

CD – Ciência de Dados

Baseado nos slides do Prof. Miguel Bozer

Informações da Disciplina

Avaliações/Projetos baseado em Competências

CAPACIDADES BÁSICAS

- Transformar dados obtidos através de cálculos matemáticos em informações pertinentes ao processo;
- Extrair informações de dados obtidos para o gerenciamento de processos industriais através de análises estatísticas;
- Processar dados para a geração de relatórios analíticos tendo em vista a visualização da informação.

Informações da Disciplina

Avaliações/Projetos baseado em Competências

CAPACIDADES SOCIOEMOCIONAIS

- Demonstrar atenção a detalhes.
- Demonstrar capacidade de síntese.
- Demonstrar capacidade de solucionar problemas.
- Demonstrar capacidade de tomar decisão.
- Demonstrar organização.
- Demonstrar raciocínio lógico

Informações da Disciplina

As notas seguirão uma tabela de nível de desempenho. Vide exemplo abaixo:

Critérios	Nota
Acertou todos os críticos e desejáveis	100
Acertou todos os críticos e 5/6 desejáveis	95
Acertou todos os críticos e 3/4 desejáveis	80
Acertou todos os críticos e 1/2 desejáveis	65
Acertou todos os críticos e 0 desejáveis	50
Não acertou todos os críticos	0

Informações da Disciplina

Teremos na nossa disciplina as avaliações **formativas**

As formativas são realizadas **durante** a evolução do conteúdo para avaliar se os alunos estão acompanhando o desenvolvimento da disciplina

- Feedback para o aluno compreender onde ele está defasado!
- Não valem nota

Ferramentas da Disciplina

- Nossas aulas serão muito HANDS ON
- Iremos utilizar a linguagem em python para utilizarmos bibliotecas de IA.



Introdução ao conceito de Dados

Dado - Definição

R\$ 45,90 > ???

Os dados são elementos que constituem a **matéria-prima** da informação. Podemos defini-los, também, como **conhecimento bruto**, ainda não devidamente tratado para prover insights para uma organização.

Assim, os dados representam um ou mais significados que, de forma isolada, não conseguem ainda transmitir uma mensagem clara.

Informação - Definição

São os dados representados de forma organizada

Total

=

R\$ 45,90

Conhecimento - Definição

É a informação com um contexto bem definido, processado de forma efetiva pelos profissionais

Compra no Aplicativo X

Total

=

R\$ 45,90

Introdução à Ciência de Dados

Ciência de Dados

Todos nós de alguma forma já tivemos algum tipo de interação com a **ciência de dados**.

Por exemplo:



Pesquisas na web



Esta Foto de Autor Desconhecido está licenciado em CC BY-SA-NC

Perguntas para o seu telefone

Ciência de Dados

Ciência de dados pode ser vista como uma intersecção entre **estatística** e **ciência da computação**

Estatística: Dessa área vem uma longa tradição de:

- Análise exploratória dos dados;
- Teste de significância;
- Visualização dos dados;

Ciência de Dados

Ciência de dados pode ser vista como uma intersecção entre **estatística** e **ciência da computação**

Ciência da computação: Dessa área temos o ***machine learning*** e a computação de alta performance para trabalhar com dados de larga escala

Obs.: *Machine Learning* será visto em mais detalhes na disciplina de inteligência artificial;

Ciência de Dados

Uma pergunta importante, por que estudar ciência de dados?

Novas tecnologias permitem que seja possível armazenarmos um grande volume de dados das pessoas, tais como:

- Redes sociais ;
- Lista de e-books lidos;
- Filmes que assistimos;
- Histórico de compras;
- Etc

Ciência de Dados

Uma pergunta importante, por que estudar ciência de dados?

Além disso, as empresas também começaram a aproveitar dados ao invés de descartá-los:

- Dados de sensores;
- Atividades em suas páginas na internet;
- Atividade na rede corporativa;
- Áudios gravados de reuniões;
- Etc;

Dados armazenados não apenas para manter a política de informação da empresa, mas para poderem ser **analisados**

Ciência de Dados

Uma pergunta importante, por que estudar ciência de dados?

Serviços em Cloud ajudaram a mudar o cenário da área de análise de dados, uma vez que permitiu o acesso a **qualquer pessoa** a realizar a análise de dados de um **grande volume de dados** em um pequeno intervalo de tempo.



Esta Foto de Autor Desconhecido está licenciado em [CC BY-NC](#)



Esta Foto de Autor Desconhecido está licenciado em [CC BY-ND](#)



Google Cloud

Esta Foto de Autor Desconhecido está licenciado em [CC BY-NC-ND](#)



Esta Foto de Autor Desconhecido está licenciado em [CC BY-SA](#)

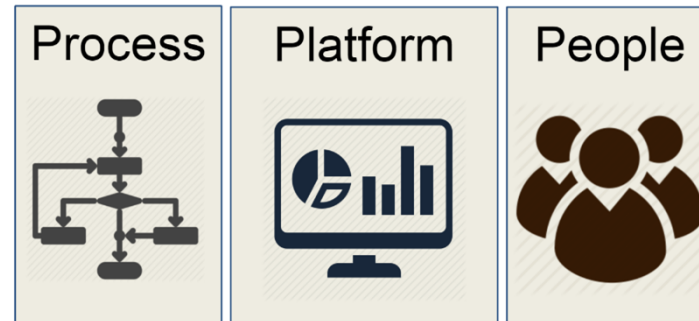


Ciência de Dados

Uma pergunta importante, por que estudar ciência de dados?

Os dados começaram a ser usados para basear as decisões corporativas

- Decisões baseadas em ciência e não apenas em intuição!



Esta Foto de Autor Desconhecido está licenciado em
CC BY-SA-NC

Ciência de Dados

Diferença da mentalidade entre um cientista de dados e um programador:

- **Programador:** Trata os dados como algo que deve ser processado. Os dados não são o foco do trabalho, mas sim o **processamento!**



Ciência de Dados

Diferença da mentalidade entre um cientista de dados e um programador:

- **Cientista de Dados:** Cientistas são obcecados em descobrir algo! Logo o cientista de dados busca usar a programação para descobrir informações nos dados.
- Nosso trabalho será **transformar números em insights**. É importante entender tanto o **porquê** quanto o **como**.

Ciência de Dados

Uma boa característica de um cientista de dados é a **curiosidade**.

Cientistas de dados usualmente fazem as seguintes perguntas para eles mesmos:

- O que eu posso aprender a partir de um conjunto de dados?
- O que eu realmente preciso saber sobre um assunto particular?
- Qual o significado daquilo que eu descobri?

Trabalhando com Dados

Trabalhando com Dados

Para podermos aprender a fazer trabalhos na área de ciência de dados, precisamos de um ambiente de programação.

Nesse cenário iremos usar um **Web Integrated Development Environment(WIDE)**

Trabalhando com Dados

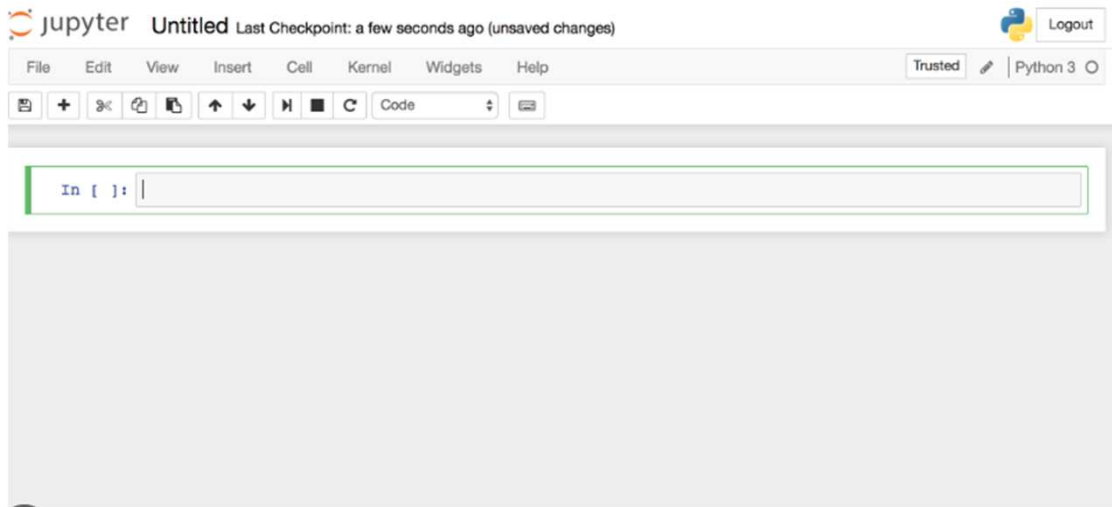
Web Integrated Development Environment (WIDE): Jupyter

Um ambiente de programação que usa o **próprio navegador** do computador para programação.

Nele podemos inserir **texto** e **células de código** com o objetivo de criar um documento interativo, com descrições a partir de resultados obtidos com as células de código.

O documento criado leva o nome de **notebook**

Para trabalhar com o Jupyter temos que instalá-lo em nosso computador



Trabalhando com Dados

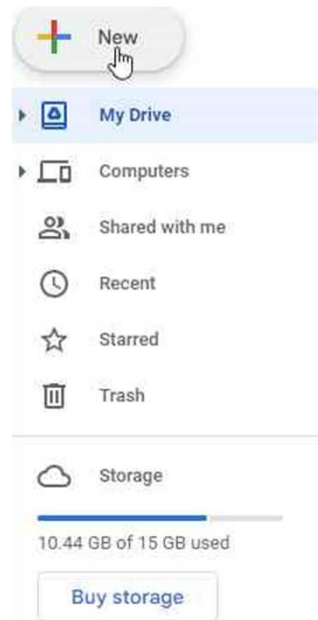
Para não termos a necessidade de instalar nada nos computadores da escola e nos computadores pessoais dos alunos iremos usar o **Google Colaboratory**



Para isso precisamos apenas ter uma **conta google**. Caso você não tenha a mesma, **crie uma agora!**

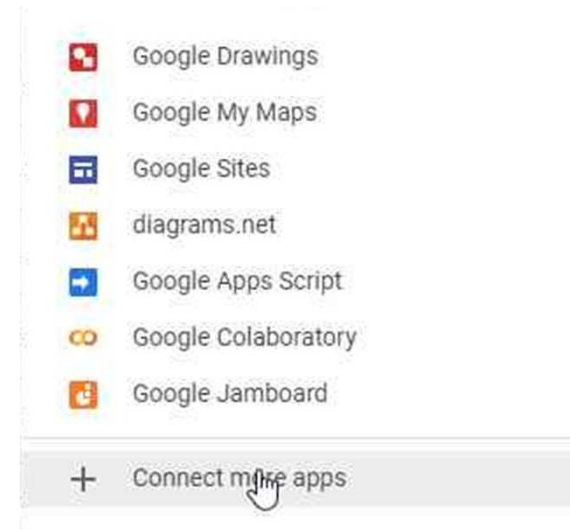
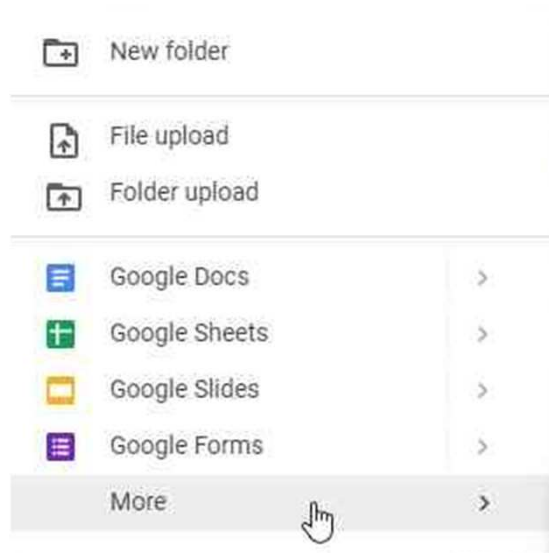
Instalando o Google Colaboratory

1. Faça o login/crie uma conta no Google Drive acessando <https://drive.google.com/>
2. Após efetuar o login no Google Drive temos que instalar a extensão do Colaboratory. Para isso clique em **+ New**



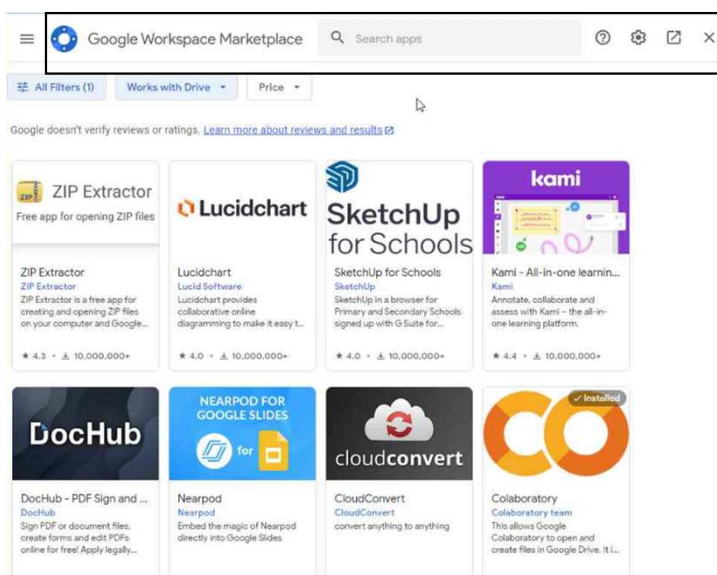
Instalando o Google Colaboratory

3. Vá até o menu More: 4. Clicar em + Connect more apps

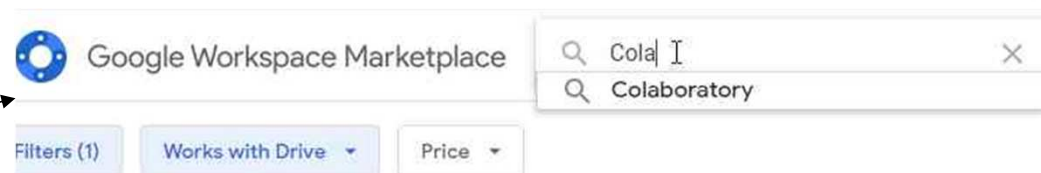


Instalando o Google Colaboratory

5. Aguarde a tela a seguir carregar

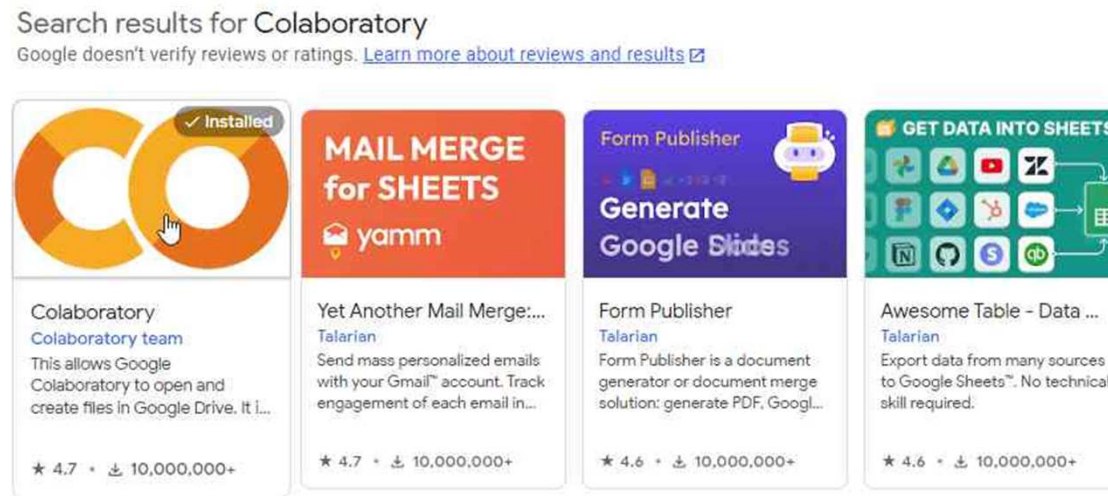


6. Na aba de pesquisa digite Colaboratory e clique no ícone dele.



Instalando o Google Colaboratory

7. Clicar no ícone do Google Colaboratory:



Instalando o Google Colaboratory

8. Clicar em **Install**

Obs.: No slide está Uninstall, pois o Colaboratory já estava instalado no Google Drive do professor



Pronto já podemos usar o Google Colaboratory nas aulas

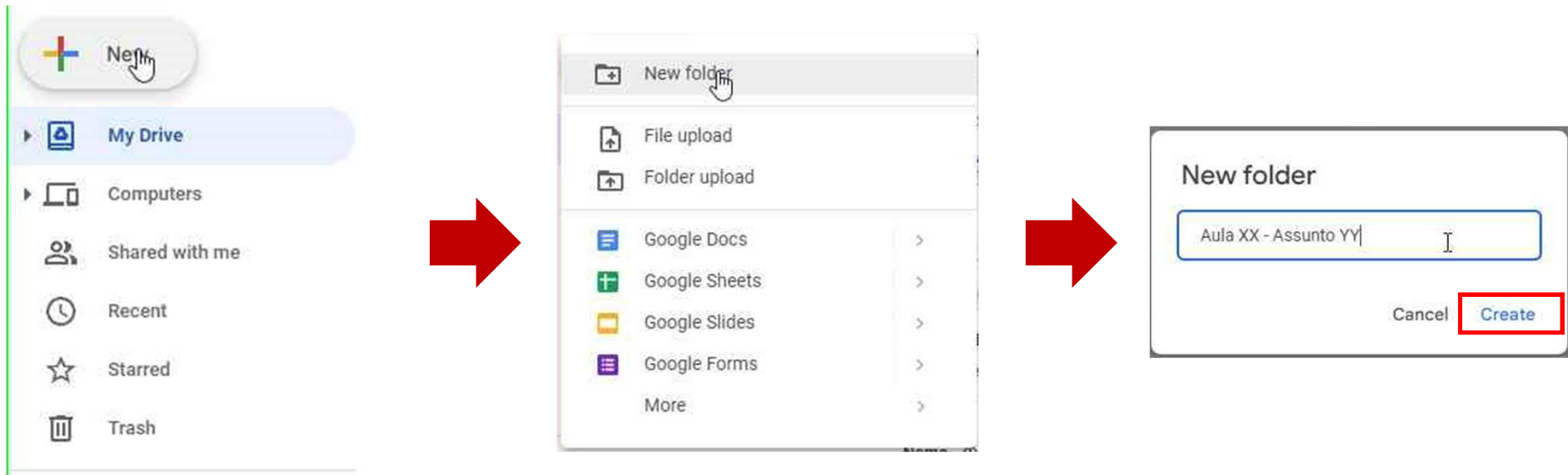
Criando um Jupyter Notebook

Para organizar os nossos notebooks podemos salvá-los em pastas do nosso Google Drive.

O professor sempre irá disponibilizar os notebooks das aulas, entretanto tente manter alguma organização dos seus arquivos para os **seus estudos**.

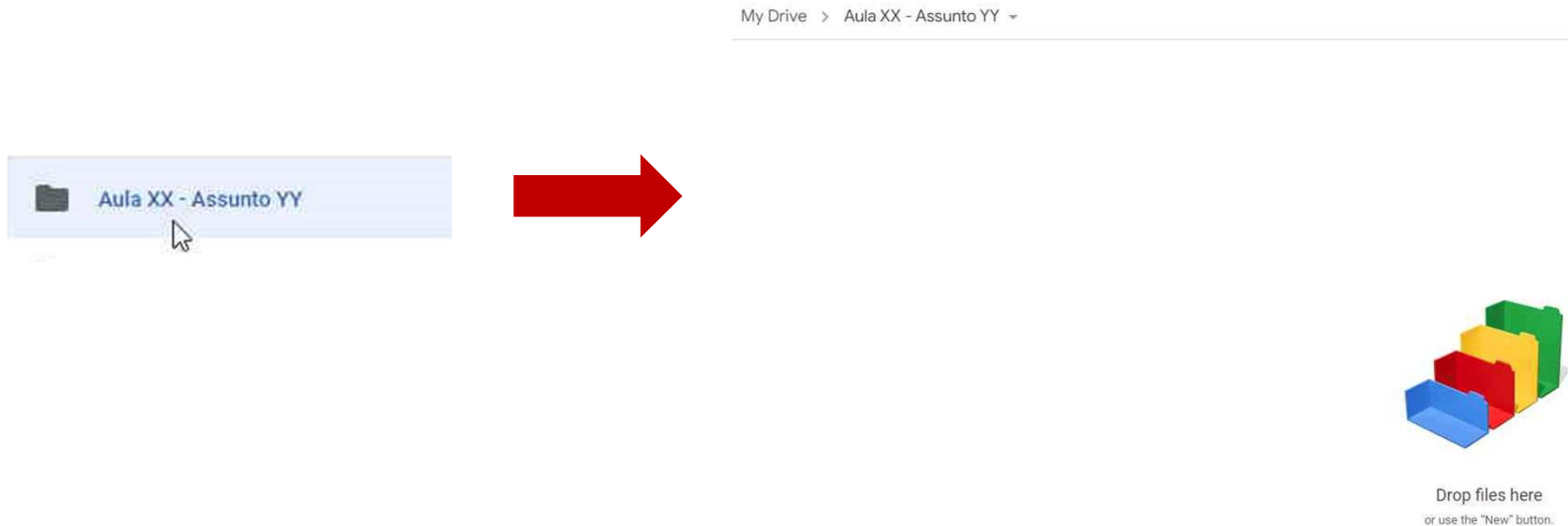
Criando um Jupyter Notebook

Para criar um arquivo Jupyter Notebook primeiro, **recomenda-se** que crie uma pasta para organizar seus arquivos



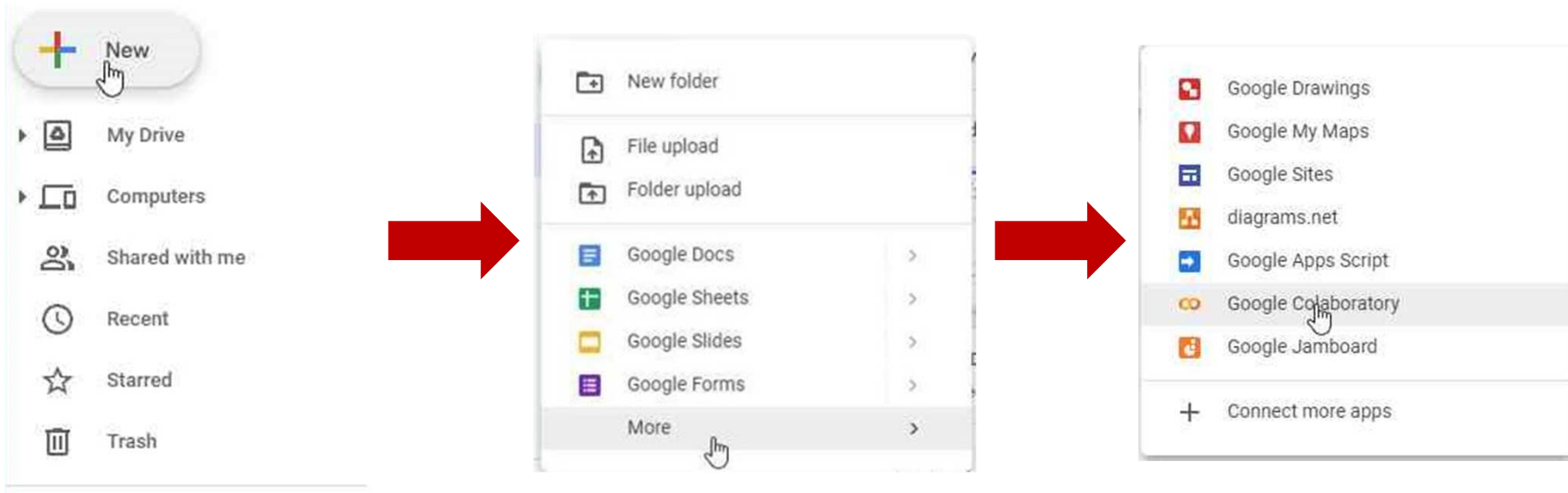
Criando um Jupyter Notebook

Para criar um arquivo Jupyter Notebook primeiro, **recomenda-se** que crie uma pasta para organizar seus arquivos



Criando um Jupyter Notebook

Após isso, dentro da pasta criada, vamos criar um arquivo Jupyter Notebook



Criando um Jupyter Notebook

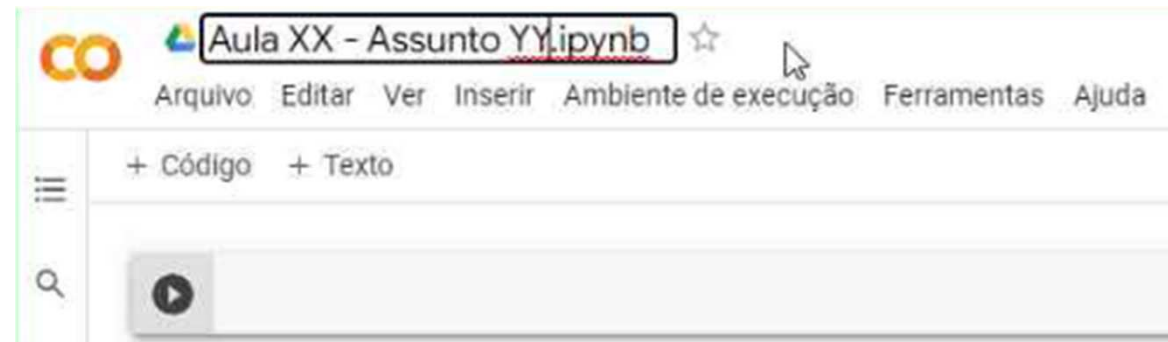
Pronto, você acabou de criar um documento Jupyter Notebook:



Vamos agora aprender a trabalhar com esse documento

Editando um Jupyter Notebook

Podemos alterar o nome do documento para organizarmos os relatórios que criarmos:



Editando um Jupyter Notebook

Podemos adicionar células de código e células de texto nesse tipo de documento.

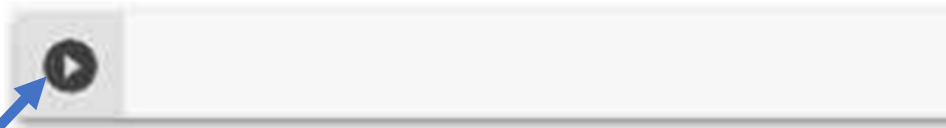
- **Texto:**



Isso é um texto qualquer

Texto apenas para
descrever algo de um
relatório

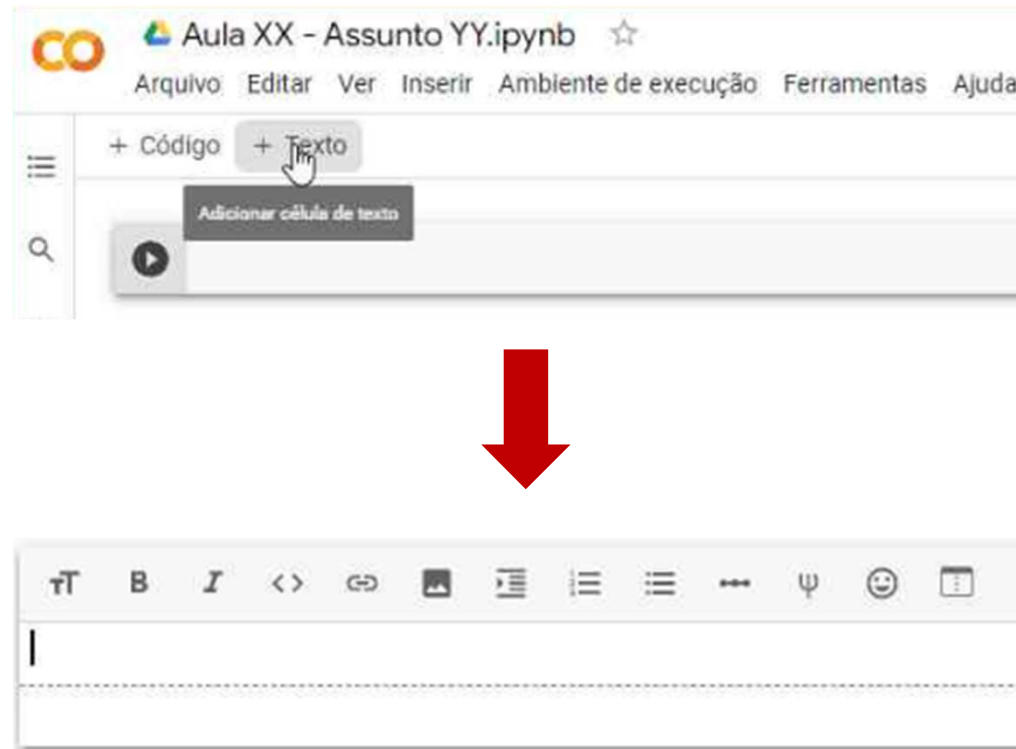
- **Código:**



Podemos executar códigos em python
pressionando esse botão

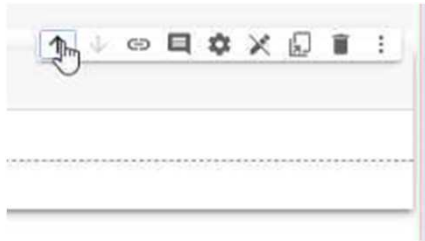
Editando um Jupyter Notebook

Para adicionar um texto podemos pressionar o botão **+Texto**



Editando um Jupyter Notebook

No canto superior podemos mover as diversas células do nosso documento:



Clique duas vezes (ou pressione "Enter") para editar

[]

Alteramos a ordem da célula de texto e da célula de código no documento.

Editando um Jupyter Notebook

Clique duas vezes (ou pressione "Enter") para editar

[]

Clique duas vezes na célula de texto



Rich text editor toolbar with icons for bold, italic, code, link, unlink, image, list, and table.

Isso é um título de nível 1

Isso é um título de nível 1

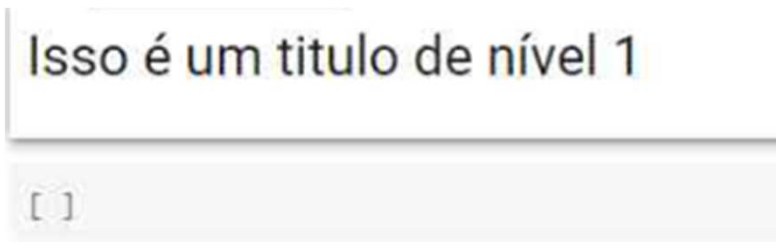
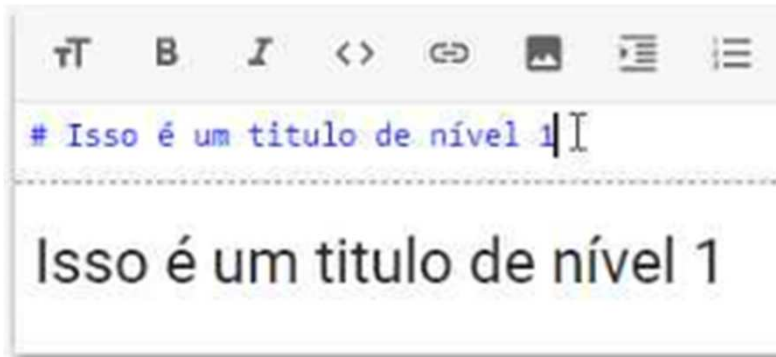
Com o # podemos criar títulos em diferentes níveis:

#: Primeiro nível

##: Segundo nível

###: Terceiro nível

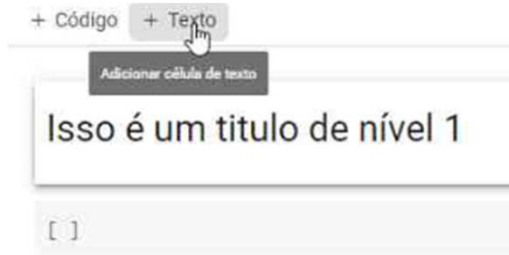
Editando um Jupyter Notebook



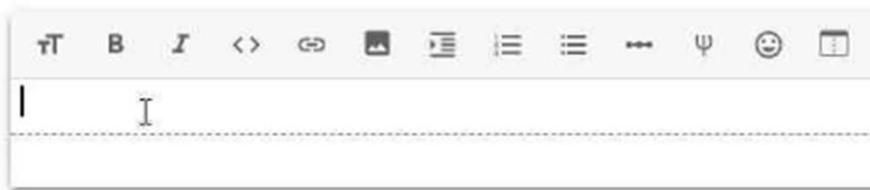
Pressione Esc ao terminar a edição do texto

Editando um Jupyter Notebook

Vamos adicionar mais uma célula:



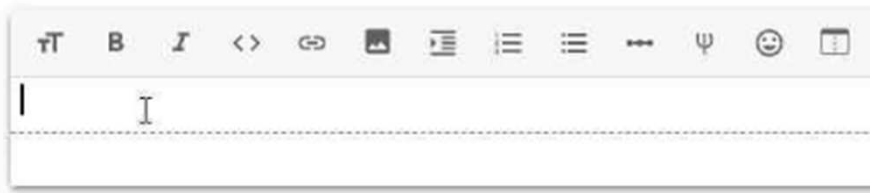
▼ Isso é um titulo de nível 1



Repare que o texto será adicionado no local que você selecionar na tela

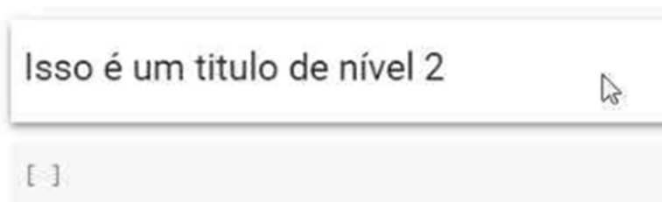
Editando um Jupyter Notebook

▼ Isso é um titulo de nível 1



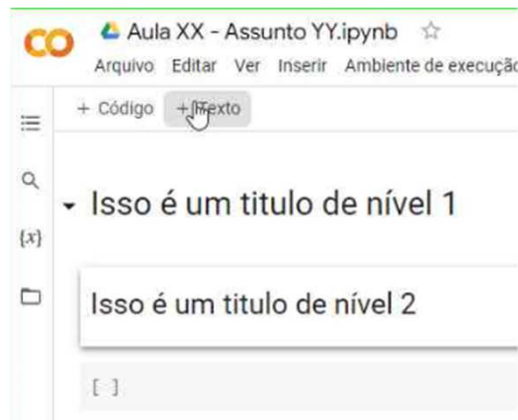
Digitar "## Isso é um título nível 2"

▼ Isso é um titulo de nível 1

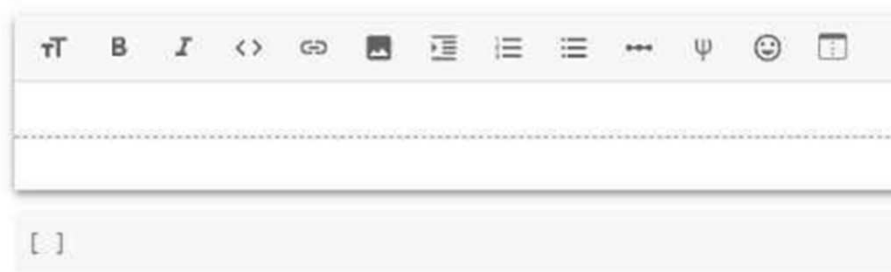


Editando um Jupyter Notebook

Vamos adicionar mais uma célula de texto:



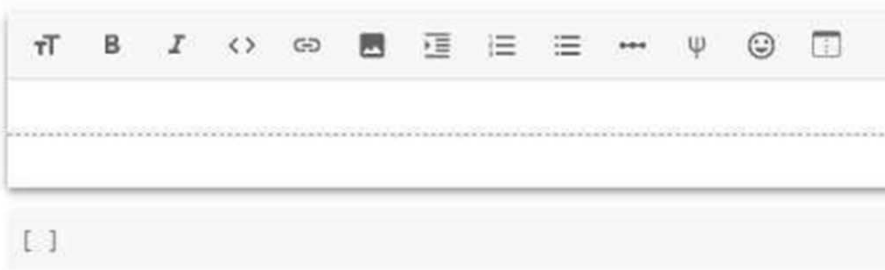
- ▼ Isso é um título de nível 1
- ▼ Isso é um título de nível 2



Editando um Jupyter Notebook

▼ Isso é um título de nível 1

▼ Isso é um título de nível 2



▼ Isso é um título de nível 1

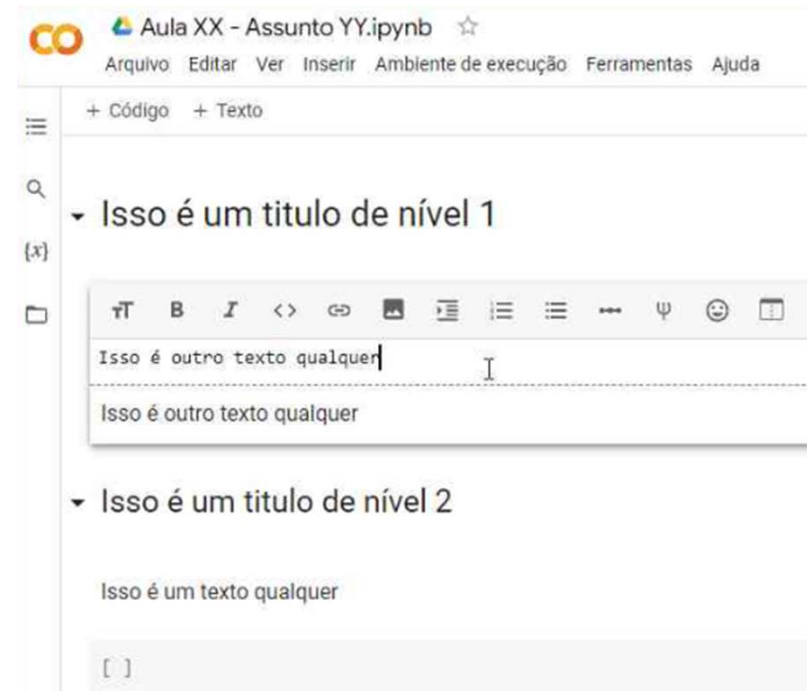
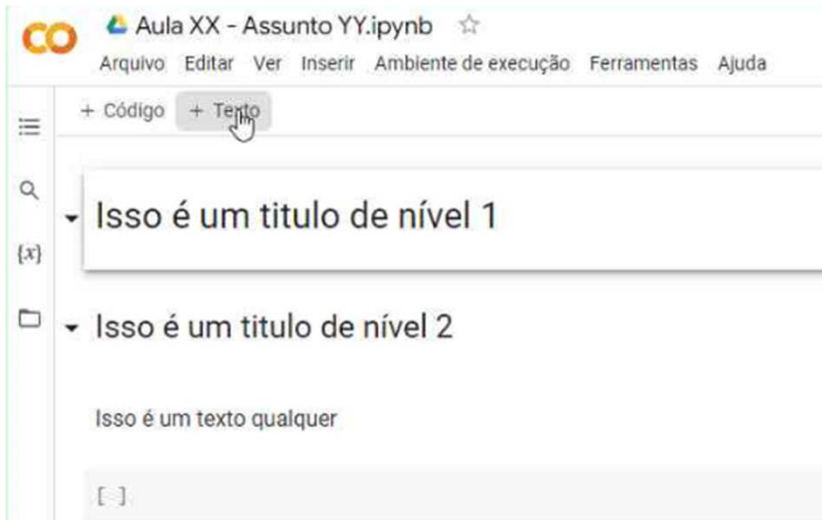
▼ Isso é um título de nível 2

Isso é um texto qualquer

Adicione um texto sem usar o "#" para inserir um texto que não seja um título

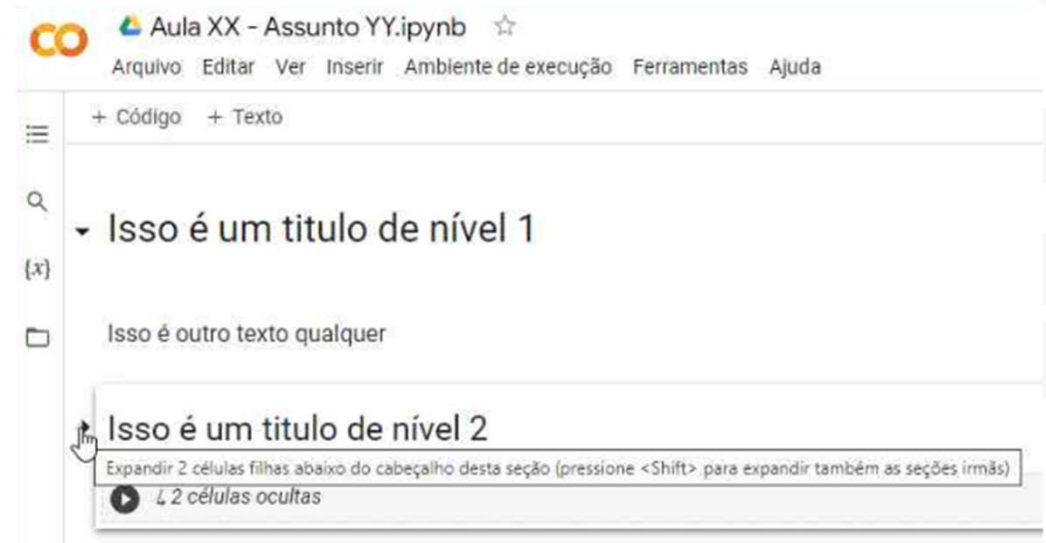
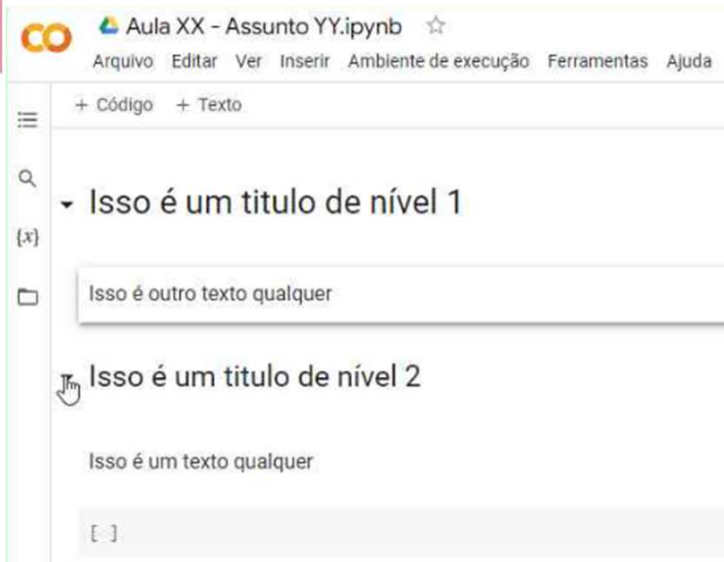
Editando um Jupyter Notebook

Clique no **"Isso é um título de nível 1"** e depois clique em **+Texto**



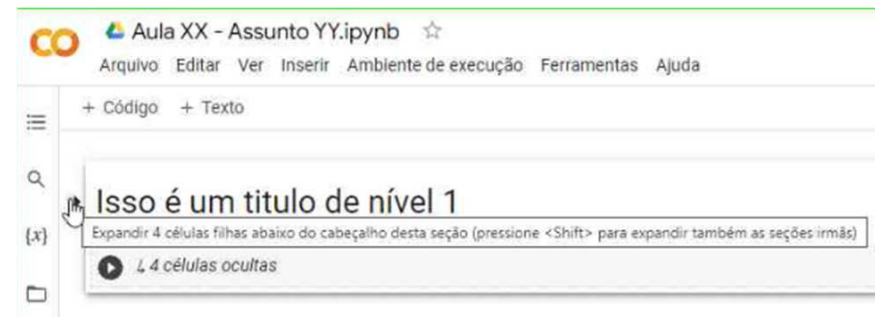
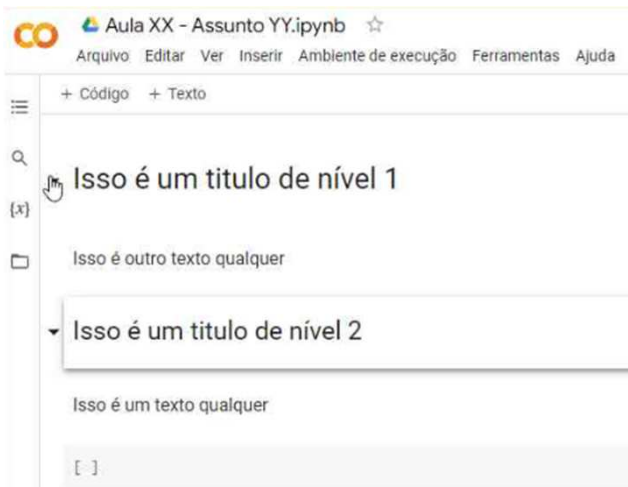
Editando um Jupyter Notebook

Podemos unir as células de acordo com seus títulos:



Editando um Jupyter Notebook

Podemos juntar o título de nível 1 e ele irá unir todos os títulos de nível 2 abaixo dele.



Editando um Jupyter Notebook

Inserir o código abaixo:

▼ Isso é um título de nível 1

Isso é outro texto qualquer

▼ Isso é um título de nível 2

Isso é um texto qualquer



▼ Isso é um título de nível 1

Isso é outro texto qualquer

▼ Isso é um título de nível 2

Isso é um texto qualquer



Editando um Jupyter Notebook

Clicar em **play** na célula de código. Note que o status do Google Colab será Conectando. Nesse instante ele está se conectando com uma máquina do Google.

▼ Isso é um título de nível 1

Isso é outro texto qualquer

▼ Isso é um título de nível 2

Isso é um texto qualquer

```
for i in range(11):  
    print(i)
```



▼ Isso é um título de nível 1

Isso é outro texto qualquer

▼ Isso é um título de nível 2

Isso é um texto qualquer

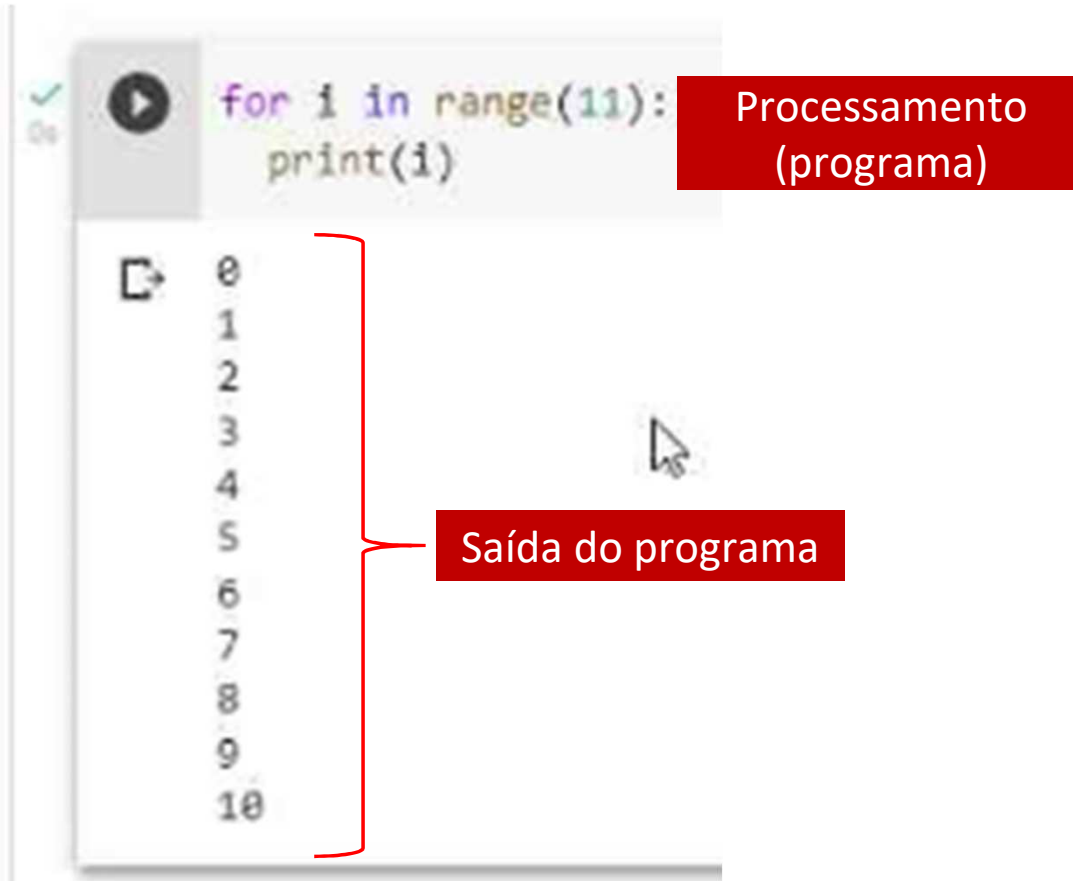
```
for i in range(11):  
    print(i)
```

Conectar ▼



... Conectando ▼

Editando um Jupyter Notebook



The image shows a Jupyter Notebook interface. At the top, there is a code cell with the following Python code:

```
for i in range(11):  
    print(i)
```

Below the code cell, the output is displayed as a list of numbers from 0 to 10, each on a new line. A red bracket on the right side of the output list groups the numbers from 0 to 10. A mouse cursor is visible over the output area.

Processamento (programa)

Saída do programa

Editando um Jupyter Notebook

No Colaboratory podemos agrupar as células a partir de seus títulos



Propriedade dos Dados

Propriedade dos Dados

Podemos classificar os dados como:

- *Estruturados;*
- *Semiestruturados*
- *Não estruturados.*

Propriedade dos Dados

*Dados **estruturados** obedecem a um esquema fixo, portanto, todos os dados têm os mesmos campos ou propriedades.*

Dados Tabulares:

Customer

CustomerID	Title	FirstName	MiddleName	LastName	Suffix	CompanyName	Phone
1	Mr.	Orlando	N.	Gee	NULL	A Bike Store	245-555-0173
2	Mr.	Keith	NULL	Harris	NULL	Progressive Sports	170-555-0127
3	Ms.	Donna	F.	Cameras	NULL	Advanced Bike Components	279-555-0130
4	Ms.	Janet	M.	Gates	NULL	Modular Cycle Systems	710-555-0173
5	Mr.	Lucy	NULL	Hamington	NULL	Metropolitan Sports Supply	828-555-0186
6	Ms.	Rosmarie	J.	Carroll	NULL	Aerobic Exercise Company	244-555-0112
7	Mr.	Dominic	P.	Gash	NULL	Associated Bikes	192-555-0173

Dados Tabulares:

Product

ProductID	Name	ProductNumber	Color	StandardCost	ListPrice	Size	Weight
680	HL Road Frame - Black, 58	FR-R92B-58	Black	1059.31	1431.50	58	1016.04
706	HL Road Frame - Red, 58	FR-R92R-58	Red	1059.31	1431.50	58	1016.04
707	Sport-100 Helmet, Red	HL-U509-R	Red	13.0863	34.99	NULL	NULL
708	Sport-100 Helmet, Black	HL-U509	Black	13.0863	34.99	NULL	NULL
709	Mountain Bike Socks, M	SO-B909-M	White	3.3963	9.50	M	NULL
710	Mountain Bike Socks, L	SO-B909-L	White	3.3963	9.50	L	NULL
711	Sport-100 Helmet, Blue	HL-U509-B	Blue	13.0863	34.99	NULL	NULL
712	AWC Logo Cap	CA-1098	Multi	6.9223	8.99	NULL	NULL
713	Long-Sleeve Logo Jersey, S	LJ-0192-S	Multi	38.4923	49.99	S	NULL

Propriedade dos Dados

*Dados **semiestruturados** são informações que têm alguma estrutura, mas que permitem alguma variação entre instâncias da entidade. Um formato comum para dados semiestruturados é o JSON (JavaScript Object Notation)*

```
// Customer 1
{
  "firstName": "Joe",
  "lastName": "Jones",
  "address":
  {
    "streetAddress": "1 Main St.",
    "city": "New York",
    "state": "NY",
    "postalCode": "10099"
  },
  "contact":
  [
    {
      "type": "home",
      "number": "555 123-1234"
    },
    {
      "type": "email",
      "address": "joe@litware.com"
    }
  ]
}
```

```
// Customer 2
{
  "firstName": "Samir",
  "lastName": "Nadoy",
  "address":
  {
    "streetAddress": "123 Elm Pl.",
    "unit": "500",
    "city": "Seattle",
    "state": "WA",
    "postalCode": "98999"
  },
  "contact":
  [
    {
      "type": "email",
      "address": "samir@northwind.com"
    }
  ]
}
```


Propriedade dos Dados

Nem todos os dados são estruturados ou até mesmo semiestruturados. Por exemplo, **documentos, imagens, dados de áudio e vídeo e arquivos binários** podem não ter uma estrutura específica. Esse tipo de dados é conhecido como dados ***não estruturados***.

Dear Joe,

Thank you for ordering your hardware supplies from our online store (order number 1000) on 1/1/2022.

Your order has been shipped and should arrive in 3-5 business days.

Contoso Hardware

Our products are of the highest quality and used by professionals.

We have amazing screwdrivers, that are really useful for tightening and loosening screws.



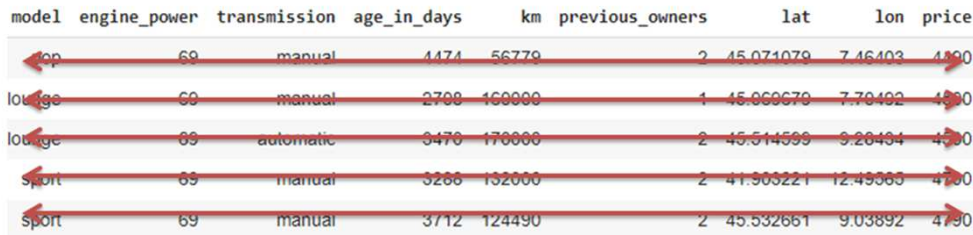
We also have wrenches (or, if you prefer, spanners)...



Propriedade dos Dados

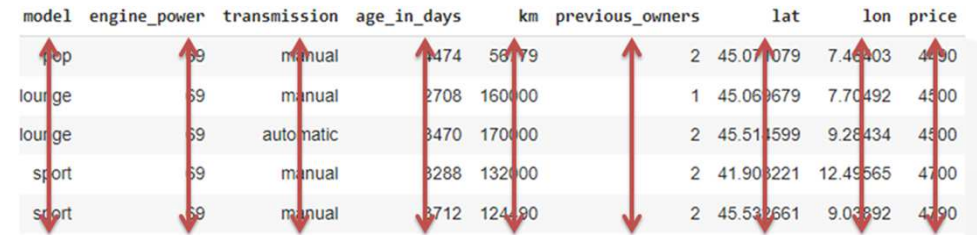
No nosso curso iremos focar em dados estruturados em forma de uma tabela, onde as **linhas** representam diferentes exemplos de um conjunto de dados e as **colunas** representam diferentes características de cada exemplo

model	engine_power	transmission	age_in_days	km	previous_owners	lat	lon	price
pop	69	manual	4474	56779	2	45.071079	7.46403	4490
lounge	69	manual	2708	160000	1	45.069679	7.70492	4500
lounge	69	automatic	3470	170000	2	45.514599	9.28434	4500
sport	69	manual	3288	132000	2	41.903221	12.49565	4700
sport	69	manual	3712	124490	2	45.532661	9.03892	4790



model	engine_power	transmission	age_in_days	km	previous_owners	lat	lon	price
pop	69	manual	4474	56779	2	45.071079	7.46403	4490
lounge	69	manual	2708	160000	1	45.069679	7.70492	4500
lounge	69	automatic	3470	170000	2	45.514599	9.28434	4500
sport	69	manual	3288	132000	2	41.903221	12.49565	4700
sport	69	manual	3712	124490	2	45.532661	9.03892	4790

Cada linha representa um carro diferente



model	engine_power	transmission	age_in_days	km	previous_owners	lat	lon	price
pop	69	manual	4474	56779	2	45.071079	7.46403	4490
lounge	69	manual	2708	160000	1	45.069679	7.70492	4500
lounge	69	automatic	3470	170000	2	45.514599	9.28434	4500
sport	69	manual	3288	132000	2	41.903221	12.49565	4700
sport	69	manual	3712	124490	2	45.532661	9.03892	4790

Cada coluna representa uma característica diferente

Exercício

Vamos juntos aprender a carregar dados no Google Colaboratory!

Seções:

Carregando arquivos de dados csv no Colaboratory:

Propriedade dos Dados

Dados quantitativos:

Dados quantitativos consiste de valores numéricos como:



Altura
Ex.: 1,87m



Medida Corporal
Ex.: Cintura 110cm



Peso(kg)
Ex.: Peso = 105kg



idade
Ex.: idade = 31 anos

Propriedade dos Dados

Dados quantitativos:

Dados quantitativos podem ser incorporados **diretamente** em formulas algébricas e modelos matemáticos e podemos criar gráficos com eles.

Propriedade dos Dados

Dados Categóricos:

Dados categóricos usualmente consistem de rótulos usados para descrever propriedades dos objetos investigados. Ex:



Mulher cis



Homem cis

...



Ruivo(a)



Loiro(a)



Cabelo cinza



Careca

...

Propriedade dos Dados

Dados Categóricos:

Dados categóricos poder ser codificados para números:

Por exemplo:



Mulher cis = 0

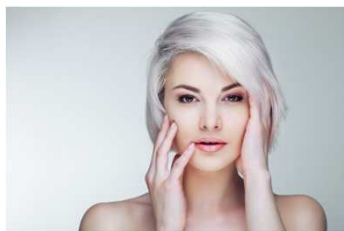


Homem cis = 1

...



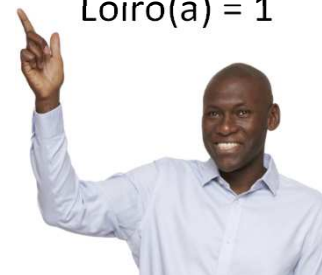
Ruivo(a) = 0



Cabelo cinza = 2



Loiro(a) = 1



Careca = 3

...

Propriedade dos Dados

Dados Categóricos:

No entanto, não podemos realmente tratar esses valores como números, para nada além de um simples teste de identidade.

Faz sentido fazer uma análise do valor máximo e mínimo da cor de cabelo?

Qual a interpretação do estilo de cabelo -1?

Futuramente iremos verificar como podemos trabalhar e interpretar valores categóricos

Propriedade dos Dados

- Dados brutos, ou dados que estão na sua forma original, precisam de uma forma para serem representados para que posteriormente seja possível aplicar ciência de dados.



Propriedade dos Dados

- Dados brutos, ou dados que estão na sua forma original, precisam de uma forma para serem representados para que posteriormente seja possível aplicar ciência de dados.
- O tratamento dos dados também será abordado na disciplina

Exercício

Vamos juntos aprender a carregar dados no Google Colaboratory!

Seções:

Comandos básicos de visualização de dados (Pandas)

Referências Bibliográficas

IGUAL, Laura; SEGUÍ, Santi. Introduction to Data Science: a python approach to concepts, techniques and applications. Gewerbestrasse: Springer, 2017.

SKIENA, Steven S.. Data Science Design Manual. Gewerbestrasse: Springer, 2017.