Term Project for: DSC530-T301 Data Exploration and Analysis

Jerica Tripp

August 14, 2021

For my analysis, I attempted to answer the question “Do US Movies with Longer Running Times Tend to get Higher Ratings on IMDB?”. The outcome of the regression analysis was that yes, there is a positive correlation between higher ratings and longer film runtimes. For every increase of 10 minutes to runtime, the average rating tends to go up one integer value. I conducted tests on the slope and was able to determine that the slope of the regression line was not likely due to chance and thus I was able to reject the null hypothesis.

For this analysis, I chose to focus on films released only in the US and with a release date between 2010 and 2020, inclusive. I also chose to not leverage variables such as language or genre. If I were to do this analysis again, I would like to see how the linear regression compares for the different genres and across languages and regions. Making the choice to only focus on 2010 to 2020 may have skewed the results. I maybe should have chosen to exclude years based on identifying outliers programmatically.

I faced challenges with importing the data from a tsv and removing characters like ‘\N’. I really felt the Pareto 80/20 rule was in full effect for this project. It took a while to get the data cleaned up and in the format I wanted, but after that it was relatively straight forward to conduct the analysis.

For the analysis, I had to do quite a lot of research to understand what the slope meant for the regression line. I had trouble identifying the relationship by units of the regression slope. I believe I made the correct assumption by using the smallest unit of average rating (0.1) and interpreting the slope to be 0.89 minutes. I was then able to convert this to an average rating of 1 and a runtime of 10 minutes to better describe the relationship identified with the regression line.