This project seeks to create a model that can accurately estimate the gross domestic box office earnings using the production budget of the movie and other descriptive data. Another question that I wanted to answer was whether or not sequels and remakes make more money than IP’s that are seeing the silver screen for the first time.



There appears to be a roughly positive relationship between production budget and earnings, but there is also a high variance. Production budget itself may not prove to be a great indicator for box-office earnings.



This plot shows that maximum number of theaters that were showing the movie at any one time. The relationship here may not be strictly linear. A transformation may be called for here.



From this box-plot, I would say that an indicator variable for whether or not a movie is animated would not make the model more accurate.



There does appear to be a difference in means for whether a movie is a sequel or not, but the variances in the two groups are also different.



There also may be a difference in means depending on if a movie is a remake or not. The variances seem closer between the two, but the non-remake group looks to have many outliers.



Here, the index is based on the movies rank. The 200th movie in the dataset is the 200th highest grossing movie in the domestic market. With that in mind, we can see a clear trend in the residuals as they start to overestimate as the rank goes up.



This is a plot to find movies with peculiar x values. To me, this would indicate that I need to clean up my data. I created a script that scraped the data from the-numbers.com. A lot of movies don’t have data for their production budget



This plot shows cook’s distance. This would indicate that there are some data points that are highly influential on my model, and I should consider removing them from the dataset.



Here we have the standardized residuals plotted against their respective phat values. This is another way to pick out highly influential points in our dataset.

Next steps: Firstly, I need to clean up my data and remove any movies that I do not have complete data for. I can also check for interaction between indicator variables and continuous variables to see if there are different slopes for different kinds of movies. I could also include the year the movie was made to see if there is any correlation there. Lastly, I could use data on the inflation adjusted gross domestic box office earnings instead of the unadjusted data.