For my second Metis project, I wanted to predict the gross domestic total revenue of films using historical data.

To collect the necessary data, I web scrapped a website call boxofficemojo.com for film data from 1980 through 2016. Through the scrapping process, I collected roughly 14,500 observations. The information collected ranged from the budget of the film to the actors and producers involved in the film.

Once the data was scrapped (a 14 hour process) I started data cleaning and formatting for analysis. I chose to drop all production budget data, as the information was only available for 4,000 of the 14,500 observations. Also I adjusted all currency related variable for inflation. The value of $1 in 1980 was roughly $2.60 in 2016, so inflation adjustment was a necessity. Finally I dropped any observations that didn’t include missing data. The final dataset contained roughly 13,000 observations.

With the dataset properly formatted, I separatee my dataset into a train and test dataset and began exploratory data analysis. The process began by identifying the correlation for every independent variable relative to gross domestic total. Of the thousands of independent variables, I identified 38 variables with a correlation above 0.10. Opening weekend revenue had the highest correlation to gross domestic total with a value of 0.875.

Using the information from the correlation analysis, I ran an ordinary least squares regression on my train data. The result was an r^2 value of 0.779. Next I ran the same model on my test dataset to determine the efficiency of the model. The model resulted in an r^2 value of 0.774. This suggested the model was a good fit for the dataset, but I wanted to double check. To verify the model, I ran a 5-fold cross validation with the model. The r^2 value for the relative folds ranged anywhere from 0.74 to 0.81. This verified the accuracy and fit of the model.