Movie Analytics

Daniel Chang, Luis Jovel

How do genre, director, or budget affect the average rating of movies? Can we ultimately predict the average rating based on these predictors? Studios could use the results to prioritize movies production. The public could also use it to gain insights into a movie.

**Data Source:**

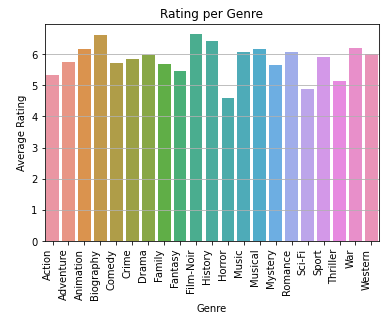
We obtained IMDb’s data from Kaggle. This dataset was one of the few datasets that included budgets. It had 85,855 observation and 22 variables. We filtered the data to movies released in the USA, and ignored movies that didn’t have a country specified. This gave a dataset with 28,511 rows.

**Data Preparation:**

The data was fairly clean, but we had to make some minor adjustments to budget. We also had to worry to about some movies that had multiple genres, directors and country. So, we had to split them to make sure everyone was included.

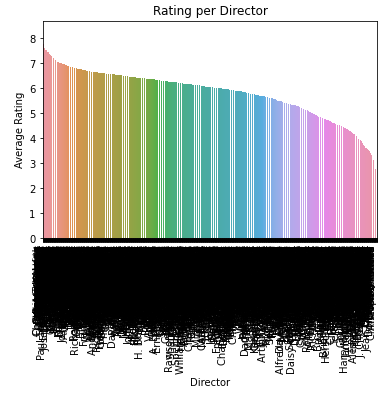
**Exploratory Data Analysis:**

We first took a look at the ratings per genre. After we averaged the movie ratings into groups of genres, we found several genres with only one movie. After removing the genres with only one movie, we plotted the following bar plot.



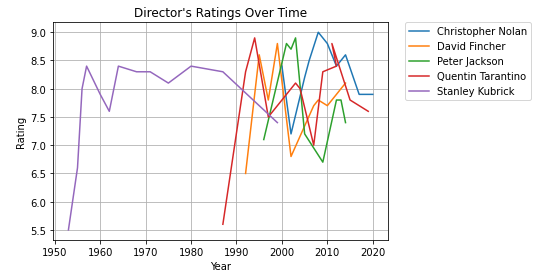
From the above plot, we can quickly see that most of the genres are fairly even, with exception to Horrors doing noticeably worse on average.

Next up we looked at directors. After taking a quick look at all the directors we found a few outlier directors that had almost perfect ratings, but only had one or two movies. We filtered out directors that had less than 5 movies to their name and sorted them in descending order. We removed them because they only created noise in our dataset and we want to be as unbiased as possible. Taking a look at the data frame, we narrowed the directors to a measly 1516 directors. Trying to plot all of them resulted into a pretty good-looking mess.



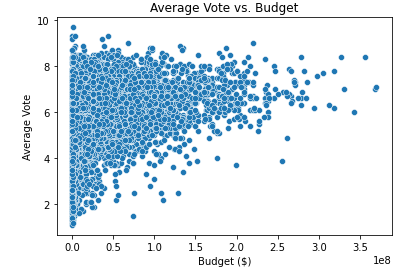
From the above mess, we took a look at the top 5 directors, and plotted their average movie ratings over their career. It does give a nice-looking picture though!

Back to a serious note, even though the above plot was awful. It did lead to some additional questions we wanted to dive into more. We started to see a trend with directors that could imply they had a big impact on what how the movie would fair in terms of average rating.

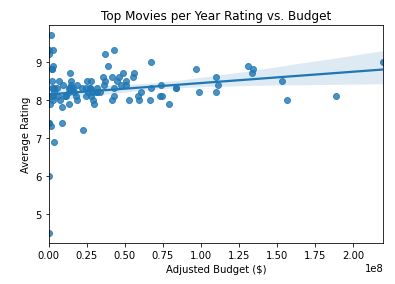


Christopher Nolan is the top-rated director according to our IMDb dataset, but had a shorter career compared to the rest of the top 5. The director’s ratings appear that their first couple of movies aren’t that great, but then they have fairly consistent results.

To see if there was a correlation between budget and movie ratings. We found a lot of movies were missing budget data so we removed them since we weren’t going to be able to find the budget. We then had to adjust for inflation using the movie release year. We used the library cpi to adjust the budget for inflation using Consumer Pricing Index data. However, the CPI data only consisted of data from 1913 to 2019, so we had to drop movies released prior to 1913, and assumed 2019 and 2020 budget to be close enough. Plotting the all the movies vs their budget resulted in the below graph.



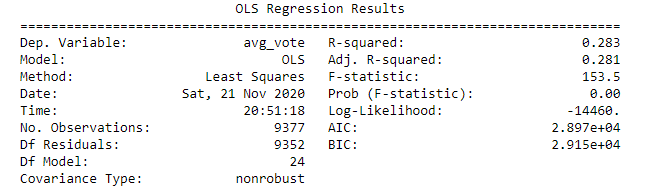
Quickly looking at the plot above, we see several movies being made for somewhat unreasonably small budgets. Like a 90-minute film with a budget of $3. We see that there isn’t a general correlation between the budget and the average rating. We thought to repeat the plot using the best movie per year from 1913 to 2020.



The regression line looks fairly flat, so we quickly computed the correlation between budget and average rating for a correlation of 0.283.

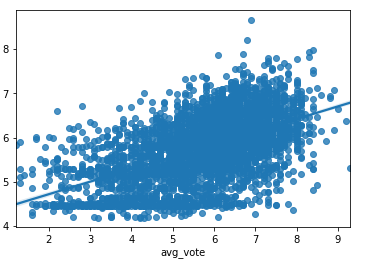
**Prediction:**

We first looked at how adjusted budget and genre would affect the average rating. In order to solve this, we created a linear regression model. We dropped the null values, created dummy variables for genre and did a 70-30 train test split. Once all the data preprocessing was completed, we implemented the linear regression model on the training data. We first how well the model the training data. We saw the r-squared value was 0.283, which says that our model only explained about 28% of the variance. It is not great but it is something we can try the model out on the test data.



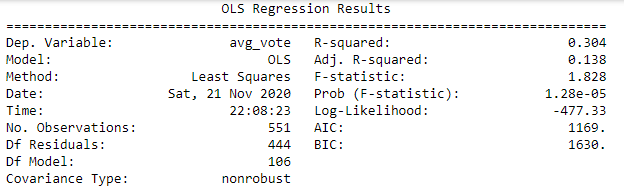
These are our results of how well our model fits the training data. It gives you some nice statistics to assess to your model further.

First, we get our predicted values from using our test data on the model generated from the training data. Afterwards we evaluate how it does compare to our actual values by calculating the root mean squared error. We get a value of 1.279 which means on average we are off by 1.279 rating scores.

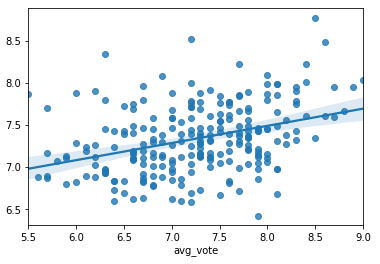


This plot is an illustration of the predicted values versus the actual average rating in our test data.

We wanted to see if we can improve our model more by adding in the top 100 directors and if it lowers our root square mean error. Our process was fairly similar to before we just added some more dummy variables for directors. Here is how well it fitted on our training data:



This model does not fit as well, but we have less data in this case so the variation would be bigger in this case. We know apply the same prediction method we did before root mean square error .459, which is definitely an improvement.



As we can see there is a significant smaller number of movies that are shown because there are less directors. Even though the model seems better we might be adding too much bias.

**Next Steps:**

In the future we might want to look at ways to reduce the dimensions by using principal component analysis or singular value decomposition, in order to see if we can use all the directors. The problem we have if we include all directors, we would have a matrix that is too large and would take days to run an algorithm. Also, we might try to include some other variables that are present in our dataset.