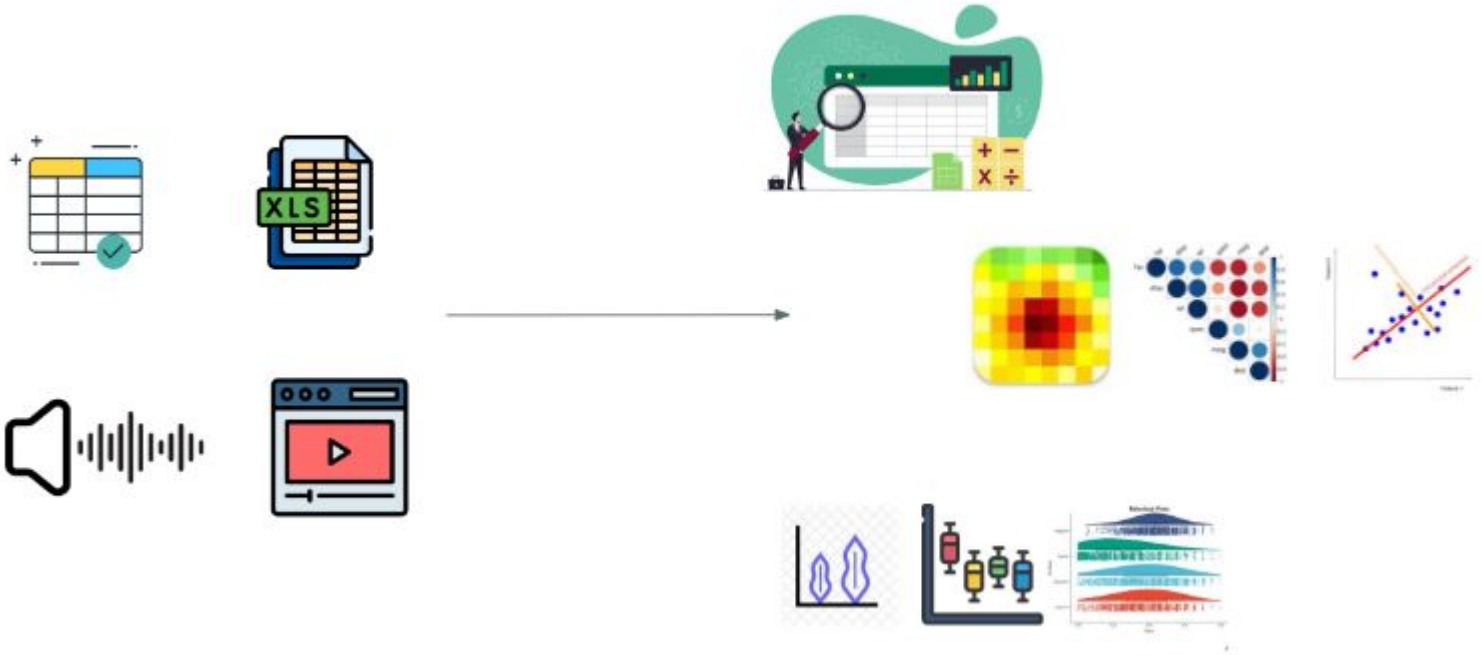
Coeficiente de Correlação Pearson

Resumo (Análise Exploratória de Dados)



Resumo (Análise Exploratória de Dados)

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

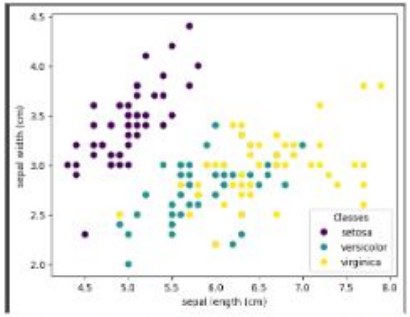
Imagem do Data Frame criado pelo Pandas.

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Principais características do Data Frame, para cada atributo

```
RangeIndex: 150 entries, 0 to 149
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   sepal length (cm)  150 non-null    float64
 1   sepal width (cm)   150 non-null    float64
 2   petal length (cm)  150 non-null    float64
 3   petal width (cm)   150 non-null    float64
dtypes: float64(4)
```

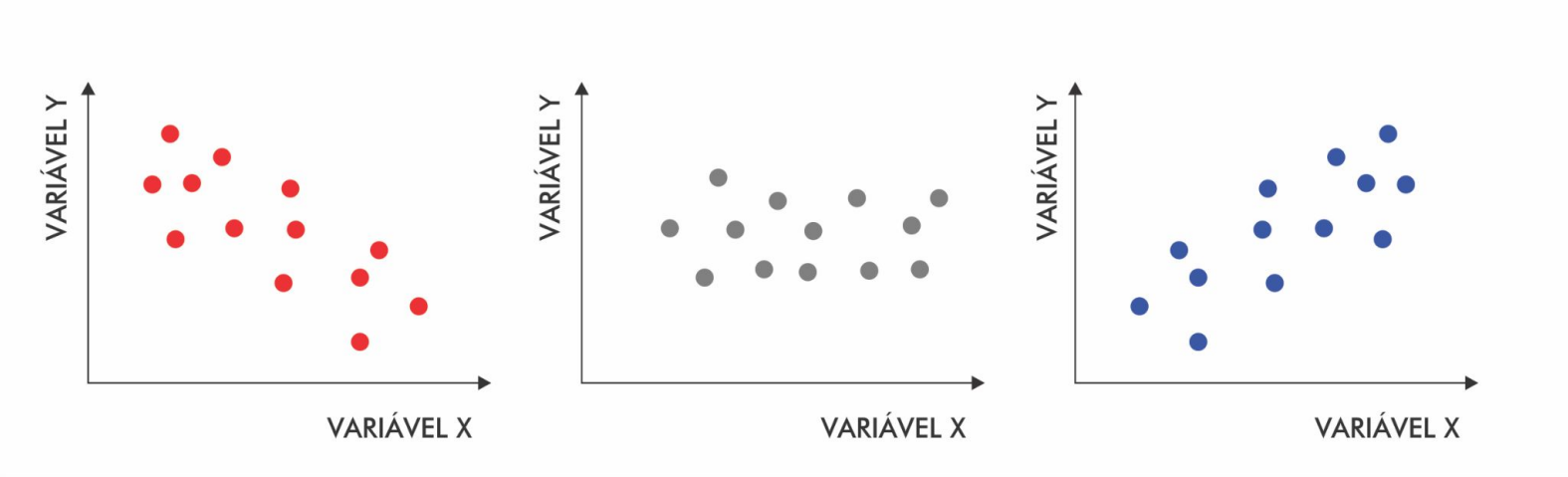
Características de cada atributo.



Visualização 2D da relação de tamanho entre largura e altura das sépalas.

Objetivo

- Entender a relação entre duas variáveis distintas de um mesmo conjunto de dados;



Cálculo do Coeficiente de Pearson

Sejam os valores $x_1,x_2,x_3,...,x_n$ e $y_1,y_2,y_3,...,y_n$, sendo i=1,2,3,...n correspondente ao dado “i” em duas variáveis distintas X e Y, temos que:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

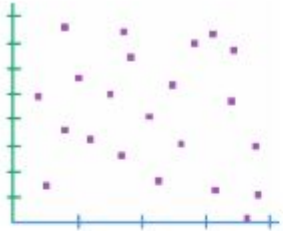
- x_i e y_i os valores do dado **i** nas variáveis X e Y;
- $\bar{x}$ e $\bar{y}$ as médias aritméticas de x_i e y_i;
- **r** é o valor do coeficiente;

Características do Coeficiente

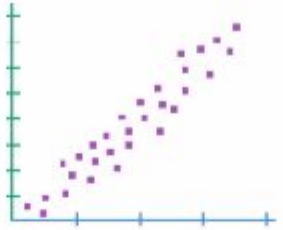
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

- $-1 \leq r \leq 1$;
- $r = 0$, indica que não há relação entre as variáveis;
- quanto maior $|r|$, maior a relação;
- $r > 0$, indica que as variáveis são diretamente proporcionais;
- $r < 0$, indica que as variáveis são inversamente proporcionais;

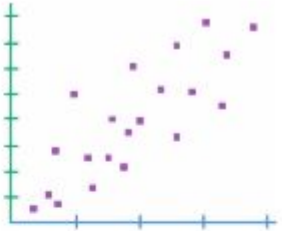
Visualização por gráficos



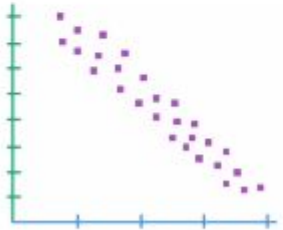
Sem correlação



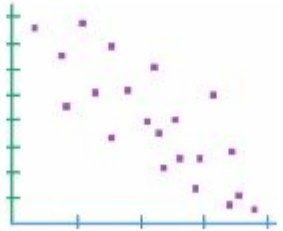
Correlação positiva forte



Correlação positiva média

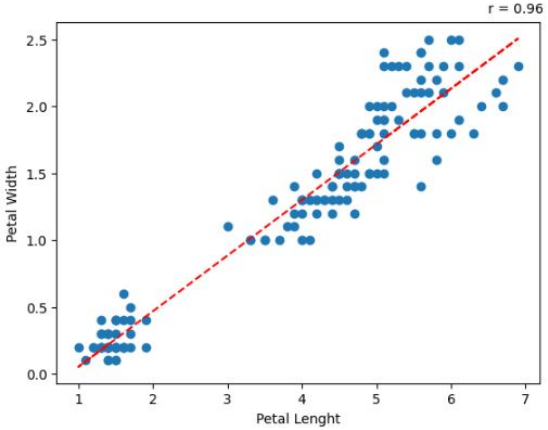


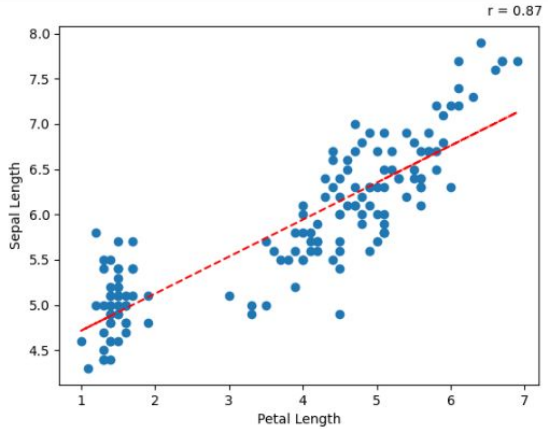
Correlação negativa forte

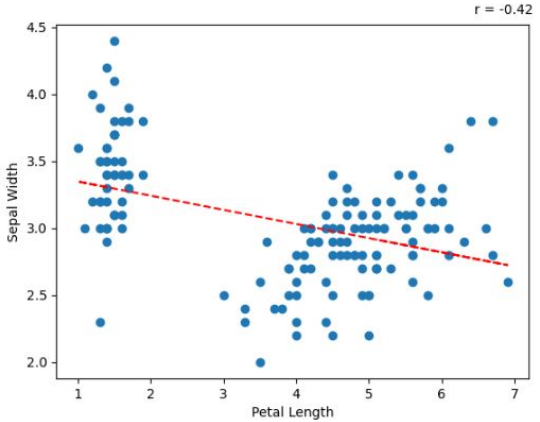


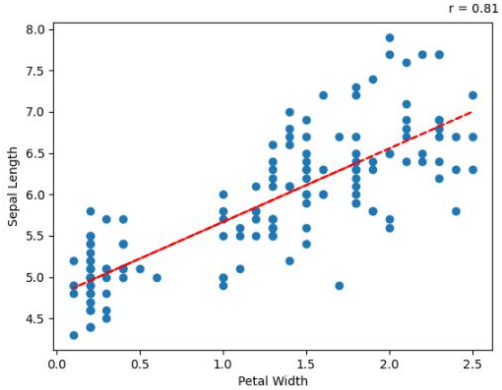
Correlação negativa média

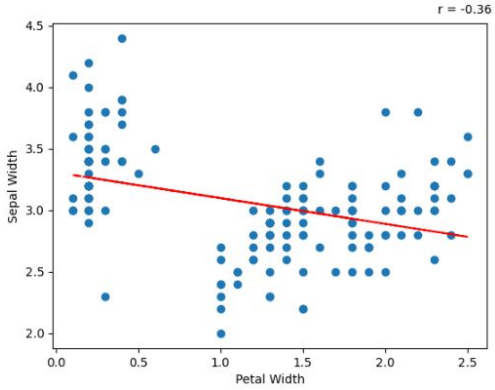
Exemplo com o Dataset Iris

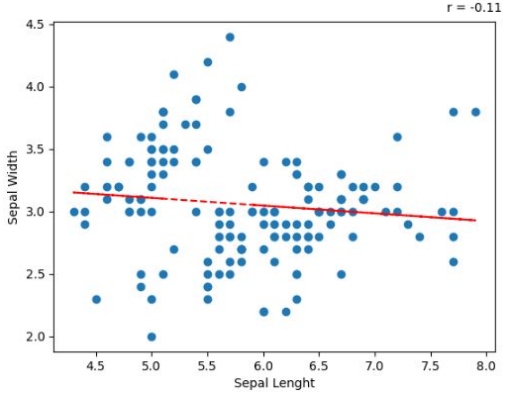












Outros cálculos para correlação

- Coeficiente de Correlação Kendall;
- Coeficiente de Correlação de Postos Spearman;