Part Two Report

For the second part of our DATA-151 project, we obtained data relevant to our research, thoroughly cleaned the data, and tested various predictor variables to determine their effects on movie earnings. This analysis left us prepared to create and test a model to predict how much money a movie would earn based on various characteristics.

We found and used a dataset containing the top 1000 ranked movies by IMDB rating on Kaggle. This dataset contains information about movies spanning from 1925 to 2019. It also includes a number of characteristics about each of these movies such as the release year, runtime, genres, director, and domestic earnings. We believed that this dataset gave a good overview of the film industry and provided sufficient information to determine what factors are most important in predicting how much money a movie will earn.

For our purposes of finding what factors are most important in determining a movie's earnings, we needed to clean our dataset in order to make it easier to perform data analysis, visualization, and modeling later on. The first thing we did was drop all rows without any recorded domestic earnings. As domestic earnings was our response variable, there was no way to use any of these data points. There were also a number of rows without a Metascore. Metascores are a rating for movies given by professional critics based on the quality of the film. After looking at the correlations between several other variables and Metascore, we found that there was no reliable way to infer a film's Metascore. Because of this, we decided to drop all rows without a Metascore. This left our dataset with 750 movies. We dropped the Poster Link, Certificate, and Overview columns, as they did not have any value in our project. The Poster Link and Overview columns had unique values for each data point, while the Certificate column was not on a common scale. We split the Genre column into three columns: Genre, Genre2, and Genre3. This change allowed us to analyze the effect of genre easily. We also removed unnecessary commas from the gross values and the "min" from the runtime column. We decided to delete the Star1, Star2, Star3, and Star4 columns. There were too many unique names for this to be a good predictor variable. Furthermore, plenty of actors appeared in multiple columns, which would have significantly complicated our data analysis. To make sure that inflation was not having an undue influence on our non-release year variables, we created a new column with the inflation-adjusted domestic gross for each movie, indexed to 1983 dollars, using a dataset on inflation from the Federal Reserve Bank of Minneapolis. Finally, we converted any numerical statistics that were in the form of strings in our data to be integers. After this, our data was fully clean and ready for analysis.

The first bit of data analysis that we performed was to examine the correlation between the release year of the movie and its non inflation-adjusted domestic gross. We expected this correlation to be quite high, however, the R^2 value was only .06. When creating a regression-line, the slope of the line was positive, which was in line with common sense and our expectations. It seems evident that movies released with a more inflated-currency would make more nominal

money. The regression-line was statistically significant despite the low R^2 value. Following this, we calculated the correlation coefficient and R^2 value for each of our quantitative predictor variables when used to predict the inflation-adjusted gross of a movie. Since we did not conduct a multiple linear regression, we did not control for the effect of all other variables. We were simply trying to see which variables seemed to correlate the most with earnings at a quick glance.

	Release Year vs Gross	Release Year vs Adjusted Gross	IMDB Rating vs Adjusted Gross	Number of Votes vs Adjusted Gross	Runtime vs Adjusted Gross	Runtime vs Adjusted Gross
Correlation Coefficient	0.236	-0.179	0.153	0.327	0.240	0.072
R^2 Value	0.055	0.031	0.023	0.106	0.057	0.0052

To our surprise, many of the predictor variables did not seem to have a large impact on domestic earnings. The only variable with an R^2 value greater than .10 was the number of votes on a film's IMDB rating.

After conducting exploratory analysis on our quantitative predictor variables, we examined whether certain characteristics for our categorical predictor variables seemed to have an effect on earnings. First, we tried to determine whether movies directed by famous directors tended to make more money than those not directed by famous directors. To define what a famous director is, we created a list of all directors who had directed at least six movies in our dataset. Using this definition of a famous director, there were 115 films directed by famous directors and 635 films made by non-famous directors. The average inflation-adjusted gross of a film made by a famous director was $58,283,529.02, while the average inflation-adjusted gross of a film made by a non-famous director was $55,540,146.95. This was not a large difference, which surprised us. To determine whether the inclusion or exclusion of a single or small number of directors was skewing our results, we decided to do the same analysis using a definition of famous director which included any director with five or more films in the dataset. The number of famous directors increased from 15 to 23. Using this new definition, there were 155 films made by famous directors compared to 595 films made by non-famous directors. The average inflation-adjusted gross of a film made by a famous director was $69,344,856.78, while the average inflation-adjusted gross of a film made by a non-famous director was $52,474,195.54. This difference made much more intuitive sense, and suggested that whether a movie was made by a famous director was an important factor in its earnings. One possible reason for the large difference between these two results is the inclusion of James Cameron, a notable highly commercially successful director, as a famous director in only the second definition.

We then continued to analyze the effect of genre upon gross earnings in an attempt to see if there was much variability based on genre. Our thoughts were that there would be a reasonable variation in gross earnings based on genre, but we were very unsure as to how much of a difference it would be. As our dataset in Excel had each movie containing up to three genres, we had to split up the genres into three different columns, and then had to search through each column in order to create dummy dataframes, which we later merged, for each individual genre. After finding our twenty-one individual variables, we then found the sum and divided that by the count in order to get the average for each genre's dataframe. Our data, thankfully, did show some significant differences in gross earning average between the genres. Family was the highest grossing at about 105 million dollars, while film-noir was the lowest at just under 20 million dollars. We then used the averages to create a bar graph containing all of the averages. Following all of this data analysis, we were ready to create a model.