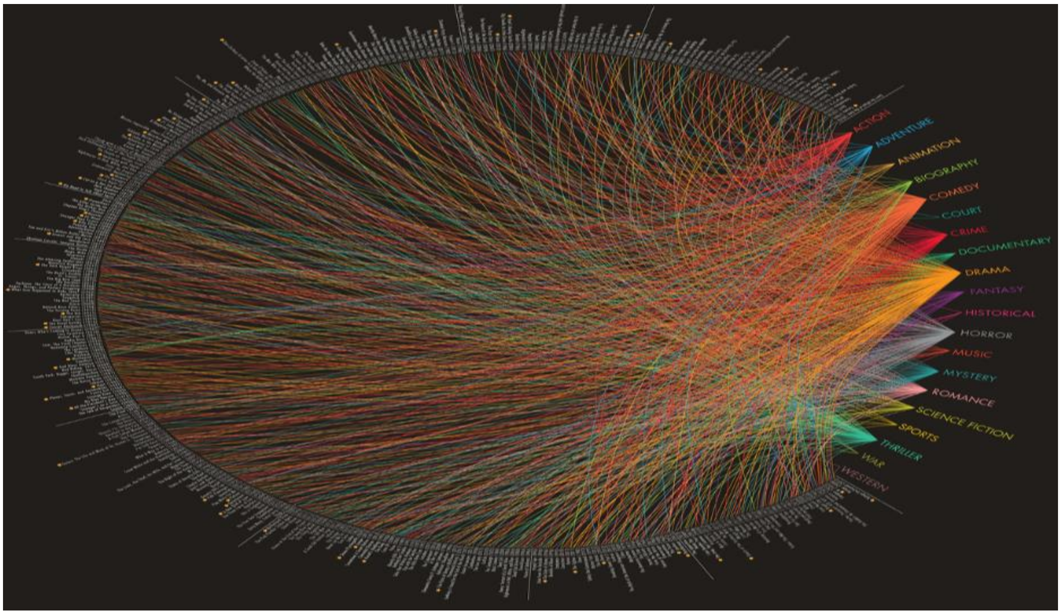
Final Project Visualisation

on

Exploring Movie Data with Interactive Visualizations.

Submitted as a part of course curriculum for TSWD



Under the guidance of

Prof Rituparna Jawahar

PES University

Submitted by

AKSHAY B K
PES2UG19CS030

Initial Analysis of the Dataset:

The data set I chose is Movie dataset

Here are some notes and comments about this datasets :

This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue.

Certain columns, like ‘cast’ and ‘genres’, contain multiple values separated by pipe (|) characters.

There are some odd characters in the ‘cast’ column. We can leave them as it is.

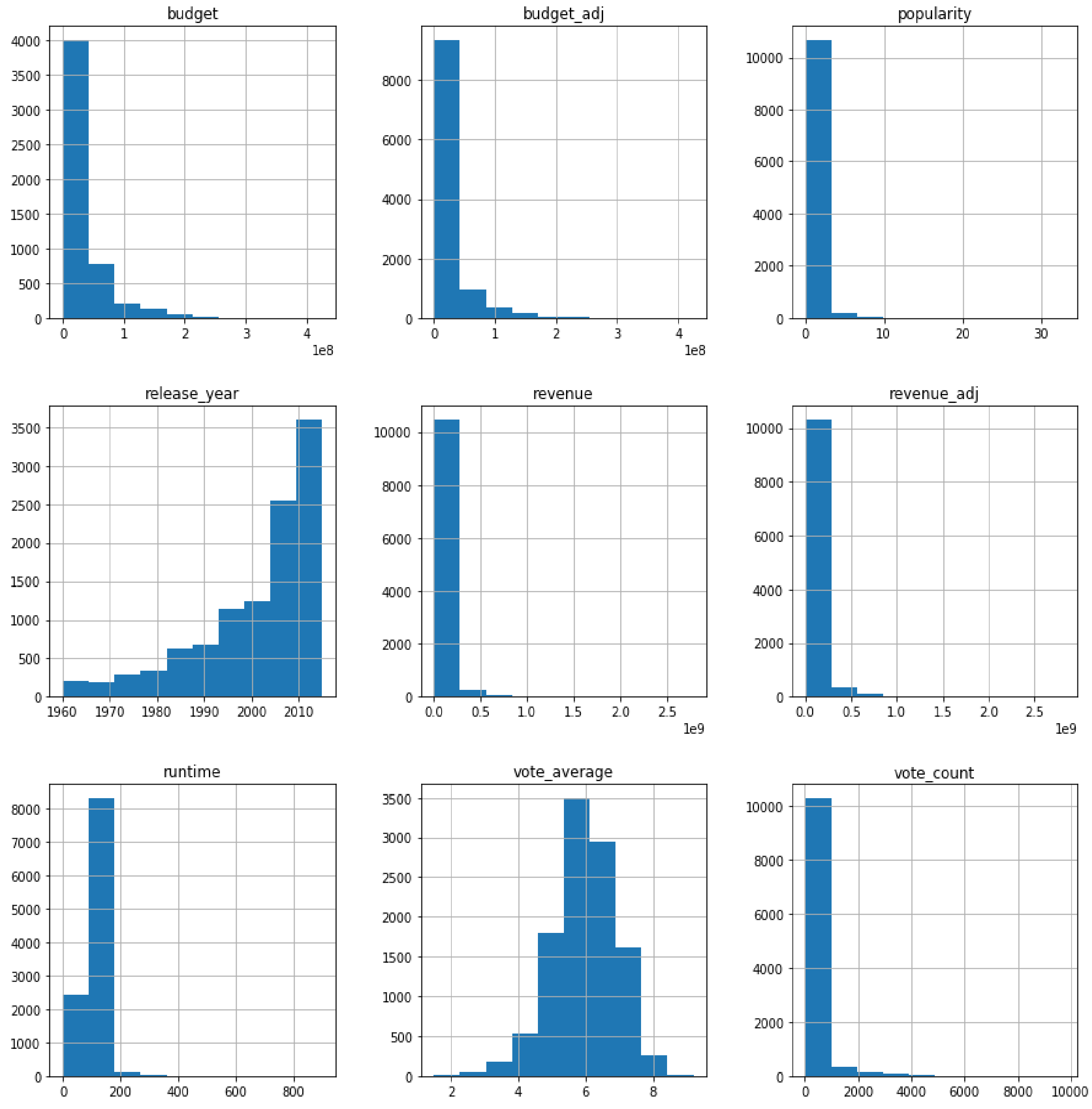
The final two columns ending with “_adj” show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time.

Get familiar with the data types, data structure. I did delete the duplicates and unuseful columns like imdb_id,homepage etc.

When handling the missing data. I use two ways: for all the missing data with data type object, i fill the null with string "missing". For budget, datatype integer,I fill 0 with np.NAN.

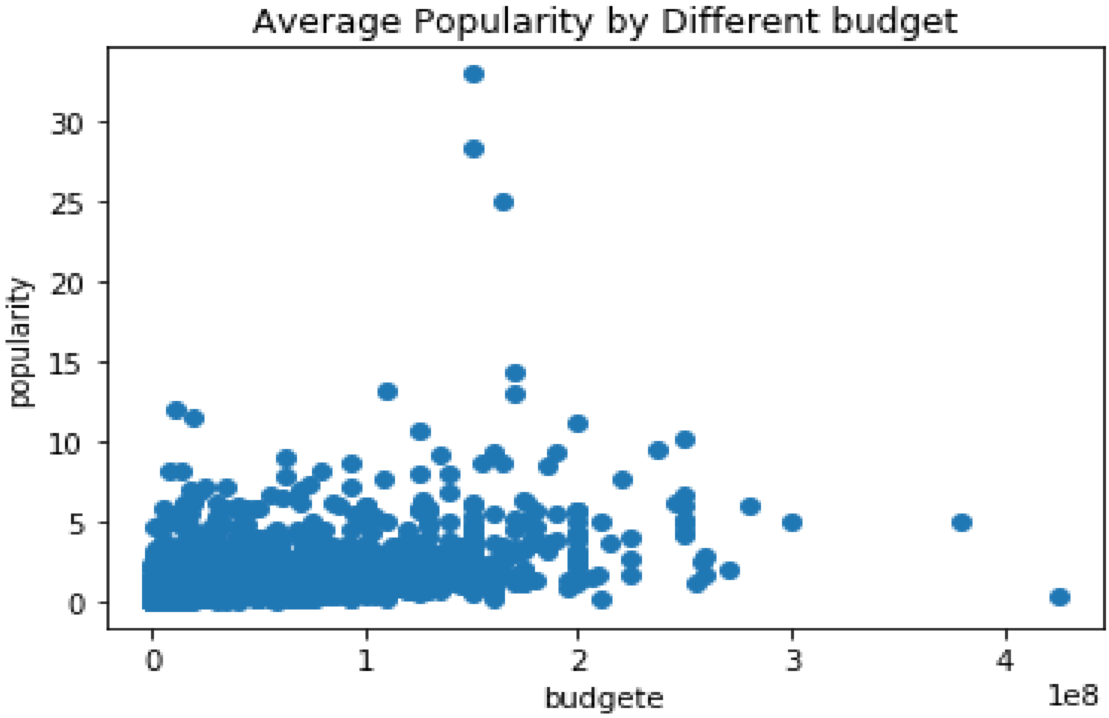
Exploration with Conclusions:

Visualisation of each variable:



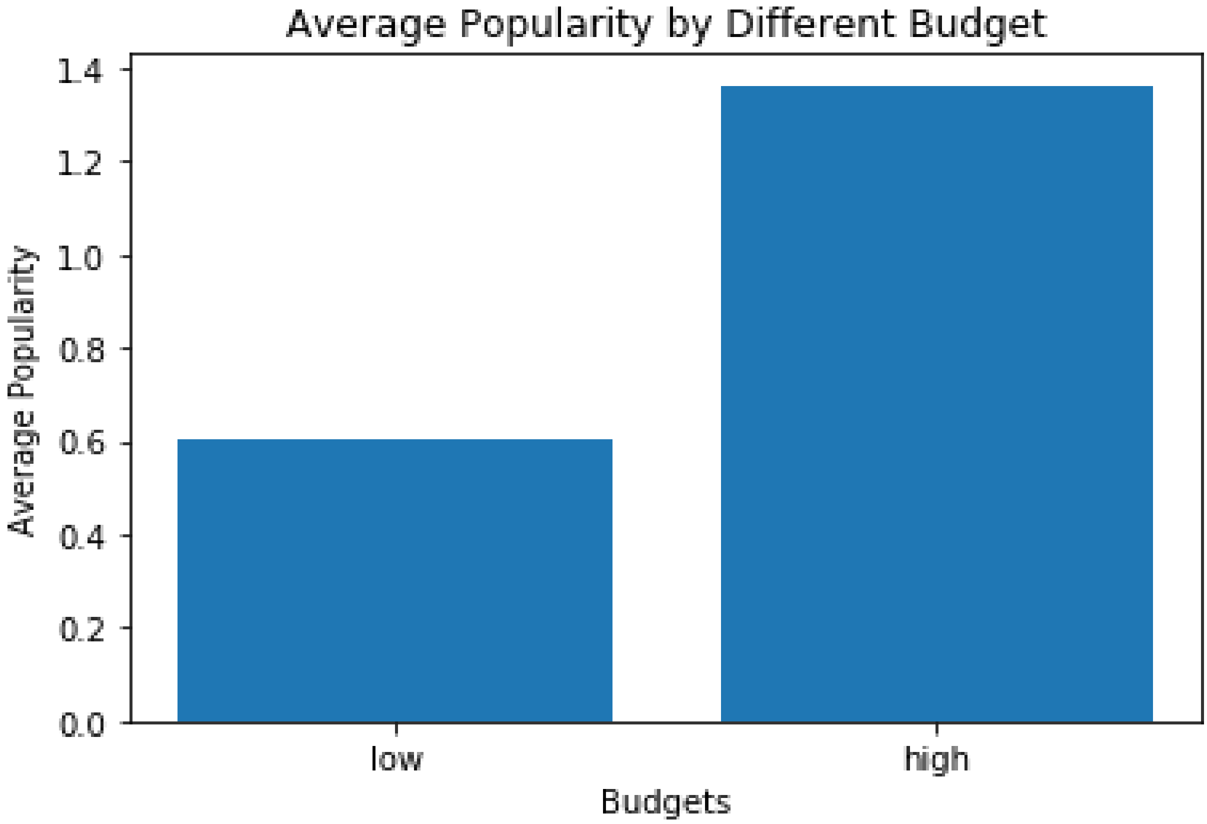
1. Does higher budget mean higher popularity? Is there a coefficient relationship?

Ploting the relation between budget and popularity



We can not see very strong relatioship between the budget and the popularity from above plot. Let's try to compare the data in another way

Creating a bar chart between 'Average Popularity by Different Budget' and 'Average Popularity'

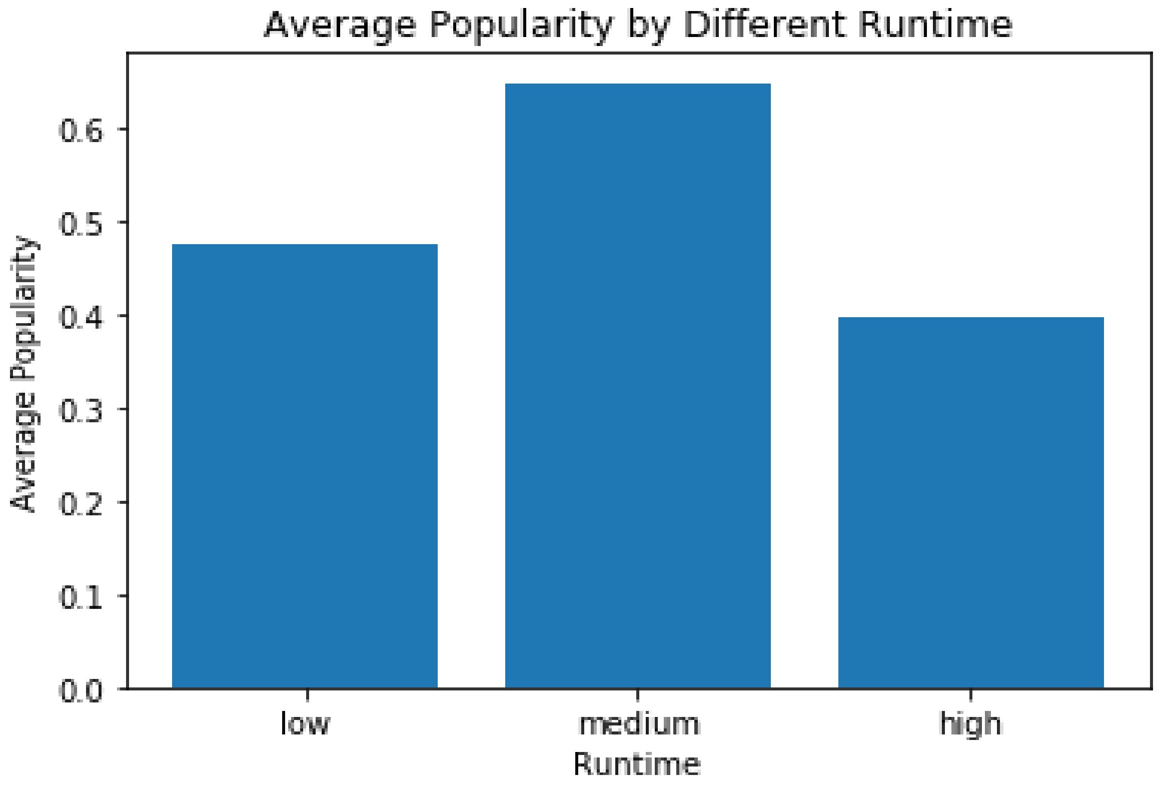


Conlcusion for Q1:

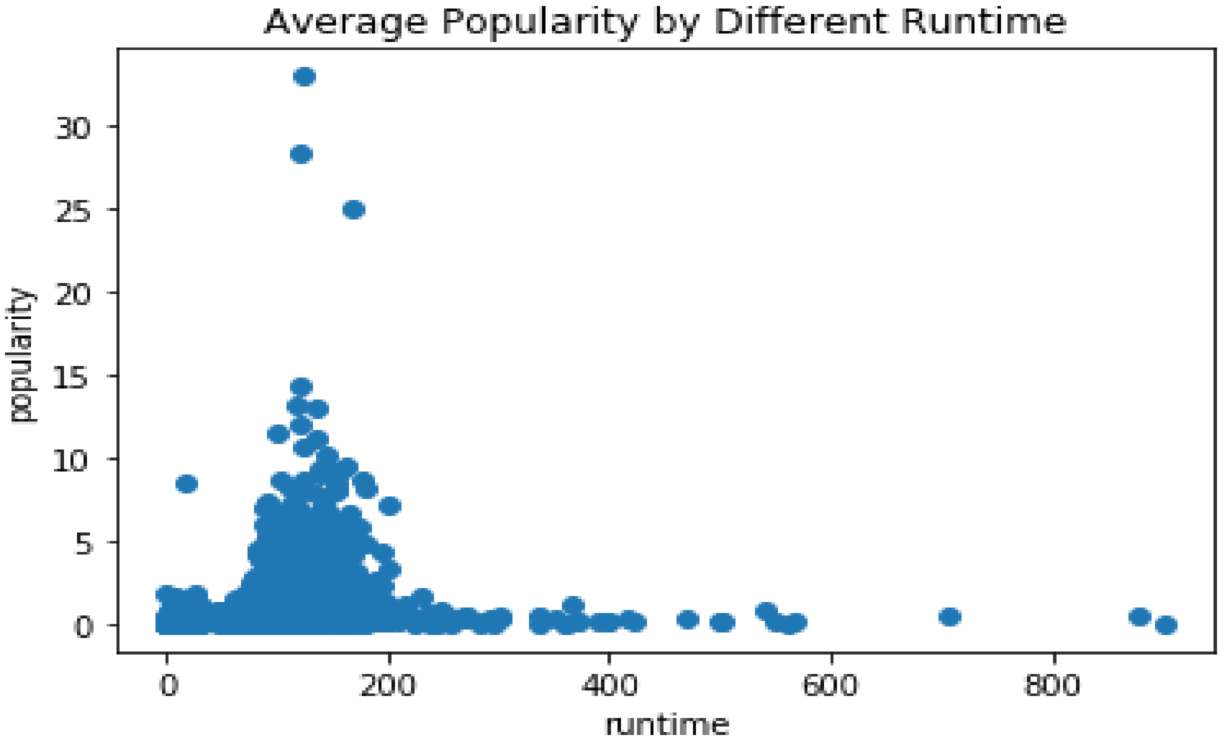
Through this bar plot, we can clearly see that high budget movies are more likely to have higher popularity. We can make this conclusion that higeher budget movies gains more than 50% higher popularity than low budget movies.

2. What length will receive the highest popularity?

Creating a bar chart between 'Average Popularity by Different Runtime' and 'Average Popularity'



Ploting the relationship between runtime and popularity

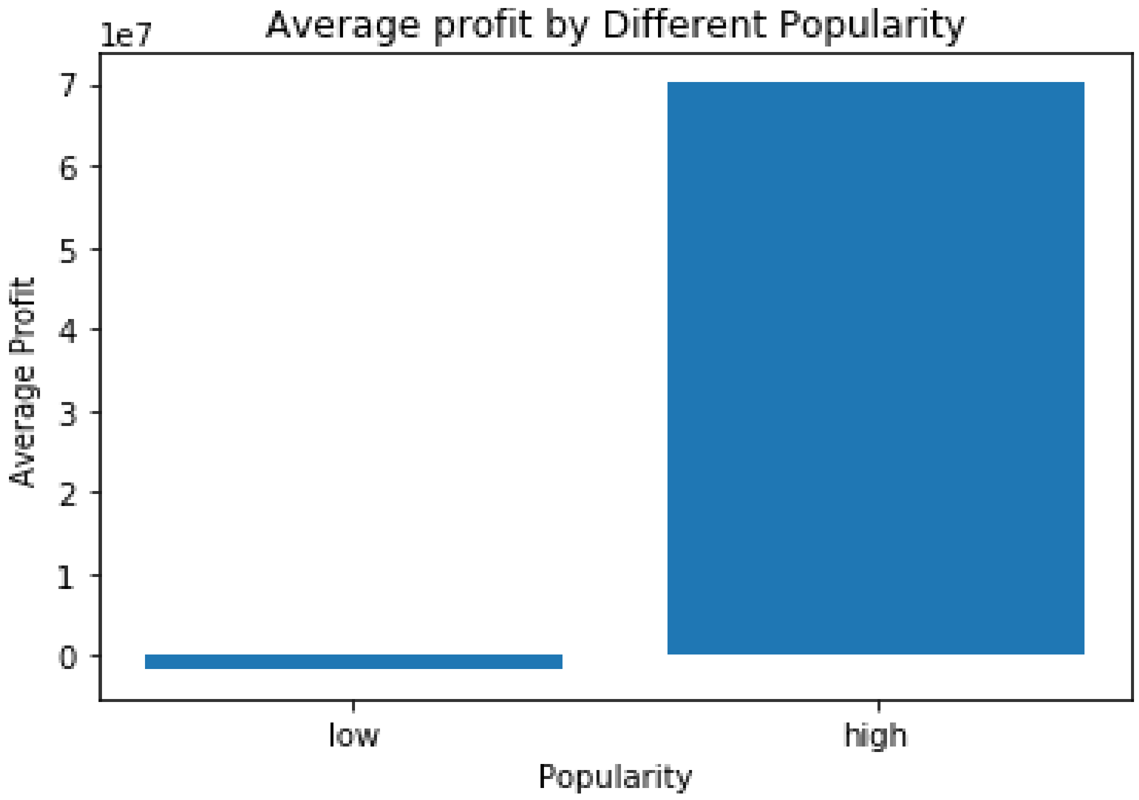


Conclusion for Q2:

Combining two plots above, we cannot simply say , the longer runtime, the more popular the movies are.
If the movies are within 200 minutes, it will be more popular. Once the movies run over 200 minutes, it's hard for them to gain high popularity

3. Higher popularity means higher profits?

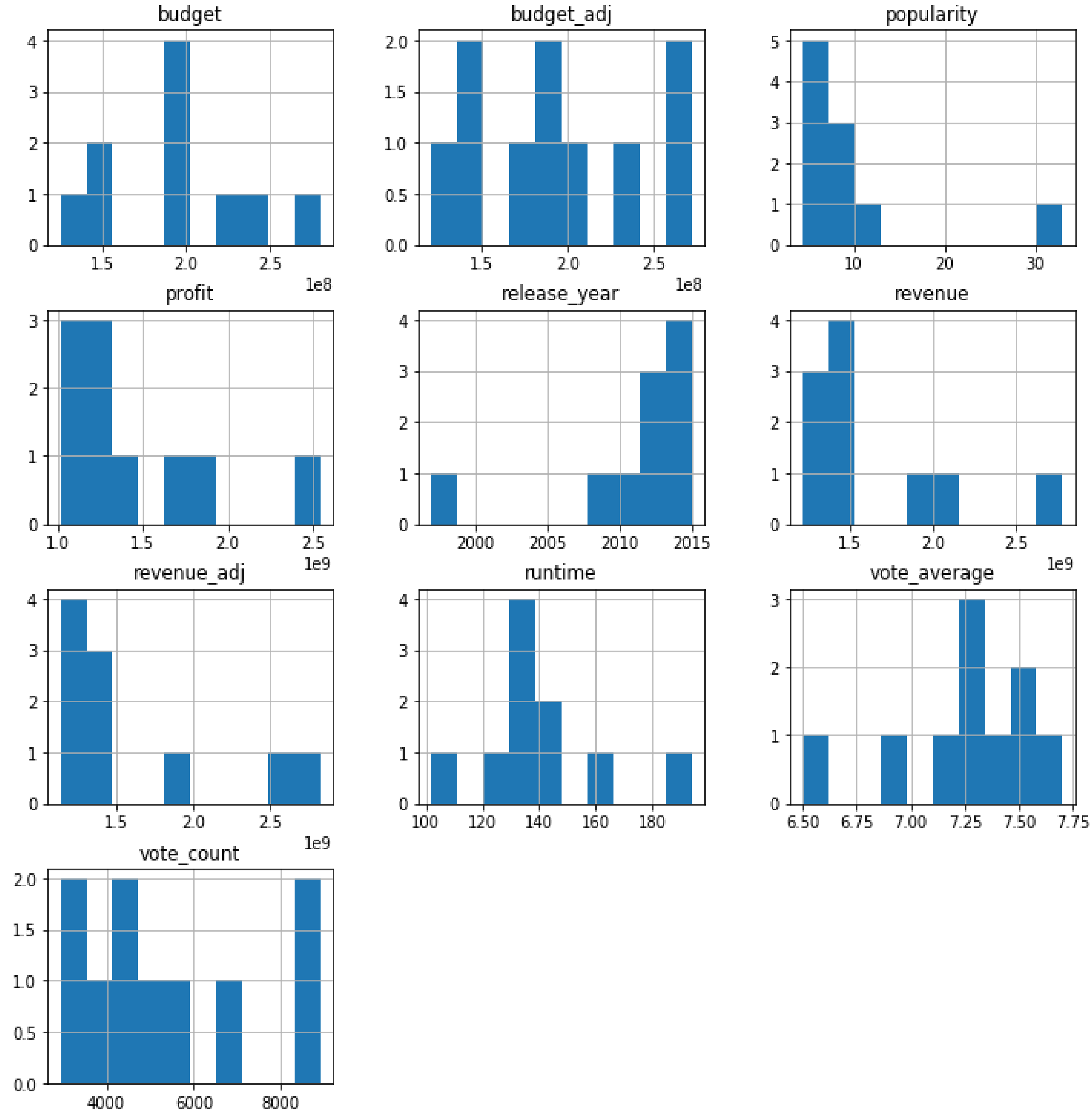
```
Creating a bar chart between 'Average
Profit by Different Popularity' and
'Average Profit'
```



Conclusion for Q3:

As we can see above, higher popularity does make much higher average profits.

4. What Features are Associate with Top 10 Revenue Movies?

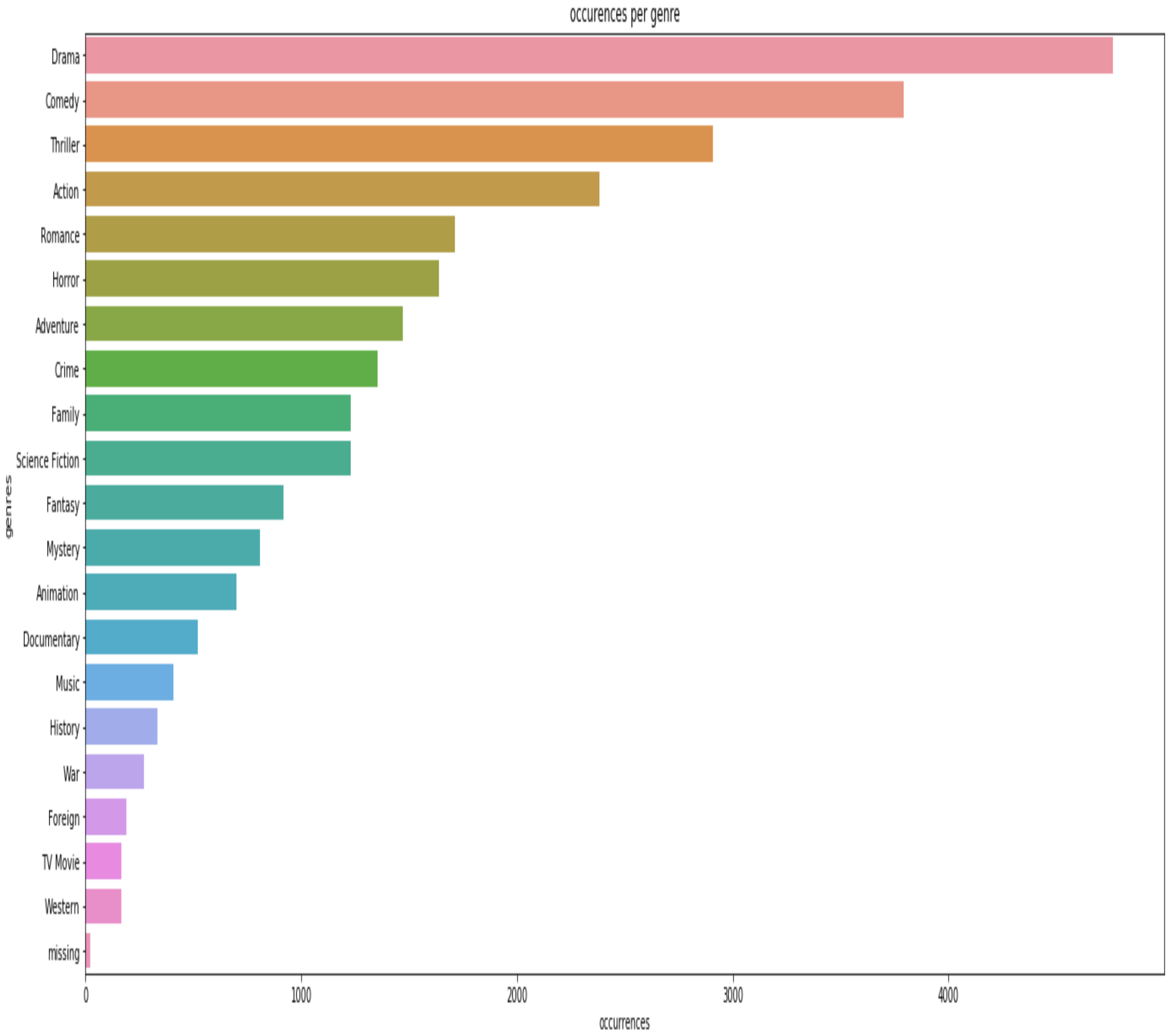


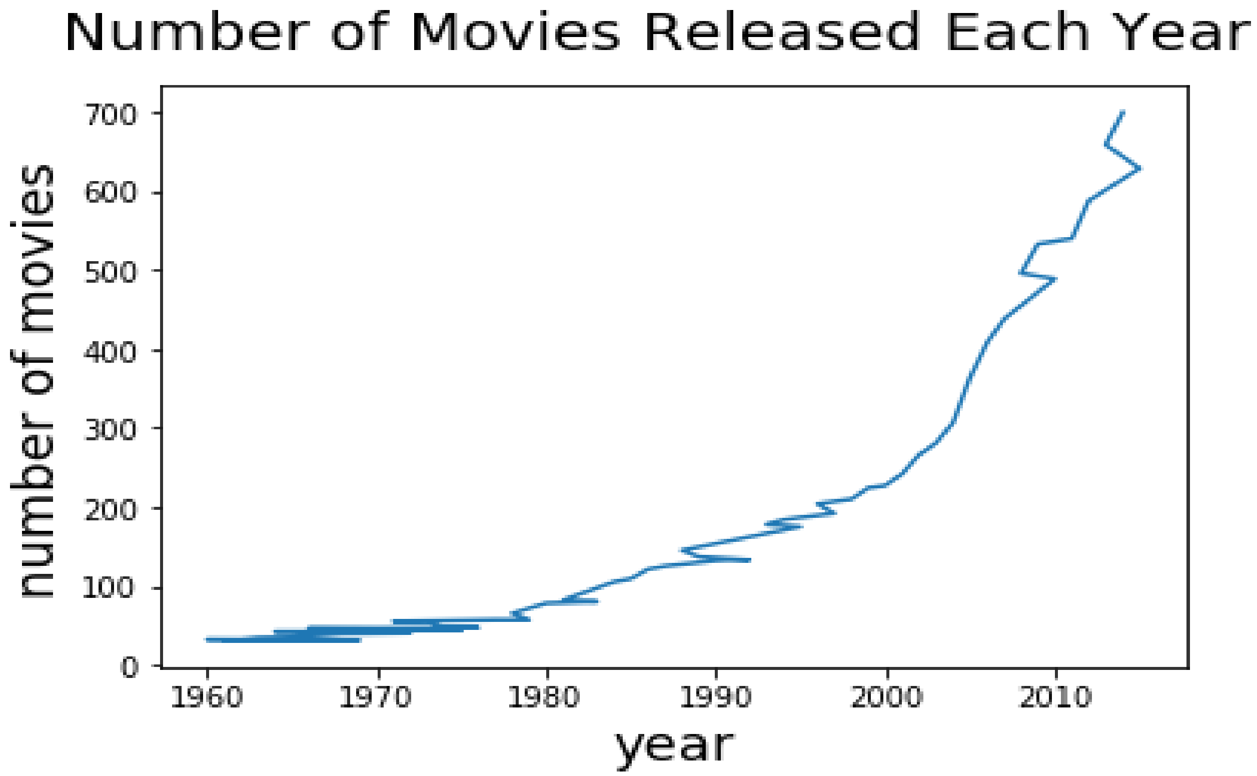
Conclusion for Q4:

There are some characteristics we can conclude from the top 10 movies. Runtime ranges from 100 mins to 200 mins. The released year are between 1995 to 2015

5. Which genres are most popular from year to year?

Splitting the genres and counting the occurrence of each one

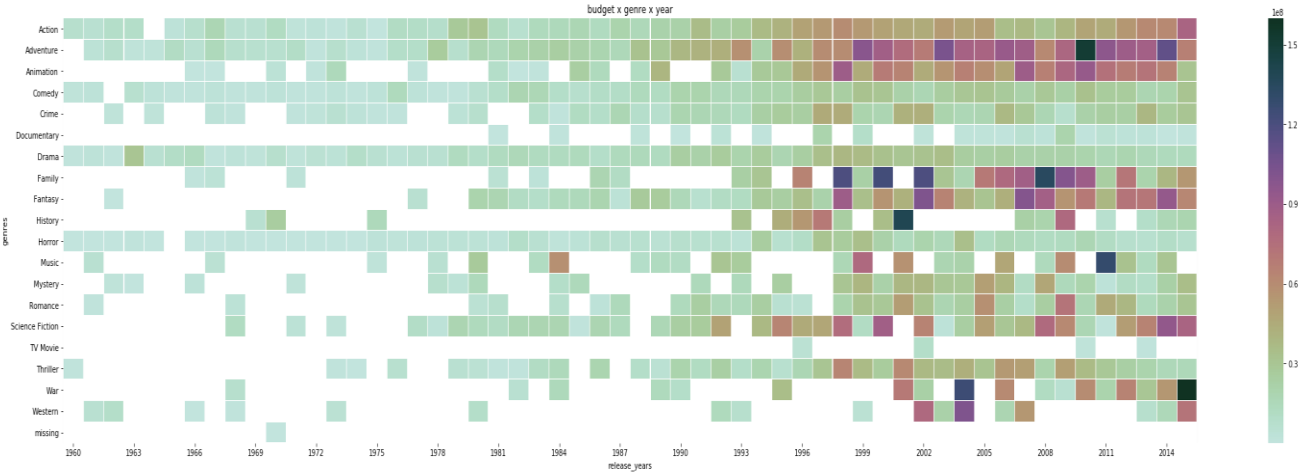


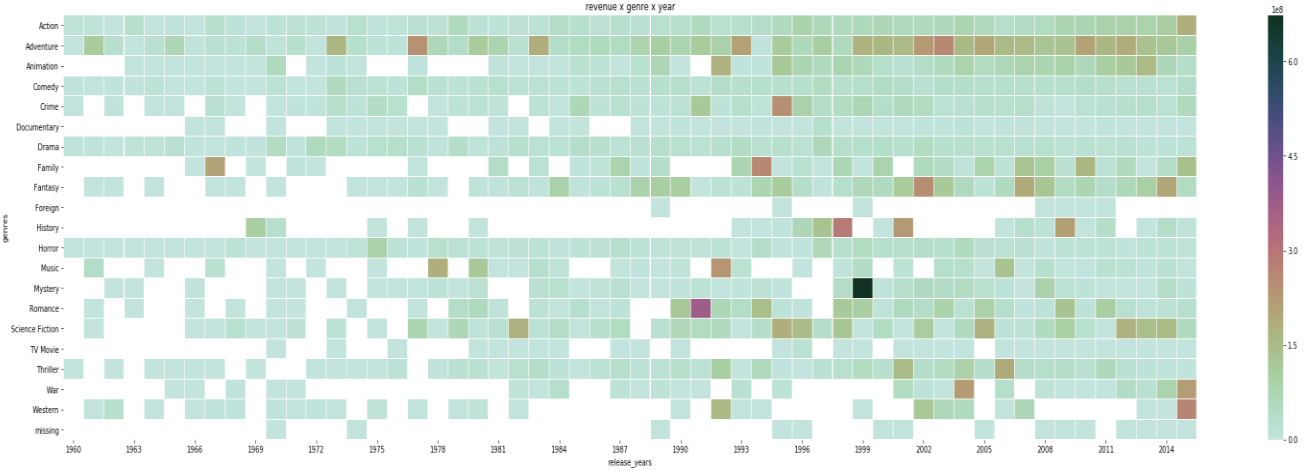


Conclusion for Q5:

From the above two plots, we can see, the top 5 genres are Drama, Comedy, Action, Horror and Adventure. The number of movies increased along the time.

The following plot shows the revenue and budget for each genre type per year genres are so specific, I will just take the first genre for each movie instead of the genre combination





Conclusion:

Based on the analysis I did above, we can make the following summarizations:

1. The quantity and range of movie gets larger. We have more choices to choose from as an audience.

2. We cannot say high budget guarantees high popularity. But for movies with higher budgets do produce higher average popularity.

3. To produce a more popular movie, the runtime should be best around 150 mins; Drama, Comedy, Action, these genres would be preferable.

Limitations:

These are factors that makes the movies become popular and successful. But we should also notice the limitations. There are some missing data and many erroreous zeros which may affect the analysis.

1. It's hard for us to know how the vote_counts and popularity are measured.

2. For foreign movies,currecy is not indicated. inflation over the years should also be taken into consideration.

**Reference:
https://www.kaggle.com/deepak525/investigate-tmdb-movie-dataset/data**