An Analysis of Film Industry Performance

Nicholas De Santos
DSC 530

Introduction and Motivation

The main purpose of this project is to do an analysis of what kind of factors and characteristics (if any) of different movies can help us determine how successful a movie will be. Specifically, two measures of success that we might be looking at during this analysis is first and foremost the amount of gross revenue the movie brought in as well as how popular the movie was with audiences. For the latter we will consider an IMDb rating as our variable to help measure this.

Background: The Data

The dataset that we are working with is a movie dataset containing information about 6820 different movie titles. The data was scraped from IMDb databases. There are movies from different years between 1986 and 2016. There are 220 movies per year and each movie has the following information included in the dataset:

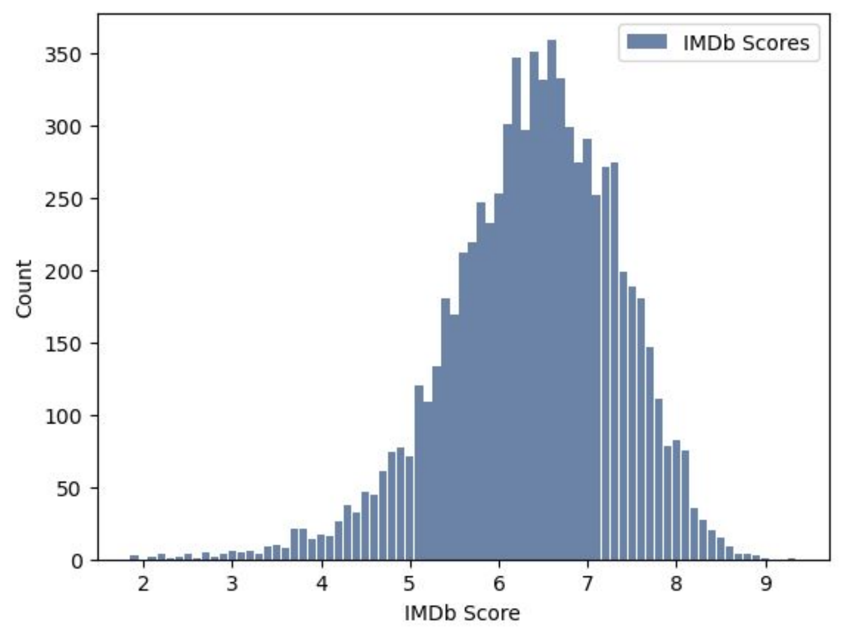
Name: name of the movie
Rating: what was the movie rated (PG, PG-13, R etc)
Genre: what genre does the movie belong to
Year: year the movie was released
Released: release date
Score: IMDb score
Votes: count of IMDb votes

Director: who directed the movie
Writer: who wrote the movie
Star: major stars in the movie
Country: what country was the movie made
Budget: budget used to make movie
Gross: how much did the movie make
Company: what production company
Runtime: how long was the movie

Variable Selection

I selected "score" and "gross" as variables because there are the variables I am planning to use as a measure of success in this analysis of the film industry. Similarly, votes could also be used as a measure of success but the amount of votes can also be an indicator on how much revenue a movie will make. Other variables that I chose to include in this analysis are the budget of the movie and the runtime for each movie. The movie genre is the only variable that is a categorical one, hence it is the only one in which we cannot create a histogram for.

Histogram: “Score” Variable Analysis

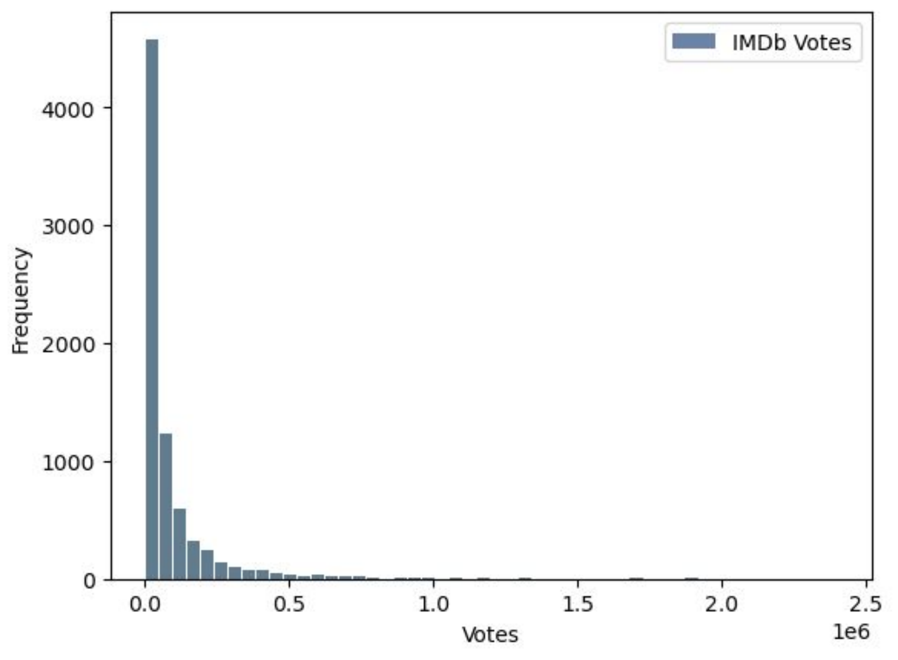



```
Score Mean:  6.390410958904109
Score Median:  6.5
Score Mode:  0    6.6
Name: score, dtype: float64
Score Standard Dev:  0.968841640253057‍6
Score Varience:  0.9386541238882352
```

Outliers: There doesn’t seem to be any extreme outliers in the distribution. Possibly on the left side.
Spread: There seems to be a slight skew to the left.
Behavior: Distribution appears to be more or less symmetrical. Majority of observations between 5 and 8.

Histogram: “Votes” Variable Analysis

```
Votes Mean:  88108.50476190477
Votes Median:  33000.0
Votes Mode:  0    13000.0
Name: votes, dtype: float64
Votes Standard Dev:  163323.7639095057
Votes Varience:  26674651857.567955
```

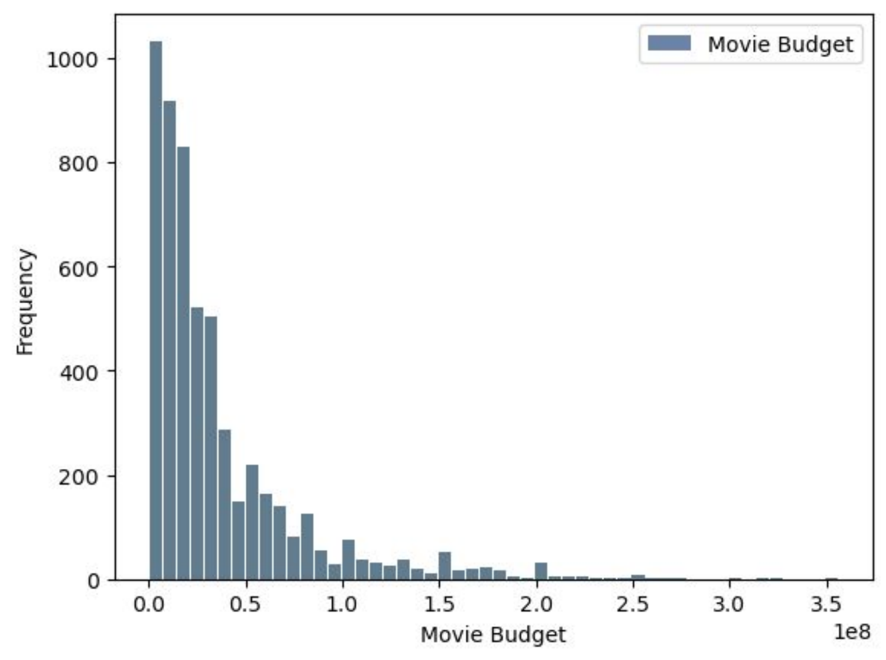


Outliers: There are definitely outliers on the right side of this distribution but it’s likely some movies are just more popular so there isn’t a reason to get rid of any outliers since the observations are possible.

Spread: Significant skew to the right

Behavior: Largest density on left side of distribution

Histogram: "Budget" Variable Analysis

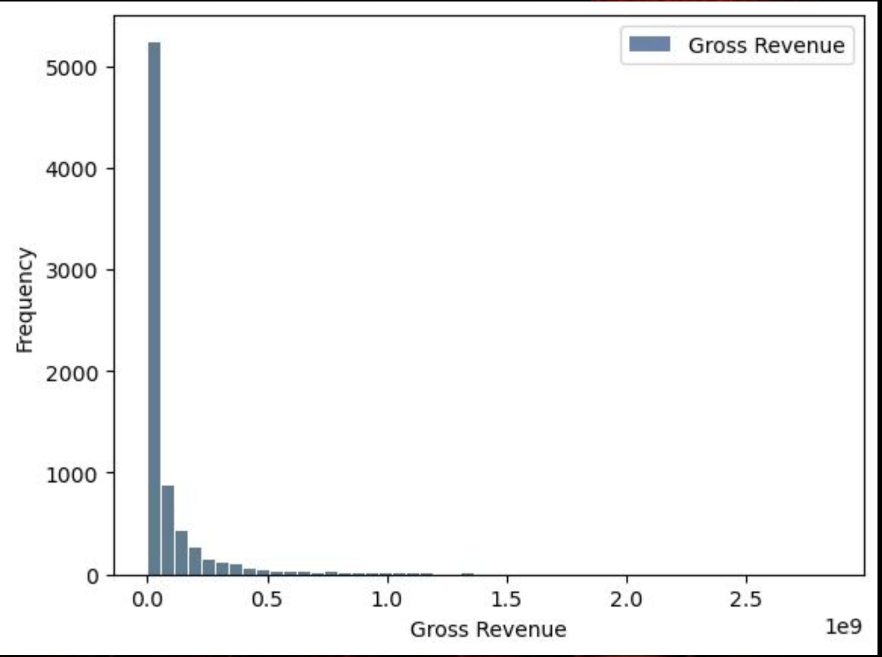



```
Budget Mean:  35589876.192650534
Budget Median:  20500000.0
Budget Mode:  0    20000000.0
Name: budget, dtype: float64
Budget Standard Dev:  41457296.60193096
Budget Varience:  1718707441540476.5
```

Outliers: Possible outliers on the right of the distribution but it's possible movies could've had a large budget so no outliers will be removed.
Spread: Significant skew to the right
Behavior: Largest density at the left side of the distribution.

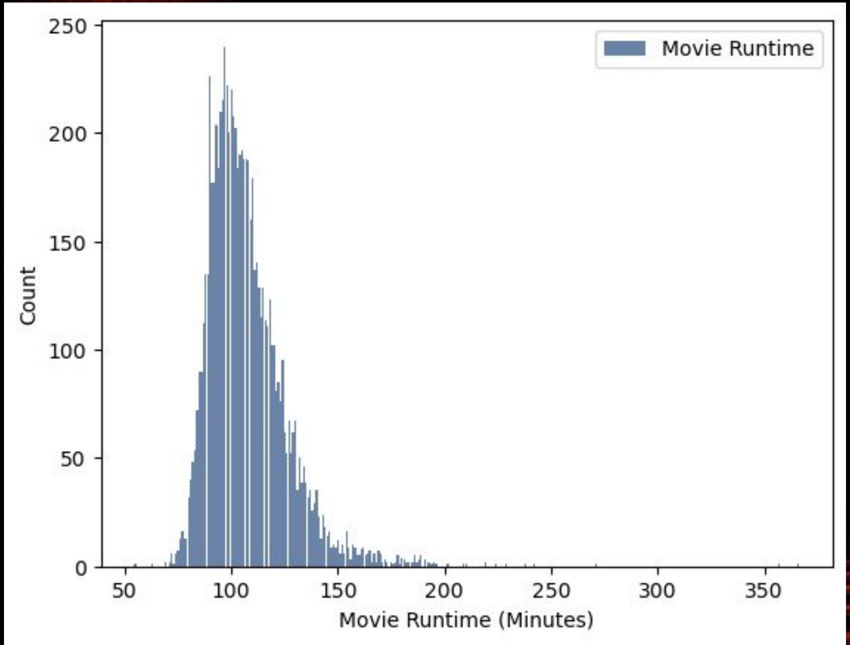
Histogram: “Gross” Variable Analysis

```
Gross Mean:  78500541.01778312
Gross Median:  20205757.0
Gross Mode:  0    14000000.0
Name: gross, dtype: float64
Gross Standard Dev:  165725124.31875733
Gross Varience:  2.7464816830467 57e+16
```



Outliers:Possible outliers to the right. Won’t be removed since it is not unlikely for some movies to make significantly more gross revenue than other movies.
Spread: Significant skew to the right
Behavior: Largest density to the left of the distribution.

Histogram: “Runtime” Variable Analysis




```
Gross Mean:  78500541.01778312
Gross Median:  20205757.0
Gross Mode:  0    14000000.0
Name: gross, dtype: float64
Gross Standard Dev:  165725124.31875733
Gross Varience:  2.746481683046757e+16
```

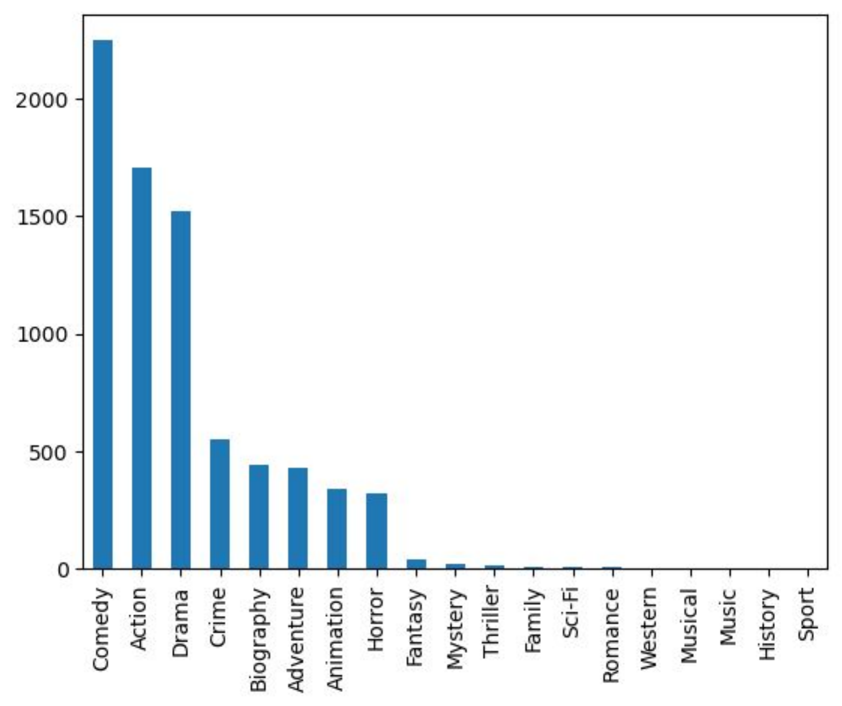
Outliers: It looks like there are some possible outliers but they won’t be removed as they are real movies. Ex: There is a movie that is actually 6 hours long (The Best of Youth).

Spread: Slightly skewed to the right

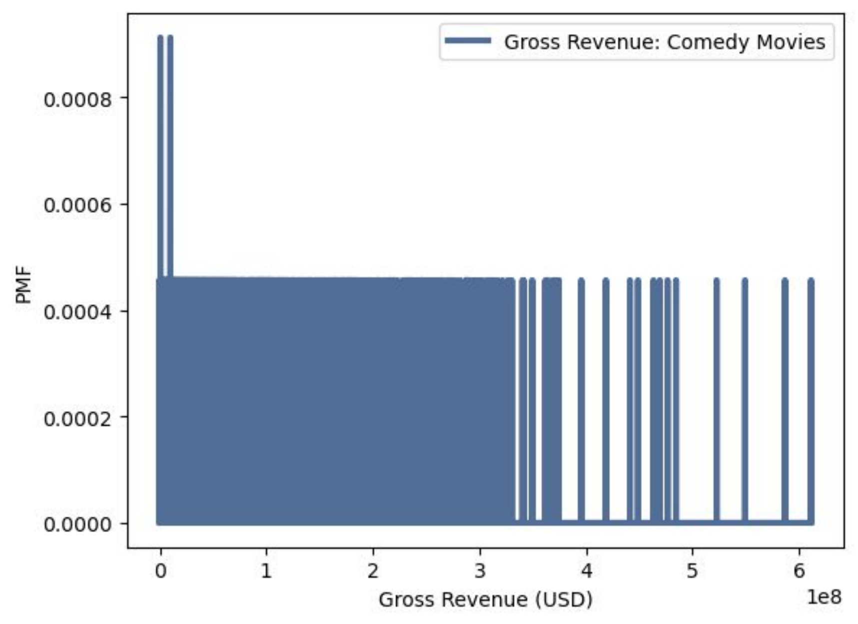
Behavior: Biggest density around 75 to 125 minutes.

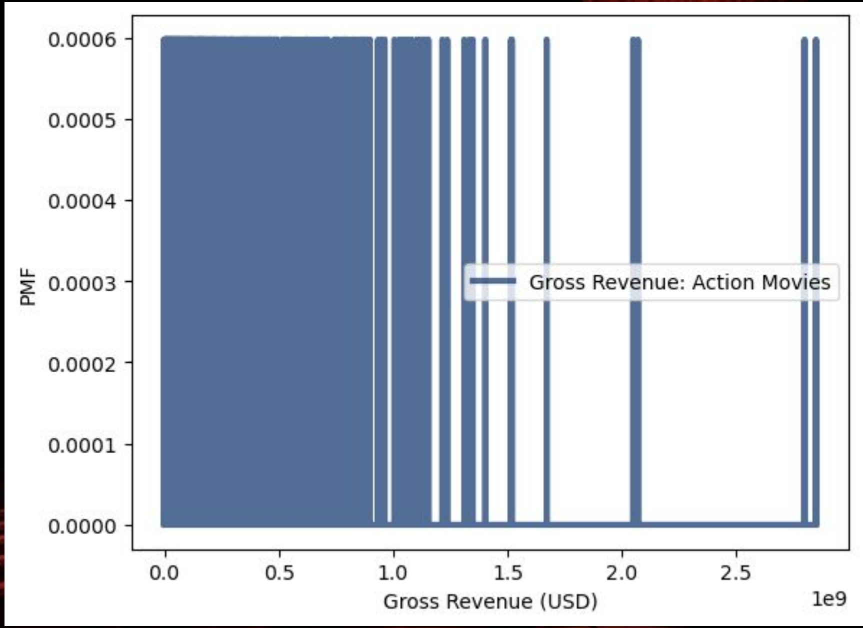
Bar Plot: “Genre” Variable Analysis

Because this is a categorical variable, there is no way to find a traditional average. We can only analyze the frequency of which each type of observation is observed. Based on the data, we see that there are more Comedy movies than any other genre. The least popular movie genre is “Sport”. In order, the top movie genres here are Comedy, Action, Drama, Crime, Biography, Adventure, Animation and Horror.

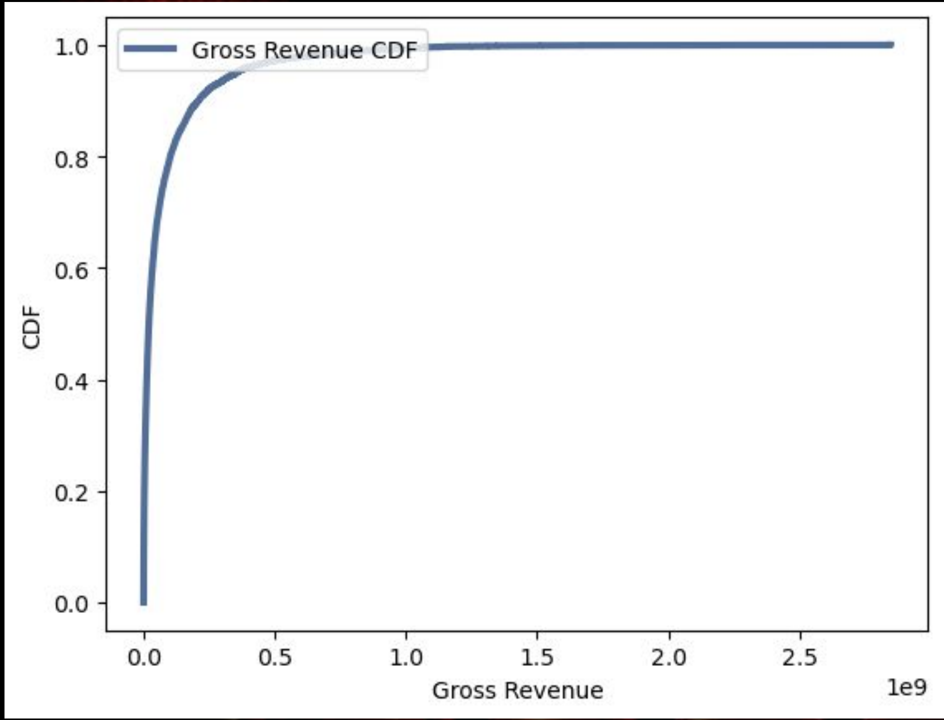


Analysis of Gross Revenue of Different Movie Genres (PMFs)





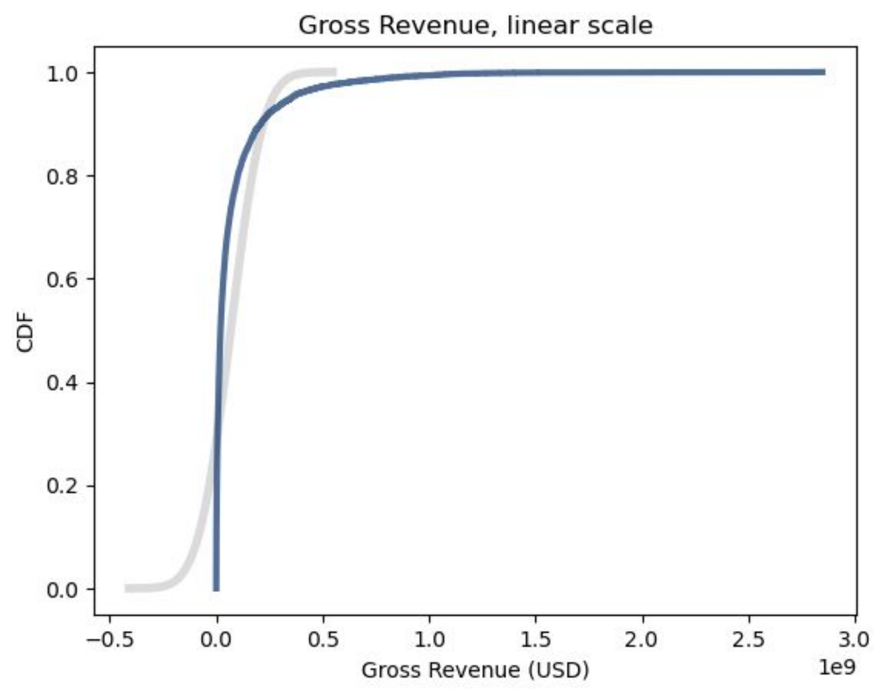
CDF of Gross Revenue

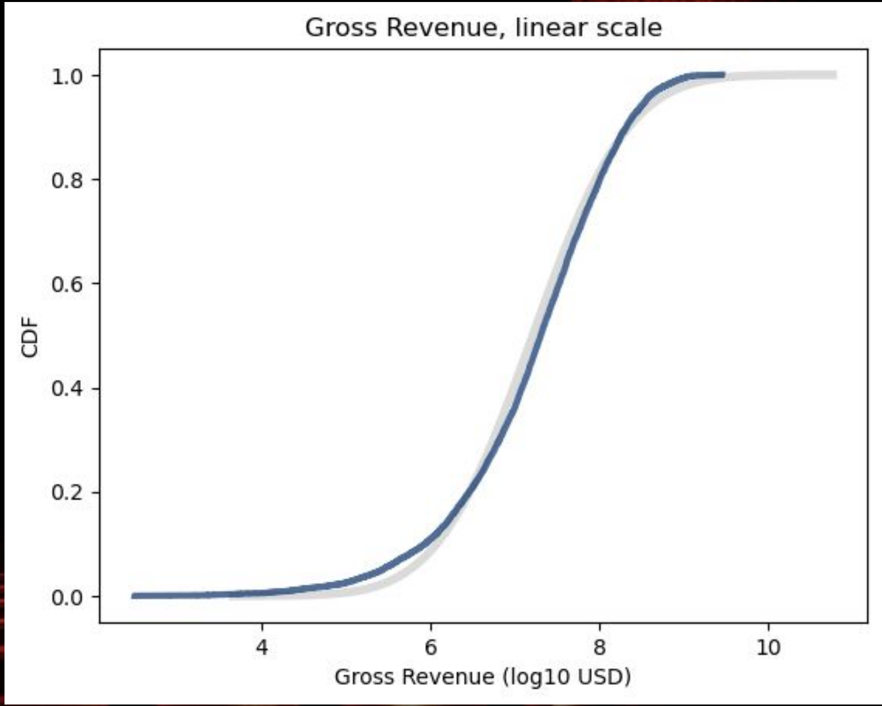


<u>What does this tell us?</u>

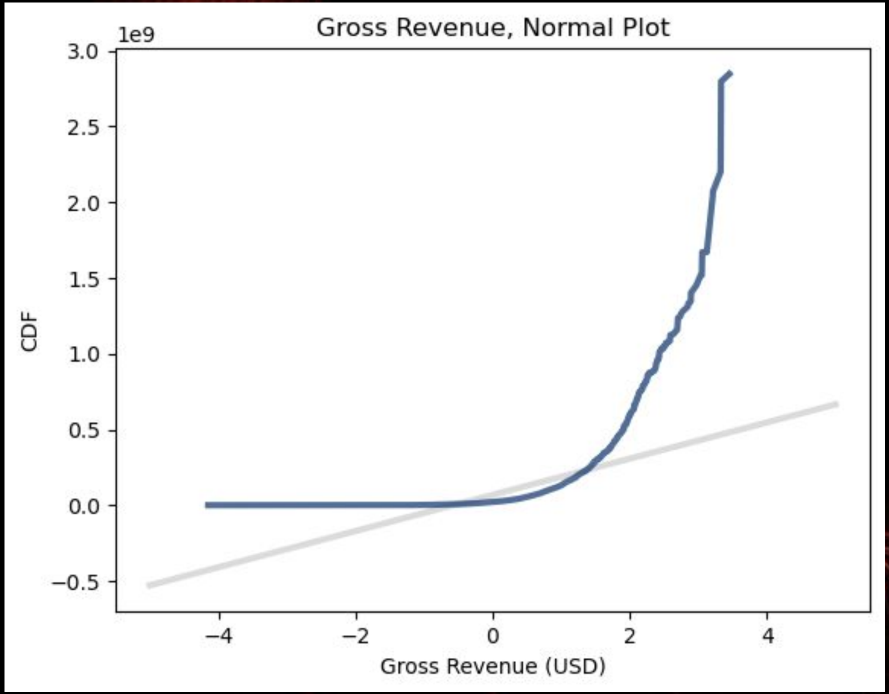
As you can see above, the CDF of the gross revenue for our data appears to follow a logarithmic path as it progresses. It's highly possible that the distribution for this variable is a lognormal distribution. If we look back to the histogram of this variable, it does appear to support this claim.

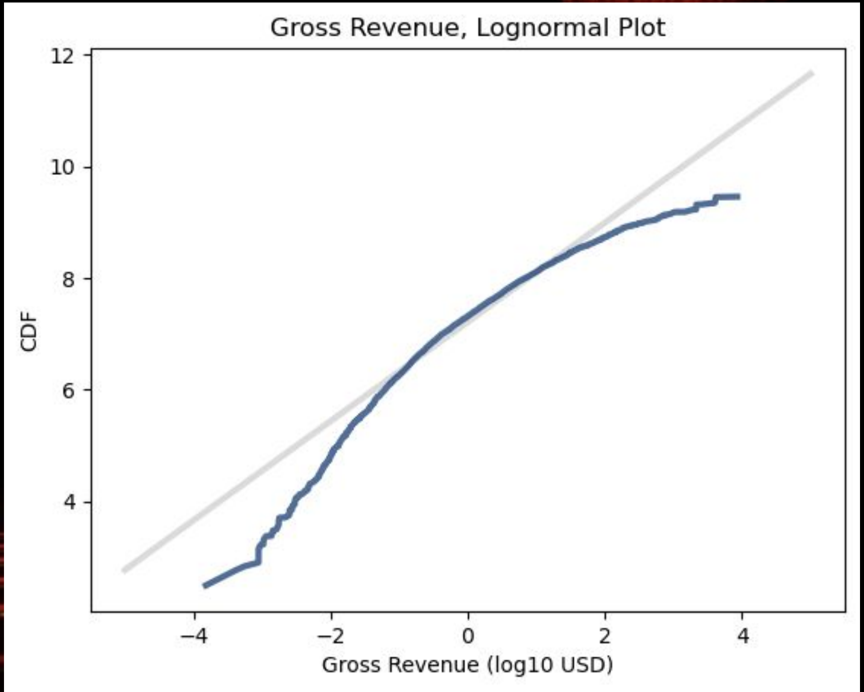
Analytical Distribution Plot





Analytical Distribution Plot

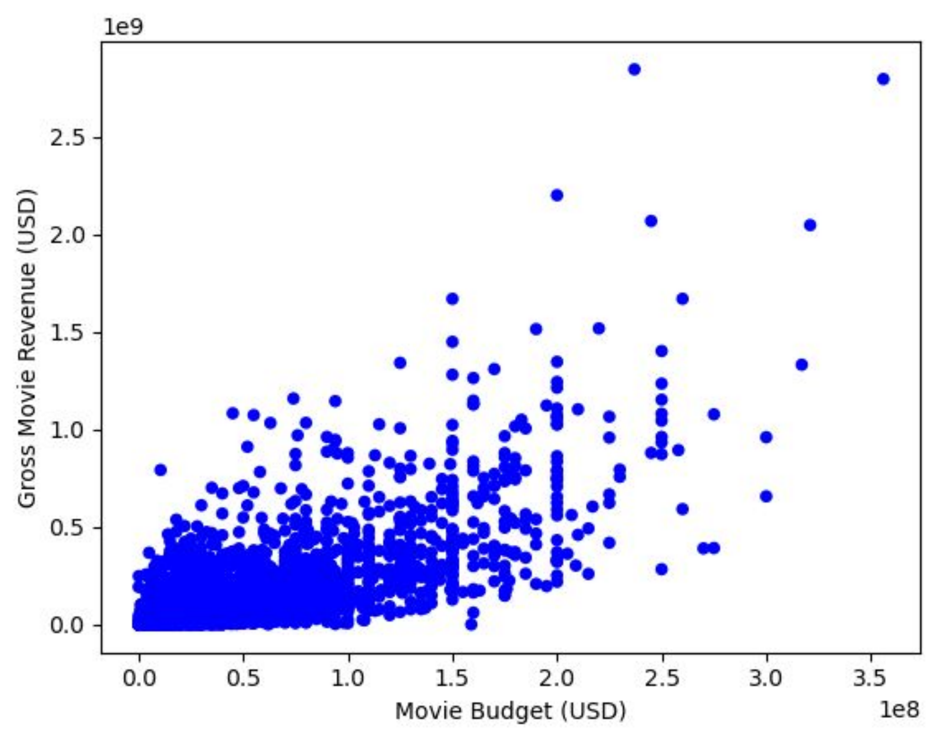




Analytical Distribution Plot

Previously I mentioned that the distribution of the Gross Revenue for the Movie Industry data set could possibly follow a log-normal distribution. Considering the figures above, it's clear to see that a log-normal distribution fits the data better than just a normal distribution. Through the graphs we see an improvement when we apply a logarithmic transformation to our data.

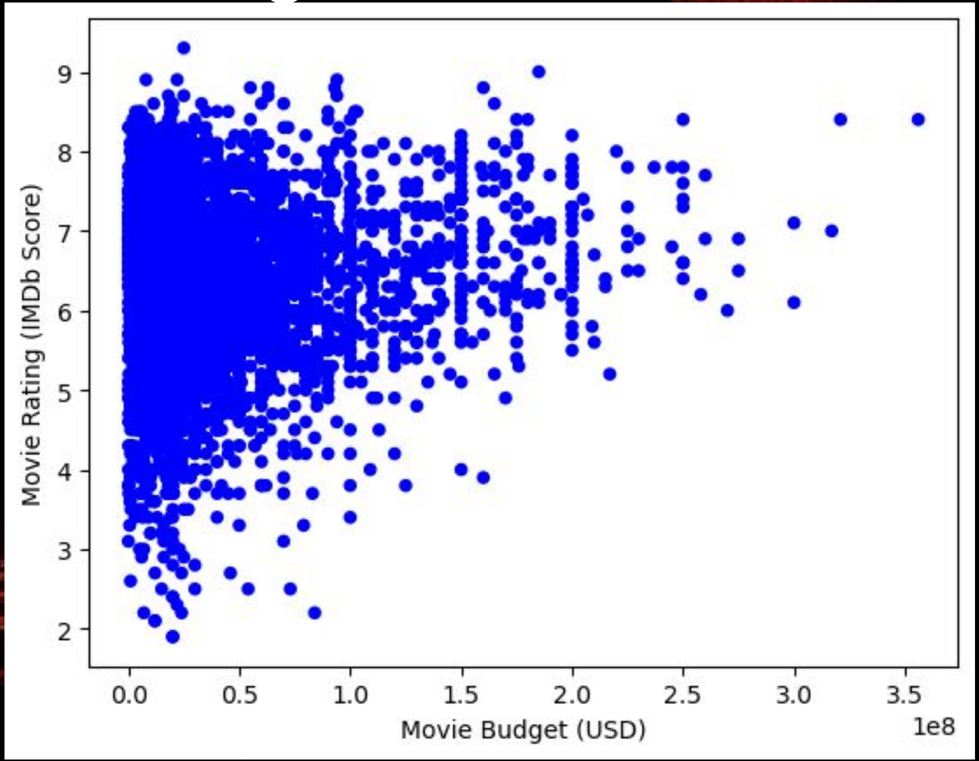
Variable Relationships: Scatter Plot of Budget and Gross Revenue




```
Variables: Budget and Gross
Covarience:  5754614010107631.0
Correlation:  0.7403948929894826
```

Variable Relationships: Scatter Plot of Budget and Movie Rating

```
Variables: Budget and Score
Covarience:  2872497.4451108794
Correlation:  0.0717919879816616
```



Hypothesis Testing

```
#separating data again
gross_comedy = moviedf.gross[moviedf.genre == "Comedy"]
gross_action = moviedf.gross[moviedf.genre == "Action"]

#cleaning data of any missing or invalid values
gross_comedy = gross_comedy[-gross_comedy.isnull()]
gross_action = gross_action[-gross_action.isnull()]

#putting data together
data = [gross_comedy, gross_action]

#Hypothesis test and result
gross_ht = DiffMeansPermute(data)
gross_pv = gross_ht.PValue()

#resulting pvalue
print(gross_pv)
```

```
0.0
```

As you can see we got a resulting p-value of 0.0. This is of course not possible for an observed p-value but this could be a minor error due to the fact that there could have been some rounding off in some of our calculations. Because of this, if we resulting in a relatively small p-value, the computer system could have rounded the small value to zero. As we recall from previous lessons, when the p-value of a hypothesis test is extremely small, it's safe to say that we can reject the null hypothesis that the means of the two samples are the same and accept the alternative hypothesis that there is a difference between the gross revenue of Comedy movies and the gross revenue of Action movies.

Regression Analysis

```
inter, slope = LeastSquares(moviedf.budget, moviedf.gross)
inter, slope
```

```
(-16825009.95297028, 3.3342796500711454)
```

```
Variables: Budget and Gross
Covarience:  5754614010107631.0
Correlation:  0.7403948929894826
```

Our model for our Gross income is a linear model in the form of (approximately) f(x) = 3.3342x - 16825009.9530. Here f(x) represents the gross revenue generated as a function of x where x is the budget the movie had. In the end, these two variables had a correlation of 0.74 which makes this a decent model.