Dataset analyzed: The TMDb movie data which can be found in this link, https://www.google.com/url?q=https://d17h27t6h515a5.cloudfront.net/topher/2017/October/59dd1c4c_tmdb-movies/tmdb-movies.csv&sa=D&ust=1532469042115000

Statement of questions posed

The following questions are poised:

1. Do movies which cost a ton gather high popularity?
2. What genres have the most popularity?
3. What kinds of properties are associated with movies that have high profit?
4. What's the relationship between vote average and popularity?
5. Does popularity influence the revenue generated?
6. Does runtime influence popularity?
7. Has it become more expensive/cheaper to produce movies?
8. What is the average popularity by month? Do movies release in some months have higher popularity score than others?

Description of investigation

The dataset was initially assessed for potential dirtiness and was found to be dirty. A few columns had missing data and some columns were not required for the investigation. For example, the documentation indicated that the revenue and budget had been re-computed as revenue_adj and budget_adj. Therefore, cleaning was done to remove unnecessary columns, null values & duplicates, and some new columns were created for easy analysis as well. Upon further inspection, the columns containing adjusted revenue and budget were found to contain several zero values. Documentation found on Kaggle revealed that these values were imputed as zeros because the details were not available. Therefore, it was decided to drop these values to make for reasonable analysis. Codes were then written to answer the questions posed. A variety of plots ranging from histograms, scatter plots, line graph, bar chart and pie chart were used for data visualizations.

Findings

1. **Distribution of movies by genres and year**

It was found that more movies have been released in recent years. There has been about 2000 % increase in the number of movies released between 1960 and 2015. This is shown in the histogram distribution of movies over the years. Also, the most popular genre seems to be Drama, closely followed by comedy.

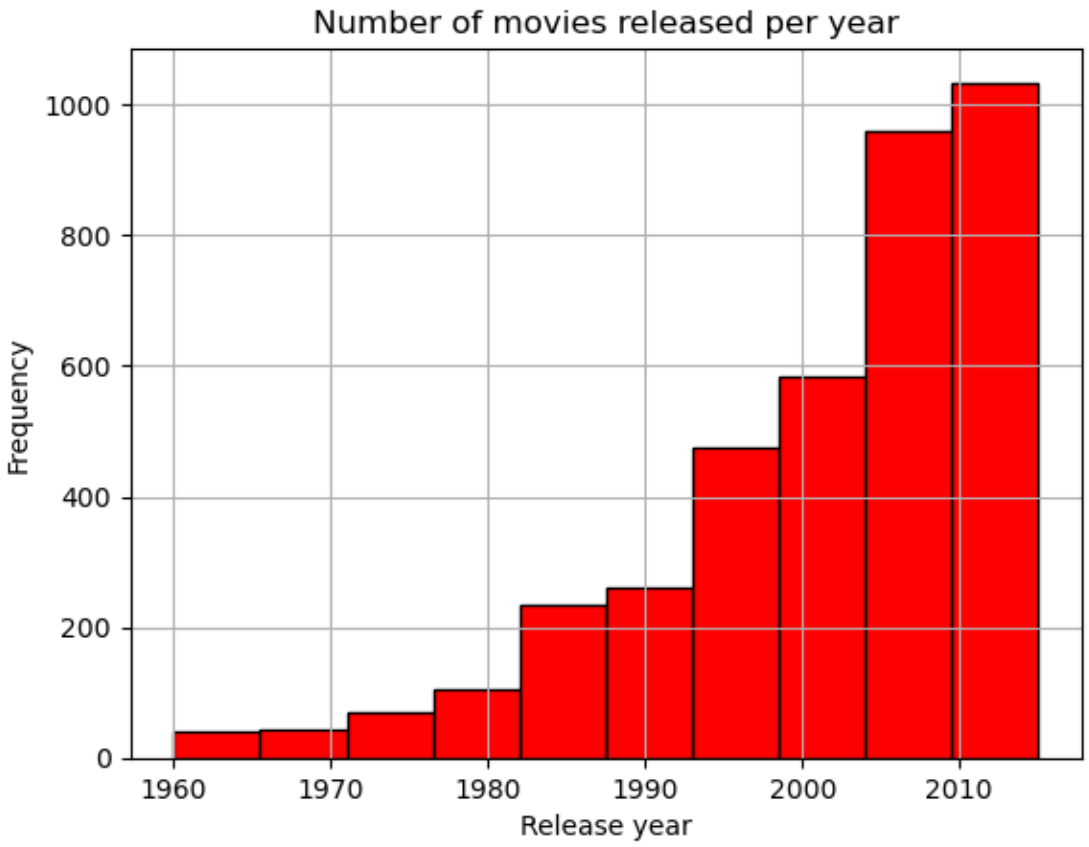


Figure 1a: Distribution of movies by year

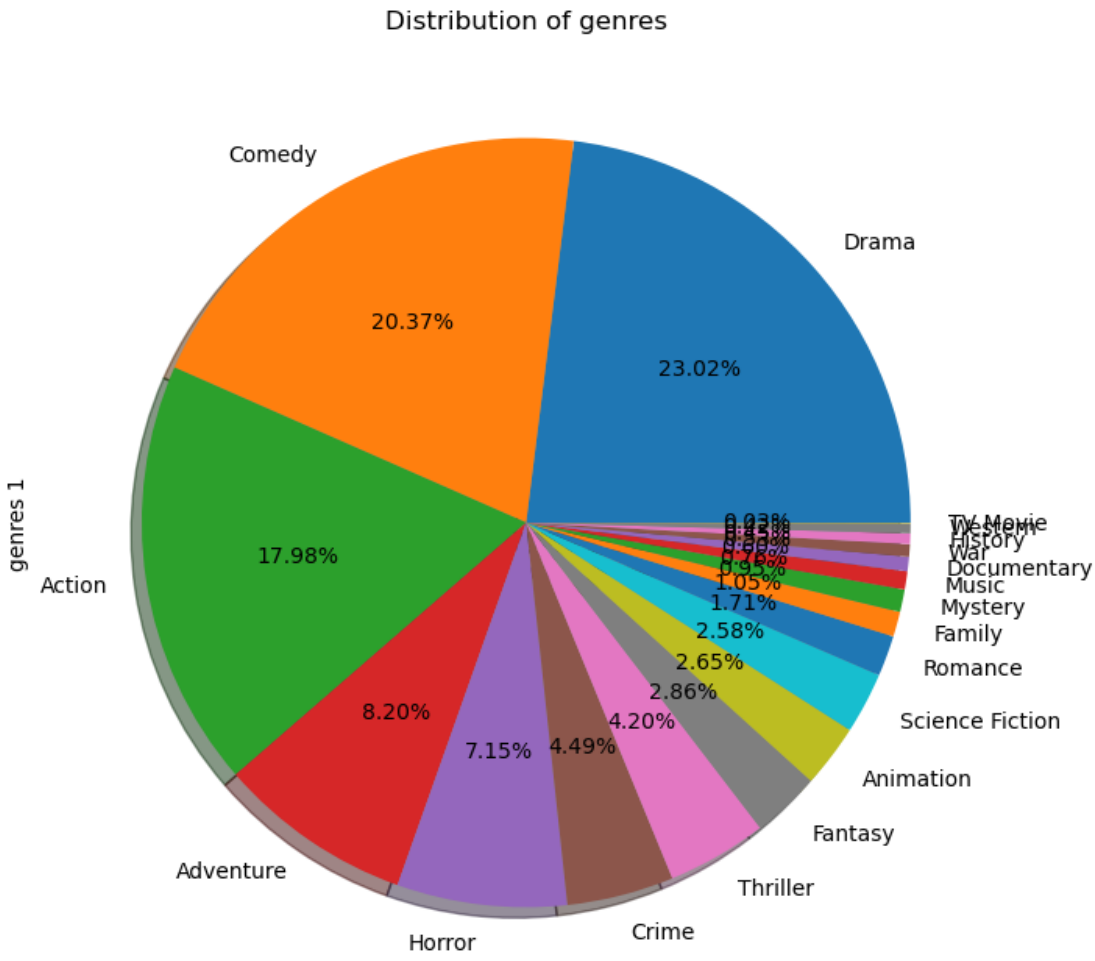


Figure 1b: Genre distribution

2. Popularity of movies by month

No astonishing information was found by comparing the popularity of the movies released across the 12 calendar months. All movies released across the months seemed to receive nearly same popularity.

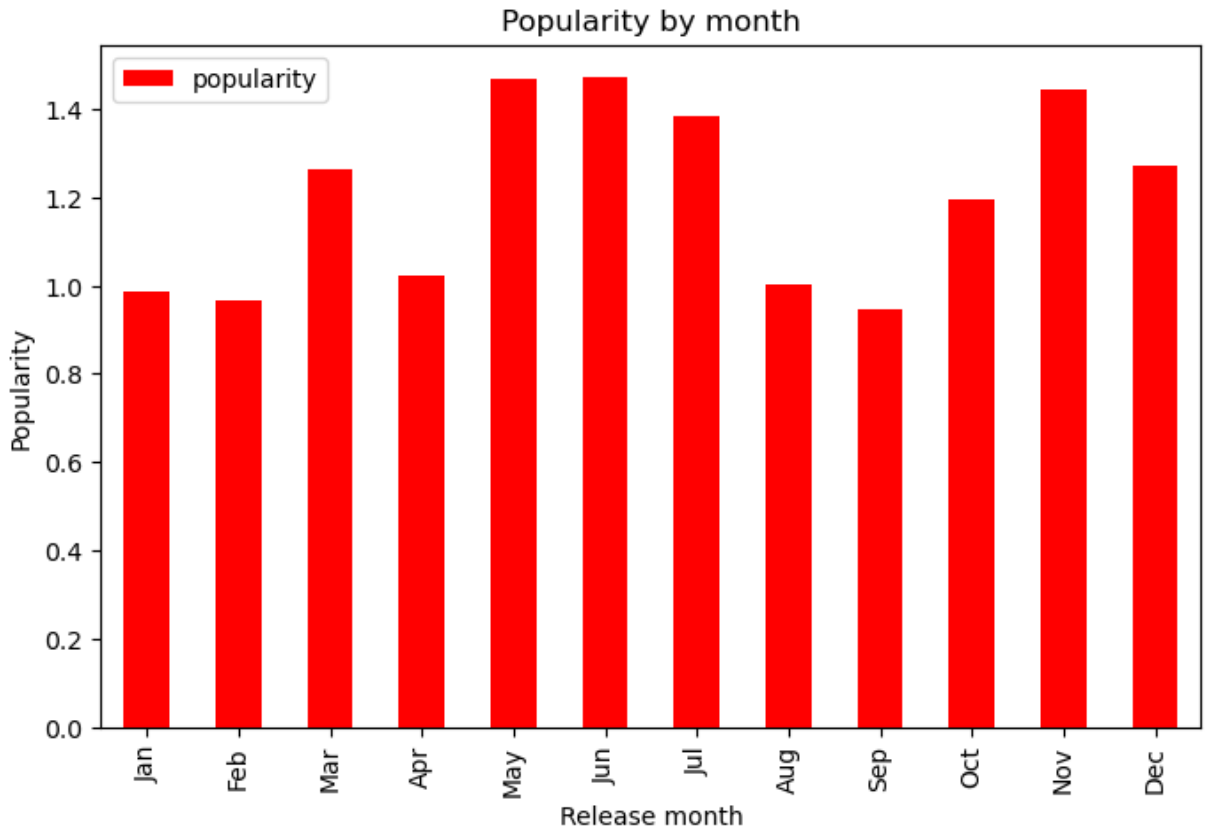


Figure 2: Plot of popularity by month

3. Cost of movie production

Visualization of the trend of mean budget spent on movies per year reveal that it has not become particularly too expensive to make movies. In fact, movies released between 1960-1970 featured a very high production cost compared to those made between 2000-2015.

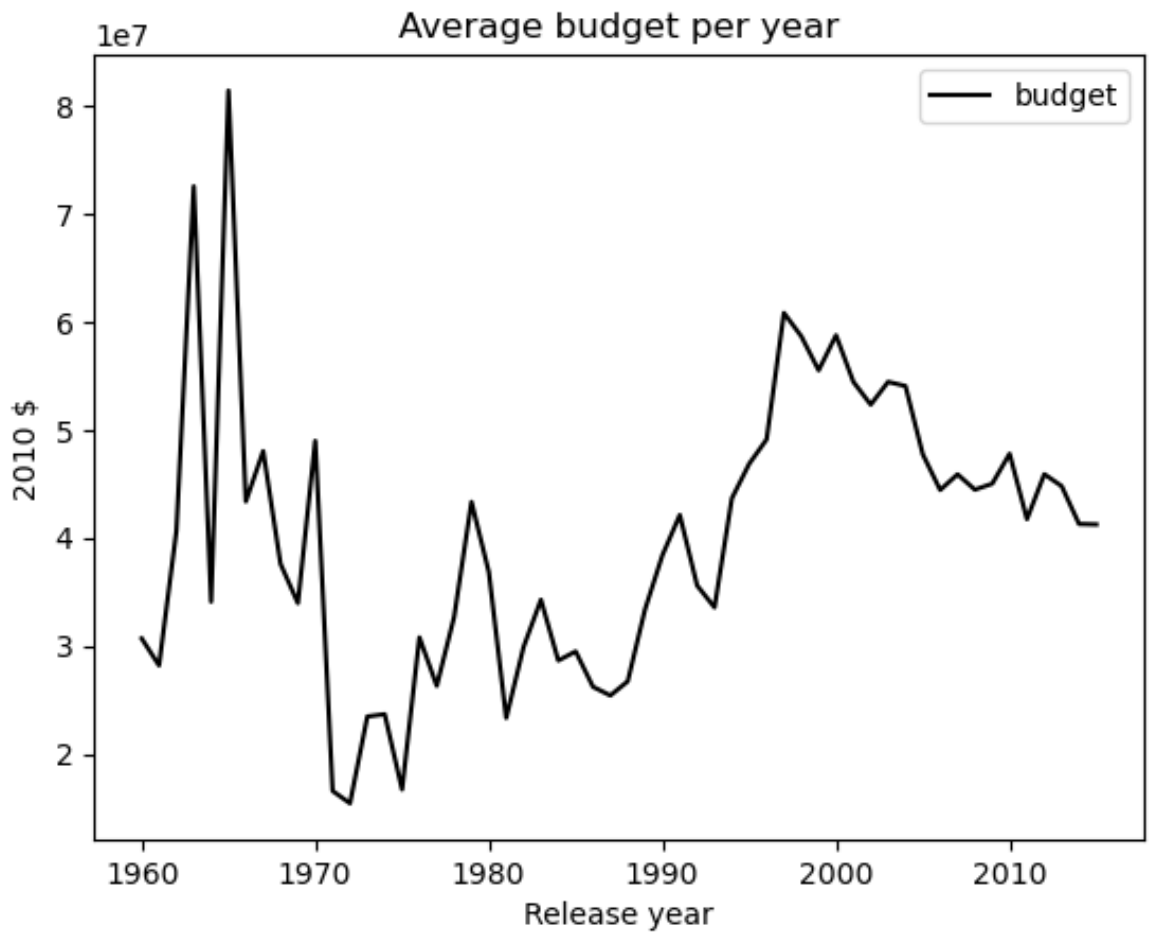


Figure 3: Line plot of average production cost (in 10 million dollars) per year

4. Vote average versus profit

A plot of the mean profit realized per vote average shows that high voters rating correspond to high profit. Movies rated >=8.0 had the highest profit margin. Not surprisingly, the movies with vote average of >=8.0 also had the most vote count. This may mean that these movies had more viewers and that the viewers found the movies to be good across board. Also, more viewers may mean more profit.

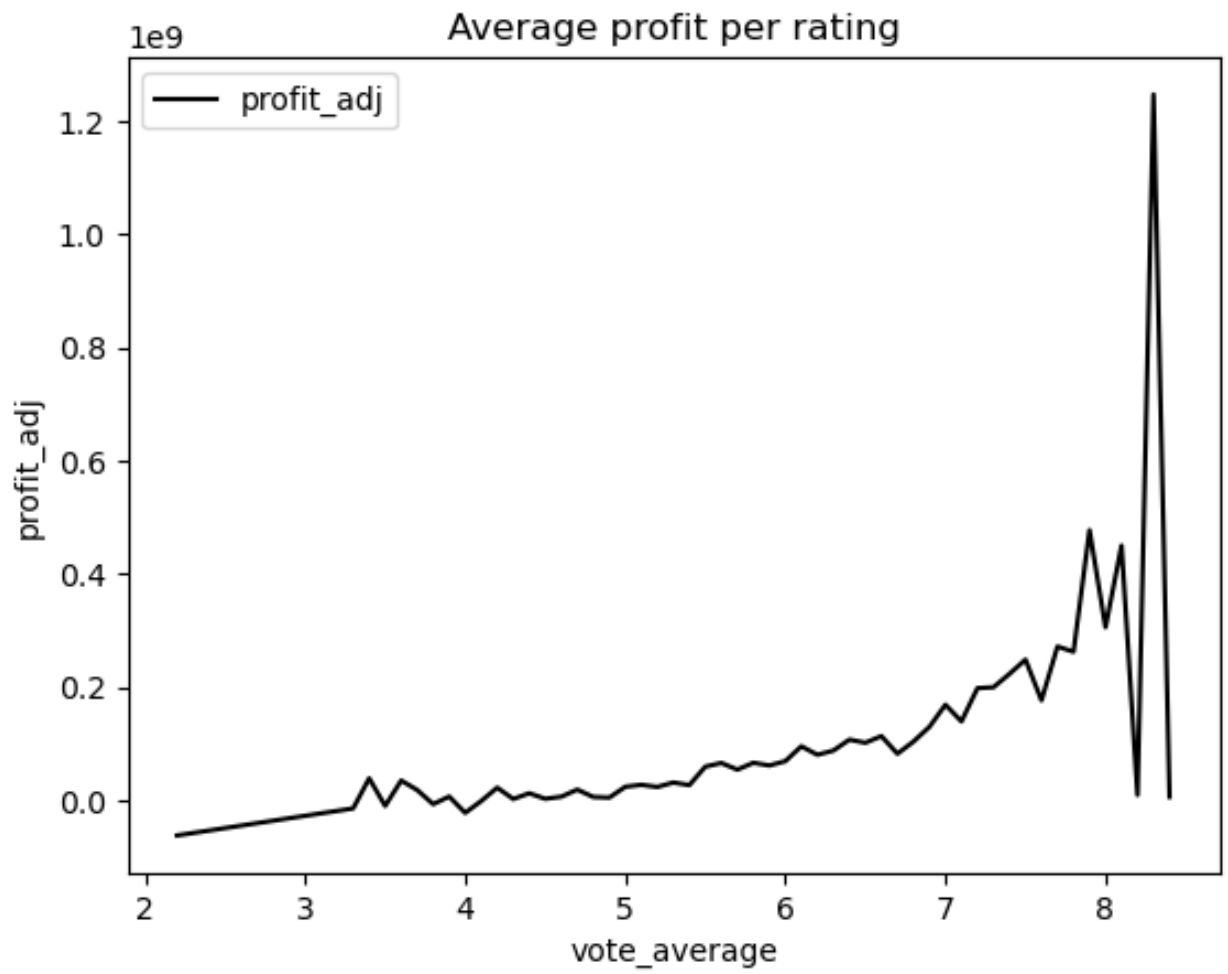


Figure 4a: Line plot of profit against average vote ratings

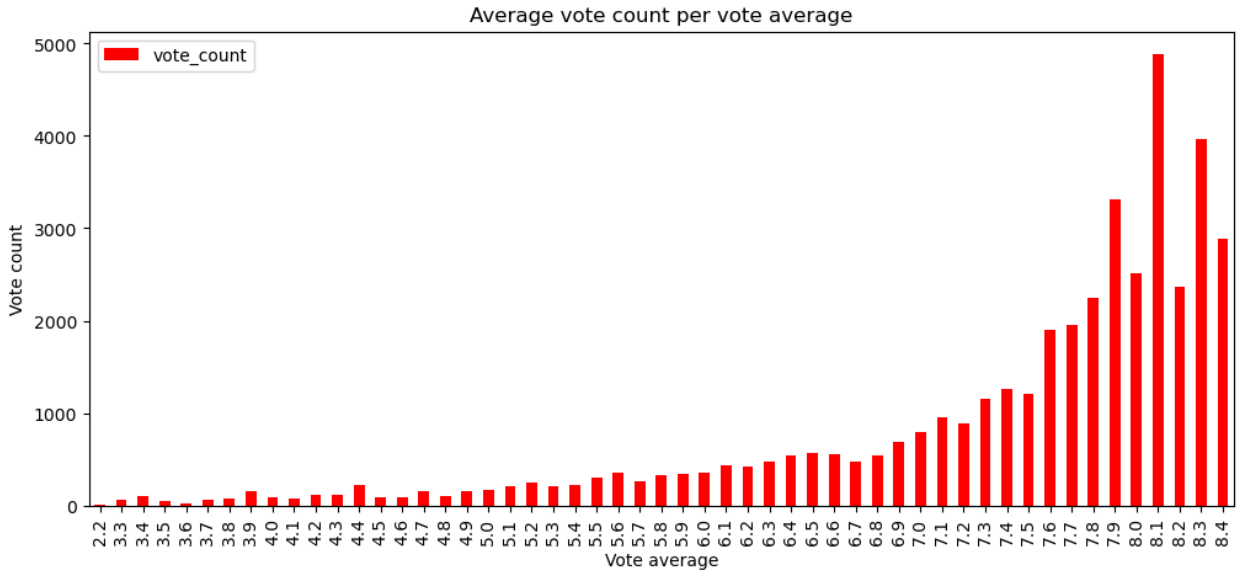
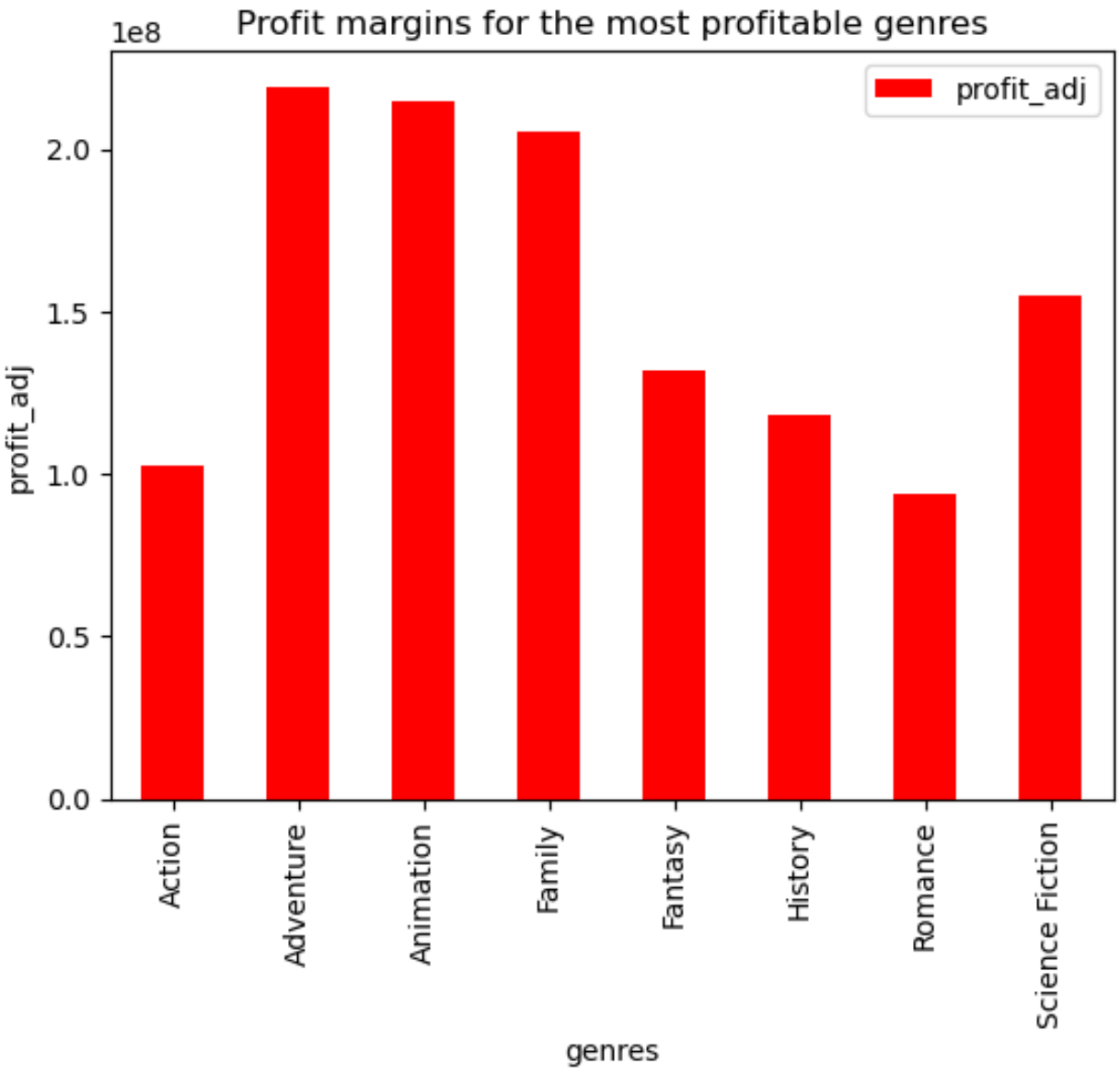


Figure 4b: Bar plot of average vote count per average vote ratings

5. Profit versus genres

The genres associated with the highest profits are action, adventure, animation, family, fantasy, history, romance, and science fiction. The mean profit margins for these genres is >= the 75th percentile of the profits generated per movie. These genres have similar ratings and runtime.



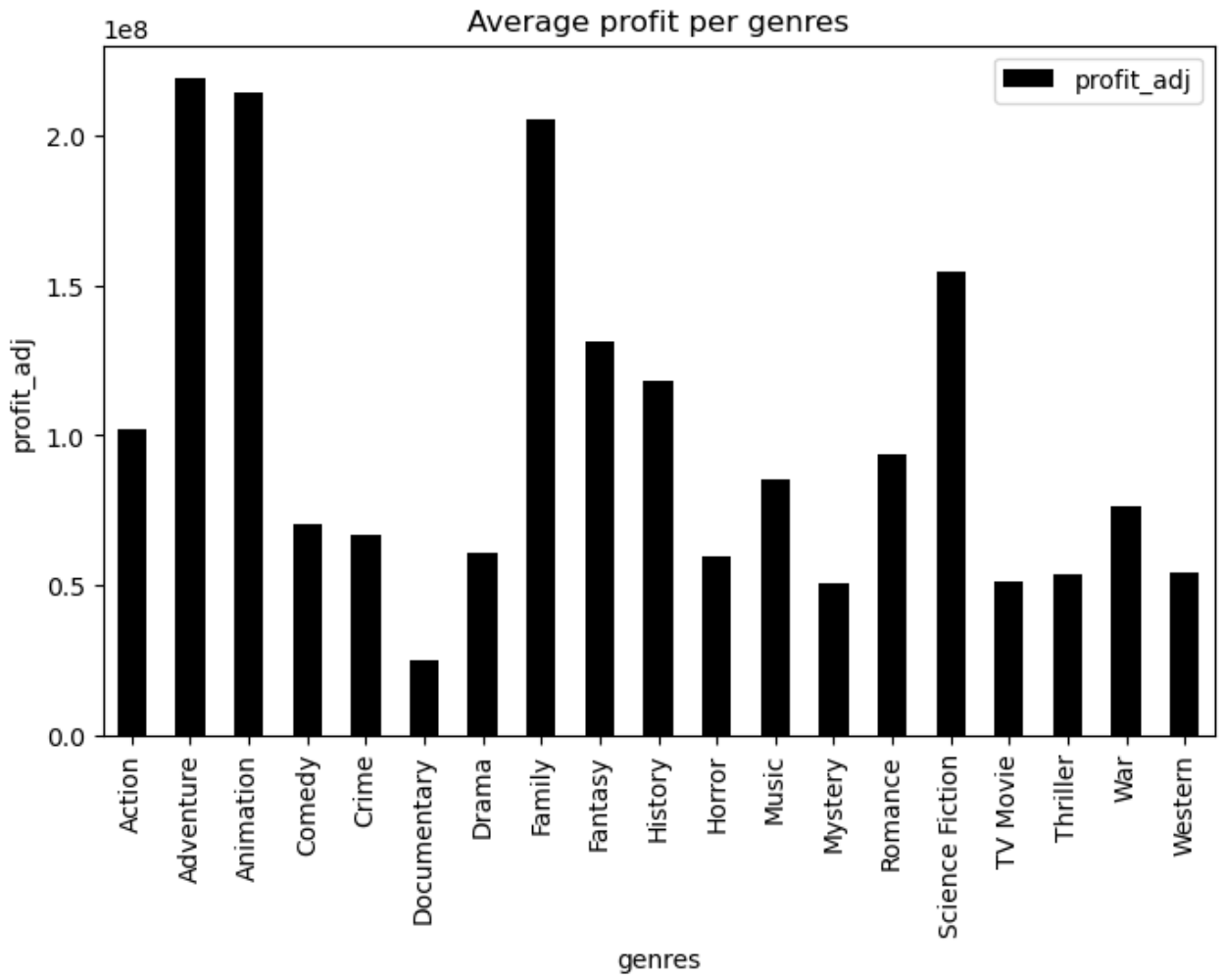
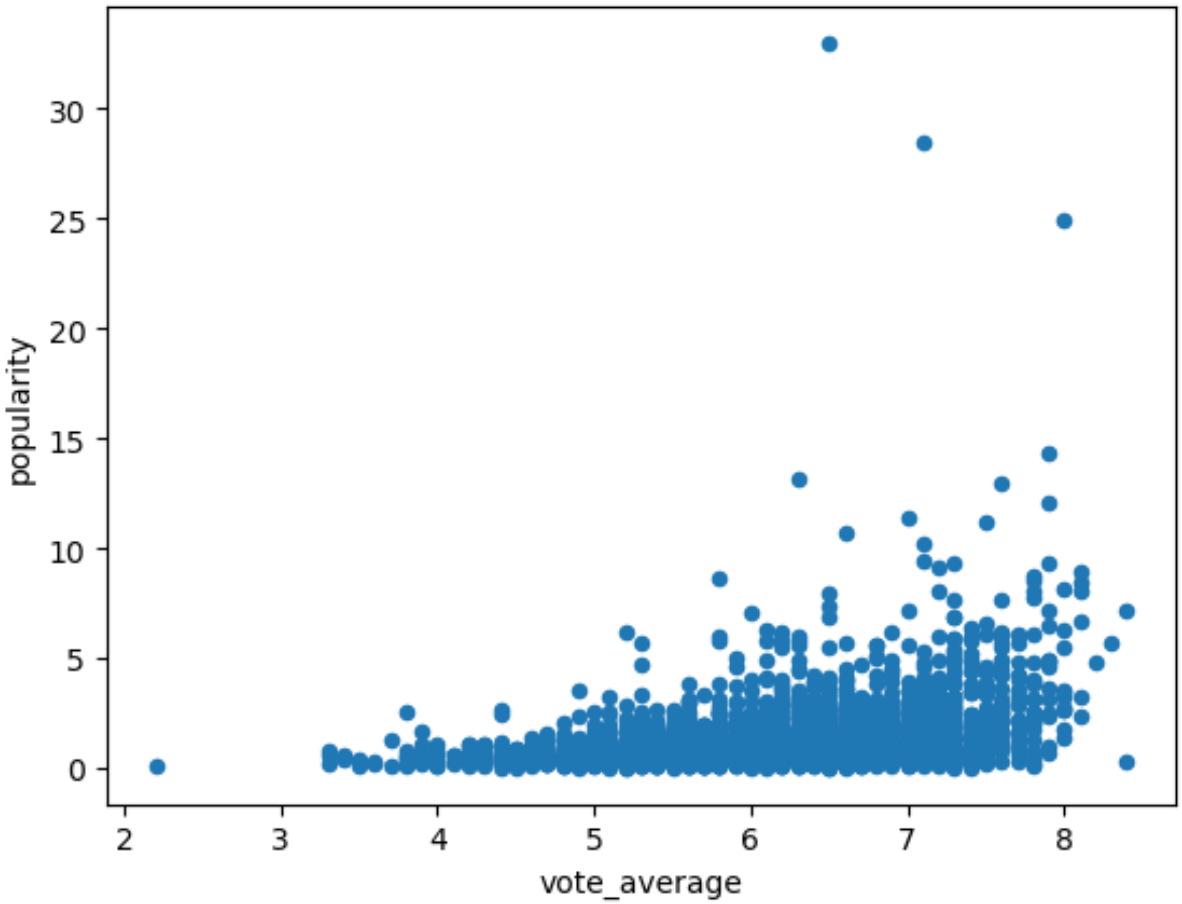
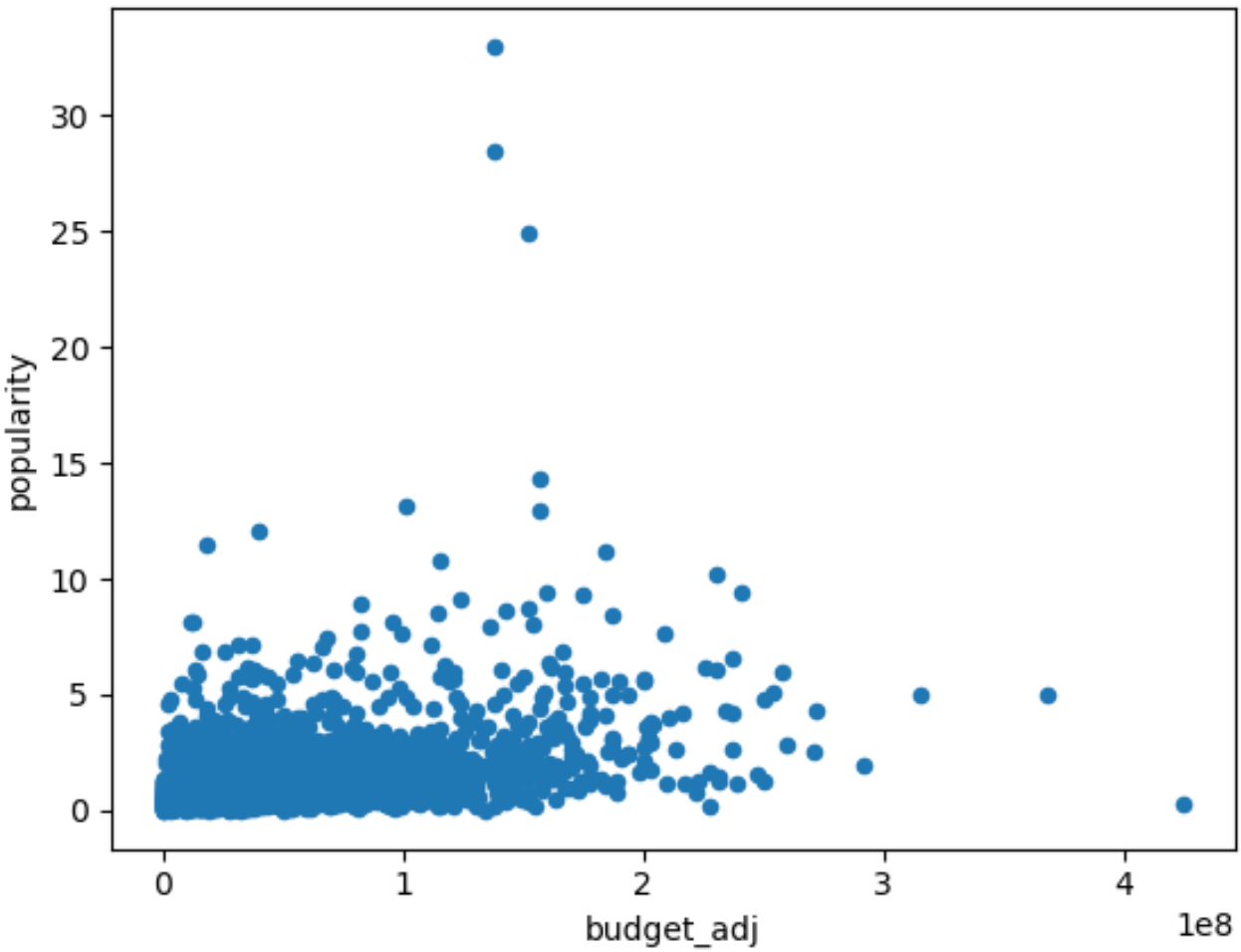


Figure 5: Plots of average profit (in 100 million dollars) made by genres.

6. Correlation of popularity with different element

There are no strong conclusions about the correlation of popularity with varying elements such as revenue, budget, and vote average.


popularity
vote_average
0
5
10
15
20
25
30
2
3
4
5
6
7
8


popularity
30
25
20
15
10
5
0
0
1
2
3
4
budget_adj
1e8

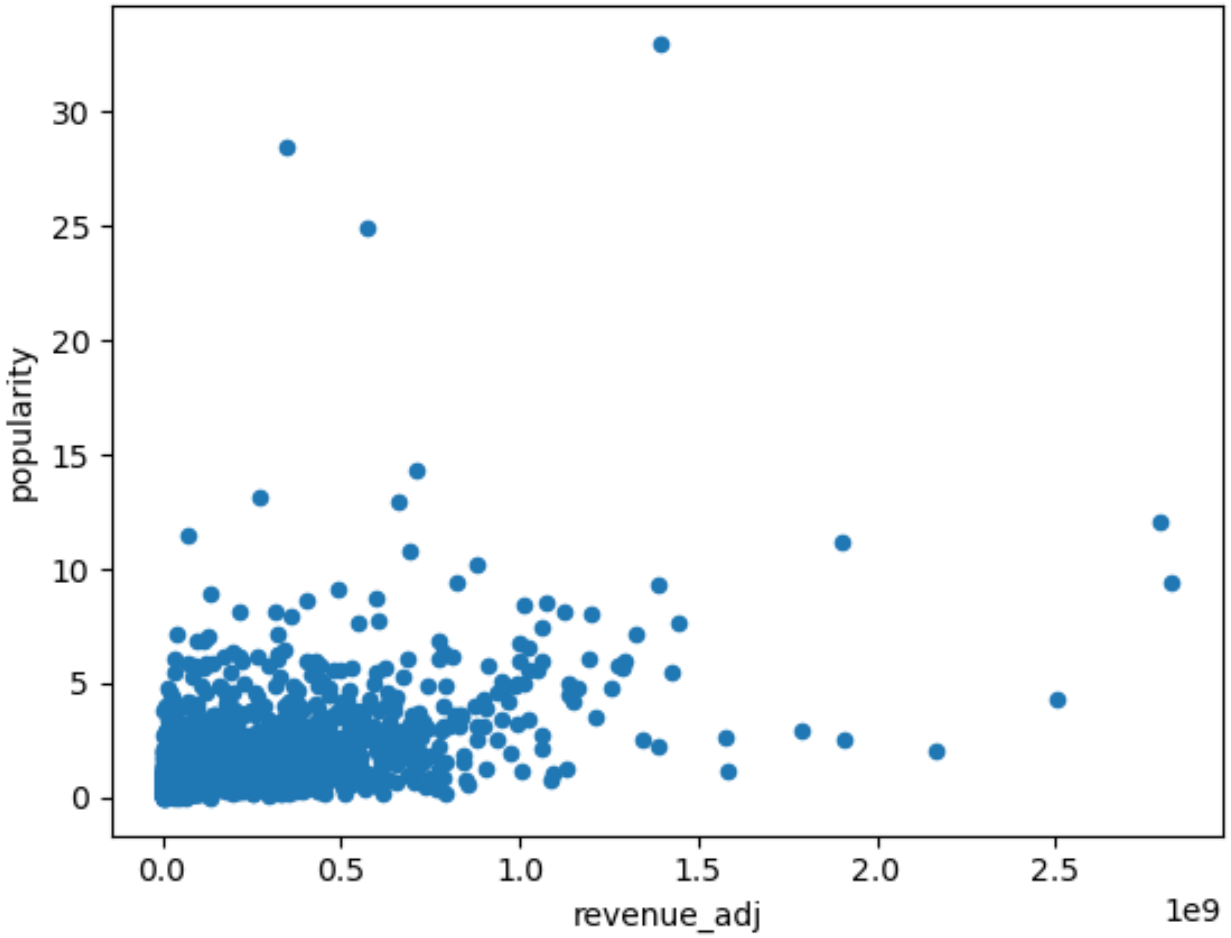


Figure 6: Scatter plots of popularity against vote average (top), budget (middle) and revenue (bottom).

These conclusions are subject to statistical tests and may not necessarily reflect the most accurate information that can be generated from this dataset. Some biases may have been introduced in the analysis especially in the case of the genres column which have different values separated by pipes. Only the genre with complete column values (genres 1) was used in the analysis and this may bias the results.