Movie Data
Exploratory Analysis

Zane Brown, zbrown2@bellarmine.edu

I. INTRODUCTION

The chosen dataset contains information on movies, including various features such as the movie title, movie id release date, budget, revenue, runtime, and genre. The data is available on The Movie Database (TMDb) and IMDb websites. The dataset provides an opportunity to analyze the factors that may affect the success or failure of a movie, including the genre, budget, runtime, and release date. As a movie fan, I found this dataset interesting to explore and gain insights into the movie industry.

II. DATA SET DESCRIPTION

This data set contains 3997 samples with 10 columns with various data types. The popularity score, vote average and vote count are categories that are provided by TMDb. Users can rate a movie on a scale from 1-10 and this is how the vote average value is acquired. The vote count is how many votes have been submitted for a movie by the users. The Title and ID variables are just identifiers and won't be included as variables in the predictive model. The dataset has no null values present. A complete listing is shown in **Table 1**.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
ID	Nominal/Int64	0%
Release Date	Nominal/Date-time64	0%
Popularity Score	Ratio/Float64	0%
Title	Nominal/Object	0%
Vote Average	Ordinal/Float64	0%
Vote Count	Ratio/Int64	0%
Genre Name	Nominal/Object	0%
Budget (USD)	Ratio/Float64	0%
Revenue (USD)	Ratio/Float64	0%
Runtime (minutes)	Ratio/Float64	0%

III. Data Set Summary Statistics

This section provides an overview of the key statistical measures and characteristics of the dataset being analyzed. These statistics help to identify the distribution of the data, the range of values, and any outliers or anomalies that may be present. Understanding these summary statistics is essential for interpreting the results of any data analysis, as they provide important insights into the nature of the data and the patterns that may exist within it.

Table 2: Summary Statistics for Movie Data

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Popularity Score	*3997*	*31.41*	*89.31*	*0.6*	*11.05*	*16.55*	*28.18*	*2,723.9*
Vote Average	*3997*	*6.52*	*0.80*	*2.4*	*6.0*	*6.54*	*7.1*	*9*
Vote Count	*3997*	*2041.66*	*3387.74*	*3*	*223*	*781*	*22,261*	*30,865*
Budget	*3997*	*30,527,462*	*43,782,606.97*	*200*	*5,000,000*	*15,000,000*	*35,000,000*	*460,000,000*

Revenue	*3997*	*92,788,798*	*199,336,091.39*	*200*	*5,200,000*	*24,261,570*	*90,000,000*	*2,920,357,000*
Runtime	*3997*	*111.19*	*21.42*	*61*	*96*	*107*	*122*	*339*

.

Table 3: Proportions for Genre (n=3997)

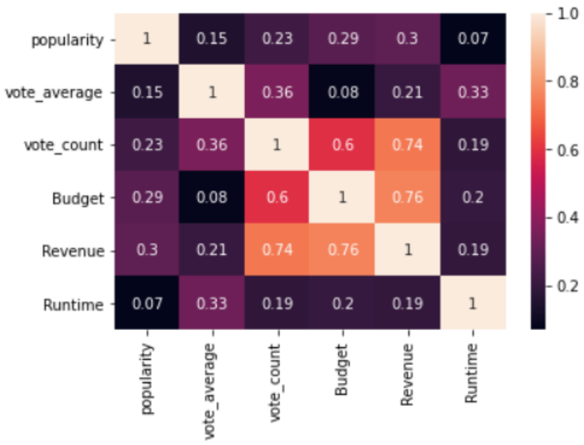
Genre Name	*Frequency*	*Proportion (%)*
Action	*674*	*16.86%*
Adventure	*221*	*5.5%*
Animation	*125*	*3.12%*
Comedy	*785*	*19.63%*
Crime	*182*	*4.5%*
Documentary	*28*	*0.7%*
Drama	*934*	*23.36%*
Family	*85*	*2.12%*
Fantasy	*90*	*2.25%*
History	*27*	*0.6%*
Horror	*285*	*7.1%*
Music	*34*	*0.85%*
Mystery	*46*	*1.15%*
Romance	*112*	*2.8%*
Sci-Fi	*97*	*2.42%*
Thriller	*200*	*5%*
War	*40*	*1%*
Western	*32*	*0.8%*

Table 4: Proportions for Month (n=3997)

Month Name	*Frequency*	*Proportion (%)*
January	*265*	*6.63%*
February	*310*	*7.75%*
March	*312*	*7.8%*
April	*266*	*6.65%*
May	*298*	*7.45%*
June	*328*	*8.2%*
July	*310*	*7.75%*
August	*339*	*8.48%*
September	*461*	*11.53%*
October	*414*	*10.35%*
November	*319*	*7.98%*
December	*375*	*9.38%*

Table 5: Correlation Table/Tables

	Popularity	*Vote Avg.*	*Vote Count*	*Budget*	*Revenue*	*Runtime*
Popularity	*1.00*	*0.149215*	*0.226094*	*0.287415*	*0.298056*	*0.069997*
Vote Avg.	*0.149215*	*1.00*	*0.359227*	*0.079788*	*0.205891*	*0.333734*
Vote Count	*0.226094*	*0.359227*	*1.00*	*0.595767*	*0.738268*	*0.188934*
Budget	*0.287415*	*0.079788*	*0.595767*	*1.00*	*0.759606*	*0.203784*
Revenue	*0.298056*	*0.205891*	*0.738268*	*0.759606*	*1.00*	*0.194891*
Runtime	*0.069997*	*0.333734*	*0.188934*	*0.203784*	*0.194891*	*1.00*



IV. DATA SET GRAPHICAL EXPLORATION

In this section, we will start by looking at the distribution of variables in the data set using distribution charts such as histograms and density plots. Next, we will explore the relationship between pairs of variables using scatter plots and visualize categorical variables in the data set using bar charts. Finally, we will use boxplots to explore the distribution of variables by different groups or categories in the data set. By using these graphical exploration techniques, we can gain a deeper understanding of the data set and identify any interesting patterns or relationships that may be useful for further analysis.

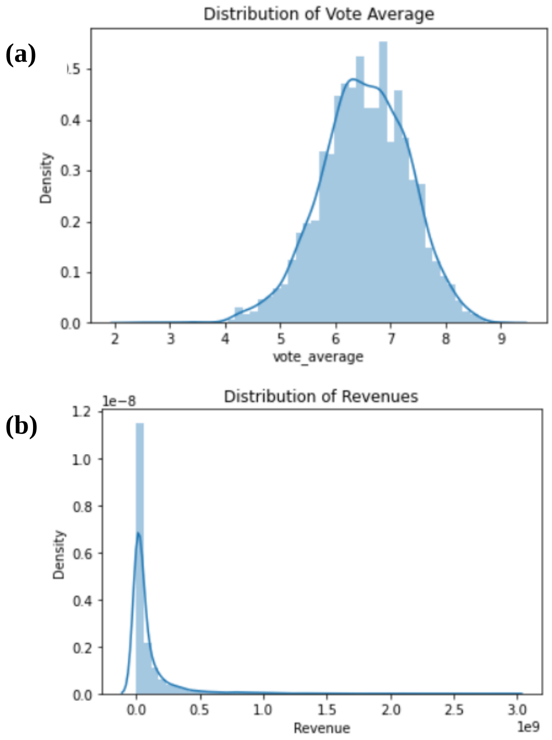
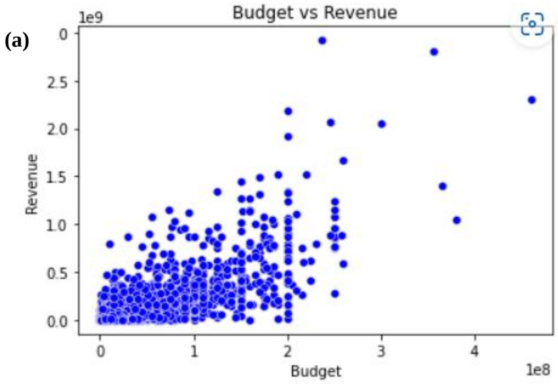


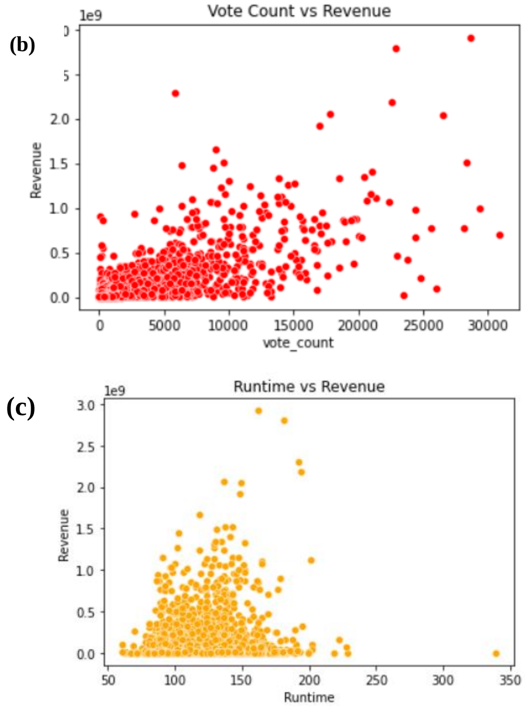
Figure 1: (a) Distribution of Vote Average from Movie dataset (b) Distribution of Vote Average value from Movie dataset

As we can see from looking at **Figure 1.a**, the distribution of vote average is relatively normal with the center being around 6.5. This tells us that the majority of the movies in this dataset have a vote average between 5.5-7.5. This turned out to be the only variable that was even relatively normally distributed.

Now looking at **Figure 1.b**, the distribution of revenues is highly skewed to the right. This is due to blockbuster movies that are usually massive box office hits that generate a lot of revenue. These movies generate a good amount more than most of the movies in the dataset and thus pull the mean of revenues up. All of the other continuous variables had a similar distribution, likely due to the same reason that the Revenue distribution is skewed.



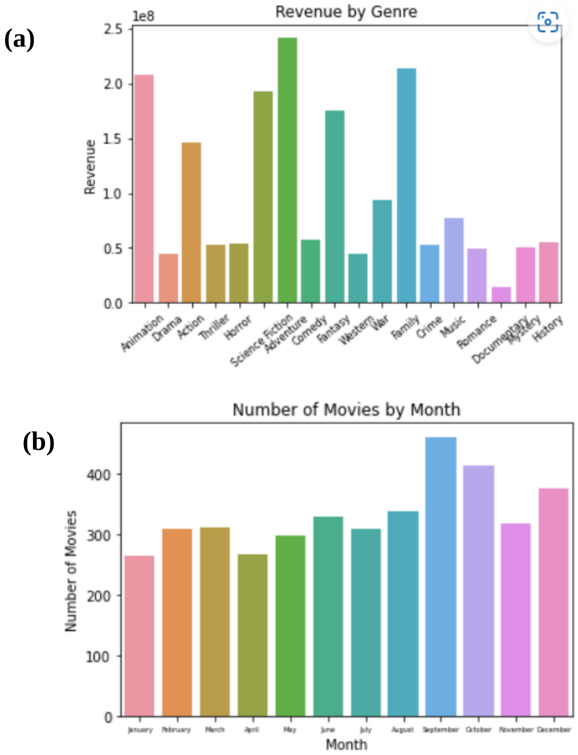
Moving on to scatter plots we are looking to see if any variables have relationships with each other. **Figure 2.a** shows budget plotted against revenue. There seems to be a generally positive relationship between budget and revenue, with most of the data clustered towards the lower left end of the scatter plot. This tells us that generally if a movie has a higher budget, the revenue may be higher as well.



Similarly, to the Budget/Revenue scatterplot, the Vote Count/Revenue scatterplot shows a weak positive relationship. The points are a little more spread out on the higher end than the previous graph, so its relationship is weaker than that of Budget/Revenue. Since TMDb API uses user input to get vote count, it would make sense that movies that are more popular and are doing well would have more people submit a rating for the movie.

The final scatterplot that is included is Runtime/Revenue. There doesn't seem to be any relationship present in the plot. However, we can see that most of the data points lie from just under 100 minutes to 200 minutes. This plot leads us to believe that the runtime of a movie likely doesn't have a big impact on the revenue of the movie.

Figure 2: (a) Comparison of Budget/Revenue from dataset, (b) Comparison of Vote Count/Revenue from dataset, (c) Comparison of Runtime/Revenue from the Movie dataset.



Looking at **Figure 3.a**, we can see that the bar plot indicates that Adventure is the genre that generated the most revenue in this dataset. Oppositely, we see that Documentaries generate the smallest revenue out of the dataset. Additionally, the animation, sci-fi, adventure, fantasy, and family genres all generate significantly more revenue than the others.

Figure 3.b is a bar plot showing the density of movies in each month for the dataset. September, October, and December have the highest number of movies released. This could be explained by the holiday season being at the end of the year. April is the month that has the fewest movies released.

Figure 3: (a) Comparison of Budget/Revenue from dataset, (b) Comparison of Vote Count/Revenue from dataset

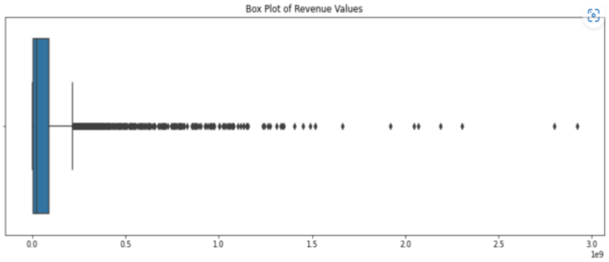


Figure 4: Boxplot of Revenue Values for Movie dataset

Lastly, **Figure 4** is a boxplot of the revenue values that shows us a distribution of the revenue values. We already knew that these values were skewed from the distribution plots, however the boxplot also shows us how many outliers there are and how far from the bulk of the data the outlier is. There are many outliers present which may give us a mean revenue that is not a good true description of the dataset or more future predictive models. This suggests that the median may be a better measure to use to get useful results.

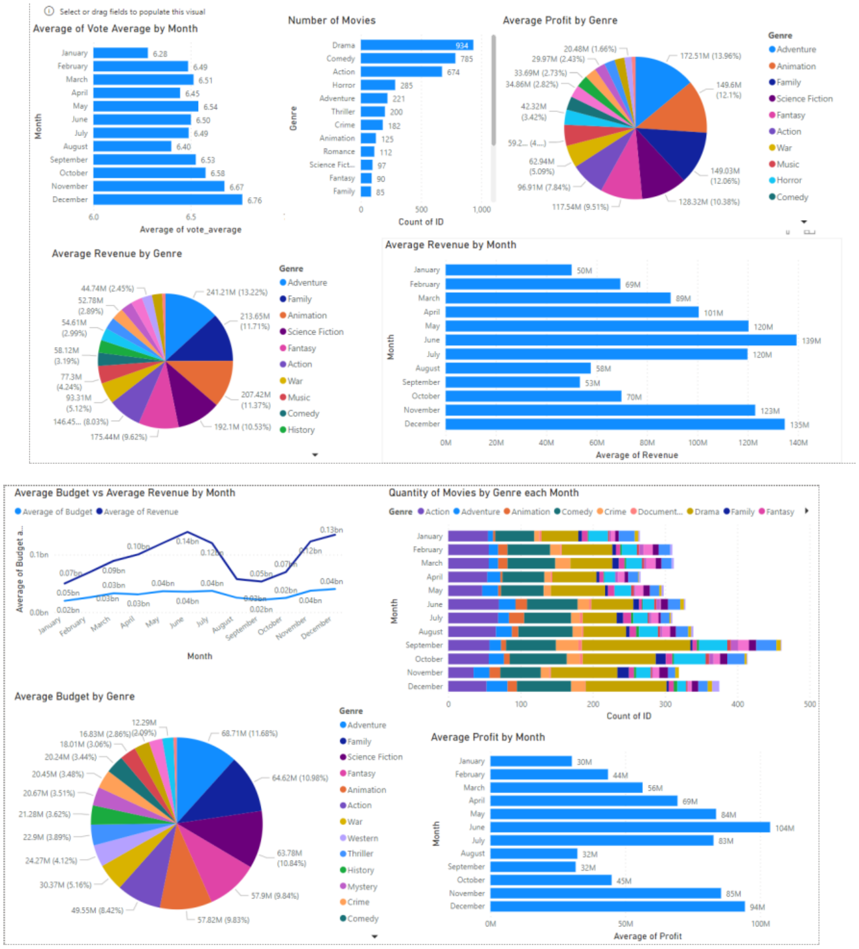


Figure 5: Power BI dashboards showing various visualizations dealing with Profit, Revenue, Genre, and Budget

V. SUMMARY OF FINDINGS

After exploring the dataset, I have found that a lot of the more successful movies are released either during the middle of the summer (July) or around holiday season (December). This is consistent with theory due to a lot of big budget movies being typically released around those times. Since I plan on using genre and release month in my model, I will have to create dummy variables for them so that they have distinct values that can be inputted into the predictive model. In the dataset, each variable has a wide range of values. For example, revenue has values that are in the millions and hundreds of millions while vote count and vote average are significantly smaller. Feature scaling will help normalize the independent variables in the dataset, that way we can fit a better model for the data. Lastly, I found that most of my variables are highly skewed with many outliers in the box plots due to the big gap in revenue, budget, vote count, popularity score values between different movies. However, this is expected because there are many movies that are released and have very little promo and would make sense that it would come nowhere near being as successful as say a Marvel movie. Since this may bias may results when creating a predictive model, I plan on adding a column that will calculate the profits of each movie and then use the median profit to build my predictive model around. That way my model will predict whether a movie will make a profit greater than 50% of the other movies present in the dataset.