Colin Michael

DSC 530: Final Project

June 4th, 2022

I have always been passionate about movies. I love the art of storytelling and immersive experiences movies create. Successful movies capture a wide variety of human emotions and present them in an entertaining fashion. I wanted to explore what makes a successful movie using data. I gathered a dataset of 1,700+ movies from IMDB, which is an online repository of movie information and reviews. User IMDB rating (1-10) was the main dependent variable in my analysis. The dataset included the following data elements I used as independent variables: year the moving was released, length of the movie in minutes, number of reviews submitted, budget, mpaa rating, and genre.

My data exploration and analysis produced some great insights. One key insight I found was that comedy movies tended to score lower than non-comedy movies. I found this particularly intriguing because of how popular I perceived comedy movies were in pop-culture. I believe my biggest discovery resulted from running an OLS regression. My regression found a positive relationship between budget and number of votes to IMDB ratings, with a R^2 of .247, meaning 24.7% of the variation in IMDB ratings can be explained by budget and number of votes. This suggests that bigger and more popular movies result in higher IMDB ratings.

In my analysis I feel like I missed the impact that actors and directors have on movie outcomes. As a movie fanatic, I anecdotally know that certain actors and directors produce high quality movies compared to lesser-known creators. In a future study, I would like to pull in variables to show if a movie included any award-winning actors or directors.

The greatest challenge I faced in this analysis was determining the size and scope of my dataset. I wanted to capture as many movies as possible to increase my sample size but struggled with determining what constitutes a movie. IMDB has dubious requirements on their definition of what a movie is. In theory, anything with a page on IMDB can be considered a movie, but I wanted to focus the scope of my analysis on notable works. I applied a filter in the original dataset requiring that a movie had at least 10 user-submitted reviews to be included in my analysis. I did this to weed out potentially illegitimate movies. Overall, I found my analysis on movie outcomes to be extremely informative and enlightening.