A
PROJECT REPORT ON

"DATA MODELING ON IMDB DATASET"

SUBMITTED BY:

Mr. AKASH BAHURE

Abstract:-

- **The IMDB dataset project involves analyzing a large dataset of movie and TV show titles, ratings, and other metadata from the popular movie and TV review website, IMDb. The dataset includes over 7 million titles, spanning a wide range of genres, languages, and release dates.**

Introduction:-

- **The IMDB dataset project is a data analysis project that involves exploring a large dataset of movie and TV show titles, ratings, and other metadata from IMDb, one of the most popular and influential movie and TV review websites. This dataset contains a wealth of information that can be used to gain insights into trends and patterns in the film and television industry, as well as to provide valuable information to filmmakers, producers, and distributors.**
- **In this project we used the conceptual,logical and physical data modelling**
- **Entities in our database are as follows**

  Titles Ratings Cast

  Companies Genres Crew

Problem Statement:-

The problem statement for the IMDb dataset project is to gain insights into trends and patterns in the movie and TV industry by analyzing a large dataset of movie and TV show titles, ratings, and other metadata from the popular movie and TV review website, IMDb. The goal is to explore the dataset and identify factors that influence movie and TV show ratings, identify popular genres and trends over time, and examine the relationship between budget and box office success.

Keywords:-

- **KEYWORDS**
  1. **Data:- Facts or information.**
  2. **Add constraint:-add a constraint after table is already created.**
  3. **Backup database:- Create a backup of existing database.**
  4. **Forward Engineering:- It is a method through which we make or create.**
  5. **Dataset:- A collection of related set of information.**

- **IMPLEMENTATION:-**

  While implementing the dataset we used different tools and Mysql Command line prompt and Mysql Workbench also for reverse and forward engineering.

Data Modeling:-

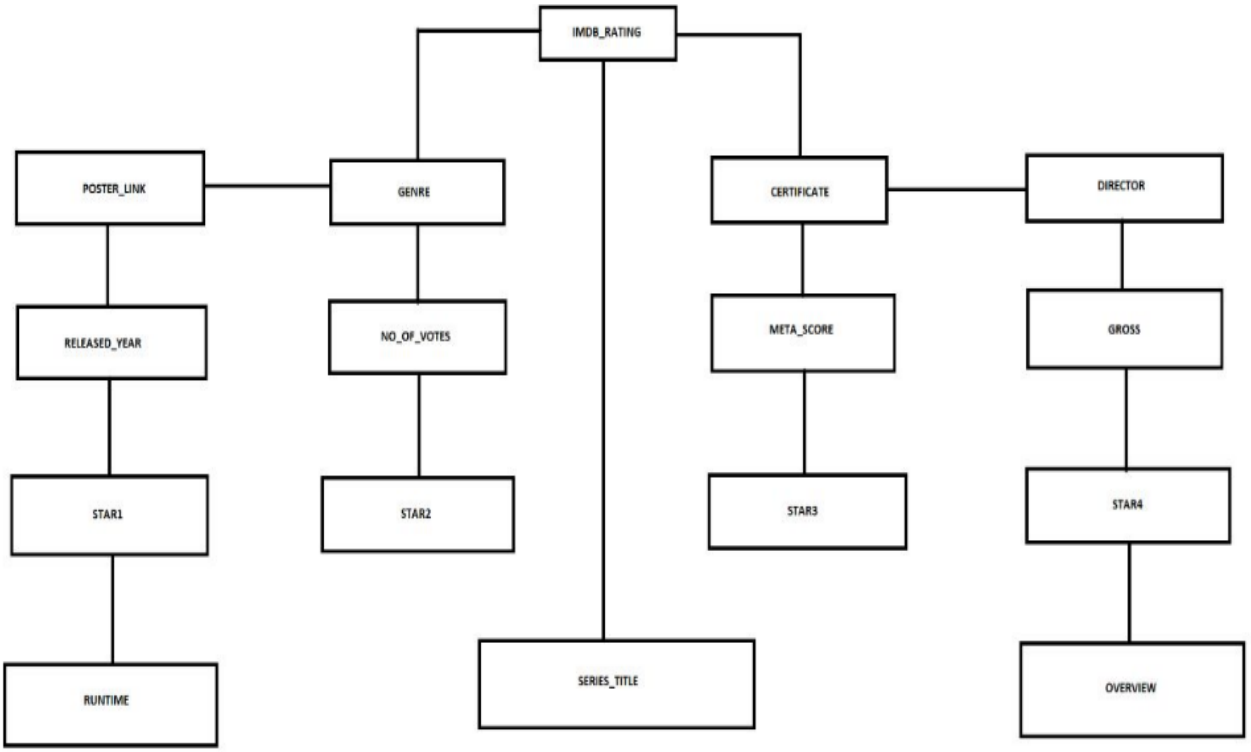
- **Data modelling is a process of creating a data model for the data to be stored in database.**
- **This data model is a conceptual representation of data object ,the association between different data objects, and the rules.**

Types Of Data Modeling:-

- ***Conceptual Data Modeling:-***
- ***Logical Data modeling:-***
- ***Physical Data Modeling:-***

Conceptual Data Modeling:-

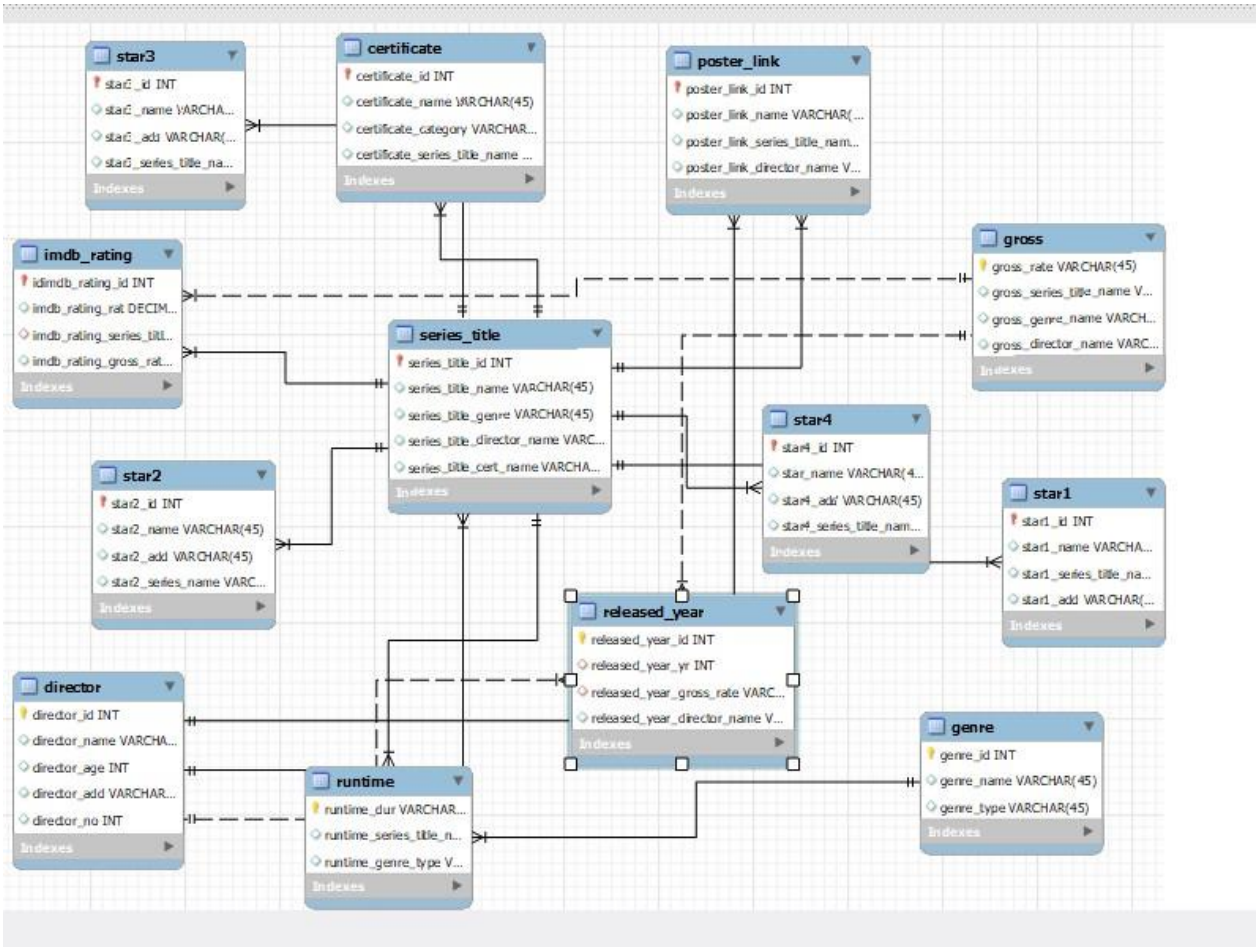
A conceptual data model is an organized view of database concepts and their relationship. The purpose of creating a conceptual data model is to establish entities,their atttributs,and relationships. In this data modelling level,there is hardly any details available on the actual database structure,business stakeholders and data architects typically create a conceptual data model.



Logical Data Modeling:-

Defines how the system should be implemented regardless of the DBMS. This model is typically created by Data Architects and Business Analyst.

The purpose is to developed technical map of rules and data structure.

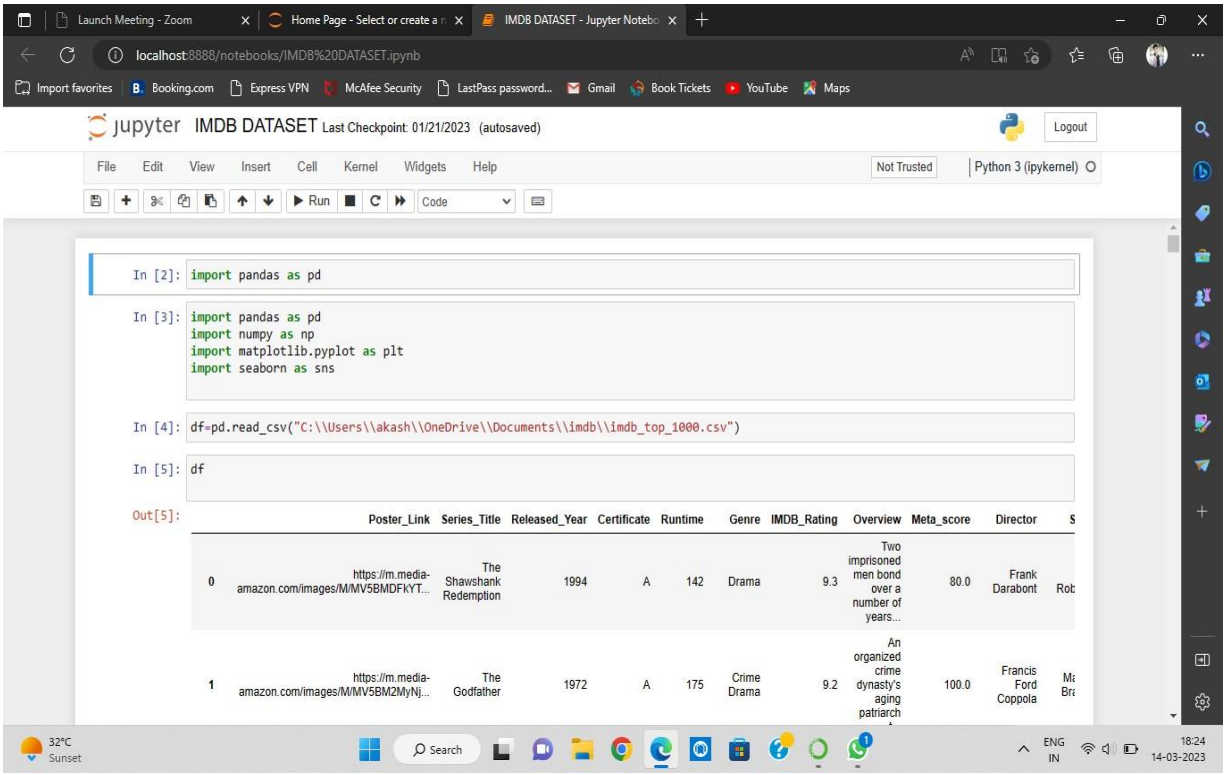


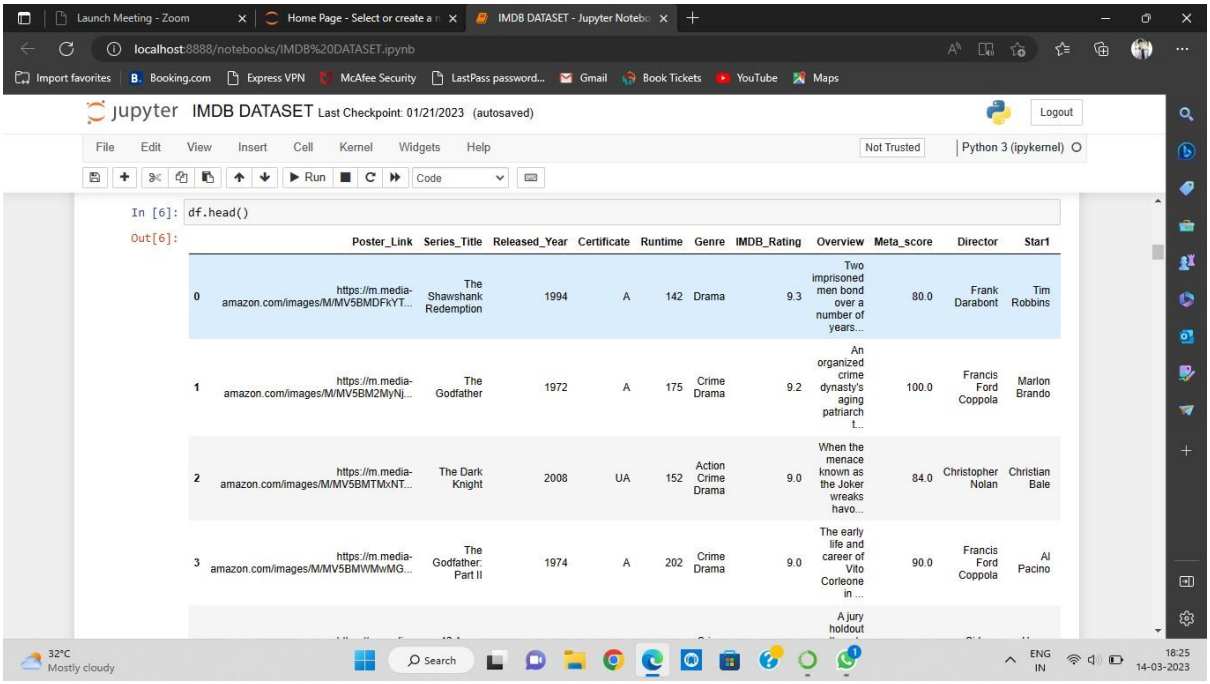
Physical Data Modeling:-

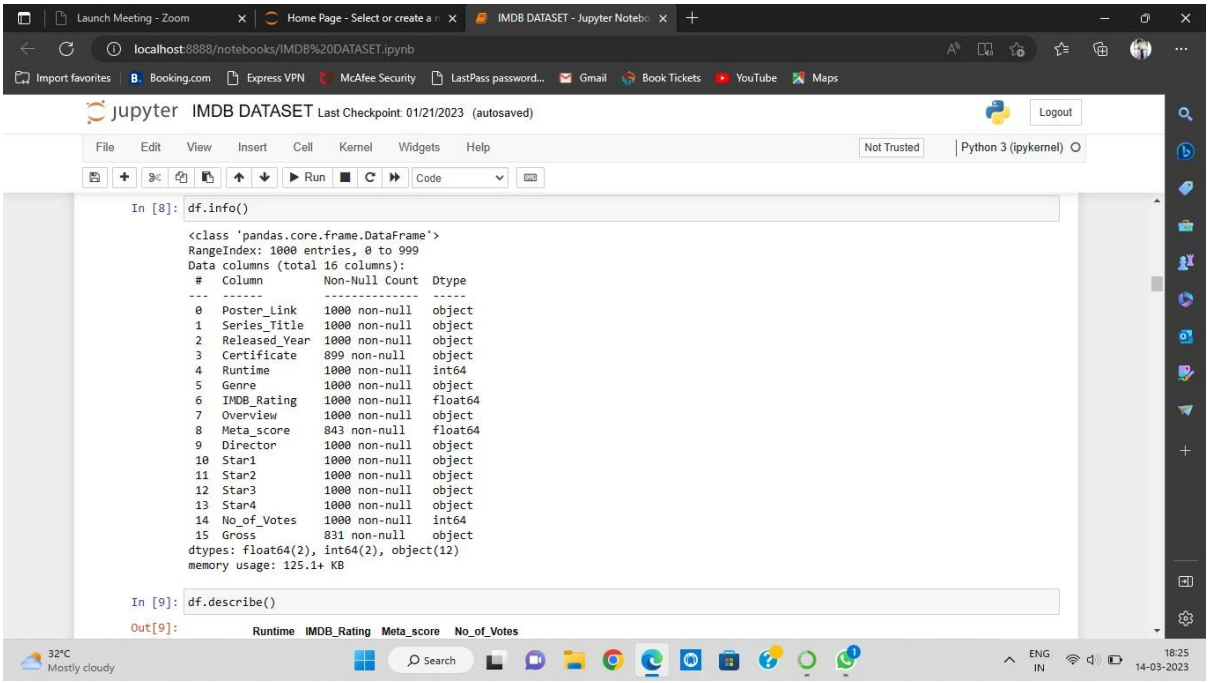
This data model describes how the system will be implemented using a specific DBMS System.

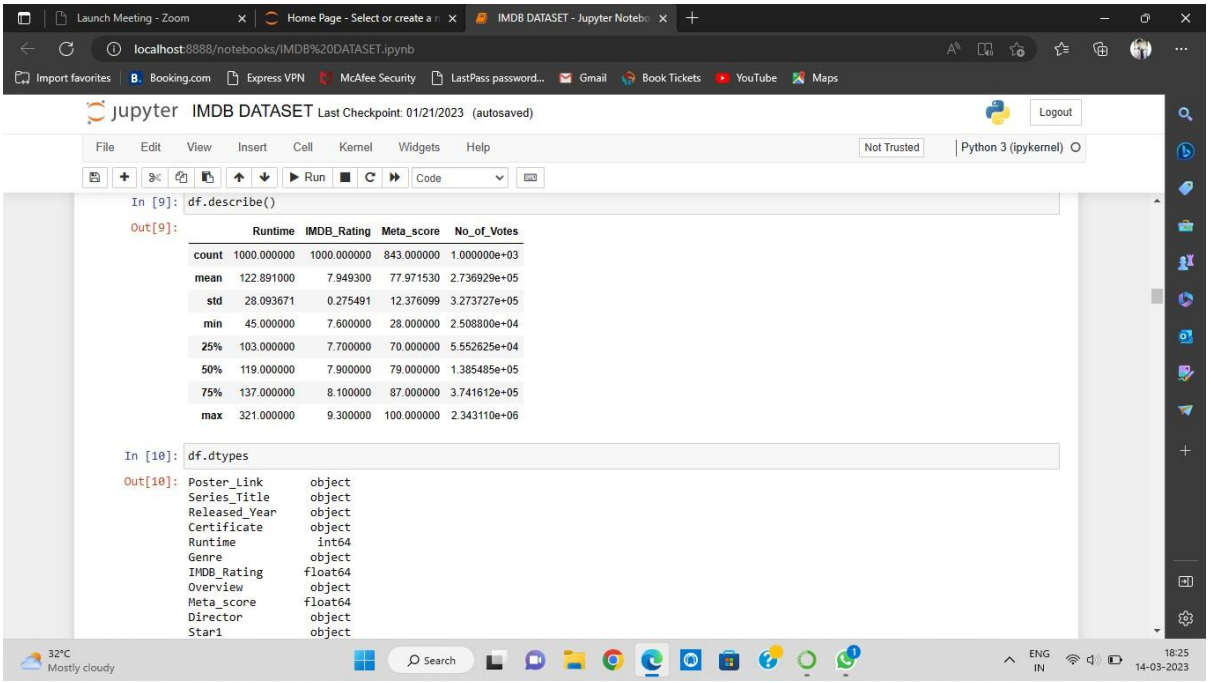
This model is typically created by DBA and devlopers.The purpose is to actual implementation of the database.

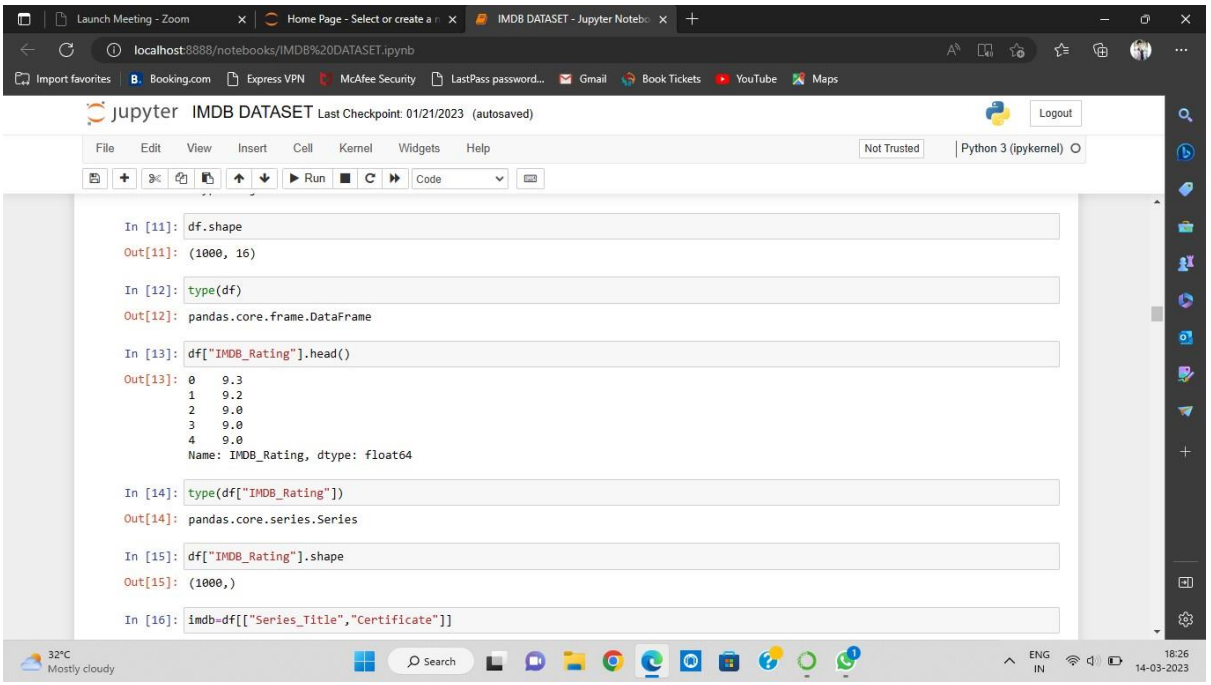
#EDA ON IMDB DTASET...

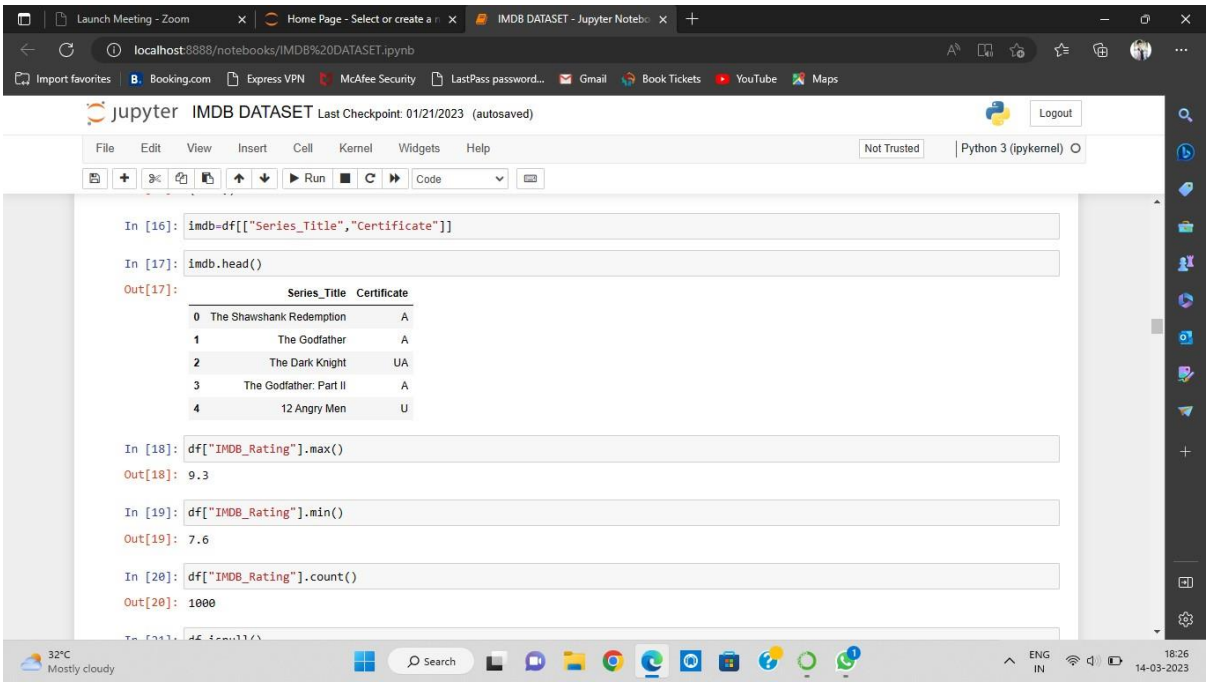


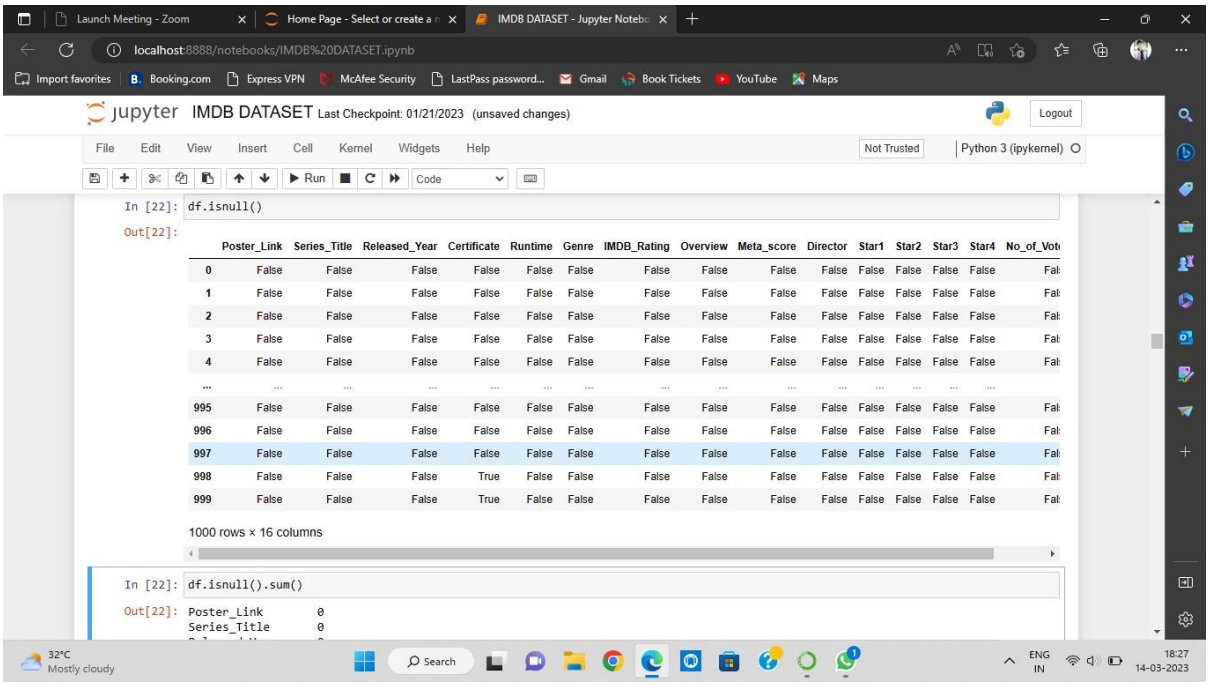


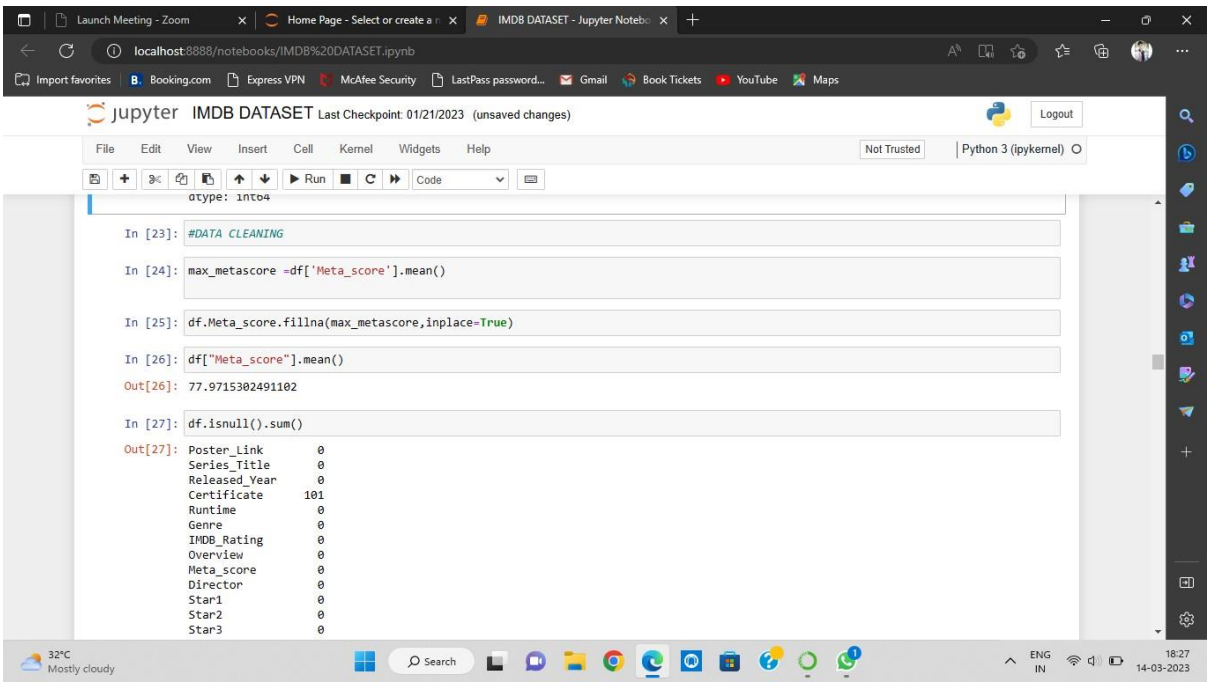

Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [8]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 16 columns):
Column Non-Null Count Dtype
--- ------ -------------- -----
0 Poster_Link 1000 non-null object
1 Series_Title 1000 non-null object
2 Released_Year 1000 non-null object
3 Certificate 899 non-null object
4 Runtime 1000 non-null int64
5 Genre 1000 non-null object
6 IMDB_Rating 1000 non-null float64
7 Overview 1000 non-null object
8 Meta_score 843 non-null float64
9 Director 1000 non-null object
10 Star1 1000 non-null object
11 Star2 1000 non-null object
12 Star3 1000 non-null object
13 Star4 1000 non-null object
14 No_of_Votes 1000 non-null int64
15 Gross 831 non-null object
dtypes: float64(2), int64(2), object(12)
memory usage: 125.1+ KB
In [9]: df.describe()
Out[9]: Runtime IMDB_Rating Meta_score No_of_Votes
32°C Mostly cloudy
Search
ENG IN
18:25 14-03-2023

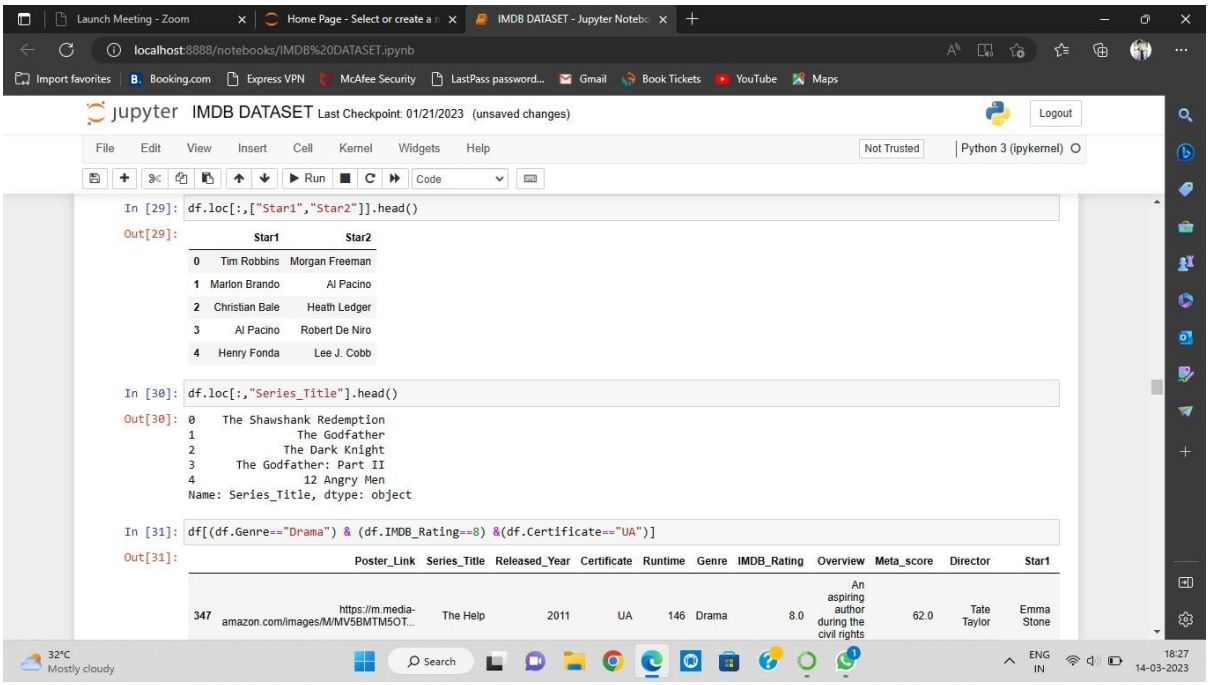

Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [9]: df.describe()
Out[9]:
Runtime IMDB_Rating Meta_score No_of_Votes
count 1000.000000 1000.000000 843.000000 1.000000e+03
mean 122.891000 7.949300 77.971530 2.736929e+05
std 28.093671 0.275491 12.376099 3.273727e+05
min 45.000000 7.600000 28.000000 2.508800e+04
25% 103.000000 7.700000 70.000000 5.552625e+04
50% 119.000000 7.900000 79.000000 1.385485e+05
75% 137.000000 8.100000 87.000000 3.741612e+05
max 321.000000 9.300000 100.000000 2.343110e+06
In [10]: df.dtypes
Out[10]: Poster_Link object
Series_Title object
Released_Year object
Certificate object
Runtime int64
Genre object
IMDB_Rating float64
Overview object
Meta_score float64
Director object
Star1 object
32°C Mostly cloudy
Search
ENG IN
18:25 14-03-2023

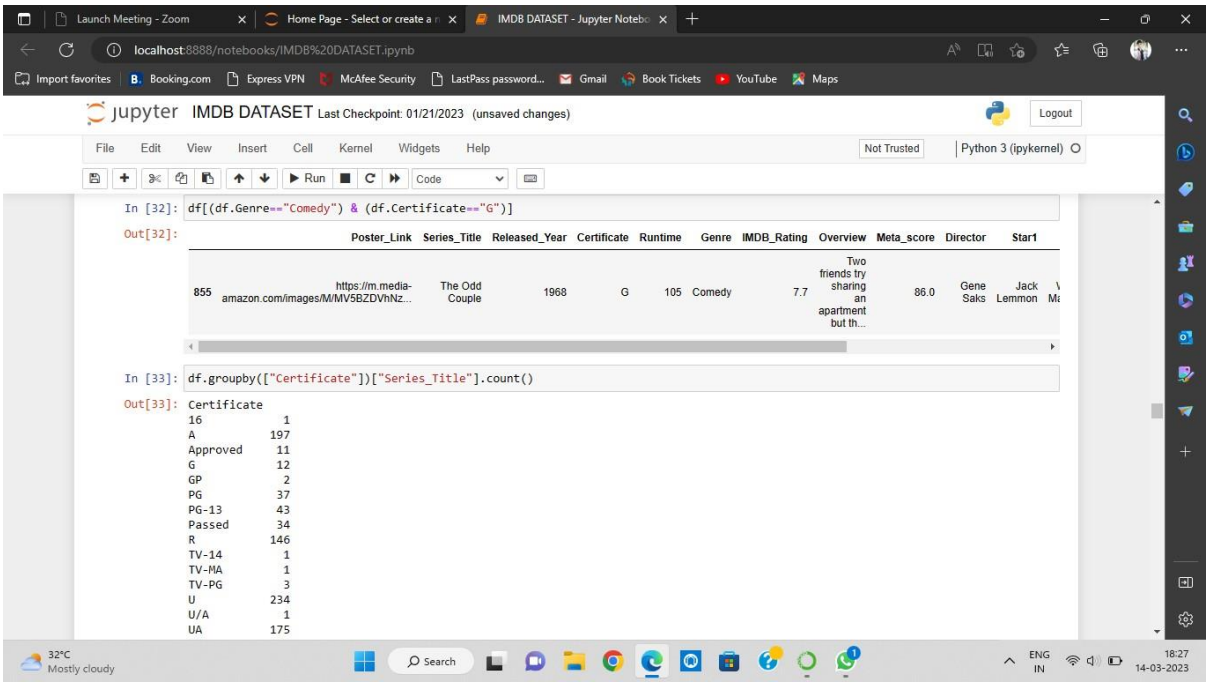

Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [11]: df.shape
Out[11]: (1000, 16)
In [12]: type(df)
Out[12]: pandas.core.frame.DataFrame
In [13]: df["IMDB_Rating"].head()
Out[13]: 0 9.3
1 9.2
2 9.0
3 9.0
4 9.0
Name: IMDB_Rating, dtype: float64
In [14]: type(df["IMDB_Rating"])
Out[14]: pandas.core.series.Series
In [15]: df["IMDB_Rating"].shape
Out[15]: (1000,)
In [16]: imdb=df[["Series_Title","Certificate"]]
32°C
Mostly cloudy
Search
ENG
IN
18:26
14-03-2023


Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [16]: imdb=df[["Series_Title","Certificate"]]
In [17]: imdb.head()
Out[17]:
Series_Title Certificate
0 The Shawshank Redemption A
1 The Godfather A
2 The Dark Knight UA
3 The Godfather: Part II A
4 12 Angry Men U
In [18]: df["IMDB_Rating"].max()
Out[18]: 9.3
In [19]: df["IMDB_Rating"].min()
Out[19]: 7.6
In [20]: df["IMDB_Rating"].count()
Out[20]: 1000
32°C
Mostly cloudy
Search
ENG
IN
18:26
14-03-2023


Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [22]: df.isnull()
Out[22]:
Poster_Link Series_Title Released_Year Certificate Runtime Genre IMDB_Rating Overview Meta_score Director Star1 Star2 Star3 Star4 No_of_Vot
0 False False False False False False False False False False False False False False Fal
1 False False False False False False False False False False False False False False Fal
2 False False False False False False False False False False False False False False Fal
3 False False False False False False False False False False False False False False Fal
4 False False False False False False False False False False False False False False Fal
995 False False False False False False False False False False False False False False Fal
996 False False False False False False False False False False False False False False Fal
997 False False False False False False False False False False False False False False Fal
998 False False False True False False False False False False False False False False Fal
999 False False False True False False False False False False False False False False Fal
1000 rows × 16 columns
In [22]: df.isnull().sum()
Out[22]: Poster_Link 0
Series_Title 0
32°C
Mostly cloudy
Search
ENG
IN
18:27
14-03-2023


Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
dtype: int64
In [23]: #DATA CLEANING
In [24]: max_metascore =df['Meta_score'].mean()
In [25]: df.Meta_score.fillna(max_metascore,inplace=True)
In [26]: df["Meta_score"].mean()
Out[26]: 77.9715302491102
In [27]: df.isnull().sum()
Out[27]: Poster_Link 0
Series_Title 0
Released_Year 0
Certificate 101
Runtime 0
Genre 0
IMDB_Rating 0
Overview 0
Meta_score 0
Director 0
Star1 0
Star2 0
Star3 0
32°C
Mostly cloudy
Search
ENG
IN
18:27
14-03-2023


Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [29]: df.loc[:,["Star1","Star2"]].head()
Out[29]:
Star1 Star2
0 Tim Robbins Morgan Freeman
1 Marlon Brando Al Pacino
2 Christian Bale Heath Ledger
3 Al Pacino Robert De Niro
4 Henry Fonda Lee J. Cobb
In [30]: df.loc[:,"Series_Title"].head()
Out[30]:
0 The Shawshank Redemption
1 The Godfather
2 The Dark Knight
3 The Godfather: Part II
4 12 Angry Men
Name: Series_Title, dtype: object
In [31]: df[(df.Genre=="Drama") & (df.IMDB_Rating==8) &(df.Certificate=="UA")]
Out[31]:
Poster_Link Series_Title Released_Year Certificate Runtime Genre IMDB_Rating Overview Meta_score Director Star1
347 https://m.media-amazon.com/images/M/MV5BMTM5OT... The Help 2011 UA 146 Drama 8.0 An aspiring author during the civil rights 62.0 Tate Taylor Emma Stone
32°C Mostly cloudy
Search
ENG IN
18:27 14-03-2023


Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [32]: df[(df.Genre=="Comedy") & (df.Certificate=="G")]
Out[32]:
Poster_Link Series_Title Released_Year Certificate Runtime Genre IMDB_Rating Overview Meta_score Director Star1
855 https://m.media-amazon.com/images/M/MV5BZDVhNz... The Odd Couple 1968 G 105 Comedy 7.7 Two friends try sharing an apartment but th... 86.0 Gene Saks Jack Lemmon
In [33]: df.groupby(["Certificate"])["Series_Title"].count()
Out[33]:
Certificate
16 1
A 197
Approved 11
G 12
GP 2
PG 37
PG-13 43
Passed 34
R 146
TV-14 1
TV-MA 1
TV-PG 3
U 234
U/A 1
UA 175
32°C Mostly cloudy
Search
ENG IN
18:27 14-03-2023

Launch Meeting - Zoom | Home Page - Select or create a | IMDB DATASET - Jupyter Notebook

localhost:8888/notebooks/IMDB%20DATASET.ipynb

Import favorites | Booking.com | Express VPN | McAfee Security | LastPass password... | Gmail | Book Tickets | YouTube | Maps

jupyter **IMDB DATASET** Last Checkpoint: 01/21/2023 (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

In [34]:
```
df["Series_Title"].shape
```

Out[34]: (1000,)

In [35]:
```
df["Series_Title"].unique().shape
```

Out[35]: (999,)

In [36]:
```
df.sort_values(by="Runtime",ascending=True).head(20)
```

Out[36]:

	Poster_Link	Series_Title	Released_Year	Certificate	Runtime	Genre	IMDB_Rating	Overview	Meta_score	Director
194	https://m.media-amazon.com/images/M/MV5BZWFhOG...	Sherlock Jr.	1924	Passed	45	Action Comedy Romance	8.2	A film projectionist longs to be a detective ...	77.97153	Buster Keaton
567	https://m.media-amazon.com/images/M/MV5BMjMyYj...	Freaks	1932	NaN	64	Drama Horror	7.9	A circus' beautiful trapeze artist agrees to m...	80.00000	Tod Browning
320	https://m.media-amazon.com/images/M/MV5BYmRiMD...	The General	1926	Passed	67	Action Adventure Comedy	8.1	When Union spies steal an engineer's beloved l...	77.97153	Clyde Bruckman
127	https://m.media-amazon.com/images/M/MV5BZjhhMT...	The Kid	1921	Passed	68	Comedy Drama Family	8.3	The Tramp cares for an abandoned child but ...	77.97153	Charles Chaplin
747	https://m.media-	Duck Soup	1933	NaN	69	Comedy Musical	7.8	Rufus T. Firefly is named	93.00000	Leo

32°C Mostly cloudy | Search | ENG IN | 18:27 14-03-2023

Launch Meeting - Zoom | Home Page - Select or create a | IMDB DATASET - Jupyter Notebook

localhost:8888/notebooks/IMDB%20DATASET.ipynb

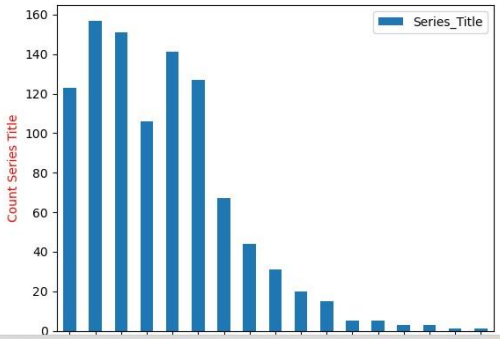
Import favorites | Booking.com | Express VPN | McAfee Security | LastPass password... | Gmail | Book Tickets | YouTube | Maps

jupyter **IMDB DATASET** Last Checkpoint: 01/21/2023 (unsaved changes) Logout

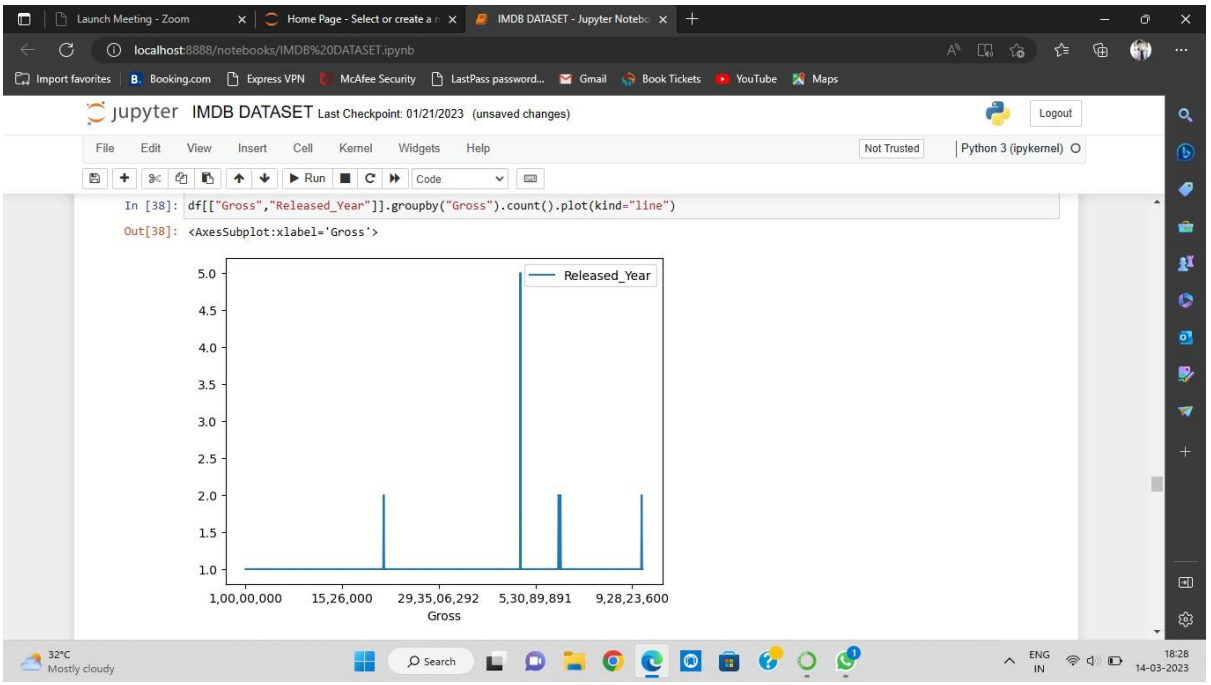
File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

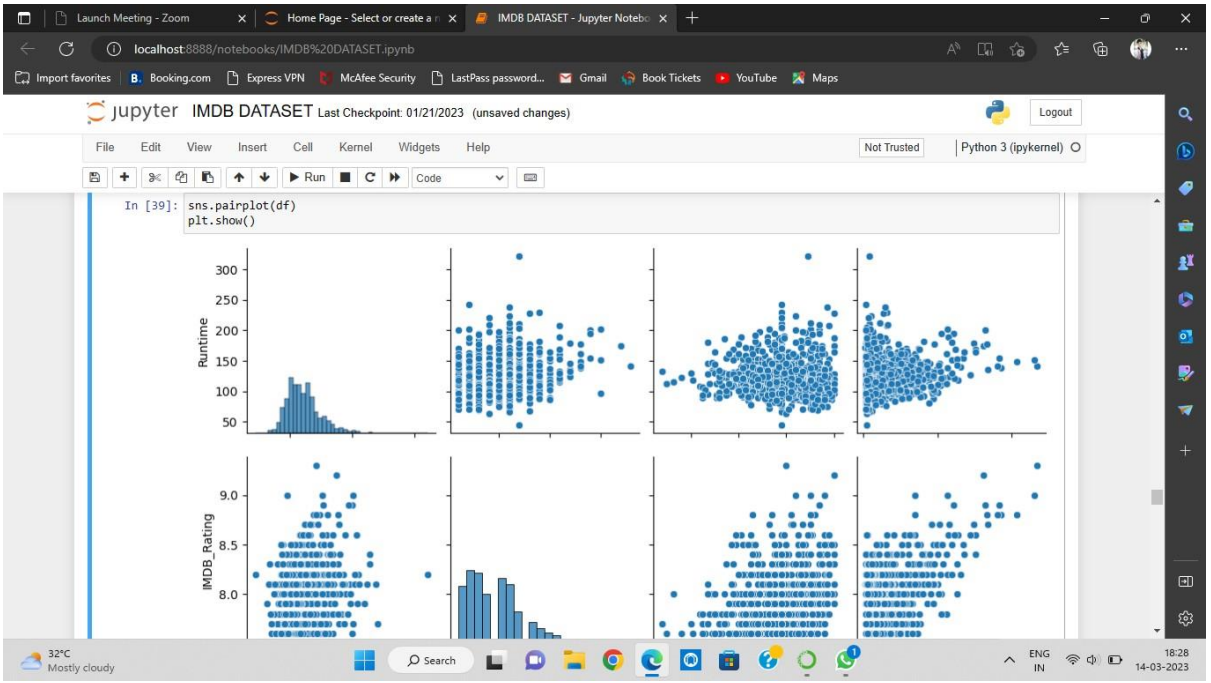
Run Code

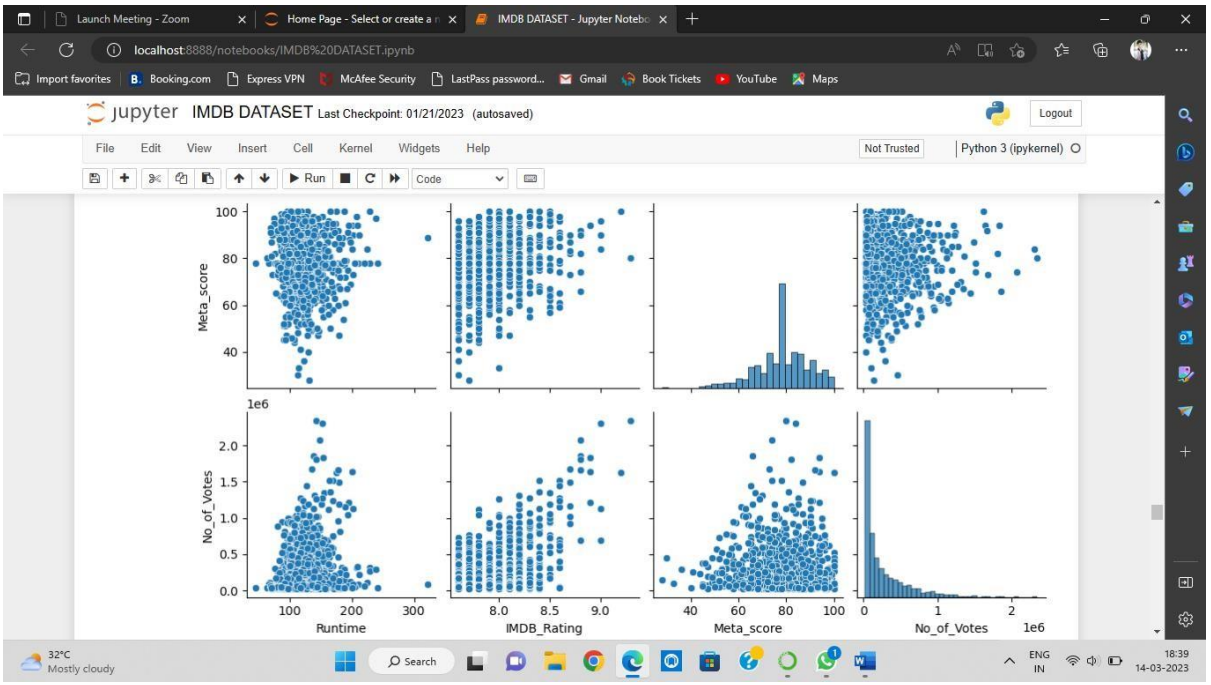
In [37]:
```
df[["IMDB_Rating","Series_Title"]].groupby("IMDB_Rating").count().plot(kind="bar")

plt.xlabel("IMDB RATING",color="r")
plt.ylabel("Count Series Title",color="r")
plt.show()
```

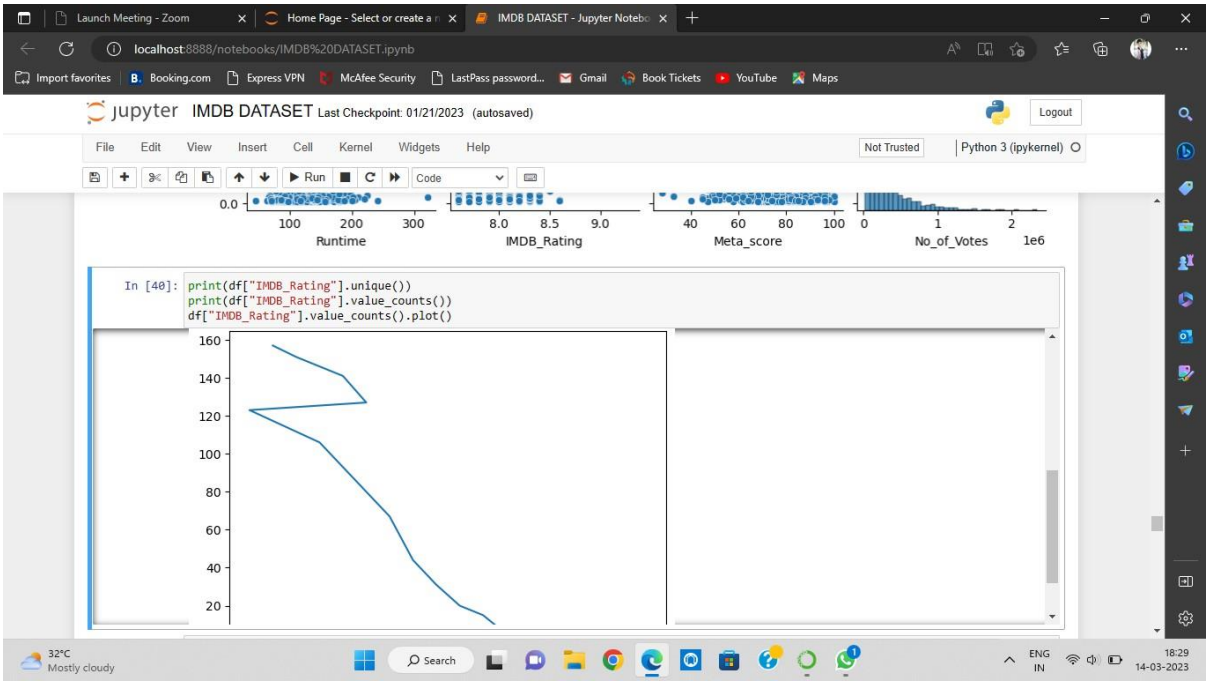


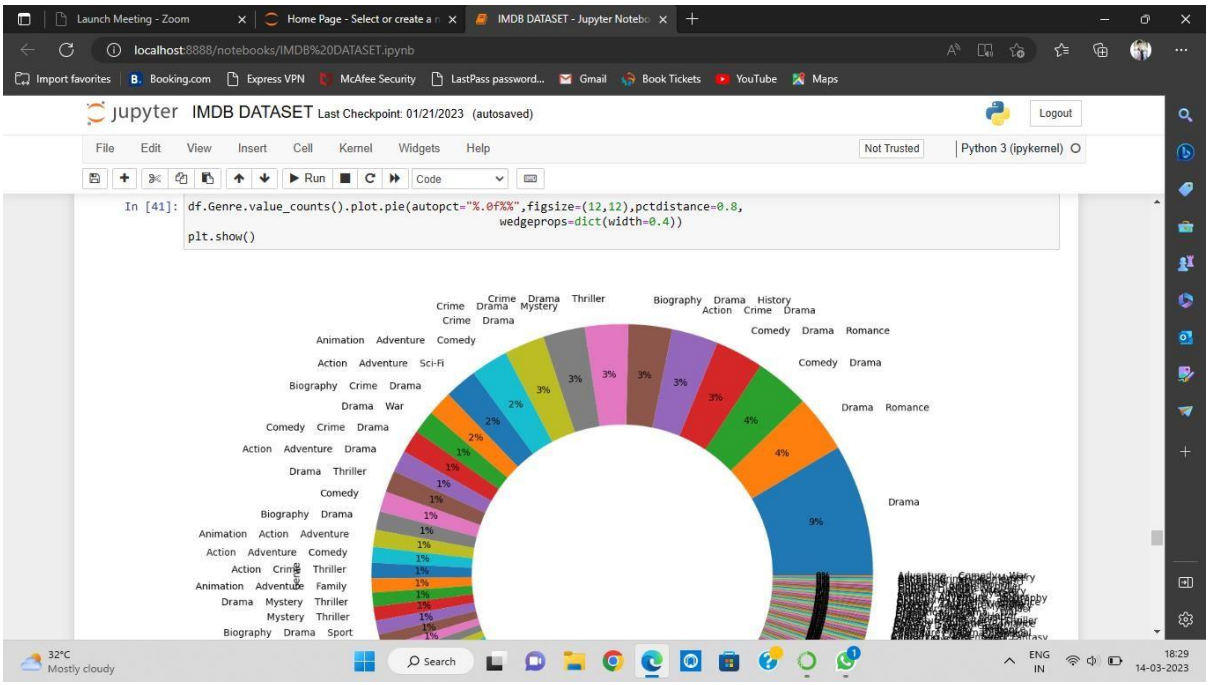
32°C Mostly cloudy | Search | ENG IN | 18:28 14-03-2023

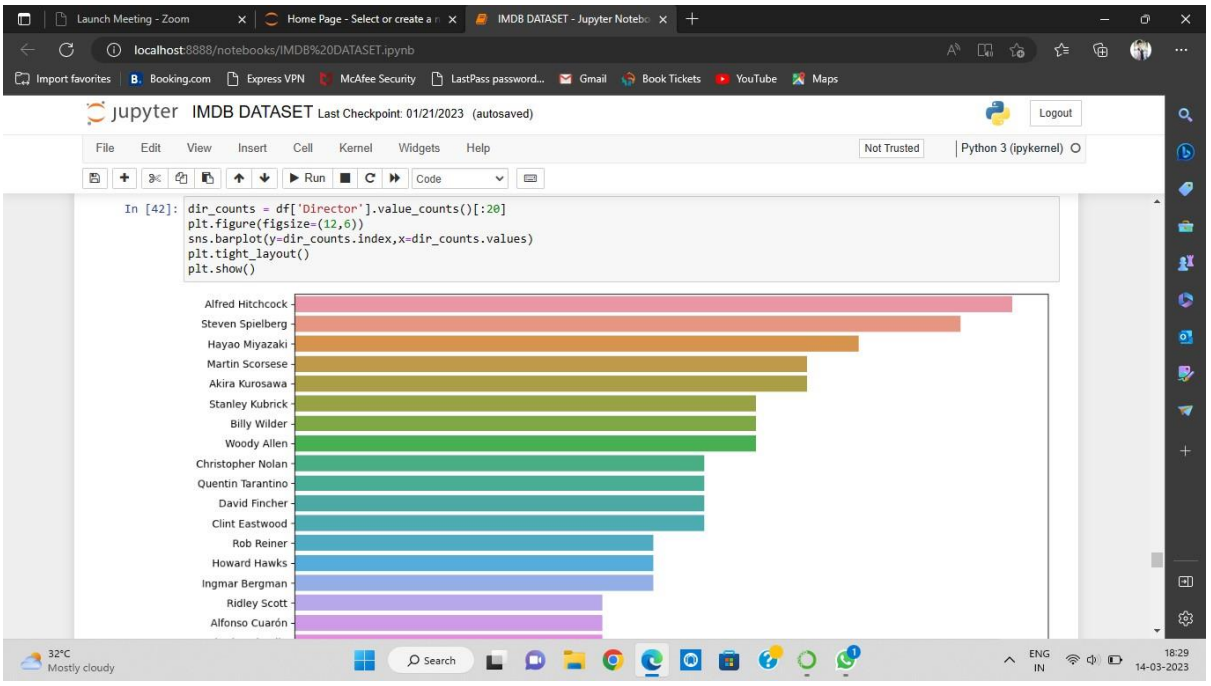

Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [38]: df[["Gross","Released_Year"]].groupby("Gross").count().plot(kind="line")
Out[38]: <AxesSubplot:xlabel='Gross'>
Released_Year
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
1,00,00,000
15,26,000
29,35,06,292
5,30,89,891
9,28,23,600
Gross
32°C
Mostly cloudy
Search
ENG
IN
18:28
14-03-2023

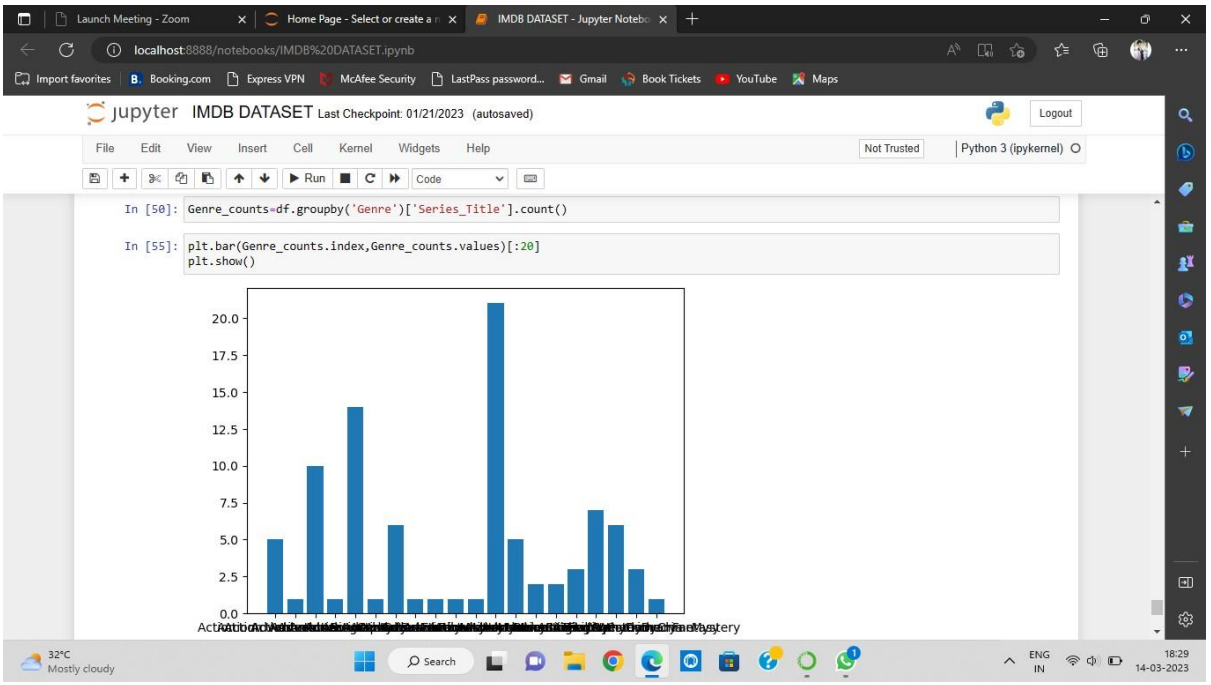

Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [39]: sns.pairplot(df)
plt.show()
Runtime
300
250
200
150
100
50
IMDB_Rating
9.0
8.5
8.0
32°C
Mostly cloudy
Search
ENG
IN
18:28
14-03-2023


Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Meta_score
No_of_Votes
1e6
Runtime
IMDB_Rating
Meta_score
No_of_Votes
1e6
32°C
Mostly cloudy
Search
ENG
IN
18:39
14-03-2023

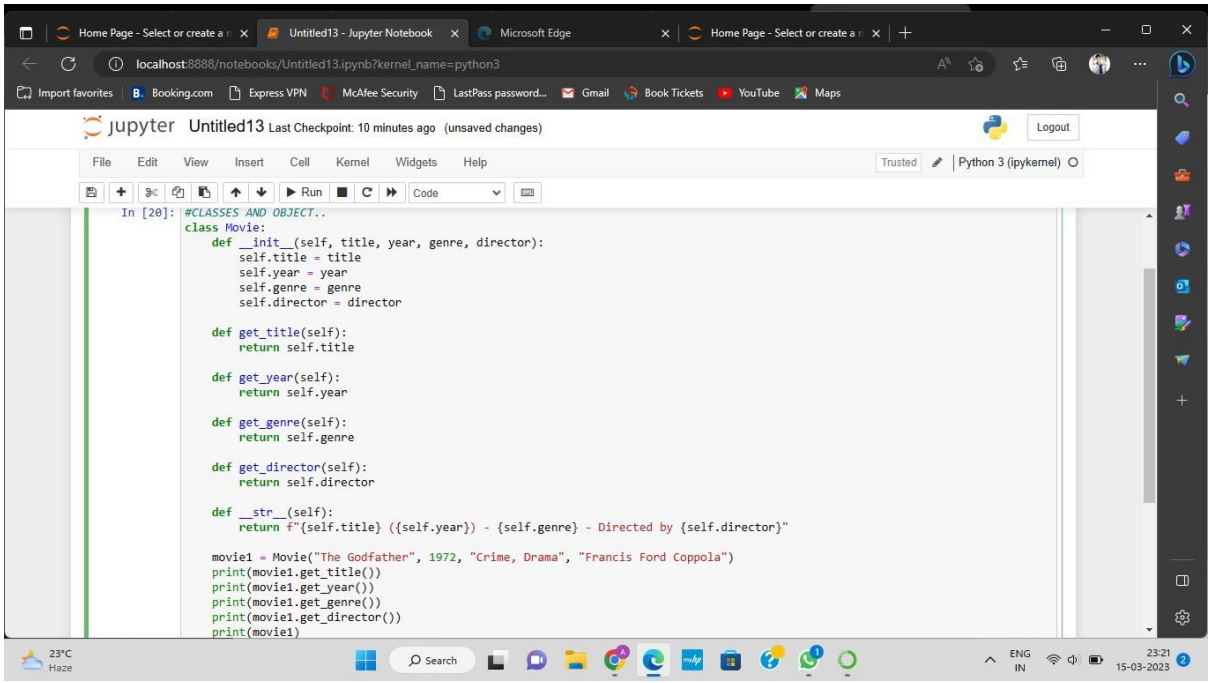

Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Runtime
IMDB_Rating
Meta_score
No_of_Votes
1e6
In [40]:
print(df["IMDB_Rating"].unique())
print(df["IMDB_Rating"].value_counts())
df["IMDB_Rating"].value_counts().plot()
32°C
Mostly cloudy
Search
ENG
IN
18:29
14-03-2023

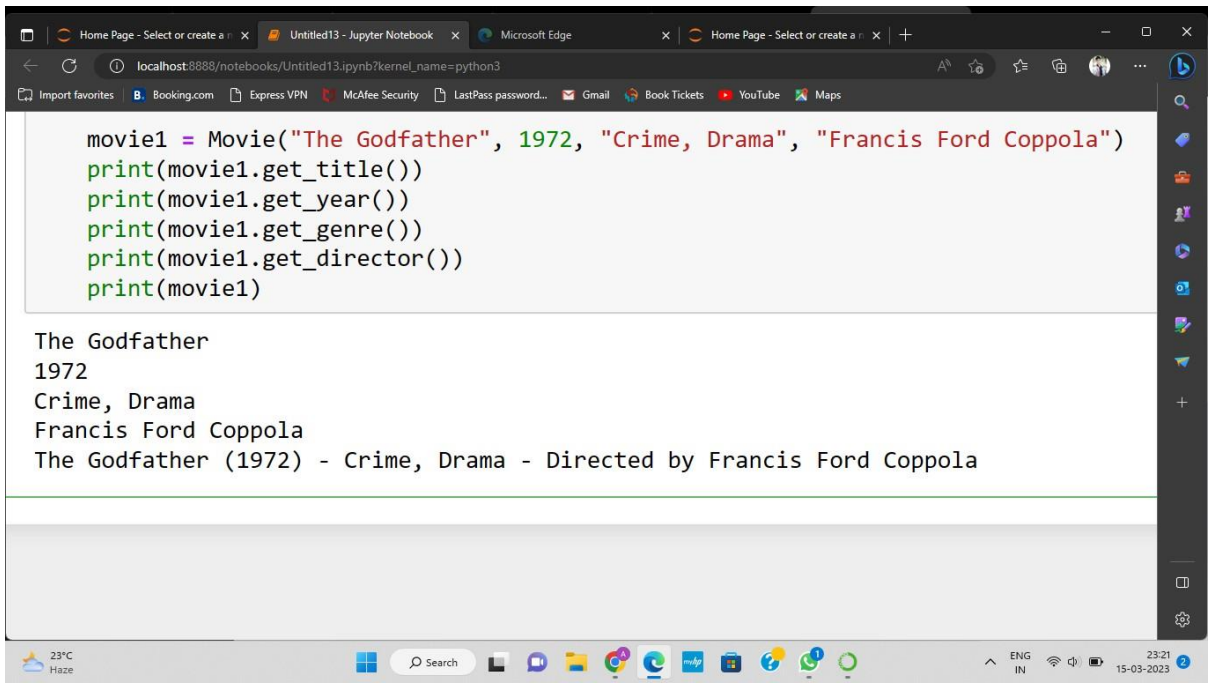

Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [41]:
df.Genre.value_counts().plot.pie(autopct="%.0f%%",figsize=(12,12),pctdistance=0.8,
wedgeprops=dict(width=0.4))
plt.show()
Crime Drama Thriller
Crime Drama Mystery
Crime Drama
Biography Drama History
Action Crime Drama
Comedy Drama Romance
Comedy Drama
Drama Romance
Drama
Animation Adventure Comedy
Action Adventure Sci-Fi
Biography Crime Drama
Drama War
Comedy Crime Drama
Action Adventure Drama
Drama Thriller
Comedy
Biography Drama
Animation Action Adventure
Action Adventure Comedy
Action Crime Thriller
Animation Adventure Family
Drama Mystery Thriller
Mystery Thriller
Biography Drama Sport
3%
3%
3%
3%
3%
4%
4%
9%
2%
2%
1%
1%
1%
1%
1%
1%
1%
1%
1%
1%
1%
1%
1%
1%
1%
32°C
Mostly cloudy
Search
ENG
IN
18:29
14-03-2023

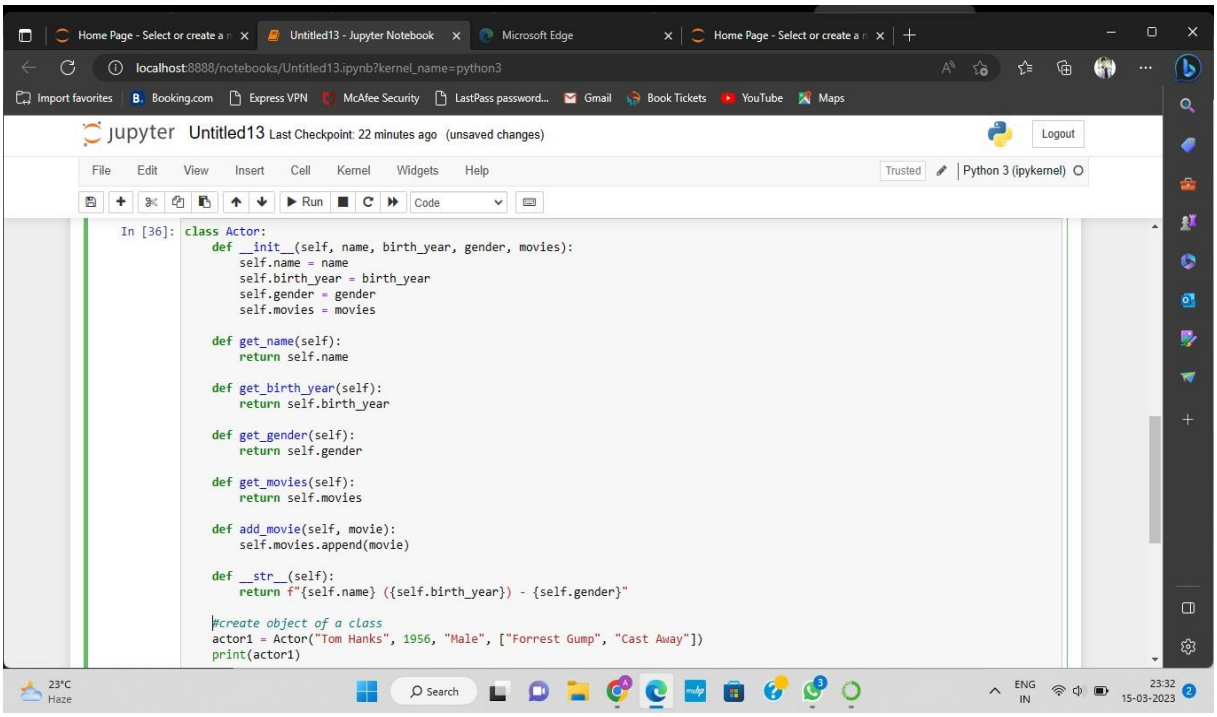

Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [42]:
dir_counts = df['Director'].value_counts()[:20]
plt.figure(figsize=(12,6))
sns.barplot(y=dir_counts.index,x=dir_counts.values)
plt.tight_layout()
plt.show()
Alfred Hitchcock
Steven Spielberg
Hayao Miyazaki
Martin Scorsese
Akira Kurosawa
Stanley Kubrick
Billy Wilder
Woody Allen
Christopher Nolan
Quentin Tarantino
David Fincher
Clint Eastwood
Rob Reiner
Howard Hawks
Ingmar Bergman
Ridley Scott
Alfonso Cuarón
32°C
Mostly cloudy
Search
ENG
IN
18:29
14-03-2023

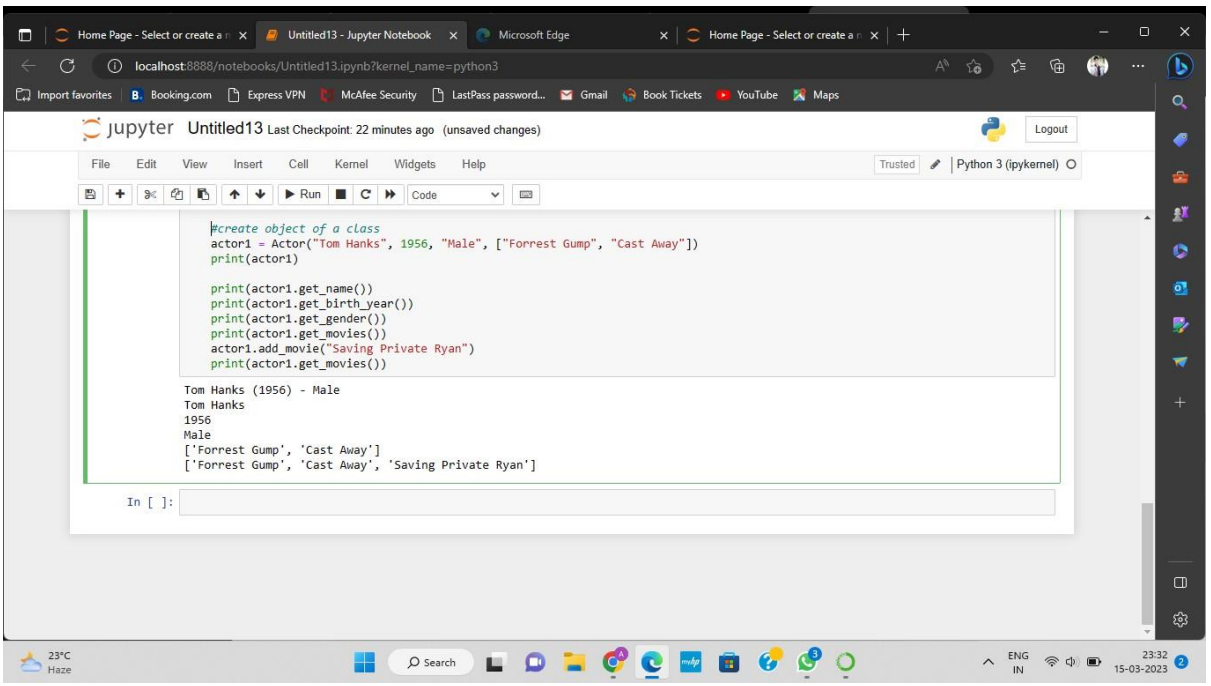

Launch Meeting - Zoom
Home Page - Select or create a
IMDB DATASET - Jupyter Notebo
localhost:8888/notebooks/IMDB%20DATASET.ipynb
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter IMDB DATASET Last Checkpoint: 01/21/2023 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [50]:
Genre_counts=df.groupby('Genre')['Series_Title'].count()
In [55]:
plt.bar(Genre_counts.index,Genre_counts.values)[:20]
plt.show()
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Mystery
32°C
Mostly cloudy
Search
ENG
IN
18:29
14-03-2023

CLASSES AND OBJECT…






Home Page - Select or create a n
Untitled13 - Jupyter Notebook
Microsoft Edge
Home Page - Select or create a n
localhost:8888/notebooks/Untitled13.ipynb?kernel_name=python3
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter Untitled13 Last Checkpoint: 22 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [36]:
class Actor:
def __init__(self, name, birth_year, gender, movies):
self.name = name
self.birth_year = birth_year
self.gender = gender
self.movies = movies
def get_name(self):
return self.name
def get_birth_year(self):
return self.birth_year
def get_gender(self):
return self.gender
def get_movies(self):
return self.movies
def add_movie(self, movie):
self.movies.append(movie)
def __str__(self):
return f"{self.name} ({self.birth_year}) - {self.gender}"
#create object of a class
actor1 = Actor("Tom Hanks", 1956, "Male", ["Forrest Gump", "Cast Away"])
print(actor1)
23°C
Haze
Search
ENG
IN
23:32
15-03-2023

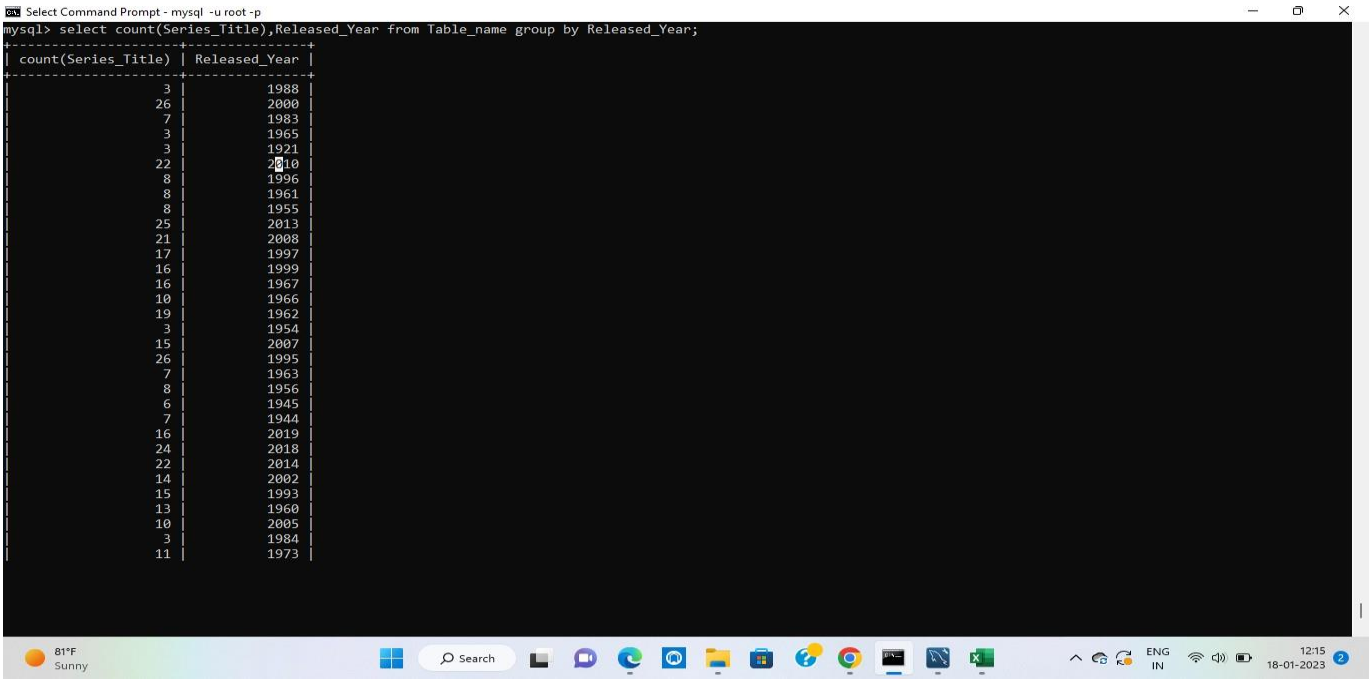

Home Page - Select or create a n
Untitled13 - Jupyter Notebook
Microsoft Edge
Home Page - Select or create a n
localhost:8888/notebooks/Untitled13.ipynb?kernel_name=python3
Import favorites
Booking.com
Express VPN
McAfee Security
LastPass password...
Gmail
Book Tickets
YouTube
Maps
jupyter Untitled13 Last Checkpoint: 22 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
#create object of a class
actor1 = Actor("Tom Hanks", 1956, "Male", ["Forrest Gump", "Cast Away"])
print(actor1)
print(actor1.get_name())
print(actor1.get_birth_year())
print(actor1.get_gender())
print(actor1.get_movies())
actor1.add_movie("Saving Private Ryan")
print(actor1.get_movies())
Tom Hanks (1956) - Male
Tom Hanks
1956
Male
['Forrest Gump', 'Cast Away']
['Forrest Gump', 'Cast Away', 'Saving Private Ryan']
In [ ]:
23°C
Haze
Search
ENG
IN
23:32
15-03-2023

Analytical queries on IMDB Dataset project

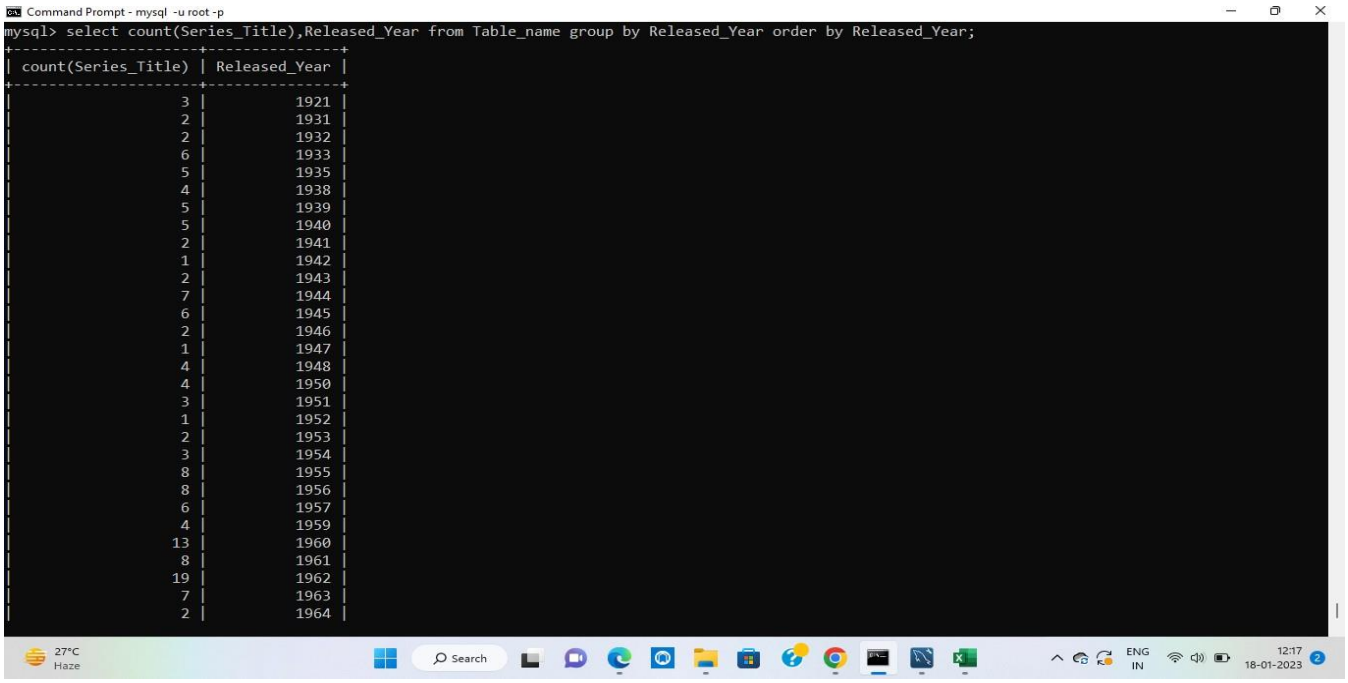
1).

```
mysql> desc table_name;
+--------------+--------------+------+-----+---------+-------+
| Field        | Type         | Null | Key | Default | Extra |
+--------------+--------------+------+-----+---------+-------+
| Poster_Link  | varchar(200) | YES  |     | NULL    |       |
| Series_Title | varchar(100) | YES  |     | NULL    |       |
| Released_Year| int          | YES  |     | NULL    |       |
| Certificate  | varchar(100) | YES  |     | NULL    |       |
| Runtime      | varchar(100) | YES  |     | NULL    |       |
| Genre        | varchar(100) | YES  |     | NULL    |       |
| IMDB_Rating  | decimal(10,2)| YES  |     | NULL    |       |
| Overview     | varchar(500) | YES  |     | NULL    |       |
| Meta_score   | int          | YES  |     | NULL    |       |
| Director     | varchar(100) | YES  |     | NULL    |       |
| Star1        | varchar(100) | YES  |     | NULL    |       |
| Star2        | varchar(100) | YES  |     | NULL    |       |
| Star3        | varchar(100) | YES  |     | NULL    |       |
| Star4        | varchar(100) | YES  |     | NULL    |       |
| No_of_Votes  | int          | YES  |     | NULL    |       |
| Gross        | varchar(100) | YES  |     | NULL    |       |
+--------------+--------------+------+-----+---------+-------+
16 rows in set (0.01 sec)
```

2).



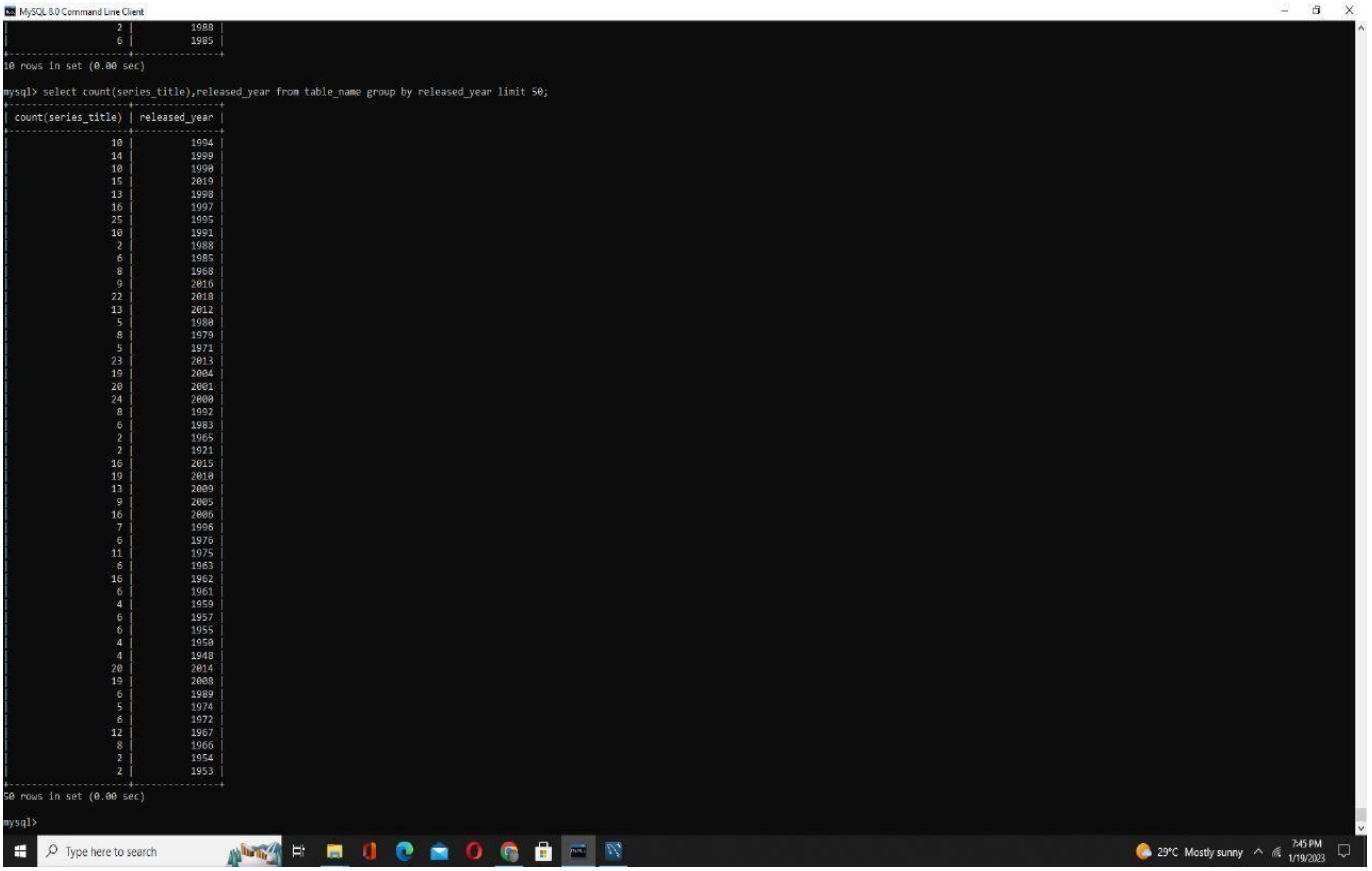
3).


Command Prompt - mysql -u root -p
mysql> select count(Series_Title),Released_Year from Table_name group by Released_Year order by Released_Year;
count(Series_Title)	Released_Year
3	1921
2	1931
2	1932
6	1933
5	1935
4	1938
5	1939
5	1940
2	1941
1	1942
2	1943
7	1944
6	1945
2	1946
1	1947
4	1948
4	1950
3	1951
1	1952
2	1953
3	1954
8	1955
8	1956
6	1957
4	1959
13	1960
8	1961
19	1962
7	1963
2	1964
27°C
Haze
Search
ENG
IN
12:17
18-01-2023

4).

```
mysql> select series_title,imdb_rating,meta_score from table_name where meta_score>95;
+---------------------------------+-------------+------------+
| series_title                    | imdb_rating | meta_score |
+---------------------------------+-------------+------------+
| Gisaengchung                    |        8.60 |         96 |
| Some Like It Hot                |        8.20 |         98 |
| Du rififi chez les hommes       |        8.20 |         97 |
| All About Eve                   |        8.20 |         98 |
| The Treasure of the Sierra Madre |        8.20 |         98 |
| 12 Years a Slave                |        8.10 |         96 |
| La battaglia di Algeri          |        8.10 |         96 |
| Ratatouille                     |        8.00 |         96 |
| The Maltese Falcon              |        8.00 |         96 |
| The Grapes of Wrath             |        8.00 |         96 |
| Boyhood                         |        7.90 |        100 |
| 4 luni, 3 saptam├óni si 2 zile  |        7.90 |         97 |
| The Lady Vanishes               |        7.80 |         98 |
| Gravity                         |        7.70 |         96 |
| Fantasia                        |        7.70 |         96 |
| Gisaengchung                    |        8.60 |         96 |
| Some Like It Hot                |        8.20 |         98 |
| Du rififi chez les hommes       |        8.20 |         97 |
| All About Eve                   |        8.20 |         98 |
| The Treasure of the Sierra Madre |        8.20 |         98 |
| 12 Years a Slave                |        8.10 |         96 |
| La battaglia di Algeri          |        8.10 |         96 |
| Ratatouille                     |        8.00 |         96 |
| The Maltese Falcon              |        8.00 |         96 |
| The Grapes of Wrath             |        8.00 |         96 |
| Boyhood                         |        7.90 |        100 |
| 4 luni, 3 saptam├óni si 2 zile  |        7.90 |         97 |
| The Lady Vanishes               |        7.80 |         98 |
| Gravity                         |        7.70 |         96 |
| Fantasia                        |        7.70 |         96 |
+---------------------------------+-------------+------------+
30 rows in set (0.00 sec)
```

5).


MySQL 8.0 Command Line Client
2 | 1988 |
6 | 1985 |
10 rows in set (0.00 sec)
mysql> select count(series_title),released_year from table_name group by released_year limit 50;
| count(series_title) | released_year |
10 | 1994 |
14 | 1999 |
10 | 1990 |
15 | 2019 |
13 | 1998 |
16 | 1997 |
25 | 1995 |
10 | 1991 |
2 | 1988 |
6 | 1985 |
8 | 1968 |
9 | 2016 |
22 | 2018 |
13 | 2012 |
5 | 1980 |
8 | 1979 |
5 | 1971 |
23 | 2013 |
19 | 2004 |
20 | 2001 |
24 | 2000 |
8 | 1992 |
6 | 1983 |
2 | 1966 |
2 | 1921 |
16 | 2015 |
19 | 2010 |
13 | 2009 |
9 | 2005 |
16 | 2006 |
7 | 1998 |
6 | 1976 |
11 | 1975 |
6 | 1963 |
16 | 1962 |
6 | 1961 |
4 | 1959 |
6 | 1957 |
6 | 1955 |
4 | 1950 |
4 | 1948 |
20 | 2014 |
19 | 2008 |
6 | 1989 |
5 | 1974 |
6 | 1972 |
12 | 1967 |
8 | 1966 |
2 | 1954 |
2 | 1953 |
50 rows in set (0.00 sec)
mysql>
Type here to search
29°C Mostly sunny
7:45 PM
1/19/2023

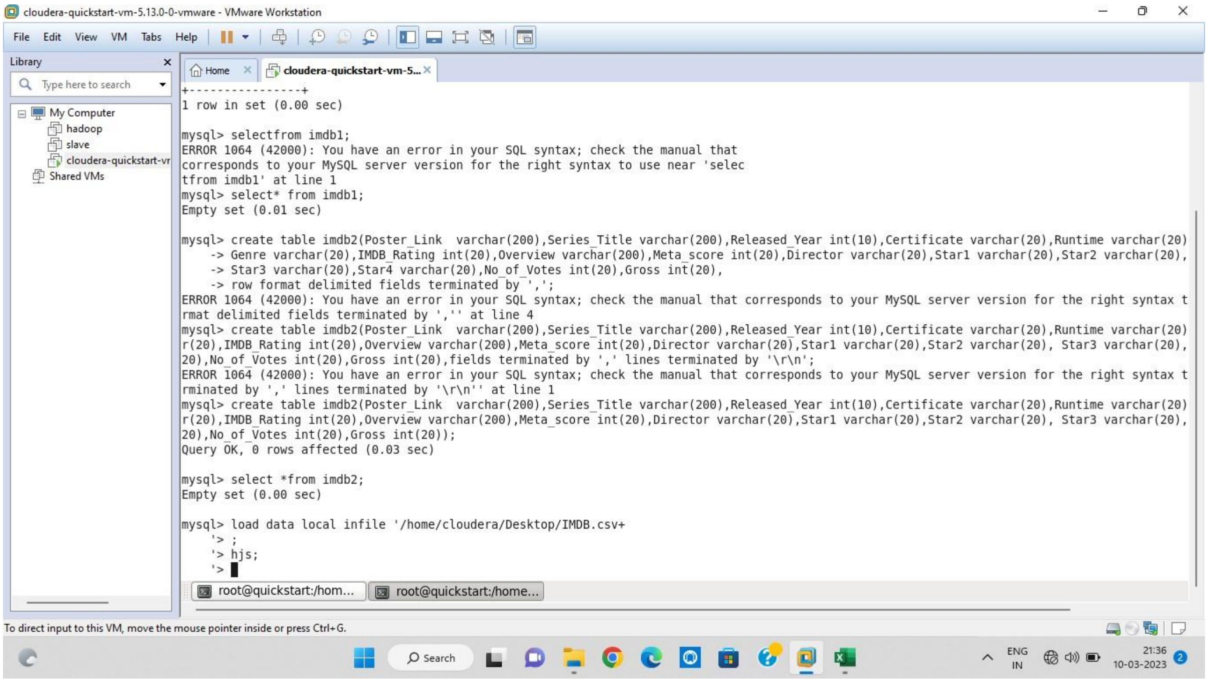
6).

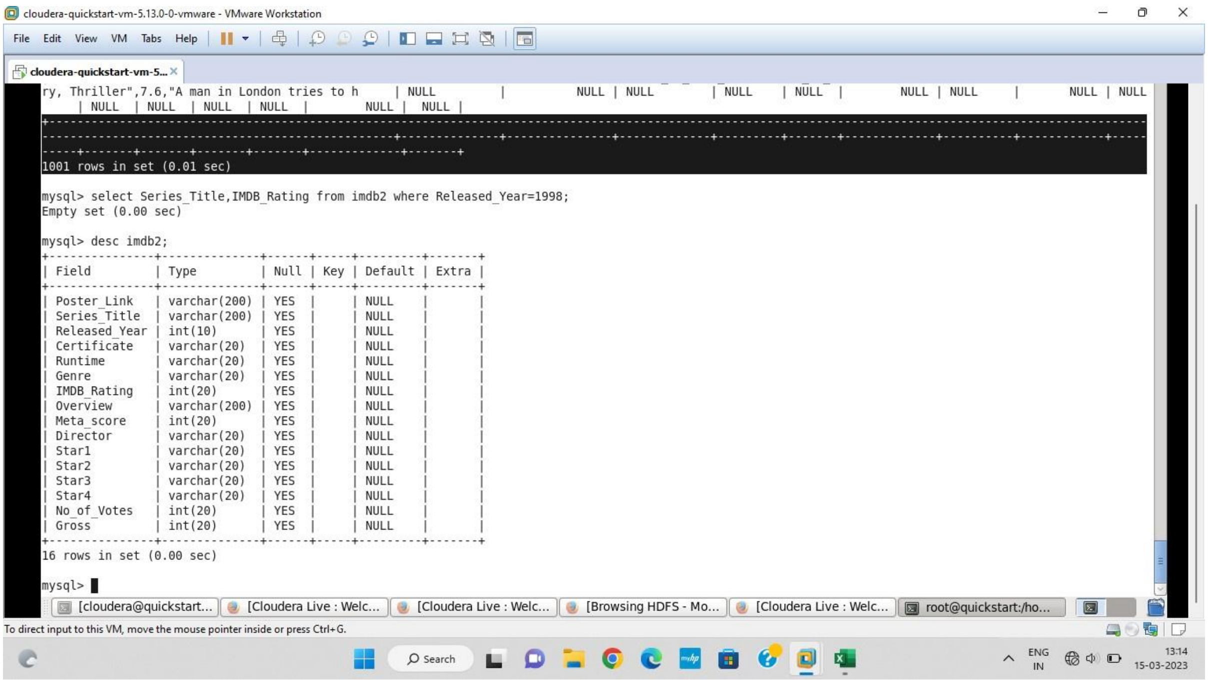
```
mysql> select count(director),director from table_name group by director limit 50;
+----------------+-------------------------+
| count(director) | director                |
+----------------+-------------------------+
|              4 | Frank Darabont          |
|             10 | Quentin Tarantino       |
|              2 | Lana Wachowski          |
|             11 | Martin Scorsese         |
|              2 | Bong Joon Ho            |
|             11 | Steven Spielberg        |
|              2 | Roberto Benigni         |
|              7 | David Fincher           |
|              4 | Jonathan Demme          |
|              4 | Bryan Singer            |
|              4 | Isao Takahata           |
|              4 | Robert Zemeckis         |
|              6 | Sergio Leone            |
|              2 | Makoto Shinkai          |
|              4 | Nitesh Tiwari           |
|              2 | Bob Persichetti         |
|              8 | Stanley Kubrick         |
|              6 | Francis Ford Coppola    |
|              4 | Ridley Scott            |
|              2 | Hrishikesh Mukherjee    |
|              2 | Rahi Anil Barve         |
|              2 | Sriram Raghavan         |
|              2 | Jeethu Joseph           |
|              2 | Michel Gondry           |
|              2 | Jean-Pierre Jeunet      |
|              5 | Guy Ritchie             |
|              4 | Darren Aronofsky        |
|              5 | Mel Gibson              |
|              3 | Brian De Palma          |
|              2 | Charles Chaplin         |
|              2 | Aditya Dhar             |
|              2 | Prashanth Neel          |
|              2 | Peter Farrelly          |
|              2 | Nishikant Kamat         |
|              2 | Vikas Bahl              |
|              4 | Rakeysh Omprakash Mehra |
|              3 | Anurag Kashyap          |
|              2 | Vikramaditya Motwane    |
|              2 | Tigmanshu Dhulia        |
|              5 | Pete Docter             |
|              2 | James McTeigue          |
|              6 | Ron Howard              |
|              2 | Priyadarshan            |
|              2 | Curtis Hanson           |
|              2 | Yavuz Turgul            |
|              2 | Michael Mann            |
|              2 | Rajkumar Santoshi       |
|             10 | Clint Eastwood          |
|              2 | Moustapha Akkad         |
|              2 | Ramesh Sippy            |
+----------------+-------------------------+
50 rows in set (0.00 sec)
```

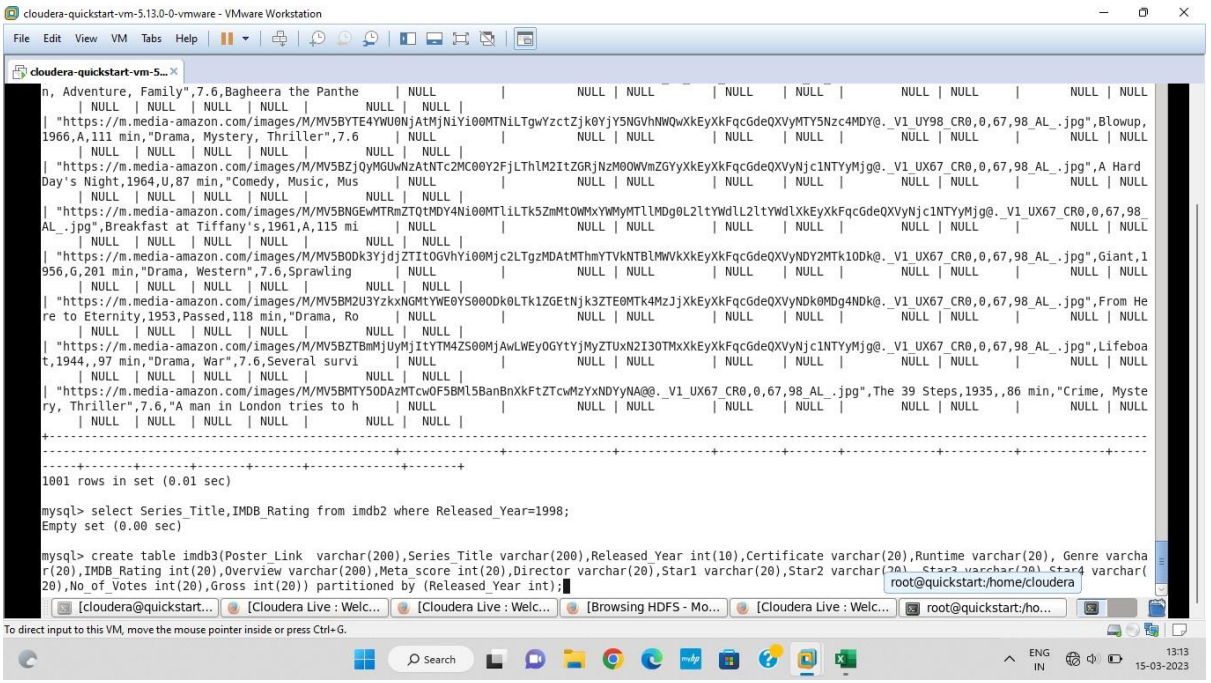
7).

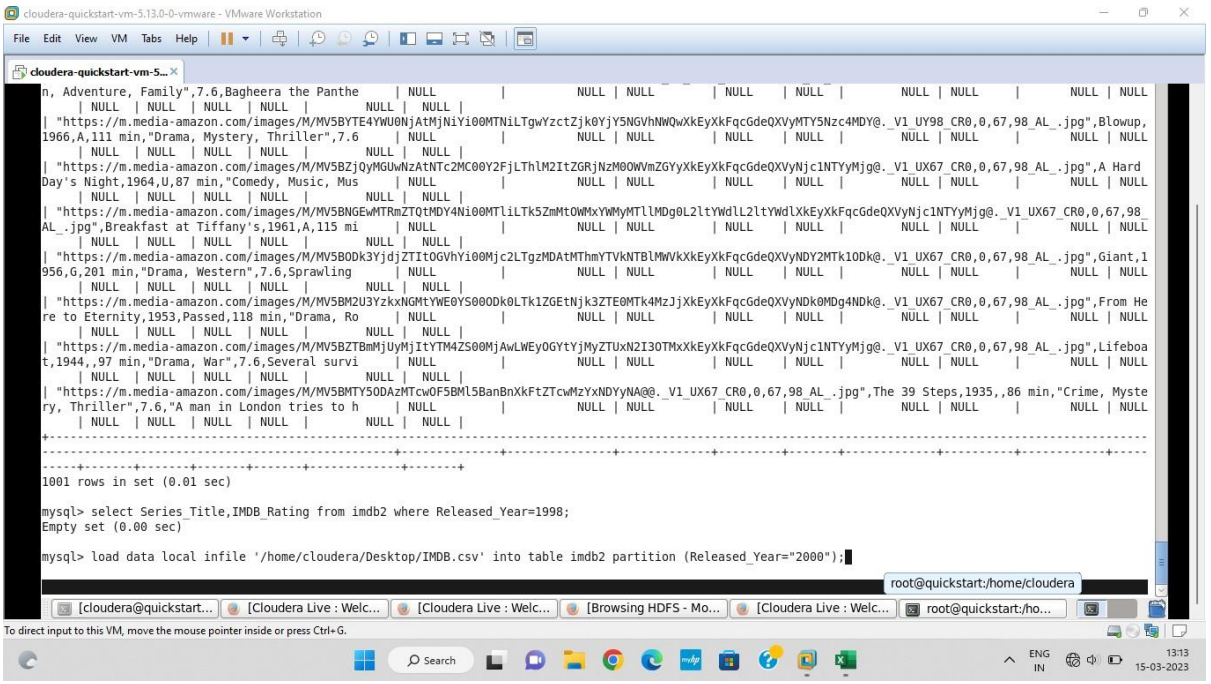
```
mysql> select count(series_title),genre from table_name group by genre  limit 50 ;
+---------------------+-----------------------------+
| count(series_title) | genre                       |
+---------------------+-----------------------------+
|                  43 | Drama                       |
|                  20 | Crime, Drama                |
|                   2 | Action, Sci-Fi              |
|                  15 | Biography, Crime, Drama     |
|                   2 | Comedy, Drama, Thriller     |
|                  14 | Drama, War                  |
|                   4 | Crime, Drama, Fantasy       |
|                  21 | Comedy, Drama, Romance      |
|                  22 | Crime, Drama, Mystery       |
|                  27 | Crime, Drama, Thriller      |
|                   3 | Crime, Mystery, Thriller    |
|                   2 | Animation, Drama, War       |
|                   4 | Adventure, Comedy, Sci-Fi   |
|                   6 | Western                     |
|                   5 | Animation, Drama, Fantasy   |
|                   8 | Action, Biography, Drama    |
|                   9 | Animation, Action, Adventure |
|                   8 | Drama, Western              |
|                   4 | Drama, Horror               |
|                   2 | Drama, Mystery, War         |
|                   4 | Horror, Sci-Fi              |
|                   2 | Drama, Musical              |
|                   2 | Drama, Fantasy, Horror      |
|                   2 | Crime, Drama, Music         |
|                   2 | Drama, Romance, Sci-Fi      |
|                   2 | Comedy, Romance             |
|                   8 | Comedy, Crime               |
|                  15 | Biography, Drama, History   |
|                   4 | Adventure, Sci-Fi           |
|                   2 | Comedy, Drama, Family       |
|                  19 | Comedy, Drama               |
|                   2 | Action, Drama, War          |
|                   2 | Action, Drama               |
|                   2 | Biography, Comedy, Drama    |
|                   6 | Adventure, Comedy, Drama    |
|                   6 | Biography, Drama, Sport     |
|                  11 | Action, Comedy, Crime       |
|                   3 | Action, Biography, Crime    |
|                   6 | Mystery, Thriller           |
|                  14 | Animation, Adventure, Comedy |
|                   2 | Action, Drama, Sci-Fi       |
|                  14 | Comedy, Crime, Drama        |
|                   9 | Biography, Drama            |
|                   2 | Action, Comedy, Romance     |
|                   4 | Action, Adventure, Comedy   |
|                   3 | Adventure, Comedy, Fantasy  |
|                   2 | Adventure, Drama, History   |
|                   6 | Action, Drama, Thriller     |
|                   2 | Comedy, Music, Romance      |
|                  27 | Drama, Romance              |
+---------------------+-----------------------------+
50 rows in set (0.00 sec)
```

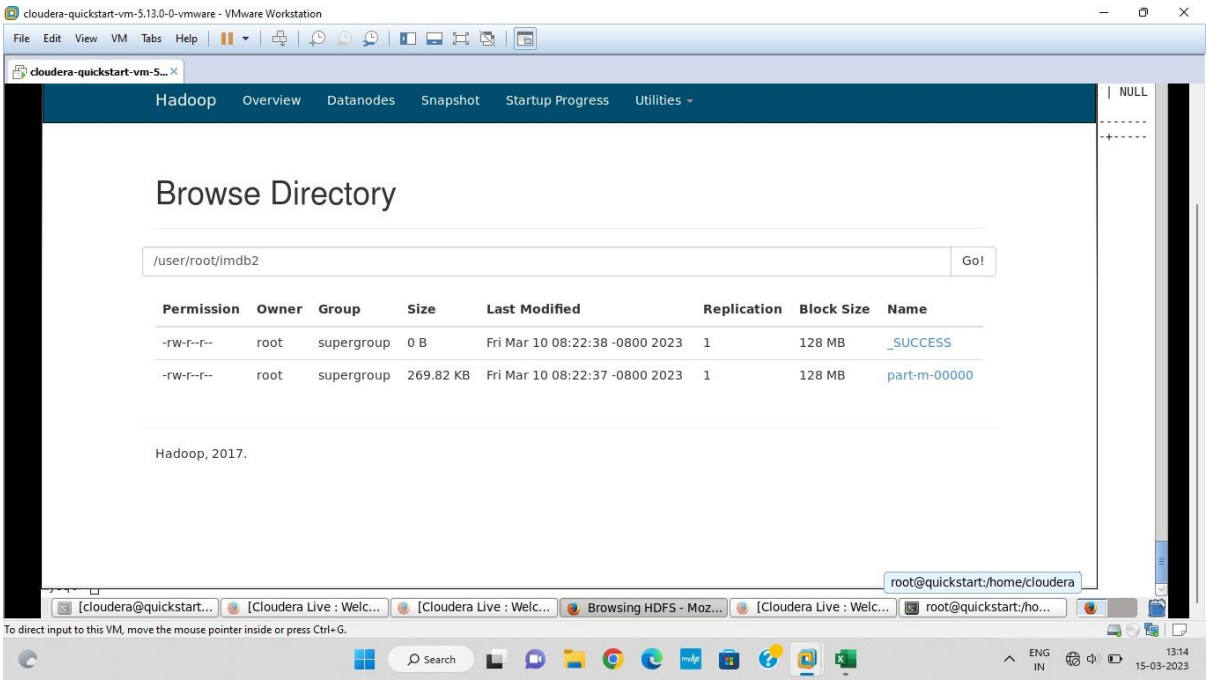
Partitioning and Bucketing:



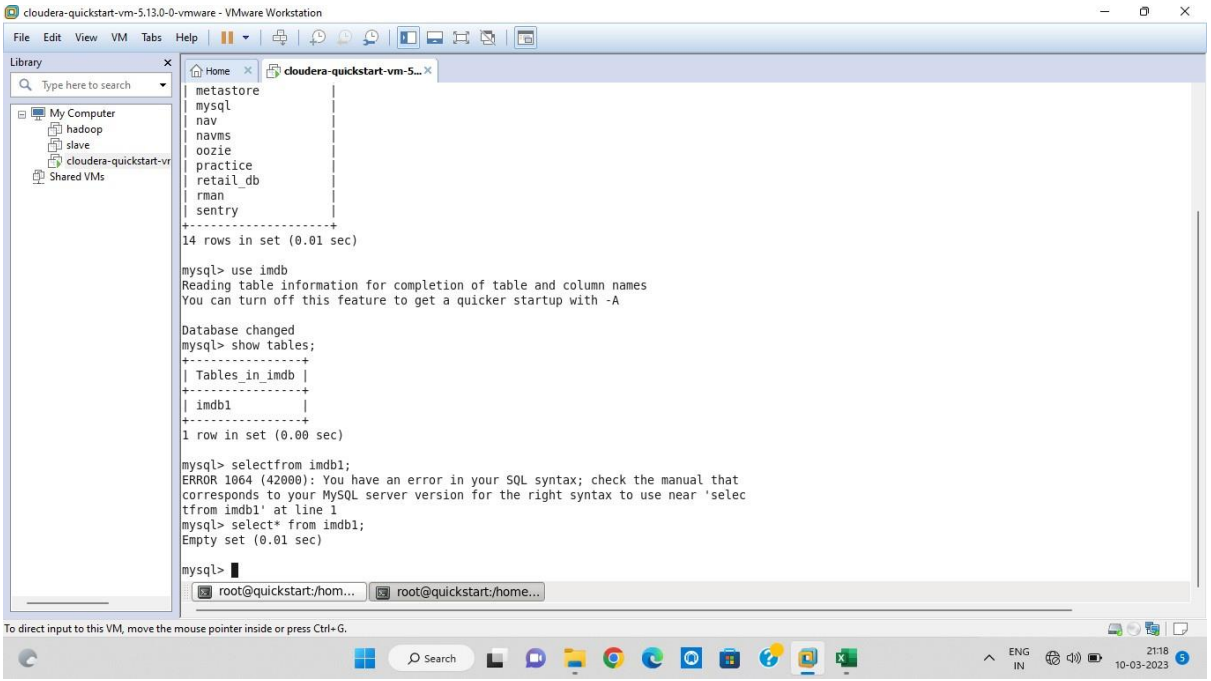


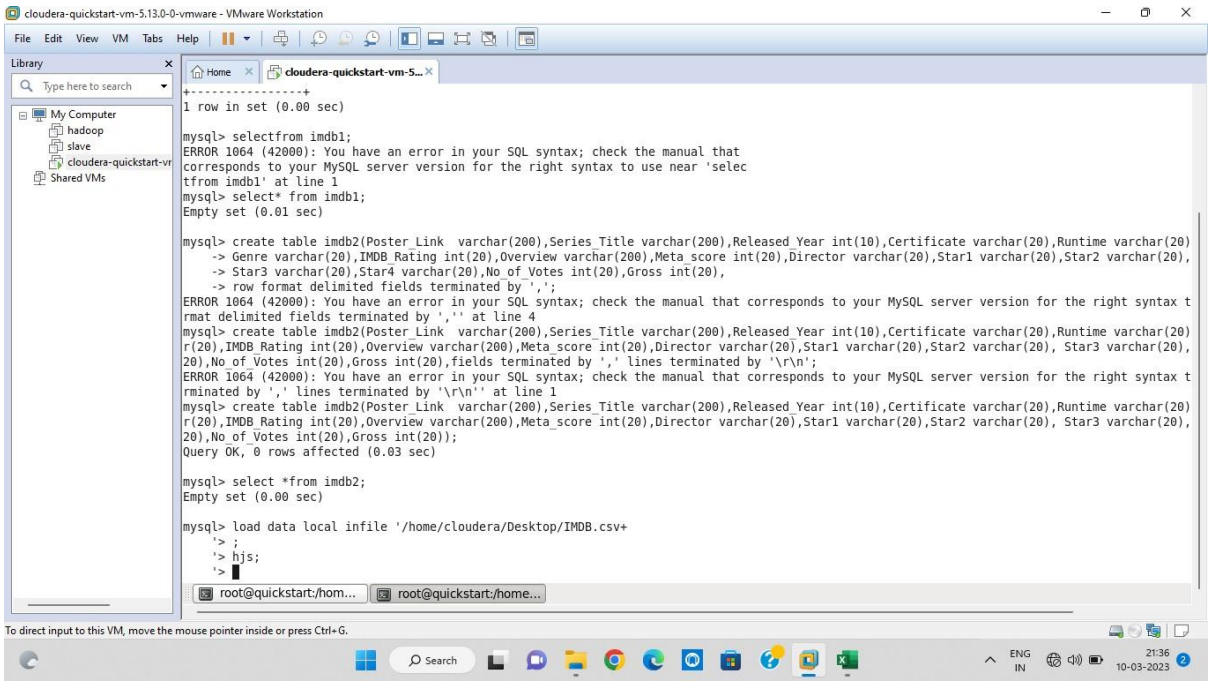

cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation
File Edit View VM Tabs Help
cloudera-quickstart-vm-5...
n, Adventure, Family",7.6,Bagheera the Panthe | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BYTE4YWU0NjAtMjNiYi00MTNiLTgwYzctZjk0YjY5NGVhNWQwXkEyXkFqcGdeQXVyMTY5Nzc4MDY@._V1_UY98_CR0,0,67,98_AL_.jpg",Blowup,
1966,A,111 min,"Drama, Mystery, Thriller",7.6 | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BZjQyMGUwNzAtNTc2MC00Y2FjLThlM2ItZGRjNzM0OWVmZGYyXkEyXkFqcGdeQXVyNjc1NTYyMjg@._V1_UX67_CR0,0,67,98_AL_.jpg",A Hard
Day's Night,1964,U,87 min,"Comedy, Music, Mus | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BNGEwMTRmZTQtMDY4Ni00MTliLTk5ZmMtOWMxYWMyMTllMDg0L2ltYWdlL2ltYWdlXkEyXkFqcGdeQXVyNjc1NTYyMjg@._V1_UX67_CR0,0,67,98_
AL_.jpg",Breakfast at Tiffany's,1961,A,115 mi | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BODk3YjdjZTItOGVhYi00Mjc2LTgzMDAtMThmYTVkNTBlMWVkXkEyXkFqcGdeQXVyNDY2MTk1ODk@._V1_UX67_CR0,0,67,98_AL_.jpg",Giant,1
956,G,201 min,"Drama, Western",7.6,Sprawling | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BM2U3YzkxNGMtYWE0YS00ODk0LTk1ZGEtNjk3ZTE0MTk4MzJjXkEyXkFqcGdeQXVyNDk0MDg4NDk@._V1_UX67_CR0,0,67,98_AL_.jpg",From He
re to Eternity,1953,Passed,118 min,"Drama, Ro | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BZTBmMjUyMjItYTM4ZS00MjAwLWEyOGYtYjMyZTUxN2I3OTMxXkEyXkFqcGdeQXVyNjc1NTYyMjg@._V1_UX67_CR0,0,67,98_AL_.jpg",Lifeboa
t,1944,,97 min,"Drama, War",7.6,Several survi | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BMTY5ODAzMTcwOF5BMl5BanBnXkFtZTcwMzYxNDYyNA@@._V1_UX67_CR0,0,67,98_AL_.jpg",The 39 Steps,1935,,86 min,"Crime, Myste
ry, Thriller",7.6,"A man in London tries to h | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
1001 rows in set (0.01 sec)
mysql> select Series_Title,IMDB_Rating from imdb2 where Released_Year=1998;
Empty set (0.00 sec)
mysql> create table imdb3(Poster_Link varchar(200),Series_Title varchar(200),Released_Year int(10),Certificate varchar(20),Runtime varchar(20), Genre varcha
r(20),IMDB_Rating int(20),Overview varchar(200),Meta_score int(20),Director varchar(20),Star1 varchar(20),Star2 varchar(20),Star3 varchar(20),Star4 varchar(
20),No_of_Votes int(20),Gross int(20)) partitioned by (Released_Year int);
root@quickstart:/home/cloudera
[cloudera@quickstart... [Cloudera Live : Welc... [Cloudera Live : Welc... [Browsing HDFS - Mo... [Cloudera Live : Welc... root@quickstart:/ho...
To direct input to this VM, move the mouse pointer inside or press Ctrl+G.
Search
ENG IN
13:13 15-03-2023

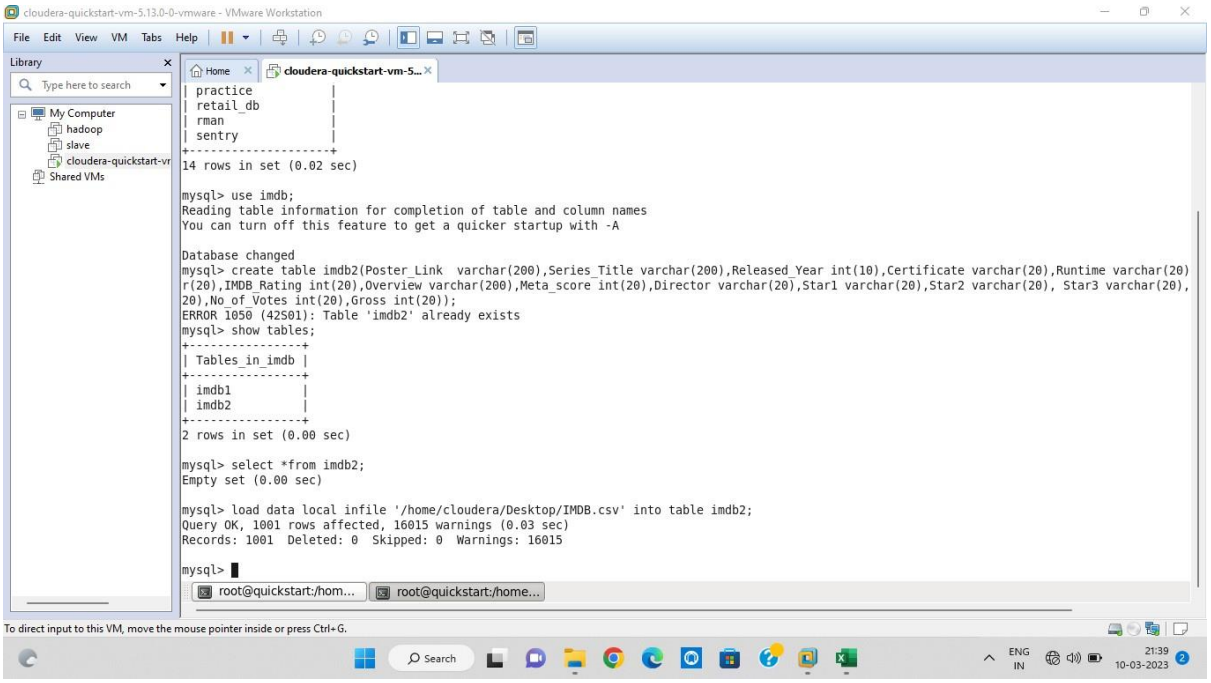

cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation
File Edit View VM Tabs Help
cloudera-quickstart-vm-5...
n, Adventure, Family",7.6,Bagheera the Panthe | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BYTE4YWU0NjAtMjNiYi00MTNiLTgwYzctZjk0YjY5NGVhNWQwXkEyXkFqcGdeQXVyMTY5Nzc4MDY@._V1_UY98_CR0,0,67,98_AL_.jpg",Blowup,
1966,A,111 min,"Drama, Mystery, Thriller",7.6 | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BZjQyMGUwNzAtNTc2MC00Y2FjLThlM2ItZGRjNzM0OWVmZGYyXkEyXkFqcGdeQXVyNjc1NTYyMjg@._V1_UX67_CR0,0,67,98_AL_.jpg",A Hard
Day's Night,1964,U,87 min,"Comedy, Music, Mus | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BNGEwMTRmZTQtMDY4Ni00MTliLTk5ZmMtOWMxYWMyMTllMDg0L2ltYWdlL2ltYWdlXkEyXkFqcGdeQXVyNjc1NTYyMjg@._V1_UX67_CR0,0,67,98_
AL_.jpg",Breakfast at Tiffany's,1961,A,115 mi | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BODk3YjdjZTItOGVhYi00Mjc2LTgzMDAtMThmYTVkNTBlMWVkXkEyXkFqcGdeQXVyNDY2MTk1ODk@._V1_UX67_CR0,0,67,98_AL_.jpg",Giant,1
956,G,201 min,"Drama, Western",7.6,Sprawling | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BM2U3YzkxNGMtYWE0YS00ODk0LTk1ZGEtNjk3ZTE0MTk4MzJjXkEyXkFqcGdeQXVyNDk0MDg4NDk@._V1_UX67_CR0,0,67,98_AL_.jpg",From He
re to Eternity,1953,Passed,118 min,"Drama, Ro | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BZTBmMjUyMjItYTM4ZS00MjAwLWEyOGYtYjMyZTUxN2I3OTMxXkEyXkFqcGdeQXVyNjc1NTYyMjg@._V1_UX67_CR0,0,67,98_AL_.jpg",Lifeboa
t,1944,,97 min,"Drama, War",7.6,Several survi | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
| "https://m.media-amazon.com/images/M/MV5BMTY5ODAzMTcwOF5BMl5BanBnXkFtZTcwMzYxNDYyNA@@._V1_UX67_CR0,0,67,98_AL_.jpg",The 39 Steps,1935,,86 min,"Crime, Myste
ry, Thriller",7.6,"A man in London tries to h | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL |
1001 rows in set (0.01 sec)
mysql> select Series_Title,IMDB_Rating from imdb2 where Released_Year=1998;
Empty set (0.00 sec)
mysql> load data local infile '/home/cloudera/Desktop/IMDB.csv' into table imdb2 partition (Released_Year="2000");
root@quickstart:/home/cloudera
[cloudera@quickstart... [Cloudera Live : Welc... [Cloudera Live : Welc... [Browsing HDFS - Mo... [Cloudera Live : Welc... root@quickstart:/ho...
To direct input to this VM, move the mouse pointer inside or press Ctrl+G.
Search
ENG IN
13:13 15-03-2023

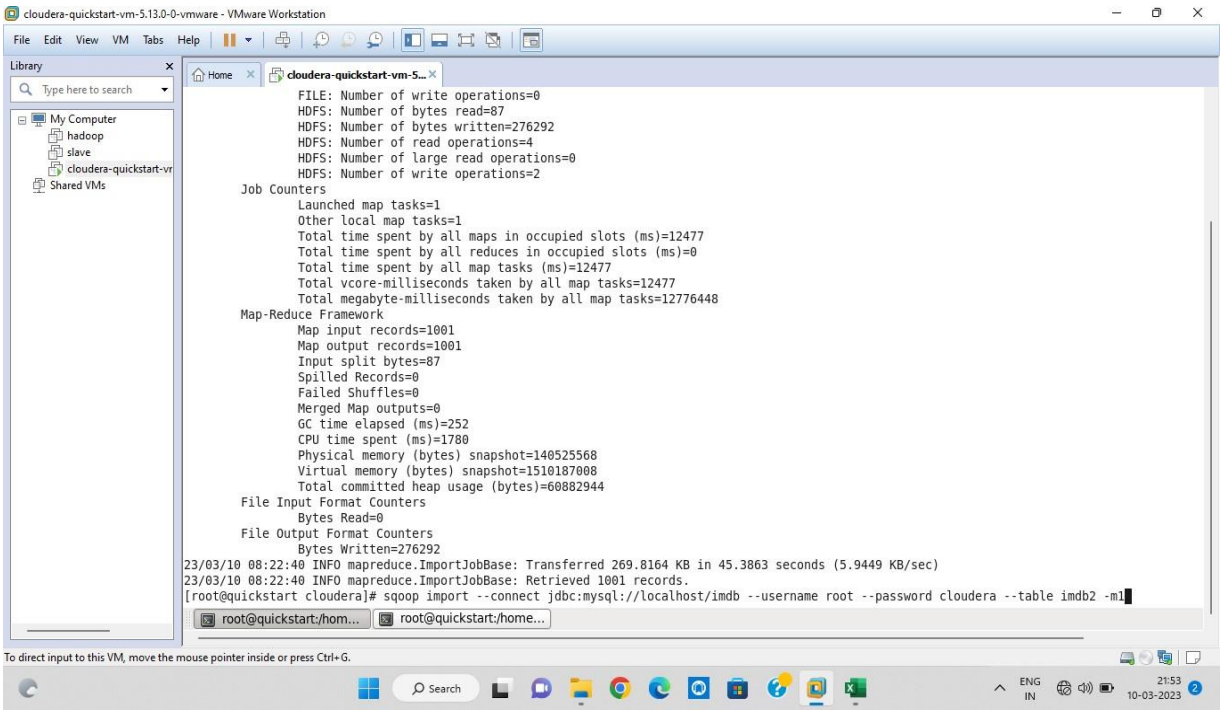

cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation
File Edit View VM Tabs Help
cloudera-quickstart-vm-5...
Hadoop Overview Datanodes Snapshot Startup Progress Utilities
Browse Directory
/user/root/imdb2
Go!
Permission Owner Group Size Last Modified Replication Block Size Name
-rw-r--r-- root supergroup 0 B Fri Mar 10 08:22:38 -0800 2023 1 128 MB _SUCCESS
-rw-r--r-- root supergroup 269.82 KB Fri Mar 10 08:22:37 -0800 2023 1 128 MB part-m-00000
Hadoop, 2017.
| NULL
root@quickstart:/home/cloudera
[cloudera@quickstart...
[Cloudera Live : Welc...
[Cloudera Live : Welc...
Browsing HDFS - Moz...
[Cloudera Live : Welc...
root@quickstart:/ho...
To direct input to this VM, move the mouse pointer inside or press Ctrl+G.
Search
ENG
IN
13:14
15-03-2023

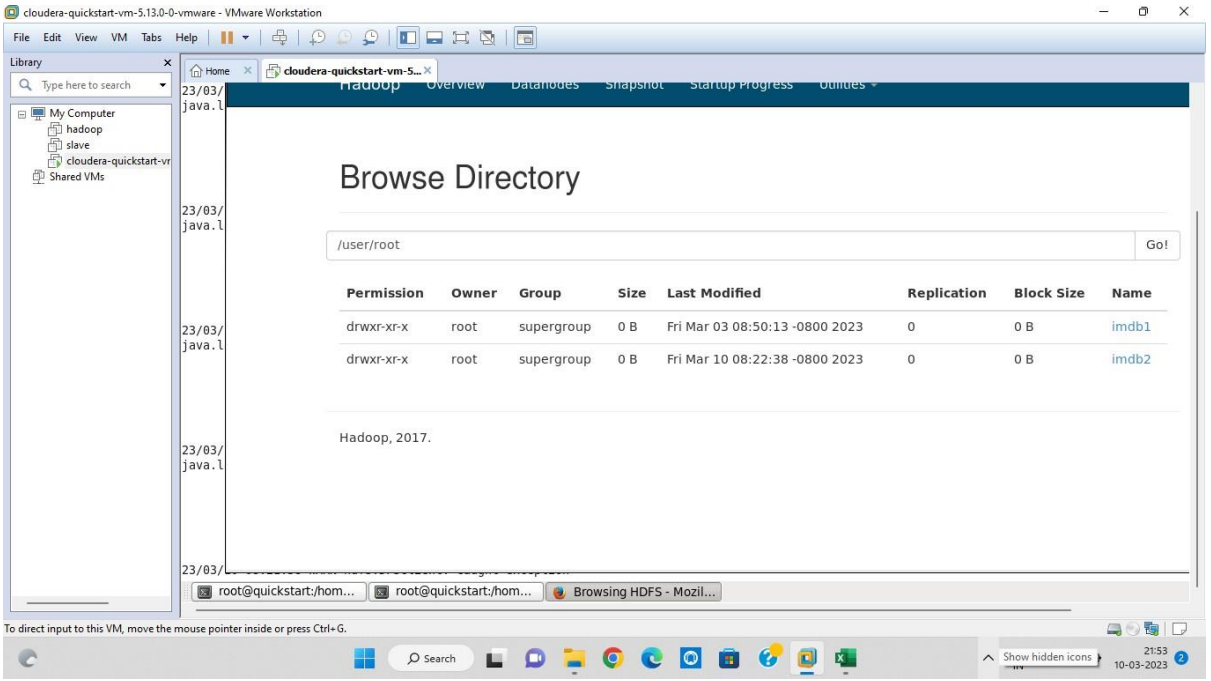
Sqoop Import and Export..

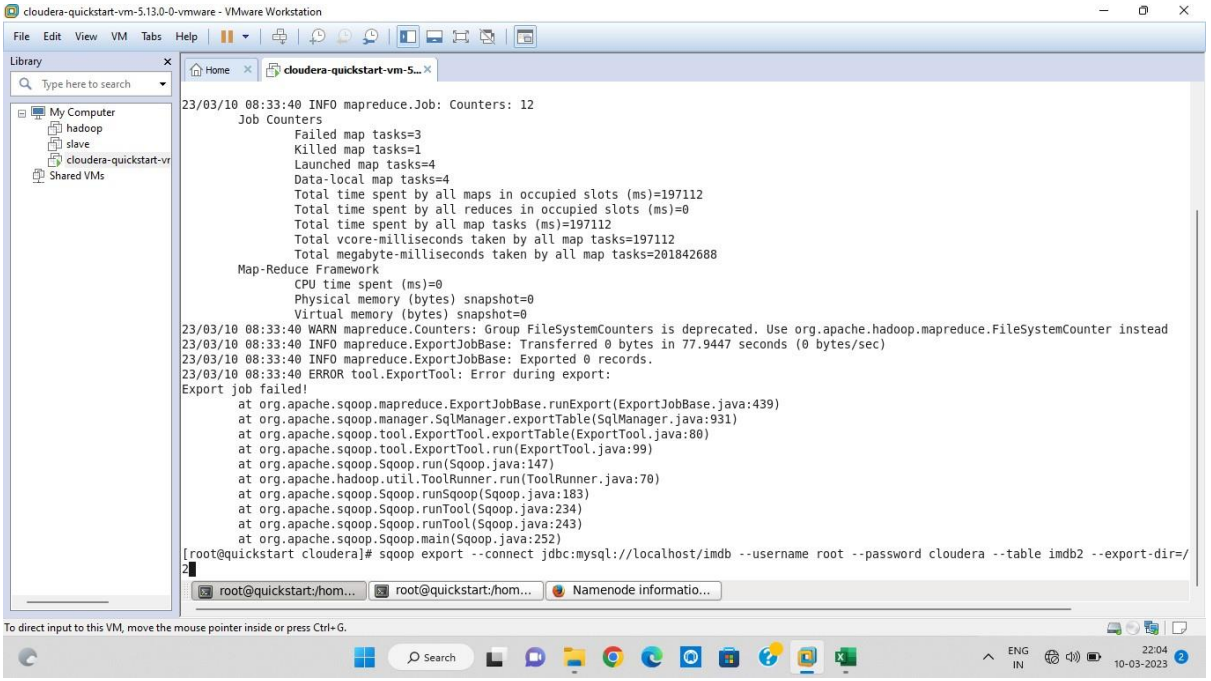



cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation
File Edit View VM Tabs Help
Library
Type here to search
My Computer
hadoop
slave
cloudera-quickstart-vr
Shared VMs
Home
cloudera-quickstart-vm-5...
+----------------+
1 row in set (0.00 sec)
mysql> selectfrom imdb1;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that
corresponds to your MySQL server version for the right syntax to use near 'selec
tfrom imdb1' at line 1
mysql> select* from imdb1;
Empty set (0.01 sec)
mysql> create table imdb2(Poster_Link varchar(200),Series_Title varchar(200),Released_Year int(10),Certificate varchar(20),Runtime varchar(20)
-> Genre varchar(20),IMDB_Rating int(20),Overview varchar(200),Meta_score int(20),Director varchar(20),Star1 varchar(20),Star2 varchar(20),
-> Star3 varchar(20),Star4 varchar(20),No_of_Votes int(20),Gross int(20),
-> row format delimited fields terminated by ',';
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax t
rmat delimited fields terminated by ',' at line 4
mysql> create table imdb2(Poster_Link varchar(200),Series_Title varchar(200),Released_Year int(10),Certificate varchar(20),Runtime varchar(20)
r(20),IMDB_Rating int(20),Overview varchar(200),Meta_score int(20),Director varchar(20),Star1 varchar(20),Star2 varchar(20), Star3 varchar(20),
20),No_of_Votes int(20),Gross int(20),fields terminated by ',' lines terminated by '\r\n';
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax t
rminated by ',' lines terminated by '\r\n'' at line 1
mysql> create table imdb2(Poster_Link varchar(200),Series_Title varchar(200),Released_Year int(10),Certificate varchar(20),Runtime varchar(20)
r(20),IMDB_Rating int(20),Overview varchar(200),Meta_score int(20),Director varchar(20),Star1 varchar(20),Star2 varchar(20), Star3 varchar(20),
20),No_of_Votes int(20),Gross int(20));
Query OK, 0 rows affected (0.03 sec)
mysql> select *from imdb2;
Empty set (0.00 sec)
mysql> load data local infile '/home/cloudera/Desktop/IMDB.csv+
'> ;
'> hjs;
'>
root@quickstart:/hom...
root@quickstart:/home...
To direct input to this VM, move the mouse pointer inside or press Ctrl+G.
Search
ENG
IN
21:36
10-03-2023


cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation
File Edit View VM Tabs Help
Library
Type here to search
My Computer
hadoop
slave
cloudera-quickstart-vr
Shared VMs
Home
cloudera-quickstart-vm-5...
| practice |
| retail_db |
| rman |
| sentry |
+--------------------+
14 rows in set (0.02 sec)
mysql> use imdb;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A
Database changed
mysql> create table imdb2(Poster_Link varchar(200),Series_Title varchar(200),Released_Year int(10),Certificate varchar(20),Runtime varchar(20)
r(20),IMDB_Rating int(20),Overview varchar(200),Meta_score int(20),Director varchar(20),Star1 varchar(20),Star2 varchar(20), Star3 varchar(20),
20),No_of_Votes int(20),Gross int(20));
ERROR 1050 (42S01): Table 'imdb2' already exists
mysql> show tables;
+----------------+
| Tables_in_imdb |
+----------------+
| imdb1 |
| imdb2 |
+----------------+
2 rows in set (0.00 sec)
mysql> select *from imdb2;
Empty set (0.00 sec)
mysql> load data local infile '/home/cloudera/Desktop/IMDB.csv' into table imdb2;
Query OK, 1001 rows affected, 16015 warnings (0.03 sec)
Records: 1001 Deleted: 0 Skipped: 0 Warnings: 16015
mysql>
root@quickstart:/hom...
root@quickstart:/home...
To direct input to this VM, move the mouse pointer inside or press Ctrl+G.
Search
ENG
IN
21:39
10-03-2023


cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation
File Edit View VM Tabs Help
Library
Type here to search
My Computer
hadoop
slave
cloudera-quickstart-vr
Shared VMs
Home
cloudera-quickstart-vm-5...
FILE: Number of write operations=0
HDFS: Number of bytes read=87
HDFS: Number of bytes written=276292
HDFS: Number of read operations=4
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
Launched map tasks=1
Other local map tasks=1
Total time spent by all maps in occupied slots (ms)=12477
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=12477
Total vcore-milliseconds taken by all map tasks=12477
Total megabyte-milliseconds taken by all map tasks=12776448
Map-Reduce Framework
Map input records=1001
Map output records=1001
Input split bytes=87
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=252
CPU time spent (ms)=1780
Physical memory (bytes) snapshot=140525568
Virtual memory (bytes) snapshot=1510187008
Total committed heap usage (bytes)=60882944
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=276292
23/03/10 08:22:40 INFO mapreduce.ImportJobBase: Transferred 269.8164 KB in 45.3863 seconds (5.9449 KB/sec)
23/03/10 08:22:40 INFO mapreduce.ImportJobBase: Retrieved 1001 records.
[root@quickstart cloudera]# sqoop import --connect jdbc:mysql://localhost/imdb --username root --password cloudera --table imdb2 -m1
root@quickstart:/hom...
root@quickstart:/home...
To direct input to this VM, move the mouse pointer inside or press Ctrl+G.
Search
ENG
IN
21:53
10-03-2023


cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation
File Edit View VM Tabs Help
Library
Type here to search
My Computer
hadoop
slave
cloudera-quickstart-vr
Shared VMs
Home
cloudera-quickstart-vm-5...
Hadoop
Overview
Datanodes
Snapshot
Startup Progress
Utilities
Browse Directory
/user/root
Go!
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
drwxr-xr-x
root
supergroup
0 B
Fri Mar 03 08:50:13 -0800 2023
0
0 B
imdb1
drwxr-xr-x
root
supergroup
0 B
Fri Mar 10 08:22:38 -0800 2023
0
0 B
imdb2
Hadoop, 2017.
root@quickstart:/hom...
root@quickstart:/hom...
Browsing HDFS - Mozil...
To direct input to this VM, move the mouse pointer inside or press Ctrl+G.
Search
Show hidden icons
21:53
10-03-2023



Observation:-

During the analysis of we come to IMDB Dataset know the following points-

>The IMDb dataset is a large collection of movie and TV show data that includes information such as title, genre, release year, cast and crew, plot summary, user ratings, and more.

>Analysis of popular genres and trends in the film industry over time.

>Comparison of ratings between different genres, directors, or actors.

>Examination of the relationship between budget and box office revenue for different films

Conclusion:-

- **The most popular genres for movies are action, drama, and comedy, while the most popular genres for TV shows are drama, comedy, and crime.**

- **There is a positive correlation between a film's budget and its box office revenue, but the relationship is not always straightforward and there are many factors that can influence a film's success.**
- **Certain directors and actors have a higher average rating for their movies compared to others, suggesting that they have a strong influence on the quality of the films they are involved in.**

Acknowledgement:-

- **The IMDb dataset was sourced from IMDb (Internet Movie Database), which is owned by Amazon.**
- **The data was obtained through the IMDb API or from one of the IMDb datasets available on the IMDb website.**
- **Any tools or libraries used for data cleaning, visualization, or analysis should be acknowledged, along with their respective authors or organizations.**

Reference:-

- **Indeed Inspiring Infotech study material**