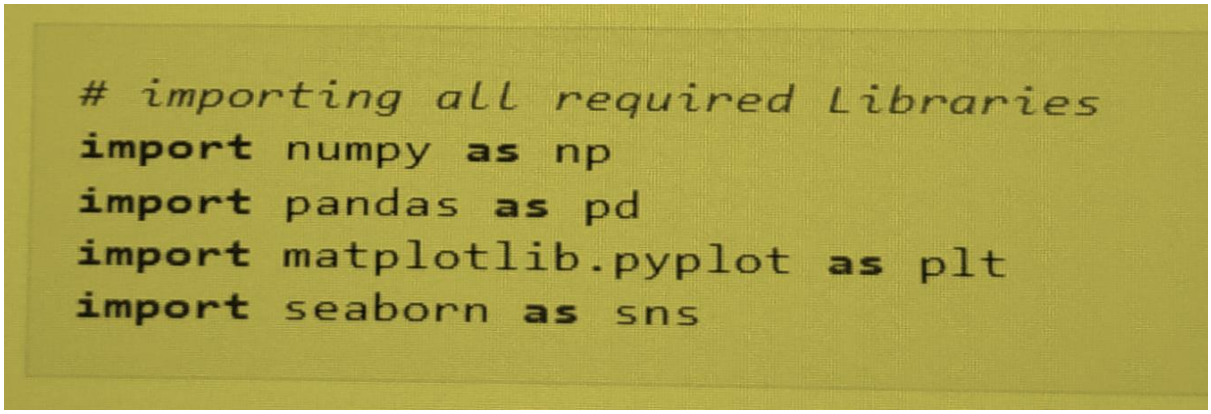
Exploratory Data Analysis of 1000 Thriller Films IMDB Rating

Step – 1: Import all required Libraries

The First Step to Perform Exploratory Data Analysis on Dataset is Importing all required Libraries such as NumPy, Pandas, Matplotlib, and Seaborn.



```
# importing all required Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Step – 2: Read the Dataset

Second Step is reading the Dataset.We are going to read the data from Excel file into a Pandas DataFrame by using read_excel () function.

Generally, The Pandas Library Provides a wide range of Probabilities for loading data into the Pandas DataFrame from files such as.xlsx, .csv, .sql, .JSON, .html, etc.

```
In [2]: # reading the Data fromexcel file
        df = pd.read_excel(r"C:\Users\mouni\Desktop\All_web_scraping_projects\1000_thrillerfilms_list.xlsx")

In [3]: df

Out[3]:
```

	S.No	Film Name	Year of Release	IMDB Rating	Story	Director Name	Film Length
0	1	The Dark Knight	2008	9.0	When the menace known as the Joker, wreaks havo...	Christopher Nolan	152
1	2	Aynabaji	2016	9.0	Ayna is an actor and the prison is his stage. ...	Amitabh Reza Chowdhury	147
2	3	Inception	2010	8.8	A thief who steals corporate secrets through t...	Christopher Nolan	148
3	4	Se7en	1995	8.6	Two detectives, a rookie and a veteran, hunt a...	David Fincher	127
4	5	The Silence of the Lambs	1991	8.6	A young F.B.I. cadet must receive the help of ...	Jonathan Demme	118
...	...	...	...	...	...	...	...
995	996	Event Horizon	1997	6.6	A rescue crew investigates a spaceship that di...	Paul W.S. Anderson	96
996	997	The Wonder	2022	6.6	A tale of two strangers who transform each oth...	Sebastián Lelio	108
997	998	The Bourne Legacy	2012	6.6	An expansion of the universe from Robert Ludlu...	Tony Gilroy	135
998	999	Jason Bourne	2016	6.6	The CIA's most dangerous former operative is d...	Paul Greengrass	123
999	1,000	Red Sparrow	2018	6.6	Ballerina Dominika Egorova is recruited to 'Sp...	Francis Lawrence	140

1000 rows × 7 columns

Step – 3: Data Cleaning

Now, we will Check for Null Values Count in Each Column of a DataFrame with the help of isnull (). Sum () method.

```
# number of null values in each column
df.isnull().sum()
```

```
S.No                0
Film Name           0
Year of Release     0
IMDB Rating         0
Story               0
Director Name       0
Film Length         0
dtype: int64
```

After checking the Null Values, we can say that there are no null values in each column showing zero count of null values.

We can also check the number of non-Null values in each column of a DataFrame by using notnull (). Sum () function.

```
# number of not null values in each column
df.notnull().sum()
```

```
S.No                1000
Film Name           1000
Year of Release     1000
IMDB Rating         1000
Story               1000
Director Name       1000
Film Length         1000
dtype: int64
```

Each column contains 1000 records.

Now, we need to Identify the number of Duplicate Records in a DataFrame. So that, we will use duplicated (). Sum ().

```
# Total number of duplicate Values in a DF
df.duplicated().sum()
```

```
0
```

There are no Duplicate Records in a DataFrame. It shows zero count of Duplicates.

Now, we are going to set an Index for the DataFrame. We consider Serial Number as an Index.

We can arange the Order of the Columns in a DataFrame.

```
# set Serial Number as an index for the DataFrame
df.set_index('S.No', inplace = True)
```

```
# Aranging the order of Columns
df = df[['Film Name', 'Director Name', 'Story', 'Film Length', 'Year of Release', 'IMDB Rating']]
df
```

S.No	Film Name	Director Name	Story	Film Length	Year of Release	IMDB Rating
1	The Dark Knight	Christopher Nolan	When the menace known as the Joker wreaks havo...	152	2008	9.0
2	Aynabaji	Amitabh Reza Chowdhury	Ayna is an actor and the prison is his stage. ...	147	2016	9.0
3	Inception	Christopher Nolan	A thief who steals corporate secrets through t...	148	2010	8.8
4	Se7en	David Fincher	Two detectives, a rookie and a veteran, hunt a...	127	1995	8.6
5	The Silence of the Lambs	Jonathan Demme	A young F.B.I. cadet must receive the help of ...	118	1991	8.6
...	...	...	...	...	...	...
996	Event Horizon	Paul W.S. Anderson	A rescue crew investigates a spaceship that di...	96	1997	6.6
997	The Wonder	Sebastián Lelio	A tale of two strangers who transform each oth...	108	2022	6.6
998	The Bourne Legacy	Tony Gilroy	An expansion of the universe from Robert Ludlu...	135	2012	6.6
999	Jason Bourne	Paul Greengrass	The CIA's most dangerous former operative is d...	123	2016	6.6
1,000	Red Sparrow	Francis Lawrence	Ballerina Dominika Egorova is recruited to 'Sp...	140	2018	6.6

1000 rows × 6 columns

We can rename column names also as we like. So that, we will rename the Year of Release column to Year of Released Column for better understanding.

```
# Renaming the Year of Release column into Year of Released column
df.rename(columns = {'Year of Release':'Year of Released'}, inplace = True)
df
```

S.No	Film Name	Director Name	Story	Film Length	Year of Released	IMDB Rating
1	The Dark Knight	Christopher Nolan	When the menace known as the Joker wreaks havo...	152	2008	9.0
2	Aynabaji	Amitabh Reza Chowdhury	Ayna is an actor and the prison is his stage. ...	147	2016	9.0
3	Inception	Christopher Nolan	A thief who steals corporate secrets through t...	148	2010	8.8
4	Se7en	David Fincher	Two detectives, a rookie and a veteran, hunt a...	127	1995	8.6
5	The Silence of the Lambs	Jonathan Demme	A young F.B.I. cadet must receive the help of ...	118	1991	8.6
...	...	...	...	...	...	...
996	Event Horizon	Paul W.S. Anderson	A rescue crew investigates a spaceship that di...	96	1997	6.6
997	The Wonder	Sebastián Lelio	A tale of two strangers who transform each oth...	108	2022	6.6
998	The Bourne Legacy	Tony Gilroy	An expansion of the universe from Robert Ludlu...	135	2012	6.6
999	Jason Bourne	Paul Greengrass	The CIA's most dangerous former operative is d...	123	2016	6.6
1,000	Red Sparrow	Francis Lawrence	Ballerina Dominika Egorova is recruited to 'Sp...	140	2018	6.6

1000 rows × 6 columns

Now, Year of Release column name changed to Year of Released.

Step – 4: Data Exploration

After completion of Data Cleaning, we will check the number of Rows and Columns in a DataFrame. For checking the number of rows and columns, we will use shape attribute.

```
# total number of Rows and Columns
df.shape
```

```
(1000, 6)
```

The DataFrame Contains 1000 records and 6 Columns.

Now, we can see all the Column names of a DataFrame by using columns attribute.

```
# Shows all the Column names in a DF
df.columns
```

```
Index(['Film Name', 'Director Name', 'Story', 'Film Length',
       'Year of Released', 'IMDB Rating'],
      dtype='object')
```

If we want to know the number of Columns, we need to perform len () on all column names like len (df.columns).

```
# number of Columns in a DF
len(df.columns)
```

```
6
```

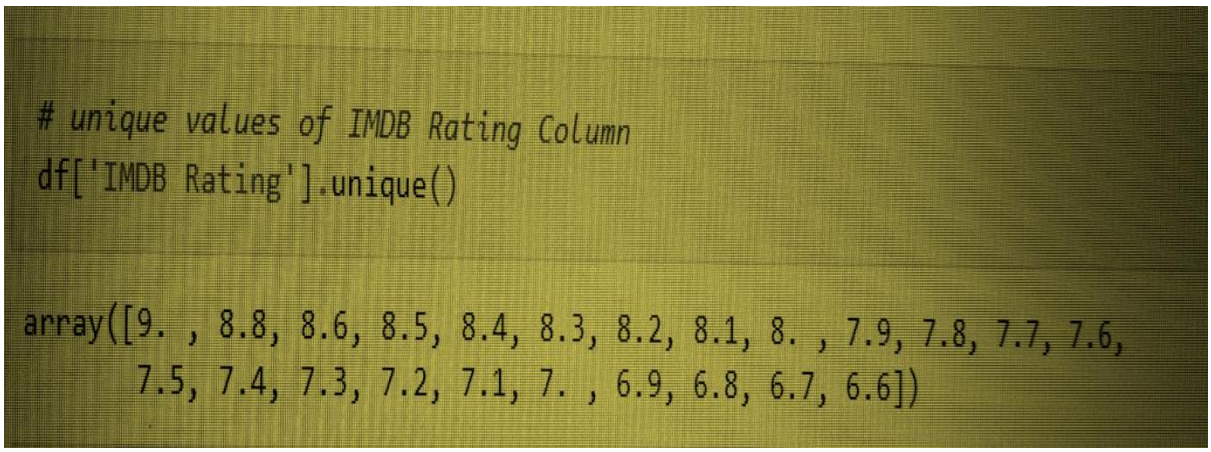
It shows that 6 columns are Present in a DataFrame.

If we want to know the data types of each column, we will use dtypes attribute.

```
# each column's Data type in a DF
df.dtypes
```

```
Film Name              object
Director Name          object
Story                  object
Film Length             int64
Year of Released        int64
IMDB Rating           float64
dtype: object
```

We can see the unique values in the IMDB Rating column of a DataFrame. For that, we will use unique () method.

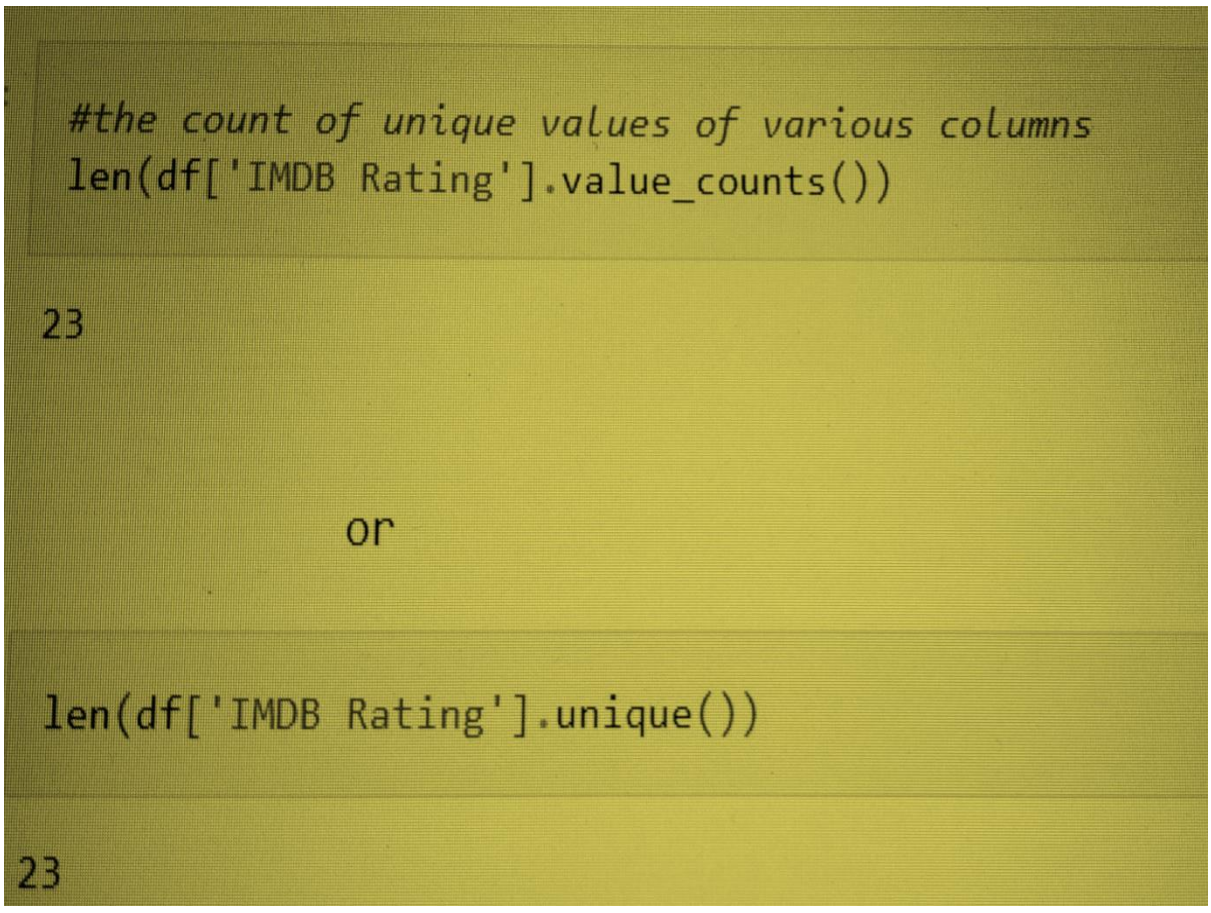


```
# unique values of IMDB Rating Column
df['IMDB Rating'].unique()

array([9. , 8.8, 8.6, 8.5, 8.4, 8.3, 8.2, 8.1, 8. , 7.9, 7.8, 7.7, 7.6,
       7.5, 7.4, 7.3, 7.2, 7.1, 7. , 6.9, 6.8, 6.7, 6.6])
```

These are the Ratings given to each thriller film.

We can count the number of Unique values in an IMDB Rating column.



```
#the count of unique values of various columns
len(df['IMDB Rating'].value_counts())

23

or

len(df['IMDB Rating'].unique())

23
```

23 unique IMDB Ratings are given to 1000 thriller films.

We will use Value_counts () function to return a Series that contain counts of Unique values. We are using this function for IMDB Rating Column in a DataFrame.

```
# returns a Series that contain counts of unique values
df['IMDB Rating'].value_counts()
```

```
6.7    104
6.8     89
7.0     85
7.1     80
7.3     79
7.2     71
7.5     64
6.9     62
7.4     56
7.6     55
7.7     50
7.8     41
8.0     34
7.9     32
8.1     30
8.2     25
6.6     10
8.3     10
8.4     10
8.5      8
8.6      2
9.0      2
8.8      1
Name: IMDB Rating, dtype: int64
```

More thriller films got 6.7 IMDB Rating that to 104 films got 6.7 IMDB Rating out of 1000 thriller films.

We can use describe () method for getting statistical information such as mean, median, mode, Percentiles, minimum values, and maximum values of all the numerical Variables have been populated.

```
# it returns the Description of the Data in a DF
df.describe()
```

	Film Length	Year of Released	IMDB Rating
count	1000.000000	1000.000000	1000.000000
mean	117.780000	2000.669000	7.315700
std	20.896992	18.512486	0.474372
min	69.000000	1920.000000	6.600000
25%	103.000000	1993.000000	6.900000
50%	115.000000	2006.000000	7.200000
75%	129.000000	2014.000000	7.600000
max	321.000000	2023.000000	9.000000

If we want to know all the Information about the DataFrame such as number of columns, number of Null Values and Non-Null Values in each column, data type of each column, and Memory usage, we can use info () function.

```
# gives all the information of the DataFrame
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1000 entries, 1 to 1,000
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Film Name           1000 non-null   object
 1   Director Name       1000 non-null   object
 2   Story               1000 non-null   object
 3   Film Length         1000 non-null   int64
 4   Year of Released    1000 non-null   int64
 5   IMDB Rating         1000 non-null   float64
dtypes: float64(1), int64(2), object(3)
memory usage: 54.7+ KB
```

We can use select_dtypes () function to get only particular data type columns in a DataFrame. If you want to get all the Numerical Columns in a DataFrame, we will use include parameter in select_dtypes () function like select_dtypes (include = 'Number'). If you want to get all the Categorical Columns in a DataFrame, we will use include parameter in select_dtypes () function like select_dtypes (include = 'object').

```
# shows only all the numeric columns in a DF
df.select_dtypes(include='number')
```

S.No	Film Length	Year of Released	IMDB Rating
1	152	2008	9.0
2	147	2016	9.0
3	148	2010	8.8
4	127	1995	8.6
5	118	1991	8.6
...	...	...	...
996	96	1997	6.6
997	108	2022	6.6
998	135	2012	6.6
999	123	2016	6.6
1,000	140	2018	6.6

1000 rows × 3 columns

```
# gives all types of data type except numeric datatype in a DF
df.select_dtypes(exclude='number')
```

S.No	Film Name	Director Name	Story
1	The Dark Knight	Christopher Nolan	When the menace known as the Joker wreaks havo...
2	Aynabaji	Amitabh Reza Chowdhury	Ayna is an actor and the prison is his stage. ...
3	Inception	Christopher Nolan	A thief who steals corporate secrets through t...
4	Se7en	David Fincher	Two detectives, a rookie and a veteran, hunt a...
5	The Silence of the Lambs	Jonathan Demme	A young F.B.I. cadet must receive the help of ...
...	...	...	...
996	Event Horizon	Paul W.S. Anderson	A rescue crew investigates a spaceship that di...
997	The Wonder	Sebastián Lelio	A tale of two strangers who transform each oth...
998	The Bourne Legacy	Tony Gilroy	An expansion of the universe from Robert Ludlu...
999	Jason Bourne	Paul Greengrass	The CIA's most dangerous former operative is d...
1,000	Red Sparrow	Francis Lawrence	Ballerina Dominika Egorova is recruited to 'Sp...

1000 rows × 3 columns

These are the Categorical Columns in a DataFrame.

If we want to get Top Five Records from a DataFrame, we will use head () function.

```
# Top 5 Records in a DF
df.head()
```

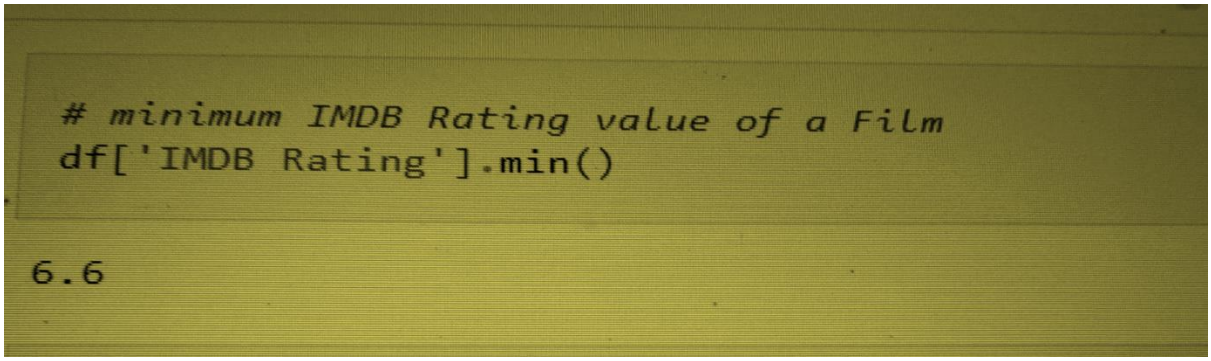
S.No	Film Name	Director Name	Story	Film Length	Year of Released	IMDB Rating
1	The Dark Knight	Christopher Nolan	When the menace known as the Joker wreaks havo...	152	2008	9.0
2	Aynabaji	Amitabh Reza Chowdhury	Ayna is an actor and the prison is his stage. ...	147	2016	9.0
3	Inception	Christopher Nolan	A thief who steals corporate secrets through t...	148	2010	8.8
4	Se7en	David Fincher	Two detectives, a rookie and a veteran, hunt a...	127	1995	8.6
5	The Silence of the Lambs	Jonathan Demme	A young F.B.I. cadet must receive the help of ...	118	1991	8.6

We can use tail () function to get Bottom Five Records from a DataFrame.

```
# Bottom 5 Records in a DF
df.tail()
```

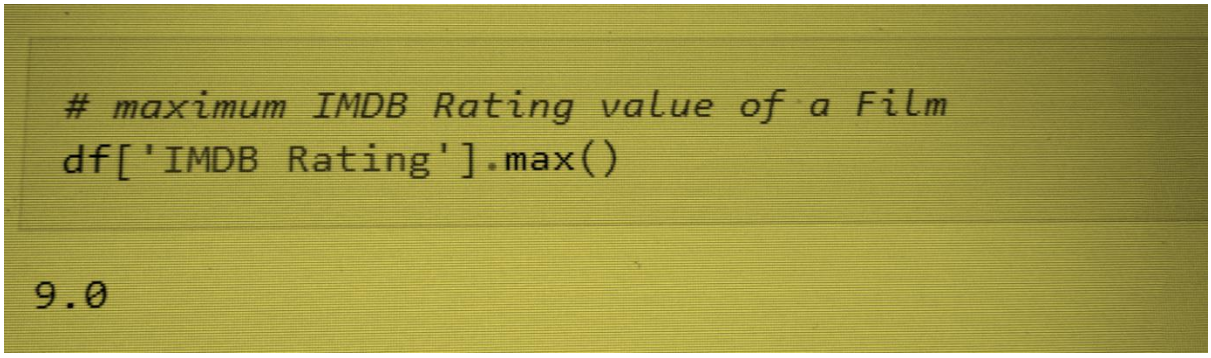
S.No	Film Name	Director Name	Story	Film Length	Year of Released	IMDB Rating
996	Event Horizon	Paul W.S. Anderson	A rescue crew investigates a spaceship that di...	96	1997	6.6
997	The Wonder	Sebastián Lelio	A tale of two strangers who transform each oth...	108	2022	6.6
998	The Bourne Legacy	Tony Gilroy	An expansion of the universe from Robert Ludlu...	135	2012	6.6
999	Jason Bourne	Paul Greengrass	The CIA's most dangerous former operative is d...	123	2016	6.6
1,000	Red Sparrow	Francis Lawrence	Ballerina Dominika Egorova is recruited to 'Sp...	140	2018	6.6

We can get Minimum Value of IMDB Rating with the help of min () function.



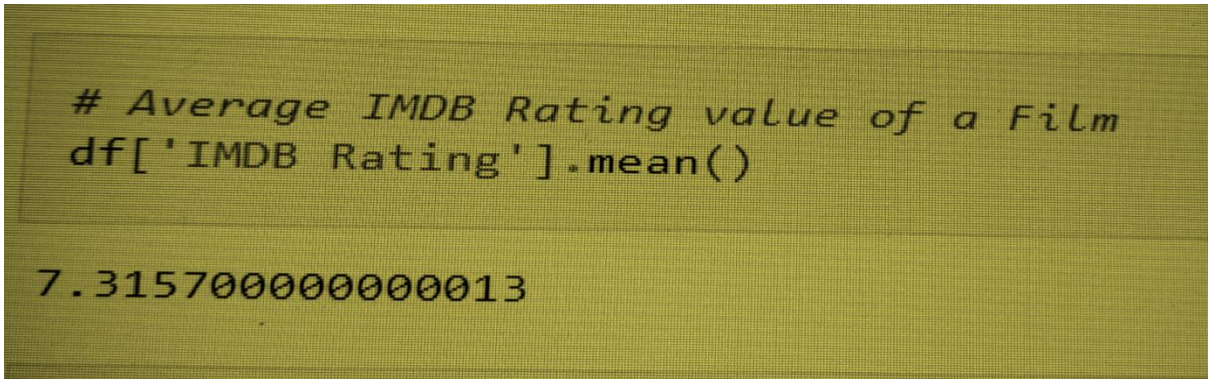
Minimum IMDB Rating given to the thriller films is 6.6.

If we want to know the Maximum Value of IMDB Rating, we will use max () function.



Highest IMDB Rating given to the thriller films is 6.6.

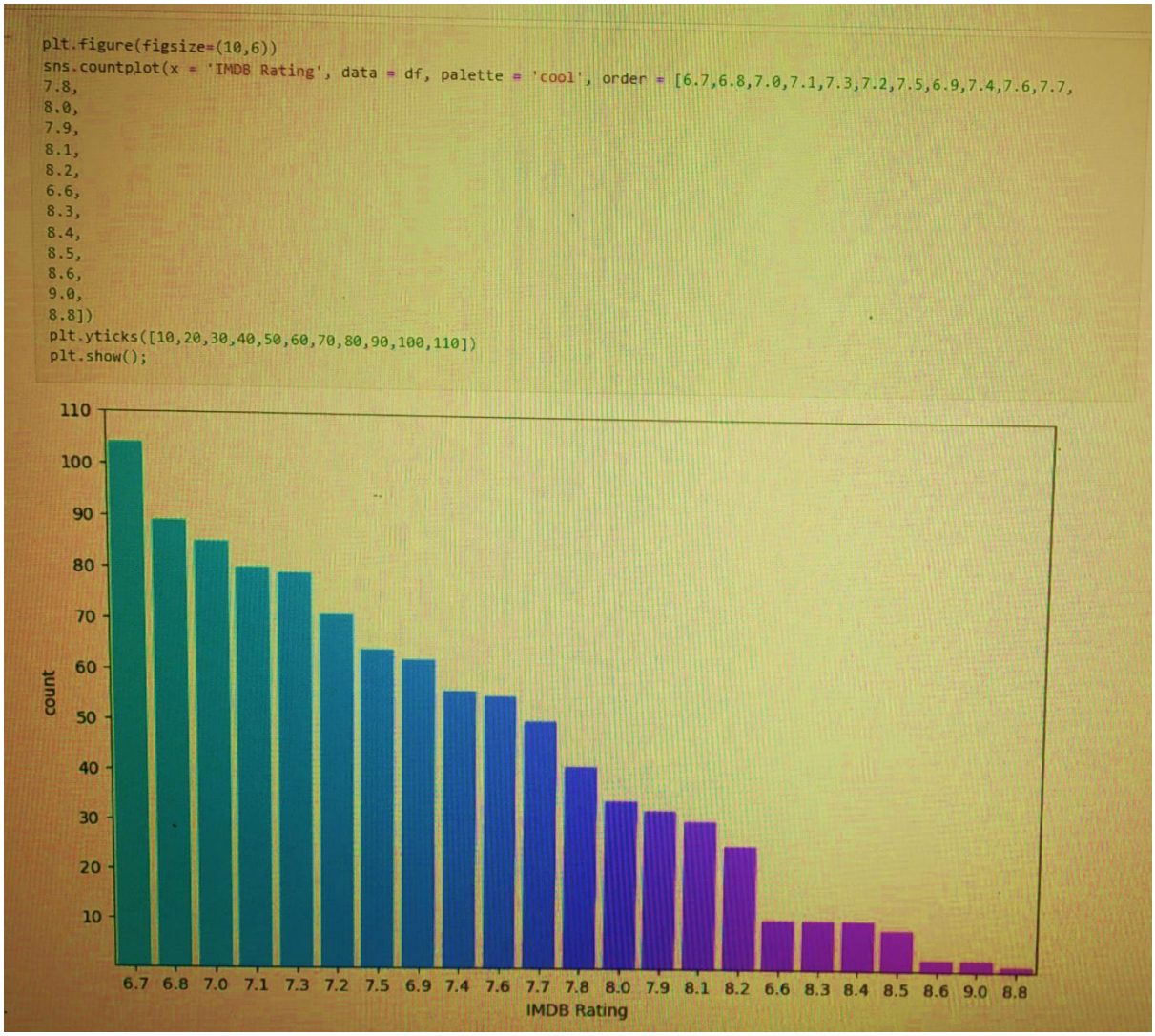
If we want to know the Average IMDB Rating of Thriller Films, we can use mean () function.



Average IMDB Rating of thriller films is 7.32.

Step – 5: Data Visualization using Matplotlib and Seaborn Libraries

If we want to Count the Number of Thriller Films get the Highest (or) Lowest IMDB Rating. We should take IMDB Rating Variable in X – Axis.

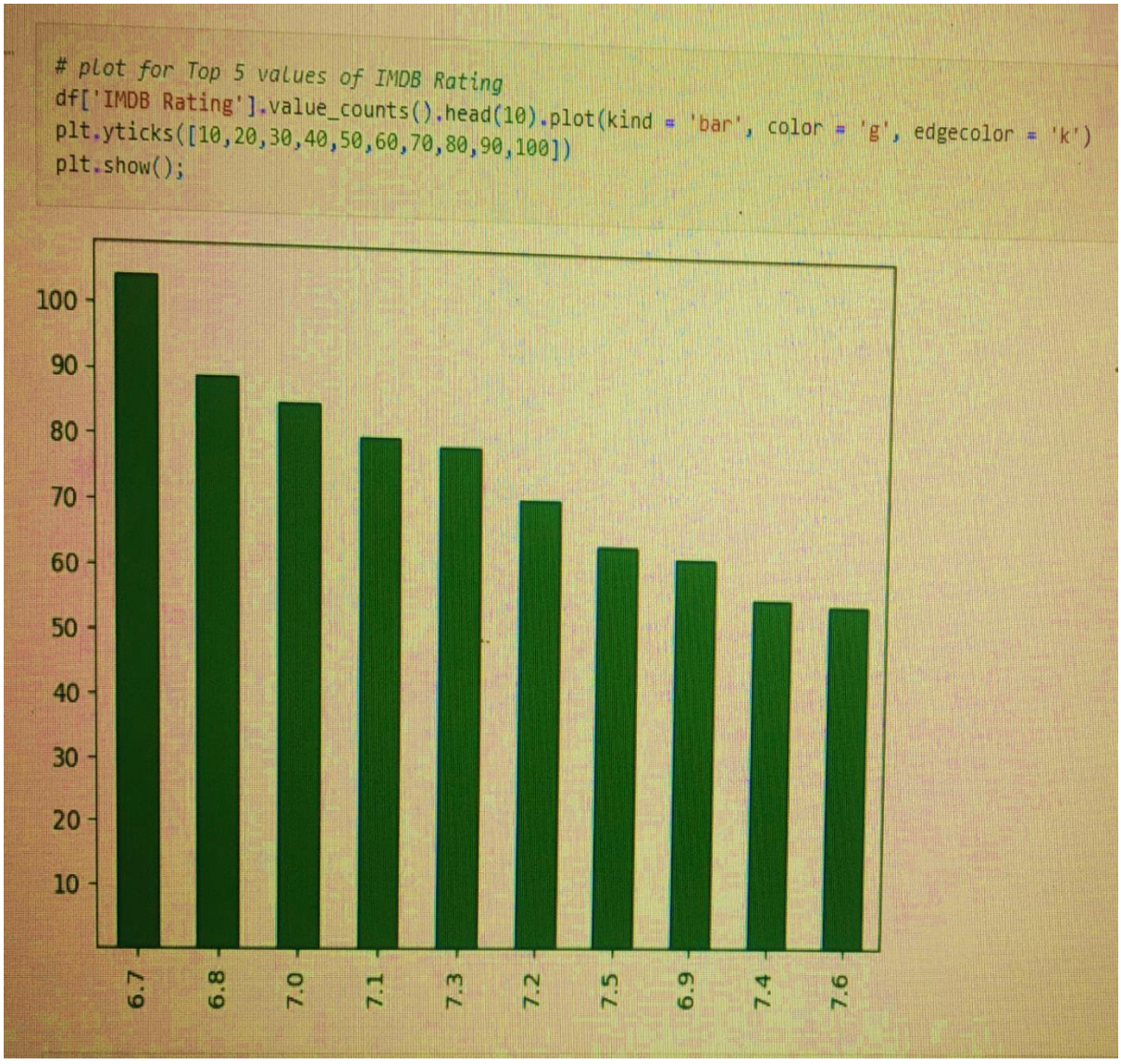


After Ploting the countplot, we can get some Insights.

Insights:

1. Approximately 105 Triller films got 6.7 IMDB Rating.
2. Less number of Thriller Films got IMDB Rating as 8.8.
3. Approximately equal number of films got 8.6 and 9.0 IMDB Ratings.

If we want to know Top 10 IMDB Ratings of Thriller Films, plot a Bar Plot.

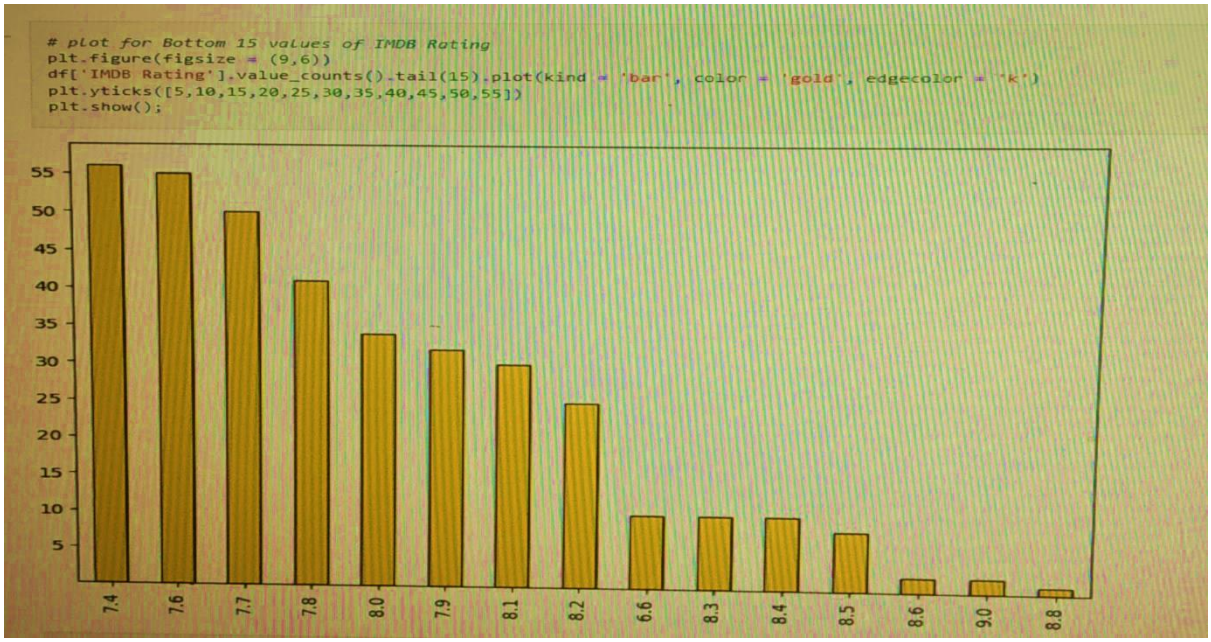


After Observing the Graph, we get some Information.

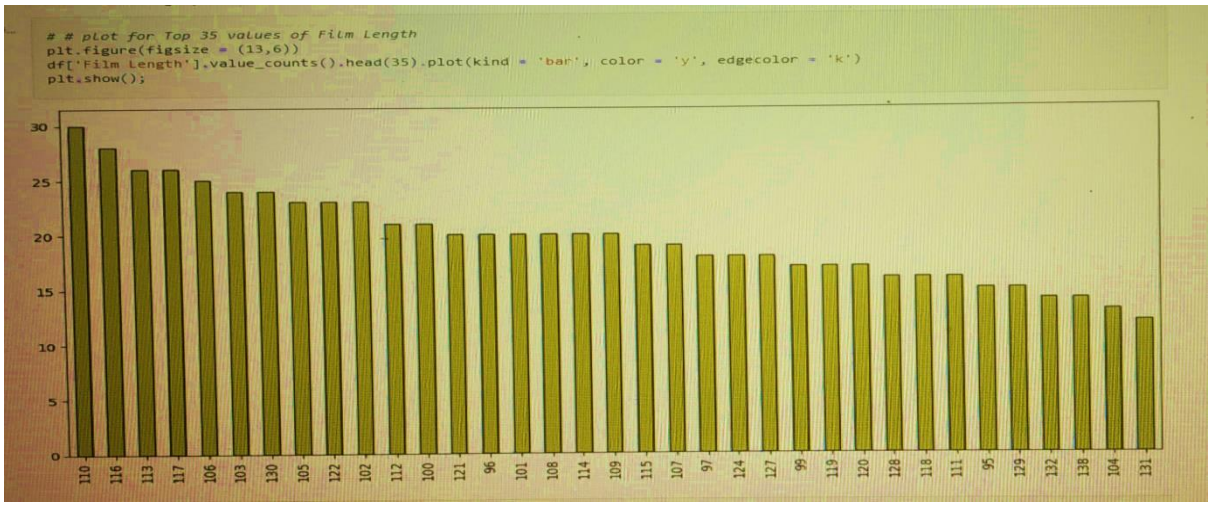
Insights:

1. Above 100 thriller Films got 6.7 Rating. And also, Highest number of Thriller Movies got 6.7 Rating as compared to other Ratings given to the Films.
2. Secondly, a greater number of Thriller Films got 6.8 IMDB Rating.
3. More number of Thriller Films got the top 10 IMDB Ratings are 6.7, 6.8, 7.0, 7.1, 7.3, 7.2, 7.5, 6.9, 7.4, 7.6.

If we want to get Bottom 15 IMDB Ratings of Thriller Films, we plot a Bar Plot.



If we want to know the Top 35 Film Durations of Thriller Films, we will plot barplot.

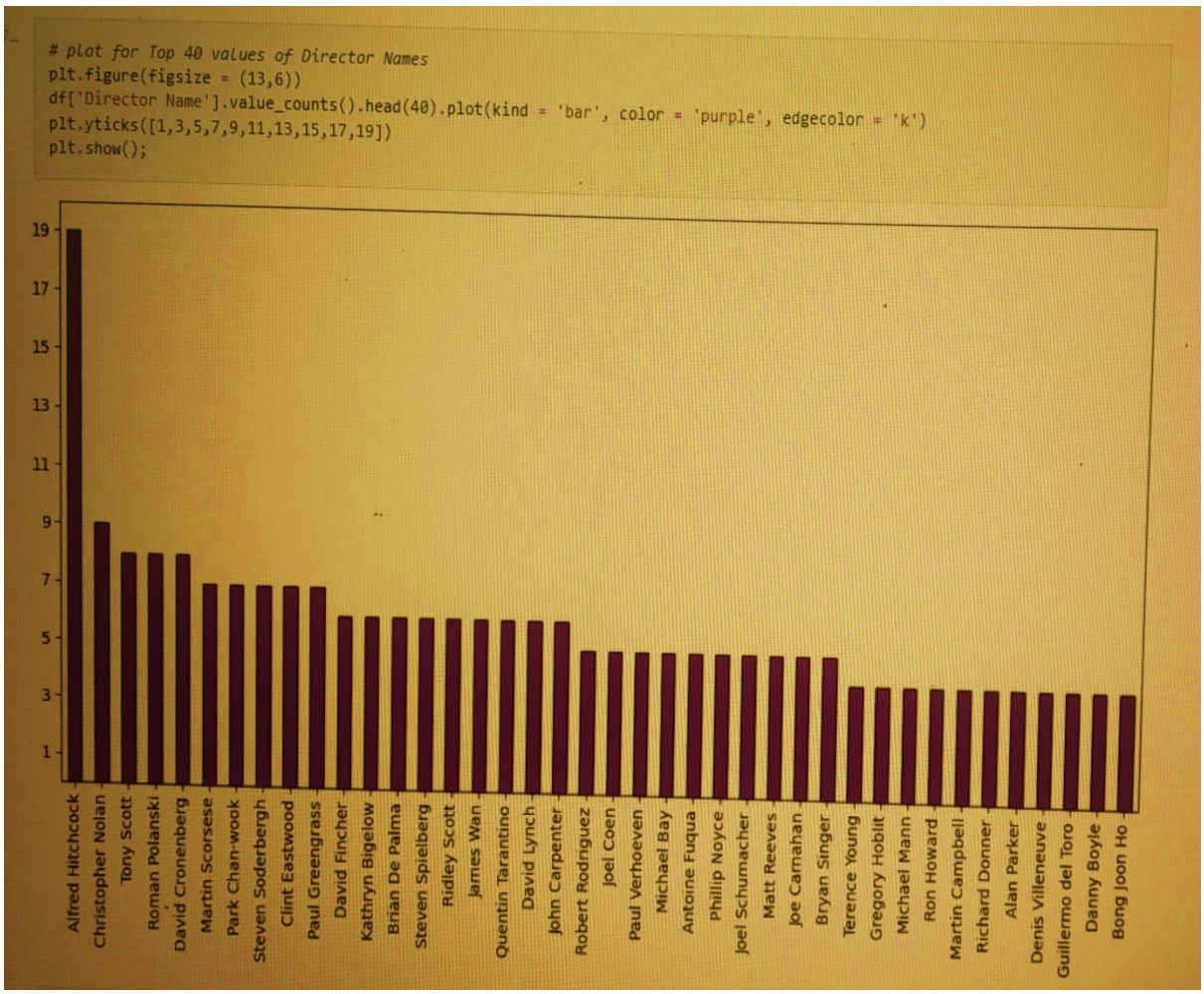


After ploting the barplot, we get to know some information.

Insights:

1. Films have the Duration 110 min is higher in number that is 30 as compared to other films Film Length

If we want to know the Top 40 Directors are directed a greater number of Films, we will plot the bar plot.



We get some Insights.

Insights:

1. Alfred Hitchcock has directed 19 films. He only directed more thriller Films as compared to other directors.
2. second, Christopher Nolan directed 9 films.