For the project, I investigated an IMDB movie set.

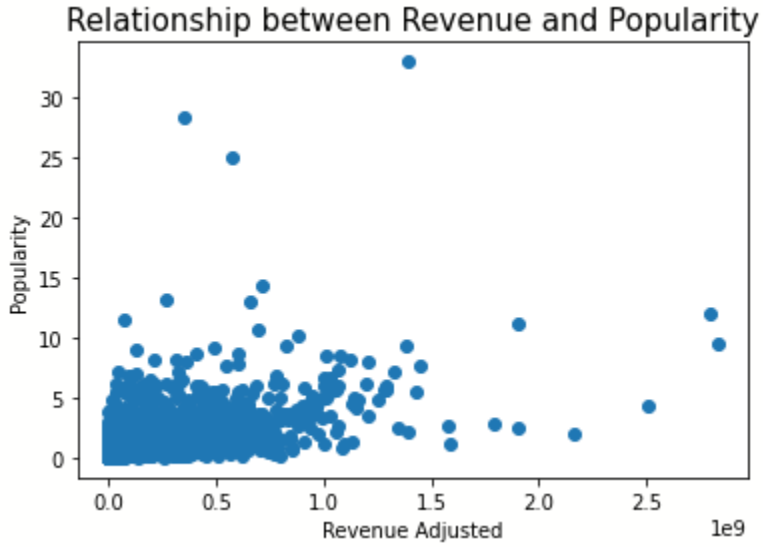
Some of the questions that I asked was the following:

1. Is there a relationship between revenue and popularity
2. Finding out if there is a relationship between runtime and popularity
3. Finding out if there is a relationship between vote average and popularity
4. Which genre of movies have the longest runtime
5. Which genres of movies have the highest average vote count
6. Finding out which directors have directed the most movies
7. Finding out the production companies that have produced the most movies based on the dataset
8. Properties of movies that have the highest popularity

To solve these questions, I used scatter plots, histograms, bar charts, and frequency charts.

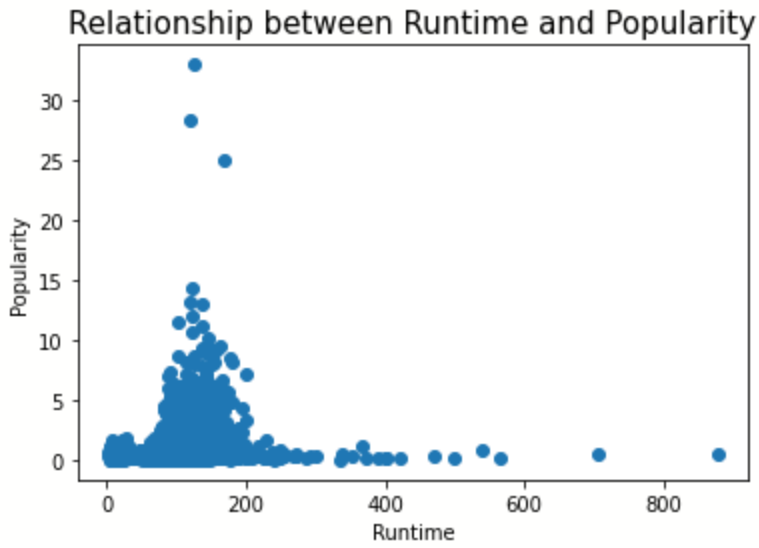
Findings for each question.

1. This is a scatter plot to find if there is a relationship between revenue and popularity



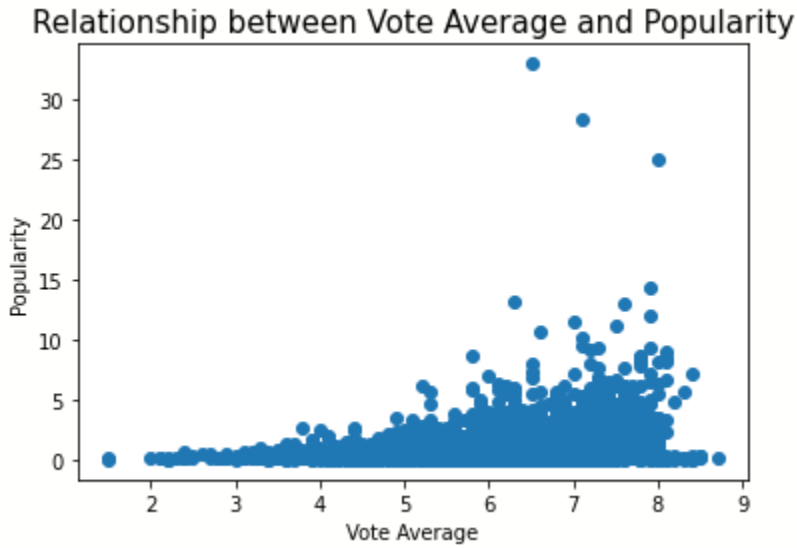

From looking at the graphs, there does not appear to be a strong relationship between the revenue of a movie and its popularity. Before performing the analysis, I thought there would be a strong relationship

2. The scatter plot below shows if there is a relationship between runtime and popularity



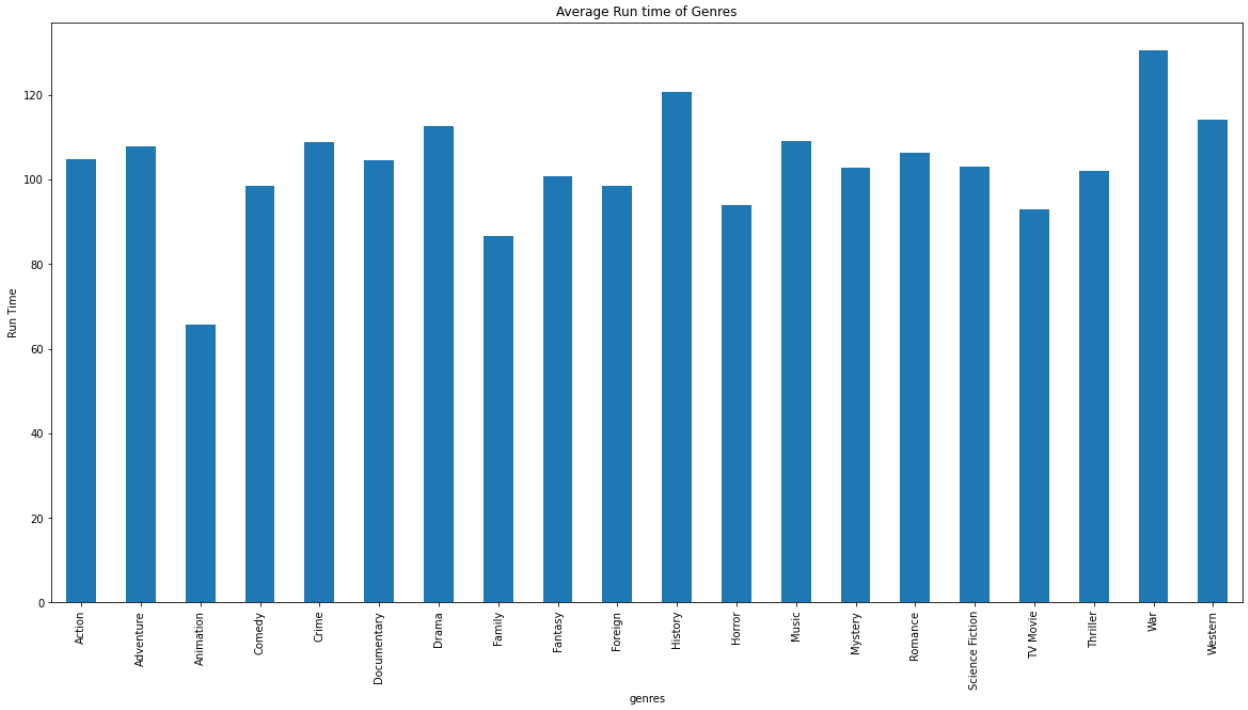

There is no relationship between runtime and popularity. Before performing the analysis, I did not think there would be a relationship

3. The scatter plot down below explores the relationship between vote average and popularity of all the movies



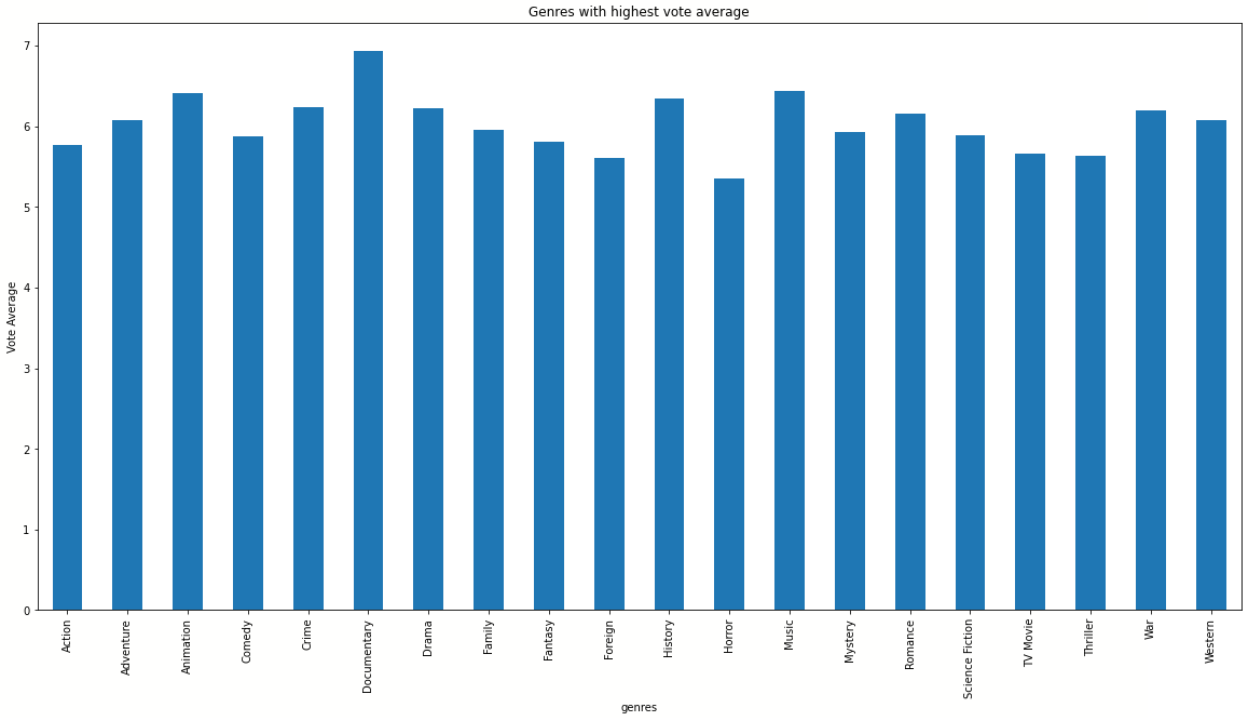

There is a strong relationship between vote average and popularity at lower values, but the relationship between the two gets weaker as the values increase

4. The bar graph down below shows the genres of movies that have the longest runtime on average




Before performing the analysis, I thought that history movies would have the longest runtime, but I was close. It is second to war movies

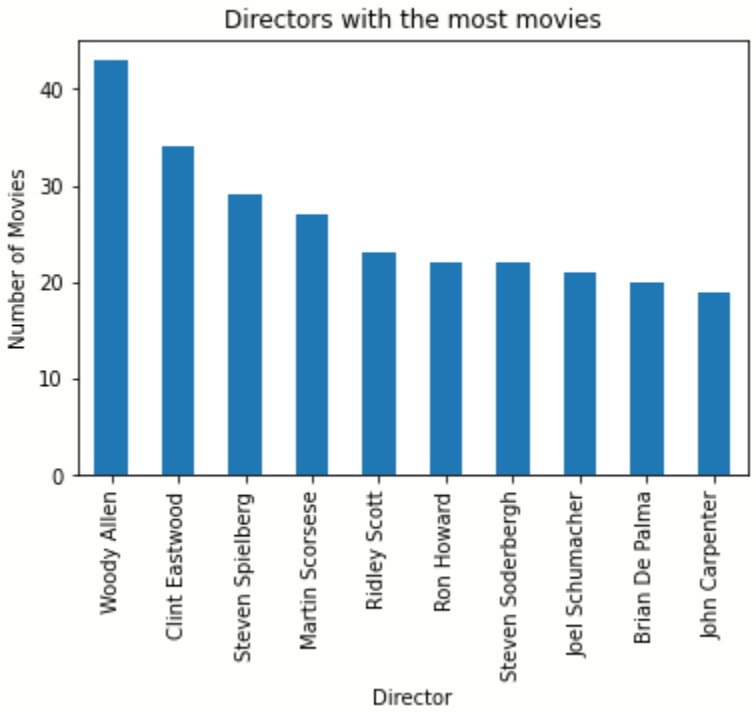
5. The graph below shows the genres with the highest vote averages




The graph shows that documentary movie genres have the highest vote average.

I did not expect that since I personally find documentaries boring, but everyone else is different

6. The graph below shows the directors who have directed the most movies found in the dataset

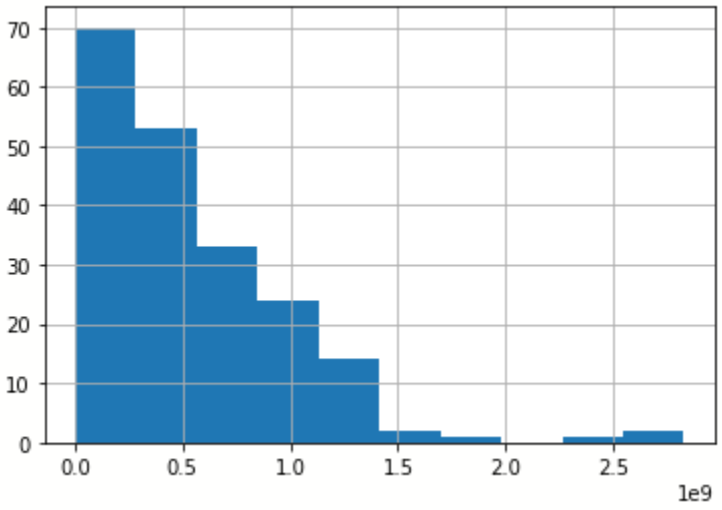


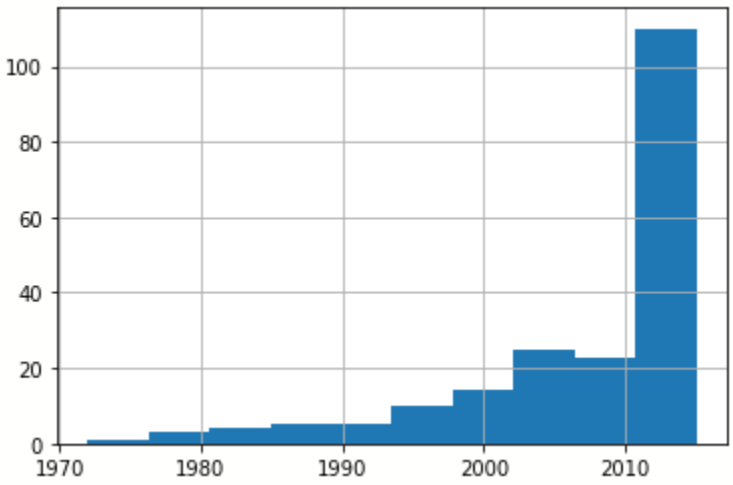

I did not think Woody Allen directed the most movies. I thought that Steven Spielberg would.

7. The next part shows the top 10 productions companies that have produced the most movies:

| | |
|---|---|
| Universal Pictures | 460 |
| Paramount Pictures | 426 |
| Columbia Pictures | 271 |
| Twentieth Century Fox Film Corporation | 242 |
| Walt Disney Pictures | 213 |
| New Line Cinema | 206 |
| Warner Bros. | 172 |
| Miramax Films | 132 |
| TriStar Pictures | 121 |

8. The charts below show the properties of the most popular movies, and further analysis




In the first chart, most movies that had the highest popularity had a revenue of 0.25e9 or $250,000,000

In the second chart, most movies that had the highest popularity were released after 2010

## Data Wrangling Process

For the data wrangling process, I did the following steps:

1. Drop duplicates
2. Drop irrelevant columns which included homepage, id, website url, keywords, overview, cast, and tagline
3. Remove extra entries in certain columns
4. Replace zeros in columns with the average values

Sources Used:


1. https://realpython.com/pandas-groupby/
2. https://www.delftstack.com/howto/matplotlib/pandas-plot-multiple-columns-on-bar-chart-matplotlib/