**Projeto da Disciplina LUI**

Prof. Evelyn Batista e Felipe Borges

**Componentes do Grupo:**

|  |  |  |
| --- | --- | --- |
| Nome | Matrícula | email |
| Renata Regina da Fonseca Santos | 192.110.195 | renatarfsantos@zipmail.com.br |
| Sérgio Gustavo Souza Guerra | 201.110.052 | sergiog.guerra@gmail.com |
| Flávia Szczerbacki | 192.110.229 | sflavia@gmail.com |
| Mariana Gonçalves Almeida Pinto | 192.110.208 | marianagoncalvesap@gmail.com |

**Sumário**

[**Introdução**](#_gjdgxs) **3**

[**Webscraping**](#_30j0zll) **4**

[**Análise dos Dados**](#_1fob9te) **7**

[**Filmes organizados por gênero dominante**](#_4d34og8) **10**

[**Filmes organizados por ano de lançamento**](#_2s8eyo1) **11**

[**Conclusão**](#_ciihott52qq5) **13**

## Introdução

O webscraping é uma ferramenta valiosa nos dias de hoje, uma vez que evita um trabalho repetitivo de copiar e colar informações e possibilita facilmente organizar os dados coletados em Excel, HTML e CSV.

No caso específico do trabalho proposto, a pesquisa sobre filmes em um dos maiores acervos da internet possibilita, além do exposto anteriormente, uma maior agilidade na coleta dos dados.

Diante disso, ao longo do desenvolvimento desse trabalho, nós utilizamos a biblioteca BeautifulSoup para trabalhar a base do IMDB a partir do URL e executar o webscraping.

Nesse processo tomamos o cuidado para modelar o browser independente do usado pelo usuário, que pode ajudar no processo de busca. Identificamos o intervalo de 100 filmes por página, além disso, definimos o número de páginas para análise como 100 (variável *n*).

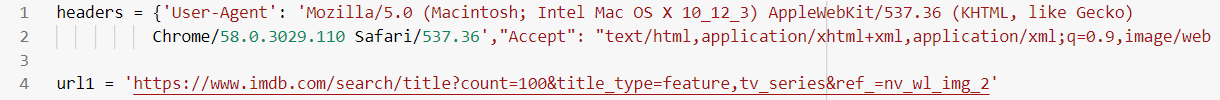
A partir disso, iniciamos a análise da base, que passou por etapas de pré-processamento; transformação; limpeza; criação de colunas e outras.

## Webscraping

Abaixo, explicamos as etapas executadas para o webscraping e criação do dataframe sem o pré-processamento do texto.

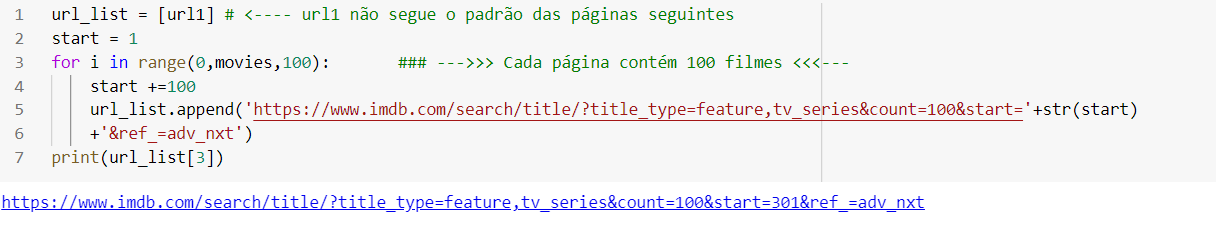
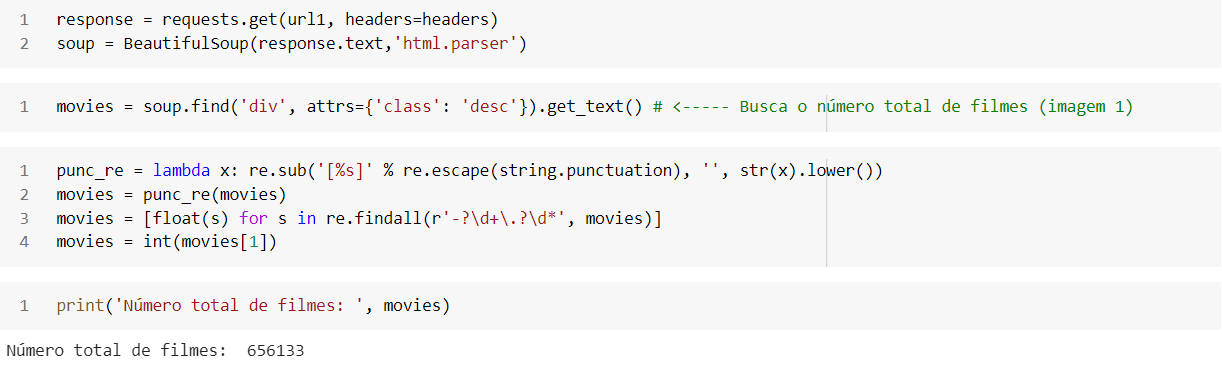
* A url da primeira página (variável *url1*) não segue o padrão das urls das páginas seguintes. Precisamos encontrar o padrão dessas páginas.



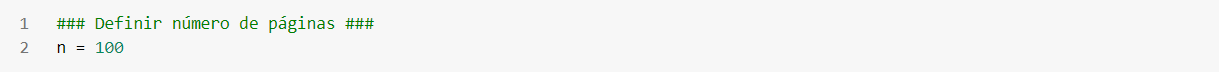


* Buscar o número total de filmes na base, para fazer um loop e construir uma lista com todas as urls seguintes à primeira.



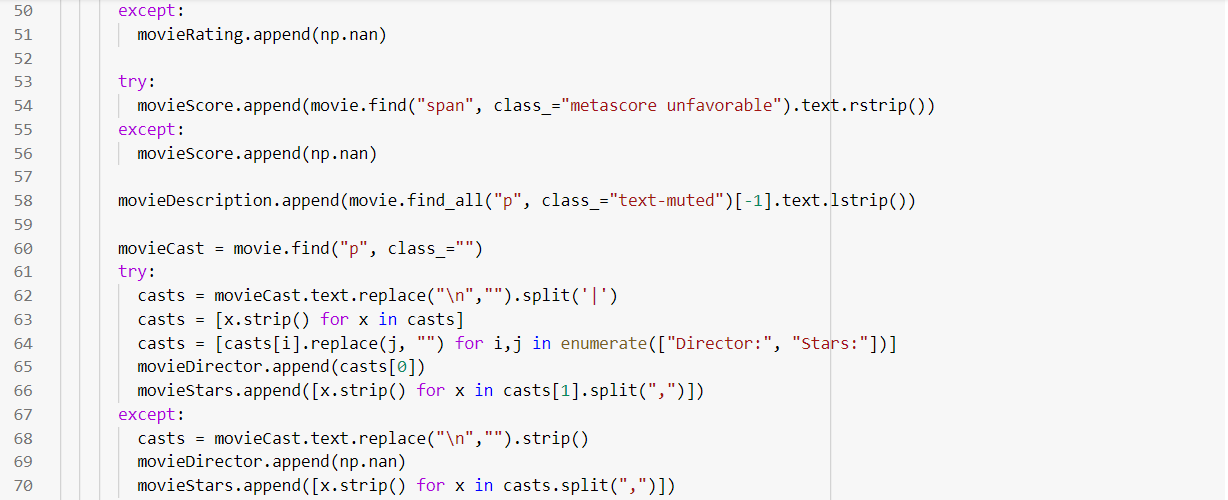


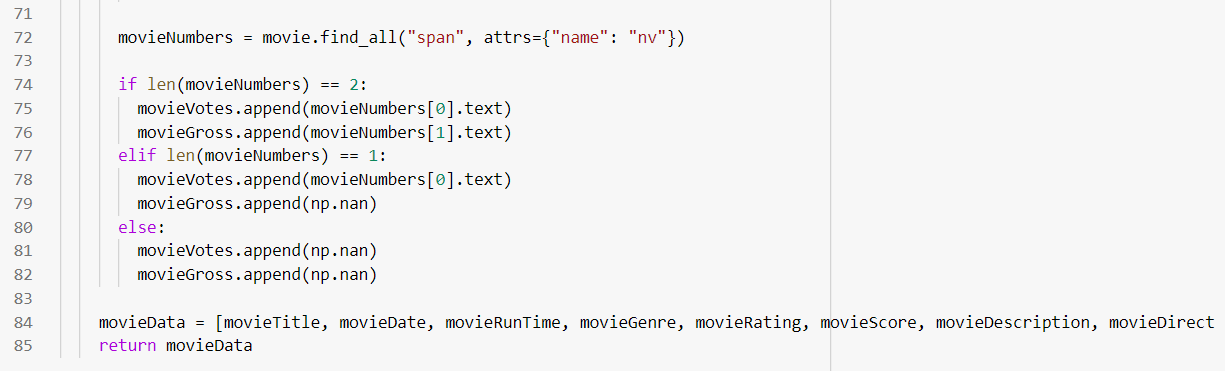
* Executar webscraping do número de páginas (variável *n*) desejado e criar dataframe com as informações

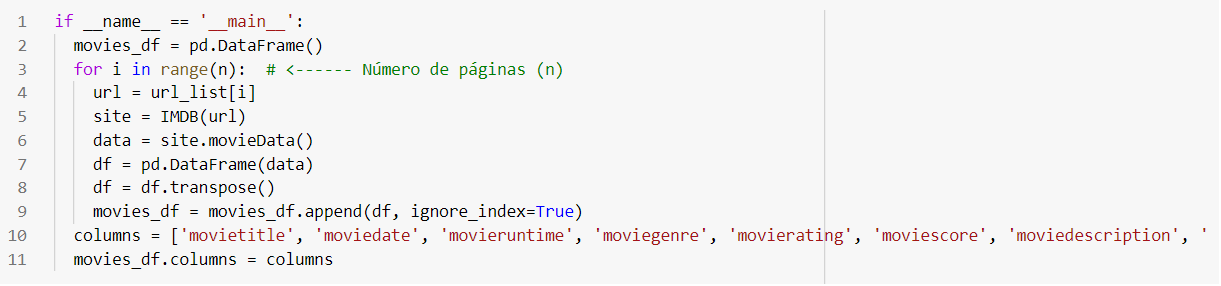




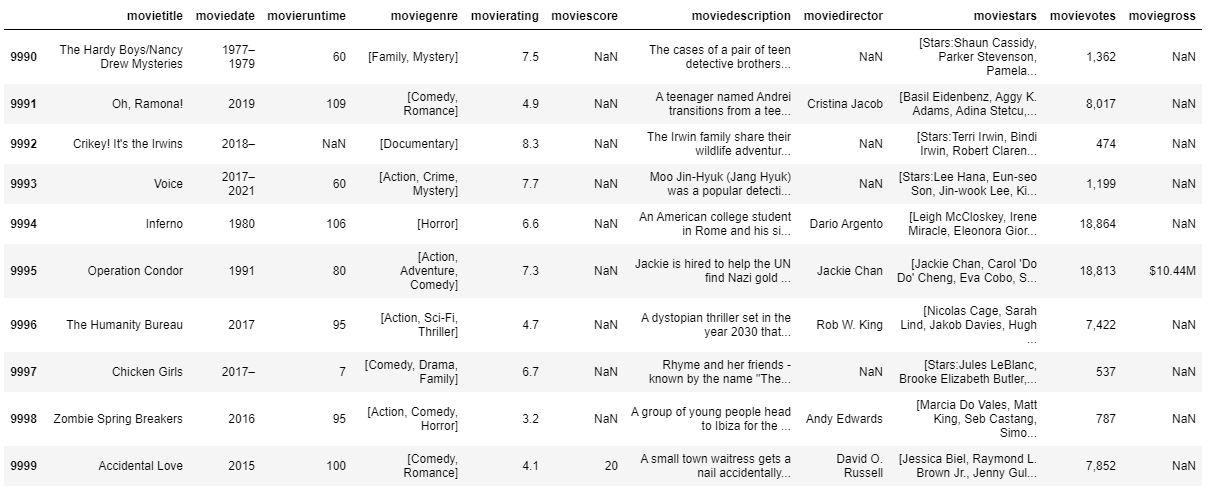








* Dataframe resultante do webscraping:



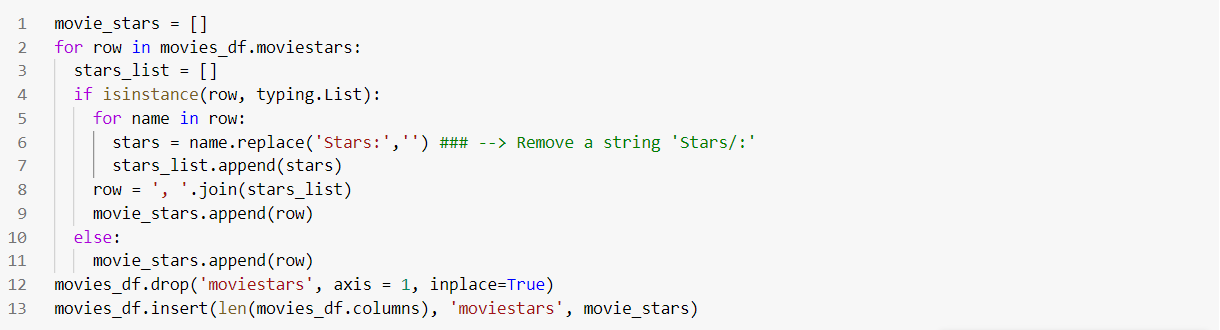
## Análise dos Dados

Identificamos que o dataframe inicial, criado a partir do webscraping, precisa passar por algumas etapas de pré-processamento de texto, para que possamos fazer a organização por gênero e por data de lançamento. Utilizamos os recursos apreendidos na disciplina de NLP para retirada da pontuação e outras sintaxes que apareciam como lixo.

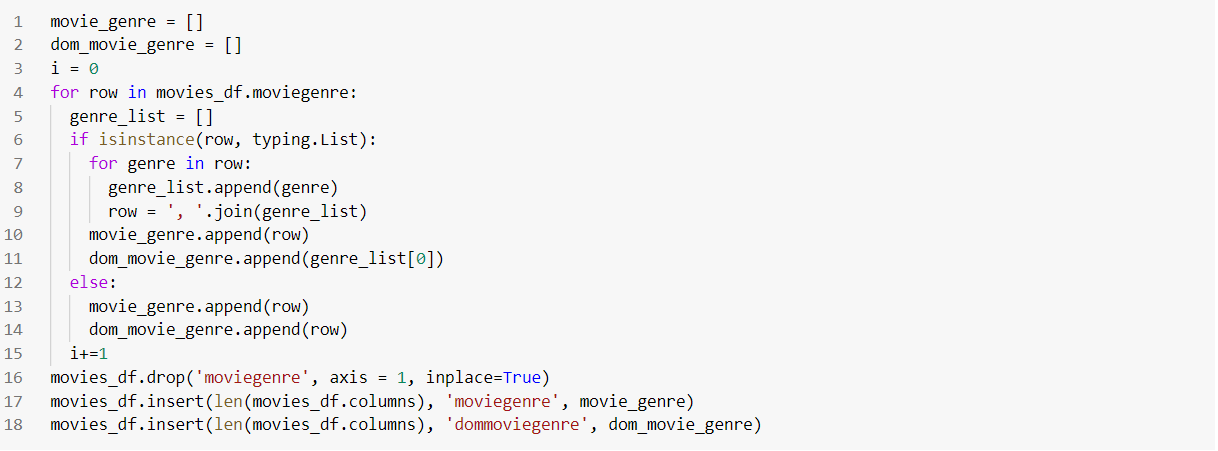
* Remover linhas onde as colunas *moviegenre* ou *moviedate* apresentam NaN.



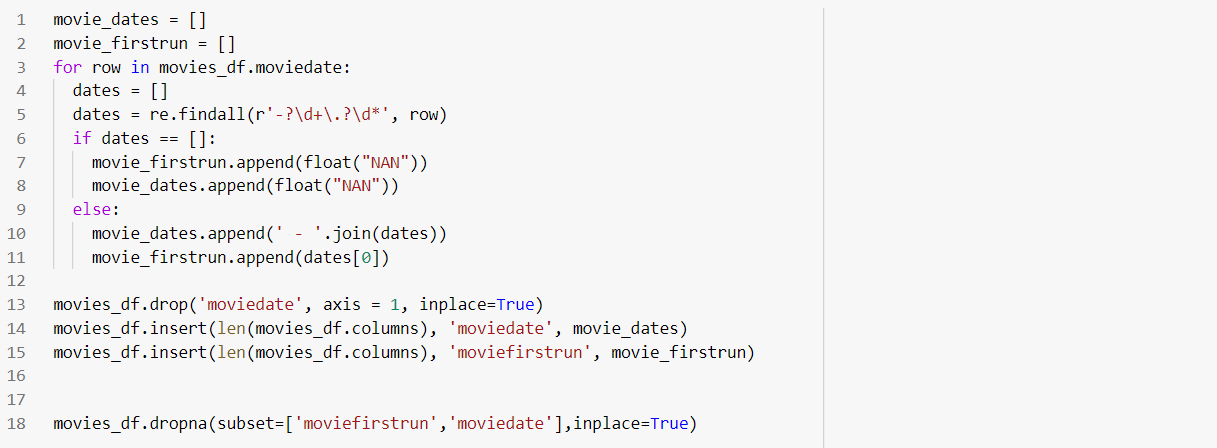
* Transformar a coluna *moviestars* string única, pois no dataframe original os nomes dos atores estão em uma lista.



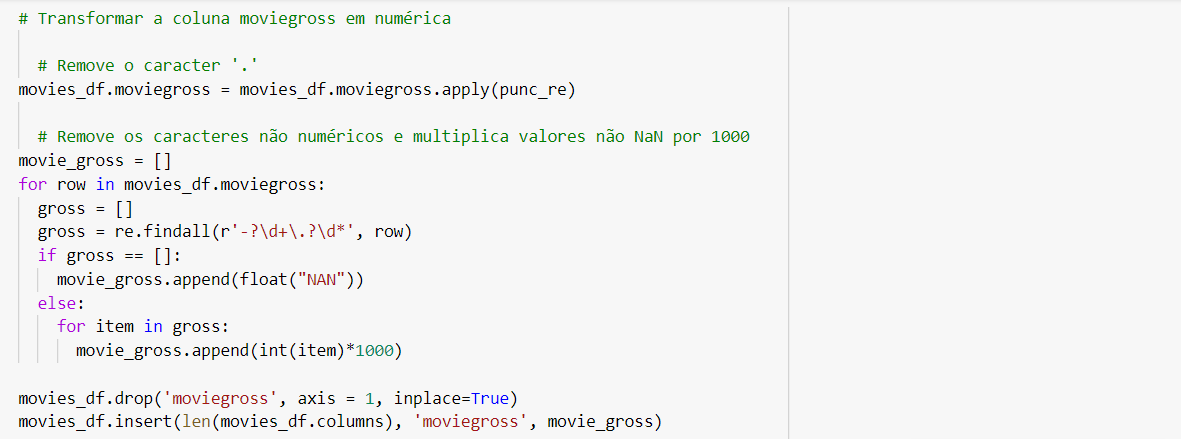
* Transformar a coluna *moviegenre* em string única,pois no dataframe original os gêneros estão em uma lista e criar a coluna *dommoviegenre* para a organização do dataframe por gênero.

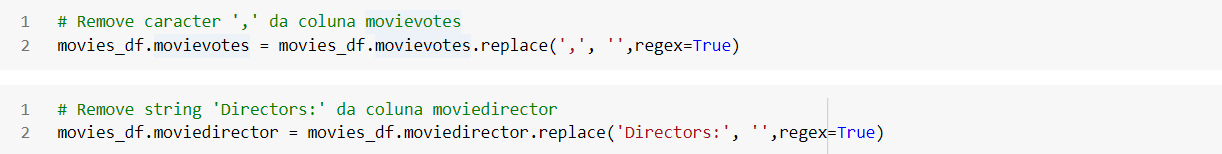


* Limpar a coluna *moviedate* e criar a coluna *moviefirstrun*. Utilizaremos a coluna *moviefirstrun* a seguir para a organização do dataframe por data.



* Limpar colunas moviegross, movievotes e moviedirector



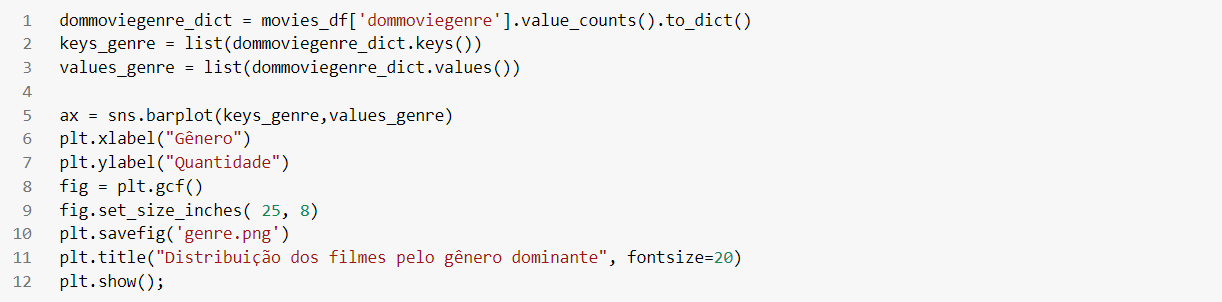


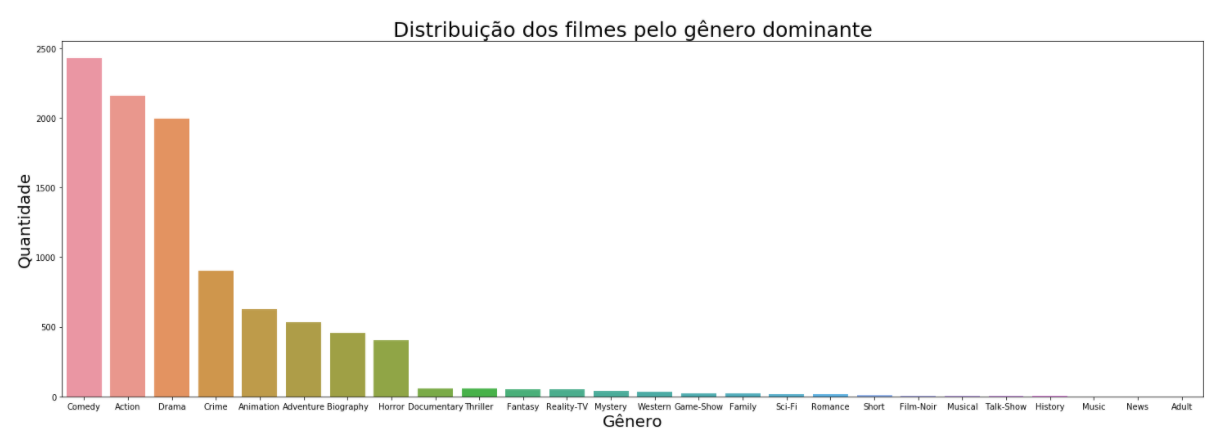
* Após o pré-processamento do texto, foi gerado o dataframe abaixo:



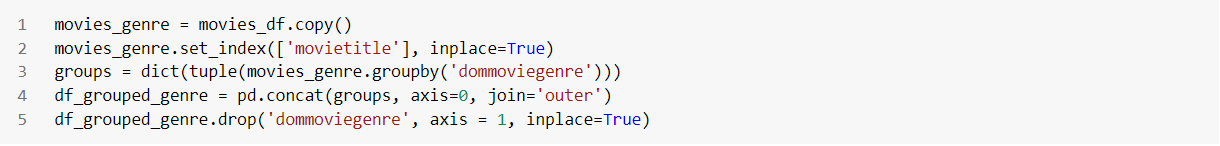
## Filmes organizados por gênero dominante

* Visualizar a distribuição dos filmes pelo gênero dominante:





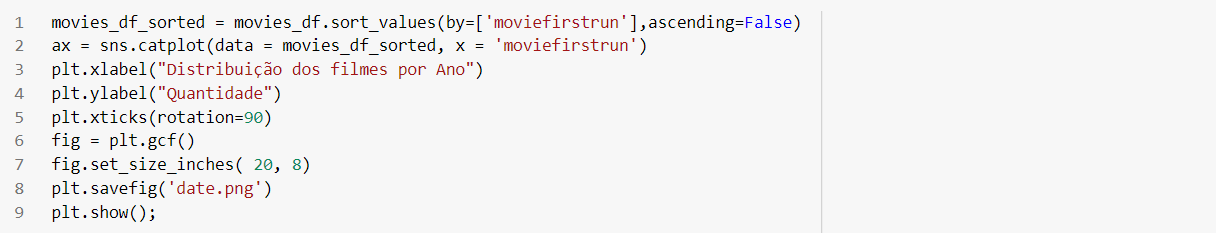
* Construir dataframe organizado pelo gênero dominante do filme:

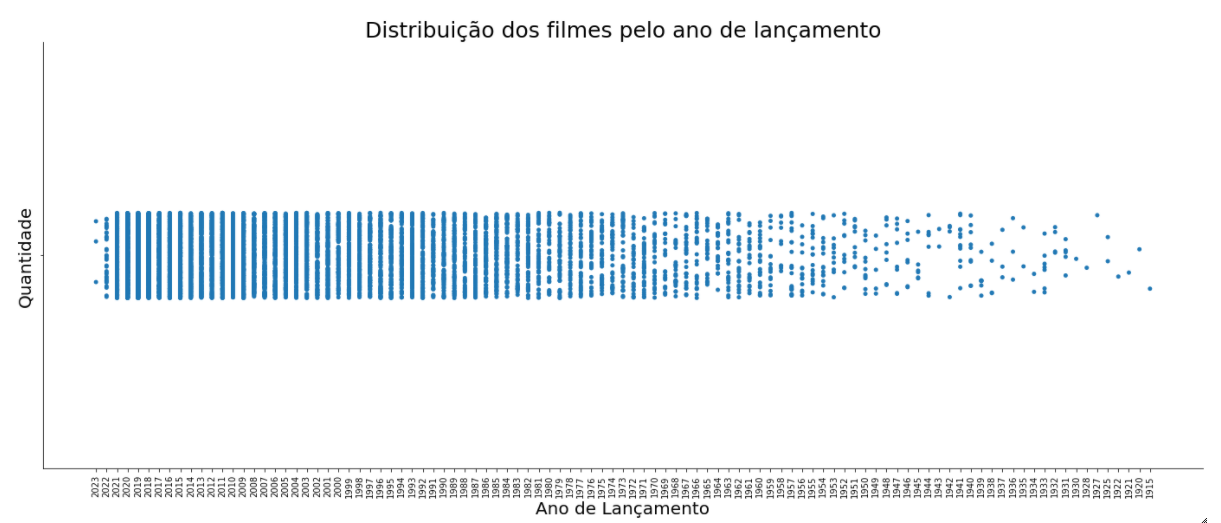




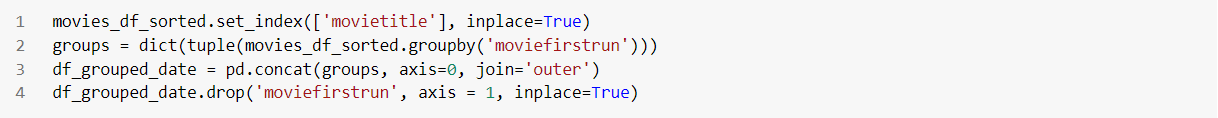
## Filmes organizados por ano de lançamento

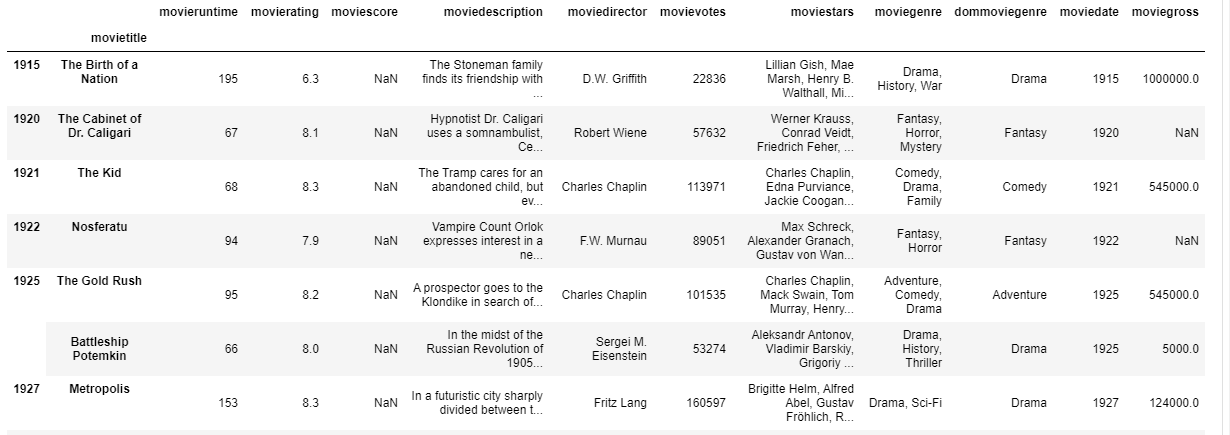
* Visualizar a distribuição dos filmes por ano de lançamento





* Construir dataframe organizado por ano de lançamento:





## Conclusão

O trabalho aqui proposto, possibilitou que trabalhássemos os conceitos absorvidos dentro da sala de aula. O que permitiu, além de uma melhor compreensão da matéria, entender os benefícios gerados por essa ferramenta, a saber:

1. Agilidade no processo de coleta de dados;

2. Economia de tempo e diminuição de custos com analistas de dados;

3. Possibilita um acompanhamento mais efetivo da concorrência.

Com o exercício, percebe-se que um número elevado de informação, nesse caso que contém a maior seleção de filmes de um site, é elaborada e analisada mais facilmente, permitindo que facilmente um analista consiga identificar a predominância, por exemplo, do gênero dos filmes e a distribuição dos anos de lançamentos. Nesse contexto, a análise sem essa ferramenta em sites que possuem um elevado número de informações, demandaria muito tempo, além de não ter a mesma confiabilidade na geração dos dados para as análises.

Diante disso, entende-se que a técnica de web scraping é fundamental nos dias de hoje, principalmente em mercados altamente competitivos, onde se há a necessidade de acompanhamento constante da concorrência e disseminação dessas informações de forma mais clara e rápida dentro da corporação.