WBA0753_v1.0

Linguagens de programação para ciência de dados (*Python* com *Spark*)

Análise estatística de dados

Principais métodos estatísticos. Utilização *Python* em métodos estatísticos. Organizar os dados para análise estatística.

Bloco 1

Yuri Sá

Introdução à análise estatística dos dados

Segundo Haslwanter (2016), estatística é a prática ou ciência de coletar e analisar dados numéricos, para extração de proporções, representação e interpretação do todo, por meio de amostras.

Algumas métricas:

- Média, mediana, moda, desvio e variância.
- Distribuição.
- Amplitude.
- Significância.

Módulos *Python*

Módulos utilizados:

- *NumPy.*
- *Pandas.*
- *SciPy.*
- *Matplotlib.*
- *Statsmodels.*

Análise estatística de dados

Principais métodos estatísticos. Utilização *Python* em métodos estatísticos. Organizar os dados para análise estatística.

Bloco 2

Yuri Sá

Métricas básicas

Métricas estatísticas básicas:

- Média.
- Mediana.
- Moda.
- Variância.
- Desvio padrão.
- Correlação:
  - Magnitude.
  - Sinal.
  - Matriz de correlação.

Um caso com dados e suas estatísticas

Caso:

Manipular e extrair estatísticas dados utilização de Internet no Brasil (IBGE, 2021).

To-do:

- Carregar os dados.
- Separar e selecionar os dados.
- Extrair métricas estatísticas básicas.
- Verificar correlação.
- Gerar gráficos.

Teoria em Prática

Bloco 3

Yuri Sá

Reflita sobre a seguinte situação

- Performance de um *e-commerce* caindo.
- Extração de dados:
  - Meses com queda.
  - Meses mais produtivos.
  - Produtos mais vendidos.
  - Dados sobre consumidores:
    - Idade.
    - Sexo.
    - Localização.

Norte para a resolução...

- Verifique como selecionar e separar dados.
- Foque na preparação dos dados.
- Crie variáveis e *DataFrames* novos com os dados selecionados.
- Agrupe dados segundo necessidade.

Dica do (a) Professor (a)

Bloco 4

Yuri Sá

Escolha do módulo e métodos

- A escolha do módulo *Python* e método estatístico para cada análise, é chave para o sucesso e eficiência do trabalho.
- A preparação dos dados ajuda, pois remove um volume morto de dados que atrapalham a visualização, consomem recursos e torna tudo mais lento.
- Separe uma pequena amostra do total dos dados para a prototipagem da análise. Eventualmente, seu *dataset* pode ser grande demais e, para pequenos testes durante a prototipagem, pode causar transtornos.

Referências

HASLWANTER, T. An introduction to Statistics with Python. With applications in the life sciences. **Switzerland: Springer International Publishing**, 2016.

IBGE. **Utilização da Internet**. Disponível em: https://ftp.ibge.gov.br/Acesso_a_internet_e_posse_celular/2015/Tabelas_de_Resultados/xlsx/01_Pessoas_de_10_Anos_ou_Mais_de_Idade/01_Utilizacao_da_Internet.xlsx. Acesso em: 29 mar. 2021.

Bons estudos!