Investigation of the Movies Dataset

Introduction:

a) **Dataset chosen for analysis:** TMDb Movies Data. The dataset contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue.

Data Wrangling:

a) **Number of rows in the dataset (before cleaning)**: 10866
b) **Number of columns in the dataset (before cleaning)**: 21
c) **Number of movie enlisted**: 10866

d) **Qualitative variables in the dataset:** Original title, director, homepage, production companies, tagline, overview, genres
e) **Quantitative variables in the dataset:** Budget, popularity, vote_count, vote_average

f) **Number of NaN values**: 13434
g) **Columns with missing values**: imdb_id, cast, homepage, director, tagline, keywords, overview, genres, production companies
h) **Number of duplicate values**: 1

Common Problems with the data:

a) Numerous rows have missing values
b) Some columns of data are not required for investigation and must be dropped
c) Duplicates must be removed

General Properties:

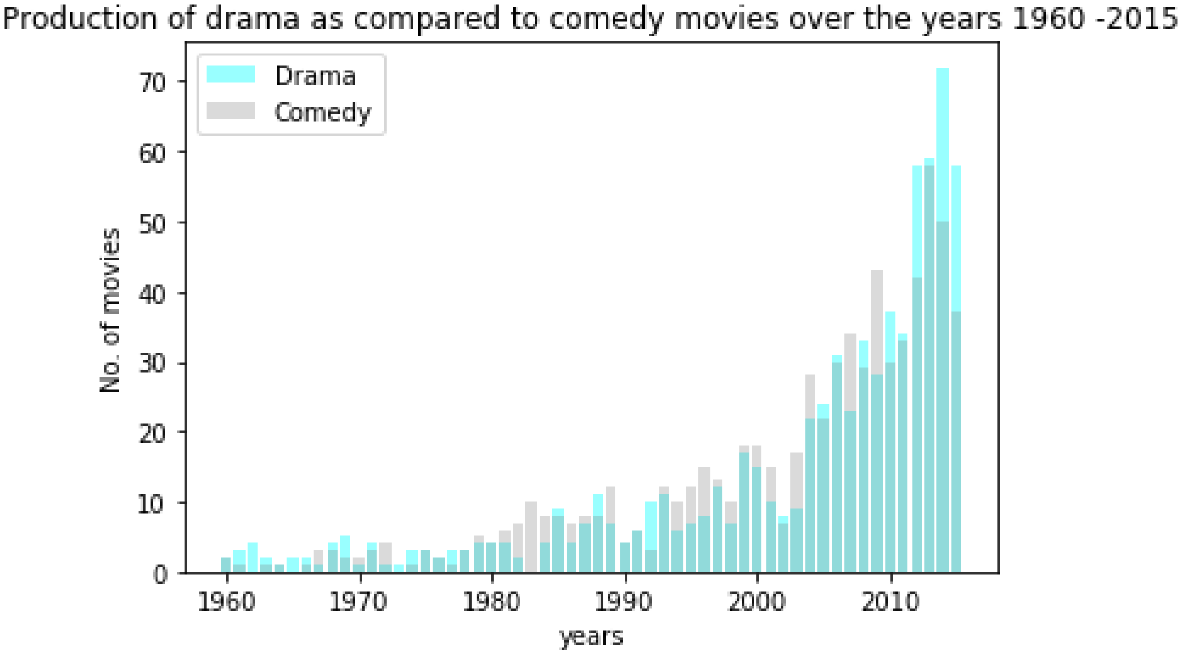
a) **Mean average vote**: 5.975011504832047
b) **Oldest Release year**: 1960
c) **Most recent Release Year**: 2015
d) **Mean Runtime of the movies**: 102.07179015186378
e) **Most Common genre**: Drama (712)
f) **Mean popularity**: 0.6464455549010583
g) **Most common producer**: Paramount Pictures (156)
h) **Most common director**: Woody Allen

Investigation of the Movies Dataset

Exploratory Data Analysis:

1) **Which genres are most popular from year to year?**

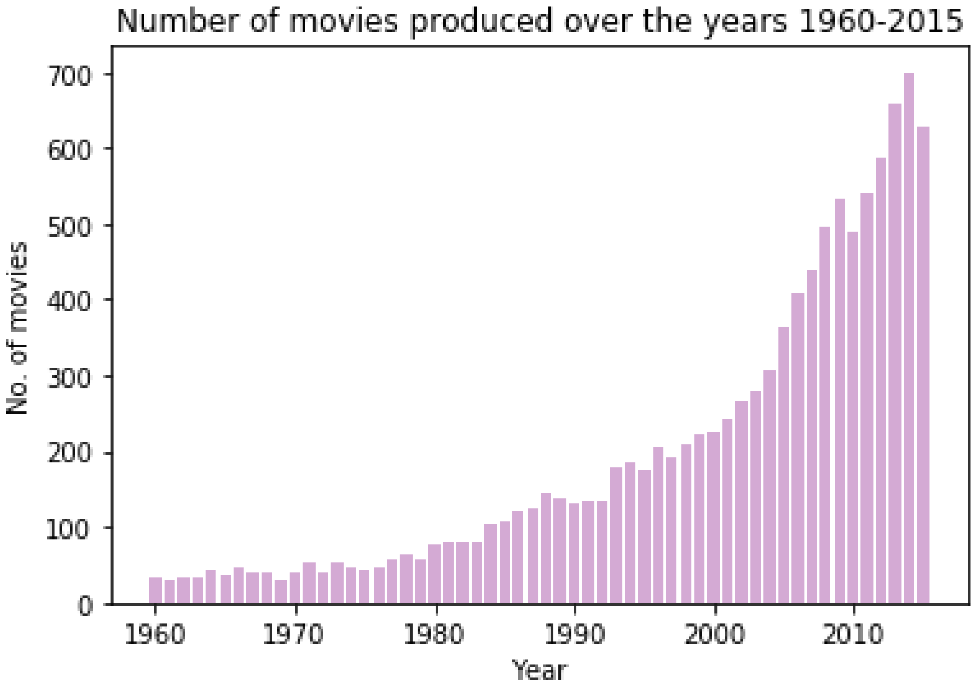
- Drama has been the most popular genre with a count of 712, followed by Comedy (with or without a supplemental genre) and Horror (with or without a supplemental genre like thriller, sci-fi etc.)
- Horror was the most popular genre in the early 1960s, 1970 and 1978
- In the 1980s and 1990s, comedy was the **most popular genre**



- It can be noted that the number of drama movies slowly increased and surpassed the number of comedy movies. The difference is significant during 2014-2015. Nonetheless, the production of both movies has increased over time

2) **It has been suspected that globalization has led to increased cross-cultural interactions. This particularly manifests in movies as different culture can be physically rendered via films. Is this true? Has there been an increase in the number of movies over the years?**

Investigation of the Movies Dataset

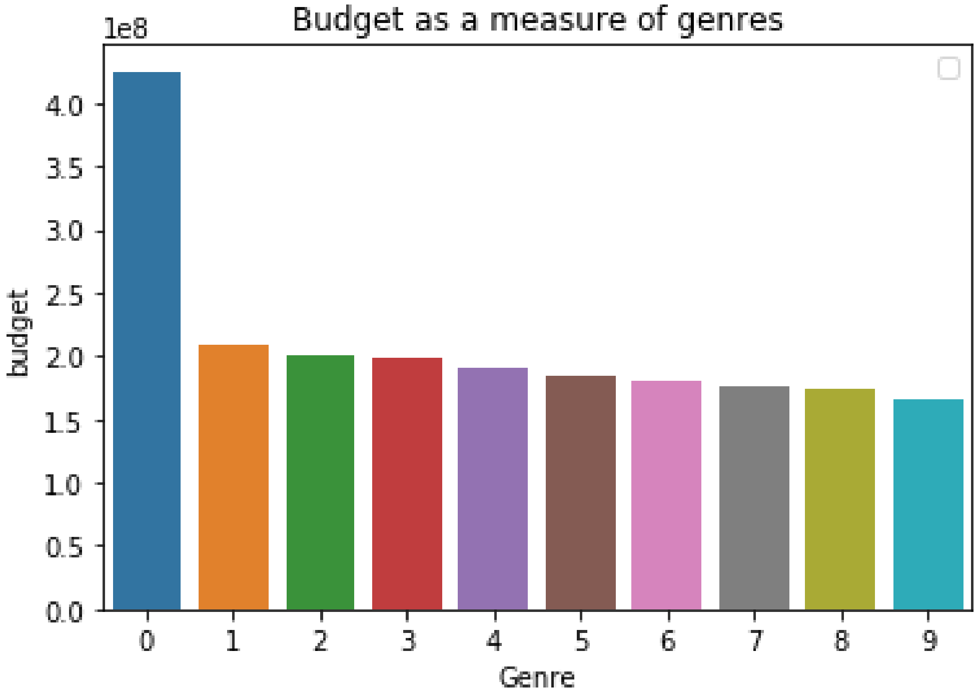


- It can be clearly seen that the number of movies being produced from 1960-2015 has increased exponentially. Thus, globalization is perhaps one of the causes of this increase.

3) How do genres relate to the budget of the movies?

X-tick label	Genres	Budget in millions of USD
0	Adventure|Fantasy|Action|Western|Thriller	425000000.0
1	Thriller|Action|Adventure|Science Fiction	209000000.0
2	Family|Fantasy|Adventure	200000000.0
3	Adventure|Action|Fantasy	198000000.0
4	Action|Family|Science Fiction|Adventure|Mystery	190000000.0
5	Animation|Adventure|Comedy|Family|Action	185000000.0
6	Fantasy|Adventure|Action|Family|Romance	180000000.0
7	Science Fiction|Fantasy|Action|Adventure	176000003.0
8	War|Adventure|Drama	175000000.0
9	Adventure|Family|Animation|Action|Comedy	165000000.0

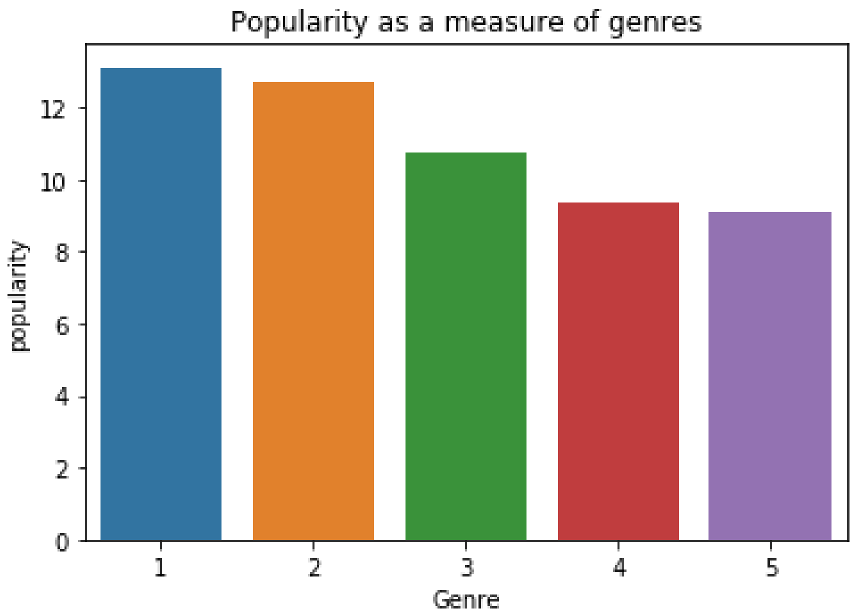
Investigation of the Movies Dataset



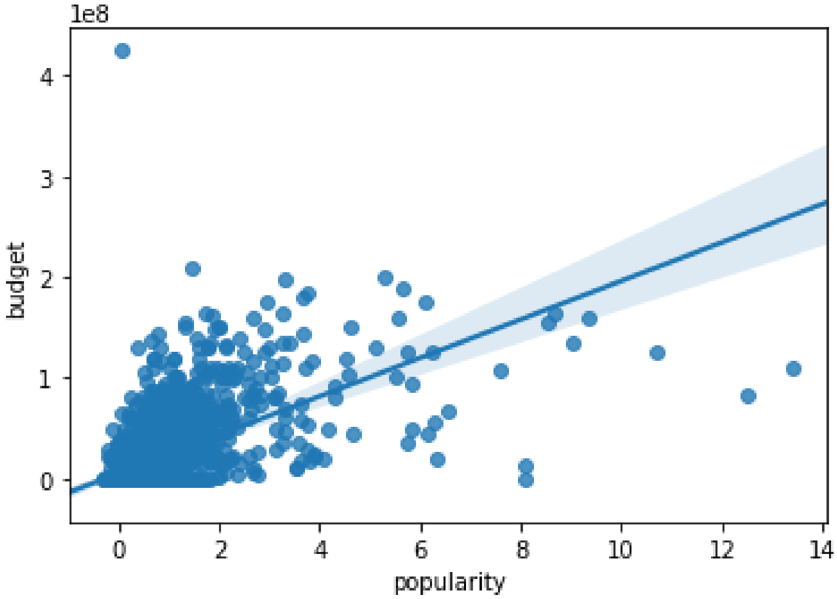
4) How do genres relate to the popularity of the movies?

Tick Number	Genre	Popularity
1	Adventure|Science Fiction|Thriller	13.112507
2	Adventure|Drama|Science Fiction	12.699699
3	Science Fiction|Adventure|Thriller	10.739009
4	Action|Thriller|Science Fiction|Mystery|Adventure	9.363643
5	Western|Drama|Adventure|Thriller	9.110700

Investigation of the Movies Dataset



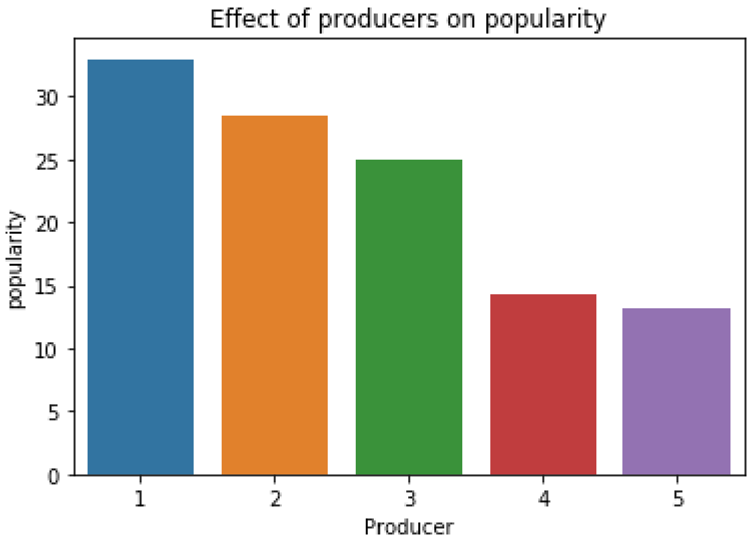
5) How does popularity correlate with budget?



It can be observed that there is a positive correlation between popularity and budget. Thus, a higher budget movies usually have higher popularity. However, there are some outliers. A few movies have had extremely high budgets but low popularity while some popular movies have been produced using low budgets

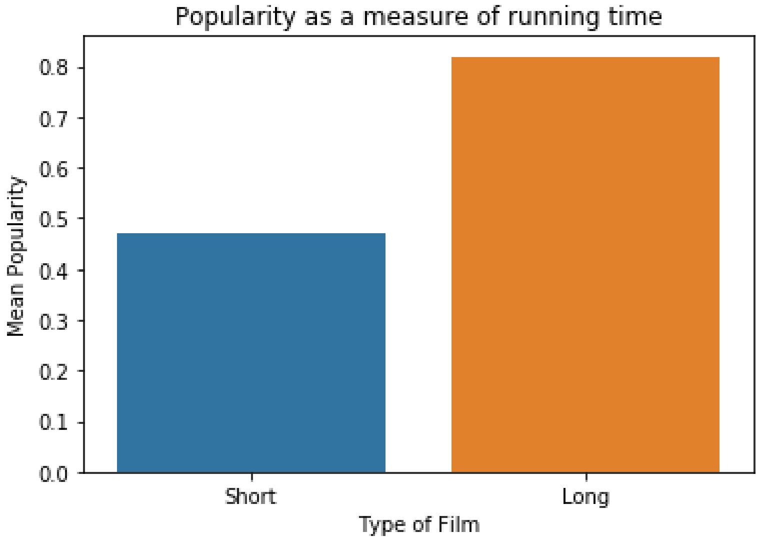
6) Producers and their effect on the popularity of a movie

Tick_label	Producers	Popularity
1	Universal Studios|Amblin Entertainment|Legendary Pictures|Fuji Television Network|Dentsu	32.985763
2	Village Roadshow Pictures|Kennedy Miller Productions	28.419936
3	Paramount Pictures|Legendary Pictures|Warner Bros.|Syncopy|Lynda Obst Productions	24.949134
4	Marvel Studios|Moving Picture Company (MPC)|Bulletproof Cupid|Revolution Sun Studios	14.311205
5	Summit Entertainment|Mandeville Films|Red Wagon Entertainment|NeoReel	13.112507



7) How does running time affect popularity?

Investigation of the Movies Dataset



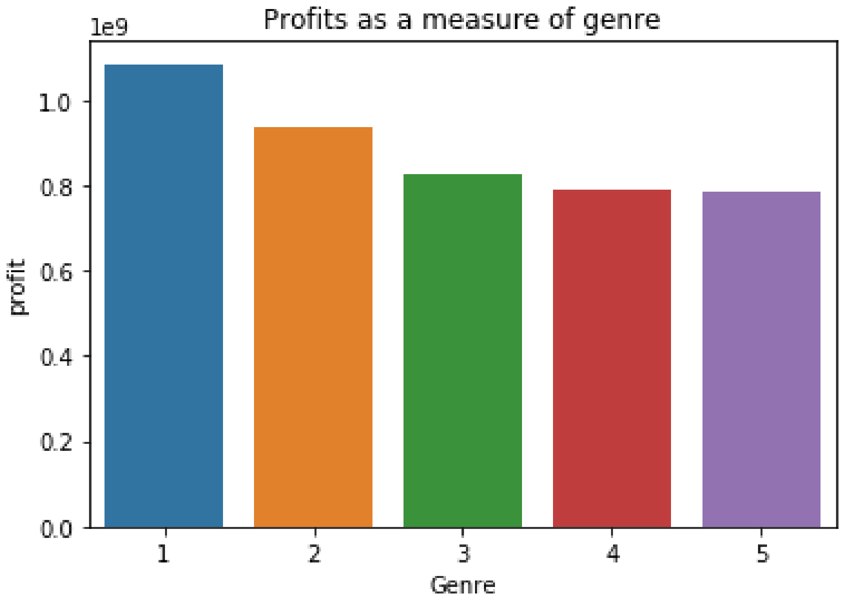
Movies were categorized into short and long ones based on whether the running time was greater than the median running time of the dataset. From the graph it can be observed that, longer movies have had a significantly higher popularity than the shorter ones

8) **Which genres have has the highest profits?**

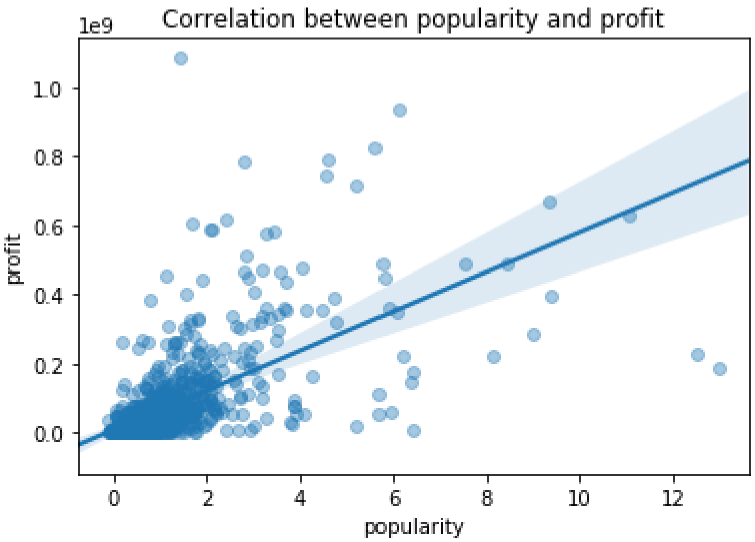
A separate column was created in the data frame which included the profits as the difference between the budget and revenue. Only the positive values were used for comparison

Tick label	Genre	Profit
1	Crime|Drama|Mystery|Thriller|Action	1.084280 e^9
2	Action|Adventure|Science Fiction|Fantasy	9.340891 e^8
3	Family|Fantasy|Adventure	8.254671 e^8
4	Adventure|Fantasy|Family|Mystery	7.882127 e^8
5	Science Fiction|Adventure|Family|Fantasy	7.824106 e^8

Investigation of the Movies Dataset



9) What is the correlation between popularity and profits?



It can be clearly observed that there is a very strong positive correlation between popularity of movies and the profits. Higher popularity has seen higher profits.

Conclusions:-

Investigation of the Movies Dataset

From the exploratory data analysis carried out and the trends identified, it can be concluded that in order to make a popular and successful movie a person:

1) Should keep an element of drama in the movie
2) Should have a theme of science-fiction, adventure or thrill in the movie because they have been found to be of high popularity
3) Should make sufficient monetary investment of at least 150 million USD in a good casting crew to reap maximum results
4) Should target a running time of more than 100 minutes
5) The movie should be produced by the production companies found to produce the most popular movies as identified in the analysis above.

With these he can expect higher profits as identified by the trends in the scatterplots.

Additional Resources:- N/A