Exploratory Data Analysis Report

IMDB MOVIES DATASET

KEY POINTS OF ANALYSIS

- IMDB Movies dataset has been taken to analyse that which movie has performed exceptionally well at the box office.
- Duration and the performance of actors has also been analysed.
- In some portions, head or sample of data has been taken as, the dataset being too large to display.

IMDB Movies Dataset

	star_rating	title	content_rating	genre	duration	actors_list
0	9.3	The Shawshank Redemption	R	Crime	142	[u'Tim Robbins', u'Morgan Freeman', u'Bob Gunt...
1	9.2	The Godfather	R	Crime	175	[u'Marlon Brando', u'Al Pacino', u'James Caan']
2	9.1	The Godfather: Part II	R	Crime	200	[u'Al Pacino', u'Robert De Niro', u'Robert Duv...
3	9.0	The Dark Knight	PG-13	Action	152	[u'Christian Bale', u'Heath Ledger', u'Aaron E...
4	8.9	Pulp Fiction	R	Crime	154	[u'John Travolta', u'Uma Thurman', u'Samuel L....

The dataset consists of 6 columns and rest statistics have been explained in the python notebook.

Plotting and Visualising data

- Most widely preferred genre.
- Most preferred content-rating.
- Most preferred genre in short movies(i.e., with duration less than 100 minutes).
- Star-rating with highest movie count.
- Which actor has done the maximum movies ?

Pre Profiling Report

Dataset info

Number of variables	6
Number of observations	979
Total Missing (%)	0.1%
Total size in memory	46.0 KiB
Average record size in memory	48.1 B

Variables types

Numeric	2
Categorical	4
Boolean	0
Date	0
Text (Unique)	0
Rejected	0
Unsupported	0

Warnings

`actors_list` has a high cardinality: 969 distinct values Warning

`title` has a high cardinality: 975 distinct values Warning

As, can be seen in pre profiling report, the data is still to be cleaned, yet it is performed in the data cleaning and pre processing phase.

Post Profiling Report

Dataset info

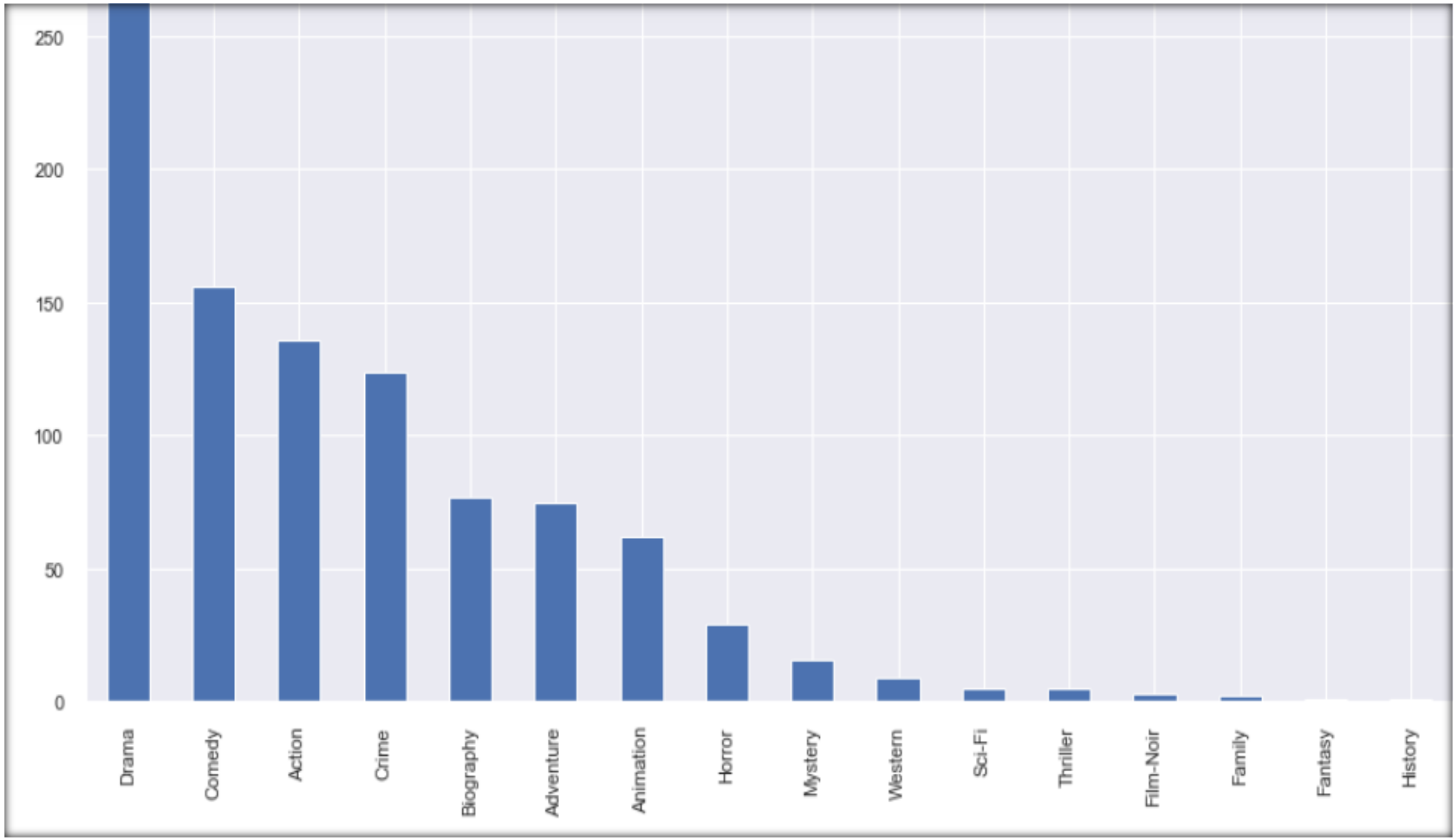
Number of variables	6
Number of observations	979
Total Missing (%)	0.0%
Total size in memory	46.0 KiB
Average record size in memory	48.1 B

Variables types

Numeric	2
Categorical	4
Boolean	0
Date	0
Text (Unique)	0
Rejected	0
Unsupported	0

As, can be seen in post profiling report, the data has been cleaned and missing values have been replaced.

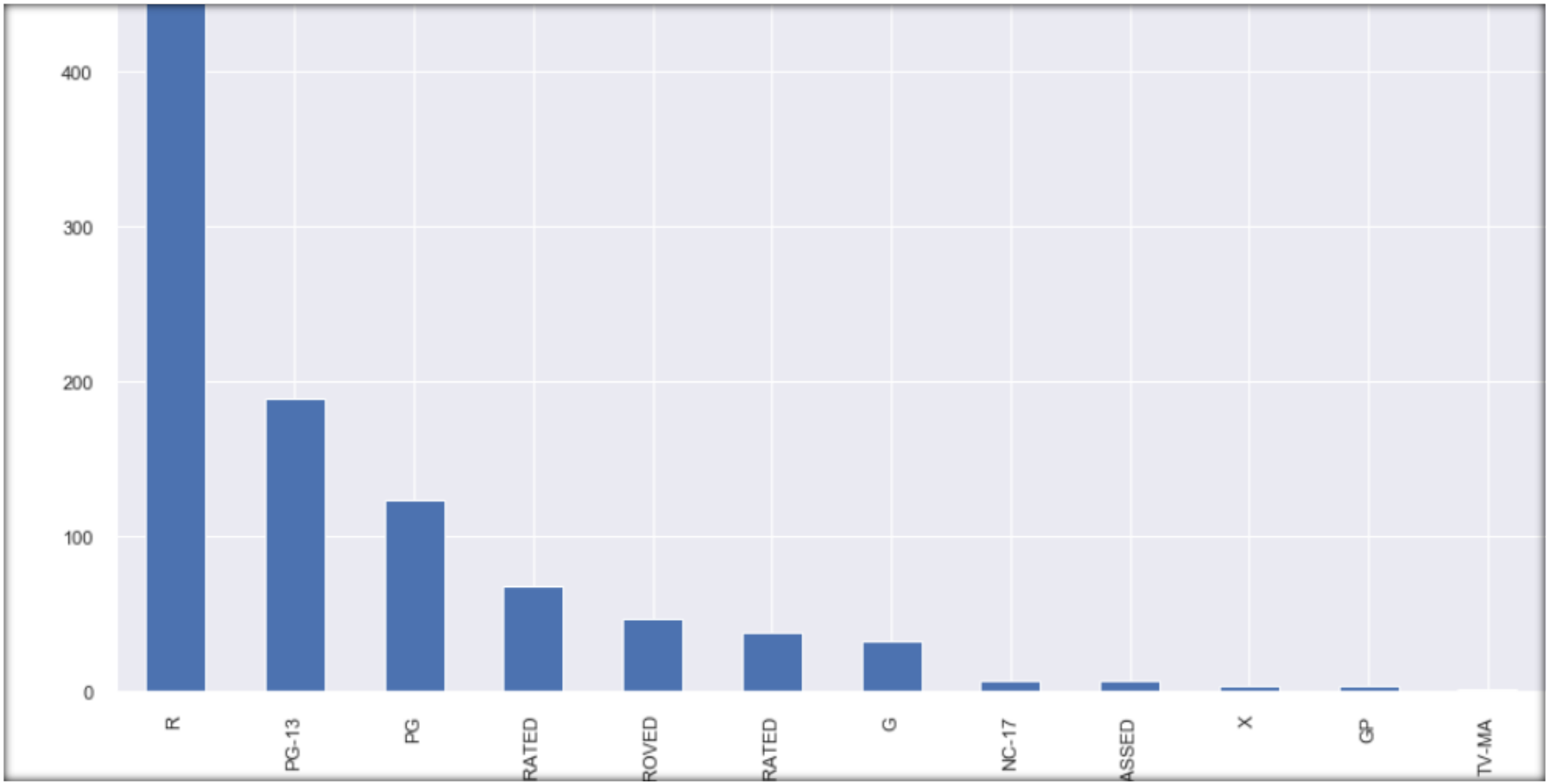
Most widely preferred Genre



Most widely preferred Genre is : “Drama”.

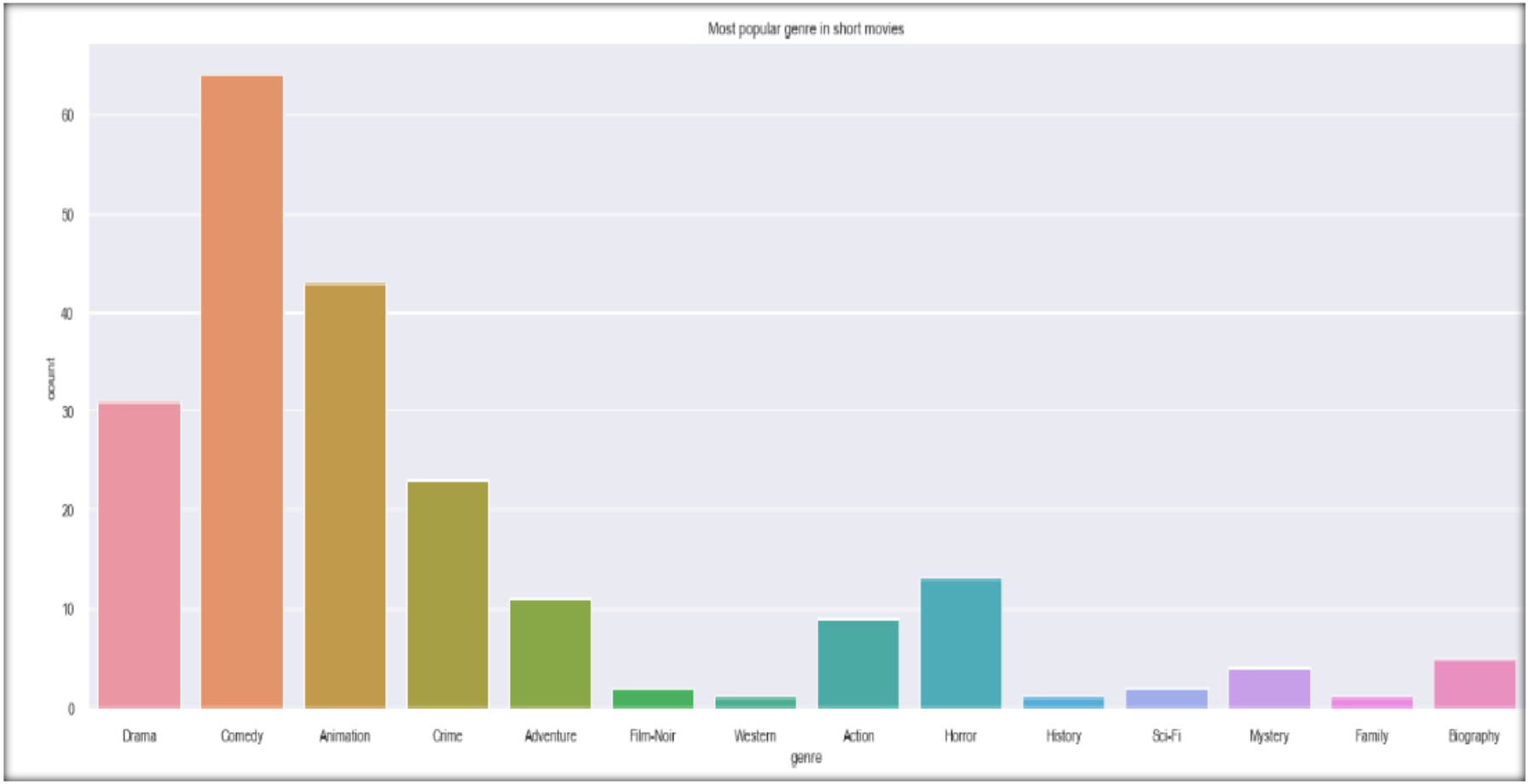
Most preferred content-rating

Most widely used content-rating among all movies.



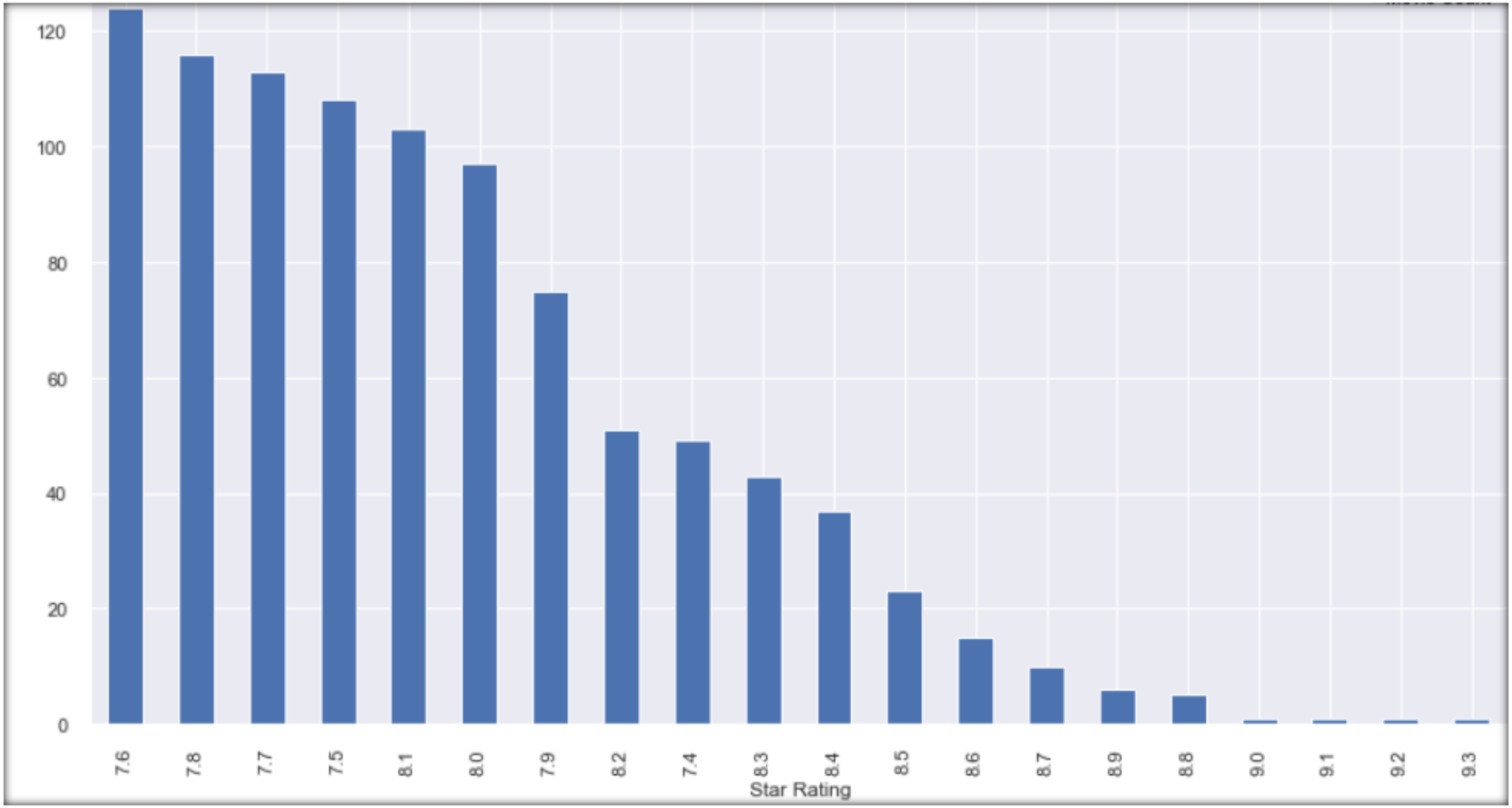
The most preferred content-rating among all movies is R rated.

Most preferred Genre in short movies.



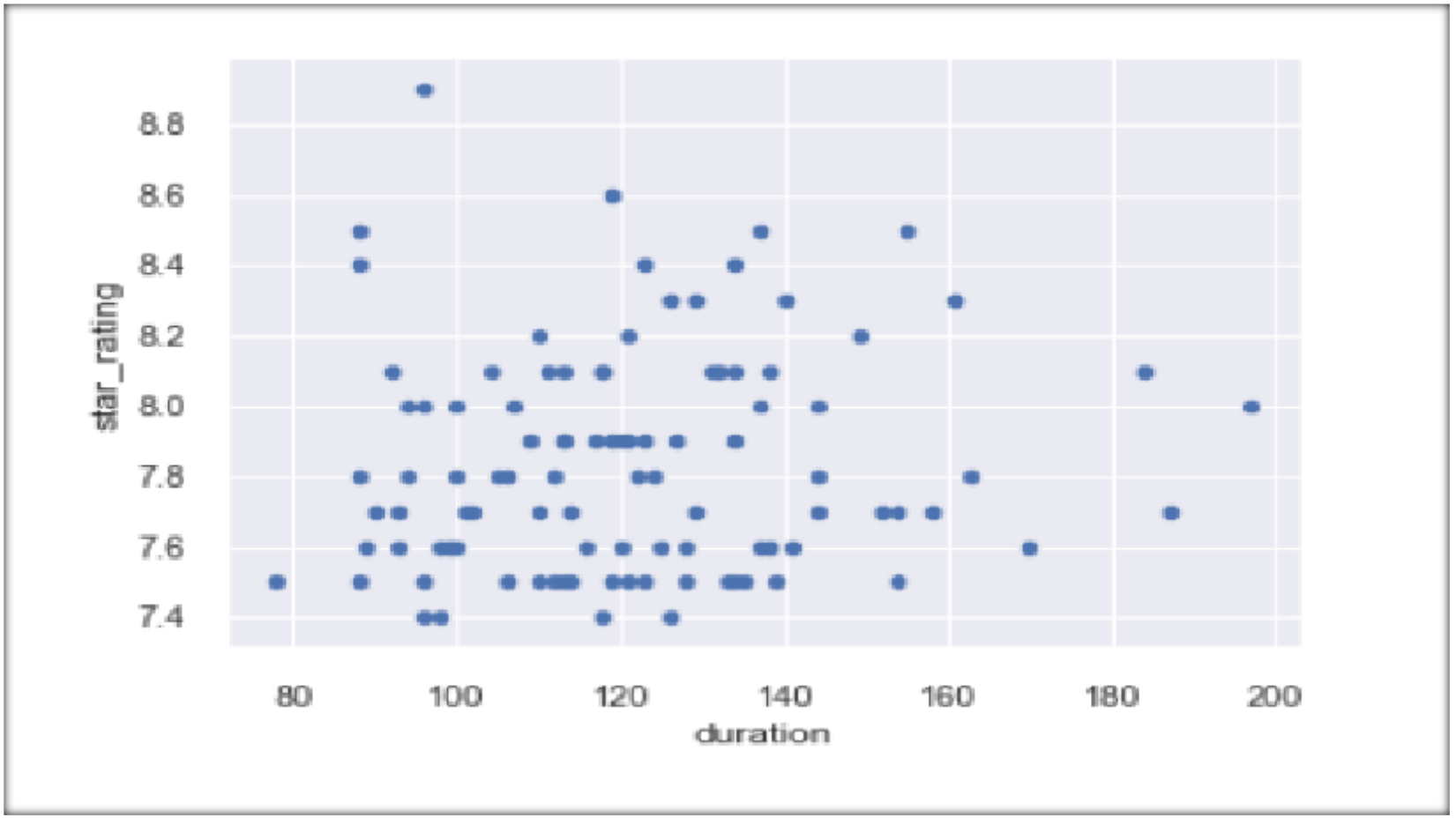
The most preferred Genre in short movies is Comedy.

Star-rating with the highest movie count



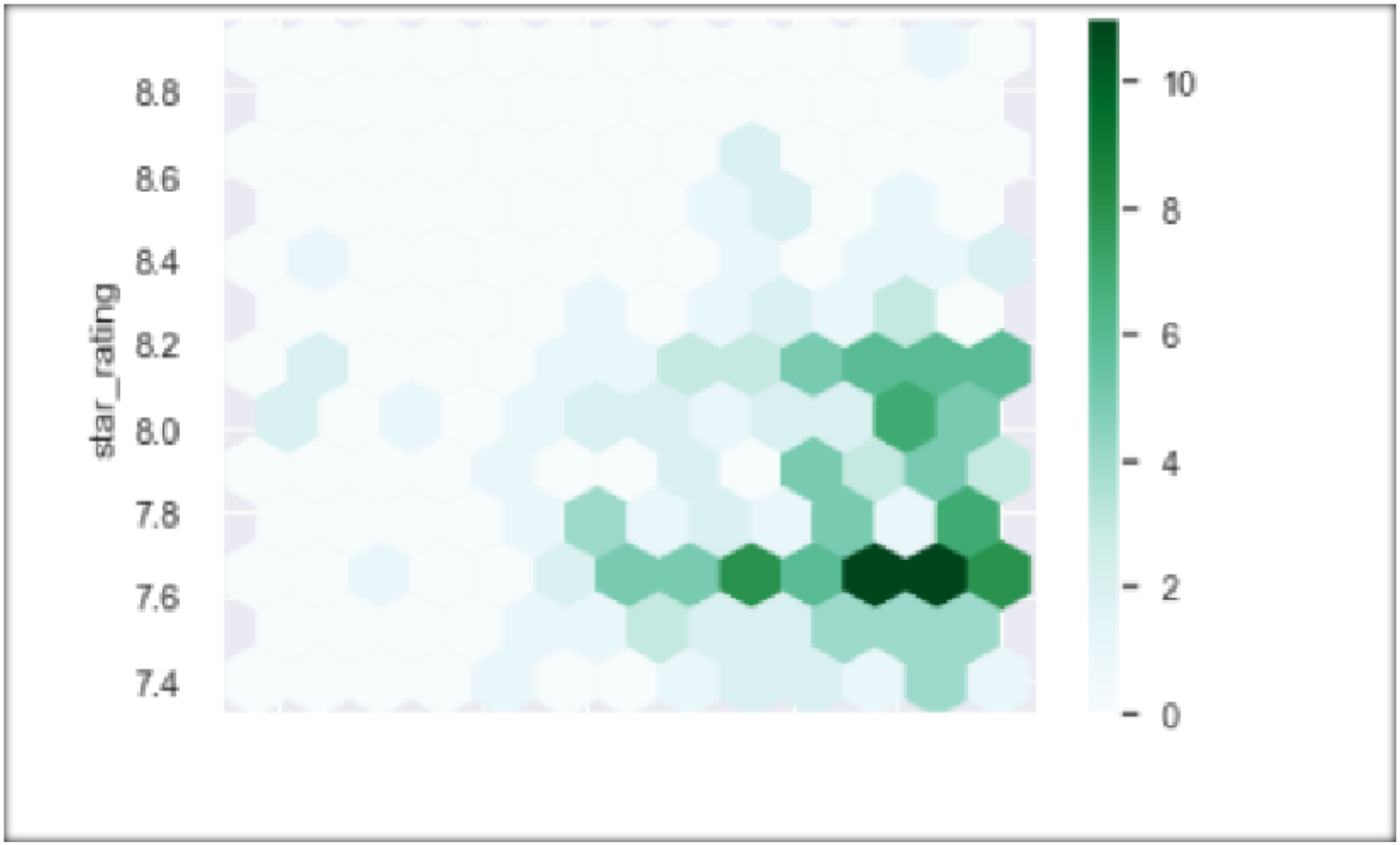
Star-rating with highest movie count is 7.6.

Sample representation using Scatter plot



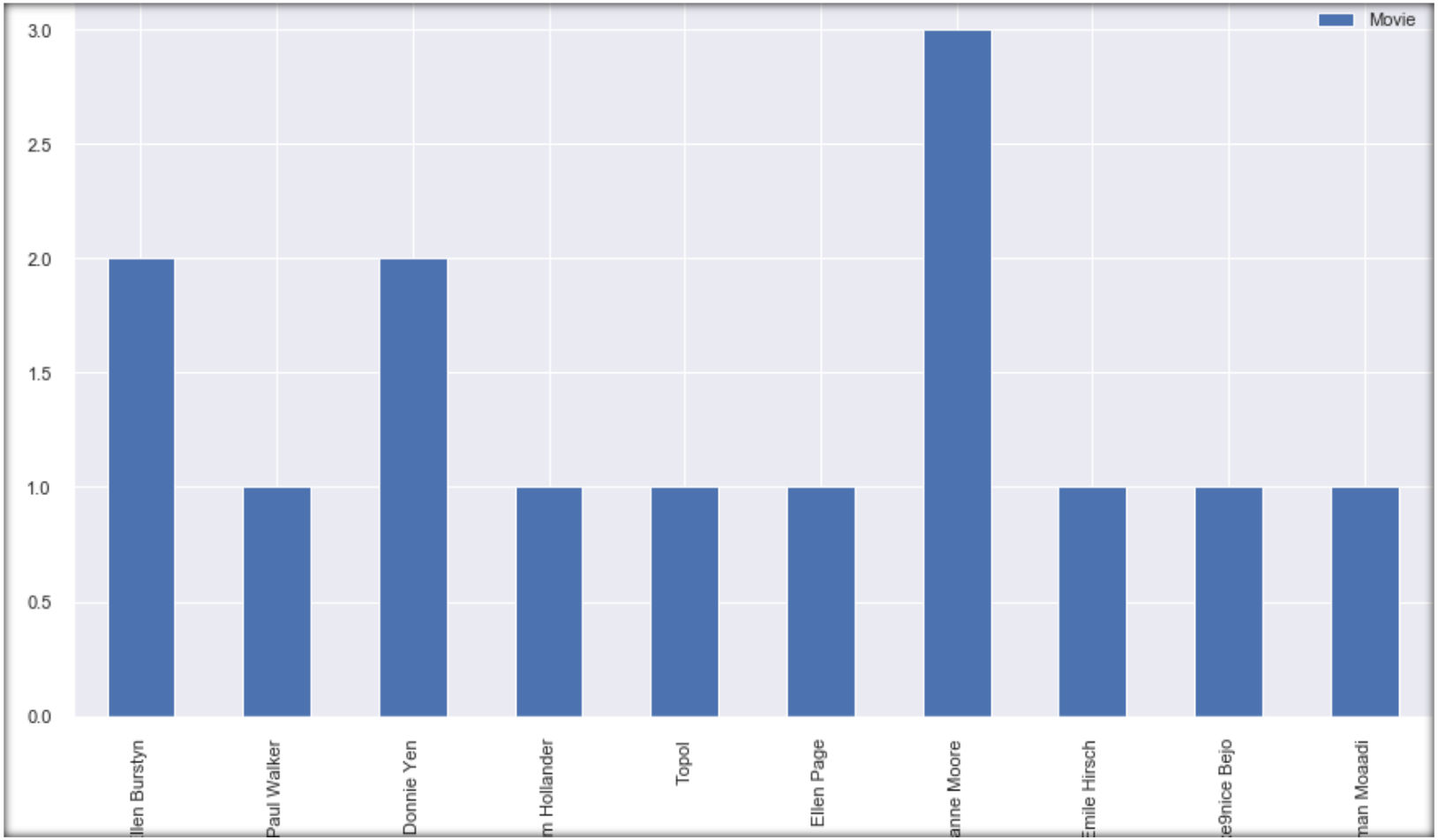
Representation of sample data using Scatter plot.

Hexplot for short movies using sample data

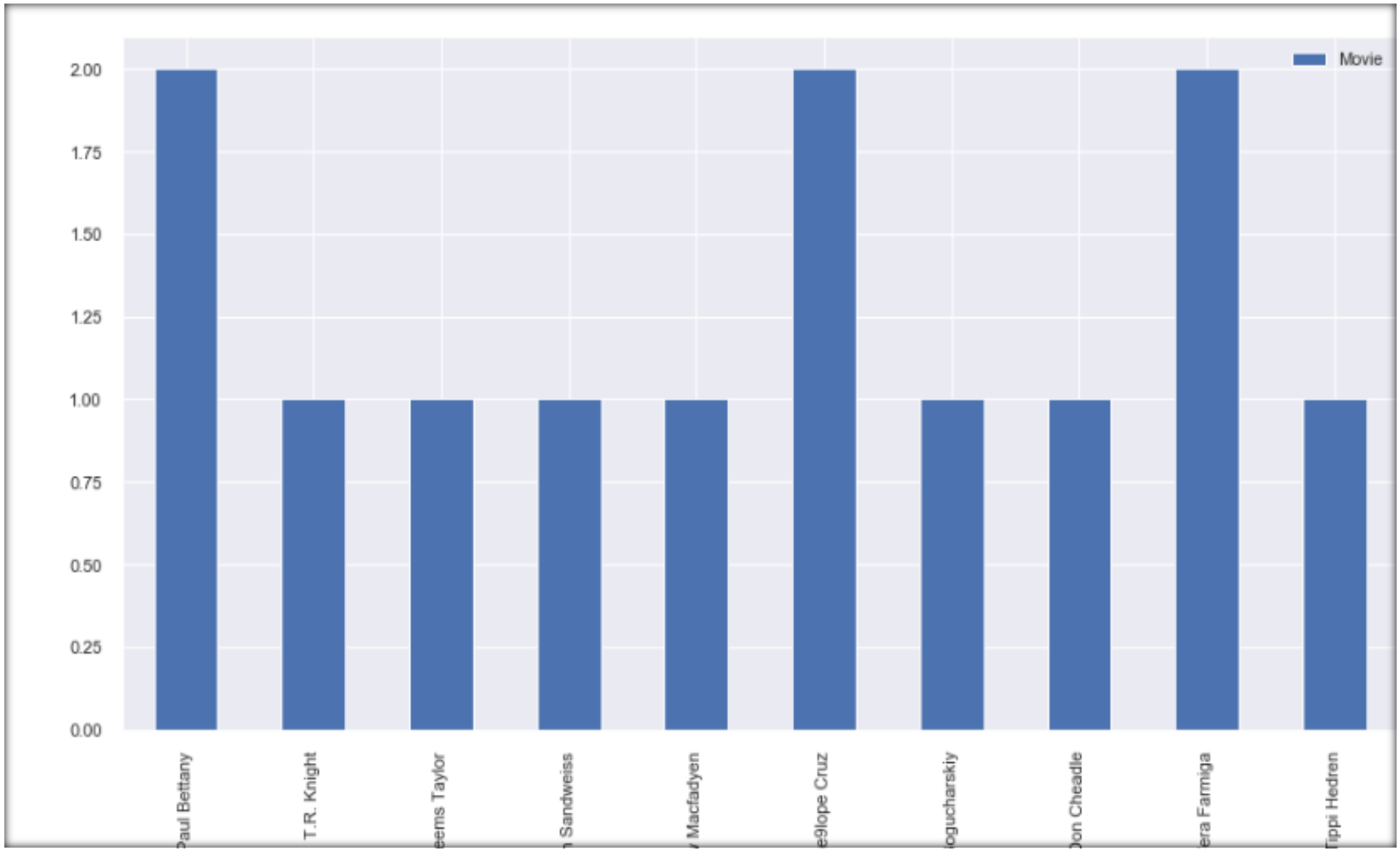


Representation of sample data from short movies using Hexplot.

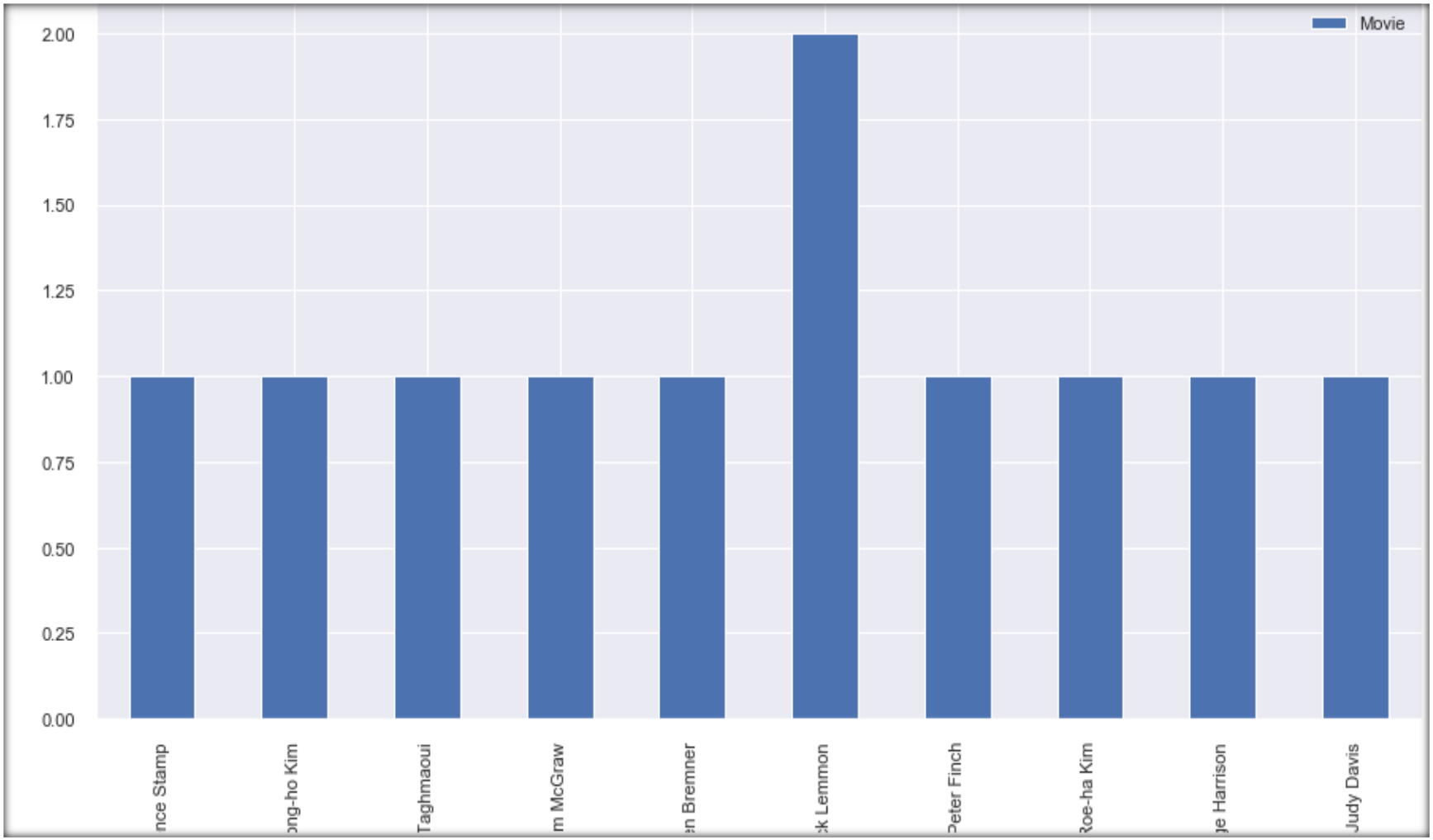
Splitting up actors_list into 3 different lists and analysing which actor did the most movies.



actor_list divided and plotted as first list of actors.



actor_list divided and plotted as second list of actors.



actor_list divided and plotted as third list of actors.

Thank You and Have a Good Day !