Data Analysis report (movies data from imdb)

By: Fayrouz Yasser

1- Introduction:

This report presents the analysis of movies data collected from IMDB. The data included each movie id, imdb_id, popularity ,budget ,evenue ,original_title ,cast, homepage, director, tagline ,keywords ,overview ,runtime ,genres ,production_companie, release_date ,vote_count ,vote_average ,release_year ,adjusted budget and adjusted revenue after the fininical cricis 2010.

For the questions I am investigating, I've only chosen the related columns that are(id', 'imdb_id', 'popularity', 'budget', 'revenue', 'director', ‘runtime', 'genres', 'release_date', 'vote_average', 'release_year', 'budget_adj', 'revenue_adj)

Notes on data:

- There are a huge number of values that are “zero” which will affect some analyses. Unrequired zero data for the proposed questions have been dropped.
- There are many illogical values in many columns like revenue, budget and runtime (I suggest excluding them in comparison)

2- Questions for analysis:

1- What is the most popular genre over years?

2- How many movies are released each year?

3- What is most common runtime?

4- What is the correlation between revenue and budget?

5-Does the runtime affect popularity?

6- Which months of the year are common in release

3- Data Wrangling:

a- Exploring data: head, shape, info, describe

b- Cleaning data:

1- Check NaN values:

>>*Decision:* drop all rows with them.

2- Check for Duplicated data:

>> *Decision:* drop unwanted columns.

3- Check for data dtype and fix if found, split strings:

>> *Decision:* #Convert release_date from string to datetime

#Extract month to new column

#Split different values of genres > Expand rows > Reset index

#Split different values of director > Expand rows > Reset index

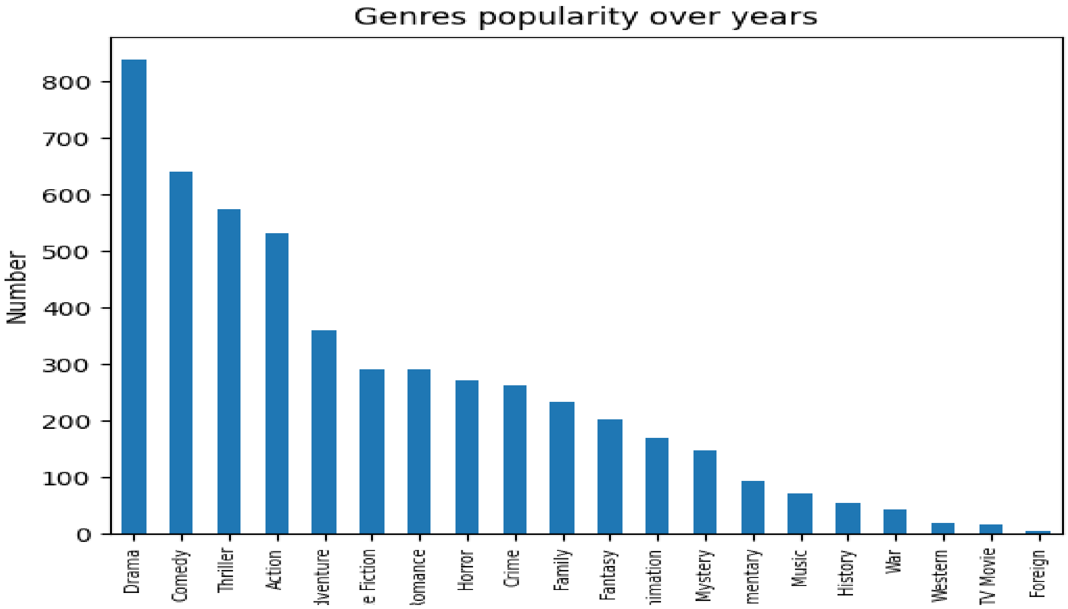
#check zero values > True > replace with NaN values to avoid miscalculation

4- Save cleaned data to new dataframe and recall new dataframe to (df_movies)

4- Exploratory data analysis:

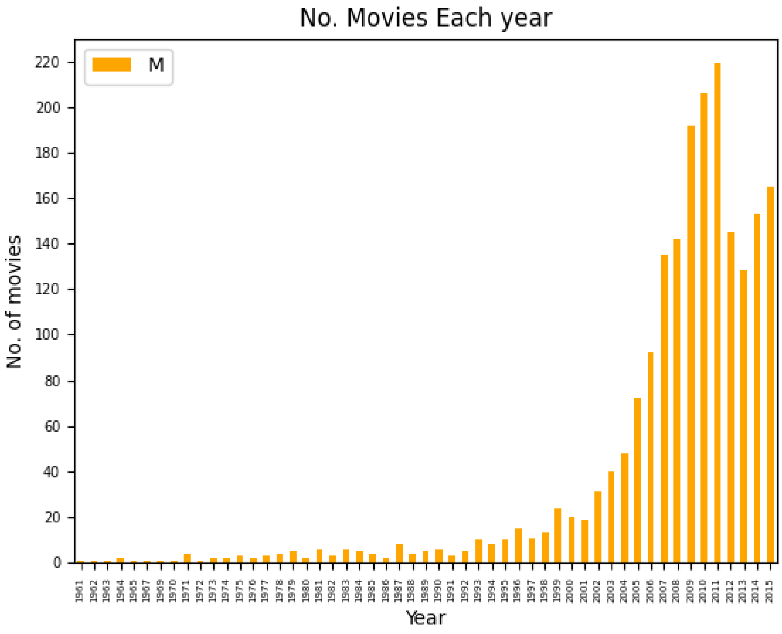
1- What is the most released genre over years?

Drama is the most released genre over years.



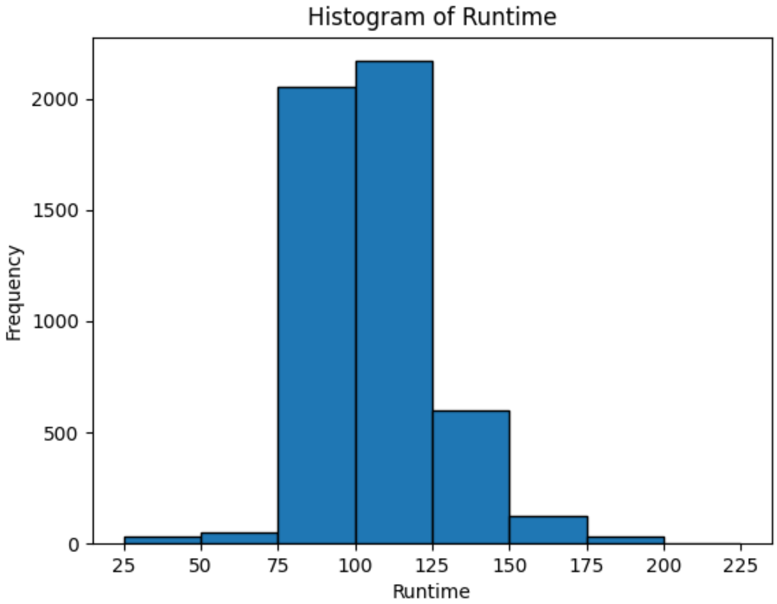
2- How movies production changed each year?

Movies production increased dramatically starting from 2000 and reached its peak in 2011 then declined to 2013 and rose again till 2015.



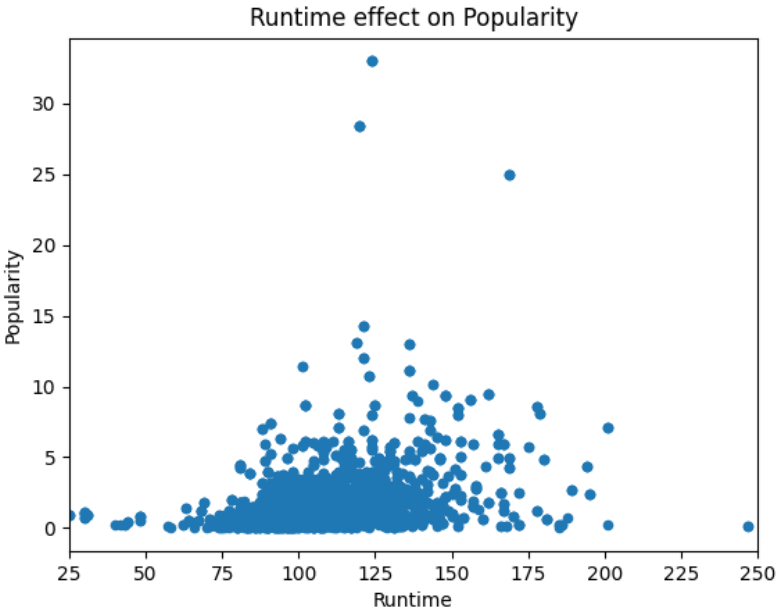
3- What is most common runtime?

Movies runtime between 100:125 min. was the most common.



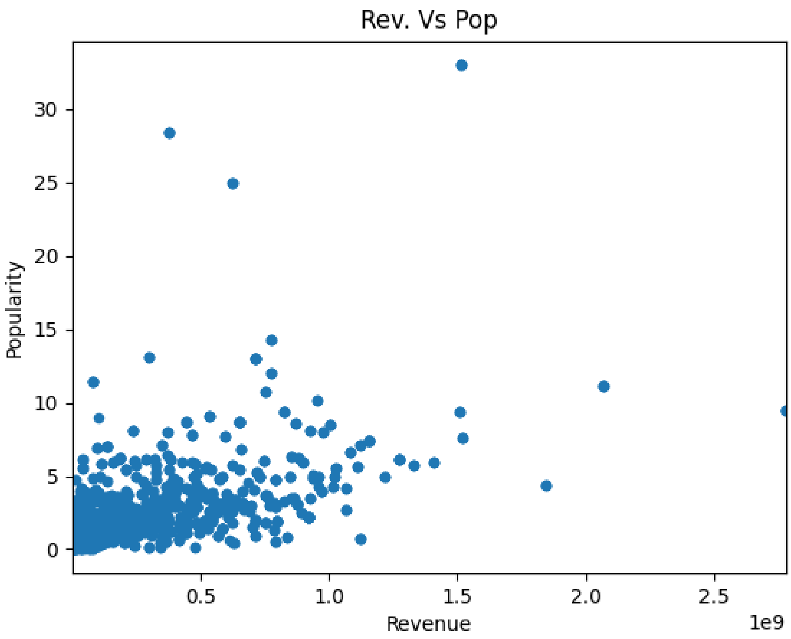
4-Does the runtime affect popularity?

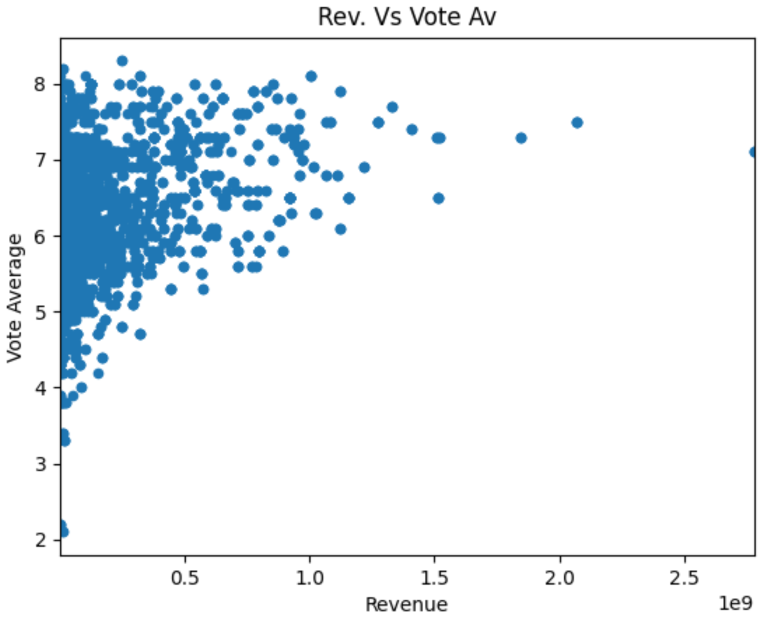
No apparent correlation can be seen between the runtime and popularity.



5- What is the correlation between revenue and (popularity/ vote average)?

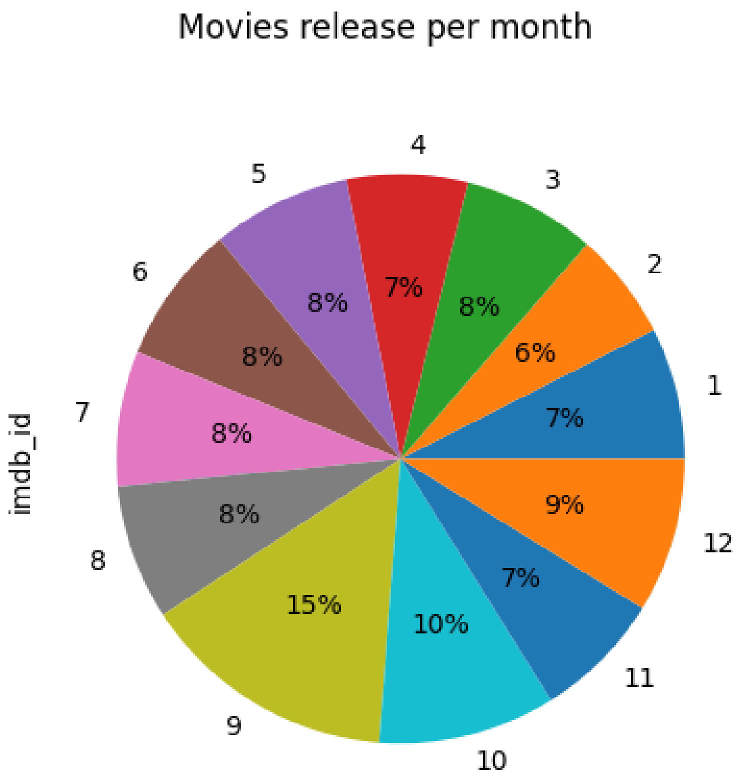
No apparent correlation can be seen between revenue and (popularity/ vote average)





6- Which months of the year are common in release?

September is the most popular month of the year for movies release, followed by October. The lowest number of movies are released in February.



5- Conclusions:

Upon exploration of the data, the main proposed questions were answered, and some limitations were addressed:

Main findings:

- The most released genre over years from1960 to 2015 was drama.
- Movies production increased dramatically starting from 2000 and reached its peak in 2011 then declined to 2013 and rose again till 2015.

- Movies runtime between 100:125 min. was the most common, but excluding the illogical values of runtime, e.g., 5 minutes and 750 min.
- No apparent correlation can be seen between the runtime and popularity.
- No apparent correlation can be seen between revenue and (popularity/ vote average), considering that there are very low and high (unrealistic) values of revenue.
- September is the most popular month of the year for movies release, followed by October. The lowest number of movies are released in February.

Limitations:

- The high values count of directors did not allow me to visualize each director number of movies.
- Other values count did not allow for clear visualization or correlations in addition to the various outliners that I could not decide which to exclude and on which criteria.
- The bandwidth of several data like popularity did not allow for perfect visualization.