# 

(https://databricks.com) Objetivo do trabalho

• objetivo é identificar qual genero tem a melhor nota IMDb e melhor Metascore. Objetivos secundários: qual ano teve a melhor média entre o genero com melhor nota; quais generos tiveram melhores notas em cada ano

#### Base de dados

 https://www.kaggle.com/datasets/parthdande/imdb-dataset-2024-updated (https://www.kaggle.com/datasets/parthdande/imdb-dataset-2024-updated)

#### Coleta de Dados

```
!pip install kaggle
Requirement already satisfied: kaggle in /local_disk0/.ephemeral_nfs/envs/pythonEnv-511eb367-a865-48fc-b3f7-4c7dd55fe8de/li
b/python3.9/site-packages (1.6.14)
Requirement\ already\ satisfied:\ python-slugify\ in\ /local\_disk0/.ephemeral\_nfs/envs/pythonEnv-511eb367-a865-48fc-b3f7-4c7dd55f
e8de/lib/python3.9/site-packages (from kaggle) (8.0.4)
Requirement already satisfied: python-dateutil in /databricks/python3/lib/python3.9/site-packages (from kaggle) (2.8.2)
Requirement already satisfied: urllib3 in /databricks/python3/lib/python3.9/site-packages (from kaggle) (1.26.9)
Requirement already satisfied: six>=1.10 in /databricks/python3/lib/python3.9/site-packages (from kaggle) (1.16.0)
Requirement already satisfied: bleach in /databricks/python3/lib/python3.9/site-packages (from kaggle) (4.1.0)
Requirement already satisfied: tqdm in /local_disk0/.ephemeral_nfs/envs/pythonEnv-511eb367-a865-48fc-b3f7-4c7dd55fe8de/lib/p
ython3.9/site-packages (from kaggle) (4.66.4)
Requirement already satisfied: certifi>=2023.7.22 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-511eb367-a865-48fc-b3f7-4c7d
d55fe8de/lib/python3.9/site-packages (from kaggle) (2024.7.4)
Requirement already satisfied: requests in /databricks/python3/lib/python3.9/site-packages (from kaggle) (2.27.1)
Requirement already satisfied: webencodings in /databricks/python3/lib/python3.9/site-packages (from bleach->kaggle) (0.5.1)
Requirement already satisfied: packaging in /databricks/python3/lib/python3.9/site-packages (from bleach->kaggle) (21.3)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /databricks/python3/lib/python3.9/site-packages (from packaging->
bleach->kaggle) (3.0.4)
Requirement already satisfied: text-unidecode>=1.3 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-511eb367-a865-48fc-b3f7-4c7
dd55fe8de/lib/python3.9/site-packages (from python-slugify->kaggle) (1.3)
Requirement already satisfied: idna<4.>=2.5 in /databricks/python3/lib/python3.9/site-packages (from requests->kaggle) (3.3)
Requirement already satisfied: charset-normalizer~=2.0.0 in /databricks/python3/lib/python3.9/site-packages (from requests-> "
```

```
def authenticate_kaggle_api(kaggle_username, kaggle_key) -> KaggleApi:
    import os

    os.environ['KAGGLE_USERNAME'] = kaggle_username
    os.environ['KAGGLE_KEY'] = kaggle_key

from kaggle.api.kaggle_api_extended import KaggleApi

api = KaggleApi()
    api.authenticate()
    print("API authenticated.")

return api
```

```
# kaggle api tokens
      import pandas as pd
      kaggle_username = 'thalesdamiletto'
      kaggle_key = '12857b21dac382a255dd1958b0928cff'
                  api = authenticate_kaggle_api(kaggle_username, kaggle_key)
       except e as Exception:
                   print(f'Erro ao autenticar a api do kaggle: {Exception}')
     dataset_name = 'parthdande/imdb-dataset-2024-updated'
      download path = 'imdb dataset 2'
      if not os.path.exists(download_path):
                  os.makedirs(download_path)
      # Baixe o dataset
      api.dataset_download_files(dataset_name, path=download_path, unzip=True)
      print(f"Os\ arquivos\ foram\ baixados\ e\ extraídos\ para:\ \{download\_path\}")
API authenticated.
Dataset URL: https://www.kaggle.com/datasets/parthdande/imdb-dataset-2024-updated (https://www.kaggle.com/datasets/parthdande/imdb-dataset-2024-updated (https://www.kaggle.com/dataset-2024-updated (https://www.kaggle.com/datas
imdb-dataset-2024-updated)
Os arquivos foram baixados e extraídos para: imdb_dataset_2
```

Гаb <b>l</b> е					Q	$\nabla$	
	∆ <sup>B</sup> c path	∆ <sup>B</sup> <sub>C</sub> name	1 <sup>2</sup> 3 size	1 <sup>2</sup> 3 modificationTime			
1	file:/databricks/driver/hadoop_accessed_config.lst	hadoop_accessed_config.lst	2755	1720647716827			
2	file:/databricks/driver/azure/	azure/	4096	1720647716823			
3	file:/databricks/driver/conf/	conf/	4096	1720647715883			
4	file:/databricks/driver/preload_class.lst	preload_class.lst	1306936	1720647716859			
5	file:/databricks/driver/imdb-dataset-2024-updated.z	imdb-dataset-2024-updated.z	335942	1720278268000			
6	file:/databricks/driver/imdb_dataset_2/	imdb_dataset_2/	4096	1720658679412			
7	file:/databricks/driver/logs/	logs/	4096	1720656819820			
8	file:/databricks/driver/path/	path/	4096	1720654458483			
9	file:/databricks/driver/ganglia/	ganglia/	4096	1720657802130			
10	file:/databricks/driver/eventlogs/	eventlogs/	4096	1720652123845			
11	file:/databricks/driver/imdb_dataset/	imdb_dataset/	4096	1720654090366			
12	file:/databricks/driver/metastore db/	metastore db/	4096	1720652209273			

Гаb <b>l</b> е							
56	<sup>B</sup> at <b>ÿiHe</b> arst	€ IMDb Rating	<b>A</b> 928∦ear	₽ <sup>B</sup> c Certificates	<b>Bo</b> Graphy	<b>P</b> au <b>Pite∉tør</b> er	
57	The Look of Silence	8.3	2014	PG-13	Documentary	Joshua Oppenheime	
58	Mesrine: Public Enemy No. 1	7.4	2008	R	Action	Jean-François Riche	
59	Walker	6.6	1987	R	Biography	Alex Cox	
60	The Sparks Brothers	7.7	2021	R	Documentary	Edgar Wright	
61	Lady Sings the Blues	7.0	1972	R	Biography	Sidney J. Furie	
62	Val	7.6	2021	R	Documentary	Ting Poo	
63	Marley	7.9	2012	PG-13	Documentary	Kevin Macdonald	
64	Prick Up Your Ears	7.1	1987	R	Biography	Stephen Frears	
65	Lisztomania	6.1	1975	R	Biography	Ken Russell	
56	Why Do Fools Fall in Love	6.4	1998	R	Biography	Gregory Nava	
67	The Glorias	6.0	2020	R	Biography	Julie Taymor	
68	The Story of Adele H	7.2	1975	PG	Biography	François Truffaut	

69	Prefontaine	6.8	1997	PG-13	Biography	Steve James	•
70	Backbeat	6.6	1994	R	Biography	lain Softley	
4,573 rd	ows						
							-

- # atualizar código para buscar os dados da internet, ao invés da memoria do databricks link: https://www.kaggle.com/datasets/p
- # df\_origem = spark.read.format("csv").option("header", True).load('dbfs:///FileStore/tables/csv/mvp/\*')
- # df\_origem.display()

#### Modelagem

- # código de modelagem
- # quebrar tabela de origem resultando em dados úteis ao nosso modelo para atender o objetivo proposto
- # objetivo é identificar qual genero tem a melhor nota IMDb e melhor Metascore. Obj secundário: qual ano teve a melhor média e
- # Colunas Necessáris: Title, IMDB Rating, Year, MetaScore
- # Criar novo dataframe apenas com as colunas necessárias

df\_cleansed = df\_origem.select('Title', 'Genre', 'IMDb Rating', 'MetaScore', "Year")

df\_cleansed = df\_cleansed.withColumnRenamed("Title", "title").withColumnRenamed("Genre", "genre").withColumnRenamed("IMDb Rati
df\_cleansed.display()

Tab <b>l</b> e						
	A <sup>B</sup> C title	₄ <sup>B</sup> c genre	ABc imdb_rating	<b>A</b> <sup>B</sup> <sub>C</sub> metascore	<b>A<sup>B</sup></b> C year	
1	End of the Spear	Adventure	6.8	45.0	2005	
2	Elvira Madigan	Biography	7.0	66.0	1967	
3	The Kid Stays in the Picture	Documentary	7.3	75.0	2002	
4	It Ain't Over	Documentary	8.2	79.0	2022	
5	Mahler	Biography	7.0	66.0	1974	
6	The Dark Horse	Biography	7.4	77.0	2014	
7	Stephen Curry: Underrated	Documentary	7.3	69.0	2023	
8	Carrington	Biography	6.8	66.0	1995	
9	Burden	Biography	6.7	57.0	2018	
10	Georgetown	Biography	6.2	49.0	2019	
11	Miles Ahead	Biography	6.4	64.0	2015	
12	The King of Kong: A Fistful of Quarters	Documentary	8.0	83.0	2007	
13	The Devil's Violinist	Biography	6.0	38.0	2013	
14	A Tale of Love and Darkness	Biography	6.0	55.0	2015	
15	Cobb	Biography	6.4	66.0	1994	

## df\_cleansed.printSchema

Out[136]: <bound method DataFrame.printSchema of DataFrame[title: string, genre: string, imdb\_rating: string, metascore: string, year: string]>

### Análise da Qualidade dos Dados

```
# Transformar imdb_rating e metascore em decimal e year em int. E dropar linahs onde a conversão der errado.
   \label{prop:condition} from \ pyspark.sql. functions \ import \ col, \ when
   from pyspark.sql.types import DecimalType, IntegerType
   # Define o esquema para as conversões
   decimal_type = DecimalType(4, 1) # Define a precisão e escala. Ajuste conforme necessário.
   integer_type = IntegerType()
   # Converta as colunas para DecimalType e remove linhas inválidas
  df transformed = (
            df_cleansed
             .withColumn("imdb_rating", when(col("imdb_rating").rlike("^[0-9]*\\.?[0-9]+$"), col("imdb_rating").cast(decimal_type)).oth
             . with Column ("metascore"), when (col("metascore").rlike("^[0-9]*\.?[0-9]+$"), col("metascore"). cast(decimal\_type)). otherwise (column ("metascore").rlike("^[0-9]*\.?[0-9]+$"), col("metascore"). cast(decimal\_type)). otherwise (column ("metascore").rlike("^[0-9]*\.?[0-9]*\.?[0-9]*\.?[0-9]*\.]), col("metascore"). cast(decimal\_type)). otherwise (column ("metascore").rlike("^[0-9]*\.?[0-9]*\.]), col("metascore"). cast(decimal\_type)). otherwise (column ("metascore").rlike("metascore"). cast(decimal\_type)). otherwise (column ("metascore").rlike("metascore"). cast(decimal\_type)). cast(decimal\_type)). cast(decimal\_type). cast
             . with Column("year", when (col("year").rlike("^[0-9]\{4\}^*), col("year").cast(integer\_type)). otherwise(None)) \\
             .dropna(subset=["imdb_rating", "metascore"])
  df_transformed.display()
  Table
                                                                                                                                                                                                                                                                                  Q 7
                                                                                                                                                                                                                                                                                                          ABU thitled Kingdom
                                                                                                                                                                                          .00 metascore 65.0
                                                                                                                                                                                                                                      1<sup>2</sup>3 year 2016
                                                                                                         allographs/
                                                                                                                                           .00 imdb rating
  104
                                                                                                                                                                              6.9
   105
                 The Dove
                                                                                                         Adventure
                                                                                                                                                                               6.3
                                                                                                                                                                                                                         66.0
                                                                                                                                                                                                                                                        1974
   106
                 Lifemark
                                                                                                         Biography
                                                                                                                                                                               6.1
                                                                                                                                                                                                                         66.0
                                                                                                                                                                                                                                                        2022
   107
                 Creation
                                                                                                         Biography
                                                                                                                                                                               6.6
                                                                                                                                                                                                                         51.0
                                                                                                                                                                                                                                                        2009
  108
                 Permanent Midnight
                                                                                                                                                                               6.2
                                                                                                                                                                                                                         57.0
                                                                                                                                                                                                                                                        1998
                                                                                                         Biography
                 Woman Walks Ahead
                                                                                                                                                                               6.7
                                                                                                                                                                                                                                                        2017
   109
                                                                                                         Biography
                                                                                                                                                                                                                         51.0
                                                                                                                                                                                                                         79.0
  110
                 An Angel at My Table
                                                                                                                                                                               7.4
                                                                                                                                                                                                                                                        1990
                                                                                                         Biography
  111
                  Won't You Be My Neighbor?
                                                                                                         Documentary
                                                                                                                                                                               8.3
                                                                                                                                                                                                                         85.0
                                                                                                                                                                                                                                                        2018
                 Mesrine: Killer Instinct
                                                                                                                                                                                                                                                       2008
  112
                                                                                                         Action
                                                                                                                                                                               7.5
                                                                                                                                                                                                                        71.0
  113
                 Ride Like a Girl
                                                                                                                                                                               7.0
                                                                                                                                                                                                                        47.0
                                                                                                                                                                                                                                                       2019
                                                                                                         Biography
                                                                                                                                                                               7.0
                                                                                                                                                                                                                         67.0
                                                                                                                                                                                                                                                       2012
  114
                 The Sapphires
                                                                                                         Biography
  115
                 Enron: The Smartest Guys in the Room
                                                                                                         Documentary
                                                                                                                                                                               7.6
                                                                                                                                                                                                                         66.0
                                                                                                                                                                                                                                                        2005
                                                                                                                                                                               7.0
                                                                                                                                                                                                                                                        2021
  116
                 Swan Song
                                                                                                         Biography
                                                                                                                                                                                                                         65.0
  117
                 Bird
                                                                                                         Biography
                                                                                                                                                                               7.1
                                                                                                                                                                                                                         78.0
                                                                                                                                                                                                                                                        1988
  118
                 Queen of Katwe
                                                                                                         Biography
                                                                                                                                                                               7.4
                                                                                                                                                                                                                         73.0
                                                                                                                                                                                                                                                        2016
                                                                                                                                                                                                                                                       2017
  119
                 Rebel in the Rye
                                                                                                        Biography
                                                                                                                                                                               6.7
                                                                                                                                                                                                                         46.0
3.573 rows
```

## Carga de dados

```
# escrever as tabelas ou dados
try:
    permanent_table_name = "imdb_final_dataset_csv"
    df_transformed.write.mode("overwrite").format("parquet").saveAsTable(permanent_table_name)

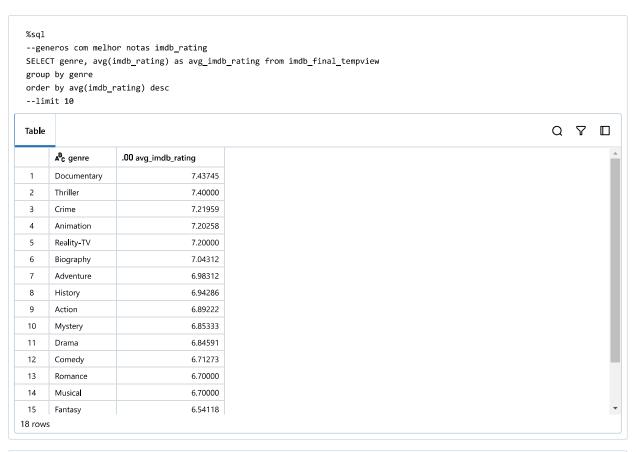
except:
    print('Tabela jé existe. Fazendo append')
    try:
        permanent_table_name = "imdb_final_dataset_csv"
            df_transformed.write.mode("append").format("parquet").saveAsTable(permanent_table_name)
    except:
        print('Erro ao fazer append.')

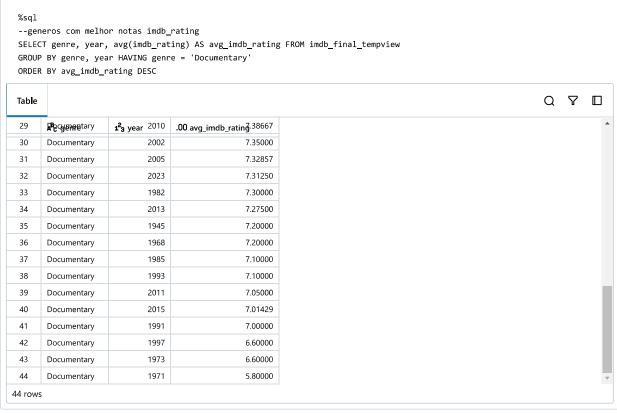
temp_view_name = "imdb_final_tempview"
    df_transformed.createOrReplaceTempView(temp_view_name)

Tabela jé existe. Fazendo append
Erro ao fazer append.
```

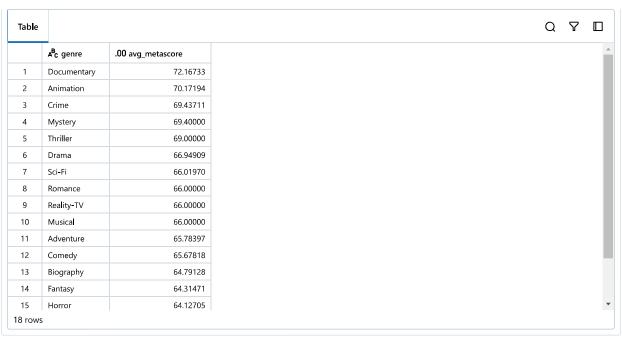
## Solução dos problemas

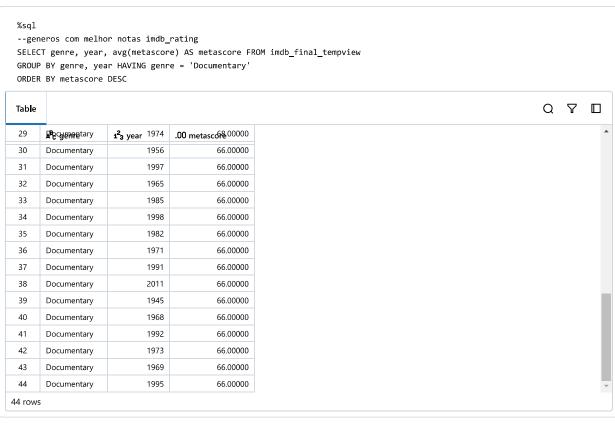
```
# códigos para soluções
# objetivo é identificar qual genero tem a melhor nota IMDb e melhor Metascore. Obj secundário: qual ano teve a melhor média e
```





```
%sql
--generos com melhor notas imdb_rating
SELECT genre, avg(metascore) as avg_metascore from imdb_final_tempview
group by genre
order by avg(metascore) desc
--limit 10
```





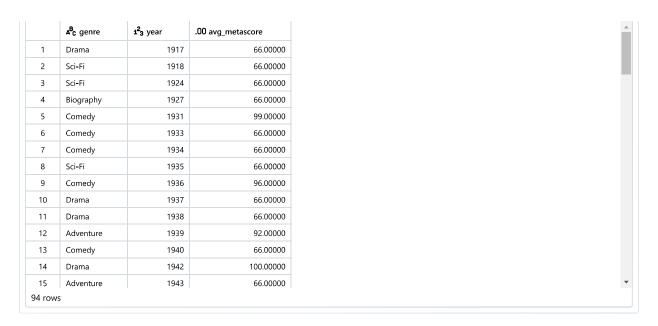
```
%sql
--Qual foi o genero com melhor média em cada ano

SELECT genre, year, avg_metascore FROM (
SELECT T1.*, ROW_NUMBER() OVER(PARTITION BY t1.year ORDER BY T1.avg_metascore desc) as rn

FROM(
SELECT genre, year, avg(metascore) AS avg_metascore FROM imdb_final_tempview

GROUP BY year, genre
) T1
)T2
WHERE rn = 1
ORDER BY YEAR ASC

Table
```



### Respostas para Objetivos

- 1. Genero com melhor avaliação imdb\_rating: Gênero Documentary
- 2. Ano com melhor nota imdb\_rating do genero com melhor média: 1974 -> 8,2
- 3. Genero com melhor avaliação metascore: Gênero Documentary
- 4. Ano com melhor nota imdb\_rating do genero com melhor média: 1970 -> 95,0
- 5. Tabela com o melhor genero e média para cada ano. Tabela gerado no ultimo select