Workshop - Análise de Estatísticas e Ciência de Dados com Python

Atividade Prática Introdutória sobre Ciência de Dados

Objetivo: Conhecer o ambiente de desenvolvimento e avaliar dados numéricos com base em cálculos estatísticos

Pré-requisitos: Linguagem de programação Python, Linux, estatística

Meta: Ao final da prática, o aluno será capaz de utilizar ferramentas de análise de dados para calcular indicadores estatísticos e comparar valores

Introdução:

A ferramenta de desenvolvimento a ser utilizada é chamada "*Jupyter Notebook*". Com ela, é possível executar comandos para análises de dados isoladamente, sem precisar executar todo o código sempre que for preciso fazer algum ajuste.

A principal API de desenvolvimento que será utilizada é chamada de Pandas. Com ela, é possível manipular estruturas de dados complexas para análises de dados.

Roteiro:

- Iniciar o Jupyter Notebook pela linha de comando do Linux e criar um projeto no Navegador
- Importar as bibliotecas Pandas e NumPy

```
In [1]:
1 import pandas as pd
2 import numpy as np
```

- Ler dados de arquivo "*series.csv*"

```
data = pd.read_csv('series.csv', index_col=False, header=None, squeeze=True);
print(data)
```

- Explorar os dados com base em estatísticas descritivas

```
# Mínimo
data.min()
```

```
3
```

```
# Máximo
data.max()
```

```
432
```

```
# Média
data.mean()
```

```
54.235294117647058
```

```
# Desvio Padrão
data.std()
```

```
101.93780543287454
```

```
# Mediana
data.median()
```

```
13.0
```

```
# Moda
data.mode()
```

```
0    5
```

- Visualização formatada das estatísticas

```
print('MIN: {}'.format(data.min()))
print('MAX: {}'.format(data.max()))
print('MÉDIA: {}'.format(data.mean()))
print('DESVIO PADRÃO: {}'.format(data.std()))
```

```
MIN: 3
MAX: 432
MÉDIA: 54.2352941176
DESVIO PADRÃO: 101.937805433
```

- Calcular os percentis

```
# 25o percentil (1o quartil)
data.quantile(.25)
```

```
6.0
```

```
# 50o percentil (2o quartil)
data.quantile(.50)
```

```
13.0
```

```
# 75o percentil (3o quartil)
data.quantile(.75)
```

```
67.0
```

```
# 95o percentil
data.quantile(.95)
```

```
156.79999999999976
```

- Calcular a tabela de frequências

```
# Tabela de Frequências
data.value_counts()
```

```
5      3
67     2
7      2
57     1
88     1
83     1
432    1
13     1
9      1
6      1
3      1
33     1
35     1
Name: 0, dtype: int64
```

- Para plotar gráficos, usar a biblioteca “*matplotlib.pyplot*”

```
import matplotlib.pyplot as plt
count, bins, ignored = plt.hist(postA, 100, normed=True, align='mid')
plt.show()
```

Atividade:

Faça um código para ler os arquivos "*altura_homens.csv*" e "*altura_mulheres.csv*". Esses arquivos contém as alturas (em cm) de 1000 homens e 1000 mulheres, respectivamente. Em seguida, responda às seguintes perguntas:

a) Qual a altura mínima e máxima dos homens e das mulheres dessas amostras?
b) Qual a média de altura dos homens e das mulheres? E qual a mediana dessas alturas?
c) Qual o desvio padrão da altura dos homens e das mulheres?
d) Qual o percentual de homens com altura menor que 160cm?
e) Qual o percentual de mulheres com altura maior que 180cm?
f) Quais as três alturas de homens que são as mais frequentes? Quantos homens possuem essas alturas?
g) Uma mulher com altura 185cm está distante quantos desvios padrões da média das mulheres?
h) É possível afirmar com determinado grau de confiança que uma pessoa com altura 150cm é um homem ou uma mulher?
i) As alturas dos homens e mulheres seguem uma distribuição Normal?
j) Execute, a partir da variável carregada com os dados, a função "*describe()*" e observe pra que serve.