Introdução à Estatística com Pandas

Organizadoras:
Agostinho, P.; Custódio, A.; Marinho, M.; Guilardi, M.; Guisordi, P.

O que é PyLadies?

PyLadies é um grupo internacional de mentoria com foco em ajudar mais mulheres a tornarem-se participantes ativas e líderes da comunidade Python.

#souPyLadiesSP

PyLadies São Paulo

O PyLadies São Paulo é um capítulo das PyLadies internacional cuja missão é incentivar quaisquer mulheres a aprenderem Python, ensinarem e motivarem outras a fazerem o mesmo.

Grupo de Estudos de Ciência de Dados

O Grupo de Estudos é um grupo de mulheres integrantes do PyLadies São Paulo que se reuniram para estudar Python e Ciência de Dados.

A sua formação ocorreu após a reunião de planejamento da comunidade em julho/2018

Sua missão é incentivar outras mulheres a formarem grupos de estudos de temas que lhe interessem e disponibilizar para a comunidade o que elas estudaram.

Agenda do Dia

Ciência de Dados

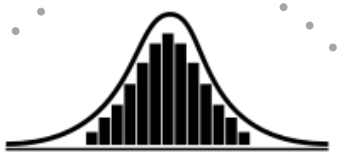
Tipos de dados

Média, Mediana e Moda

Variabilidade, Dispersão e Quartis

A importância dos dados

Pandas



Visualização dos dados

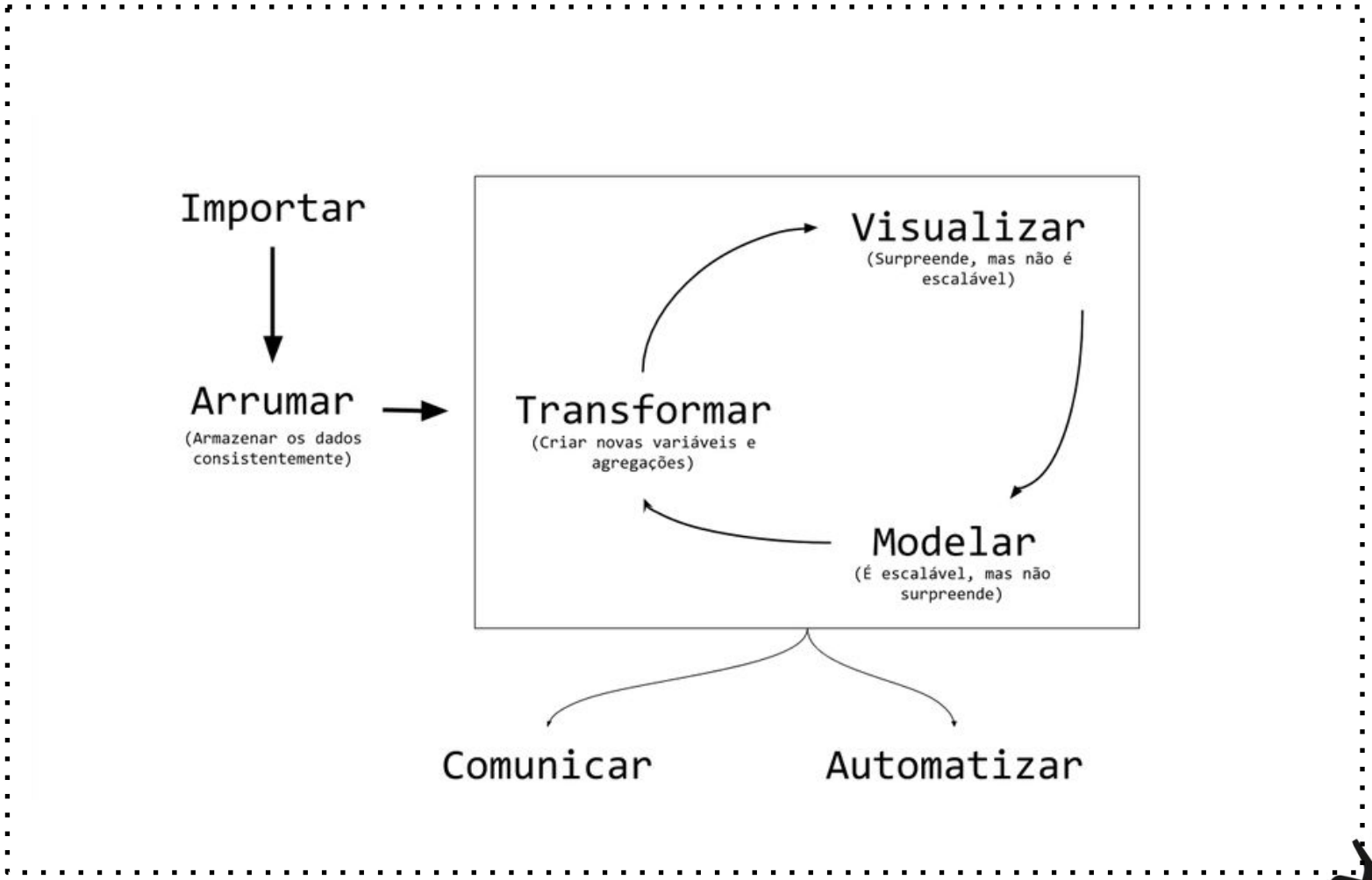
pyladies São Paulo

Ciência de Dados

Ciência de Dados, ou Data Science, é uma ciência interdisciplinar que processa grandes conjuntos de dados, geralmente brutos, usando métodos estatísticos, ou matemáticos, para extrair idéias sobre determinado assunto.

pyladies São Paulo

Ciclo da Ciência de Dados



Fonte: Curso - R (www.cursor.com.br)

pyladies São Paulo

Ferramentas para Ciência de Dados

BANCO DE DADOS

LINGUAGENS

ANÁLISES EXPLORATÓRIAS

AMBIENTES

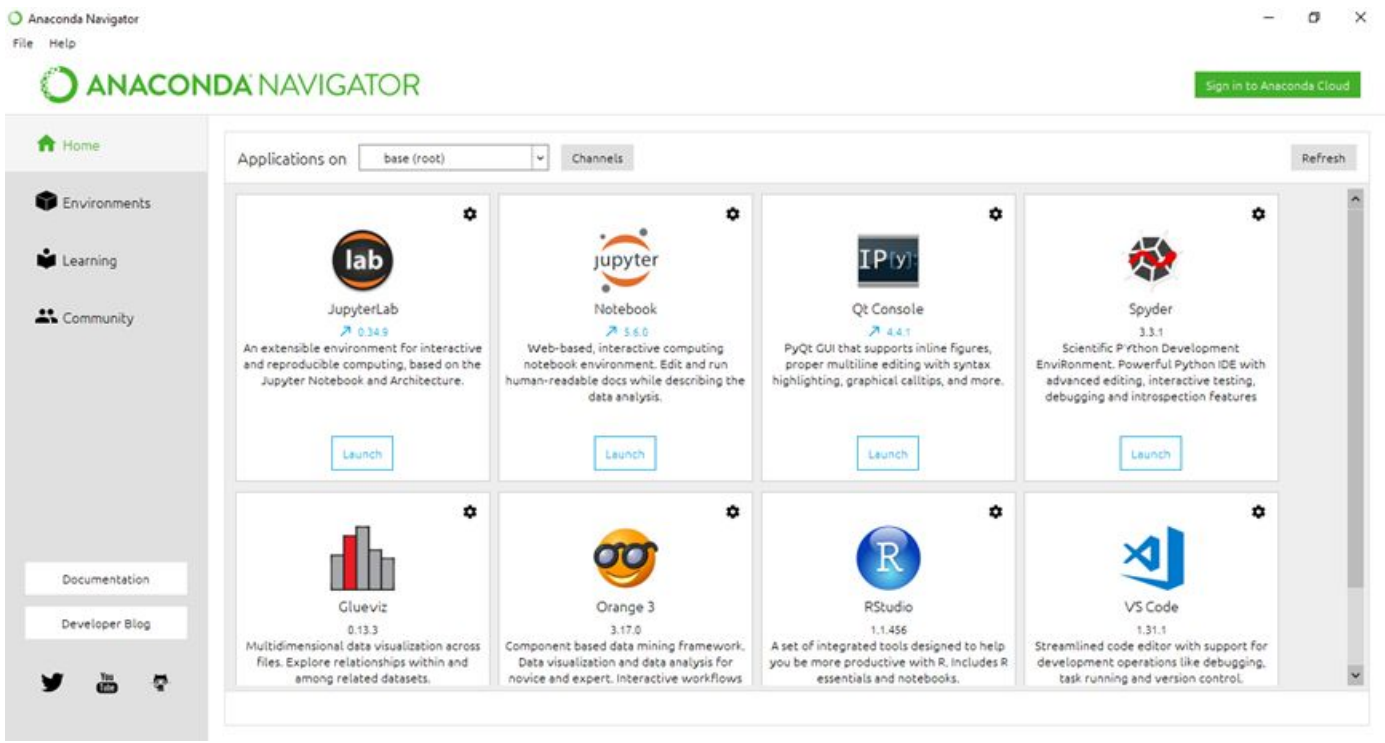
GERENCIADOR DE PACOTES

VISUALIZAÇÃO

pyladies São Paulo

Anaconda

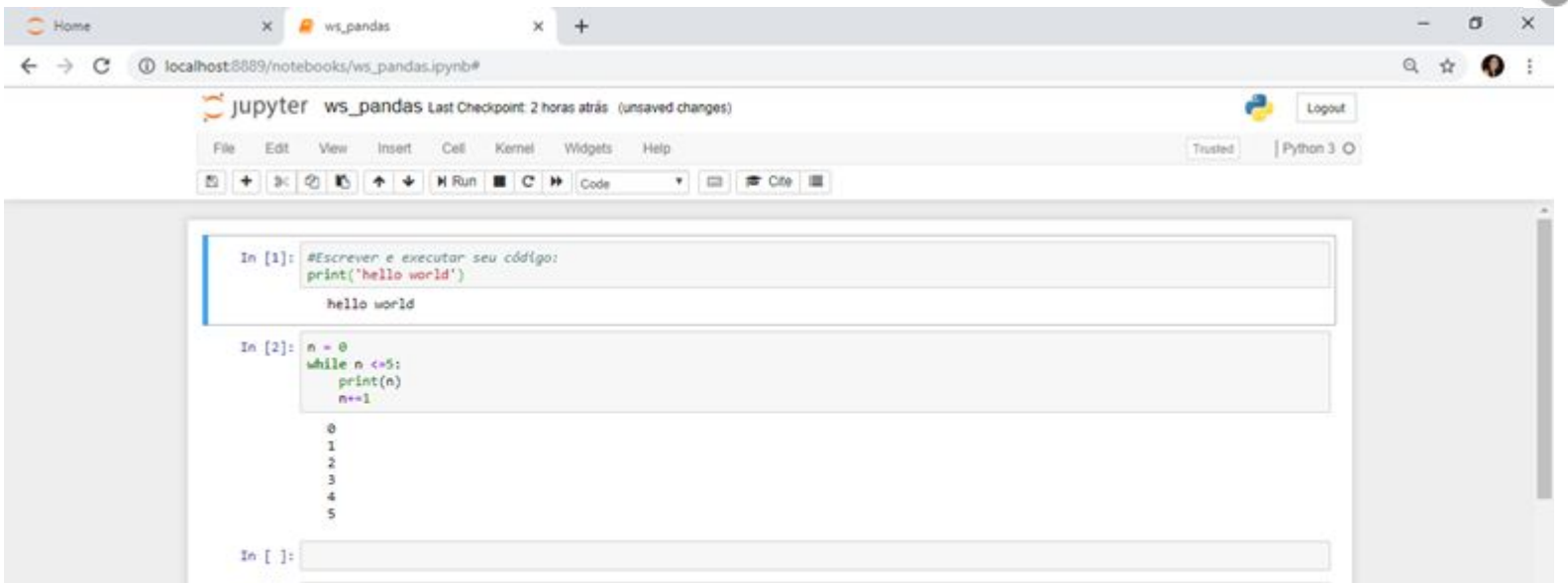
Distribuição gratuita de código aberto das linguagens Python e R. Permite gerenciar pacotes via interface gráfica e linha de comando (conda prompt).



pyladies São Paulo

Jupyter Notebook

Notebooks são documentos que contém códigos, textos, gráficos, links, equações etc. e por conta da quantidade de recursos são muito utilizados em análises exploratórias.



Jupyter Notebook

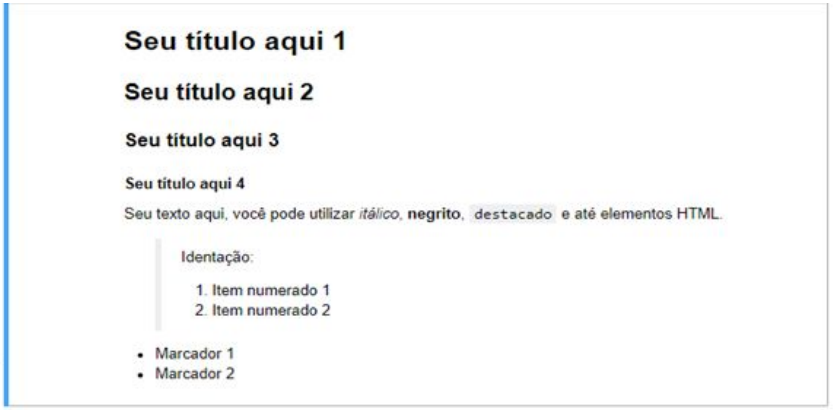
As células configuradas como “Markdown” permite escrever textos simples e formatados.

Modo edição

```
# Seu título aqui 1
## Seu título aqui 2
### Seu título aqui 3
#### Seu título aqui 4
Seu texto aqui, você pode utilizar _itálico_, **negrito**, `destacado` e até elementos HTML.
> Identação:
1. Item numerado 1
2. Item numerado 2

- Marcador 1
- Marcador 2
```

Célula executada



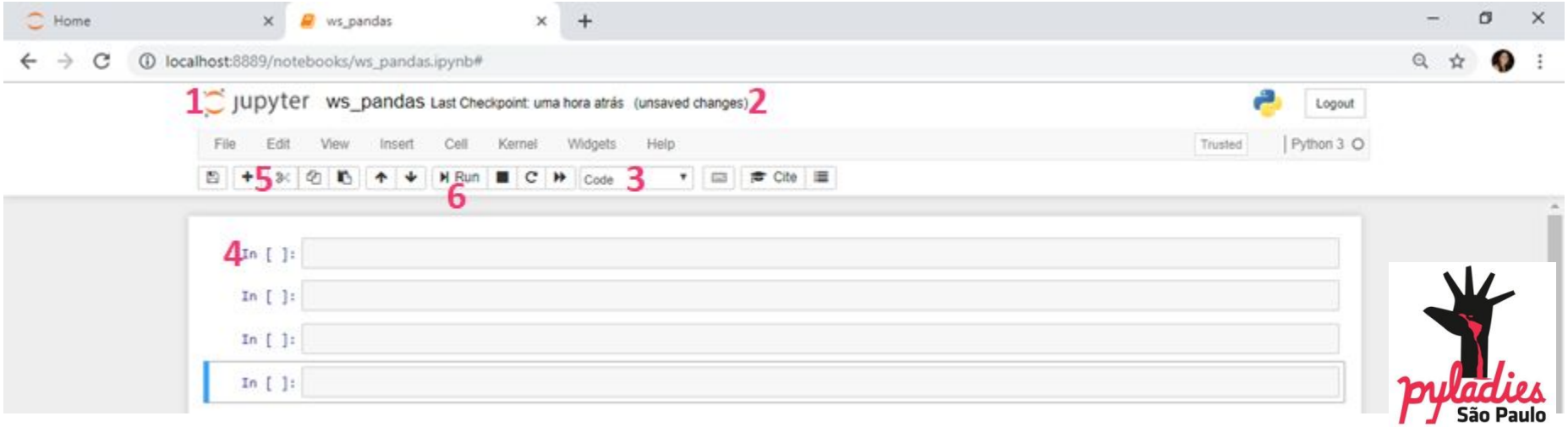

```
<h3> Markdown com HTML </h3>

<p>Estou criando um link para
<a href="https://github.com/angelica93/GEDS">Grupo de estudos Data Science</a>.
</p>
```

pyladies São Paulo

Jupyter Notebook

1. Visualização dos arquivos do diretório atual;
2. Nome do arquivo e quando foi salvo;
3. Configuração da célula, pode ser: code, markdown ou Raw NBConvert;
4. Célula para digitar texto ou código;
5. Adicionar célula abaixo;
6. Executar a célula atual.

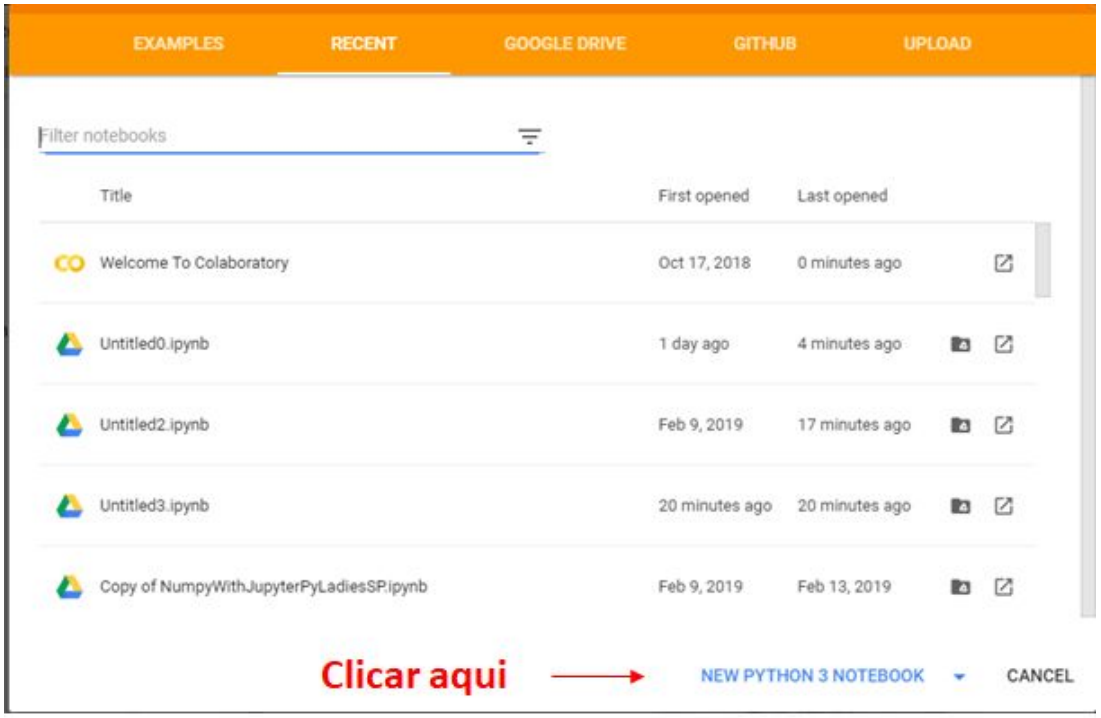


pyladies São Paulo

Google Colaboratory

O Colab segue o mesmo padrão do Jupyter Notebook. Nele é possível adicionar células de código, texto, importar arquivos etc. Para ter acesso ao Google Colab, basta logar na sua conta Google e acessar o link https://colab.research.google.com;

Aparecerá uma tela como essa:



pyladies São Paulo

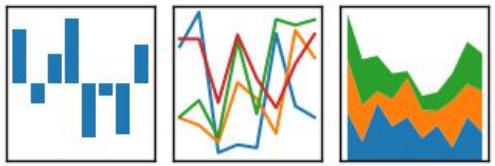
Bibliotecas para Ciência de Dados em Python

pyladies São Paulo

As que utilizaremos hoje

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Pandas é uma biblioteca de código aberto que fornece estruturas de dados de alto desempenho e fáceis de usar e ferramentas de análise de dados para a linguagem de programação Python.

https://pandas.pydata.org/

matplotlib

O Matplotlib é uma biblioteca de plotagem 2D em Python que produz números de qualidade de publicação em uma variedade de formatos impressos e ambientes interativos entre plataformas.

https://matplotlib.org/

pyladies São Paulo

Importância dos Dados

Dados são **importantes** para qualquer organização, seja ela pública ou privada. A partir dos dados é possível fazer planejamento de metas, estratégias e tomadas de decisão.

pyladies São Paulo

Tipos de dados

Quantitativos - Quantifica ou mede

Discretos:
Assumem valores em um conjunto especificado de números.

Contínuos:
Assumem valores em um intervalo contínuo de números.

Qualitativos - Característica ou qualidade

Nominal:
Característica que não possui ordem.

Ordinal:
Característica que possui uma ordem de grandeza.

pyladies São Paulo

Tipos de Dados

Quantitativos

Discretos:
- Quantidade de pessoas na sala?
- Quantidade de dias no mês?

Contínuos:
- Qual sua altura?
- Qual a distância daqui até sua casa?

Qualitativos

Nominal:
- Qual seu Estado? (SP, MG, RJ, Outros)
- Qual gênero você se identifica?

Ordinal:
- Avaliação curso (Ruim a Ótimo).
- Qual seu nível de escolaridade?

pyladies São Paulo

Tipos de Dados - Exemplo PyLadies

ID	Estado Origem	Idade	Escolaridade	Trabalha como Programadora	Renda Mensal
1	SP	36	4	S	3737,52
2	SP	25	2	N	400,00
3	MG	34	3	S	2366,14
4	RJ	23	3	S	2841,29
5	SP	31	4	N	800
6	SP	34	5	S	3433,02
7	SP	39	5	S	2752,74
8	PE	24	3	S	3682,33
9	RJ	29	3	S	2359,28
10	SP	27	3	S	2119,15
11	SP	30	3	S	3326,79
12	SP	25	4	S	2684,05
13	SP	23	2	S	3507,84
14	SP	16	1	N	0
15	SP	36	4	N	800

Legenda Escolaridade

1	Ensino Medio Completo
2	Graduanda
3	Graduação Completa
4	Pós graduanda
5	Pós graduação completa

- Dados meramente ilustrativos.

pyladies São Paulo

Tipos de Dados - Dados Qualitativos

ID	Estado Origem	Idade	Escolaridade	Trabalha como Programadora	Renda Mensal
1	SP	36	4	S	3737,52
2	SP	25	2	N	400,00
3	MG	34	3	S	2366,14
4	RJ	31	4	S	2841,29
5	SP	31	4	N	800
6	SP	34	5	S	3433,02
7	SP	39	5	S	2752,74
8	PE	24	3	S	3682,33
9	RJ	29	3	S	2359,28
10	SP	27	3	S	2119,15
11	SP	30	3	S	3326,79
12	SP	25	4	S	2684,05
13	SP	23	2	S	3507,84
14	SP	16	1	N	0
15	SP	36	4	N	800

Dados Qualitativos

Legenda Escolaridade

1	Ensino Medio Completo
2	Graduanda
3	Graduação Completa
4	Pós graduanda
5	Pós graduação completa

pyladies São Paulo

Tipos de Dados - Categóricos

ID	Estado Origem	Idade	Escolaridade	Trabalha como Programadora	Renda Mensal
1	SP	36	4	S	3737,52
2	SP	25	2	N	400,00
3	MG	34	3	S	2366,14
4	SP	23	4	S	2841,29
5	SP	31	4	N	800
6	SP	34	5	S	3433,02
7	SP	39	5	S	2752,74
8	PE	24	3	S	3682,33
9	RJ	29	3	S	2359,28
10	SP	27	3	S	2119,15
11	SP	30	3	S	3326,79
12	SP	25	4	S	2684,05
13	SP	23	2	S	3507,84
14	SP	16	1	N	0
15	SP	36	4	N	800

Nominal **Ordinal** **Nominal**

Legenda Escolaridade

1	Ensino Medio Completo
2	Graduanda
3	Graduação Completa
4	Pós graduanda
5	Pós graduação completa

pyladies São Paulo

Tipos de Dados - Dados Quantitativos

ID	Estado Origem	Idade	Escolaridade	Trabalha como Programadora	Renda Mensal
1	SP	36	4	S	3737,52
2	SP	25	2	N	400,00
3	MG	34	3	S	2366,14
4	RJ	23			1,29
5	SP	31			00
6	SP	34	5	S	3433,02
7	SP	39	5	S	2752,74
8	PE	24	3	S	3682,33
9	RJ	29	3	S	2359,28
10	SP	27	3	S	2119,15
11	SP	30	3	S	3326,79
12	SP	25	4	S	2684,05
13	SP	23	2	S	3507,84
14	SP	16	1	N	0
15	SP	36	4	N	800

Dados Quantitativos

Legenda Escolaridade

1	Ensino Medio Completo
2	Graduanda
3	Graduação Completa
4	Pós graduanda
5	Pós graduação completa

pyladies São Paulo

Tipos de Dados - Numéricos

ID	Estado Origem	Idade	Escolaridade	Trabalha como Programadora	Renda Mensal
1	SP	36	4	S	3737,52
2	SP	25	2	N	400,00
3	MG	34	3	S	2366,14
4	RJ	23	3	S	1220,10
5	SP	31	4	N	800
6	SP	34	5	S	3433,02
7	SP	39	5	S	2752,74
8	PE	24	3	S	3682,33
9	RJ	29	3	S	2359,28
10	SP	27	3	S	2119,15
11	SP	30	3	S	3326,79
12	SP	25	4	S	2684,05
13	SP	23	2	S	3507,84
14	SP	16	1	N	0
15	SP	36	4	N	800

Discreto

Contínuo

Legenda Escolaridade

1	Ensino Medio Completo
2	Graduanda
3	Graduação Completa
4	Pós graduanda
5	Pós graduação completa

pyladies São Paulo

Trabalhando com Pandas

O acrônimo Pandas vem da combinação de *Panel Data* e *Python Data Analysis**.

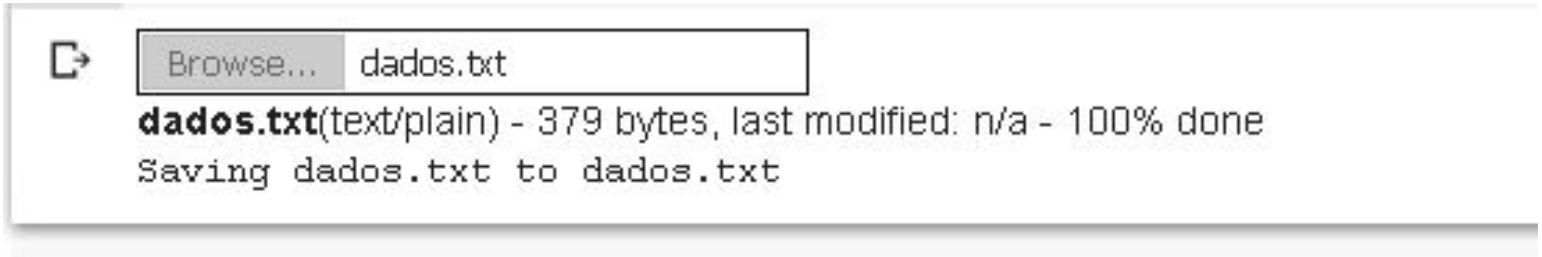
- Dados de Painel - Python para Análise de Dados

pyladies São Paulo

Primeiro Passo:

1. Abrir as bibliotecas que você utilizará
2. Subir o arquivo que possui seus dados

```
[ ]  #abrindo as bibliotecas que serão utilizadas
     import pandas as pd
     import matplotlib.pyplot as plt
     from google.colab import files
     uploaded = files.upload()
```



pyladies São Paulo

Abrindo o arquivo:

Importando um CSV para o Colab:

CSV - Comma-Separated Values

vírgula separando valor

pd.read_csv('nome_arquivo', sep = ';', decimal = ',')

```
# Transformando o arquivo importando em um dataframe
  dados_pyladies = pd.read_csv('dados.txt', sep=';', decimal = ',')
```

Argumentos	
separador	;
decimal	,

pyladies São Paulo

Visualizando o arquivo:

nome_dataframe.head()

```
#para ver as cinco primeiras linhas
dados_pyladies.head()
```

	Estado Origem	Idade	Escolaridade	Trabalha_como_Programadora	Renda_Mensal
0	SP	36	4	S	3737.52
1	SP	25	2	N	400.00
2	MG	34	3	S	2366.14
3	RJ	23	3	S	2841.29
4	SP	31	4	N	800.00

pyladies São Paulo

Visualizando o arquivo:

nome_dataframe.tail()

```
[22] # para ver as cinco últimas linhas
     dados_pyladies.tail()
```

	Estado Origem	Idade	Escolaridade	Trabalha_como_Programadora	Renda_Mensal
10	SP	30	3	S	3326.79
11	SP	25	4	S	2684.05
12	SP	23	2	S	3507.84
13	SP	16	1	N	0.00
14	SP	36	4	N	800.00

pyladies São Paulo

As Estruturas dos Dados:

O que é um dataframe??

DataFrame é uma estrutura de dados bidimensional - parecida com uma tabela de excel ou um banco de dados.

pyladies São Paulo

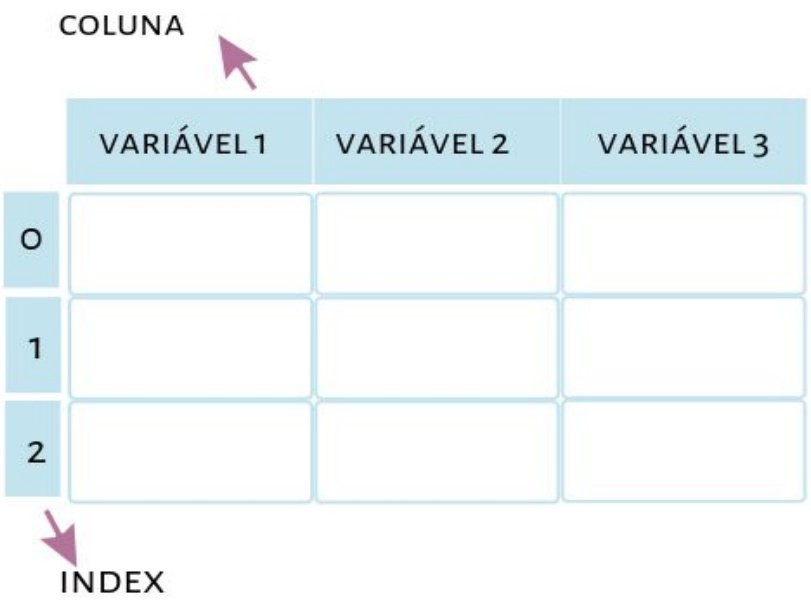
As Estruturas dos Dados:

Estrutura de dados bidimensional (colunas e linhas) cujo índice começa no **zero.**

O dataframe contém colunas que armazenam diferentes tipos de informações (string, float, integer e etc)

Ele é uma classe de objeto da biblioteca Pandas.

dataframe



pyladies São Paulo

As Estruturas dos Dados:

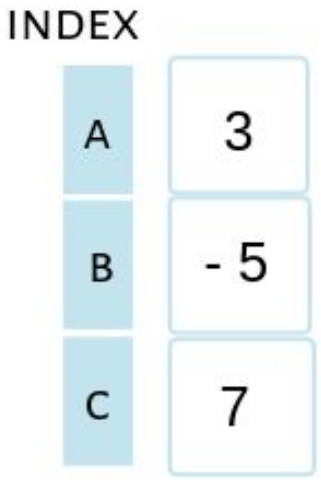
E o series ??

DataSerie é estrutura unidimensional - como uma coluna do excel

pyladies São Paulo

As Estrutura dos Dados

Series



Um array unidimensional e rotulado capaz de armazenar qualquer tipo de dado.

```
s = pd.Series([3,-5,7,4], index = ['a','b','c','d'])
print(s)
```

```
a    3
b   -5
c    7
d    4
dtype: int64
```

pyladies São Paulo

As Estrutura dos Dados:

Linhas e Colunas

nome_dataframe.shape

```
dados_pyladies.shape
```

```
(15, 5)
```

Variáveis (colunas)

nome_dataframe.columns

```
dados_pyladies.columns
```

```
Index(['Estado Origem ', 'Idade', 'Escolaridade', 'Trabalha_como_Programadora',
       'Renda_Mensal'],
      dtype='object')
```

pyladies São Paulo

Conhecendo os Dados:

Informações Gerais

nome_dataframe.info()

```
dados_pyladies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15 entries, 0 to 14
Data columns (total 5 columns):
Estado Origem                  15 non-null object
Idade                          15 non-null int64
Escolaridade                   15 non-null int64
Trabalha_como_Programadora     15 non-null object
Renda_Mensal                   15 non-null float64
dtypes: float64(1), int64(2), object(2)
memory usage: 680.0+ bytes
```

pyladies São Paulo

Selecionando uma Variável (coluna):

nome_dataframe['coluna']

```
dados_pyladies['Estado Origem ']
```

```
0     SP
1     SP
2     MG
3     RJ
4     SP
5     SP
6     SP
7     PE
8     RJ
9     SP
10    SP
11    SP
12    SP
13    SP
14    SP
Name: Estado Origem , dtype: object
```

Lembrando que uma coluna de dataframe é uma series.

pyladies São Paulo

Filtrando um dataframe:

nome_dataframe[nome_dataframe['coluna'] == condição]

```
dados_pyladies[dados_pyladies['Trabalha_como_Programadora'] == 'S']
```

	ID	Estado Origem	Idade	Escolaridade	Trabalha_como_Programadora	Renda_Mensal
0	1	SP	36	4	S	3737,52
2	3	MG	34	3	S	2366,14
3	4	RJ	23	3	S	2841,29
5	6	SP	34	5	S	3433,02
6	7	SP	39	5	S	2752,74
7	8	PE	24	3	S	3682,33
8	9	RJ	29	3	S	2359,28
9	10	SP	27	3	S	2119,15
10	11	SP	30	3	S	3326,79
11	12	SP	25	4	S	2684,05
12	13	SP	23	2	S	3507,84

Aqui você insere a condição para o filtro que você quer. Se a condição for um texto, não se esqueça das aspas!

Aqui você coloca o operador lógico que atende o filtro que você precisa.

pyladies São Paulo

Dataset Tips

O dataset **Tips** tem informações sobre os clientes de restaurantes, valores pagos, dia das refeições, entre outras informações.

Ele está disponível em inglês na biblioteca Seaborn. O Grupo de Estudos de Ciência de Dados das PyLadies São Paulo tratou o dataset (tradução e inserção da coluna tempo_permanencia).

O dataset Tips será a sua base de dados de trabalho nesse workshop.

Colunas

1. **total conta:** valor gasto na refeição. Variável Numérica contínua.
2. **gorjeta:** valor dado como gorjeta. Variável Númerica contínua.
3. **genero:** feminino ou masculino. Variável Categórica.
4. **fumante:** se fuma ou não. Variável Categórica.
5. **dia:** dia da semana. Variável Categórica.
6. **pessoas_mesa:** quantas pessoas havia em cada mesa. Variável Numérica discreta.
7. **tempo_permanencia:** o tempo que as pessoas ficaram no restaurante. Variável Numérica discreta.

pyladies São Paulo

Agora é com você!

Com um novo notebook aberto é hora de começar a trabalhar. Execute as linhas abaixo e quando abrir a janela de seleção de arquivos, escolha o arquivo a ser importado.

https://colab.research.google.com;

```
# Abrindo as bibliotecas que serão utilizadas
import matplotlib.pyplot as plt
import pandas as pd
from google.colab import files
uploaded = files.upload()
```

Arquivo importado com sucesso.

```
# Transformando o arquivo importado em um dataframe
tips_data = pd.read_csv('tips.csv')
```

Célula executada sem erros.

pyladies São Paulo

Desafio Conhecendo os Dados

Agora é a hora de explorar o nosso dataset. Cada um dos exercícios podem ser resolvidos com apenas uma instrução do Pandas.

1. Mostre as **7 primeiras linhas** do dataset que você importou.

2. Mostre as **9 últimas linhas** do dataset.

3. **Quantas colunas e linhas** tem o dataframe?

4. Quais são os **nomes das colunas**?

5. **Mostre de uma vez só**: quantas linhas e colunas tem o dataset, além do nome das colunas, o tipo delas e se elas apresentam valores nulos.

6. **Liste**: a coluna **genero**, depois somente a coluna **gorjeta**. Por último, liste as **duas colunas**: gorjeta e genero.

7. **Filtre** o dataframe com as linhas das **clientes mulheres**.

pyladies São Paulo

Resposta Desafio Conhecendo os Dados

1. Mostre as **7 primeiras linhas** do dataset que você importou.

   tips_data.head(7)

2. Mostre as **9 últimas linhas** do dataset.

   tips_data.tail(9)

3. **Quantas colunas e linhas** tem o dataframe?

   tips_data.shape
   O dataframe tem 244 linhas e 8 colunas.

4. Quais são os **nomes das colunas**?

   tips_data.columns

pyladies São Paulo

Resposta Desafio Conhecendo os Dados

5. **Mostre de uma vez só**: quantas linhas e colunas tem o dataset, além do nome das colunas, o tipo delas e se elas apresentam valores nulos.

   tips_data.info()

6. **Liste**: a coluna **genero**, depois somente a coluna **gorjeta**. Por último, liste as **duas colunas**: gorjeta e genero.

   tips_data['genero']

   tips_data['gorjeta']

   tips_data[['genero', 'gorjeta']]

7. **Filtre** o dataframe com as linhas das **clientes mulheres**.

   tips_data[tips_data['genero'] == 'Feminino']

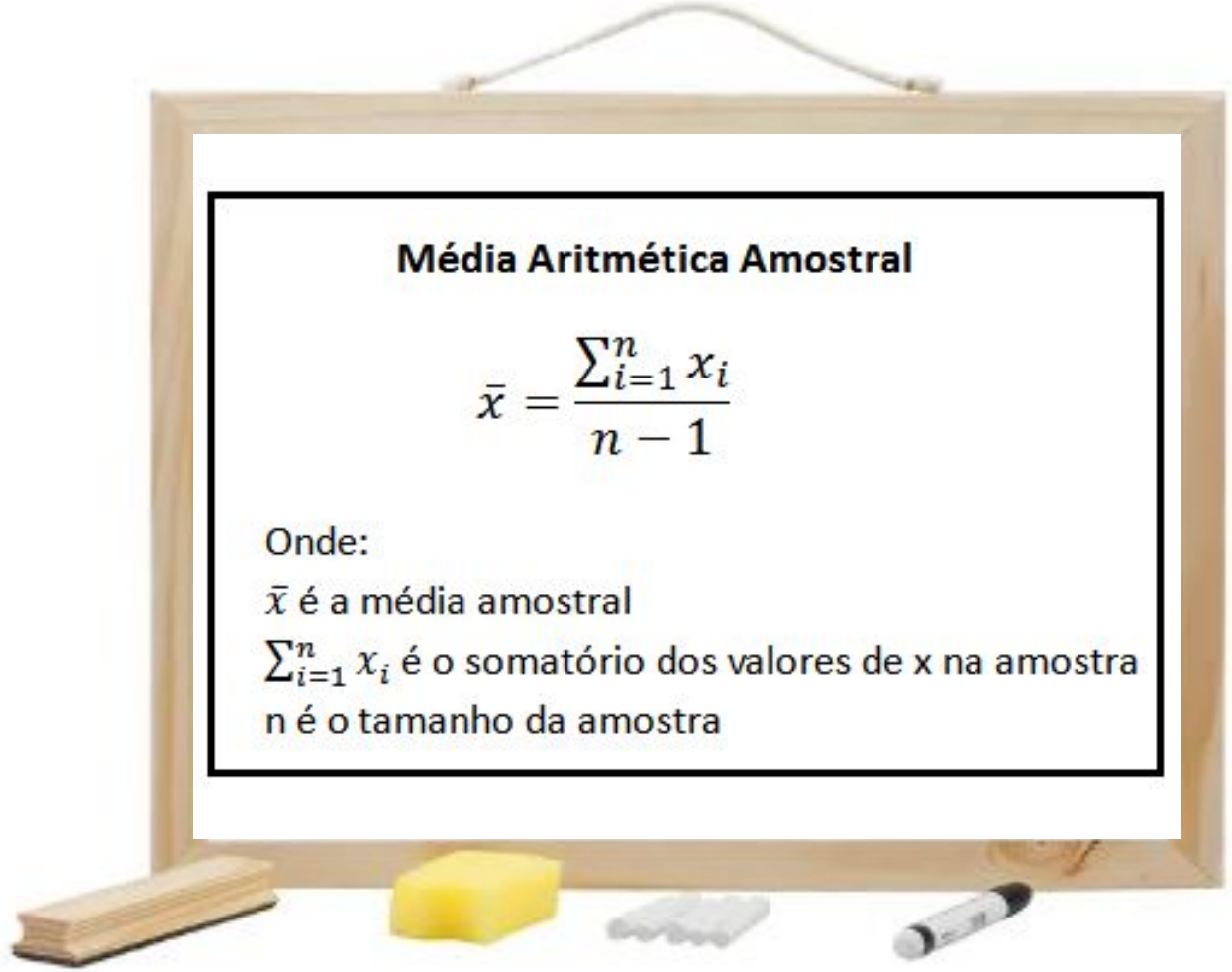
pyladies São Paulo

Estatística

pyladies São Paulo

Média

A **média** é a soma de todos os elementos dividido pelo número de elementos.



pyladies São Paulo

Média

Qual a renda média das meninas no dataframe dados_pyladies?

Para responder essa pergunta precisamos:

1º) Somar a renda mensal de todas as meninas;

```
valor = (3737.52 + 400.00 +  2366.14 +  2841.29 + 800.00 + 3433.02 + 2752.74 + 3682.33 + 2359.28 +
         2119.15 + 3326.79 + 2684.05 + 3507.84 + 0.00 + 800.00)
```

2º) Dividir o valor obtido pelo total de meninas.

```
media = valor / 15
print(media)
```

```
2320.6766666666667
```

Ou seja, em média a Renda Mensal é de R$2.320,68.

pyladies São Paulo

Média

Codando fica:

nome_dataframe['coluna'].mean()

Média Renda Mensal:

```
dados_pyladies['Renda_Mensal'].mean()
```

```
2320.6766666666667
```

Média Idade:

```
dados_pyladies['Idade'].mean()
```

```
28.8
```

pyladies São Paulo

Desafio Média

Que tal utilizar a média para conhecermos ainda mais os dados que estamos analisando? Agora é com você!

1. Mostre a **média** da coluna **gorjetas**.

2. **Escolha** uma coluna quantitativa e mostre a média.

3. Como fazer para mostrar a **média** de **todas as colunas** quantitativas ao mesmo tempo? Mostra pra gente!

pyladies São Paulo

Resposta Desafio Média

1. Mostre a **média** da coluna **gorjetas**.

   tips_data['gorjeta'].mean()

2. **Escolha** uma coluna quantitativa e mostre a média.

   tips_data['total_conta'].mean()

3. Como fazer para mostrar a **média** de **todas as colunas** quantitativas ao mesmo tempo? Mostra pra gente!

   tips_data.mean()

pyladies São Paulo

Moda

A **Moda** é aquele elemento que mais se repete na distribuição dos dados.

Estado Origem	Frequência
SP	11

Qual será a UF que mais se repete?

nome_dataframe['coluna'].mode()

```
[48]  dados_pyladies['Estado Origem '].mode()

      0    SP
      dtype: object
```

pyladies São Paulo

Desafio Moda

Vimos que a moda é o valor que mais se repete em um conjunto de dados, daí vem o nome dela. Sendo assim, queremos saber:

1. **Quem mais frequenta** o restaurante que estamos analisando, homens ou mulheres?

2. Qual é o **dia preferido** para os clientes irem ao restaurante?

3. Na maioria das vezes os clientes vão ao restaurante para **almoçar** ou para **jantar**?

4. **Faça uma pergunta** e responda utilizando a moda. Conta pra gente o que descobriu! ;)

pyladies São Paulo

Resposta Desafio Moda

1. **Quem mais frequenta** o restaurante que estamos analisando, homens ou mulheres?
   tips_data['genero'].mode()

2. Qual é o **dia preferido** para os clientes irem ao restaurante?
   tips_data['dia'].mode()

3. Na maioria das vezes os clientes vão ao restaurante para **almoçar** ou para **jantar**?
   tips_data['horario'].mode()

4. **Faça uma pergunta** e responda utilizando a moda. Conta pra gente o que descobriu! ;)

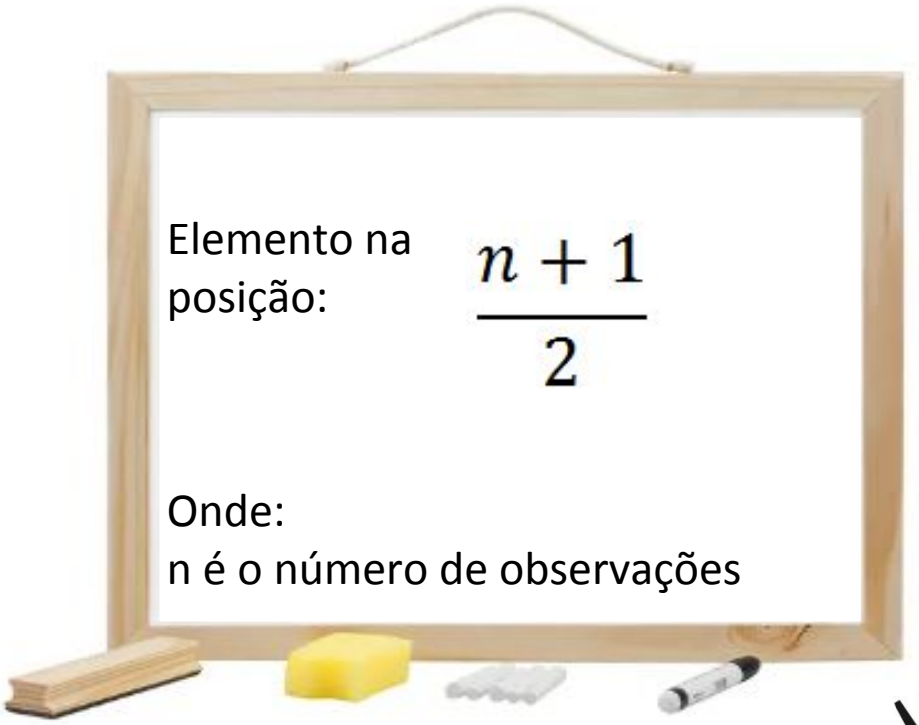
pyladies São Paulo

Mediana

Mediana é o valor do meio de um conjunto de dados ordenados.

- Para um conjunto com número ímpar de observações: é o valor que divide exatamente na metade esse conjunto.

- Para um conjunto com número par de observações: é a média dos dois valores do meio.



pyladies São Paulo

Mediana

Qual a Mediana quando observamos a idade das meninas?

Para responder essa pergunta precisamos:

1º) Ordenar os dados do menor para o maior valor;

2º) Selecionar o valor mediano dos dados.

Idade													
16	23	23	24	25	25	27	29	30	31	34	34	36	36	39

Observamos que a Mediana não é influenciada pelo valor baixo de idade.

pyladies São Paulo

Mediana

Codando fica:

nome_dataframe['coluna'].median()

Idade Mediana:

```
[50]  dados_pyladies['Idade'].median()

      29.0
```

Salário Mediano:

```
dados_pyladies['Renda_Mensal'].median()

2684.05
```

pyladies São Paulo

Desafio Mediana

Vimos que a mediana é o valor que divide ao meio um conjunto de dados ordenados. Então conta pra gente:

1. Qual é a **mediana** das **contas** que os clientes pagam no restaurante?

2. Qual é a **mediana** das **gorjetas** que os clientes pagam no restaurante?

3. A **mediana** das contas é **igual, maior ou menor** do que a **média** das contas?

4. **Compare** também a mediana e a média das **gorjetas**.

pyladies São Paulo

Resposta Desafio Mediana

1. Qual é a **mediana** das **contas** que os clientes pagam no restaurante?
   tips_data['total_conta'].median()

2. Qual é a **mediana** das **gorjetas** que os clientes pagam no restaurante?
   tips_data['gorjeta'].median()

3. A **mediana** das contas é **igual, maior ou menor** do que a **média** das contas?
   tips_data['total_conta'].mean()
   A mediana é menor do que a média do total das contas.

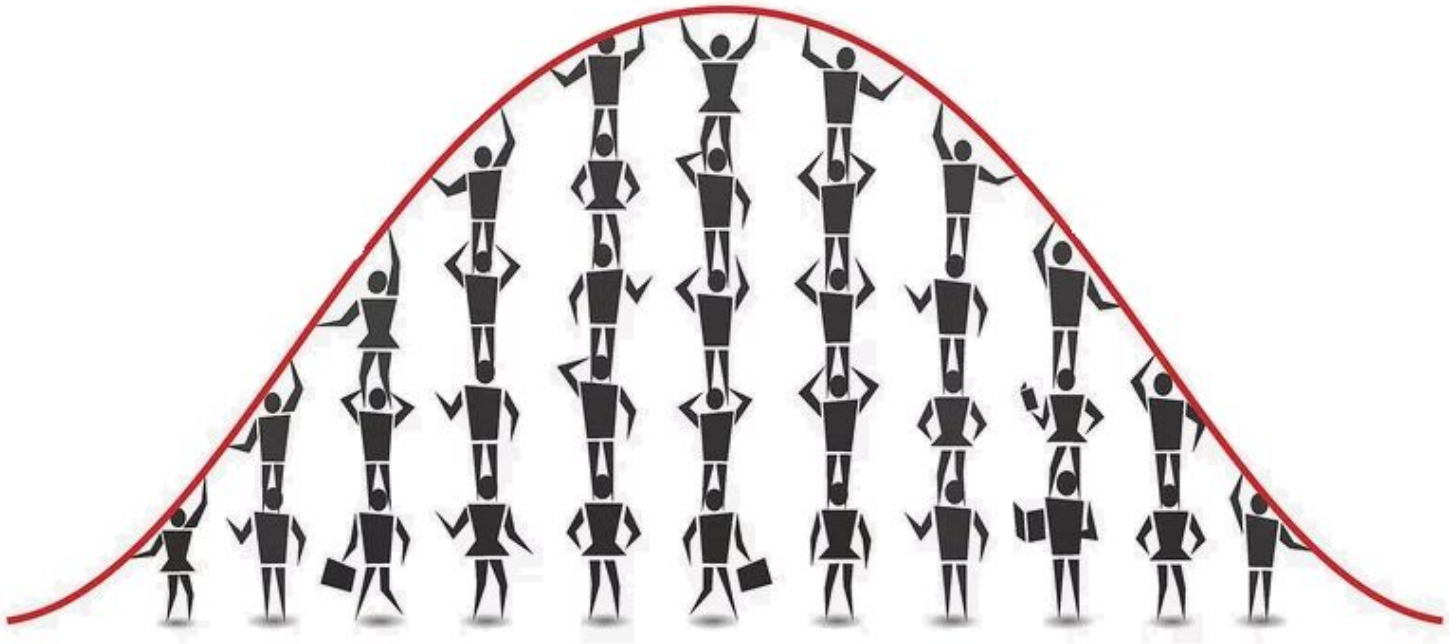
4. **Compare** também a mediana e a média das **gorjetas**.
   tips_data['gorjeta'].mean()
   A mediana e a média das gorjetas são iguais.

pyladies São Paulo

Mas média e mediana são as mesmas coisas???

Como Média, Moda e Mediana se relacionam?

Falando brevemente sobre distribuições, há vários tipos de comportamento natural das medidas que observamos, um deles é a distribuição Normal.



pyladies São Paulo

Como Média, Moda e Mediana se relacionam?

Em amostras normalmente distribuídas a Média, a Mediana e a Moda possuem valores próximos!

Se observarmos a renda mensal das meninas que trabalham com programação temos que a média e a mediana são muito próximas mesmo.

Média	R$ 2982,74
Mediana	R$ 2841,29

Se observarmos a idade das meninas, também obtemos valores de média e mediana próximos!

Média	Aprox. 29 anos
Mediana	29 anos

pyladies São Paulo

Desafio Média, Moda e Mediana

Vimos como as medidas se relacionam, agora conta pra gente:

1. Qual é a **média** do tempo que os clientes ficam no restaurante?
2. Qual é a **mediana** do tempo que os clientes ficam no restaurante?
3. Qual é a **moda** do tempo que os clientes ficam no restaurante?

pyladies São Paulo

Resposta Desafio Média, Moda e Mediana

1. Qual é a **média** do tempo que os clientes ficam no restaurante?

   tips_data['tempo_permanencia'].mean()

2. Qual é a **mediana** do tempo que os clientes ficam no restaurante?

   tips_data['tempo_permanencia'].median()

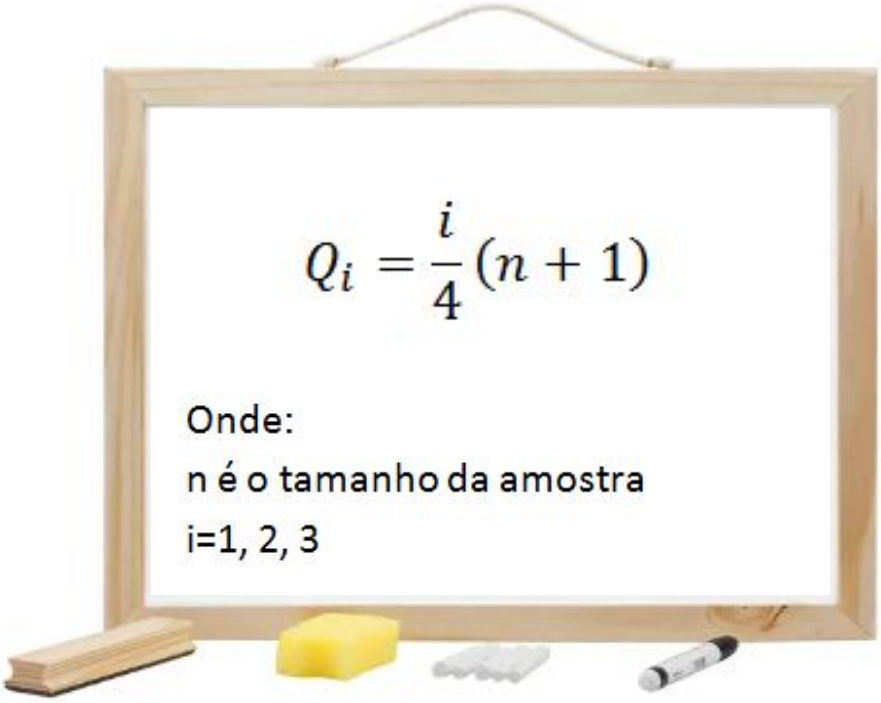
3. Qual é a **moda** do tempo que os clientes ficam no restaurante?

   tips_data['tempo_permanencia'].mode()

pyladies São Paulo

Quartis

Quartis são valores que dividem uma amostra de dados ordenados em quatro partes iguais.
Com eles você pode rapidamente avaliar a dispersão e a tendência central de um conjunto de dados, que são etapas importantes na compreensão dos seus dados.



pyladies São Paulo

Quartis

Importante: para encontrar os quartis, os dados devem estar ordenados!

- 1º Quartil (Q1) - é onde estão 25% dos valores do conjunto de dados.
- 2º Quartil (Q2) - é onde estão até 50% dos valores, ou seja, a mediana!
- 3º Quartil (Q3) - é onde estão até 75% dos valores.

Isso quer dizer que:

✓ 25% dos valores do conjunto de dados são menores ou iguais ao Q1 e 75% dos valores são superiores ao Q1.

✓ 25% dos valores são superiores ou iguais ao Q3 e 75% dos valores são menores que Q3.

✓ 50% dos valores estão entre o 1º e o 3º Quartil.

pyladies São Paulo

Quartis

Codando fica:

nome_dataframe['coluna'].quantile()

25%

```
dados_pyladies.quantile(.25)
```

```
Idade           24.500
Escolaridade     3.000
Renda_Mensal  1459.575
Name: 0.25, dtype: float64
```

75%

```
[21] dados_pyladies.quantile(.75)
```

```
Idade           34.000
Escolaridade     4.000
Renda_Mensal  3379.905
Name: 0.75, dtype: float64
```

pyladies São Paulo

Desafio Quartis

Vimos que os quartis dividem o conjunto de dados em 4 partes iguais e podem nos dar insights sobre os dados.

1. Quais são os **quartis** das **contas**? (coluna total_contas)

2. Calcule a **mediana** das contas . Existe algum **quartil igual** à mediana? Por que?

3. Qual é o **máximo** do valor que classifica uma conta entre as **25% contas mais baratas**?

4. Qual é o **mínimo** do valor que classifica uma conta entre as **25% contas mais caras**?

pyladies São Paulo

Resposta Desafio Quartis

1. Quais são os **quartis** das **contas**? (coluna total_contas)

   tips_data['total_conta'].quantile(0.25)

   tips_data['total_conta'].quantile(0.50)

   tips_data['total_conta'].quantile(0.75)

2. Calcule a **mediana** das contas . Existe algum **quartil igual** à mediana? Por que?

   tips_data['total_conta'].median()

pyladies São Paulo

Resposta Desafio Quartis

3. Qual é o **máximo** do valor que classifica uma conta entre as **25% contas mais baratas**?

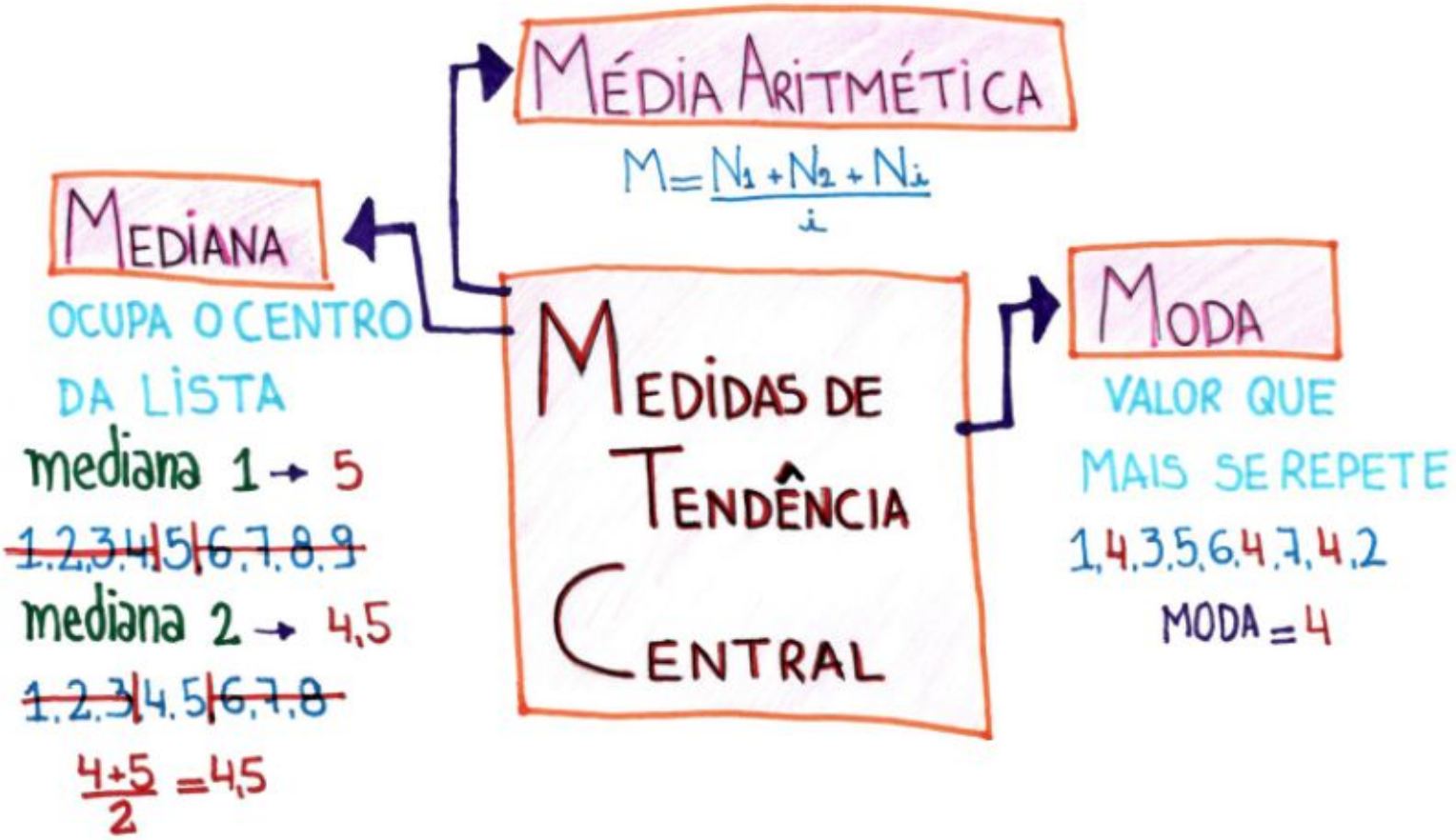
   tips_data['total_conta'].quantile(0.25)
   Contas até 13.35 estão entre as contas mais baratas do dataset tips.
   Utilizamos o 1o quartil (Q1) para responder essa questão pois 25% dos dados são iguais ou menores que ele.

4. Qual é o **mínimo** do valor que classifica uma conta entre as **25% contas mais caras**?

   tips_data['total_conta'].quantile(0.75)
   Contas a partir de 24.13 estão entre as contas mais caras do dataset tips.
   Utilizamos o 3o quartil (Q3) para responder essa questão pois 25% dos dados são iguais ou maiores que ele e 75 % dos dados são menores que ele.

pyladies São Paulo

O que vimos até aqui!



Fonte: https://brasilescola.uol.com.br/matematica/moda-media-mediana.htm

pyladies São Paulo

Tabela de Frequência

Mas será que a renda mensal média varia com relação às demais características?

Para respondermos isso podemos criar uma tabela de frequência que nos mostrará a variação dos dados um pelo outro

Trabalha como Programadora	Soma Renda Mensal	Quantidade Meninas	Renda Mensal Média
S	32810,15	11	2982,74
N	2000	4	500,00

Característica: Trabalhar ou não com programação!

pyladies São Paulo

Tabela de Frequência - Usando o Groupby

O Pandas possui a função groupby que nos permite agrupar dados, como o exemplo anterior.

Ele nos permite visualizar rapidamente uma tabela de frequência.

nome_dataframe.groupby('coluna').método()

```
dados_pyladies.groupby('Trabalha_como_Programadora').count()
```

Trabalha_como_Programadora	Estado Origem	Idade	Escolaridade	Renda_Mensal
N	4	4	4	4
S	11	11	11	11

- alguns métodos não funcionam com o groupby, para saber mais consulte a documentação da biblioteca Pandas

pyladies São Paulo

Tabela de Frequência - Groupby

Podemos agrupar separando os dados por outras variáveis (colunas) como no exemplo abaixo

nome_dataframe.groupby('coluna')['coluna'].método()

```
dados_pyladies.groupby('Trabalha_como_Programadora')['Escolaridade'].value_counts()
```

```
Trabalha_como_Programadora  Escolaridade
N                           4               2
                            1               1
                            2               1
S                           3               6
                            4               2
                            5               2
                            2               1
Name: Escolaridade, dtype: int64
```

Como podemos ler a tabela acima?

pyladies São Paulo

Tabela de Frequência - Groupby

Há vários métodos que podem ser utilizados com o groupby.

Para contar os valores

```
nome_dataframe.groupby('coluna')['coluna'].value_counts()
```

Para somar valores

```
nome_dataframe.groupby('coluna')['coluna'].sum()
```

qual outro?

```
nome_dataframe.groupby('coluna')['coluna'].método()
```

pyladies São Paulo

Tabela de Frequência - Groupby + Agg

Podemos utilizar um Groupby com uma função Para isso utilizamos a função aggregation.

nome_dataframe.groupby('coluna')['coluna'].agg(['método', 'método'])

```
dados_pyladies.groupby('Trabalha_como_Programadora')['Renda_Mensal'].agg(['count','mean','median'])
```

Trabalha_como_Programadora	count	mean	median
N	4	500.000000	600.00
S	11	2982.740909	2841.29

pyladies São Paulo

Desafio Tabela de Frequência - Groupby + Agg

Vimos que o Groupby agrupa valores. Sendo assim, utilize o GroupBy para descobrir:

1. Temos **mais homens** ou **mulheres fumantes** no dataset? **Quantas** mulheres e quantos homens são fumantes?

2. Existe algum **dia na semana** que **há mais mulheres do que homens** no restaurante? Se sim, qual é o dia?

3. Qual é o **número de pessoas nas mesas** que é **mais comum** durante o almoço e durante o jantar?

4. Em **qual refeição** o restaurante **mais fatura**?

5. Mostre a **soma** total e a **média** das contas por gênero com um único comando do Pandas.

pyladies São Paulo

Resposta Desafio Tabela de Frequência

1. Temos **mais homens** ou **mulheres fumantes** no dataset? **Quantas** mulheres e quantos homens são fumantes?

   tips_data.groupby('fumante')['genero'].value_counts()
   Analisando os dados vemos que o dataset tem 33 mulheres fumantes e 60 homens que fumam. Portanto os dados mostram mais homens fumantes.

2. Existe algum **dia na semana** que **há mais mulheres do que homens** no restaurante? Se sim, qual é o dia?

   tips_data.groupby('genero')['dia'].value_counts()
   Sim, na quinta-feira há mais mulheres que homens.

pyladies São Paulo

Resposta Desafio Tabela de Frequência

3. Qual é o **número de pessoas nas mesas** que é **mais comum** durante o almoço e durante o jantar?

   tips_data.groupby('horario')['pessoas_mesa'].value_counts()
   Durante o almoço e o jantar é mais comum ter 2 pessoas na mesa.

4. Em **qual refeição** o restaurante **mais fatura**?

   tips_data.groupby('horario')['total_conta'].sum()
   O restaurante fatura mais durante o horário do jantar.

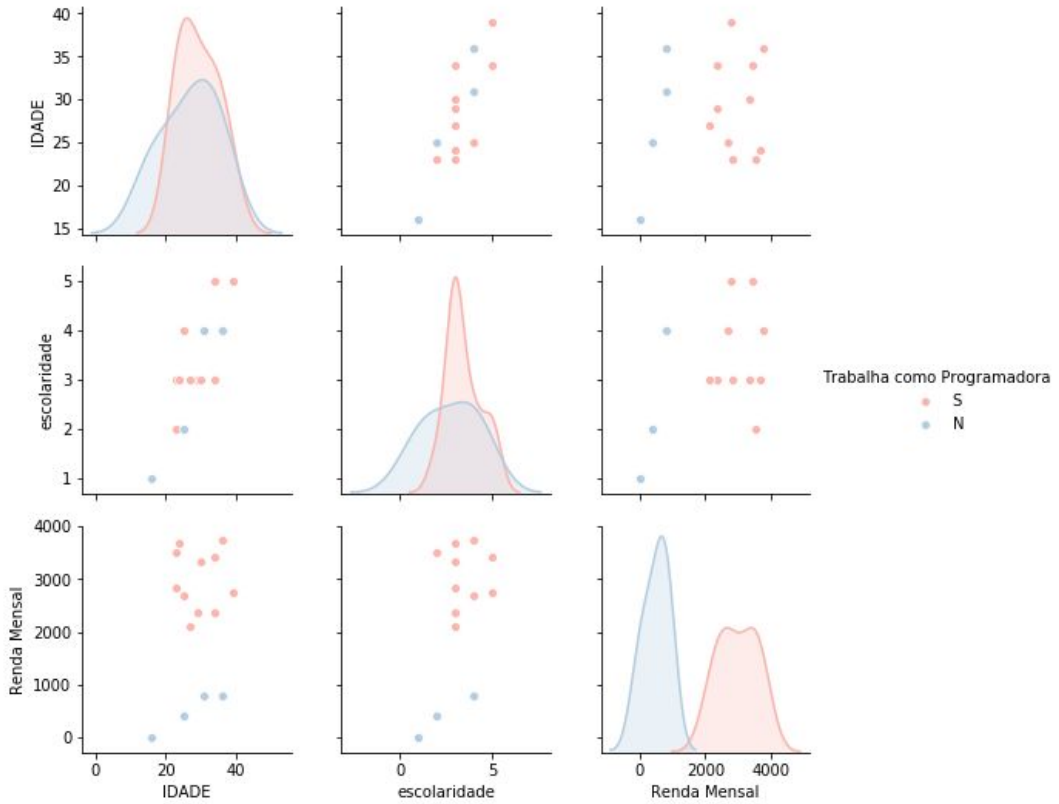
5. Mostre a **soma** total e a **média** das contas por gênero com um único comando do Pandas.

   tips_data.groupby('genero')['total_conta'].agg(['sum', 'mean'])

pyladies São Paulo

Visualização de Dados - Gráficos

Para identificarmos como as variáveis estão distribuídas e tornar as informações de um conjunto mais visuais nós plotamos gráficos.



pyladies São Paulo

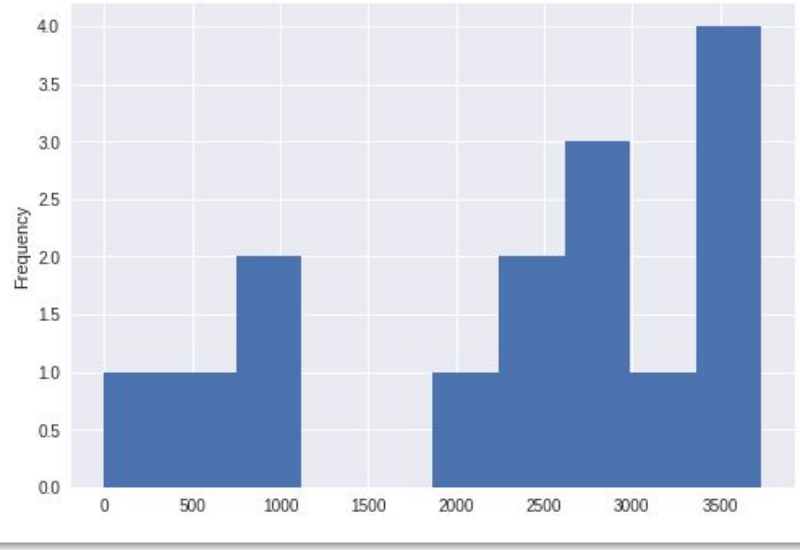
Visualização de Dados - Gráficos

Histograma

O Histograma é um gráfico que representa a distribuição de frequências de uma variável numérica contínua.

nome_dataframe['coluna'].plot.hist()

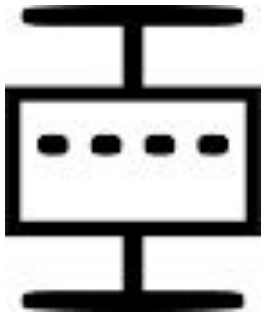
```
dados_pyladies['Renda_Mensal'].plot.hist();
```



pyladies São Paulo

Visualização de Dados - Gráficos

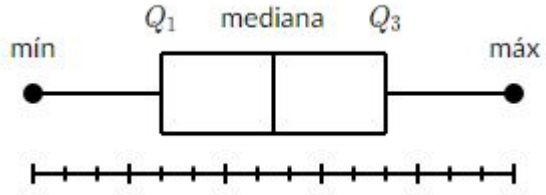
BoxPlot



O boxplot nos permite avaliar a distribuição do conjunto de dados, utilizando como referência os quartis.

A “caixa principal” é formada pelo primeiro quartil, a mediana e terceiro quartil.

As hastes inferior e superior são os limites e podem ser calculadas da seguinte forma:



$$Limite\ inferior: Q_1 - 1{,}5(Q_3 - Q_1)$$

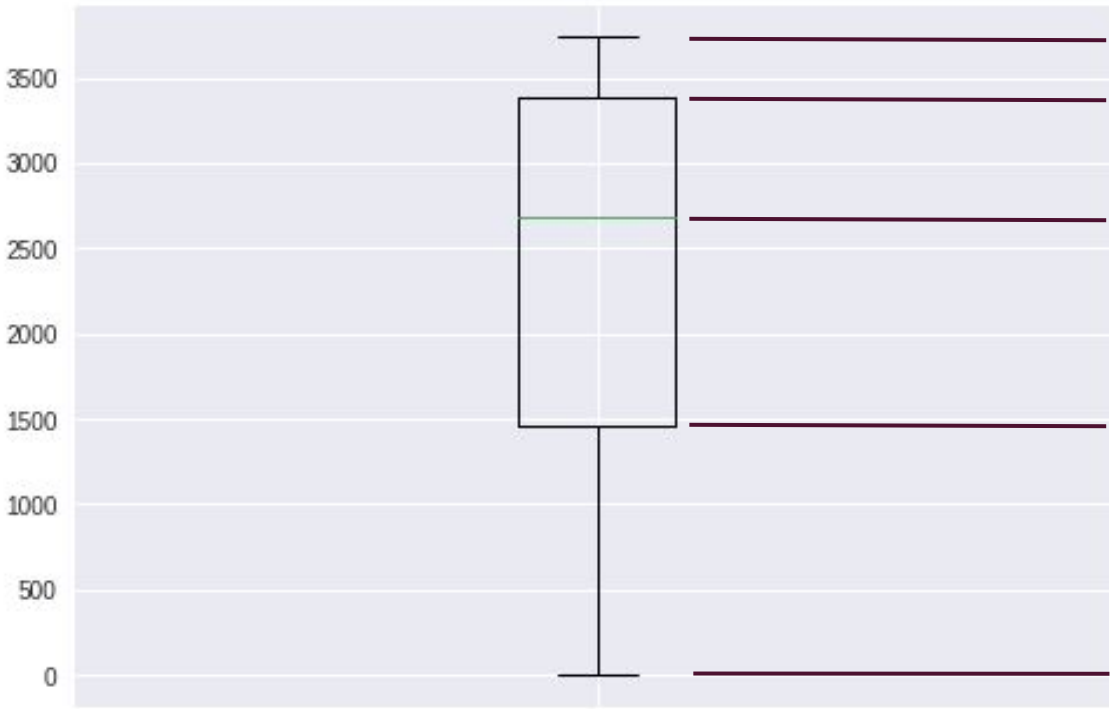
$$Limite\ superior: Q_3 + 1{,}5(Q_3 - Q_1)$$

pyladies São Paulo

Visualização de Dados - Gráficos

plt.boxplot(nome_dataframe['coluna'])

```
plt.boxplot(dados_pyladies['Renda_Mensal']);
```



Limite Superior

3º Quartil

Mediana

1º Quartil

Limite Inferior

pyladies São Paulo

Desafio Visualização de Dados - Gráficos

1. Explore a visualização do tempo de permanência (coluna **tempo_permanencia**) utilizando um histograma.

2. Observe a distribuição do valor da conta (coluna **total_conta)** utilizando um **boxplot**.

pyladies São Paulo

Resposta Desafio Visualização de Dados

1. Mostre o **histograma** da coluna **tempo_permanencia**.

```
plt.hist(tips_data['tempo_permanencia'])
plt.title('Histograma: Tempo de Permanência')
plt.xlabel('tempo_permanencia')
plt.ylabel('Frequência');
```

2. Plote o **boxplot** da coluna **total_conta**.

```
plt.boxplot(tips_data['total_conta'])
plt.title('Boxplot: Total da Conta')
plt.ylabel('total_conta')
```

pyladies São Paulo

Dispersão dos Dados

Quando comparamos a média com o restante dos valores de uma váriavel, nós queremos entender o quanto aquele valor está distante da média.

ID	Estado Origem	Idade	Escolaridade	Trabalha como Programadora	Renda Mensal
1	SP	36	4	S	3737,52
2	SP	25	2	N	400,00

A média da Renda Mensal é de: **R$ 2320,68**

Se compararmos os valores da tabela acima percebemos o quanto eles variam em relação a média

pyladies São Paulo

Variância

Uma medida muito interessante para avaliarmos a dispersão dos dados é a **variância**!
Vimos que a média nos informa sobre a tendência central, mas a variância que indica como esses dados variam dentro de uma distribuição.

pyladies São Paulo

Variância

Será que as meninas que trabalham como programadora tem rendas parecidas? E as meninas que não trabalham como programadoras?

Para responder essa pergunta precisamos:

1º) Selecionar separadamente as meninas que trabalham ou não, como programadoras;
2º) Avaliar a soma dos desvios ao quadrado;
3º) Dividir essa soma pelo total de meninas considerado.

pyladies São Paulo

Variância

Renda Mensal (x)
3737,52
400
2366,14
2841,29
800
3433,02
2752,74
3682,33
2359,28
2119,15
3326,79
2684,05
3507,84
0
800
2684,05

$$s = \frac{(x1 - \text{média})^2 + (x2-\text{média})^2 + (x3-\text{média})^2}{(n-1)}$$

$$s = \frac{(3737,52-26805)^2 + (400-2684,05)^2 + \ldots\ldots}{(15-1)}$$

$$s = 1560989.622$$

pyladies São Paulo

Variância

Codando fica:

nome_dataframe['coluna'].var()

Variância Renda:

```
dados_pyladies['Renda_Mensal'].var()
```

```
1560989.6225666667
```

Variância Idade:

```
dados_pyladies['Idade'].var()
```

```
39.60000000000001
```

pyladies São Paulo

Desvio Padrão

O **desvio padrão** é uma medida que expressa o grau de dispersão de um conjunto de dados

Ele é a raiz quadrada da variância e a vantagem de utilizarmos esta medida é que o desvio padrão é expresso na mesma unidade dos dados, o que facilita a comparação.

pyladies São Paulo

Desvio Padrão

Codando fica:

nome_dataframe['coluna'].std()

Desvio Padrão Renda:

```
dados_pyladies['Renda_Mensal'].std()
```

```
1249.3957029567
```

Desvio Padrão Idade:

```
dados_pyladies['Idade'].std()
```

```
6.292853089020 91
```

pyladies São Paulo

Desafio - Variabilidade

Vimos que podemos usar duas medidas de **variabilidade** para entender o quanto os valores variam em relação a média. Sendo assim, responda:

1. Qual a **variância do tempo de permanência** dos clientes no restaurante ? Ela varia muito em **relação a média**? E o desvio padrão? O que podemos entender com isso?

2. Será que o **valor da fatura varia** muito entre homens e mulheres? Selecione a **melhor medida** para comparar um grupo ao outro e verifique.

pyladies São Paulo

Resposta Desafio - Variabilidade

Vimos que podemos usar duas medidas de **variabilidade** para entender o quanto os valores variam em relação a média. Sendo assim, responda:

1. Qual a **variância do tempo de permanência** dos clientes no restaurante ? Ela varia muito em **relação a média**? E o desvio padrão? O que podemos entender com isso?
   tips_data['tempo_permanencia'].var()
   tips_data['tempo_permanencia'].std()

2. Será que o **valor da fatura varia** muito entre homens e mulheres? Selecione a **melhor medida** para comparar um grupo ao outro e verifique.
   tips_data['total_conta'].std()
   tips_data[tips_data['genero']=='Feminino']['total_conta'].std()
   tips_data[tips_data['genero']=='Masculino']['total_conta'].std()
   Outra solução: tips_data.groupby('genero')['total_conta'].std()

pyladies São Paulo

Por fim! Describe

Para conseguirmos visualizar as medidas centrais e de dispersão de um conjunto de dados, nós podemos utilizar o método describe.

nome_dataframe.describe()

```
[14] dados_pyladies.describe()
```

	Idade	Escolaridade	Renda_Mensal
count	15.000000	15.000000	15.000000
mean	28.800000	3.266667	2320.676667
std	6.292853	1.099784	1249.395703
min	16.000000	1.000000	0.000000
25%	24.500000	3.000000	1459.575000
50%	29.000000	3.000000	2684.050000
75%	34.000000	4.000000	3379.905000
max	39.000000	5.000000	3737.520000

pyladies São Paulo

Desafio Describe

Vimos que o Describe traz todas as medidas de uma variável, agora utilizando este método verifique:

1. As informações do **faturamento** do restaurante.

2. As informações de todo o **dataset**.

3. Que tipos de variáveis ele mostra?

pyladies São Paulo

Resposta Desafio Describe

Vimos que o Describe traz todas as medidas de uma variável, agora utilizando este método verifique:

1. As informações do **faturamento** do restaurante.

   tips_data['total_conta'].describe()

2. As informações de todo o **dataset**.
   tips_data.describe()

3. Que tipos de variáveis ele mostra?
   O describe mostra resultados para variáveis numéricas.

pyladies São Paulo

O que vimos hoje

- O valor médio de uma distribuição, calculando a **média**;
- Vimos como identificar qual dado mais se repete na distribuição, observando a **moda**;
- Aprendemos que a **mediana** separa os dados no meio, 50% 50% - e que ela é o 2º quartil;
- O conjunto de dados pode ser dividido em 4 **Quartis**, 25% em cada um;
- Podemos construir **Tabelas de Frequência** para os dados categóricos;
- Conversamos que visualizar os dados é um jeito de entender a distribuição das variáveis e para isso plotamos o **histograma** e o **boxplot**;
- Vimos como os dados podem ser dispersos e que podemos avaliar essa dispersão pela **variância** e o **desvio padrão**.

pyladies São Paulo

Não Entre em Pânico!

CIÊNCIA DE DADOS

01	FAÇA UMA PERGUNTA INTERESSANTE	• O que você quer prever? • Qual o objetivo científico? • O que você faria se tivesse todos esses dados? • O que pode acontecer?
02	OBTENHA OS DADOS	• Como os dados foram amostrados? • Quais dados são relevantes? • Existem problemas de privacidade?
03	EXPLORE OS DADOS	• Existem anomalias? • Existem padrões?
04	MODELE OS DADOS	• Construa um modelo; • Encaixe o modelo; • Valide o modelo.
05	VISUALIZE E DIVULGUE OS RESULTADOS	• O que aprendemos? • Os resultados fazem sentido? • Podemos contar uma história?

pyladies São Paulo

Para saber mais

- Livro Guia Mangá de Estatística - Shin Takahashi
- Plataforma Kaggle - https://www.kaggle.com/
- Podcast Pizza de Dados - https://pizzadedados.com/
- Documentação Pandas - https://pandas.pydata.org/pandas-docs/stable/index.html
- Udacity - https://www.udacity.com/
- Canal EstaThiFisco https://www.youtube.com/channel/UC4jROkPjTvnXRkuo2GAwKXw
- Minerando Dados - http://minerandodados.com.br/
- Cientista de Dados com GIFs - https://paulovasconcellos.com.br/
- Data Hackers https://datahackers.com.br/
- Estatística Básica - P. A. Bussab, W. de O. Moretin - https://edisciplinas.usp.br/mod/resource/view.php?id=2425203

pyladies São Paulo

E por hoje é só pessoal!


PyLadiesSP


PyLadiesSãoPaulo


PyLadiesSP


@PyLadiesSP


PyLadiesSP


saopaulo@pyladies.com

Mulheres que amam programar e ensinar Python