



LADE

LABORATÓRIO DE ANÁLISES DE DADOS
EDUCACIONAIS E ESTATÍSTICA APLICADA

IFCE - CAMPUS FORTALEZA

Estatística Descritiva com Python

Análise de dados : Teoria e Tecnologias

1. Importação de Dados
2. Limpeza e Preparação dos Dados
3. Análise Exploratória de Dados (EDA)
4. Modelagem de Dados
5. Visualização de Resultados

Fonte: <https://www.cursospm3.com.br/blog/python-para-analise-de-dados/>

1. Planejamento
2. Coleta de Dados
3. Crítica dos dados
4. Apuração dos dados
5. Análises de Dados
6. Emissão do Relatório Final
7. Comunicação dos Resultados

Fonte: Costa, Giovani. Estatística Aplicada A Educacao Com Abordagem Além da análise descritiva

1. Olhar para o quadro geral;
2. Obter os dados;
3. Descobrir e visualizar os dados para obter informações;
4. Preparar os dados para os algoritmos do Aprendizado de Máquina;
5. Selecionar e treinar um modelo;
6. Ajustar o seu modelo;
7. Apresentar sua solução;
8. Lançar, monitorar e manter seu sistema.

Fonte :Géron, Aurélien. Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow

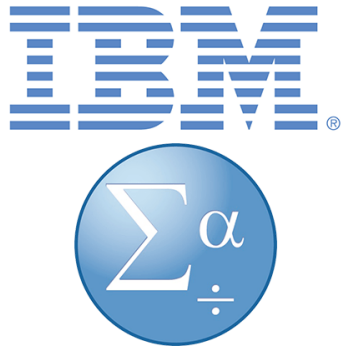


- 1. Defina seus objetivos:** Antes de começar a análise, é crucial entender o que você quer alcançar. Qual é a pergunta ou problema que você está tentando resolver?
- 2. Coleta de dados:** Reúna os dados relevantes para sua análise. Isso pode envolver a obtenção de dados de fontes diferentes, como bancos de dados, arquivos CSV, APIs da web, etc.
- 3. Limpeza de dados:** Os dados geralmente precisam ser limpos e preparados para análise. Isso pode incluir a remoção de dados duplicados, tratamento de valores ausentes, padronização de formatos de dados, entre outras tarefas.
- 4. Exploração de dados:** Nesta etapa, você explora seus dados para entender melhor sua estrutura e identificar padrões, tendências e relações. Isso pode envolver a criação de visualizações de dados, como gráficos e tabelas.
- 5. Análise estatística:** Use técnicas estatísticas para analisar seus dados de maneira mais aprofundada. Isso pode incluir medidas de tendência central, dispersão, correlação, regressão, entre outras.
- 6. Modelagem de dados (se aplicável):** Se você estiver realizando análises preditivas ou inferenciais, pode ser necessário criar modelos estatísticos ou de aprendizado de máquina para fazer previsões ou extrair insights dos dados.
- 7. Interpretação dos resultados:** Analise os resultados de sua análise e interprete o que eles significam em relação aos seus objetivos iniciais. Isso pode envolver a elaboração de conclusões e recomendações com base nos insights obtidos.
- 8. Comunicação dos resultados:** Comunique seus resultados de forma clara e compreensível para as partes interessadas relevantes. Isso pode incluir a criação de relatórios, apresentações ou visualizações de dados para transmitir suas descobertas.
- 9. Iteração e refinamento:** Às vezes, é necessário revisitar etapas anteriores, refinar suas análises ou coletar mais dados para obter uma compreensão mais completa do problema em questão.

(ChatGPT 3.5)



Análise de dados : Teoria e Tecnologias



Usaremos :



Pandas - <https://pandas.pydata.org/>

Uma das bibliotecas mais conhecidas e mais usadas por profissionais de dados, ela permite que o usuário **manipule, transforme e analise dados** de maneira muito otimizada.

A **Pandas** possibilita a leitura em vários formatos, como SQL, CSV, Excel, etc., além de funcionar, principalmente, com dois tipos de estrutura de dados: *Series* e *DataFrames*.

DataFrames seguem uma estrutura semelhante a uma planilha de **Excel**.

Series, se refere a um **array unidimensional**, que pode ser entendido como uma lista simples de valores.

NumPy - <https://numpy.org/>

A biblioteca **Numpy compila funções** relacionadas à álgebra linear e computação numérica, trabalhando com *arrays* multidimensionais, cálculos rápidos, entre outras funcionalidades.

Além disso, a biblioteca **NumPy está no núcleo de basicamente todos** os programas e bibliotecas que lidam com operações matemáticas e usam a linguagem de programação Python.

Como por exemplo, a própria biblioteca Pandas baseia sua estrutura de dados (*DataFrames* e *Series*) em *arrays* de NumPy.

Fonte: <https://www.cursospm3.com.br/blog/python-para-analise-de-dados/>



Gráficos - Visualização de Dados

Quantos números 9?

2	2	5	6	7	1	1	6	9	1
9	1	7	5	5	5	6	2	5	9
4	5	2	9	6	9	7	6	4	6
8	1	5	7	8	5	6	6	6	7
7	2	3	6	8	9	1	7	9	1
3	8	6	8	4	5	6	9	4	5
4	9	9	2	3	7	1	9	1	2
3	7	8	1	6	1	5	6	1	6
5	6	6	8	6	6	9	1	2	6
3	2	4	2	6	9	4	2	7	1

FIGURE 1.3 How many 9s are there?

E agora, quantos
números 9?

2	2	5	6	7	1	1	6	9	1
9	1	7	5	5	5	6	2	5	9
4	5	2	9	6	9	7	6	4	6
8	1	5	7	8	5	6	6	6	7
7	2	3	6	8	9	1	7	9	1
3	8	6	8	4	5	6	9	4	5
4	9	9	2	3	7	1	9	1	2
3	7	8	1	6	1	5	6	1	6
5	6	6	8	6	6	9	1	2	6
3	2	4	2	6	9	4	2	7	1

FIGURE 1.4 Now it's easy to count the 9s.

E em um gráfico, como ficaria?

2	2	5	6	7	1	1	6	9	1
9	1	7	5	5	5	6	2	5	9
4	5	2	9	6	9	7	6	4	6
8	1	5	7	8	5	6	6	6	7
7	2	3	6	8	9	1	7	9	1
3	8	6	8	4	5	6	9	4	5
4	9	9	2	3	7	1	9	1	2
3	7	8	1	6	1	5	6	1	6
5	6	6	8	6	6	9	1	2	6
3	2	4	2	6	9	4	2	7	1

FIGURE 1.4 Now it's easy to count the 9s.

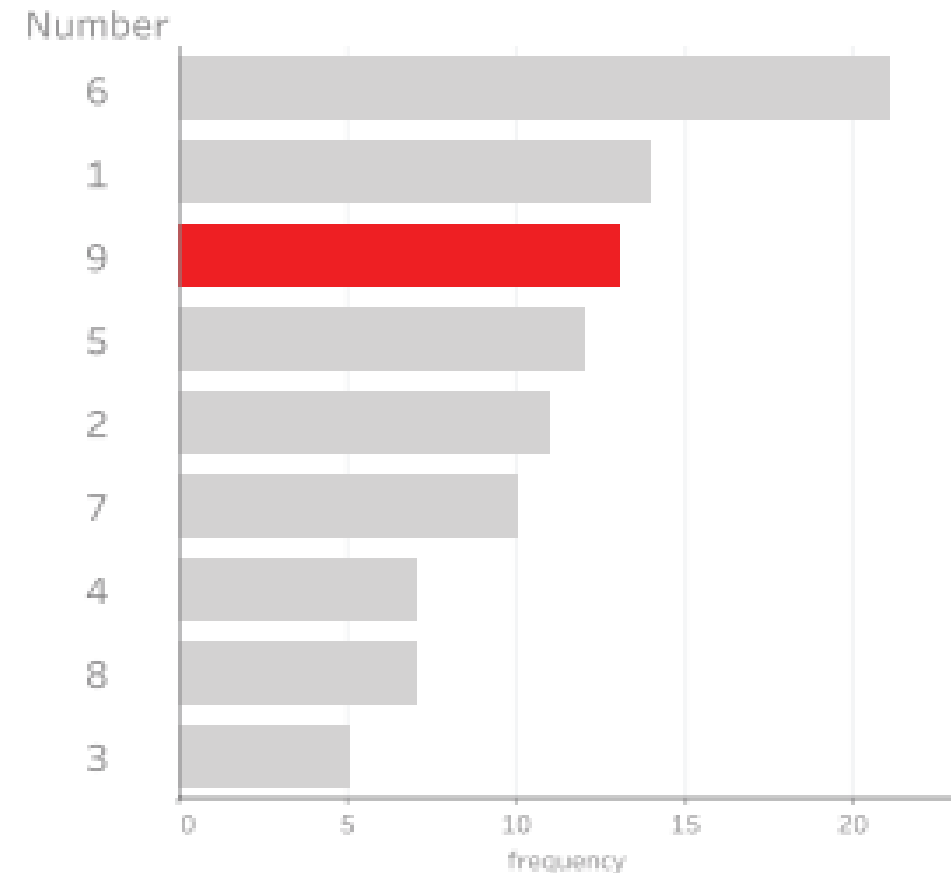


FIGURE 1.9 Sorted bar chart using color and length to show how many 9s are in our table.

TABLE 1.2 What are the trends in sales?

Category	2013 Q1	2013 Q2	2013 Q3	2013 Q4	2014 Q1	2014 Q2	2014 Q3	2014 Q4
Furniture	\$463,988	\$352,779	\$338,169	\$317,735	\$320,875	\$287,934	\$319,537	\$324,319
Office Supplies	\$232,558	\$290,055	\$265,083	\$246,946	\$219,514	\$202,412	\$198,268	\$279,679
Technology	\$563,866	\$244,045	\$432,299	\$461,616	\$285,527	\$353,237	\$338,360	\$420,018
Category	2015 Q1	2015 Q2	2015 Q3	2015 Q4	2016 Q1	2016 Q2	2016 Q3	2016 Q4
Furniture	\$307,028	\$273,836	\$290,886	\$397,912	\$337,299	\$245,445	\$286,972	\$313,878
Office Supplies	\$207,363	\$183,631	\$191,405	\$217,950	\$241,281	\$286,548	\$217,198	\$272,870
Technology	\$333,002	\$291,116	\$356,243	\$386,445	\$386,387	\$397,201	\$359,656	\$375,229

O que você vê?

Há algo perceptível, no que tange a análise de dados?

É possível analisar esses dados, do jeito que estão dispostos?

É possível responder a pergunta: Quais são as tendências de vendas para tecnologia?



FIGURE 1.2 Now can you see the trends?

A visualização de dados é de FATO o que chamamos de “entrega”. É o produto final!

Gráficos

Os métodos gráficos têm encontrado um uso cada vez maior devido ao seu forte apelo visual. Normalmente, é mais fácil para qualquer pessoa entender a mensagem de um gráfico do que aquela embutida em tabelas ou sumários numéricos.

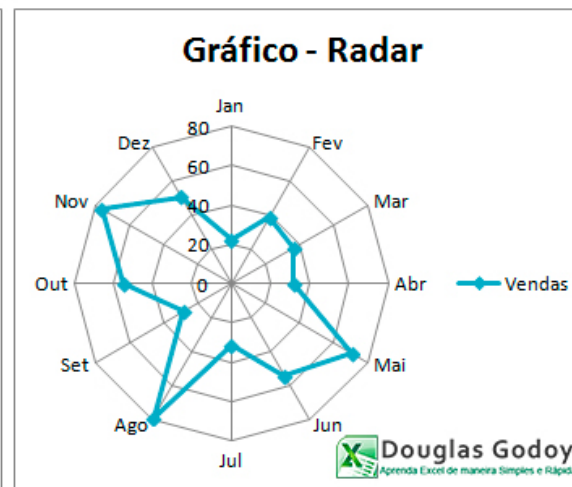
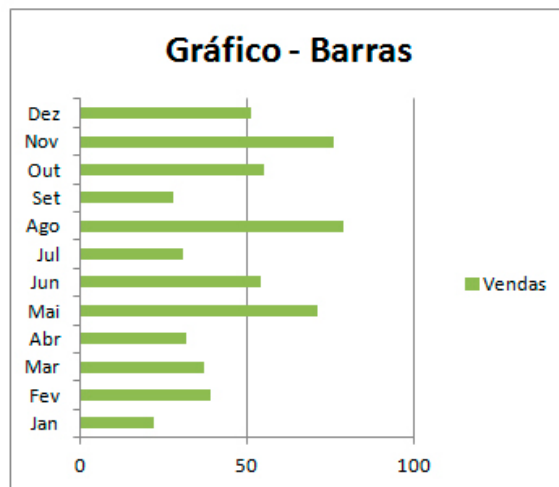
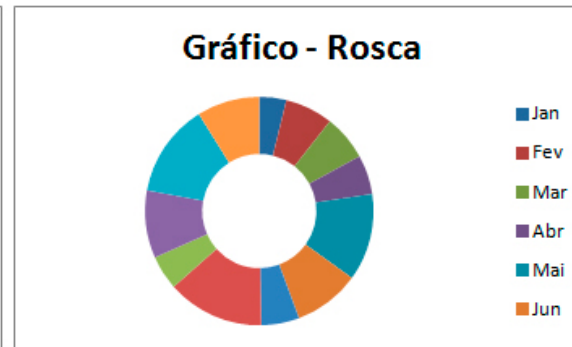
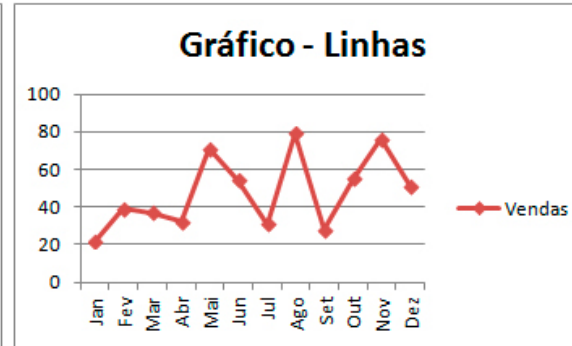
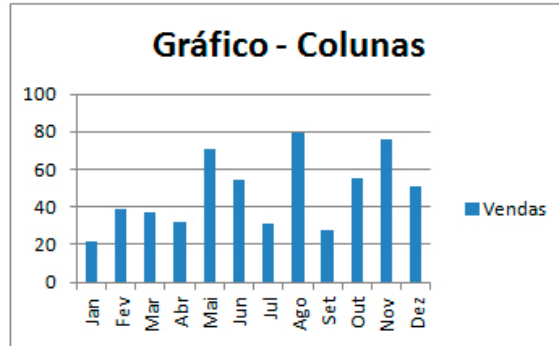
Os gráficos são utilizados para diversos fins:

- (a) buscar padrões e relações;**
- (b) confirmar (ou não) certas expectativas que se tinha sobre os dados**
- (c) descobrir novos fenômenos;**
- (d) confirmar (ou não) suposições feitas sobre os procedimentos estatísticos usados; e**
- (e) apresentar resultados de modo mais rápido e fácil.**

O que é gráfico?

É uma representação de informações obtidas em pesquisas por meio de formas geométricas para facilitar a leitura dos dados.

	Vendas
Jan	22
Fev	39
Mar	37
Abr	32
Mai	71
Jun	54
Jul	31
Ago	79
Set	28
Out	55
Nov	76
Dez	51



Gráficos para Variáveis Qualitativas

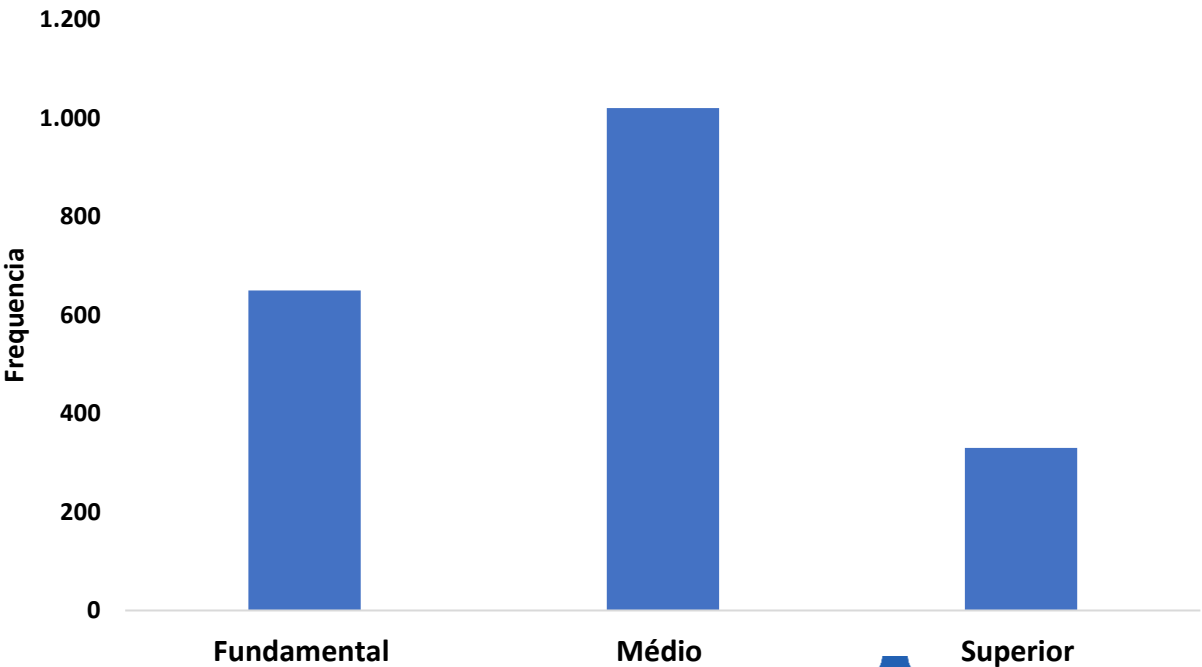
Informações sobre estado civil, grau de instrução, número de filhos, salário (expresso como fração do salário mínimo), idade (medida em anos e meses) e procedência de 36 empregados da seção de orçamentos da Companhia MB.

Nº	Estado civil	Grau de instrução	Nº de filhos	Salário (× sal. mín.)	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	—	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	—	5,73	20	10	outra
5	solteiro	ensino fundamental	—	6,26	40	07	outra
6	casado	ensino fundamental	0	6,66	28	00	interior
7	solteiro	ensino fundamental	—	6,86	41	00	interior
8	solteiro	ensino fundamental	—	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	capital
10	solteiro	ensino médio	—	7,44	23	06	outra
11	casado	ensino médio	2	8,12	33	06	interior
12	solteiro	ensino fundamental	—	8,46	27	11	capital
13	solteiro	ensino médio	—	8,74	37	05	outra
14	casado	ensino fundamental	3	8,95	44	02	outra
15	casado	ensino médio	0	9,13	30	05	interior
16	solteiro	ensino médio	—	9,35	38	08	outra
17	casado	ensino médio	1	9,77	31	07	capital
18	casado	ensino fundamental	2	9,80	39	07	outra
19	solteiro	superior	—	10,53	25	08	interior
20	solteiro	ensino médio	—	10,76	37	04	interior
21	casado	ensino médio	1	11,06	30	09	outra
22	solteiro	ensino médio	—	11,59	34	02	capital
23	solteiro	ensino fundamental	—	12,00	41	00	outra
24	casado	superior	0	12,79	26	01	outra
25	casado	ensino médio	2	13,23	32	05	interior
26	casado	ensino médio	2	13,60	35	00	outra
27	solteiro	ensino fundamental	—	13,85	46	07	outra
28	casado	ensino médio	0	14,69	29	08	interior
29	casado	ensino médio	5	14,71	40	06	interior
30	casado	ensino médio	2	15,99	35	10	capital
31	solteiro	superior	—	16,22	31	05	outra
32	casado	ensino médio	1	16,61	36	04	interior
33	casado	superior	3	17,26	43	07	capital
34	solteiro	superior	—	18,75	33	07	capital
35	casado	ensino médio	2	19,40	48	11	capital
36	casado	superior	3	23,30	42	02	interior

Fonte: Dados hipotéticos.

Grau de instrução	Frequência n_i	Porcentagem $100f_i$
Fundamental	650	32,50
Médio	1.020	51,00
Superior	330	16,50
Total	2.000	100,00

Gráfico em Barras



OBS.: Apresentação gráfica é um complemento importante da apresentação tabular.

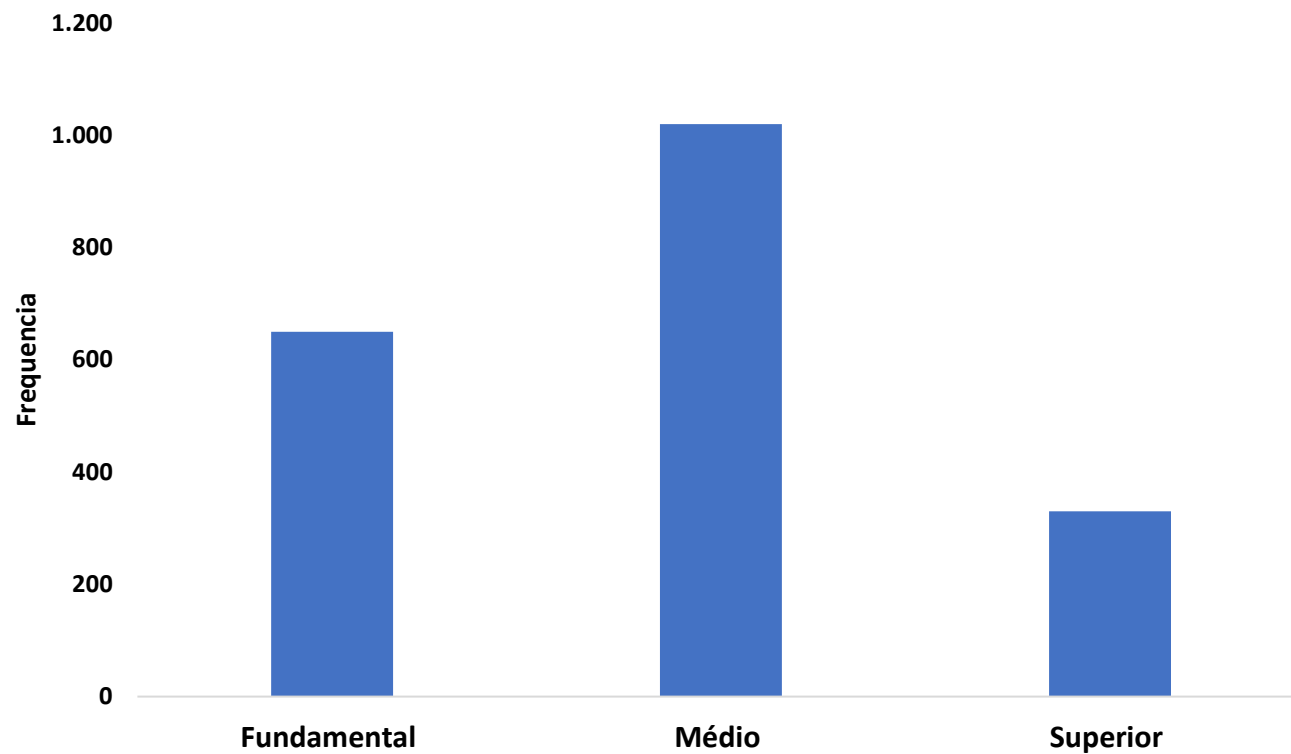
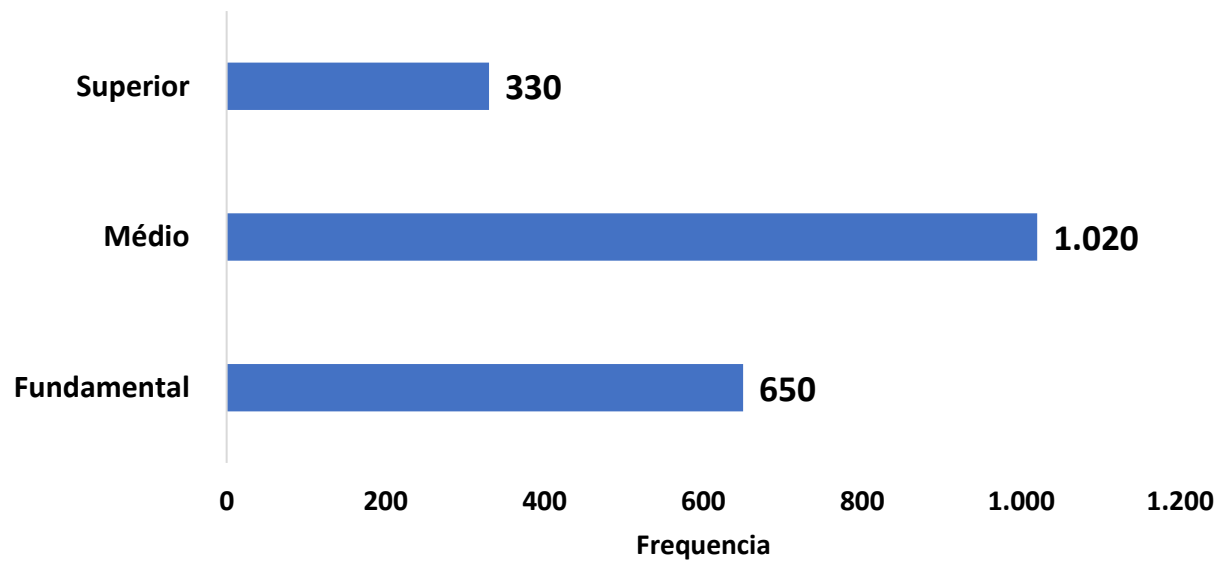
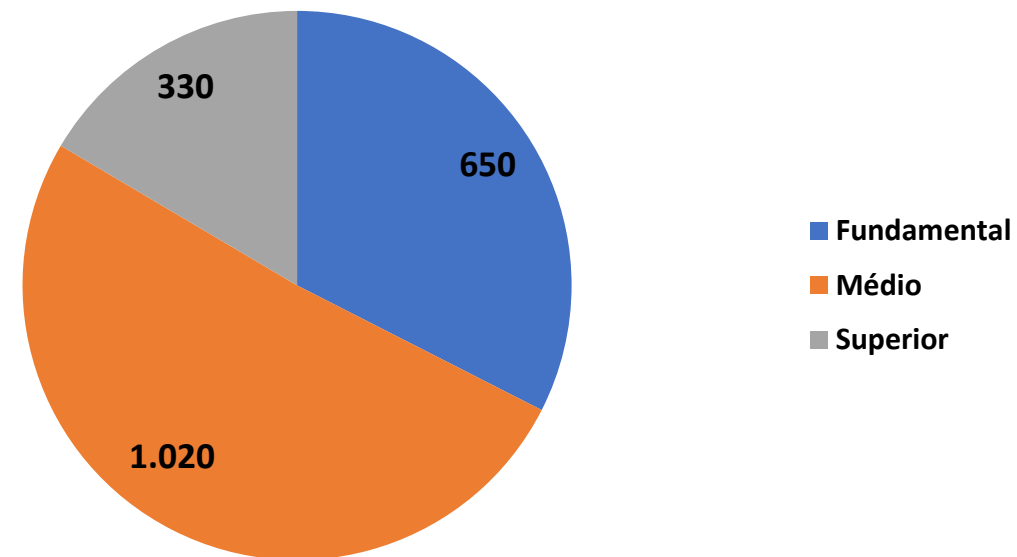


Gráfico em Setores



Gráficos para Variáveis Quantitativas

Gráfico de dispersão unidimensional

Gráficos de dispersão unidimensionais para a variável Z: número de filhos.

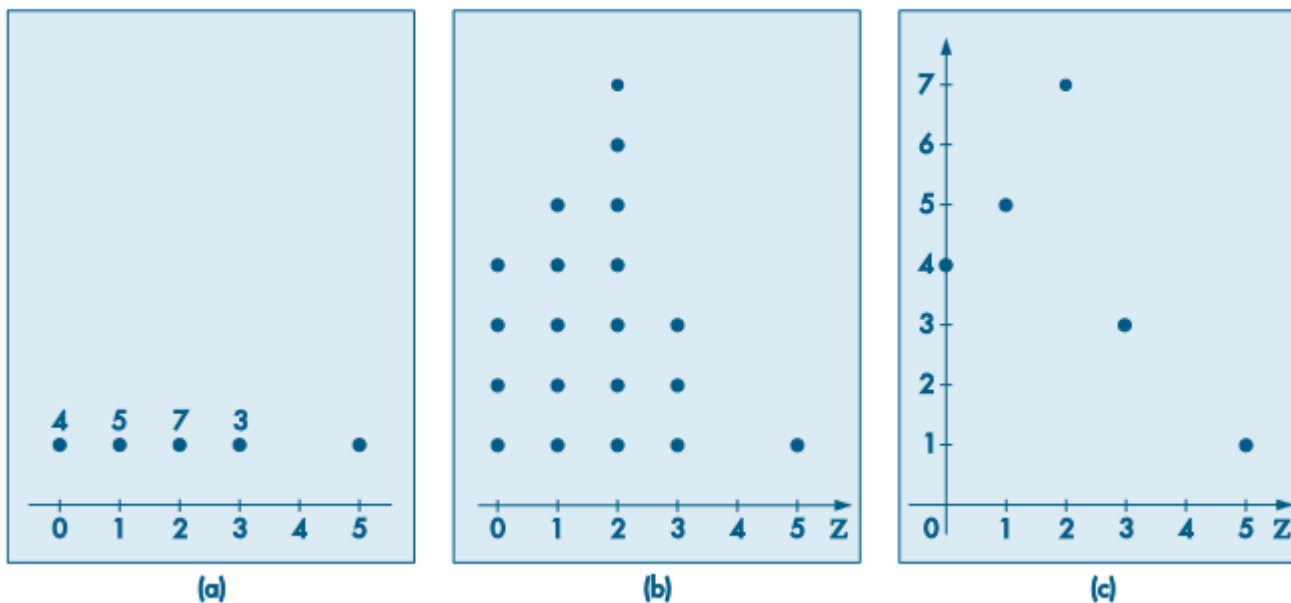
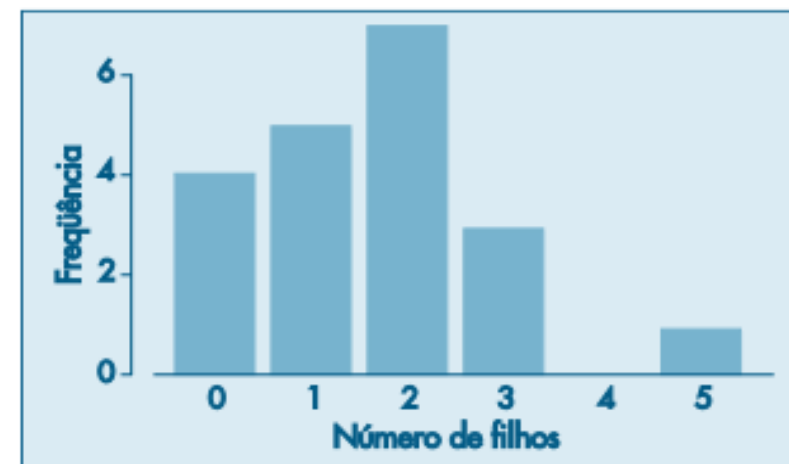


Gráfico em Barras

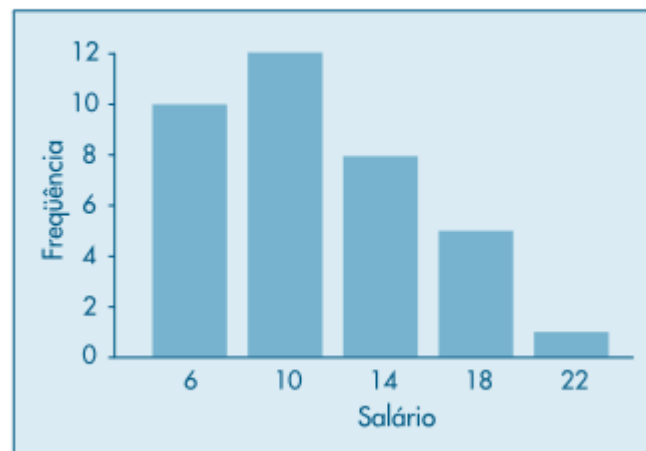
Gráfico em barras para a variável Z: número de filhos.



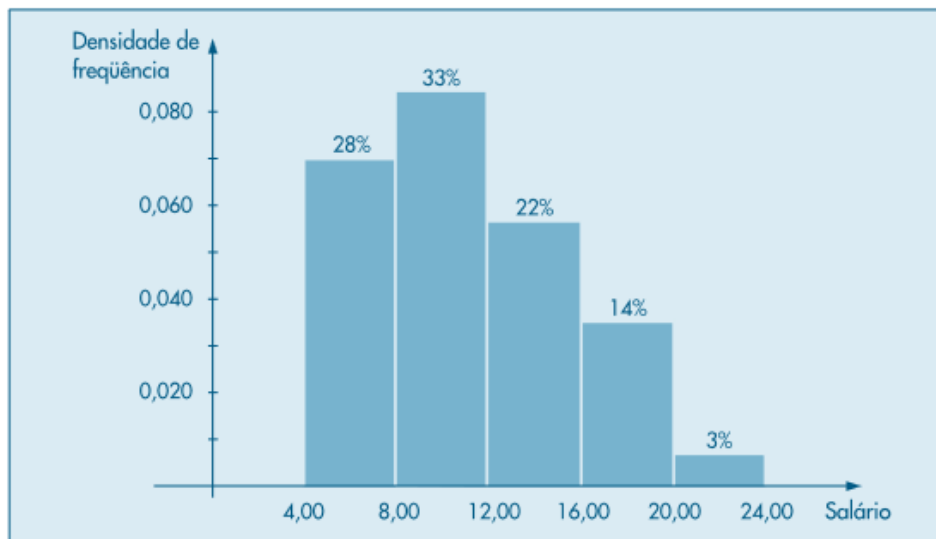
Distribuição de freqüências da variável S , salário dos empregados da seção de orçamentos da Companhia MB.

Classes de salários	Ponto médio s_i	Freqüência n_i	Porcentagem $100 f_i$
4,00 – 8,00	6,00	10	27,78
8,00 – 12,00	10,00	12	33,33
12,00 – 16,00	14,00	8	22,22
16,00 – 20,00	18,00	5	13,89
20,00 – 24,00	22,00	1	2,78
Total	—	36	100,00

Gráfico em barras para a variável S : salários.



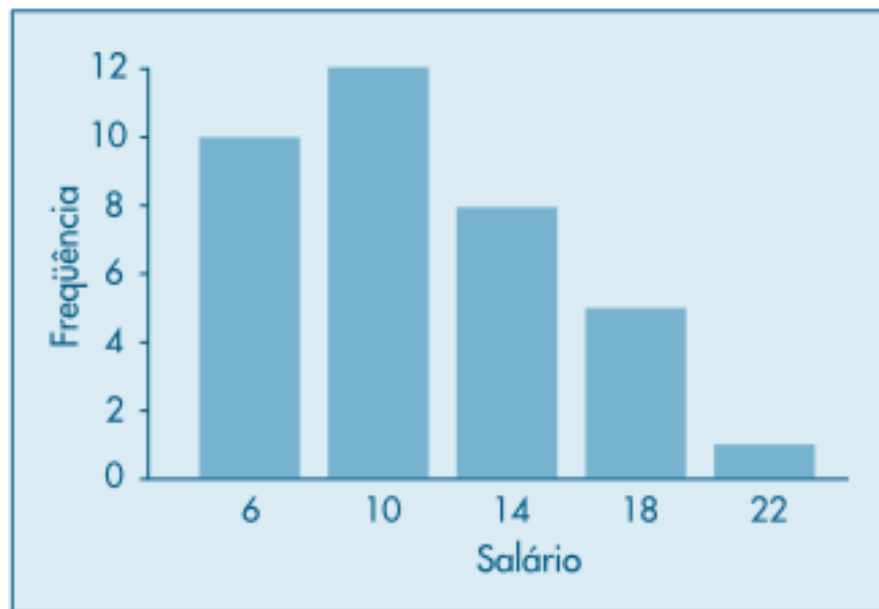
Histograma da variável S : salários.



Distribuição de freqüências da variável S , salário dos empregados da seção de orçamentos da Companhia MB.

Classes de salários	Ponto médio s_i	Frequência n_i	Porcentagem $100 f_i$
4,00 – 8,00	6,00	10	27,78
8,00 – 12,00	10,00	12	33,33
12,00 – 16,00	14,00	8	22,22
16,00 – 20,00	18,00	5	13,89
20,00 – 24,00	22,00	1	2,78
Total	—	36	100,00

Gráfico em barras para a variável S : salários.



b = base do retângulo = amplitude do intervalo

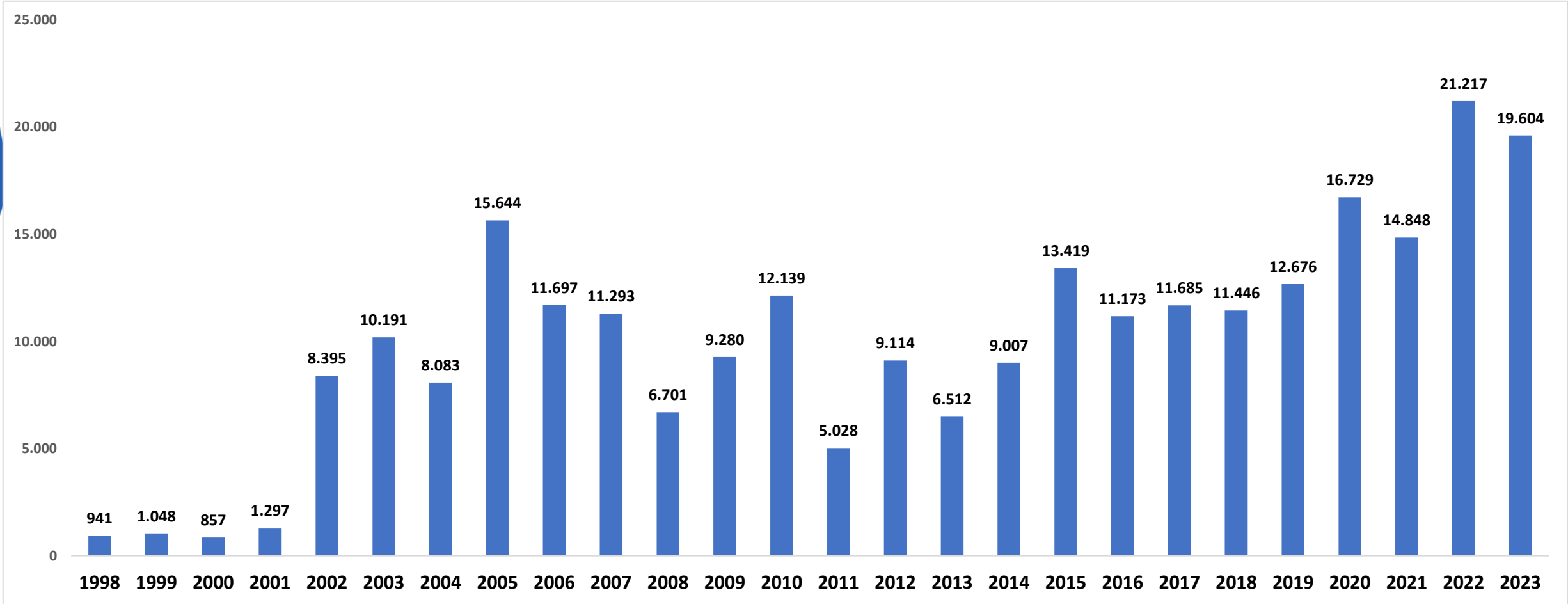
H = altura do retângulo

S = área do retângulo = frequência da classe

$$H = S/b$$

Gráfico em Linha

Os gráficos em linhas são frequentemente usados para representação de séries de tempo (quando um dos fatores for o tempo), isto porque quando a série cobre um grande número de períodos de tempo, a representação dos valores através de colunas pode conduzir a uma excessiva concentração de dados.



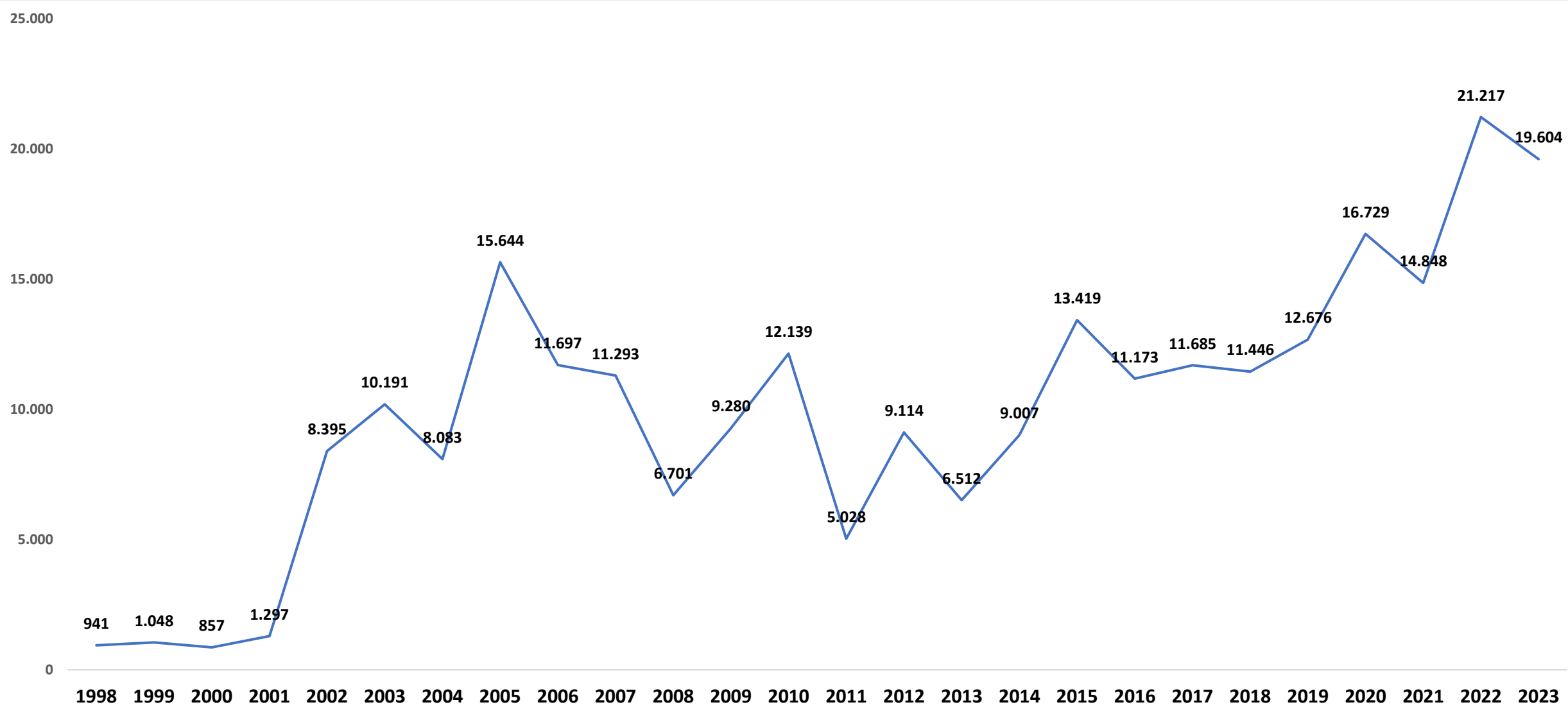
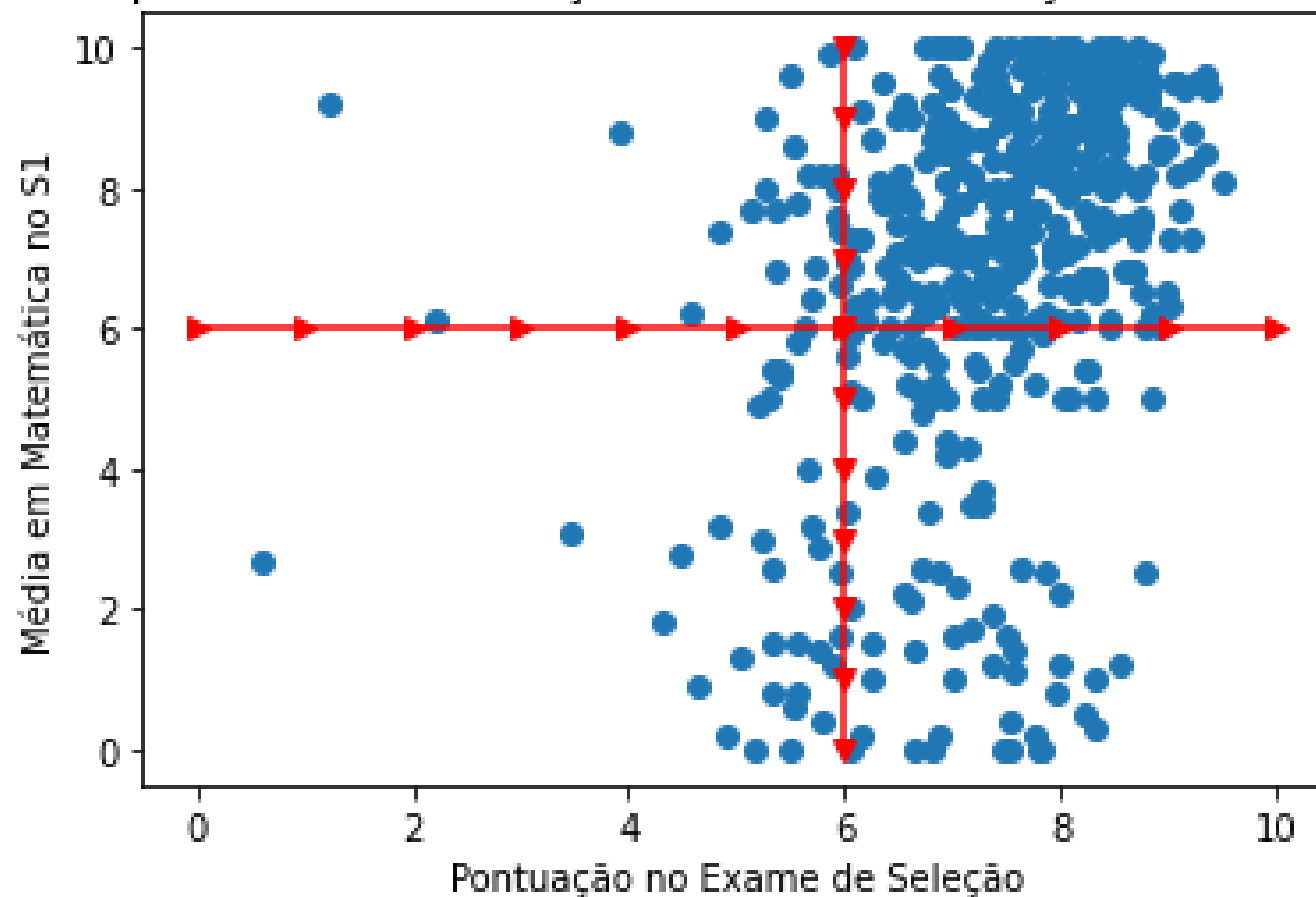


Gráfico de Dispersão entre Pontuação no Exame de Seleção e Média em Matemática



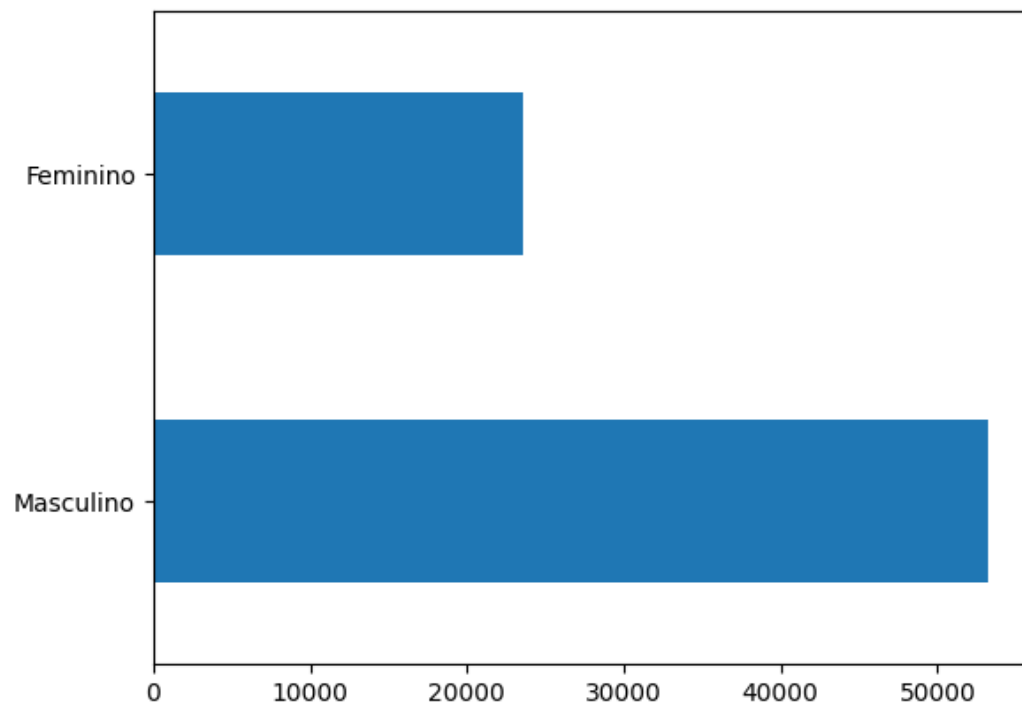
Usando o Python

```
✓ [14] #Frequencia Absoluta do Sexo  
0s FreuenciaSexo = PesquisaNacional['Sexo'].value_counts()  
FrequenciaSexo
```

```
Masculino    53250  
Feminino     23590  
Name: Sexo, dtype: int64
```

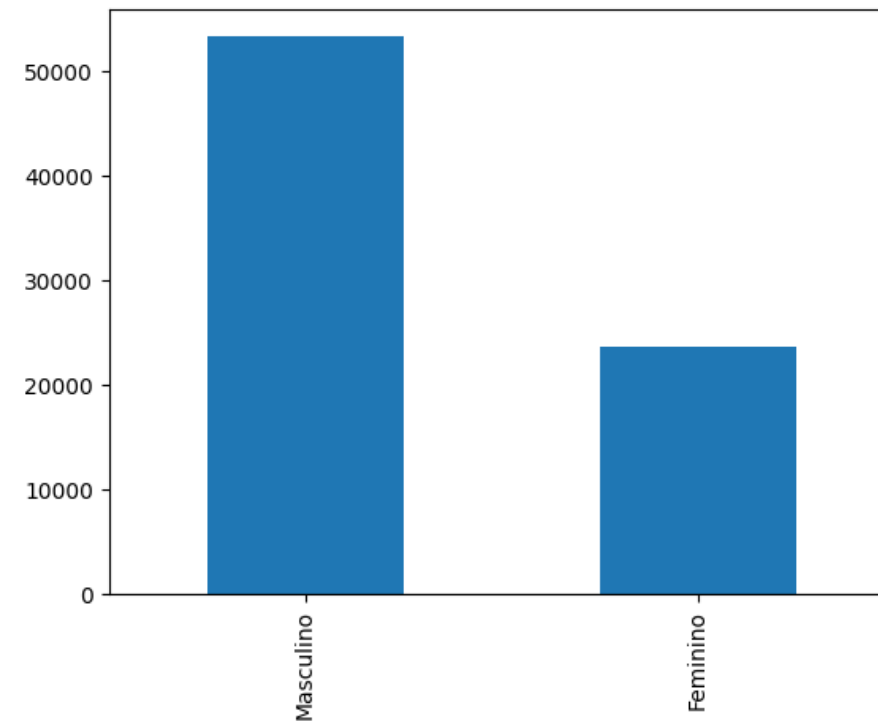
```
✓ [16] FreuenciaSexo.plot(kind='barh')
```

<Axes: >



```
✓ FreuenciaSexo.plot(kind='bar')
```

<Axes: >



[matplotlib.pyplot.step](#)
[matplotlib.pyplot.loglog](#)
[matplotlib.pyplot.semilogx](#)
[matplotlib.pyplot.semilogy](#)
[matplotlib.pyplot.fill_between](#)
[matplotlib.pyplot.fill_betweenx](#)
[matplotlib.pyplot.bar](#)
[matplotlib.pyplot.barh](#)
[matplotlib.pyplot.bar_label](#)
[matplotlib.pyplot.stem](#)
[matplotlib.pyplot.eventplot](#)
[matplotlib.pyplot.pie](#)
[matplotlib.pyplot.stackplot](#)
[matplotlib.pyplot.broken_barh](#)
[matplotlib.pyplot.vlines](#)
[matplotlib.pyplot.hlines](#)
[matplotlib.pyplot.fill](#)
[matplotlib.pyplot.polar](#)
[matplotlib.pyplot.axhline](#)
[matplotlib.pyplot.axhspan](#)
[matplotlib.pyplot.axvline](#)
[matplotlib.pyplot.axvspan](#)
[matplotlib.pyplot.axline](#)
[matplotlib.pyplot.acorr](#)

[Home](#) > [API Reference](#) > [matplotlib.pyplot](#) > **[matplotlib.pyplot.bar](#)**

matplotlib.pyplot.bar

matplotlib.pyplot.bar(*x*, *height*, *width=0.8*, *bottom=None*, ***, *align='center'*, *data=None*, ***kwargs*) [\[source\]](#)

Make a bar plot.

The bars are positioned at *x* with the given *align*. Their dimensions are given by *height* and *width*. The vertical baseline is *bottom* (default 0).

Many parameters can take either a single value applying to all bars or a sequence of values, one for each bar.

Parameters:

***x* : float or array-like**

The x coordinates of the bars. See also *align* for the alignment of the bars to the coordinates.

***height* : float or array-like**

The height(s) of the bars.

Note that if *bottom* has units (e.g. datetime), *height* should be in units that are a difference from the value of *bottom* (e.g. timedelta).

***width* : float or array-like, default: 0.8**

The width(s) of the bars.

Note that if *x* has units (e.g. datetime), then *width* should be in units that are a difference (e.g. timedelta) around the *x* values.

On this page

[bar\(\)](#)

Examples using

[matplotlib.pyplot.bar](#)

```
fig, ax = plt.subplots()

Sexo = ['Masculino', 'Feminino']
counts = [53250, 23590]
bar_labels = ['M', 'F']
bar_colors = ['tab:red', 'tab:blue']

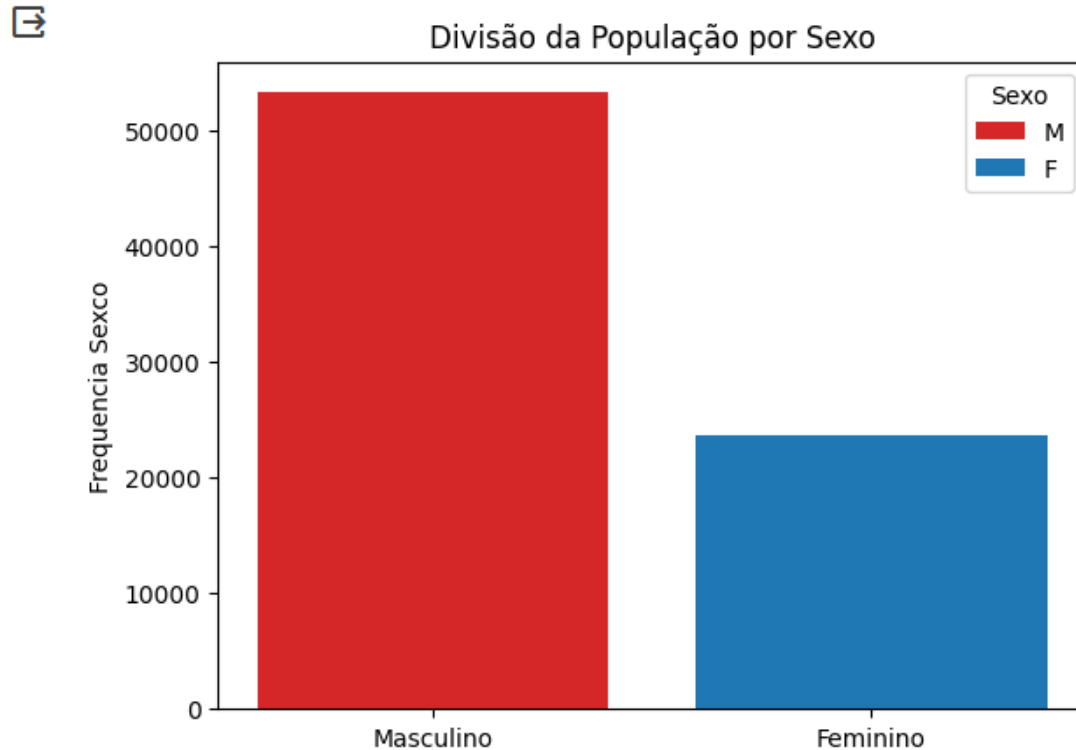
ax.bar(Sexo, counts, label=bar_labels, color=bar_colors)

ax.set_ylabel('Frequencia Sexco')
ax.set_title('Divisão da População por Sexo')
ax.legend(title='Sexo')

plt.show()
```

Estudar a documentação :

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.bar.html



```
[ ] import matplotlib.pyplot as plt
import numpy as np
import random      # Importando o módulo random para gerar números aleatórios
import pandas as pd
```


```
▶ # Criando um vetor Aleatorio de Idades, variação da idade de 20 a 78
Idade=[]                # Idade=[]: Inicializando uma lista vazia chamada "Idade"
for i in range(20):     # range(): gerar números aleatórios , Iterando 20 vezes usando um loop for
    Idade.append(random.randint(20,78)) # random.randint(20,78) : Gerando um número inteiro aleatório no intervalo de 20 a 78 ,
                                     # Idade.append: Adicionando o número gerado à lista "Idade"
```

```
[ ] # Criando um vetor de instrução com Em - Ensino Médio, Ef - Ensino Fundamenatal, Es - Ensino Superior
Categorias=["Ef","Em","Es"]
Instrucao=[]
for i in range(20):
    Instrucao.append(random.choice(Categorias))
```



```
[ ] # Montando a base de dados
BancoInsIdade=pd.DataFrame({'Idade':Idade,'Instrução':Instrucao})
```

▶ # Mostrando a base de dados
BancoInsIdade



	Idade	Instrução
0	34	Es
1	21	Es
2	43	Em
3	56	Es
4	77	Es
5	44	Es

▶ #Tabela de frequencia para variavel Instrução
TabelaFrequenciaInstrucao=BancoInsIdade['Instrução'].value_counts()
TabelaFrequenciaInstrucao

```
Es    10
Ef     7
Em     3
Name: Instrução, dtype: int64
```

```
import matplotlib.pyplot as plt # Importando a biblioteca matplotlib.pyplot para visualização de dados
import numpy as np
import random # Importando o módulo random para gerar números aleatórios
import pandas as pd
```

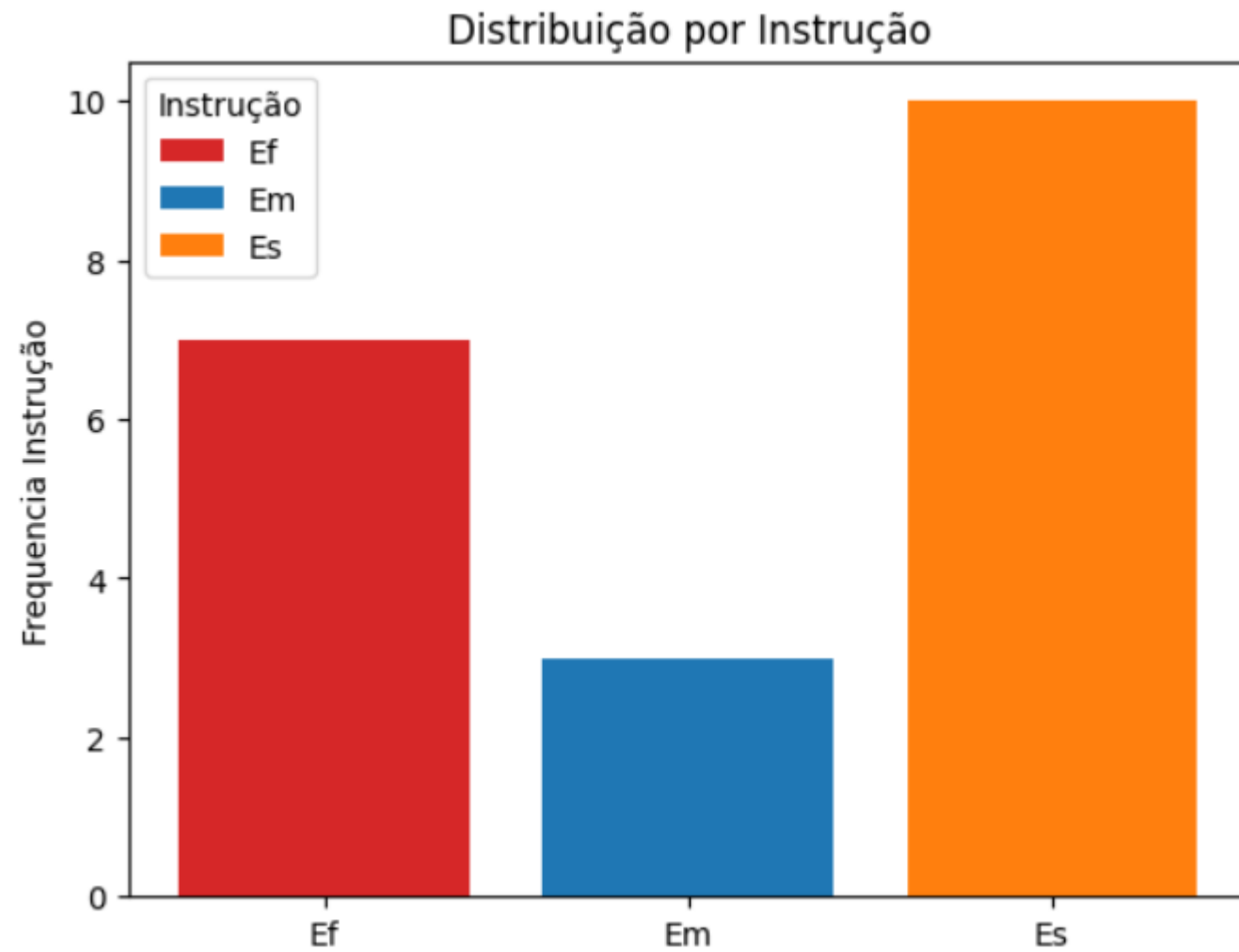
```
# Gráfico
# Criando uma figura e eixos para o gráfico
fig, ax = plt.subplots()


# Definindo as categorias de instrução
Categorias=["Ef","Em","Es"]
# Definindo a contagem de frequência para cada categoria
counts = [7,3,10]
# Definindo rótulos das barras
bar_labels = ['Ef', 'Em', 'Es']
# Definindo cores das barras
bar_colors = ['tab:red', 'tab:blue', 'tab:orange']

# Criando o gráfico de barras
ax.bar(Categorias, counts, label=bar_labels, color=bar_colors)

# Definindo o rótulo do eixo y
ax.set_ylabel('Frequencia Instrução')
# Definindo o título do gráfico
ax.set_title('Distribuição por Instrução')
# Adicionando a legenda ao gráfico
ax.legend(title='Instrução')

# Exibindo o gráfico
plt.show()
```




 DadosOrange



	Class	x	y
0	C1	0.154042	0.660900
1	C1	0.146099	0.655597
2	C1	0.077670	0.714094
3	C1	0.173892	0.742617
4	C1	0.116827	0.734952
...
257	C2	0.309123	0.161230
258	C2	0.256587	0.102027
259	C2	0.283186	0.200190
260	C2	0.309723	0.191331
261	C2	0.312523	0.252244

262 rows × 3 columns

[6] DadosOrange=DadosOrange.sample(frac=1)

 DadosOrange



	Class	x	y
34	C1	0.243057	0.836333
124	C2	0.520553	0.565756
252	C2	0.279452	0.115375
144	C2	0.374506	0.351138
105	C2	0.411299	0.261298
...
232	C2	0.433528	0.132172
151	C2	0.380396	0.284143
261	C2	0.312523	0.252244
82	C2	0.752182	0.610067
93	C2	0.535714	0.373286

262 rows × 3 columns

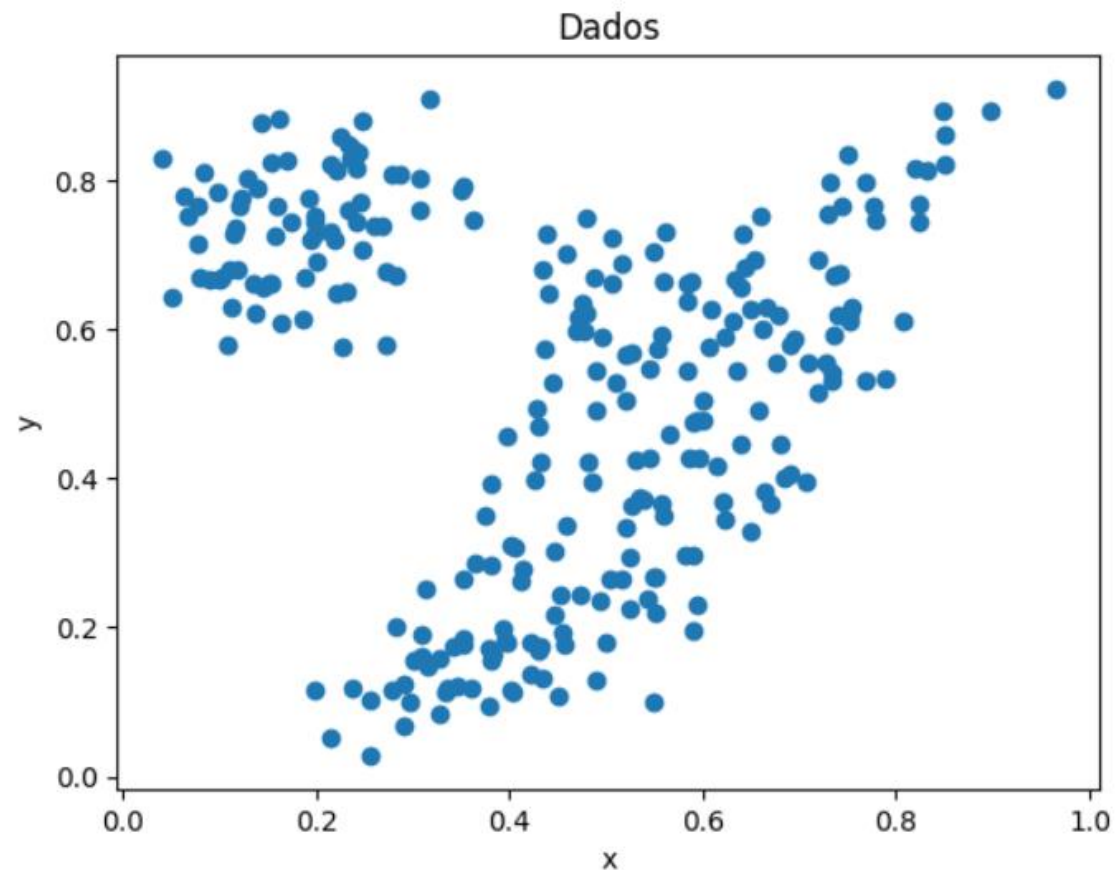
```
# Plota um gráfico de dispersão usando os dados da coluna 'x' como coordenadas x e os dados da coluna 'y' como coordenadas y
plt.scatter(DadosOrange['x'],DadosOrange['y'])

# Define o título do gráfico como 'Dados'
plt.title('Dados')

# Define o rótulo do eixo x como 'x'
plt.xlabel('x')

# Define o rótulo do eixo y como 'y'
plt.ylabel('y')

# Mostra o gráfico
plt.show()
```

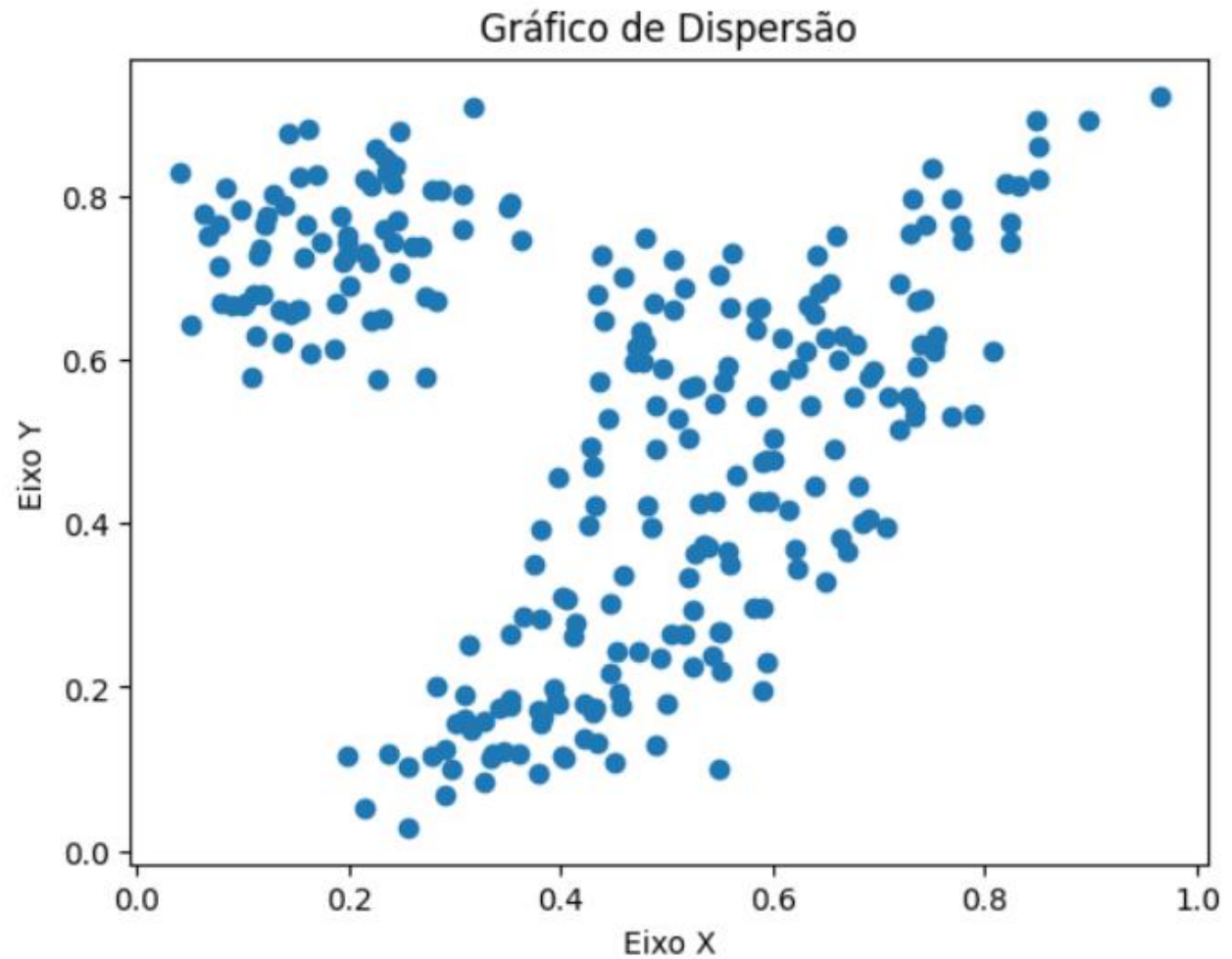


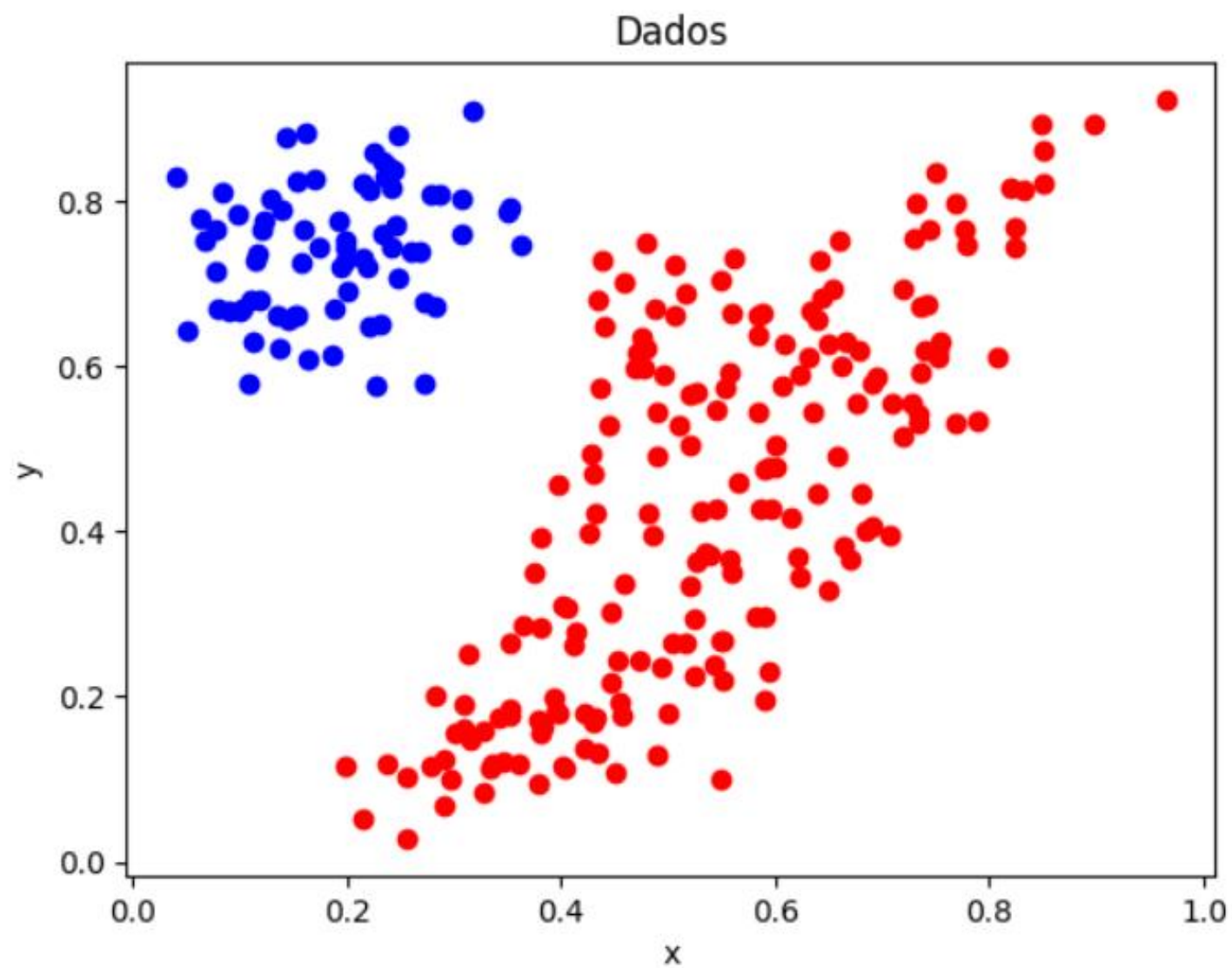
```
# Gráfico
```

```
fig, ax = plt.subplots()  
ax.scatter(DadosOrange['x'],DadosOrange['y'])
```

```
ax.set_ylabel('Eixo Y')  
ax.set_xlabel('Eixo X')  
ax.set_title('Gráfico de Dispersão')
```

```
plt.show()
```





Tipos de Variáveis

Qualitativas

Nominal

- Profissão
- Sexo
- Religião

Ordinal

- Escolaridade
- Estágio da doença
- Classe social

Quantitativas

Discreta

- Nº de filhos
- Nº de acessos à plataforma

Contínua

- Altura
- Peso
- Salário

Uma **medida de tendência central** é um valor no centro ou no meio de um conjunto de dados

Medidas de Tendência Central

MODA:(Mo) Denominamos moda o valor que ocorre com maior frequência em uma série de valores.

Exemplo: Conjunto de valores 1,2,2,5,6,7,7,7,7,20 tem como Moda:

Mo = 7

MÉDIA: A média (M_e) é calculada somando-se todos os valores de um conjunto de dados e dividindo-se pelo número de elementos deste conjunto.

Exemplo: Conjunto de valores 3,2,4,4,6,7,11

$$Me = \frac{3+2+4+4+6+7+11}{7} = 5,28$$

$$Me = \frac{x_1 + x_2 + \dots + x_n}{n}$$

MEDIANA: A Mediana (M_d) representa o valor central de um conjunto de dados. Para encontrar o valor da mediana é necessário colocar os valores em ordem crescente ou decrescente.

Quando o número elementos de um conjunto é par, a mediana é encontrada pela média dos dois valores centrais.

MEDIANA: Quando o número elementos de um conjunto é ímpar, a mediana é o termo de ordem $(n+1)/2$.

Exemplo: Conjunto de valores 1,2,2,5,**6,7**,7,7,7,20 tem como mediana:

Temos dez termo, então a mediana será a média aritmética dos dois termos centrais.

$$Md = \frac{6+7}{2} = 6,5$$

Conjunto de valores 1,2,2,5,**6**,7,7,7,20 tem como mediana:

Temos nove termo, então a mediana será o termo central.

$$Md = 6$$

Ponto médio é o valor que está a meio caminho entre o valor maior e o menor valor. Para obtê-lo, somamos esses valores extremos e dividimos o resultado por 2, como na fórmula a seguir:

$$\text{ponto médio} = \frac{\text{maior valor} + \text{menor valor}}{2}$$

Exemplo: Determinar o ponto médio do conjunto de dados: 10 29 26 15 23 17 25 0 20

$$\text{Ponto médio} = \frac{29 + 0}{2} = 14,5$$

A **média ponderada** é conhecida também como média aritmética simples ponderada.

A média é ponderada quando se atribui peso a cada um dos valores.

O peso faz com que alguns valores tenham mais impactos no cálculo da média.

A fórmula para calcular a média entre os valores $x_1, x_2, x_3, \dots, x_n$ com pesos $p_1, p_2, p_3, \dots, p_n$, respectivamente, é

$$x = \frac{x_1 \cdot p_1 + x_2 \cdot p_2 + x_3 \cdot p_3 + \dots + x_n \cdot p_n}{p_1 + p_2 + p_3 + \dots + p_n}$$

Função	Quantidade	Salário
Auxiliar administrativo	5	R\$ 1100
Atendente	16	R\$ 2000
Gerente	3	R\$ 5500
Diretor	1	R\$ 12.500

$$x = \frac{5 \cdot 1100 + 16 \cdot 2000 + 3 \cdot 5500 + 1 \cdot 12.500}{5 + 16 + 3 + 1}$$

$$x = \frac{5500 + 32.000 + 16.500 + 12.500}{25}$$

$$x = \frac{66.500}{25}$$

$$x = 2660$$

Média de uma distribuição de frequência.

Ponto médio (x)	Frequência (f)	<u>x.f</u>
12,5	6	75
24,5	10	245
36,5	13	474
48,5	8	388
60,5	5	302
72,5	6	435
84,5	2	169
Soma	n = 50	2089

$$\bar{X} = \frac{\sum x.f}{n}$$

x = ponto médio.

f = frequência da classe.

$$\bar{X} = \frac{2089}{50}$$

$$\bar{X} = 41,78$$

TABELA 2-6 Comparação entre Média, Mediana, Moda e Ponto Médio

Medida	Definição	Quão Freqüente?	Existência	Leva em Conta todos os Valores?	Afetada pelos Valores Extremos?	Vantagens e Desvantagens
Média	$\bar{x} = \frac{\sum x}{n}$	“média” mais familiar	existe sempre	sim	sim	usada em todo este livro; funciona bem com muitos métodos estatísticos
Mediana	valor do meio	usada comumente	existe sempre	não	não	costuma ser uma boa escolha se há alguns valores extremos
Moda	valor mais freqüente	usada às vezes	pode não existir; pode haver mais de uma moda	não	não	apropriada para dados ao nível nominal
Ponto médio	$\frac{\text{alto} + \text{baixo}}{2}$	raramente usada	existe sempre	não	sim	muito sensível a valores extremos

Comentários gerais:

- Para um conjunto de dados aproximadamente simétrico com uma moda, a média, a mediana, a moda e o ponto médio tendem a coincidir.
- Para um conjunto de dados obviamente assimétrico, convém levar em conta a média e a mediana.
- A média é relativamente *confiável*; ou seja, quando as amostras são extraídas da mesma população, as médias tendem a ser mais constantes do que outras medidas (constantes no sentido de que as médias amostrais extraídas da mesma população não variam tanto quanto as outras medidas).

Medidas de Dispersão: amplitude, variância e desvio padrão

AMPLITUDE: A amplitude de um conjunto de dados é a diferença entre o valor máximo e mínimo destes valores.

VARIÂNCIA: Variância é uma medida de dispersão e é usada também para expressar o quanto um conjunto de dados se desvia da média.

$$V = \frac{\sum_{i=1}^n (x_i - Me)^2}{n}$$

DESVIO PADRÃO: O desvio padrão de um conjunto de dados é uma medida da variação dos valores em relação a media.

$$\text{Desvio Padrão} = \sqrt{V}$$

Entendo o desvio Padrão

Caso 01

Regra Prática (desvio-padrão em termos da amplitude)

Para conjuntos de dados típicos, a amplitude mede aproximadamente 4 desvios-padrão ($4s$), de forma que podemos aproximar como segue o desvio-padrão:

$$\text{desvio-padrão} \approx \frac{\text{amplitude}}{4} \quad \text{regra prática}$$

Esta expressão dá uma estimativa razoável para o desvio-padrão, quando conhecemos os valores mínimo e máximo. Desde que conheçamos o desvio-padrão, podemos utilizá-lo para entender melhor os dados, fazendo estimativas dos valores mínimo e máximo como se segue:

$$\begin{aligned} \text{mínimo} &\approx (\text{média}) - 2 \times (\text{desvio-padrão}) \\ \text{máximo} &\approx (\text{média}) + 2 \times (\text{desvio-padrão}) \end{aligned}$$

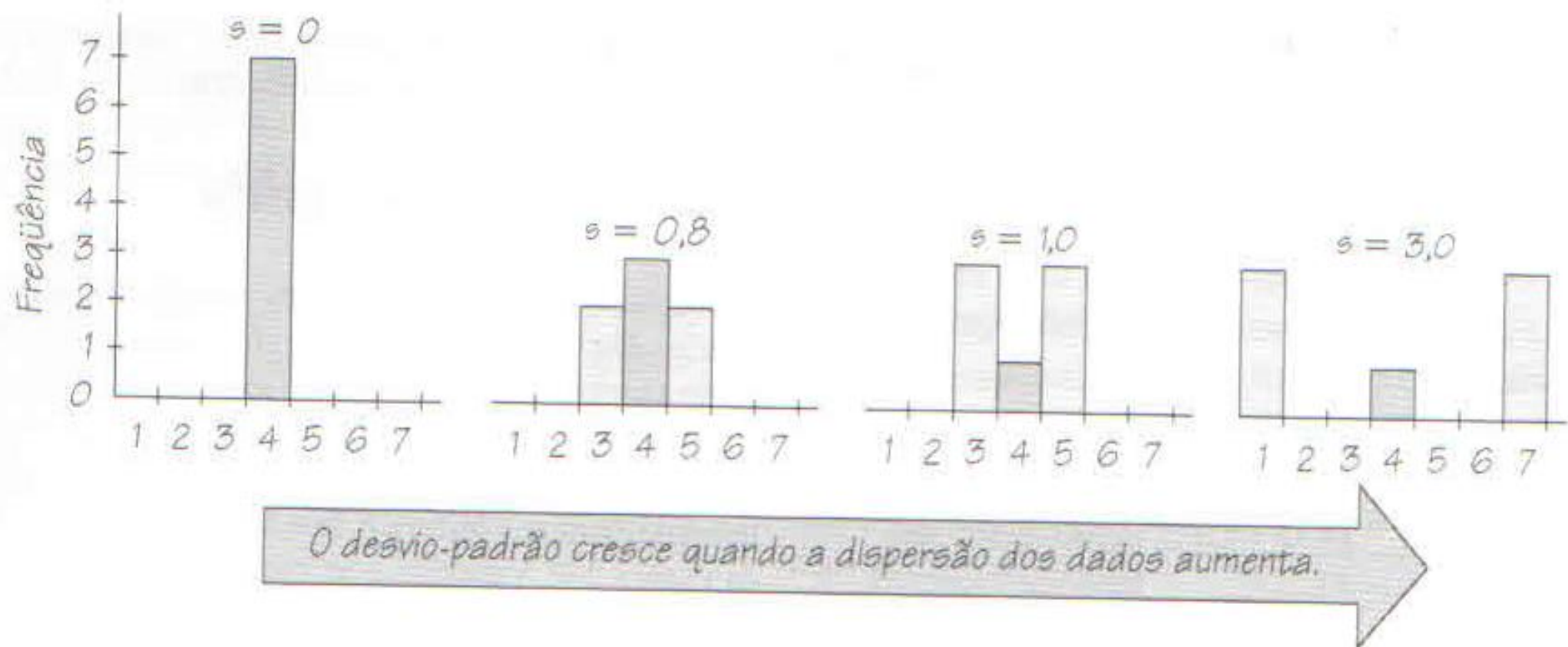


Fig. 2-9 Média idêntica, desvios-padrão diferentes.

Caso 02

Regra Empírica (ou Regra 68-95-99) para os Dados

Outra regra que auxilia a interpretação do valor de um desvio-padrão é a **regra empírica**, aplicável *somente a conjuntos de dados com distribuição aproximadamente em forma de sino*, conforme a Figura 2-10. Essa figura mostra como a média e o desvio-padrão estão relacionados com a proporção dos dados que se enquadram em determinados

limites. Assim é que, com uma distribuição em forma de sino, temos 95% dos seus valores a menos de dois desvios-padrão da média. A regra empírica costuma ser designada abreviadamente como a **regra 68-95-99**.

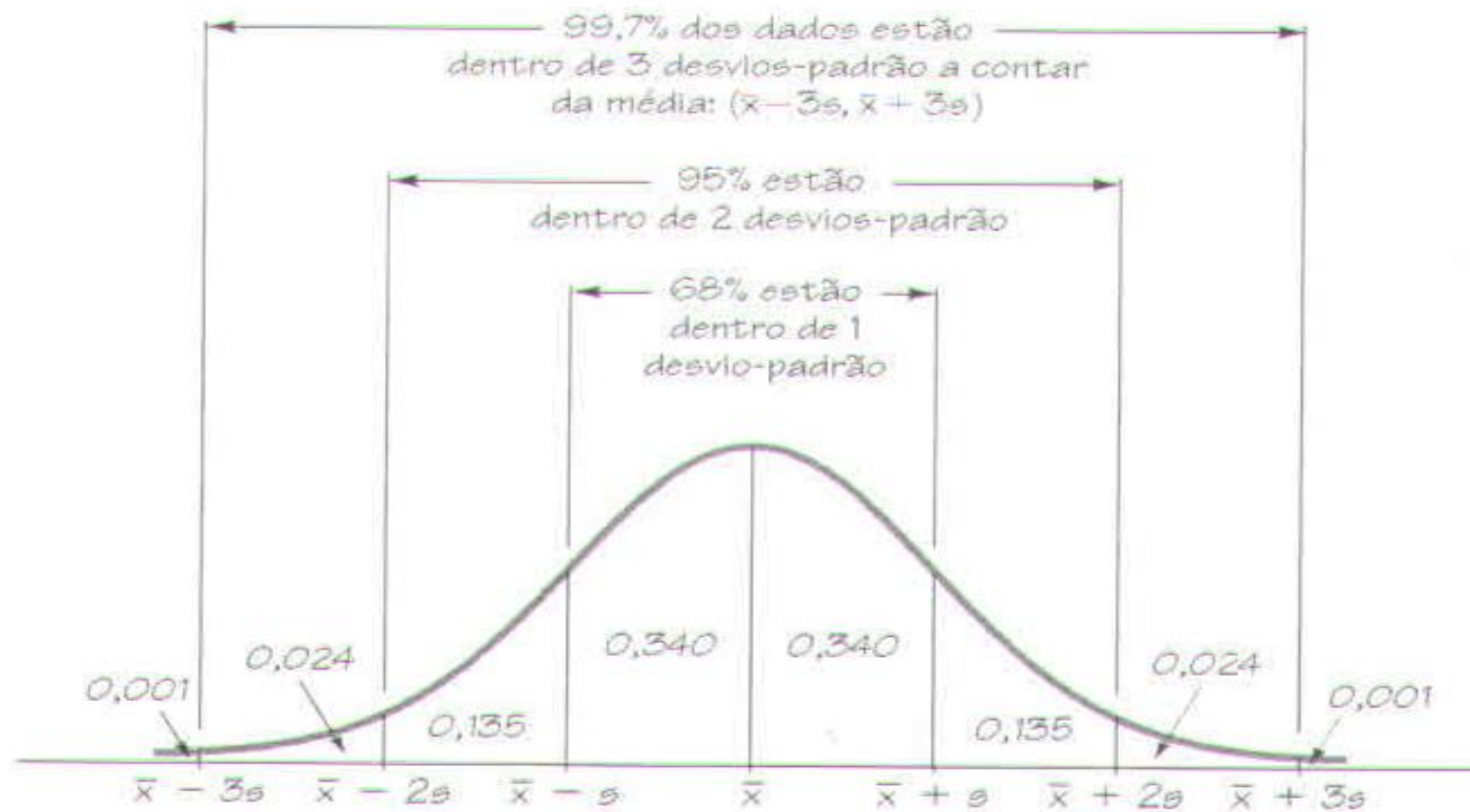


Fig. 2-10 A regra empírica.

A Regra 68-95-99 para Dados com Distribuição em Forma de Sino

- Cerca de 68% dos valores estão a menos de 1 desvio-padrão a contar da média.
- Cerca de 95% dos valores estão a menos de 2 desvios-padrão a contar da média.
- Cerca de 99,7% dos valores estão a menos de 3 desvios-padrão a contar da média.

Teorema de Tchebichev

A proporção (ou fração) de *qualquer* conjunto de dados a menos de K desvios-padrão a contar da média é sempre *ao menos* $1 - 1/K^2$, onde K é um número positivo maior do que 1. Para $K = 2$ e $K = 3$, temos os seguintes resultados específicos:

- Ao menos $3/4$ (ou 75%) de todos os valores estão no intervalo que vai de 2 desvios-padrão abaixo da média a 2 desvios-padrão acima da média ($\bar{x} - 2s$ a $\bar{x} + 2s$).
- Ao menos $8/9$ (ou 89%) de todos os valores estão no intervalo que vai de 3 desvios-padrão abaixo da média até 3 desvios-padrão acima da média ($\bar{x} - 3s$ a $\bar{x} + 3s$).

Curiosidade sobre as medidas de variabilidade

Um Bom Conselho aos Jornalistas

O colunista Max Frankel escreveu no *The New York Times*: "As escolas de jornalismo não dão a devida importância à estatística, e algumas permitem que seus estudantes se formem sem qualquer treinamento com números. Como podem tais repórteres escrever conscientemente sobre comércio, bem-estar social, crime, ou tarifas aéreas, saúde e nutrição? O uso descuidado pela mídia de números sobre a incidência de

acidentes ou de doenças assusta o povo, deixando-o vulnerável aos truques jornalísticos, à demagogia política, e à fraude comercial." O colunista cita diversos casos, inclusive o exemplo de um artigo de página inteira sobre o déficit da cidade de Nova York, com uma promessa do prefeito daquela cidade de cobrir um déficit orçamentário de \$2,7 bilhões; mas em todo o artigo não se menciona uma vez sequer o *total* do orçamento, de modo que a cifra de \$2,7 bilhões por si só pouco significa.

Banco Jefferson Valley (Fila única)	6,5	6,6	6,7	6,8	7,1	7,3	7,4	7,7	7,7	7,7
Banco da Providência (Fila múltipla)	4,2	5,4	5,8	6,2	6,7	7,7	7,7	8,5	9,3	10,0

Os clientes do Jefferson Valley Bank entram em uma fila única que é atendida por três caixas. Os clientes do Bank of Providence podem entrar em qualquer uma de três filas que conduzem a três guichês. Se fizermos o Exercício 5 da Seção 2-4, veremos que ambos os bancos têm a mesma média de 7,15, a mesma mediana de 7,20, a mesma moda de 7,7 e o mesmo ponto médio de 7,10. Com base apenas nestas medidas de tendência central, poderíamos admitir que os tempos de espera nos dois bancos fossem praticamente os mesmos. Todavia, esquadrinhando os tempos de espera originais, constataríamos uma diferença fundamental: O Jefferson Valley Bank tem tempos de espera com muito menos *variação* do que o Bank of Providence. Mantidas todas as outras características, os clientes provavelmente preferirão o Jefferson Valley Bank, onde não correm o risco de entrar em uma fila muito mais lenta do que as outras.

Fazendo uma comparação subjetiva dos tempos de espera nos dois bancos, podemos ver a característica da variação. Passemos agora a algumas formas específicas de *medir* efetivamente a variação. Começaremos com a amplitude.

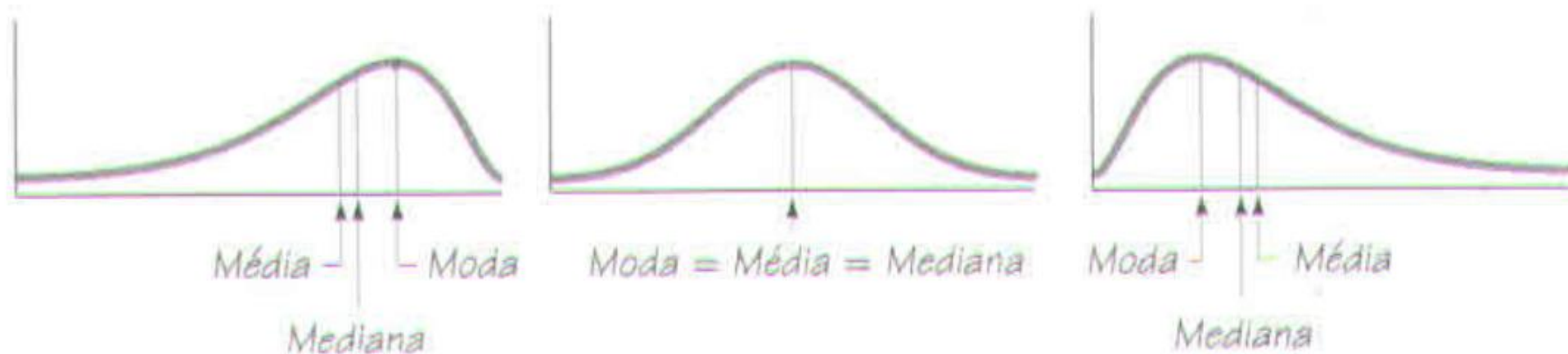
Assimetria

A comparação da média, mediana e moda pode nos dizer algo sobre a característica da assimetria, definida a seguir e ilustrada na Figura 2-8.

DEFINIÇÃO

Uma distribuição de dados é **assimétrica** quando não é simétrica, estendendo-se mais para um lado do que para o outro. (Uma distribuição de dados é **simétrica** quando a metade esquerda do seu histograma é aproximadamente a imagem-espelho da metade direita.)

Os dados assimétricos *para a esquerda* dizem-se **negativamente assimétricos**; a média e a mediana estão à esquerda da moda. Embora nem sempre previsíveis, os dados negativamente assimétricos têm em geral a média à esquerda da mediana. (Veja Figura 2-8(a).) Os dados assimétricos *para a direita* dizem-se **positivamente assimétricos**; a média e a mediana estão à direita



(a) Assimétrica para a esquerda (negativamente assimétrica): A média e a mediana estão à esquerda da moda.

(b) Simétrica (assimetria zero): A média, a mediana e a moda coincidem.

(c) Simétrica para a direita (positivamente assimétrica): A média e a mediana estão à direita da moda.

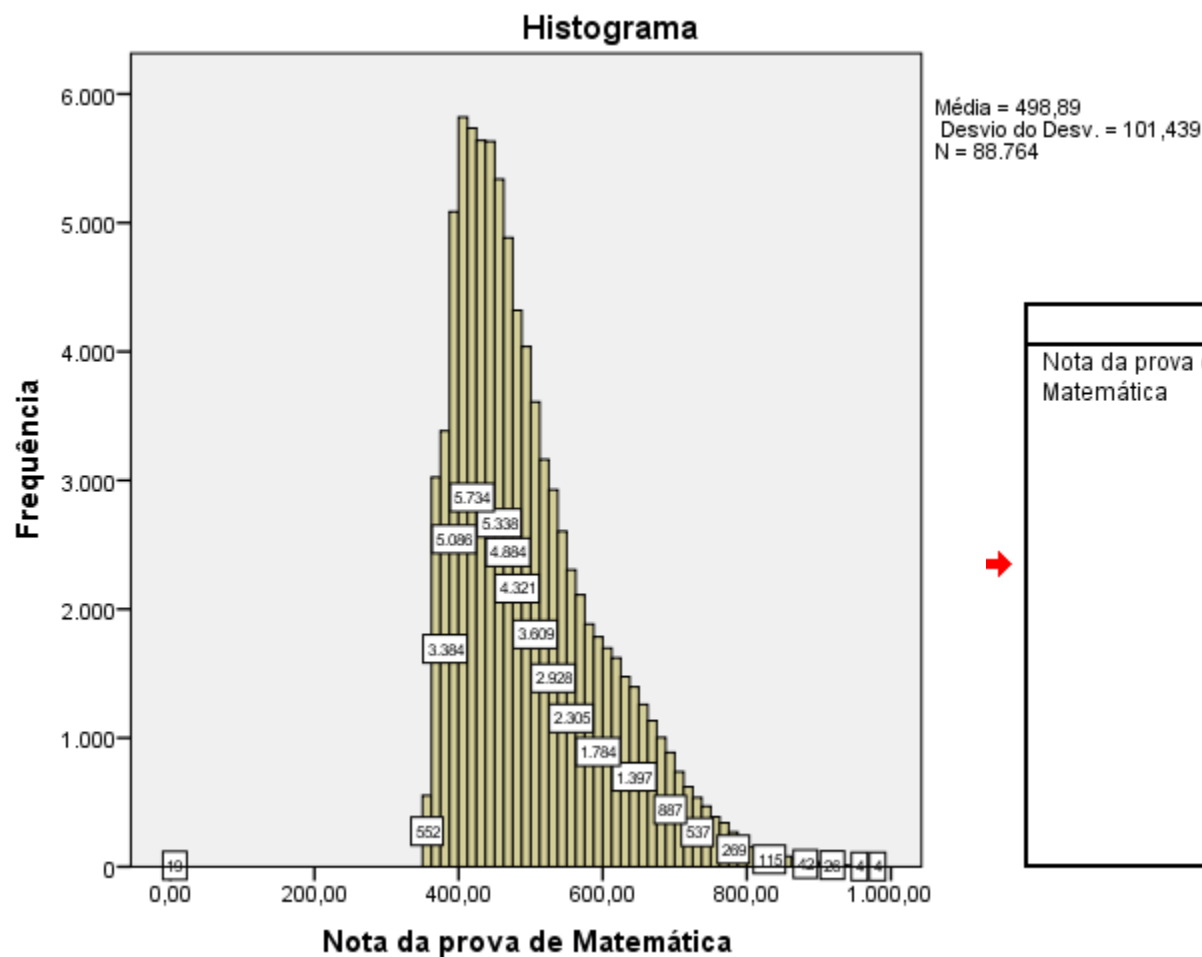
Primeiro Coeficiente de Pearson:

$$AS = \frac{\text{Média} - \text{Moda}}{\text{Desvio Padrão}}$$

$AS = 0$, Média=Moda, Distribuição Simétrica

$AS < 0$, Média<Moda, Distribuição assimétrica à esquerda ou negativa

$AS > 0$, Média>Moda, Distribuição assimétrica à direita ou positiva



Descritivas

			Estatística	Erro Padrão
Nota da prova de Matemática	Média		498.8895	.34047
	95% Intervalo de Confiança para Média	Limite inferior	498.2222	
		Limite superior	499.5568	
	5% da média aparada		491.6880	
	Mediana		473.2000	
	Variância		10289,796	
	Desvio Padrão		101.43863	
	Mínimo		.00	
	Máximo		984.70	
	Amplitude		984.70	
	Amplitude interquartil		132.00	
	Assimetria		1,030	,008
	Curtose		,901	,016

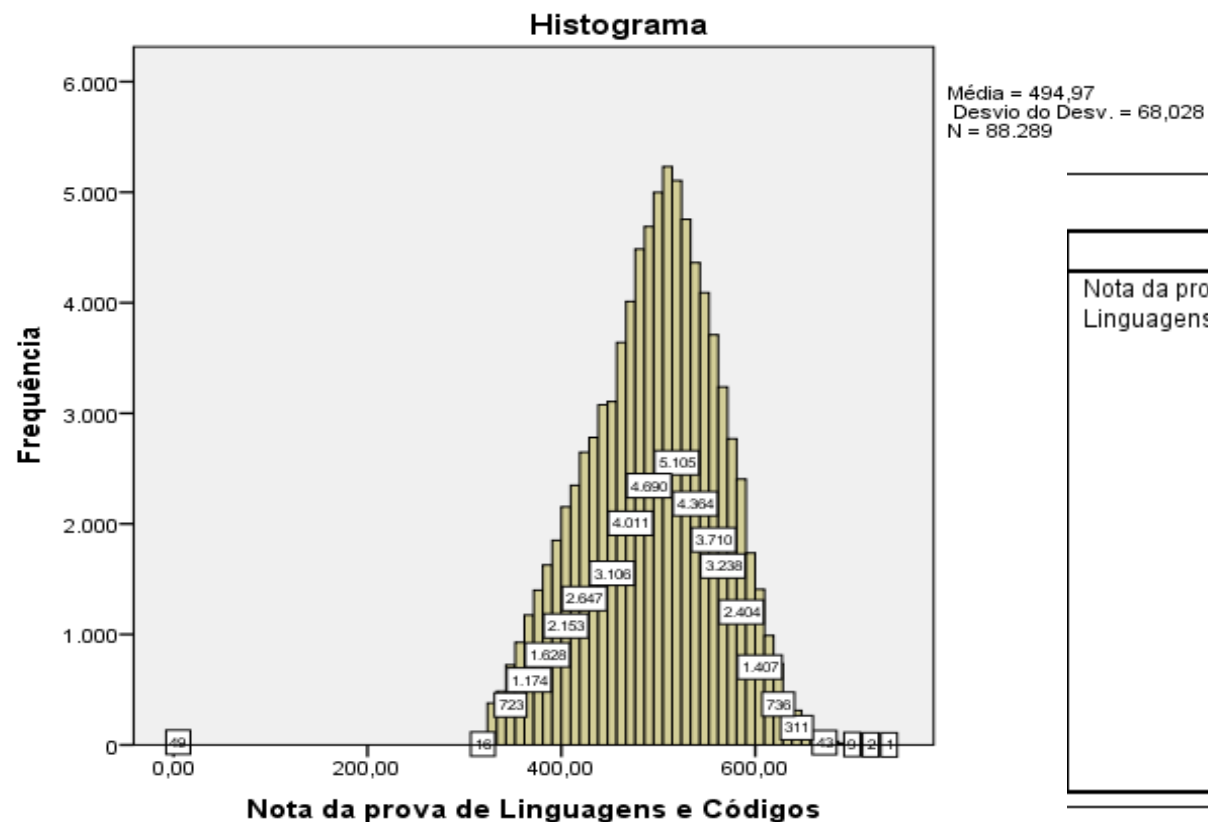
Primeiro Coeficiente de Pearson:

$$AS = \frac{Média - Moda}{Desvio Padrão}$$

AS= 0 , Média=Moda, Distribuição Simétrica

AS<0 , Média<Moda, Distribuição assimétrica à esquerda ou negativa

AS>0, Média>Moda, Distribuição assimétrica à direita ou positiva



Descritivas			
		Estatística	Erro Padrão
Nota da prova de Linguagens e Códigos	Média	492.9083	.22674
	95% Intervalo de Confiança para Média		
	Limite inferior	492.4639	
	Limite superior	493.3527	
	5% da média aparada	493.8472	
	Mediana	498.1000	
	Variância	4736,884	
	Desvio Padrão	68.82503	
	Mínimo	.00	
	Máximo	736.20	
	Amplitude	736.20	
	Amplitude interquartil	95.40	
	Assimetria	-.449	.008
	Curtose	1,506	.016

Primeiro Coeficiente de Pearson:

$$AS = \frac{Média - Moda}{Desvio Padrão}$$

AS= 0 , Média=Moda, Distribuição Simétrica

AS<0 , Média<Moda, Distribuição assimétrica à esquerda ou negativa

AS>0, Média>Moda, Distribuição assimétrica à direita ou positiva

COEFICIENTE DE VARIAÇÃO (CV)

Usado para descrever o nível de variabilidade dentro de uma população, independentemente dos valores absolutos das observações. Define-se como:

$$CV = \frac{s}{\bar{x}}$$

em que s é o desvio-padrão da amostra e \bar{x} é a média da amostra.

Habitualmente, o CV é apresentado sob a forma de percentagem:

$$CV = \frac{s}{\bar{x}} \times 100$$

Assim, uma possível classificação é dada por:

$CV \leq 15\%$	Fraca dispersão
$15\% < CV \leq 30\%$	Moderada dispersão
$CV > 30\%$	Forte dispersão

Coeficiente de Variação de *Pearson*

Seu resultado não é da mesma grandeza da escala (p. ex.: kg, cm, ton etc...), ao contrário, é o resultado que expressa em porcentagem a fração que o desvio-padrão é da média.

Esse coeficiente é dado pela razão entre o desvio-padrão e a média referentes a dados de uma mesma série:

$$CV = \frac{S}{\bar{X}} \cdot 100$$

Tem-se que:

$CV \leq 15\%$, *baixa dispersão (dados homogêneos)*;

$15\% < CV \leq 30\%$, *média dispersão*;

$CV > 30\%$, *alta dispersão (dados muito heterogêneos)*.

As medidas de variabilidade que vimos, anteriormente, somente são comparáveis quando se referem a uma mesma escala de medidas, com a mesma unidade, e, ainda, quando os grupos têm médias não muito diferentes.

Exemplo:

- Não tem sentido comparar a variabilidade de crianças, em altura e peso, usando o desvio-padrão. Isto porque as escalas são de unidades diferentes: altura, em centímetros; peso, em gramas;
- Não tem sentido comparar desvios-padrão de adultos e crianças ou grupos essencialmente diferentes, embora a unidade de escala seja a mesma em um teste de inteligência.

Para esses casos em que são diferentes as medidas em comparação ou os grupos, usa-se o coeficiente de variação.

Usando o Python

```
# Calcular Média, Mediana, Moda, Amplitude, Variância, Desvio Padrão, Coeficiente de Variação
Amplitude=PesquisaNacional['Renda'].max()-PesquisaNacional['Renda'].min()
Variancia=PesquisaNacional['Renda'].var()
DesvioPadrao=PesquisaNacional['Renda'].std()
Media=PesquisaNacional['Renda'].mean()
Moda=PesquisaNacional['Renda'].mode()
Mediana=PesquisaNacional['Renda'].median()
CVRenda=(PesquisaNacional['Renda'].std()/PesquisaNacional['Renda'].mean())*100
print(Amplitude)
print(Variancia)
print(DesvioPadrao)
print(Media)
print(Moda)
print(Mediana)
print(CVRenda)
```

```
#Calculando a média por grupo
```

```
PesquisaNacional['Renda'].groupby(PesquisaNacional['Sexo']).mean()
```

```
Sexo
Feminino    1566.847393
Masculino    2192.441596
Name: Renda, dtype: float64
```

```
#Calculando a média por mais de um grupo
```

```
PesquisaNacional['Renda'].groupby([PesquisaNacional['Cor'],PesquisaNacional['Sexo']]).mean()
```

```
Cor      Sexo
Amarela  Feminino    3027.341880
         Masculino    4758.251064
Branca   Feminino    2109.866750
         Masculino    2925.744435
Indígena Feminino    2464.386139
         Masculino    1081.710938
Parda    Feminino    1176.758516
         Masculino    1659.577425
Preta    Feminino    1134.596400
         Masculino    1603.861687
Name: Renda, dtype: float64
```

```
▶ PesquisaNacional.groupby(['UF'])['Renda'].mean().sort_values()
```

```
➡ UF  
Maranhão          1019.432009  
Piauí             1074.550784  
Sergipe           1109.111111  
Alagoas           1144.552602  
Ceará             1255.403692  
Paraíba           1293.370487  
Rio Grande do Norte 1344.721480  
Pará              1399.076871  
Bahia             1429.645094  
Amazonas          1445.130100  
Acre              1506.091782  
Pernambuco        1527.079319  
Tocantins         1771.094946  
Roraima           1783.588889  
Rondônia          1789.761223  
Amapá             1861.353516  
Goiás             1994.580794  
Espírito Santo    2026.383852  
Minas Gerais      2056.432084  
Mato Grosso       2130.652778  
Mato Grosso do Sul 2262.604167  
Rio Grande do Sul 2315.158336
```

```
#Moda
PesquisaNacional['Renda'].groupby(PesquisaNacional['Cor']).mode()
# Dá erro
# A função mode() não está diretamente disponível como mean() ou median().
# Você precisará aplicar uma função personalizada para calcular a moda em cada grupo.
```

```
# Definir uma função personalizada para calcular a moda
def calcular_moda(group):
    if not group.empty:
        moda=group.mode()
        if not group.empty:
            return moda.iloc[0]
    return None

# Versão mais simples
# def calcular_moda(group):
#     return group.mode().iloc[0] if not group.empty else None

# Aplicar a função de cálculo da moda após agrupar os dados por 'grupo' e aplicar a função personalizada
moda_por_grupo=PesquisaNacional.groupby('UF')['Renda'].apply(calcular_moda)

# .sort_values() combinação

print(moda_por_grupo)
```

	Nota_Matemática	Pontuação	Reserva_vaga	Curso	Semestre
0	6.7	8.22	Ampla Concorrência	Técnico Integrado em Edificações - Campus Fort...	2018.2
1	7.5	7.96	Ampla Concorrência	Técnico Integrado em Edificações - Campus Fort...	2018.2
2	9.8	7.79	Ampla Concorrência	Técnico Integrado em Edificações - Campus Fort...	2018.2
3	8.4	7.52	Ampla Concorrência	Técnico Integrado em Edificações - Campus Fort...	2018.2
4	6.7	7.32	Ampla Concorrência	Técnico Integrado em Edificações - Campus Fort...	2018.2
...
463	6.0	7.00	L6	Técnico Integrado em Telecomunicações - Campus...	2020.1
464	6.3	6.83	L6	Técnico Integrado em Telecomunicações - Campus...	2020.1
465	7.4	6.79	L6	Técnico Integrado em Telecomunicações - Campus...	2020.1
466	8.1	6.33	L8	Técnico Integrado em Telecomunicações - Campus...	2020.1
467	6.6	5.97	L8	Técnico Integrado em Telecomunicações - Campus...	2020.1

468 rows × 5 columns



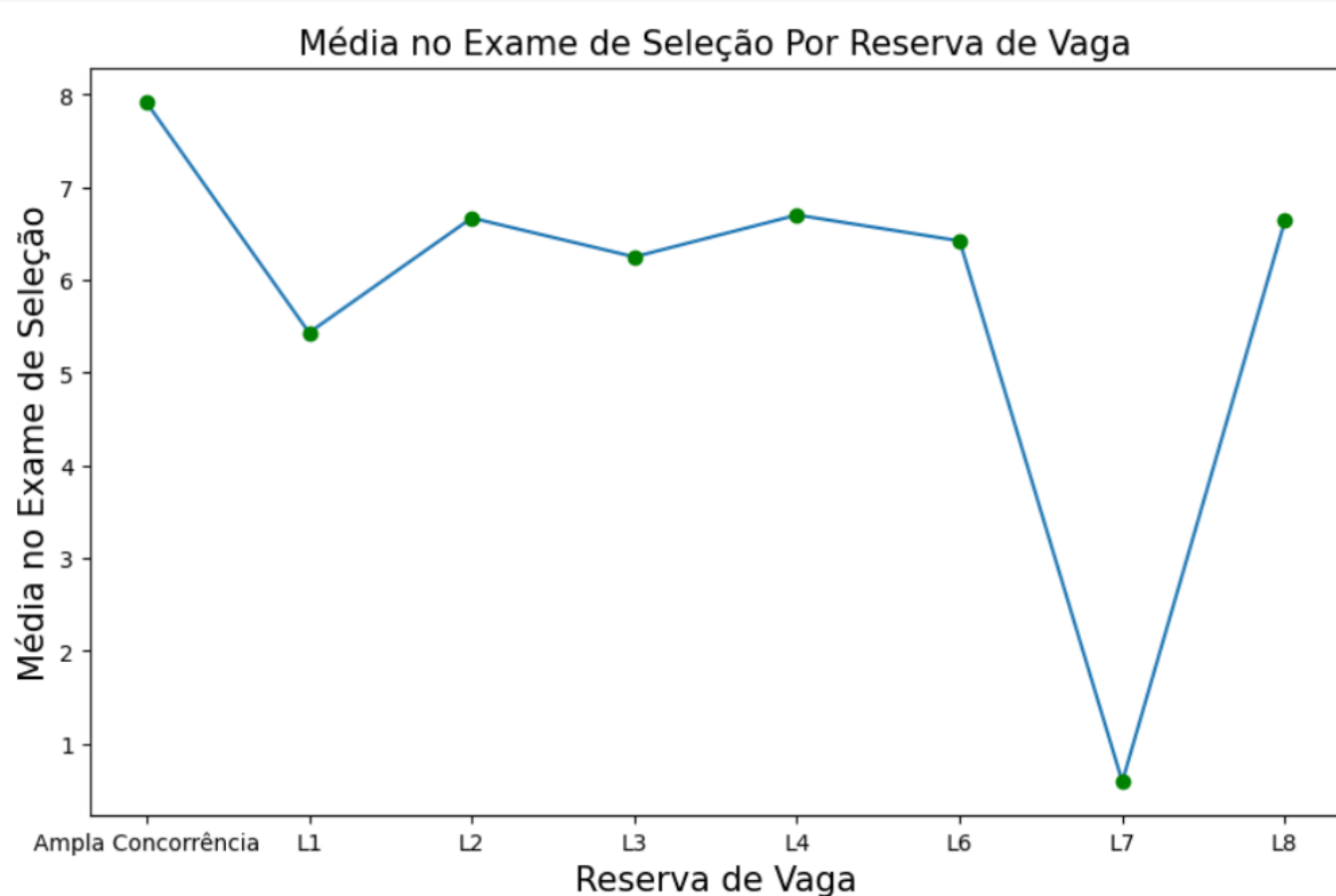
```
#Estatística descritiva da pontuação final no exame de seleção por reserva de vaga.
grup_reserva=Integrado['Pontuação'].groupby(Integrado['Reserva_vaga'])
des=grup_reserva.describe()
des
```

	count	mean	std	min	25%	50%	75%	max
Reserva_vaga								
Ampla Concorrência	273.0	7.916484	0.706946	6.16	7.3700	8.00	8.440	9.50
L1	4.0	5.432500	0.365183	5.16	5.2425	5.30	5.490	5.97
L2	81.0	6.669877	1.040162	3.47	6.0200	6.78	7.470	9.07
L3	1.0	6.250000	NaN	6.25	6.2500	6.25	6.250	6.25
L4	19.0	6.701579	1.604779	1.22	6.0850	6.96	7.680	8.81
L6	75.0	6.422400	1.074192	2.21	5.7300	6.59	7.235	8.76
L7	1.0	0.590000	NaN	0.59	0.5900	0.59	0.590	0.59
L8	14.0	6.639286	0.754631	5.68	5.9825	6.50	7.330	7.76

```
mean_scores = grup_reserva.mean()
mean_scores
```

```
Reserva_vaga
Ampla Concorrência    7.916484
L1                    5.432500
L2                    6.669877
L3                    6.250000
L4                    6.701579
L6                    6.422400
L7                    0.590000
L8                    6.639286
Name: Pontuação, dtype: float64
```

```
# Representação gráfica da pontuação final no exame de seleção por reserva de vaga.  
plt.figure(figsize=(10, 6))  
plt.plot(des['mean'])  
plt.plot(des['mean'],'go')  
plt.xlabel('Reserva de Vaga',fontsize=15)  
plt.ylabel('Média no Exame de Seleção', fontsize=15)  
plt.title('Média no Exame de Seleção Por Reserva de Vaga',fontsize=15 )  
plt.show()
```



```
#Estatística descritiva da média em matemática no S1 por semestre.
grup_curso=Integrado['Nota_Matemática'].groupby(Integrado['Semestre'])
des_sem_mat=grup_curso.describe()
des_sem_mat
```

	count	mean	std	min	25%	50%	75%	max
Semestre								
2018.2	103.0	6.088350	3.037585	0.0	4.700	6.90	8.400	10.0
2019.1	110.0	6.730909	2.461772	0.0	5.825	7.10	8.700	10.0
2019.2	118.0	6.774576	2.662310	0.0	6.000	7.25	8.975	10.0
2020.1	137.0	7.741606	2.133784	0.0	7.300	8.30	9.000	10.0


```
# Define o tamanho da figura do gráfico como 10 polegadas de largura por 6 polegadas de altura
plt.figure(figsize=(10, 6))

# Plota um gráfico de linha da média dos dados contidos na coluna 'mean' do DataFrame 'des_sem_mat'
plt.plot(des_sem_mat['mean'])

# Plota pontos verdes sobre a linha do gráfico de linha para representar os valores individuais da média
plt.plot(des_sem_mat['mean'], 'go')

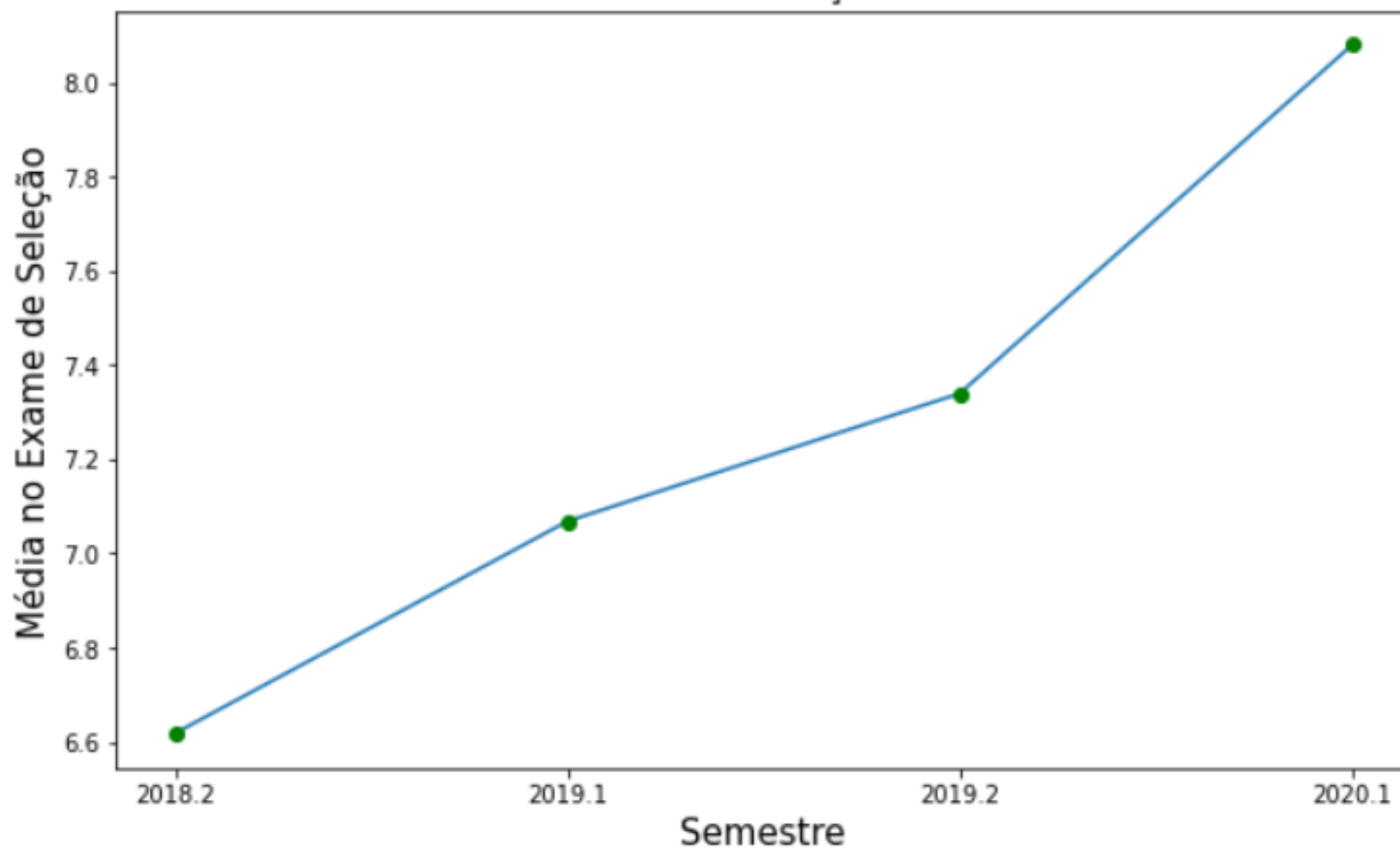
# Define o rótulo do eixo x como 'Semestre' com tamanho de fonte 15
plt.xlabel('Semestre', fontsize=15)

# Define o rótulo do eixo y como 'Média em Matemática' com tamanho de fonte 15
plt.ylabel('Média em Matemática', fontsize=15)

# Define o título do gráfico como 'Média em Matemática Por Semestre' com tamanho de fonte 15
plt.title('Média em Matemática Por Semestre', fontsize=15)

# Mostra o gráfico
plt.show()
```

Média no Exame de Seleção Por Semestre



Exercícios pesquisa outras bibliotecas Estatística para Python e estudar groupby, apply.

Exercício :

Na Base de dados Pesquisa Nacional por Amostra de Domicílios.csv faça:

- a) Calcular média, moda, mediana, amplitude, variância, desvio padrão e CV para variáveis quantitativas**
- b) Construir gráficos adequados para os pares de variáveis**

Exercício para Casa :

Na Base de dados Enem_Fortaleza_2021

(OBS.: Use sua Imaginação/Criatividade e o Bom Senso)

- a) Calcular média, moda, mediana, amplitude, variância, desvio padrão e CV para variáveis quantitativas**
- b) Construir gráficos adequados para os pares de variáveis**



LADE

LABORATÓRIO DE ANÁLISES DE DADOS
EDUCACIONAIS E ESTATÍSTICA APLICADA
— IFCE - CAMPUS FORTALEZA —

OBRIGADO!!!