

# LIMPEZA DE DADOS

# AGENDA

01

**Introdução**

02

**Estatística**

03

**Análise Descritiva**

04

**Testes de Hipóteses**

05

**Limpeza de dados**





# Limpeza de Dados

## 05 Limpeza de dados

---

Primeiro de tudo, esse treinamento é **mais focado no ferramental** necessário para que se possa **analisar dados**, incluindo principalmente **ferramentas de Excel avançado**, ignorando um pouco o Python pois este já foi passado em **outro treinamento** e tem o *script* explicado.

Alguns **comandos importantes** são necessários para **alterar dados**, são eles as **funções**: `se()`, `filtro()`, `seerro()`, `procv()`, `proch()`, `índice()`, `corresp()`, `média()`, `médiase()`, `somase()`, `cont.valores()`, `cont.se()`, `cont.ses()`, `esquerda()`, `direita()`, `arrumar()`.

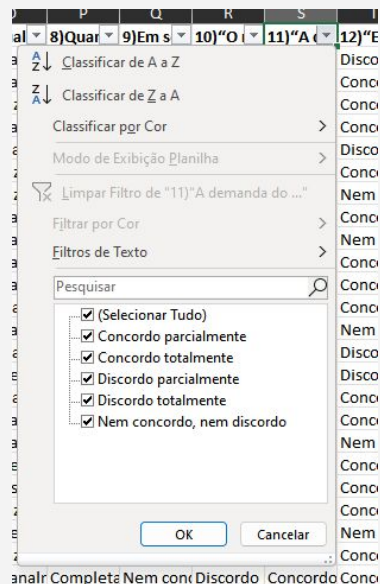
Outras **ferramentas** também são abordadas para **alterar os dados no Excel**, que são: **filtro**, **tabela**, **tabela dinâmica**, **texto para colunas**, **preenchimento relâmpago (Ctrl + E)**, **remover duplicados**, **validação de dados** e **Power Query**. O Power Query será abordado em outro treinamento.





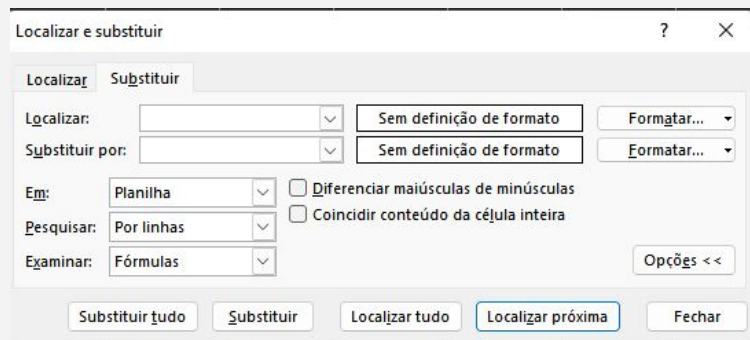
## 05 Limpeza de dados

**Filtro (Ctrl + Shift + L):** é uma das **funções mais importantes do Excel** e auxilia muito na análise de dados. Em uma tabela, normalmente pode-se **criar filtros** para todas as **variáveis** para alterar os dados dela, observando **quais os valores únicos** presentes e também podendo **ordenar os elementos na coluna**, além de filtrá-los.



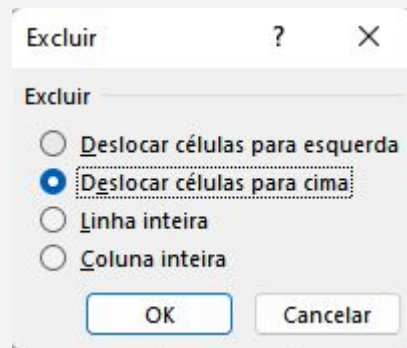
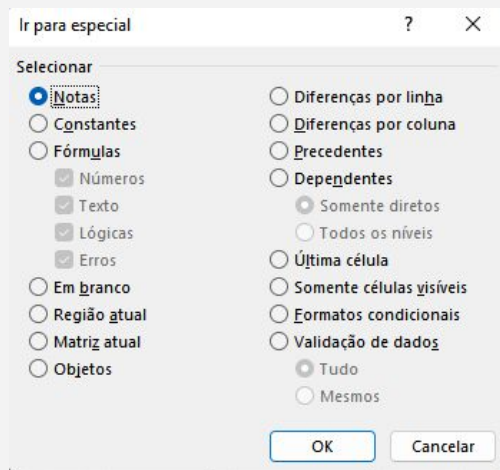
## 05 Limpeza de dados

**Localizar e Substituir:** esse comando é **bem instrutivo** e bem **importante** para a **limpeza de dados**, sua função é **substituir algum termo por outro** e pode ser usado em várias coisas, por exemplo **remover alguns espaços indesejados** da base (o código do Python de clusterização também faz isso automaticamente). É a função para substituir os **“Concordo totalmente” por 5** por exemplo.



## 05 Limpeza de dados

**Ir para especial:** tem **funções parecidas do que o localizar**, mas tem alguns objetivos diferentes. O **principal uso** dessa ferramenta é **remover células em branco** de uma tabela. Por exemplo, se quiser **remover células em branco**, vá para **Ir para especial** -> **Em branco** -> **Botão direito** -> **Excluir** -> **Deslocar células para cima**.



## 05 Limpeza de dados

---

**=SE(teste\_lógico; valor\_se\_verdadeiro: valor\_se\_falso):** a **função SE** primeira testa **algum argumento lógico** e você diz o que **ele deve fazer caso o teste seja verdadeiro** e caso seja **falso**. Nesse sentido, ele é muito bom para diversas situações, até mesmo para **alterar as bases de dados**.

**=FILTRO(matriz; incluir; [se\_vazia]):** essa função serve para **filtrar uma tabela** e, por isso, também **retorna uma tabela**. Primeiro, você **insere a tabela dentro de matriz** e dá a **condição** para a função na parte de incluir, por exemplo **=FILTRO(C1:D6;D1:D6 > 4)** filtra a tabela apenas com **valores acima de 4** para a variável na coluna D.

**=SEERRO(valor; valor\_se\_erro):** essa função **também é muito parecida com a função se**, porém você diz um valor para ela, normalmente uma **função** que pode dar erro e, se der erro, você também diz o que **ela deve substituir caso contrário**.





## 05 Limpeza de dados

---

**=PROCV(valor\_procurado; matriz\_tabela; núm\_índice\_tabela):** essa é uma das **funções mais conhecidas** e serve para **cruzar dados** entre tabelas. Primeiro, quando você tem uma tabela e **quer procurar algum valor específico** nela, você **diz primeiro o valor** que está correlacionado com aquele que você procura, depois a **matriz** que estão todos os valores e depois o **índice da variável procurada** (possível colocar **=CORRESP()** aqui).

**=PROCH(valor\_procurado; matriz\_tabela; núm\_índice\_tabela):** é muito parecido com o **procv**, mas é utilizada para **cruzar dados na horizontal**, em que o número índice é uma das linhas da tabela.

**=ÍNDICE(matriz; núm\_linha; núm\_coluna):** a **função índice** retorna o **elemento de linha e coluna a serem definidos**. Ela funciona muito bem com o **corresp** que diz a posição de algum termo.



## 05 Limpeza de dados

---

**=CORRESP(valor\_procurado; matriz\_procurada):** é uma função que **casa muito com o procv e com o índice**, pois diz a **exata posição** de algum valor procurado dentro de uma matriz. Nesse sentido, pode utilizar o corresp dentro de **núm\_índice\_tabela** no **procv** para **automatizar a entrada** (quando o valor do índice muda) ou usar índice e corresp.

**=ESQUERDA(texto; núm\_caract):** retorna os **n caracteres** à **esquerda** do texto. Serve muito para **alterar textos** para uma outra forma.

**=DIREITA(texto; núm\_caract):** retorna os **n caracteres** à **direita** do texto. Também é uma função muito boa para **alterar textos**.

**=ARRUMAR(texto):** função muito importante, pois ela **retira todos os espaços em branco** de um texto (muitas vezes em pesquisa quanti estão cheios de espaços). Também podem ser removidos por **Localizar e Substituir**.

---



## 05 Limpeza de dados

---

**=MÉDIA(matriz):** retorna a **média de uma sequência de números**.

**=MÉDIASE(range; critério):** retorna a **média de uma distribuição** de números porém dá um **critério** para cada um dos números, por exemplo, **médiase(D1:D9; "> 5")** calcula a média de um intervalo **desconsiderando números menores que 5**

**=SOMASE(intervalo; critério):** **soma** os números porém com **algum critério** para cada um dos números, do mesmo modo que ocorre a **médiase**.

**=CONT.VALORES(matriz):** **conta quantos valores não vazios** estão na base de dados. Também há um **cont.se** que diz também o **critério** para contar os valores.

**=SOMASES(intervalo\_soma; intervalo\_critério1; critério1; ...):** é uma **soma com mais restrições para colocar**, primeiro colocando o **intervalo a ser somado e depois critérios**.



## 05 Limpeza de dados

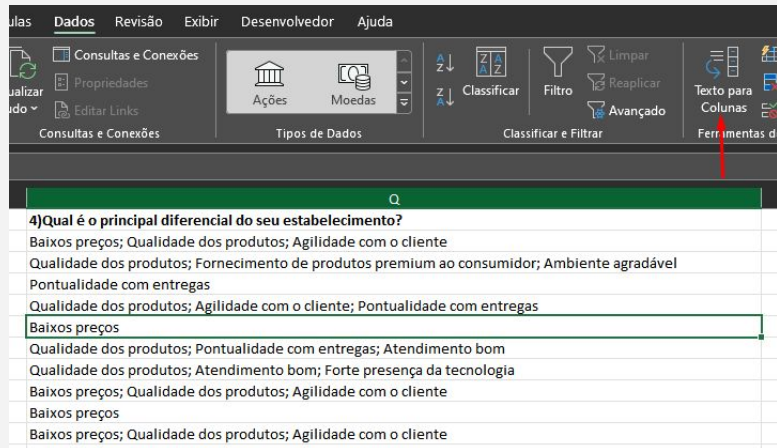
**Preenchimento Relâmpago (Ctrl + E):** é uma função muito importante no Excel que **ajuda na automatização de preenchimento de planilhas**. Nesse caso, você diz um padrão com base nos outros dados da tabela e aperta **Ctrl + E** ou vai em **Dados -> Preenchimento Relâmpago** que ele completa todos os outros dados. Isso é muito bom para **limpar dados** sem o uso de fórmulas.

A	B	C	D
Nome	Nascimento	Mês Nascimento	Ano Nascimento
Esthevão	03/ago	ago	2002
Enzo	12/set	set	2002
Manô	22/out		
Kike	01/dez		
Deco	10/jan		
Paraíba	19/fev		
Beto	31/mar		
Ana	10/mai		
Carol	19/jun		
Manu	29/jul		
Jorge	07/set		
Rafa	17/out		
Valen	26/nov		

A	B	C	D
Nome	Nascimento	Mês Nascimento	Ano Nascimento
Esthevão	03/ago	ago	2002
Enzo	12/set	set	2002
Manô	22/out	out	2002
Kike	01/dez	dez	2002
Deco	10/jan	jan	2002
Paraíba	19/fev	fev	2002
Beto	31/mar	mar	2002
Ana	10/mai	mai	2002
Carol	19/jun	jun	2002
Manu	29/jul	jul	2002
Jorge	07/set	set	2002
Rafa	17/out	out	2002
Valen	26/nov	nov	2002

## 05 Limpeza de dados

**Texto para colunas:** é uma **ferramenta** que **separa uma coluna** em várias bifurcando o texto presente. Vá em **“Dados” -> “Texto para colunas” -> “delimitado”** (para dizer algum separador que deve separar os textos) e seleciona qual esse **delimitador** (**ponto e vírgula, espaço, barra** são os mais comuns, também possível colocar o desejado em “outros”), dando **ok** depois (é importante não ter colunas do lado para ele conseguir separar).



Q	R	S	T
4)Qual é o principal diferencial do seu estabelecimento?			
Baixos preços	Qualidade dos	Agilidade com o cliente	
Qualidade dos produtos	Fornecimento c	Ambiente agradável	
Pontualidade com entregas			
Qualidade dos produtos	Agilidade com c	Pontualidade com entrega	
Baixos preços			
Qualidade dos produtos	Pontualidade c	Atendimento bom	
Qualidade dos produtos	Atendimento b	Forte presença da tecnolo	
Baixos preços	Qualidade dos	Agilidade com o cliente	
Baixos preços	Qualidade dos	Fornecimento de produtos	
Baixos preços	Agilidade com c	Ambiente agradável	
Qualidade dos produtos	Ambiente agrac	Atendimento bom	
Baixos preços			
Baixos preços			
Meu estabelecimento não possui diferenciais			
Baixos preços	Qualidade dos	Atendimento bom	
Baixos preços	Qualidade dos	Ambiente agradável	
Baixos preços	Atendimento b	Meu estabelecimento não	



## 05 Limpeza de dados

**Remover duplicadas:** outra **ferramenta muito importante** para descobrir quais os elementos presentes em uma tabela. O **filtro** funciona muito bem nesses casos, mas muitas vezes é necessário **saber quais os elementos únicos nessa tabela**.

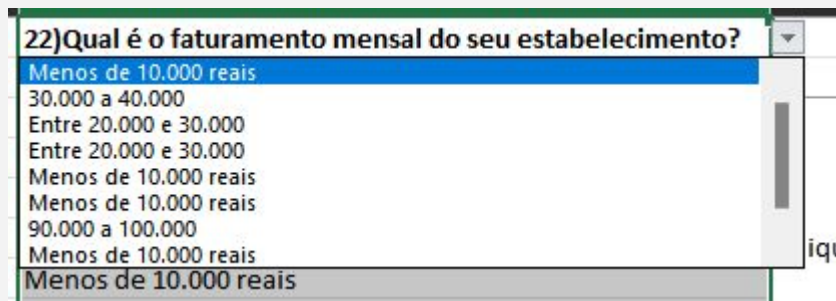
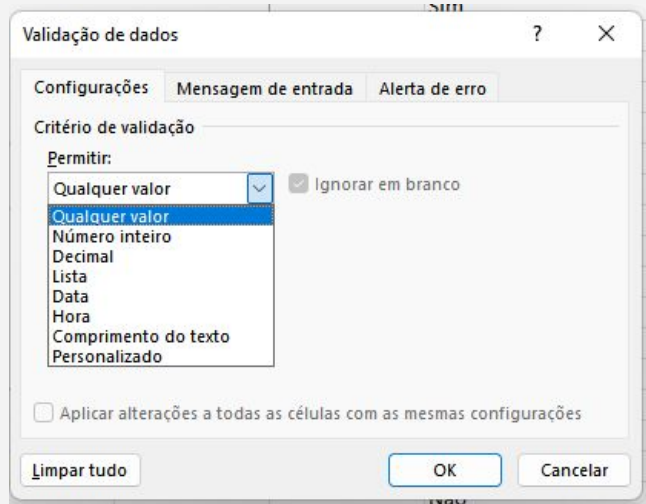
Depois de **removidas** as **duplicadas**, é possível **juntar isso com um cont.se** para saber **quantos valores têm em cada elemento**. Essa é uma das funções mais utilizadas, visto que para todas as **variáveis não clusterizáveis** é bom ter uma noção macro de distribuição.

Q	R	S	T
4)Qual é o principal diferencial do seu estabelecimento?			
Baixos preços	Qualidade dos p	Agilidade com o cliente	
Qualidade dos produtos	Fornecimento c	Ambiente agradável	
Pontualidade com entregas			
Qualidade dos produtos	Agilidade com c	Pontualidade com entrega	
Baixos preços			
Qualidade dos produtos	Pontualidade c	Atendimento bom	
Qualidade dos produtos	Atendimento b	Forte presença da tecnologia	
Baixos preços	Qualidade dos p	Agilidade com o cliente	
Baixos preços			
Baixos preços	Qualidade dos p	Agilidade com o cliente	
Baixos preços	Qualidade dos p	Fornecimento de produtos	
Qualidade dos produtos	Agilidade com c	Ambiente agradável	
Baixos preços	Ambiente agrac	Atendimento bom	
Baixos preços			
Meu estabelecimento não possui diferenciais			
Baixos preços	Qualidade dos p	Atendimento bom	
Baixos preços	Qualidade dos p	Ambiente agradável	
Ambiente agradável	Atendimento b	Meu estabelecimento não	

Q
4)Qual é o principal diferencial do seu estabelecimento?
Baixos preços
Qualidade dos produtos
Pontualidade com entregas
Meu estabelecimento não possui diferenciais
Ambiente agradável
Fornecimento de produtos premium ao consumidor
Atendimento bom
Agilidade com o cliente

## 05 Limpeza de dados

**Validação de dados:** essa é uma ferramenta de **formatação das células**. Normalmente, essa **validação de dados** normalmente é utilizada para **transformar células** em listas que serão preenchidas pelo cliente depois. É muito semelhante com a **caixa de seleção do VBA**, porém no próprio Excel.



## 05 Limpeza de dados

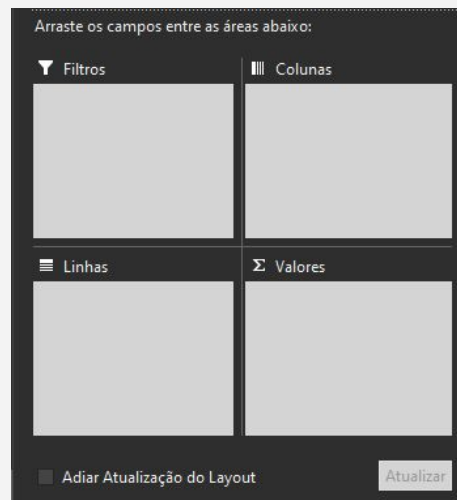
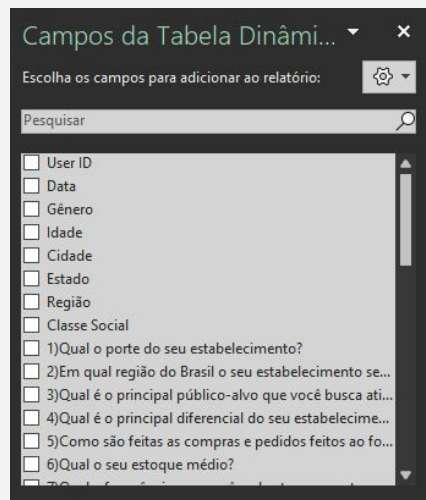
**Formatar como tabela:** é algo **muito importante no Excel**, pois facilita a **manipulação de dados**. Depois de **formatar como tabela**, é possível, por exemplo, **adicionar novas linhas** na tabela mantendo a formatação da tabela como um todo. Outro fator importante é que a **tabela puxa também a fórmula usada nas outras linhas** também para a nova linha da tabela, como mostrado abaixo. Ela permite criar **gráficos automáticos** que mudam quando mudar o filtro de alguma coluna.

Nome	=ESQUERDA("A1"; 3)
Esthevão	Est
Enzo	Enz
Manô	Man
Kike	Kik
Deco	Dec
Paraíba	Par
Beto	Bet
Ana	Ana
Carol	Car
Manu	Man
Jorge	Jor
Rafa	Raf
Valen	Val

Nome	=ESQUERDA("A1"; 3)
Esthevão	Est
Enzo	Enz
Manô	Man
Kike	Kik
Deco	Dec
Paraíba	Par
Beto	Bet
Ana	Ana
Carol	Car
Manu	Man
Jorge	Jor
Rafa	Raf
Valen	Val
Eduardo	Edu

## 05 Limpeza de dados

**Tabela dinâmica:** também muito importante para **analisar dados**, foi explicada no **treinamento de dashboard** mas ela também tem muitas funções para a análise em **pesquisas quantitativas**. Após formatar como tabela, vá em **Design da Tabela -> Resumir em Tabela Dinâmica -> Planilha Existente** e escolher uma célula em uma outra aba.



## 05 Limpeza de dados

**Tabela dinâmica:** a **tabela** tem dois principais campos, um com todas as **variáveis** e outro com **4 campos para criar a tabela**. Normalmente adiciona-se variáveis em **linhas, colunas e valores** para poder se analisar como estão distribuídas, além de **criar relações** entre variáveis. Ela faz o processo de **remover duplicatas e cont.se** de um jeito automático mas nem sempre é possível fazê-la.

**Campos da Tabela Dinâmi...** ✕

Escolha os campos para adicionar ao relatório:

Pesquisar

- ☐ User ID
- ☐ Data
- ☐ Gênero
- ☐ Idade
- ☐ Cidade
- ☐ Estado
- ☐ Região
- ☐ Classe Social
- ☐ 1)Qual o porte do seu estabelecimento?
- ☐ 2)Em qual região do Brasil o seu estabelecimento se...
- ☐ 3)Qual é o principal público-alvo que você busca ati...
- ☐ 4)Qual é o principal diferencial do seu estabelecime...
- ☐ 5)Como são feitas as compras e pedidos feitos ao fo...
- ☐ 6)Qual o seu estoque médio?

Arraste os campos entre as áreas abaixo:

**Filtros**

**Colunas**

**Linhas**

**Valores**

Adiar Atualização do Layout

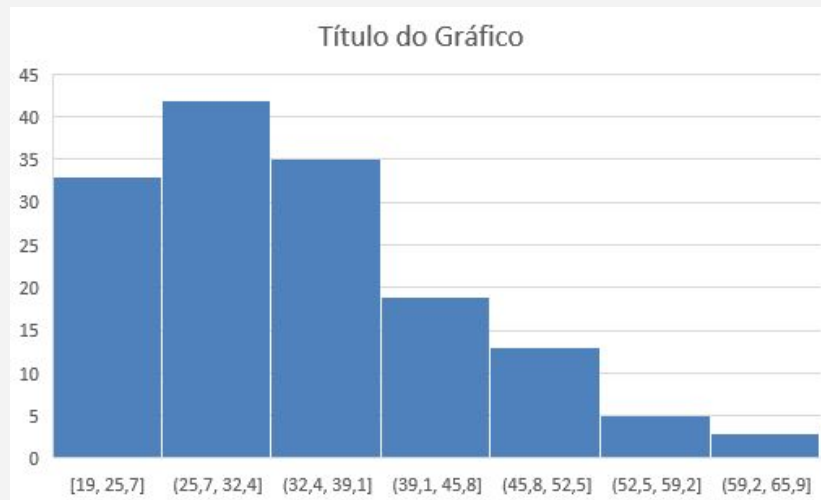
Atualizar

Contagem de Gênero			
Rótulos de Linha	Rótulos de Coluna		Total Geral
	Feminino	Masculino	
Centro Oeste	8	5	13
Nordeste	20	22	42
Norte	7	5	12
Sudeste	44	27	71
Sul	8	4	12
Total Geral	87	63	150



## 05 Limpeza de dados

**Histograma:** para **análise de dados numéricos**, é bom fazer um **histograma** para entender o **comportamento** deles. O histograma funciona como um **complemento** para os **boxplots** na análise de como estão **compartimentados**. No exemplo abaixo, é possível perceber que as **pessoas estão mais presentes em anos mais novos**.



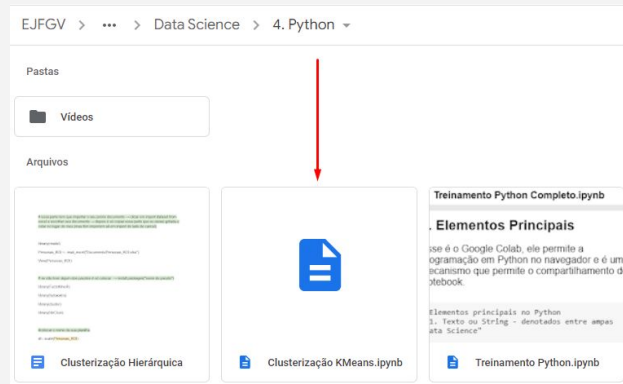
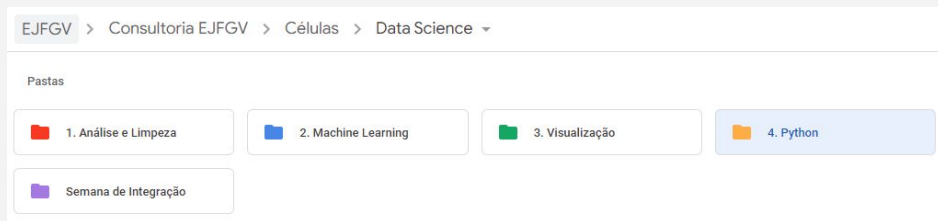


# Clusterização

## 06 Clusterização

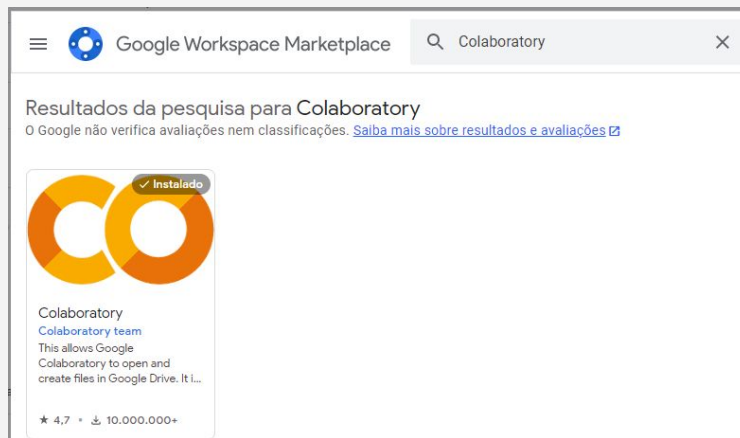
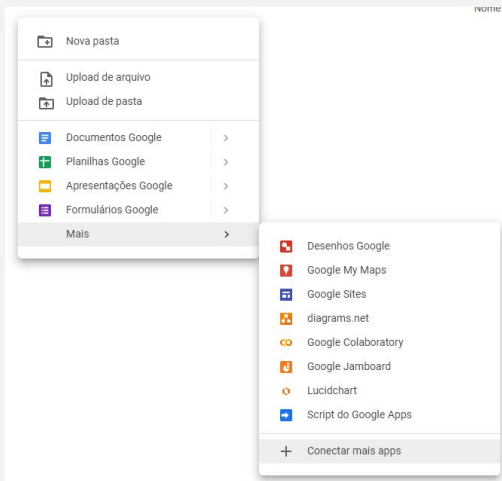
Primeiramente, deve-se abrir o **drive da célula de Data Science** e entrar em **Python** para encontrar onde está o **código da clusterização**. O código está como **“Clusterização KMeans”** e, para utilizá-lo, primeiro deve-se **criar uma cópia** dele e alterar a cópia para que outros possam usar o **código no futuro**.

Após tudo ser feito, você pode colocar o **código na pasta do seu projeto** para não ficar no **Drive** da célula.



## 06 Clusterização

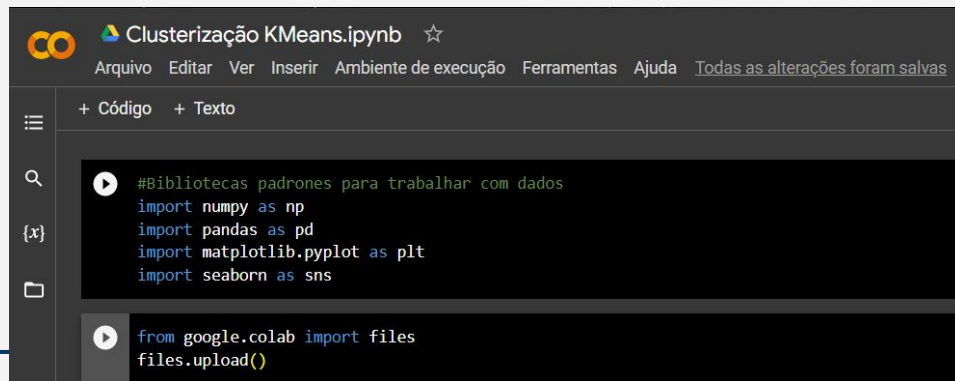
Na primeira vez que for utilizar o **Google Colab**, é importante **importá-lo para o Drive** para que seja possível utilizá-lo. Para isso, clique no **botão direito** em qualquer lugar do Drive, vá em **“Mais”** e **“Conectar mais apps”**. Depois disso, basta procurar **Colaboratory** e instalar o **Google Colab** no seu **Drive** clicando nele.



## 06 Clusterização

O **Google Colab** é uma ferramenta que permite utilizar **Python** no **navegador** e é muito útil. Algumas coisas é importante saber: o **Colab é dividido em textos e códigos** e, para criar algum dos dois, basta clicar em “+” em cada um que queira **adicionar** e **movê-lo** pelas ferramentas que estão **à direita** do texto ou código.

Para rodar algum código, basta **clicar no play** ou apertar **Ctrl + Enter** e, para rodar tudo, tudo antes do código selecionado ou tudo após, basta ir em “**Ambiente de execução**” e selecionar a opção desejada.





## 06 Clusterização

Ao rodar o código, a primeira coisa que temos que fazer é **importar o arquivo** em **“files.upload()”** e, para isso, basta ir em **“Escolher arquivos”** e selecionar o **Excel** de preferência. Após rodado, aparecerá um **texto enorme** e basta clicar no **“x”** para eliminá-lo (**não fará diferença**, esse texto é apenas para **visualização**).

Após isso, deve escrever o **nome do arquivo** depois de `read_excel` e colocar o **nome da aba a ser escolhida** (se for a primeira, pode retirar “`, sheet_name = ...`”).

```
from google.colab import files
files.upload()
```

```
[ ] dados = pd.read_excel("Cópia de Gráficos.xls", sheet name = "Respondentes Válidos")
```

## 06 Clusterização

O **próximo trabalho** a ser feito é alterar os **termos de escala Likert** que estão na lista para como está no **Excel** (deve ser escrito exatamente igual, senão o código dará erro). Isso fará com que o **Python filtre as perguntas** que são Likert e **também substitua** cada uma por **números**.

Apenas **essas perguntas serão tidas na clusterização** pois são as **variáveis numéricas** da base de dados. Depois disso, é feita uma **limpeza na base** e será importado o algoritmo de **clusterização**. Caso queira mais detalhes, há um treinamento de clusterização disponível no Drive, em **“Treinamento de Machine Learning”**.

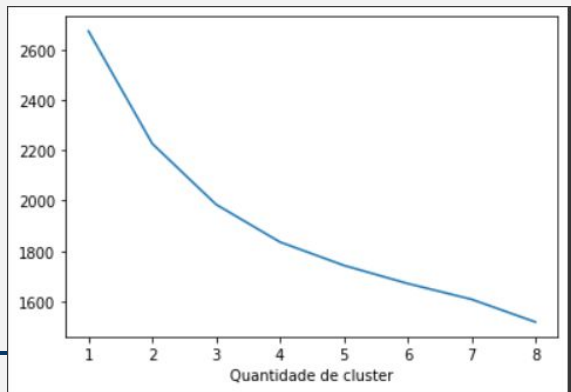


```
# Primeiro coloque as opções de resposta dentro dessa lista em aspas
# Eu indicaria copiar cada uma do Excel para o Python para ficar exatamente igual a não dar erro
# Esse código coloca em uma tabela todas as respostas desse tipo
list = ["Concordo totalmente", "Concordo parcialmente", "Nem concordo, nem discordo", "Discordo parcialmente", "Discordo totalmente"] # Deve manter a ordem
x = []
```

## 06 Clusterização

O gráfico seguinte é muito importante: ele mede a **soma da distância ao quadrado de cada dado do centro do *cluster*** para diferentes **quantidades de *clusters***. Ao que se pode perceber, o aumento do número de *clusters* diminui a distância dos dados até o centro do *cluster*, porém deve-se **escolher o número ótimo** de *clusters* com cuidado.

Em geral, o ótimo é na **menor diminuição dessa distância** (calculado pelo Python e posto no código), porém é possível também **alterar o número de *clusters*** desejados no código abaixo (normalmente, o valor escolhido pode ser **1 a mais do que o número ótimo**).



```
[19] # Essa é a quantidade ideal de clusters, escolhemos com base no menor decaimento de distância do gráfico
      # Se decair muito pouco ou o número de cluster for muito alto sem diminuir muito a distância, o número ideal é 1 (não há clusters)
      diff = []
      for i in range(0, len(wcss)-1):
          diff.append(wcss[i]-wcss[i+1])
      n = diff.index(max(diff)) + 2
      n

      2

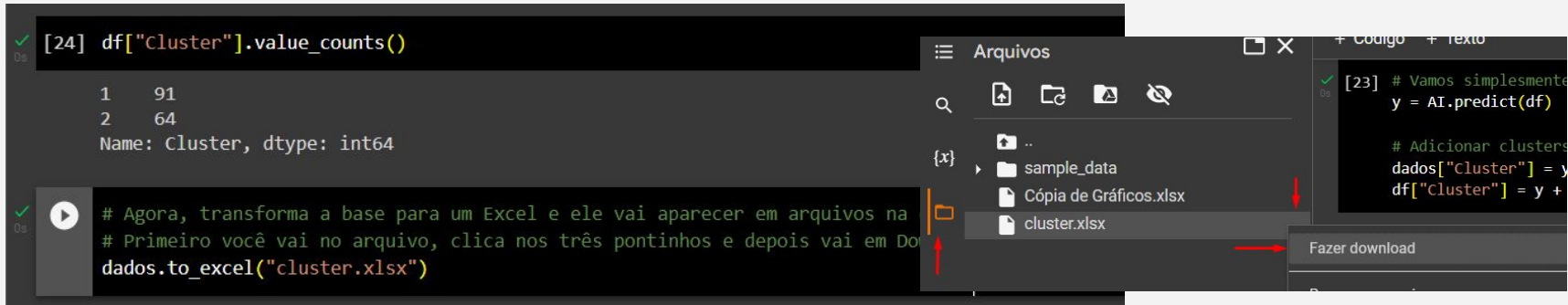
      [ ] # Criando nossa AI
          AI = KMeans(n_clusters = n)
```

← alterar n pelo número de clusters desejado

## 06 Clusterização

Por fim, o **primeiro código** abaixo mostra a **quantidade de indivíduos** por *cluster* e o de baixo é a **exportação da base** com os *clusters* para o Excel. Para **exportar o arquivo**, deve ir em **“Arquivos”**, à direita do Colab, ir em **“cluster.xlsx”**, ir nos **três pontos** e clicar em fazer **download**.

A base exportada já possui cada *cluster* que cada indivíduo pertence e permitirá a **análise quantitativa** mais facilmente por *cluster*.



The screenshot shows the Google Colab interface. On the left, a code cell [24] displays the output of `df["Cluster"].value_counts()`, showing two clusters: 1 with 91 individuals and 2 with 64 individuals. Below it, a code cell [23] shows the command `dados.to_excel("cluster.xlsx")` with a comment in Portuguese. On the right, the 'Arquivos' (Files) panel is open, showing a folder named 'sample\_data' and two files: 'Cópia de Gráficos.xlsx' and 'cluster.xlsx'. A red arrow points to the three-dot menu next to 'cluster.xlsx', and another red arrow points to the 'Fazer download' (Download) option in the dropdown menu.

```
[24] df["Cluster"].value_counts()

1    91
2    64
Name: Cluster, dtype: int64

# Agora, transforma a base para um Excel e ele vai aparecer em arquivos na
# Primeiro você vai no arquivo, clica nos três pontinhos e depois vai em Do
dados.to_excel("cluster.xlsx")

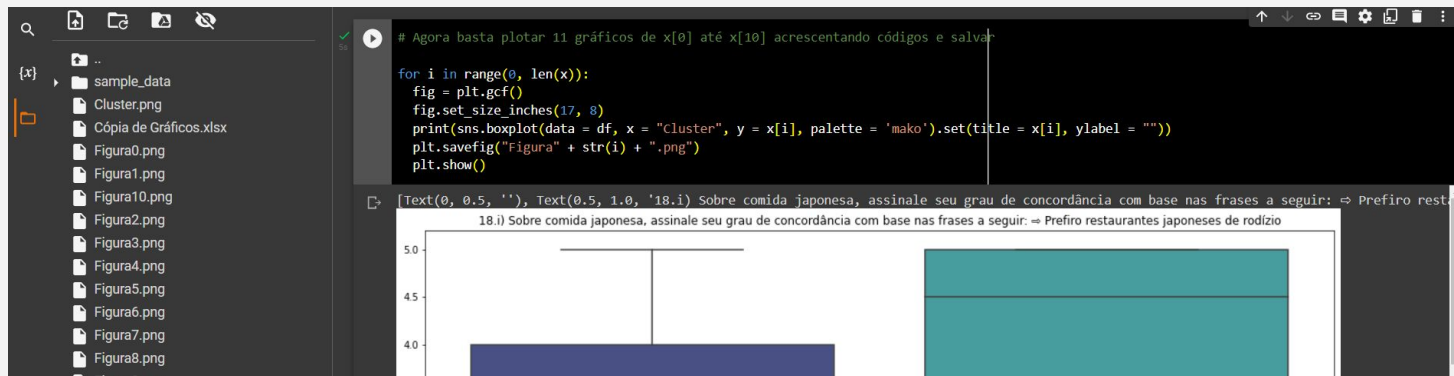
[23] # Vamos simplesmente
y = AI.predict(df)

# Adicionar clusters
dados["Cluster"] = y
df["Cluster"] = y +
```

## 06 Clusterização

Ainda, o código **também faz todos os boxplots sozinho** e deixa eles salvos na **esquerda** para **baixar**. Cada título está de acordo com a pergunta feita e os **clusters estão dispostos lado a lado** para poder comparar.

Desse modo, basta analisá-los e **atribuir uma persona a cada um**.







Obrigado!