# Aula 01 - Minicurso Python

| Property | Ítalo Lima |
| --- | --- |

## Introdução

Para aplicar os conceitos aprendidos durante a disciplina de Probabilidade e Estatística iremos utilizar algumas bases de dados distintas durante o decorrer das aulas e criar notebooks com estudos sobre as mesmas.

Para esta aula, a base de dados utilizada é retirada do site Movie Lens, onde as pessoas podem votar nos filmes disponíveis que desejarem. Lembrando que os dados em específicos que iremos utilizar são apenas um sample, ou seja: uma amostra da base de dados original.

Link para o site: https://movielens.org/

Link para os dados em CSV: https://raw.githubusercontent.com/alura-cursos/introducao-a-data-science/master/aula0/ml-latest-small/ratings.csv

## Prática

### Importando bibliotecas e dados

Para começar o estudo, iremos abrir a ferramenta Colab e importar as bibliotecas que serão utilizadas.

```
import pandas as pd
import numpy as np
```

Feito isso, queremos criar um estrutura para guardar os nossos dados, podemos fazer isso de duas formas:

```
#Forma mais detalhada e indireta
url = 'https://raw.githubusercontent.com/alura-cursos/introducao-a-data-science/master/aula0/ml-latest-small/ratings.csv'
df = pd.DataFrame()
df = pd.read_csv(url)
```

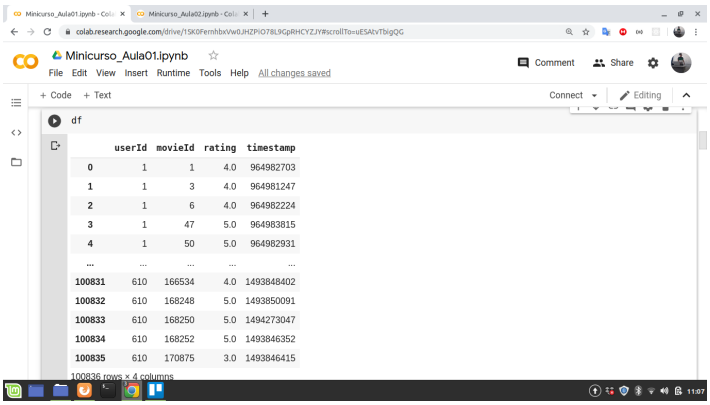
```
#Forma mais simples e direta
df = pd.read_csv('https://raw.githubusercontent.com/alura-cursos/introducao-a-data-science/master/aula0/ml-latest-small/ratings.csv'
```

Como os dados já estão em formato CSV (Formato utilizado pelas planilhas), podemos utilizar o método pd.read_csv( ) da biblioteca pandas que lê esses dados, e então passá-los para a nossa estrutura, o Data Frame.

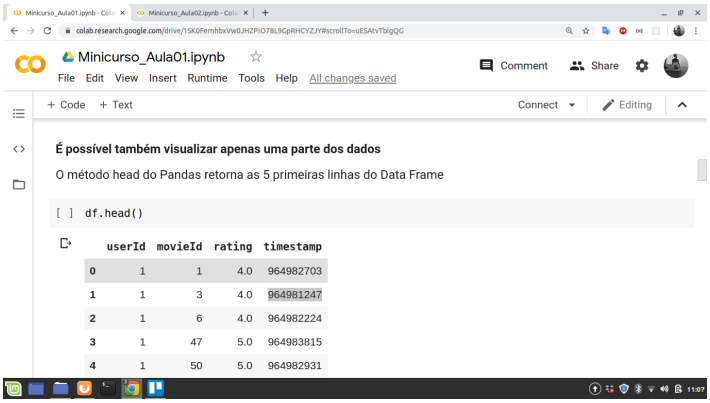
### Visualizando os dados

Agora, podemos visualizar esses dados, também é possível fazer isso de mais de uma forma.

```
#A primeira e mais simples é apenas digitar o nome respectivo do Data Frame e executar a célula
df
```

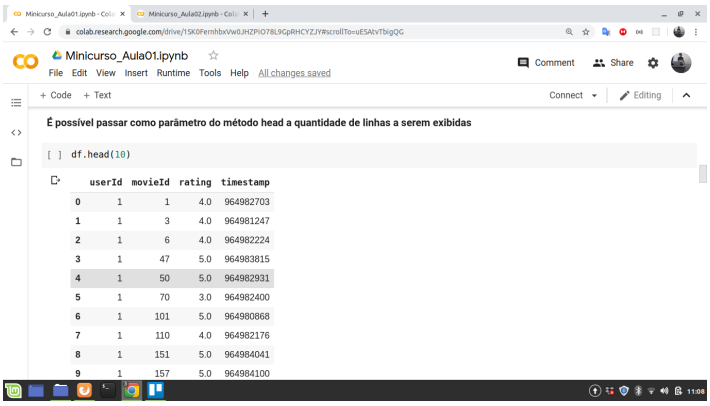



```
#A segunda é utilizar o método head() do Pandas que nos permite visualizar somente às n primerias linhas do Data Frame
df.head()
```




Por padrão, o método head( ) exibe as 5 primeiras linhas, mas podemos passar por parâmetro um número respectivo a quantidade de linhas que desejamos visualizar.

```
#Por exemplo, visualizando as 10 primeiras linhas
df.head(10)
```




## Visualizando a Estrutura dos Dados

```
df.shape
```

O método shape da biblioteca Pandas informa a estrutura do Data Frame, ou seja, a quantidade de linhas e colunas, respectivamente.

**Os dados possuem 4 colunas, são essas:**

**userId**→Id para o usuário cadastrado no site.

**movieId**→Id do filme, baseado em outros dados que contém a lista de filmes do site.

**rating**→ Notas atribuídas por um usuário associada a um filme.

**timestamp**→Tempo em segundos desde a meia-noite, horário universal coordenado (UTC) de 1 de janeiro de 1970.

## Explorando os Dados

Já vimos que a nossa base de dados se trata de votos atribuídos por usuários, para os filmes desejados, assim, seria interessante conhecermos a escala utilizada para as notas.

Podemos começar verificando os valores máximo e mínimo possível.

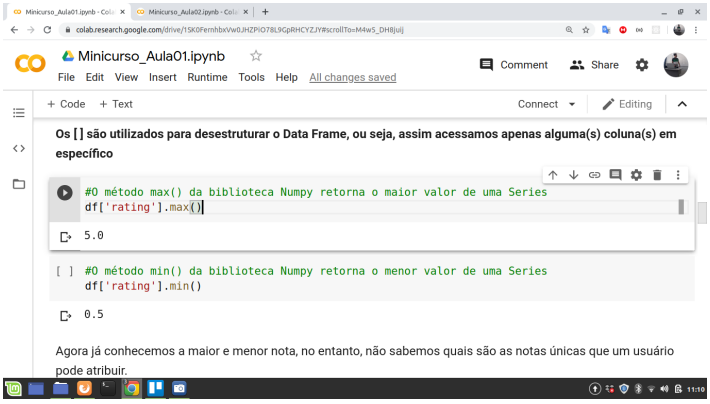
Se o filme for muito ruim, qual a menor nota possível que posso atribuir ?

```
#Os [] são utilizados para desestruturar os dados, acessando nesse caso somente a coluna 'rating'.
df['rating'].max()
```

Se o filme for muito bom, qual a maior nota possível que posso atribuir ?

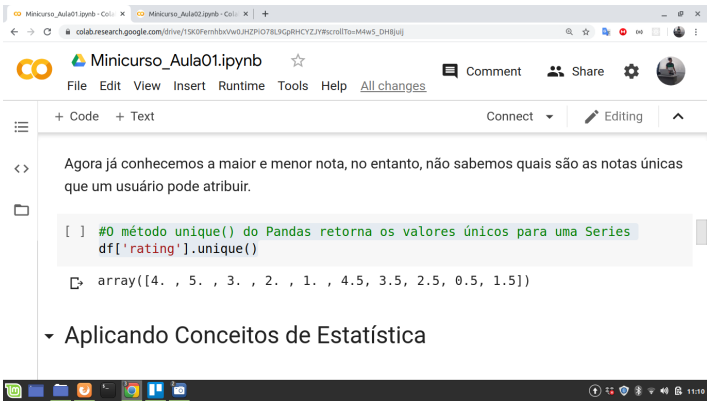
```
df['rating'].min()
```

Os métodos max( ) e min( ) pertencem a biblioteca Numpy e como vimos, retornam o maior e o menor valor de uma base de dados, respectivamente.




Agora já conhecemos a maior e a menor nota, mas não sabemos qual a forma precisa da escala, posso por exemplo atribuir um 3.73 a um filme que vi ?

```
#O método unique() do Pandas retorna os valores únicos para uma Series
df['rating'].unique()
```



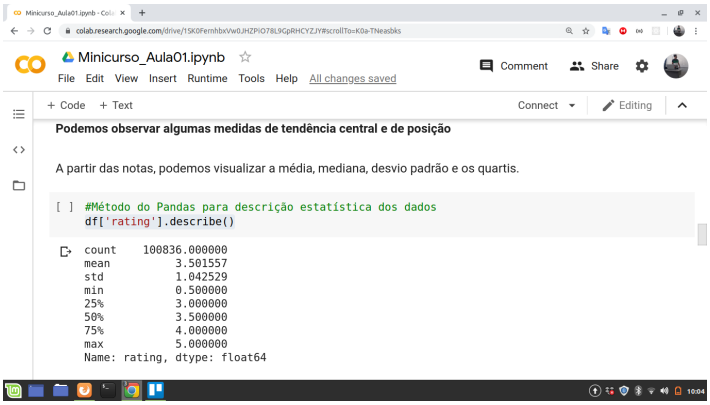

## Aplicando Conceitos de Estatística

Baseando-se nas primeiras aulas da disciplina, poderemos explorar algumas informações que aprendemos.

Para verificar uma descrição estatística nos nossos dados, é possível utilizar um método do Pandas, esse é o describe( )

```
df['rating'].describe()
```

Apenas a coluna de notas foi selecionada, devido ao fato de que observações estatísticas não são úteis para todos os tipos de dados, como é o caso por exemplo das outras colunas.




## Entendendo o Describe

**O método describe retorna uma série de medidas estatísticas, medidas gerais, medidas de posição e medidas de tendência central.**

****count**** = Quantidade de linhas da coluna 'rating'.

****mean**** = Média de todas as notas.

****std**** = Desvio padrão para todas as notas.

****min**** = Entrada miníma, ou seja: menor nota atribuída.

****25****% = 1º Quartil dos dados.

****50****% = 2º Quartil dos dados e respectivamente a mediana.

****75****% = 3º Quartil dos dados.

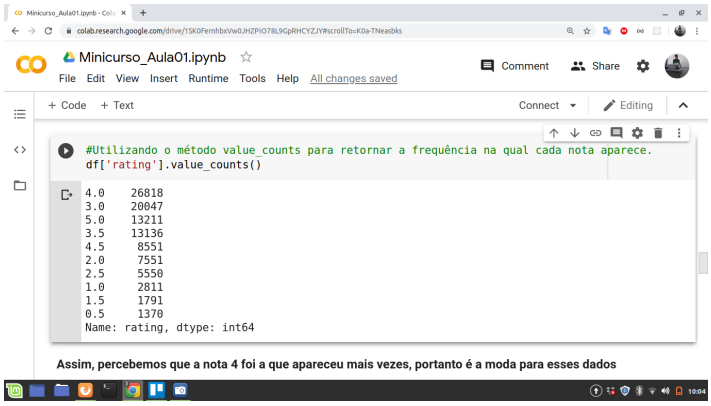
****max**** = Entrada máxima, ou seja: maior nota atribuída.

Podemos começar a fazer novas observações sobre os nossos dados a partir dessas medidas, por exemplo a "baixa" variação apresentada pelas notas, destacada pelo valor do desvio padrão.

Uma possível explicação é que o site apresenta uma escala delimitada e relativamente pequena (0.5 - 5).

**Observamos a média e a mediana, é possível descobrir a moda ?**

```
#Utilizando o método value_counts() para retornar a frequência na qual cada nota aparece.
df['rating'].value_counts()
```




O método apresenta o formato: Valor:Frequência e já exibe os dados de forma ordenada, pelo 2º atributo.