**DSC 530 Final EDA Project**

Zach Quinn

Bellevue University

As conventional movie theaters experience declining attendance as a result of the coronavirus, streaming giants like Disney Plus, Netflix, Hulu and Amazon Prime have capitalized on national quarantine orders and made theatrical releases available on their respective platforms. The mechanisms behind such platforms are largely review-driven in an effort to compel users to select content based upon a social principle of coercion. Therefore, the research question for this exploratory data analysis is: Does the release year and runtime of a particular film selection have a significant impact on how it is rated and, therefore, promoted on Disney Plus, Hulu, Netflix and Amazon Prime? Ultimately, this project was unable to reject the null hypothesis of the statistical question, as a correlation plot and ensuing linear regression model revealed little to no significant correlation between the numeric explanatory variables ‘Year’ and ‘IMDB’ as well as ‘Year’ and ‘Runtime.’ However, a correlation heat plot suggested that the correlation score among the aforementioned variables is 0.40. The other variables examined in this analysis: ‘Rotten Tomatoes’ and ‘Netflix’, displayed moderate correlations. A histogram plotting the value count of IMDB concluded that the majority of content earned middling IMDB scores of 6-7, with a mean value of 6.4. IMDB values were plotted on a scatter plot to test and visualize the hypothetical causal relationship of year vs. IMDB. A distribution plot revealed that the data demonstrated a normal distribution. The probability mass function was utilized to supplement this analysis to determine the probability of obtaining an IMDB score above 5. The mean Rotten Tomatoes score was 61 percent ‘fresh.’ The content across platforms sought to appeal largely to mature audiences with the majority share of content falling within the age group of 18 and older. The minimum runtime for content was 11 minutes while the maximum runtime for streaming content in the data set was 260 minutes, with a mean runtime of 102 minutes, suggesting that the majority of consumers preferred feature-length content that was under two hours in length. A PMF was conducted and plotted to determine the probability of a movie running over 100 minutes. A Cumulative distribution function (CDF) compared the expected vs. observed values of the year variable. Outlying values occurred primarily in the runtime column. For instance, in this column, a film had a purported runtime of 1200 minutes, or 20 hours. To avoid skewing the data, outliers were left intact, which may have resulted in inflated descriptive statistics like mean and max/min values. Since regression models only accommodate numeric variables, this analysis likely missed trends that could have been uncovered among categorical variables, missing insights such as ‘Top performing titles’ or ‘Most popular actors’, which a target audience of digital content production executives could leverage to gain valuable perspectives on their subscriber bases in order to feed their respective machine learning algorithms.

The variables supplied in this data set all concerned the content and not the audience. To build a more comprehensive linear model and, consequently, produce a more insightful overall narrative, it would have been helpful to have access to a data set consisting of basic demographic insights such as the gender, age and location of individual users. Initially, I made incorrect assumptions about the correlation levels of the variables and the nature of the relationship between the variables (i.e. whether they were predictors/explanatory variables, and to what degree they impact the outcome of the experiment). I struggled for a while with converting column types with special characters like percentages signs, specifically for the Rotten Tomatoes column. I faced challenges that were rooted, primarily, in a misunderstanding of computational logic as it related to the implementation of Python functions for the purpose of statistical analysis. In reexamining some of the *Think Stats* lessons, I noticed that Downey produced his own libraries and pre-defined functions to aid in the calculation of PMF and CDF, among other calculations. However, since Thinkplot is not a universally applicable library, I struggled to convert Downey’s principles and processes to other libraries like Numpy, Scipy and Pandas. Additionally, in seeking out solutions to errors, I struggled with phrasing my inquiries and narrowing them down to one of the aforementioned libraries. Fundamentally, I am still struggling with the process involved in constructing a linear model, as this process is considerably more involved than the linear regression functionality in R. Overall, while I believe I am beginning to understand the EDA process and have improved upon my Python coding ability over the semester, I am not entirely confident in the outcome of my analysis and am working to improve my statistical analysis skill set using supplemental resources.