Microsoft

MICROSOFT MOVIE STUDIO PROJECT

UNDERSTANDING THE PROBLEM

- Microsoft is seeking to open up a new movie studio based on its competitors who are all focusing on creating original video content. The role as a data scientist is to review the different types of films and access the films that are doing best at the box office. This question will be answered by reviewing the ratings, directors, genre, language among other attributes of these movies.

PROBLEM STATEMENT

- The problem statement is to determine the movies that are doing best at the box office by reviewing various attributes of the movies. This include the movie rating, the genre, number of votes, among other attributes.

DATA UNDERSTANDING

DATA COLLECTION

- The data was collected from imdb, box office mojo, rotten tomatoes, themoviedb and the numbers.

Microsoft

DATA PREPARATION

SELECTING DATA

We’ll use data from four tables in the imdb data set: movie_ratings, movie_basics, directors and persons.

DATA PREPARATION

DATA CLEANING

- This was done to ensure the Validity, Accuracy, Completeness, Consistency and Uniformity of the Data.
- Missing values in the datasets were checked for and were found to be none. The data was found to be consistent there being no duplicated data.

EXPLORATORY DATA ANALYSIS

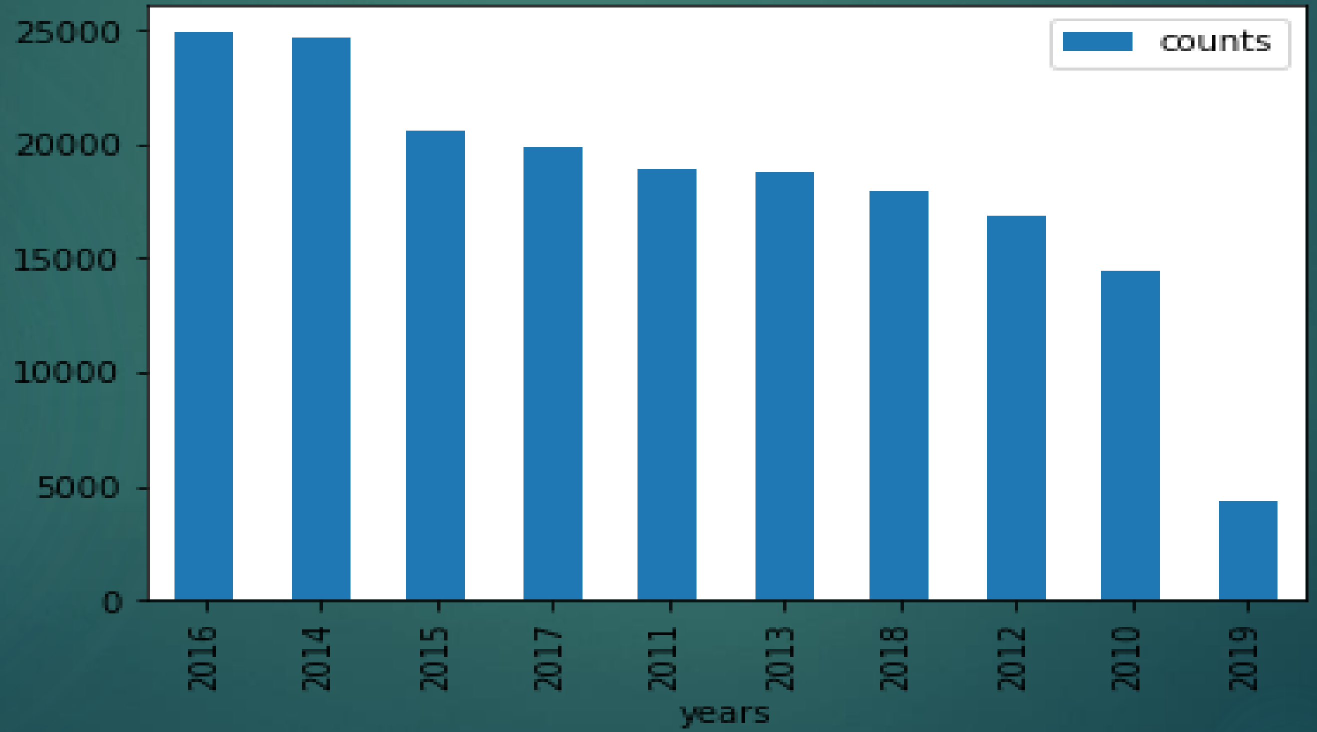
UNIVARIATE DATA ANALYSIS

Numerical Data

There were a number of numerical data that we worked on, this include the number of movies produced per year, the number of movies produced by the top directors, the rank of the movies based on the movie ratings.

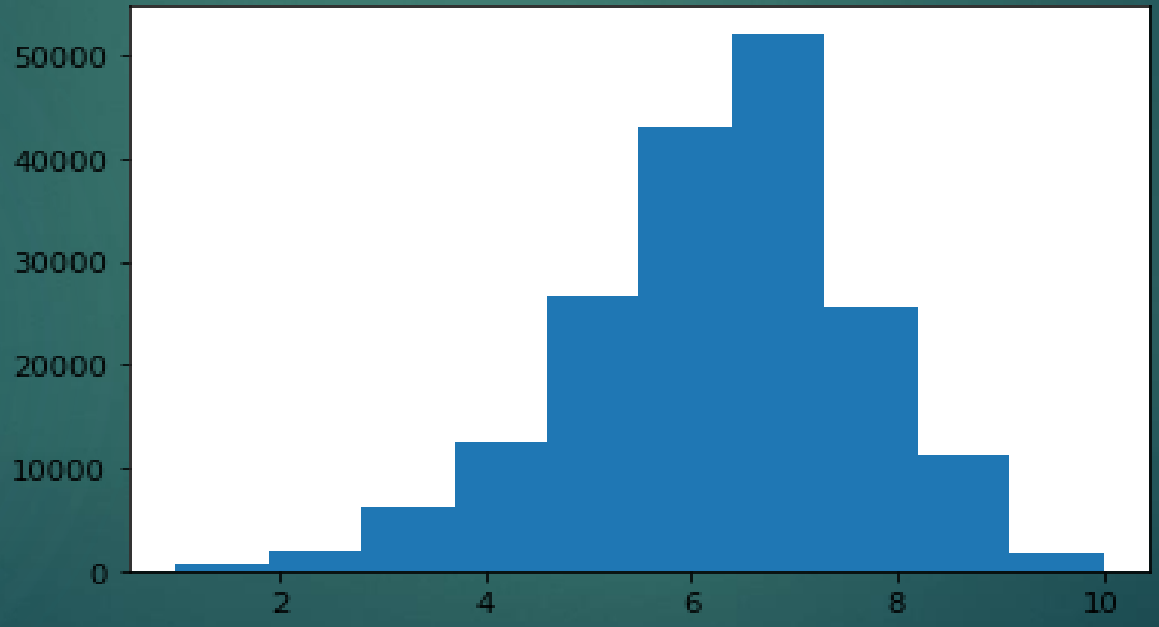
EXPLORATORY DATA ANALYSIS

Number of Movies per year



EXPLORATORY DATA ANALYSIS

IMDB Rating Distribution



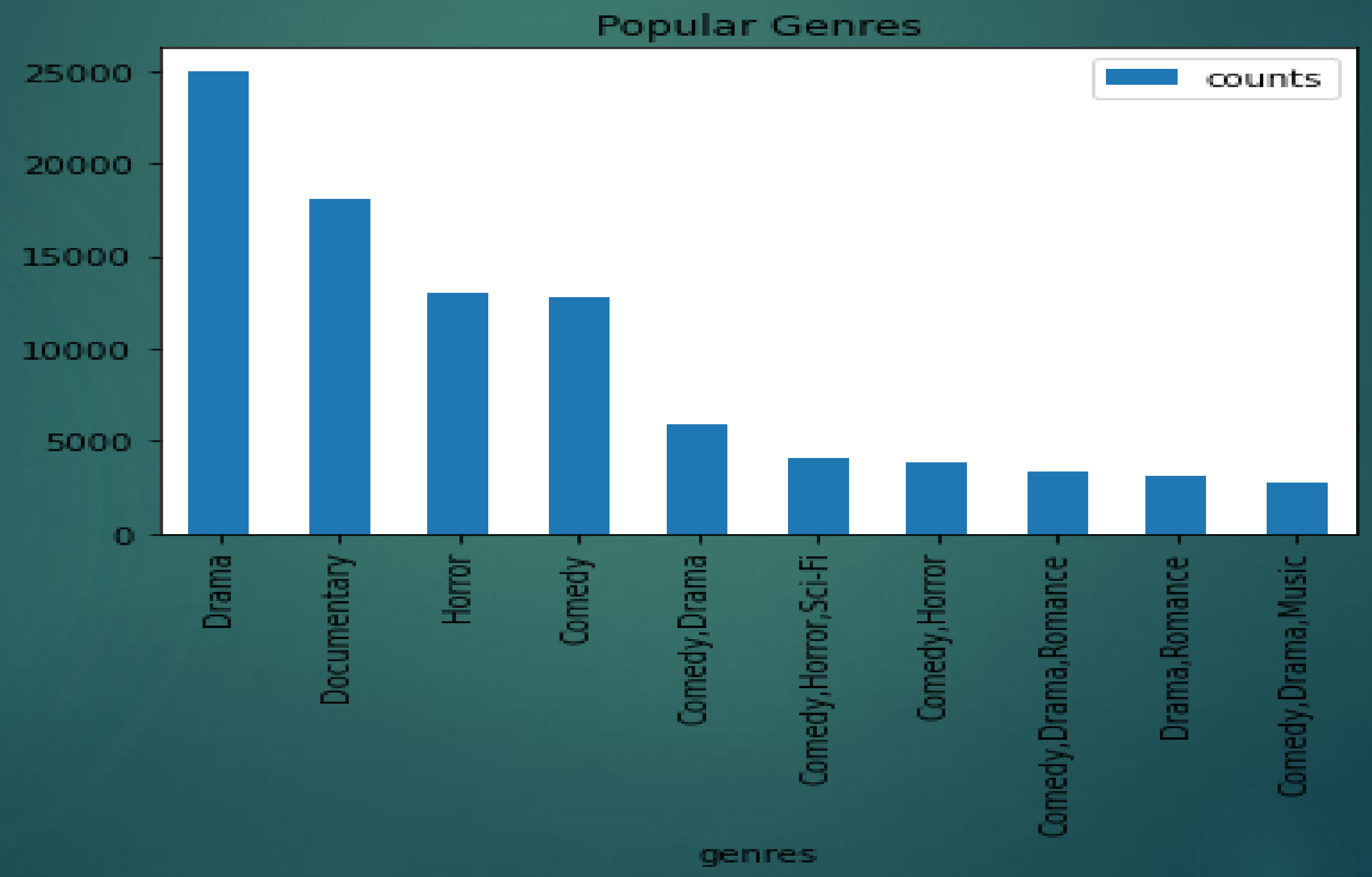
EXPLORATORY DATA ANALYSIS

Categorical Data

- Most of the analysis was done using categorical data where we the categorized the movies based on the popular genres by reviewing the number of movies in each genre. We also identified the top directors in the various categories. In addition, we reviewed the distribution of the movie ratings in the top five genres in order to see whether there is a direct correlation between the ratings and popular movies based on the number of movies.

EXPLORATORY DATA ANALYSIS

Popular Genres



Microsoft

Summary Statistics

	start_year	runtime_minutes	averagerating
count	181387	163584	181387
mean	2014.309802	97.789484	6.217683
std	2.536111	194.434689	1.388026
min	2010	3	1
25%	2012	84	5.4
50%	2014	94	6.3
75%	2016	107	7.2
max	2019	51420	10

Univariate Analysis Recommendation

Based on our findings we would recommend Microsoft to produce drama, documentary, horror and comedy and comedy_drama as this were the most popular genres based on the movie count. Among these documentary, drama and comedy_drama are highly rated.

RECOMMENDATION

Based on the analysis done we recommend Microsoft to take up the documentary, drama and comedy_drama genres and consider basing their production on these top three genres.