**Summarization of EDA - IMDB Movie Dataset**

Statistical/Hypothetical Question

After careful examination of the dataset, the statistical question that I came up with was “Which variable or variables in this dataset will have a strong driving influence when choosing the most popular movies?”.

Outcome of the EDA

I believe analyzing a dataset is like telling a story. There is always a beginning, middle and end. Similarly, I would like to describe my EDA like a story. For this movie dataset, an indepth analysis conducted to analyze what variable or variables would have strong influencing choosing the most popular movies. Initially I had to clean the data freeing it from all the null values, NA’s and blank values. Then to check out the data there were 5 histograms for the 5 variables that I had chosen. Most of the histograms were skewed, except for one. The Revenue and Runtime variables were skewed to the right and the release date and the rating variables were to the left. The one variable Metascore was an exception. It is almost normally distributed with slight skew to the left.

Furthermore, I conducted other form of analysis such as Probability mass functions on the rating function and actually had to break down the rating function further for this concept by splitting into two categories high and low rating and I had created a sperate column in my dataset as portrayed in my code. In addition, I also conducted an exponential analytical distribution to check out the Runtime variable.

One of the most important analysis I felt in this EDA was the correlation analysis. This clearly showed my that the variables in this dataset don’t have much of relationship between one another. Most of variables either have a negative or a weak relationship, barring a few. I had chosen Revenue and Runtime to be compared against each other and also after conducting Pearson and co-variance analysis both had negative correlations and negative co-variances. Those few variables also did not have the strongest of relationships. The few exception relationships were Vote and Revenue, which had a correlation of 0.64, Votes and Rating of 0.52 and Metascore and Rating of 0.67.

During last stages of analysis, I had conducted a chi-square test on the variables. My main variable of the test was Rating. The last important test, I had to create was a linear regression model for this dataset, where I had to test and train the model. The dependent variables I chose was Votes and the explanatory variables were Revenue, Rank and Year. The best the model could with best variables could provide was a R-squared of 0.58, which is moderate and by no means great.

Finally based on the indepth analysis conducted, coming back to my statistics question and answering it, I believe the Votes and Rating are two best columns in this dataset which have strong influence on the popular choice of movies from 2006-2016.

What do you feel was missed during the analysis?

I felt more movies/data and variables should have been provided for analysis than just the ones given. I thought it was little less. With more movies and variables included, there would also has a higher chance of having positive and strong correlation/relationship with more variables.

Were there any assumptions made you felt were incorrect?

While examining the data before starting the analysis I assumed Revenue and Runtime would be variables that could have a strong influence, but I was totally incorrect based on the analysis conducted.

What challenges did you face, what did you not fully understand?

I had difficult time finding suitable dependent and explanatory variables for the linear regression model due to the weak and negative correlations. Eventually, I found those variables which gave me a decently moderate model with an average r square score.