Movie Industry Exploratory Analysis

Ty Johnson, Joey Guthrie, Reese Bottorff

1. Data Description

Dataset Name: Movie Industry

Link: Movie Industry (kaggle.com)

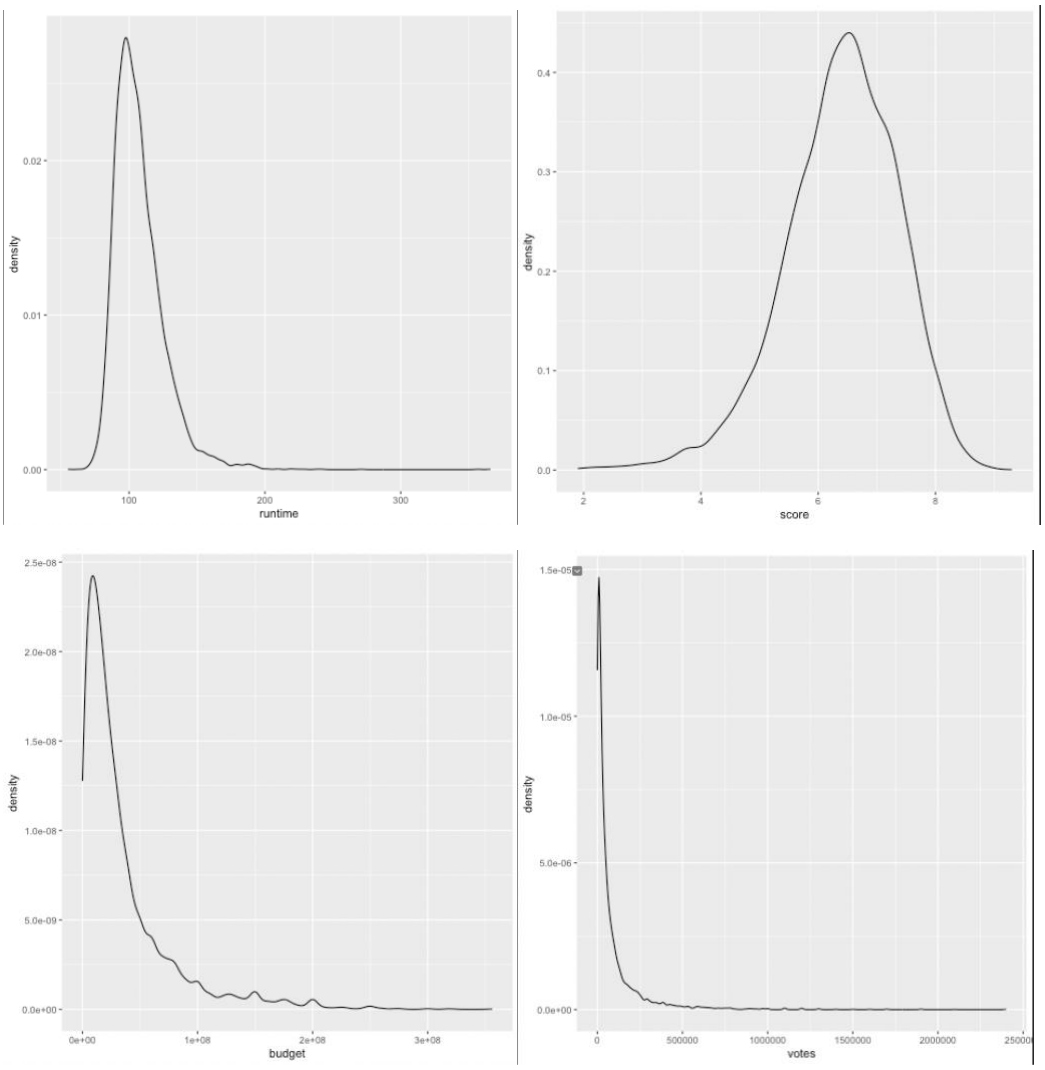
The dataset that we chose was Movie Industry from Kaggle. There are 7668 total entries with movies from 1980 – 2020. The dataset includes 15 variables that track various information about movies. The distribution of discrete and categorical variables goes as follows

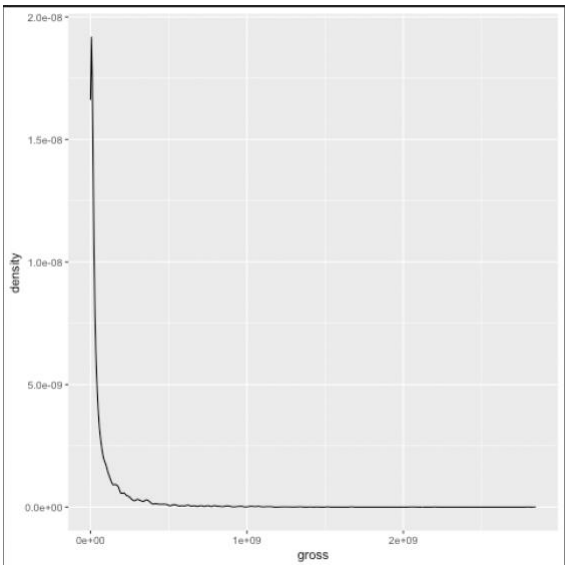
- Discrete: name, rating, genre, year, released, votes, director, writer, star, country, company
- Continuous: score, budget, gross, runtime

2. EDA Findings

- Using the head() function, we were able to see the top 6 movies which include the rating, genre, year, release date and much more.
- Using the glimpse() function, we were able to see each column and the data within the column. We are also able to see what type of data the column contains (character, int, object)
- Using the summary() function, we were able to see the statistics of each column. We found it interesting that the mean release year was 2000 and the mean score was 6.39. This shows that most movies were released in the year 2000 and the mean score was positive/high
- Using colSums() function, we were able to see that score, votes, budget, gross, and runtime had missing values. Using mean and median, we were able to impute the missing values.

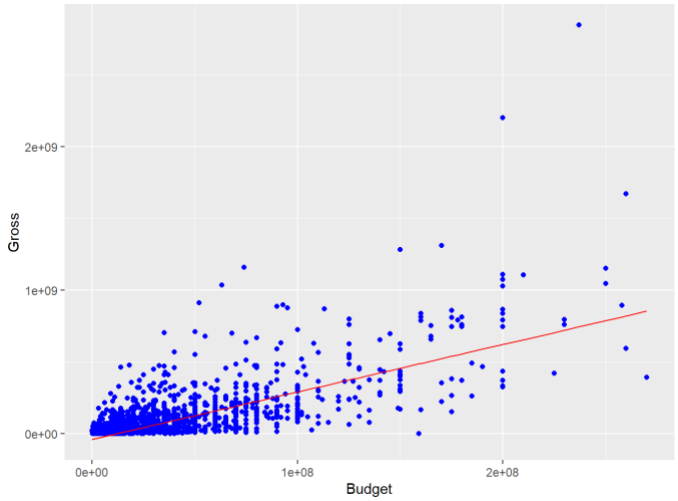
- Below are the density graphs showing the data from each missing column, showing the data before the missing values are imputed.





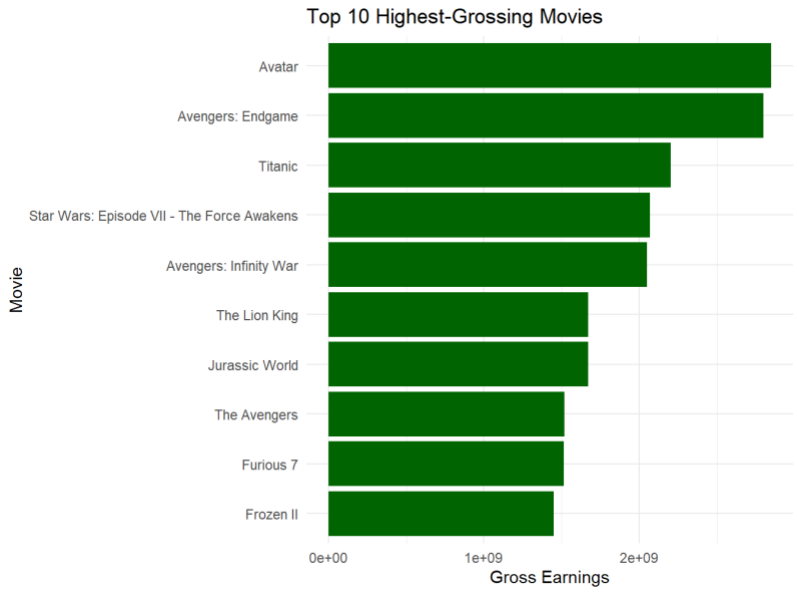
3. Regression Analysis

For our regression analysis, we chose to use the budget of a movie to predict the gross. We found that our model failed to predict the potential gross of a movie based on its budget with accuracy. This tells us that budget is not a strong factor in the potential gross of a movie. The MAE in our dataset was 69 million which could suggest that our model's predictions were off by that number on average. The MSE was to e^16 in scientific notation, suggesting that there were errors out large outliers in our dataset that were affecting the model. The RMSE was 116 million which again shows the inaccuracy of the model.

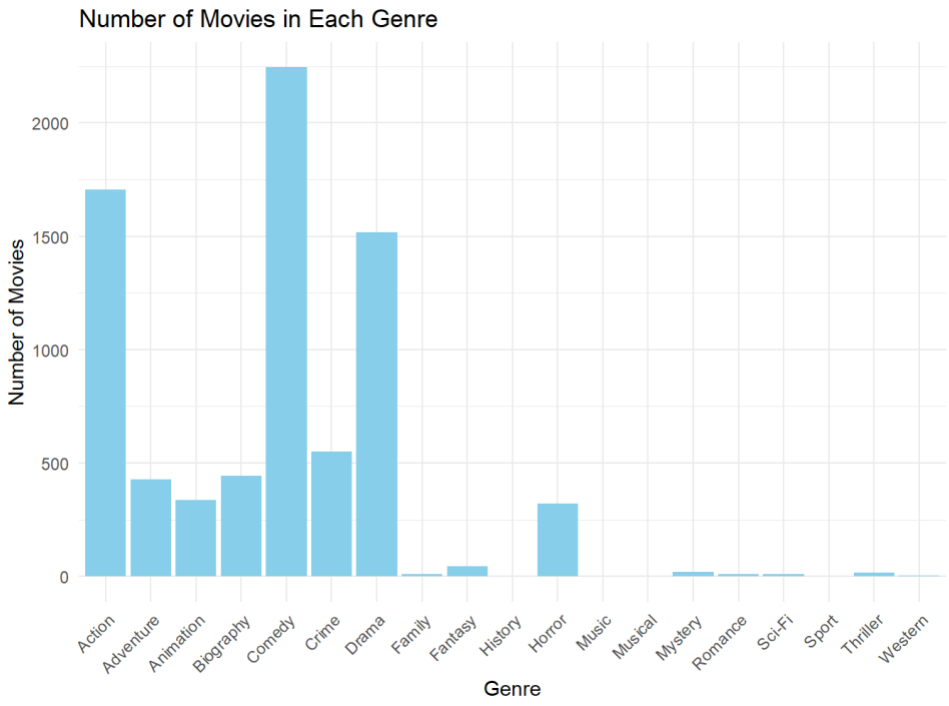


4. EDA Findings Continued

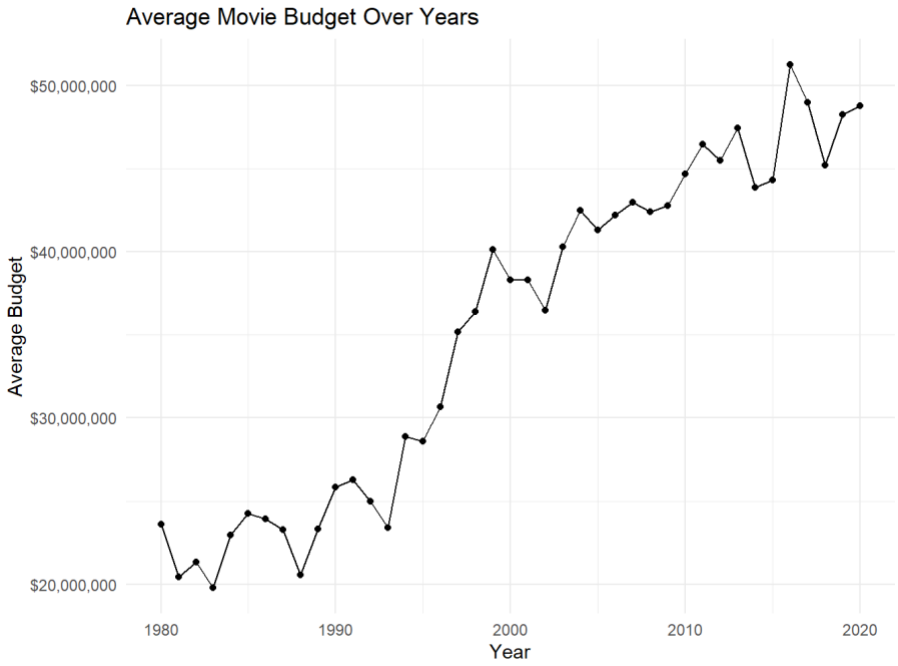
Graph A tells us the top 10 grossing movies. Avatar, Avengers: Endgame, and Titanic hold the top three spots.



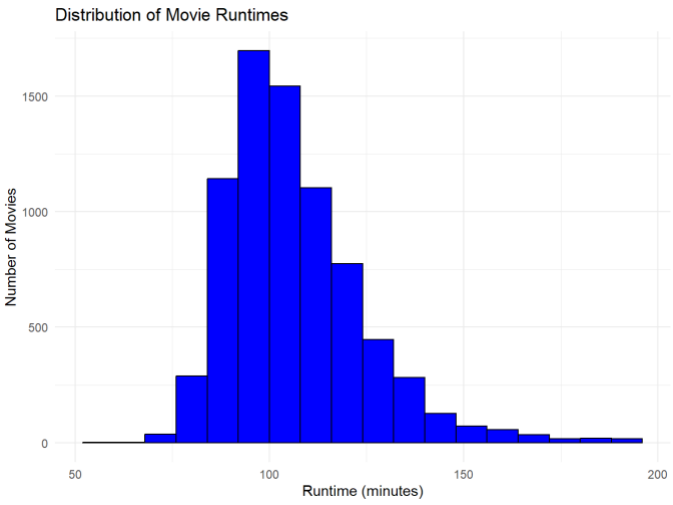
Graph B tells us the number of movies that are made in each genre. Comedy appears to be the most popular genre for movies, as it holds the most movies. Action and Drama follow as second and third.



Graph C tells us about the average budget for movies over the years since 1980. Companies are spending increasingly more on movies each year, with it not being in the 40–50-million-dollar range for movies today.



Graph D tells us the distribution of movie runtimes. Most movies fall into the 75–120-minute range. The data in the histogram is skewed is skewed to the right.



5. Conclusion and Recommendations

In conclusion, there were several takeaways that we got from this data exploration. Year-by-year, the average movie budget continues to grow. We found that the gross from movies grew

year by year until peaking in 2010, dropping off significantly year-by-year since then (which can likely be explained by the rise of streaming services). The top three grossing movies since 1980 were Avatar, Avengers: Endgame, and Titanic. We found that most movies fall into the comedy, action, and drama categories. Overall, our project was a success, and we were able to gain insight into the movie industry. The dataset was an easy one to work with. If we were to further analyze, some of the things we would look into are comparing trends of the box office and streaming growth patterns over the past several years and seeing how movies perform in respect to both. Additionally, we would try to find more factors that may explain the potential gross of a movie.