Course Project

Caitie Schrotberger

Bellevue University

DSC 530: Data Exploration and Analytics

Course Project

For this project my hypothesis is that the most profitable movies come only from those sponsored by popular production companies. To evaluate this, I reviewed a data set from The Movie Database and compared budget, revenue, and Profit Rate (budget/revenue) between the top 10 production companies (As determined by the top 10 most frequent production companies in the databased. This analysis was done in excel outside of the programming.) compared to those that were not.

Based on the results of my exploratory data analysis, those movies made by the top 10 have higher profit rates. Additionally, when doing a linear regression model using budget and revenue, where budget is the independent variable and revenue is the dependent variable, those movies produced by the top 10 have a model that starts which a higher revenue stream.

To improve this analysis, additional segmentation could have been done by movie genre and year of production. The genre of movie may be predictive as well. Could comedies and action movies have a different rate of return than a drama or horror movie? Also, the values provided for budget and revenue were not normalized for inflation so separating the analysis based on year or decade of production may provide additional insight.

The amount of money specifically spent on marketing out of the budget would have been a variable that could have helped the analysis. There are so many costs that go into movie creation, knowing the amount of it spent on marketing or what percentage of the budget was spent on marketing could have a huge influence, especially in today’s modern environment with social media.

Having looked at the data in detail through excel to drive the direction of my analysis, I assumed that the data was accurate. I did notice that sometimes the budget or revenue amount was significantly different than the other. For example it would provide a budget of 1,000,000 but a revenue of 12. I looked some of these up online to find that the revenue was really 12,000,000. I left the values as is but could have done additional data cleansing to make the results more accurate.

The biggest challenge for me was using Jupyter as the tool for the data exploration part of this project. I’m much more familiar with SQL and could have easily determined how to code what I needed but it took more work to review and update the data prior to the statistical summary and analysis. Jupyter was much easier to do the statistical analysis such as PMF, CDF, correlation, and regression.