente contra pontos aberrantes**.

- **Para calcular a moda:**
  1. Conta-se a quantidade de vezes que cada elemento apareceu.
  2. A moda é dada pelo elemento mais frequente.
  3. **Obs:** um conjunto é **bimodal** quando apresenta duas modas.

Máximo e Mínimo

- **Medidas de amplitude** como máximo e mínimo ajudam a **entender a variabilidade dos dados.**

- **Ajudam a entender a faixa de valores** que inclui todos os valores do conjunto de dados.

Variância

- A variância é uma **medida de dispersão** que mostra o **quão distante cada valor** do conjunto de dados **está** do valor central (**média**).
- Quanto **menor é a variância, mais próximo** os valores estão da média.
- Quanto **maior a variância, mais distantes** estão os valores da média.

- $\sigma^2 = \frac{\sum(x_i - \mu)^2}{n}$
- n: é a quantidade de elementos.
- $x_i, \ldots x_n$: elementos do conjunto.
- μ: média do conjunto.

Desvio Padrão

- O desvio padrão indica o “**erro**” de cada elemento **em relação a média.**

- Em outras palavras, podemos dizer que o desvio padrão **minimiza a variabilidade** dos dados e apresenta uma medida mais direta em relação a média.

- $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$

- n: é a quantidade de elementos.
- $x_i, \ldots x_n$: elementos do conjunto.
- μ: média do conjunto.

Quartis

- Quartil é uma medida que divide os dados em **quatro partes**, correspondendo aos percentuais de **25%, 50%, 75% e 100%.**

- Os dados são **ordenados do menor para o maior** e são divididos 4 partes exatas.

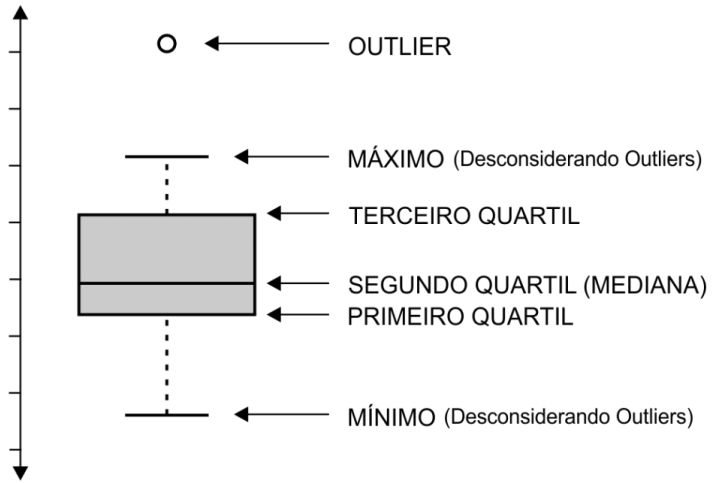
- Essas medidas ajudam a **entender** como estão **os grupos do conjunto de dados.**

Apresentação de Gráficos Univariados

Boxplot

- O boxplot fornece uma **análise visual** da posição, dispersão, simetria, caudas e valores discrepantes do conjunto de dados.

- O boxplot é interessante para **analisar o comportamento de vários grupos lado a lado.**



Histograma

- O histograma é conhecido como o gráfico de **distribuição de frequências.**

- **Cada retângulo representa** a quantidade ou **a frequência** com que cada valor ocorre no conjunto de dados.

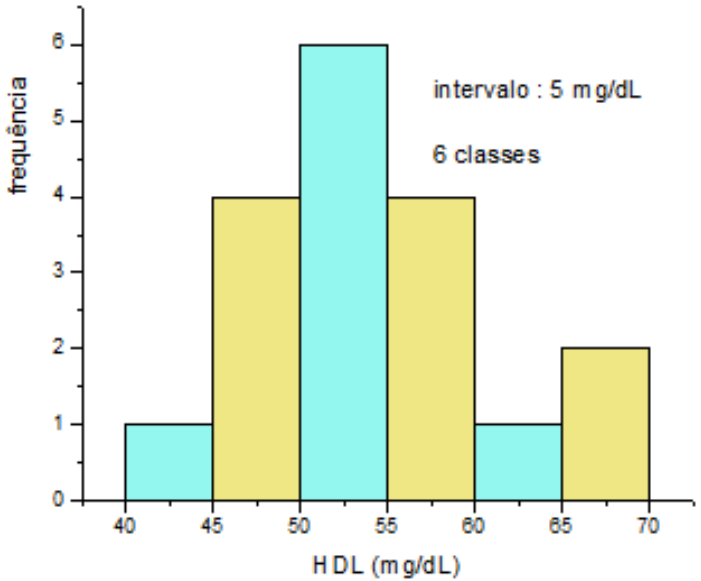


Gráfico de Dispersão (Scatterplot)

- O scatterplot ou **gráfico de dispersão** possibilita plotar duas variáveis distintas e ver a relação delas no espaço.

- O gráfico de regressão possibilita ver a **relação linear** entre as duas variáveis.

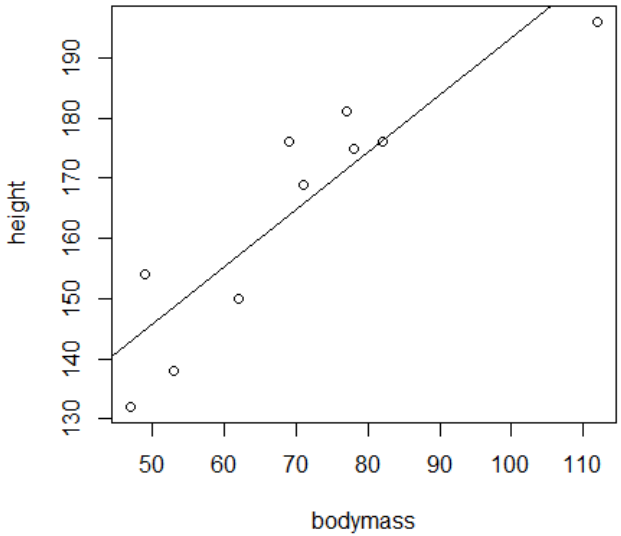
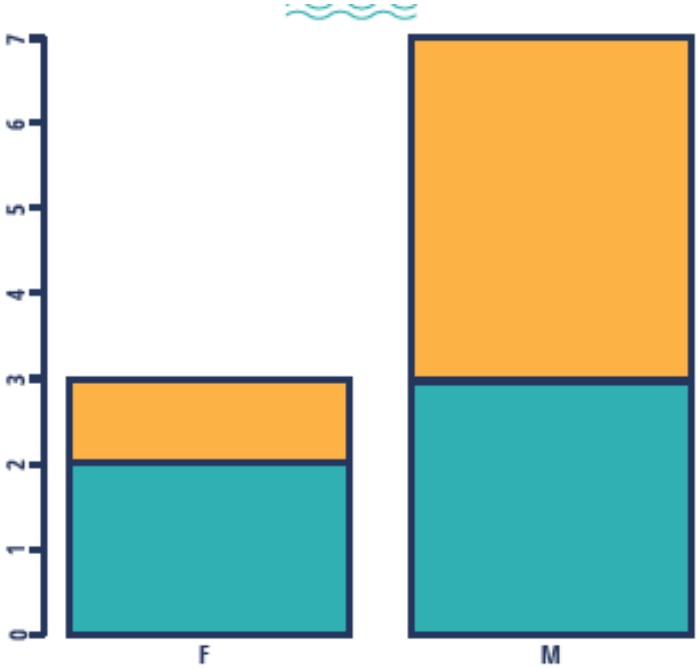


Gráfico de Barras da Tabela de Contingência

- O gráfico de barras apresenta a **quantidade de ocorrências** para as combinações de variáveis qualitativas analisadas.

- Dentro do mesmo gráfico é possível **verificar a frequência de diferentes grupos.**

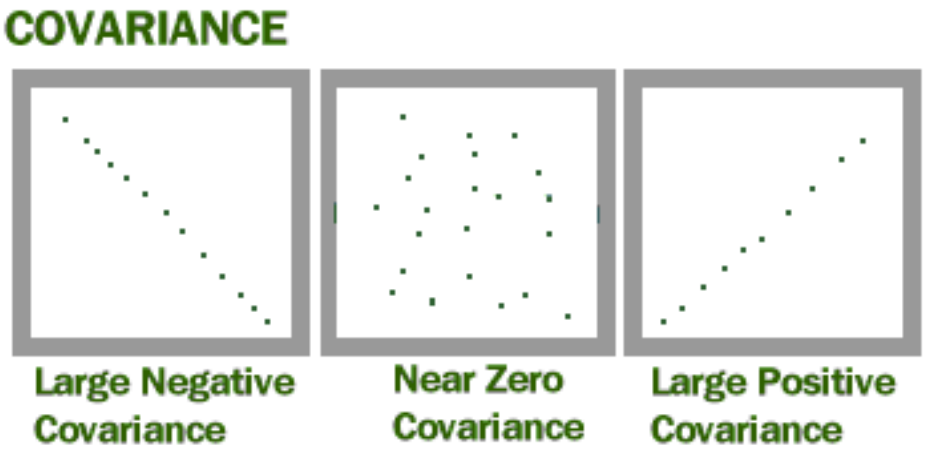


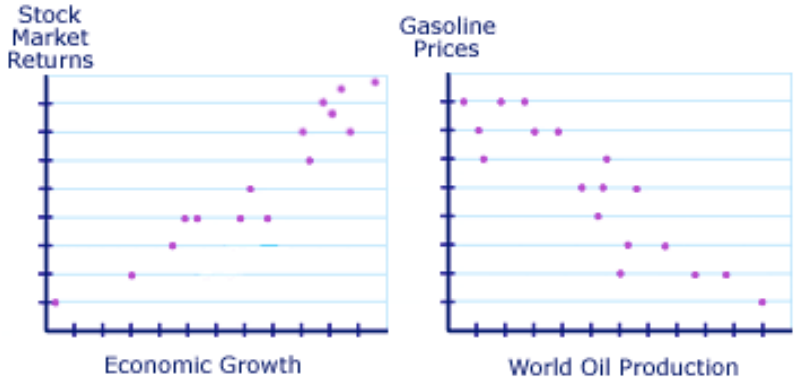
Estatísticas Bivariadas

Medidas de Variabilidade: Covariância

- Avalia a **variância conjunta de dois atributos** (bivariada).
- Se a variação dos valores de um atributo **acompanha a do outro.**
- **Covariância positiva:** valores altos para um atributo X estão associados a valores altos para outro atributo Y
- **Covariância negativa:** X aumenta Y diminui
- **Zero:** ausência de relação

Covariância





Covariância

- A **covariância (amostral)** entre dois atributos X eY:

$$cov(X,Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$$

Economic Growth % (x_i)	S & P 500 Returns % (y_i)
2.1	8
2.5	12
4.0	14
3.6	10

Exemplo

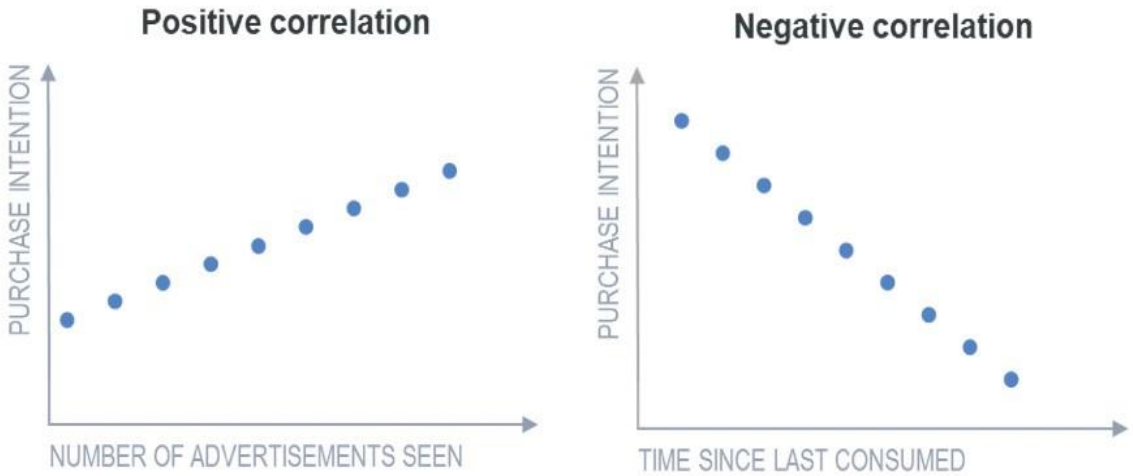
Economic Growth % (x_i)	S & P 500 Returns % (y_i)
2.1	8
2.5	12
4.0	14
3.6	10

$$\bar{x} = 3.1$$
$$\bar{y} = 11$$

$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$
$$= \frac{(2.1 - 3.1)(8 - 11) + \ldots}{4\text{-}1}$$
$$= \frac{(-1)(-3) + (-0.6)(1) + (0.9)(3) + \ldots}{3}$$
$$= \frac{3 + (-0.6) + 2.7 + (-0.5)}{3}$$
$$= \frac{4.6}{3}$$
$$= 1.53$$

Correlação

- Correlação **padroniza a medida de relação entre os atributos**:
  - **Valores entre 1 e -1**
  - 0: sem correlação



Correlação de Pearson

- **Normaliza a covariância pelo desvio padrão dos atributos**
- Suposições:
  - Variáveis seguem uma gaussiana
  - Variáveis contínuas
  - Linearidade
- Quantifica a existência de uma **relação linear entre as variáveis.**

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Correlação de Spearman

- **Não paramétrico:** atributos são relacionados por qualquer função monotônica
- **Variáveis podem ser ordinais**

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

- d2: quadrado da diferença entre os ranqueamentos dos atributos
- n: número de instâncias

Exemplo: Correlação de Spearman

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

Correlação de Spearman

- Ordena-se cada atributo e gero um ranking

	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2-1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

Exercícios

1. Calcule a média (mean), mediana (median), moda (mode), variância (var), desvio padrão (std) para preço e área de imóveis à venda.
2. Calcule os quartis da área (dica: describe)
3. Calcule o preço médio do metro quadrado de venda para Recife e adicione ao dataframe
4. Calcule o preço médio do metro quadrado de venda para apartamentos
5. Encontre os bairros com maior e menor valor de venda por metro quadrado
6. Encontre a variável que tem maior correlação de spearman com o preço de imóveis à venda

Análise Descritiva de Dados com Python
CAIC 2023

Profº Gustavo Miranda
gustavo.oliveira@penedo.ufal.br