Exploratory Data Analysis Term Project:

The Movie Dataset

Jonathon Scroggins

DSC 530

Professor Parajulee

After performing an EDA on the Movie Industry Dataset found at <https://www.kaggle.com/danielgrijalvas/movies>, I would say that I feel that the outcome overall would be inconclusive. I feel like based on the data that I obtained some of the variables did have an impact on being able to predict the gross of a movie, but I was unable to fully prove that during this analysis. Based on all the analysis I did on the individual variables through histograms, PDFs, and CDFs, I feel like the data had good distributions of certain variables but not of others. The scores and the runtimes of the data were normally distributed, but the budget, gross and number of votes was not and in fact the budget and gross were very skewed to the right. I was however pleased with some of the correlation analysis that I received. Based on that analysis it indicated that there is a positive relationship between the budget and then gross that would be considered significant. There were positive correlations between the gross and the score and the gross and the runtime, but those would not be considered significant correlations. Due to the significant positive correlation between budget and gross, that would indicate that the more money you spend on a movie the more money it will make. It makes me wonder if the advertising money spent or the number of theaters released in would have been a good variable to be included as well in the data set to build a better regression model.

I was hoping to be able to determine the relationship between the rating and the score or the gross, but there did not seem to be a definitive answer to the question. It seems that R rated and PG13 rated movies overall were the top grossers, but I am not sure I had enough data or information to make that assertion. It could also be due to the challenges I was facing when coding in Python. I was unable to do any kind of clustering or knn analysis on the data to see if that might be a good way to see how the ratings effected the gross or the score of a movie. I think overall that is some of the biggest challenges I faced and will have to work on moving forward. I wanted to try and build a full regression model and make predictions in Python but were unable to do so at this time due to a lack of Python experience I feel. I also feel like that bleeds into the assumptions that I made that I feel were incorrect. I thought that with this data I could make a good model and determine the things that most influenced the gross earnings of a movie. It is possible with some more analysis of the data even so far as to add in the directors, actors and other things involved in the movie it might build a better model. For now, the best conclusion I can make on the data would be that there is a significant relationship between the budget of a movie and the amount the movie makes.