Final Report

IMDB & Box Office Movie Rating

Blue Bai, Chan Khine, James Provanzano, Tess Serrato

Table of Content

1. Background

Our team is composed of four avid movie watchers. Since the start of the pandemic at the beginning of 2020, we found ourselves trapped inside with few things to keep us entertained. This led to us spending more time watching movies than ever before. However, due to the excess of movie streaming services available these days such as Netflix, Hulu, HBO Max, and more it is almost impossible to make a decision on what movie to watch. Since then, we have seen theaters begin to reopen, leading us to wonder what direction movie directors might go with new movies. Hence, our team is interested in learning what directly impacts the box office performance of a movie. This is useful to determine which films moviegoers prefer and why. This would benefit both movie studios and theaters, allowing them to better tailor their movies to the audience's preferences.

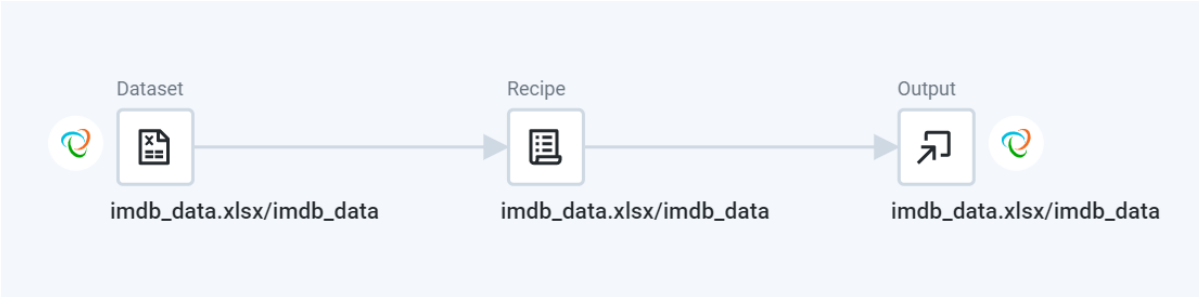
2. Datasource

We first used a dataset from IMDb for the years 2000 to 2019. We chose to use this dataset because IMDb is a popular source for users to explore more about movies. Moreover, IMDb only offers user reviews which is one of our team's focuses. This is because critics might have a more objective review based on soundtracks, cinematography, etc., whereas normal users might have more subjective opinions. Therefore, analyzing normal user reviews might give us more insight into what attracts an everyday movie viewer. We downloaded a second data set from Box Office Mojo for the same time period of 2000 to 2019. We decided to use Box Office Mojo because it is a reliable source owned by Amazon, that provides information on movies such as the studio in which it was filmed, the box office year, the box office rank, and the box office revenue of the movie both domestically and internationally. Analyzing the box office revenue of a movie might give us more insight into which movies attract a larger audience versus movies which have a lower box office gross revenue.

After much consideration, we concluded that the two datasets are credible for our analysis. On the one hand, most data on IMDb such as ratings and synopsis are added by users. So, there might be user bias, especially in movie ratings. For instance, *Avengers: End Game* might be popular because it is part of the Marvel franchise rather than being an Action movie. On the other hand, this user bias might give us more insight into what features moviegoers are interested in. Similarly with Box Office Mojo, there might be movies with higher or lower gross revenue due to variables such as the marketing behind the movie, or the level of movie competition at the time of the release. Furthermore, it does not take into consideration that movies can perform better years down the line or movies that are available outside of the box office. However, with all this taken into consideration, being able to determine which movies viewers are willing to go see in the theater could give us a more in-depth perception of the movies viewers prefer. Therefore, these two datasets are quite reliable for our group's analysis.

3. Data Preparation

Upon viewing the data, we discovered that many categories had missing or invalid data. The first step was to delete irrelevant categories. For this, we used Microsoft Excel to delete columns titled: endYear, editor, producer, cinematographer, composer, actor, actress, writer, and production_design. From here, the row titled "startYear" was renamed to "Year Released". However, there was still a lot of missing data. The photo below shows the flow and steps taken in Trifacta Wrangler to clean the remaining missing data.



1 **Delete rows** where runtimeMinutes is '\\N'

2 **Delete rows** where genres is '\\N'

3 **Rename** isAdult to 'Adult'

4 **Remove symbols** from director

5 **Delete rows** where ISMISSING([director])

6 **Delete** tconst

7 **Change** Adult **type** to Boolean

The missing values in runtimeMinutes and genres were specified with "\N". To remove the observations with this value, we delete rows with that specific value. The name for the variable confirming if the movie is rated for adults, was changed from "isAdult" to "Adult". The values in the director column were surrounded by square brackets, so we removed these symbols. The observations missing information about the director are left blank when the brackets were removed, so all observations with missing director values were deleted. The tconst column was deleted because it does not have relevance to our desired dataset. Lastly, on Trifacta Wrangler, the Adult values are changed to binary categories of 0 and 1, 1 indicating an adult film.

To ensure we had enough data to make predictions, we added a second dataset to the IMDb dataset. We chose to use the dataset from box office mojo because it is one of the leading box office reporting services that tracks box office receipts both domestically and internationally.

This dataset has information about the movie's box office profit and ranking which can be a target variable for our project. Since the box office revenue and ranking can help show the popularity of a movie we decided that this dataset would be useful in attention to the data from IMDb. The data sets were joined by title and year to accurately match the movies with similar names which were released in different years. This provided us with 2,708 observations with no missing data.

Our joint dataset from IMDb and Box Office Mojo turned out to be very strong. We were accurately able to combine the two datasets into one large dataset. Furthermore, through our wrangling and cleaning process, we were able to either remove or fix all incorrect and null variables. This gave us a final dataset which included variables such as the movie titles, the year that the movie was released, the runtime length of the movie, the average rating, the box office rank, the number of votes the movie received, the worldwide gross revenue, and lastly the domestic gross and international gross revenue. Our final dataset had more than 2,708 observations from a period of 19 years with zero errors. Providing us with variables needed to make accurate predictions on the popularity of a movie. We did believe that genres were going to play an important role in making our predictions, however, due to the wide range of values under genre, it was too difficult to manually observe the impact of each genre. Our data is pre Covid19 and does not take into account the performance of any movies during or after the pandemic.

4. Descriptive Statistics and Data Visualization

4.1. Descriptive Statistics

```
 genresSplit.1      genresSplit.2       bo_year_rank      title              studio          worldwidegross
 Length:2708        Length:2708        Min.   :  1.0   Length:2708        Length:2708        Min.   :3.000e+03
 Class :character   Class :character   1st Qu.: 51.0   Class :character   Class :character   1st Qu.:1.250e+07
 Mode  :character   Mode  :character   Median :100.0   Mode  :character   Mode  :character   Median :4.510e+07
                                       Mean   :117.5                                         Mean   :1.065e+08
                                       3rd Qu.:164.0                                         3rd Qu.:1.170e+08
                                       Max.   :466.0                                         Max.   :2.779e+09
 domesticgross        domesticpct     overseasgross        overseas.pct       bo_year     primaryTitle           Adult
 Min.   :      300   Min.   :  0.00   Min.   :3.000e+02   Min.   :  0.00   Min.   :2000   Length:2708        Min.   :0
 1st Qu.:  4400000   1st Qu.: 33.70   1st Qu.:3.600e+06   1st Qu.: 28.88   1st Qu.:2004   Class :character   1st Qu.:0
 Median : 24200000   Median : 51.30   Median :1.840e+07   Median : 48.70   Median :2008   Mode  :character   Median :0
 Mean   : 46764927   Mean   : 51.49   Mean   :5.976e+07   Mean   : 48.51   Mean   :2009                      Mean   :0
 3rd Qu.: 58225000   3rd Qu.: 71.12   3rd Qu.:5.972e+07   3rd Qu.: 66.30   3rd Qu.:2012                      3rd Qu.:0
 Max.   :749800000   Max.   :100.00   Max.   :2.029e+09   Max.   :100.00   Max.   :2019                      Max.   :0
  YearReleased  runtimeMinutes    genres            director         averageRating      numVotes
 Min.   :2000   Min.   : 25    Length:2708        Length:2708        Min.   :1.900   Min.   :      7
 1st Qu.:2004   1st Qu.: 95    Class :character   Class :character   1st Qu.:5.800   1st Qu.:  16433
 Median :2008   Median :105    Mode  :character   Mode  :character   Median :6.500   Median :  51147
 Mean   :2009   Mean   :109                                          Mean   :6.371   Mean   : 106731
 3rd Qu.:2012   3rd Qu.:119                                          3rd Qu.:7.100   3rd Qu.: 126512
 Max.   :2019   Max.   :219                                          Max.   :9.000   Max.   :2099320
'data.frame':   2708 obs. of  19 variables:
```

Figure 4.1.1: Summary Statistics

The above figure is the summary statistics for the original data. There are 19 variables and 2,708 observations. We can observe that for averageRating the maximum number is 9, the minimum is 1.9 and the mean value is 6.37; for runtimeMinutes the maximum is 219, the minimum is 25 and the mean value is 109; for numVotes, the maximum number is 2,099,320, the minimum is 7 and the mean is 106,731.

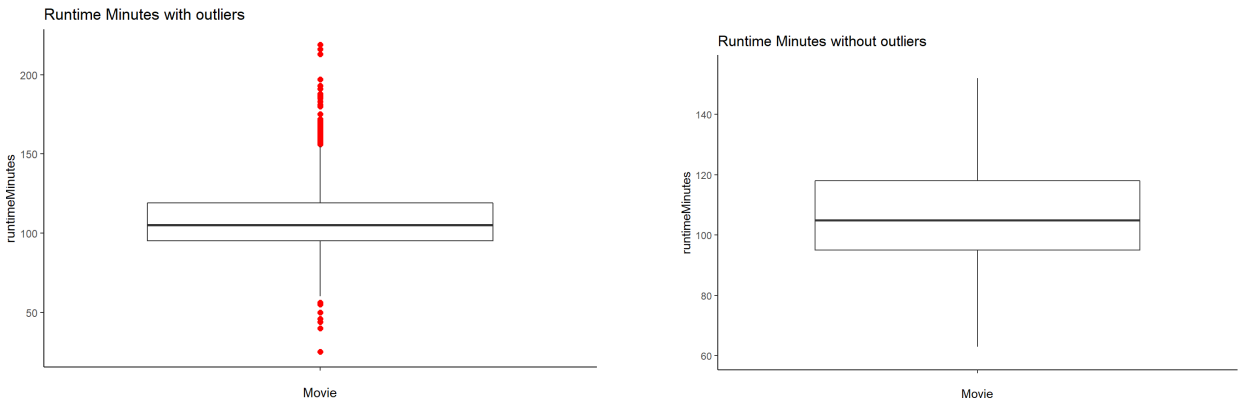


Figure 4.1.2: Box Plots of Runtime Minutes with and without Outliers

The box plot on the left which includes outliers shows Runtime Minutes on the Y-Axis and Movies on the X-Axis. As you can see the 86 outliers mainly lie below 50 minutes in run time or above 150 minutes in run time. The box plot on the right with Runtime Minutes on the Y-axis and Movies on the X-axis removes all 86 outliers. This produces a clean graph which shows that the majority of movies have a length of 90 to 120 minutes, with the most common length being 105 minutes.

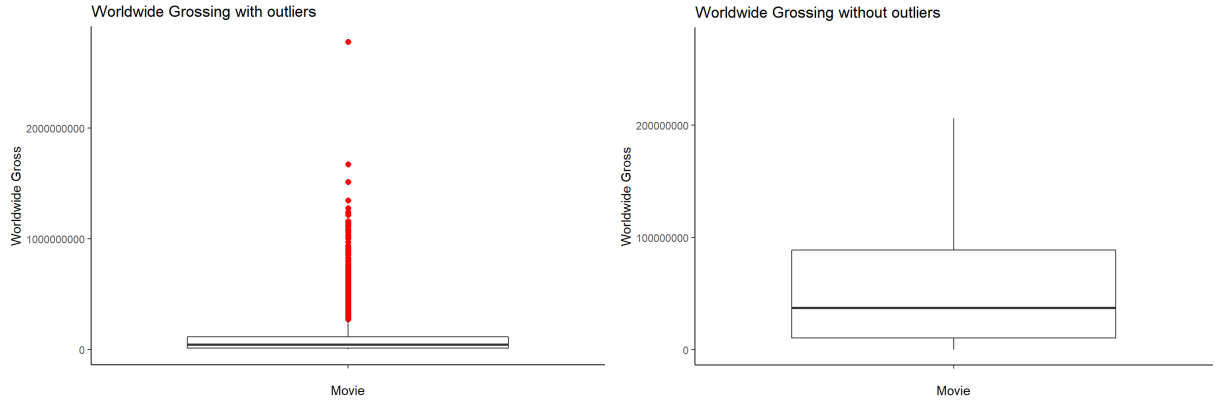


Figure 4.1.3: Box Plots of Worldwide Gross with and without Outliers

The box plot on the left shows Worldwide gross on the Y-axis and Movies on the X-Axis. It includes 267 outliers of movies that grossed well above the median of 40 million dollars. The box plot on the right shows Worldwide gross on the Y-axis and Movies on the X-Axis. Excluding

the 267 outliers, it shows the highest destiny of movies grossing between 10 million and 100 million dollars, with the median being about 40 million dollars.

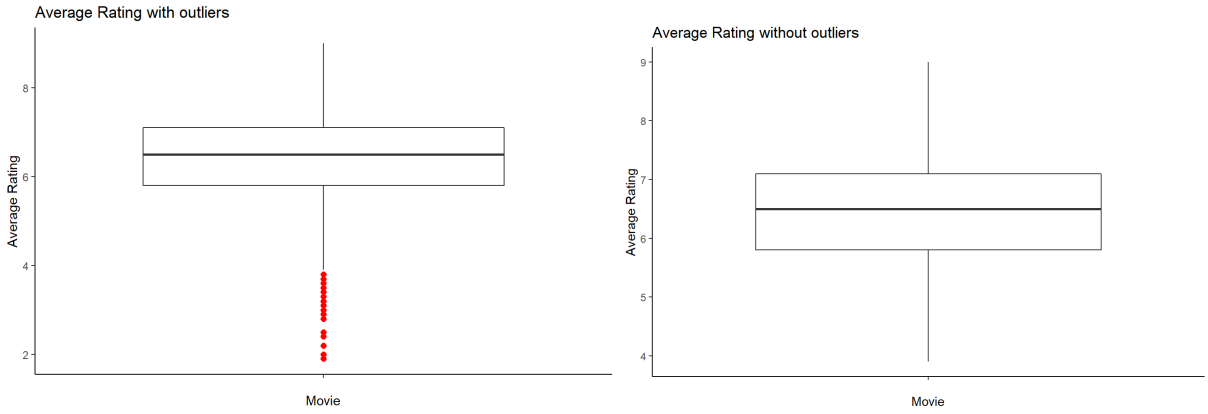


Figure 4.1.4: Box Plots of Average Rating with and without Outliers

This box plot on the left with Average Rating on the X-axis and Movies on the Y-axis includes 57 outliers that all fall below a rating of 4. This box plot on the right shows the Average rating of movies while excluding the 57 outliers. You can see that the most frequent average rating that a movie receives is 6.75. This makes sense since most movies fall between a rating between 6 and 7. This is interesting because on a scale of 0 through 10 you would expect the highest density of movies to be rated at the scale average of 5.

4.2. Exploratory Analysis

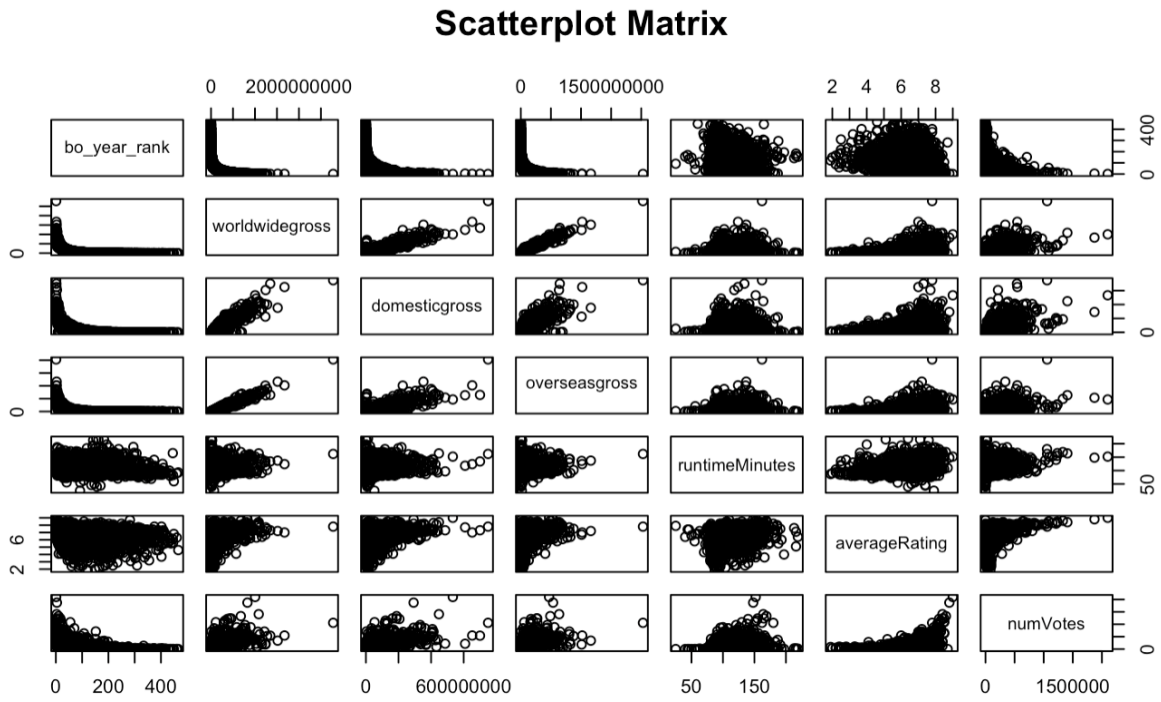


Figure 4.2: Scatterplot Matrix

In figure 3.2.1, we have a scatterplot matrix with seven different variables. The seven variables are; bo_year_rank, worldwidegross, domesticgross, overseasgross, runtimeMinutes, averageRating, and numVotes. As you can see above, each of the graphs shows a correlation between each of the variables. The graphs with the strongest correlations between variables are worldwidegross and domesticgross, overseasgross and domesticgross, and averageRating and numVotes. The correlation between worldwidegross, domesticgross, and overseasgross is not surprising since worldwidegross is the sum of domesticgross and overseasgross.

Worldwide gross and bo_year_rank scatterplot shows a strong curvilinear relationship. This may be reasonable since if a movie receives a high worldwide gross, it will rank higher at the box office. The plot comparing Worldwide Gross and runtime minutes has a high density in the space between 75 and 175 minutes. This is reasonable considering most movies are made within this range. When comparing worldwide gross and average rating, this graph shows a positive correlation which would indicate that higher grossing movies are more highly rated. The plot with Worldwide gross and numVotes shows an interesting pattern. However, it does have a positive correlation, so this would indicate movies getting more votes are popular, so they would have a higher gross.

Bo year rank and avg rating show a negative correlation, meaning the closer to the top box office rank, the higher the rating. However, there are many outliers within this plot.

AverageRating and numVotes show an interesting correlation that the more votes a movie receives the higher the average rating the movie has. This makes sense because if a movie is gathering a lot of votes it can be expected that the movie is being well received by the audience.

5. Clustering

5.1. Hierarchical cluster

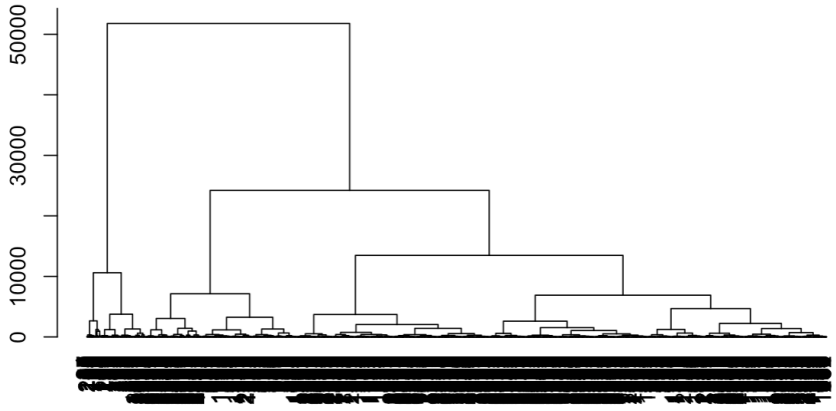


Figure 5.1.1: Ward.D Dendrogram Plot

We choose 6 variables as different features (worldwidegross, domesticgross, overseasgross, runtimeMinutes, averageRating, and numVotes) to initiate the hierarchical cluster. Then, we normalize the data and select the Euclidean method to calculate the Intra-Cluster distance. After trying all the linkage parameters, we select the "ward.D" linkage as the best parameter to calculate the inter-Cluster distance since it demonstrates the clearest and balanced cluster distribution.

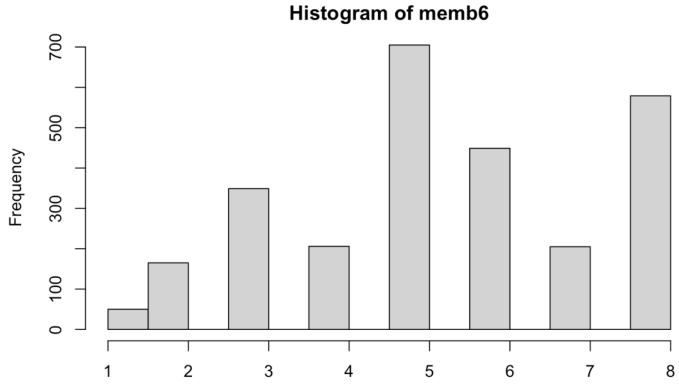


Figure 5.1.2: Histogram with k=8

The histogram above shows the number of observations allocated to each cluster. To keep the number of observations in each cluster balance, we cut the tree at a height of about 500 and received 8 clusters. From the histogram, we can observe that the number of observations in each cluster is comparably similar with k equal to 8.

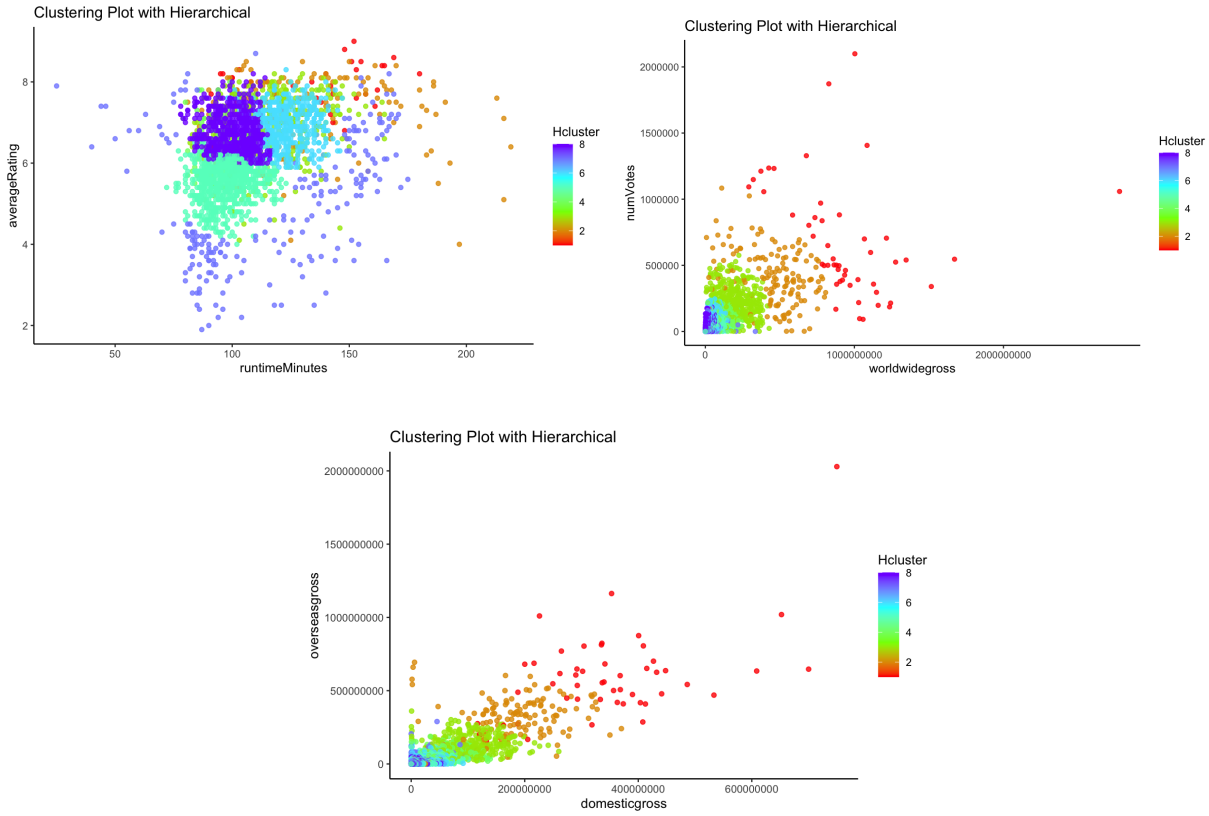


Figure 5.1.3: Scatter plots of Hierarchical clusters

To observe the characteristics of each cluster, we use the three scatter plots above and set 6 different features as X and Y axis separately. In the figure above, the red dots represent cluster 1, which has the highest average IMDb rating, medium runtime minutes, highest worldwide gross, highest number of votes, highest overseas gross, and highest domestic gross. The brown dots represent cluster 2, which has a high IMDb average rating, high runtime minutes, high worldwide gross, high number of votes, high domestic gross, and high oversea gross. The grass green dot represents cluster 3, which indicates the movies with high IMDb average rating, medium runtime minutes, medium number of votes, medium worldwide gross, medium overseas gross, and medium domestic gross. The brighter green dots represent cluster 4, which has a high IMDb average rating, low runtime minutes, low number of votes, low worldwide gross, low overseas gross, and low domestic gross. The cyan dots represent cluster 5, which has a medium IMDb average rating, low runtime minutes, low number of votes, low worldwide gross, low

overseas gross, and low domestic gross. The blue dots represent cluster 6, which indicates the movies with high IMDb average rating, medium runtime minutes, low number of votes, low worldwide gross, low overseas gross, and low domestic gross. The light purple dots represent cluster 7, which indicates the movies with low IMDb average rating, medium runtime minutes, low number of votes, low worldwide gross, low overseas gross, and low domestic gross. The dark purple dots represent cluster 8, which indicates the movies with high IMDb average rating, low runtime minutes, low number of votes, low worldwide gross, low overseas gross, and low domestic gross.

5.2. K-means Cluster

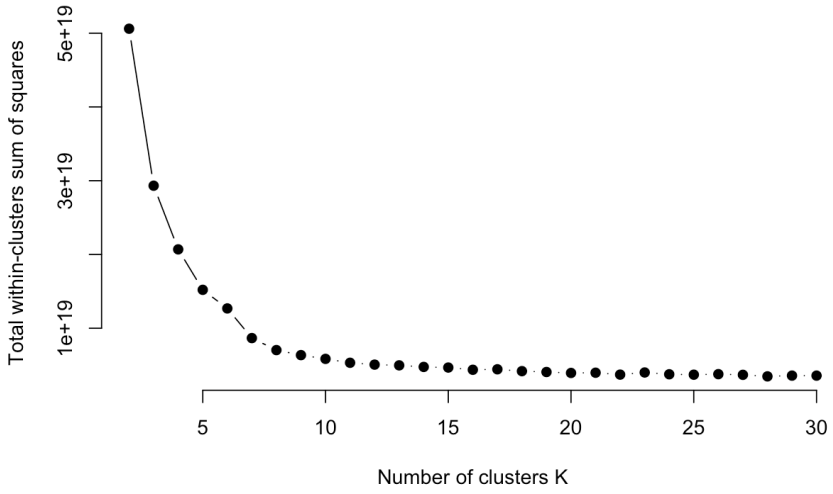


Figure 5.2.1: Elbow plot

The figure above is an elbow plot showing when k = 8, the total within-cluster sum gradually becomes flat. It can be seen as a bend (or “elbow”) at k = 8. This bend indicates that additional clusters beyond the eighth have little value. Thus, we pick the k = 8 for our model.

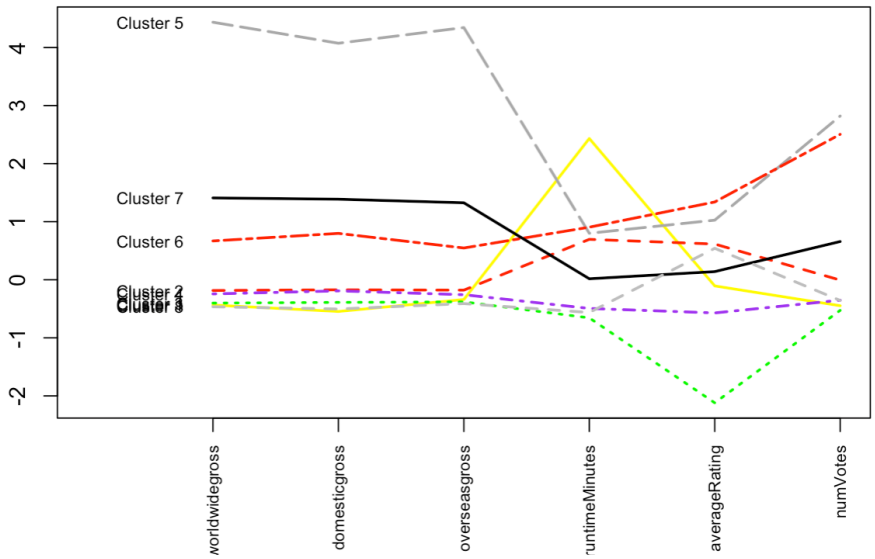


Figure 5.2.2: Centroid Plot

From the figure above, we can see the characterization of each cluster. The yellow line indicates cluster 1 with low worldwide gross, low domestic gross, low oversea gross, medium average rating, highest runtime minutes, and low runtime minutes. The bottom red line indicates cluster 2 which is low in worldwide gross, domestic gross, oversea gross, and high in average rating, runtime minutes, and medium in number of votes. The green line indicates cluster 3 with low worldwide gross, low domestic gross, low oversea gross, lowest average rating, low runtime minutes, and low number of votes. The purple line indicates cluster 4, which is low in all of these 6 features. The upper gray line indicates cluster 5, which is the highest in worldwide gross, domestic gross, oversea gross, high in average rating, runtime minutes, and numbers of votes. The upper red line indicated cluster 6 which is medium in worldwide gross, domestic gross, overseas gross, and has high runtime minutes, average rating, and numbers of votes. The black line indicates cluster 7 which has a high worldwide gross, domestic gross, oversea gross, and medium runtime minutes, average rating, and numbers of votes. The bottom gray line indicates cluster 8 with low worldwide gross, low domestic gross, low overseas gross, low run-time minutes, medium average rating, and low number of votes.

6. Regression Analysis

6.1. Target Variable Description

Our target variable is Worldwide Gross (worldwidegross). We think this is a good target variable because it can provide guidance for the movie industry to better understand what type of movies are earning the most. The explanatory variables are bo_year_rank, runtimeMinutes, averageRating, and numVotes. Using these explanatory variables, we will be able to predict a movie's earnings based on its runtime, average rating on IMDb, and the number of votes on

IMDb. An accurate prediction of worldwide gross matters because the production company will be able to estimate the profit the movie will earn based on the data they have within the explanatory variables. Then, the company will be able to gauge how much advertising they need to promote the movie. They will also be able to decide if they want to make a sequel or prequel of the movie or create merchandise relating to the movie.

6.2. Initial Hypothesis

We theorize some initial hypotheses to guide us with the regression analysis. We hypothesize that runtime minutes will have a negative impact on worldwide gross since people would not want to watch long movies in theatres unless it is very popular. We also hypothesize that the average rating and number of votes will have a positive impact on worldwide gross. Our assumption is that if a movie has a high average rating, it means that people who watched the movie enjoy it. Similarly, if a movie has a high number of votes, we can say that the movie is popular. There are also a few questions that guide us in our regression analysis: What are the factors that influence worldwide gross?; Will adding clusters as features improve the accuracy of the models?

6.3. Data Preparation

We add number of votes squared as higher polynomial variables since we think there is a non-linear relationship between worldwide gross and number of votes. We also add an interaction term named bo_and_num, which is a product of box office rank and number of votes. This is because we are assuming that everyone who watches the movie in theatres will write a review afterward. These reviews will then influence more people to watch the movie which will affect worldwide gross. Thus, our explanatory variables are runtime minutes, average rating, number of votes, number of votes squared, and the interaction term between box office rank and number of votes.

Since we plan to compare and contrast the models we build in Rstudio to DataRobot, we also remove the outliers that dataRobot suggested until it reaches 10% of the data. Thus, we remove 270 rows out of 2,708 rows from the dataset. The data is then partitioned to training and validation by 60:40 for the regression analysis. After the iterative process of running different regression analyses, we discover that the multiple regression model gives us the lowest RMSE values. Thus, we focus on multiple regression models with cluster and non-cluster data.

6.4. Manual Multiple Regression without Clustering

```
Call:
lm(formula = worldwidegross ~ ., data = train.t)

Residuals:
       Min         1Q     Median         3Q        Max
-103047683  -18512299   -8393404    6761555  303028585

Coefficients:
                     Estimate      Std. Error t value             Pr(>|t|)
(Intercept)    46488196.378435  8363975.488961   5.558         0.0000000324 ***
runtimeMinutes  -130551.879994    58340.045087  -2.238               0.0254 *
averageRating   -796809.206350  1184467.277271  -0.673               0.5012
numVotes           2250.560402       49.064144  45.870 < 0.0000000000000002 ***
numVotes_sq          -0.003794        0.000164 -23.139 < 0.0000000000000002 ***
bo_and_num          -14.792426        0.333930 -44.298 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40210000 on 1456 degrees of freedom
Multiple R-squared:  0.7702,    Adjusted R-squared:  0.7694
F-statistic: 976.2 on 5 and 1456 DF,  p-value: < 0.00000000000000022

runtimeMinutes  averageRating       numVotes    numVotes_sq     bo_and_num
      1.088631       1.236915      12.005578       9.594647       1.976759


     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
     3000  10650000  38800000  68740872  93950000 462200000
             ME     RMSE      MAE   MPE MAPE
Test set 1835684 40504048 24639849 -2512 3033
```

Figure 6.4.1: Model Result

The best model before we put the clustering results as predictors is the result from the multiple regression model as shown in the above figure. This is because after going through the iterative process of checking RMSE for different regression models, we concluded that the multiple regression model gives the lowest RMSE. We decided to use the lowest RMSE as our best model since simply adding more predictors can increase the adjusted R-squared.

The RMSE value is 40,504,048 and the adjusted R-squared is 0.769. When checking the variance inflation factor, we can see that except for numVotes and numVotes_sq, the variance inflation factor for all variables is less than 5. We ignored the VIF values for the variables and their squared terms values as they will have collinearity since they are the squared terms.

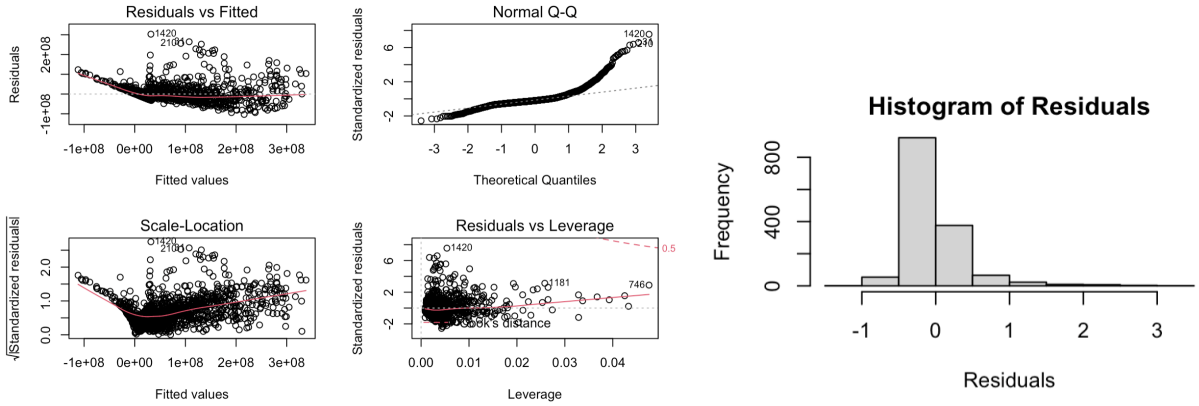


Figure 6.4.2: Diagnostic Plots and Histogram of Residuals

As we can see in the above plots, the Residuals vs. Fitted plot in the upper left does not look quite randomly distributed. The Normal Q-Q plot in the upper middle does not track closely to the diagonal line for the right side. The histogram of residuals in the right does not look like a bell-shaped curve. It is skewed to the right. Hence, we added hierarchical and k-mean clusters to improve our model.

6.5. Manual Multiple Regression with Clustering

```
Call:
lm(formula = worldwidegross ~ ., data = train.tC)

Residuals:
       Min         1Q     Median         3Q        Max
-126370134  -14956933   -4656003    9435721  197612090

Coefficients:
                      Estimate        Std. Error t value             Pr(>|t|)
(Intercept)     196324668.007803   19267989.130420   10.19 < 0.0000000000000002 ***
runtimeMinutes    -221958.739335      71074.302760   -3.12               0.0018 **
averageRating    -4560586.873706    1601965.291972   -2.85               0.0045 **
numVotes             1527.582810         48.716547   31.36 < 0.0000000000000002 ***
Hcluster3         -62708449.871729    8642067.855395   -7.26     0.00000000000065 ***
Hcluster4         -95666886.261297    9539966.158382  -10.03 < 0.0000000000000002 ***
Hcluster5        -125993855.985736    9715222.064587  -12.97 < 0.0000000000000002 ***
Hcluster6        -115074851.529588    9317475.389024  -12.35 < 0.0000000000000002 ***
Hcluster7        -120840494.600561    9609278.129692  -12.58 < 0.0000000000000002 ***
Hcluster8        -125858071.715077    9693905.309946  -12.98 < 0.0000000000000002 ***
km.cluster2         3469532.945051    6013099.967606    0.58               0.5640
km.cluster3       -10123837.914482    6785247.550290   -1.49               0.1359
km.cluster4         8377441.528139    6269236.054710    1.34               0.1817
km.cluster6        64620747.096442   11262641.364419    5.74     0.00000001168019 ***
km.cluster7       107855197.325340    7563254.695197   14.26 < 0.0000000000000002 ***
km.cluster8        -1212682.810481    6706925.883556   -0.18               0.8565
numVotes_sq              -0.003420          0.000143  -23.99 < 0.0000000000000002 ***
bo_and_num               -8.532412          0.316885  -26.93 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0 ‘**’ 0 ‘*’ 0 ‘.’ 0 ‘ ’ 1

Residual standard error: 29500000 on 1444 degrees of freedom
Multiple R-squared:  0.878,     Adjusted R-squared:  0.876
F-statistic:  609 on 17 and 1444 DF,  p-value: <0.0000000000000002

               GVIF Df GVIF^(1/(2*Df))
runtimeMinutes    3  1               2
averageRating     4  1               2
numVotes         22  1               5
Hcluster        107  6               1
km.cluster      232  6               2
numVotes_sq      14  1               4
bo_and_num        3  1               2
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
     3000  10650000  38800000  68740872  93950000 462200000
             ME     RMSE      MAE   MPE MAPE
Test set 1256100 29406899 19023472 -2076 2301
```

Figure 6.5.1: Model Result

After adding hierarchical and k-mean clusters to the previous model, we can see that RMSE decreases to 29,406,899, and R-squared increases to 0.876. We can also see that VIF values for all the predictors are less than 5 except for numVotes. However, the VIF value is still acceptable since it is less than 10.

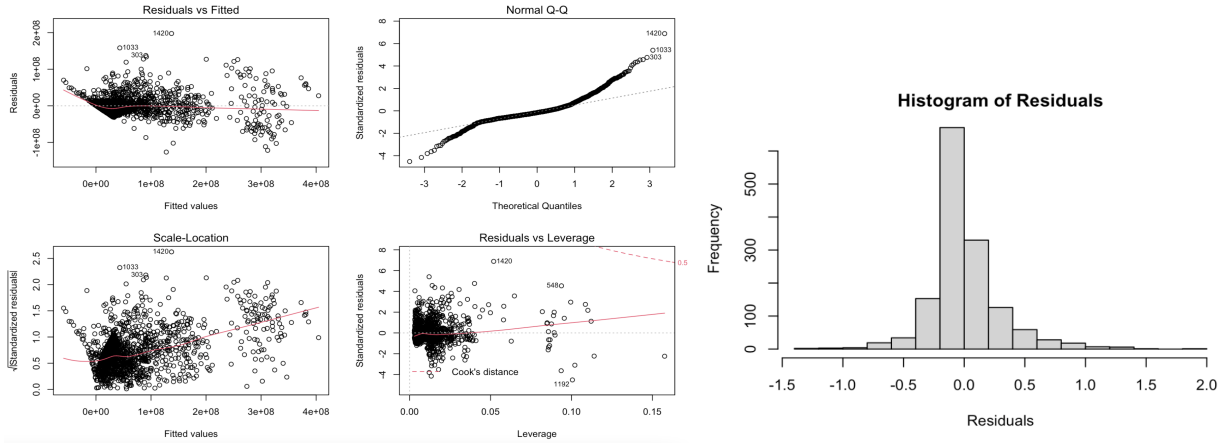


Figure 6.5.2: Diagnostic Plots and Histogram of Residuals

As we can see in the above plots, the Residuals vs. Fitted plot in the upper left shows a bit more random distribution than the previous plot. The Normal Q-Q plot in the upper-middle tracked closely to the diagonal line. The histogram of residuals has a bell-shaped distribution.

6.6. Compare and Contrast Between Manual and DataRobot Models

Regression	Manual	DataRobot
Without Clusters	RMSE: 40,504,048 R-squared: 0.769	RMSE: 43,289,000 R-squared: 0.7422
Cluster	RMSE: 29,406,899 R-squared: 0.876	RMSE: 31,419,000 R-squared: 0.8642

Table 6.6.1: Model Accuracy Comparison

To make sure all our models are consistent, we use the same features as well as clusters when searching for the models on DataRobot. We pick multiple regression models from DataRobot because we want to make a fair comparison between each model. As shown in the above table, the best model is the manual model with clusters since it has the lowest RMSE value which is 29,406,899. By picking the model with the lowest RMSE value, we hope to minimize the overfitting issues in our model.

Regression	Manual	DataRobot
Without Clusters	RuntimeMinutes numVotes numVotes_sq bo_and_num	numVotes numVotes_sq bo_and_num
Cluster	RuntimeMinutes averageRating numVotes Hcluster 3,4,5,6,7, and 8 Km.cluster 6 and 7 numVotes_sq bo_and_num	numVotes numVotes_sq bo_and_num

Table 6.6.2: Important Features Comparison

Furthermore, we compare the important features between the manual and DataRobot models. We choose variables with p-values lower than 0.05 as important features for the manual model. For DataRobot models, we choose features that have an impact of more than 30%. Surprisingly, almost all of the cluster dummy variables are significant for the manual model. However, for DataRobot, the cluster dummy variables do not have much impact. We can also see that only a few features have a significant impact on the model for DataRobot while a majority of the features are significant for the manual model.

6.7. Best Model Interpretation

```
Call:
lm(formula = worldwidegross ~ ., data = train.tC)

Residuals:
       Min         1Q     Median         3Q        Max
-126370134  -14956933   -4656003    9435721  197612090

Coefficients:
                      Estimate        Std. Error t value             Pr(>|t|)
(Intercept)     196324668.007803   19267989.130420   10.19 < 0.0000000000000002 ***
runtimeMinutes    -221958.739335      71074.302760   -3.12               0.0018 **
averageRating    -4560586.873706    1601965.291972   -2.85               0.0045 **
numVotes             1527.582810         48.716547   31.36 < 0.0000000000000002 ***
Hcluster3        -62708449.871729    8642067.855395   -7.26     0.00000000000065 ***
Hcluster4        -95666886.261297    9539966.158382  -10.03 < 0.0000000000000002 ***
Hcluster5       -125993855.985736    9715222.064587  -12.97 < 0.0000000000000002 ***
Hcluster6       -115074851.529588    9317475.389024  -12.35 < 0.0000000000000002 ***
Hcluster7       -120840494.600561    9609278.129692  -12.58 < 0.0000000000000002 ***
Hcluster8       -125858071.715077    9693905.309946  -12.98 < 0.0000000000000002 ***
km.cluster2        3469532.945051    6013099.967606    0.58               0.5640
km.cluster3      -10123837.914482    6785247.550290   -1.49               0.1359
km.cluster4        8377441.528139    6269236.054710    1.34               0.1817
km.cluster6       64620747.096442   11262641.364419    5.74     0.00000001168019 ***
km.cluster7      107855197.325340    7563254.695197   14.26 < 0.0000000000000002 ***
km.cluster8       -1212682.810481    6706925.883556   -0.18               0.8565
numVotes_sq            -0.003420          0.000143  -23.99 < 0.0000000000000002 ***
bo_and_num             -8.532412          0.316885  -26.93 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0 ‘**’ 0 ‘*’ 0 ‘.’ 0 ‘ ’ 1

Residual standard error: 29500000 on 1444 degrees of freedom
Multiple R-squared:  0.878,     Adjusted R-squared:  0.876
F-statistic:  609 on 17 and 1444 DF,  p-value: <0.0000000000000002
```

Figure 6.7.1: Best Model Result

The above results show the multiple regression formula of the best model after the comparison between all of the models. The RMSE 29,406,899 is the lowest RMSE out of all the models. We believe that we minimized overfitting in this model because we select our best model by choosing the lowest RMSE. Since adjusted R-squared will keep improving if we eliminate more outliers or add more features, by choosing the lowest RMSE, we try to manage our model from overfitting.

The intercept is 196,324,668.007803. If the runtime minutes of the movie (runtimeminutes) increase by 1 minute, the worldwide gross will decrease by 221,958.739335. If the user’s average rating of the movie (averageRating) increases by 1 rank, the worldwide gross

will decrease by 4,560,586.873706. If the number of user votes (numVotes) increases by 1 vote, the worldwide gross will increase by 1,527.582810.

If the movie is in group 3 of the hierarchical cluster, the worldwide gross will decrease by 62,708,449.871729. If the movie is in group 4 of the hierarchical cluster, the worldwide gross will decrease by 95,666,886.261297. If the movie is in group 5 of the hierarchical cluster, the worldwide gross will decrease by 125,993,855.985736. If the movie is in group 6 of the hierarchical cluster, the worldwide gross will decrease by 115,074,851.529588. If the movie is in group 7 of the hierarchical cluster, the worldwide gross will decrease by 120,840,494.600561. If the movie is in group 8 of the hierarchical cluster, the worldwide gross will decrease by 125,858,071.715077.

If the movie is in group 2 of the k-means cluster, the worldwide gross will increase by 3,469,532.945051. If the movie is in group 3 of the k-means cluster, the worldwide gross will decrease by 10,123,837.914482. If the movie is in group 4 of the k-means cluster, the worldwide gross will increase by 8,377,441.528139. If the movie is in group 6 of the k-means cluster, the worldwide gross will increase by 64,620,747.096442. If the movie is in group 7 of the k-means cluster, the worldwide gross will increase by 107,855,197.325340. If the movie is in group 8 of the k-means cluster, the worldwide gross will decrease by 1,212,682.810481.

If the number of user votes squared (numVotes) increases by 1 vote, the worldwide gross will decrease by 0.003420. If the interaction term of box office and number of votes increases by 1 unit, the worldwide gross will decrease by 8.532412.

From the figure above, we can also see that most of the significant variables are negatively impacting worldwide gross. From this result, we can conclude that our hypothesis that average rating will positively impact worldwide gross is incorrect. This result is surprising since an increase in average rating should help movies to earn more at the box office. This may be explainable if the people who watch movies in theaters have quite drastic preferences from people who provide ratings on IMDb.

Among all of the predictors, the significant variables that positively affect worldwide gross are the number of votes, and k means clusters 6 and 7. K means clusters 6 and 7 might be impacting positively since according to figure 5.2.2, the movies in these clusters have around average runtime minutes, medium or high average ratings, and medium or high number of votes. It seems that having average runtime minutes is an important factor for predicting a high worldwide gross.

7. Classification Analysis

7.1. Target Variable Description

We are trying to predict whether the movie will receive a high rating from its viewers. Our target categorical variable is popularmovie which is determined by the average rating of the movie. We think it is a good target variable because it can provide guidance for the movie industry to better know the customers' preferences. We divide the popularmovie into two categories by its mean of 6.37 since it can make a more balanced dataset if we use the mean as the cutoff. So, if the movie rating is above 6.37 then the popularmovie variable will be equal to 1. Otherwise, it will be 0. Then, the data is cleaned by removing the outliers outside the 1.5 times IQR range to receive better performance and meaningful prediction results.

7.2. Initial Hypothesis

We believe the variables runtimeMinutes, Worldwide gross, and numVotes will be significant in determining if the movie is popular. Movies within the range of 60 to 90 mins are an ideal length. Worldwide Gross will help determine a movie's popularity because a higher gross would indicate more people watched the movie. Number of votes, similar to worldwide gross, would determine popularity because a higher number would indicate many saw the movie and were interested in the movie enough to write a review.

7.3. Data Preparation

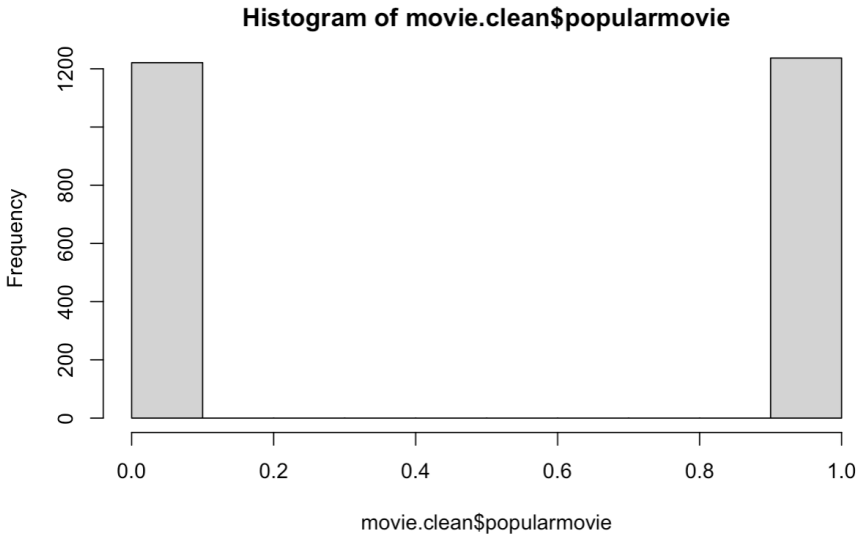


Figure 7.3.1: Histogram of Popular Movie without Outliers

We removed the outliers outside the 1.5 times IQR range, and the figure above shows the distribution of the target variable after cleaning. This results in 2,458 observations after removing

244 data points. From the histogram above, we can see the amount of both categories is almost at the same level. Thus, we can conclude the data is balanced and start the modeling process.

7.4. Manual Classification without Clustering Interpretation

```
best F1= 0.7258178 ; best_coeflearn= Zhu ; best_mfinal= 100 ; best_maxdepth = 3
```

Figure 7.4.1: Boosted Trees Grid Search Results

Our best manual classification model without clustering is the Boosted model. After data partitioning, by using the grid search we found the best values for coeflearn, mfinal, and maxdepth. The best value for coeflearn is Zhu, mfinal is 100, and maxdepth is 3. Lastly, we also found the best F1 score to be 0.7258.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 342 143
         1 162 437

               Accuracy : 0.7186
                 95% CI : (0.6908, 0.7452)
    No Information Rate : 0.5351
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4331

 Mcnemar's Test P-Value : 0.3027

            Sensitivity : 0.6786
            Specificity : 0.7534
         Pos Pred Value : 0.7052
         Neg Pred Value : 0.7295
             Prevalence : 0.4649
         Detection Rate : 0.3155
   Detection Prevalence : 0.4474
      Balanced Accuracy : 0.7160

       'Positive' Class : 0
```

```
          Reference
Prediction   0   1
         0 568 194
         1 156 706

               Accuracy : 0.7845
                 95% CI : (0.7637, 0.8043)
    No Information Rate : 0.5542
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.5661

 Mcnemar's Test P-Value : 0.04796

            Sensitivity : 0.7845
            Specificity : 0.7844
         Pos Pred Value : 0.7454
         Neg Pred Value : 0.8190
             Prevalence : 0.4458
         Detection Rate : 0.3498
   Detection Prevalence : 0.4692
      Balanced Accuracy : 0.7845
```

Figure 7.4.2: Confusion Matrix

We set the values from grid search to be the new parameters, which gave us the confusion matrices above (left using valid set and right using train set). From the left confusion matrix, we can see that the accuracy is 0.7186, sensitivity is 0.6786 and specificity is 0.7534. This model can correctly identify 67.86% of a movie that has an average rating higher than 6.37 as a movie with a rating higher than 6.37. This model can also correctly identify 75.34% of movies that have an average rating lower than 6.37 as a movie with a rating lower than 6.37. The overall accuracy is 0.7186 which is good, however, it can be improved further. Comparing the matrix using the valid set to the matrix using the train set, we see the accuracy, specificity, and sensitivity are all

higher when using training data to make predictions. This tells us the model suffers from over-fitting. This, like the bagging method, shows over-fitting which can potentially cause problems when transferring the model to other data. We also calculate the F1 score for this classification, which is 0.7141. F1 score is the weighted average of sensitivity and specificity values with 1 being the best and 0 being the worst. We think it is reasonable to get a 0.7141 for the F1 score, although it is noticeably lower than the F1 score of 0.7258 returned from the grid search.

7.5. DataRobot Best Model without Clustering Interpretation

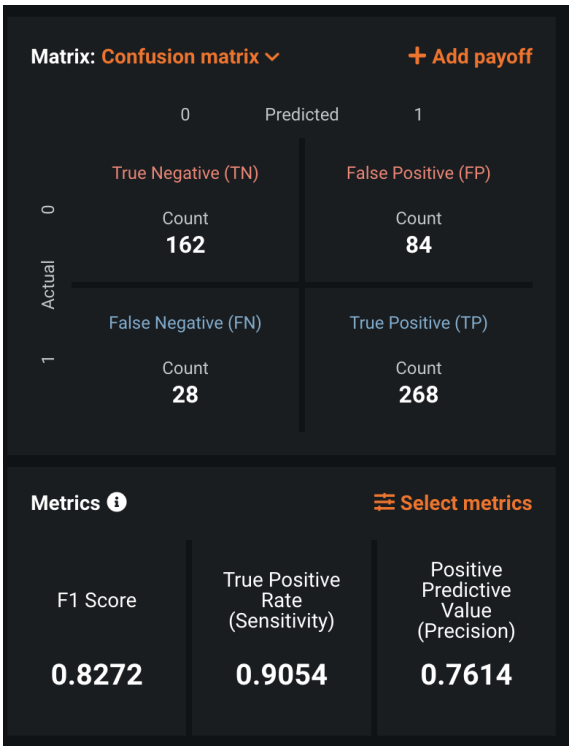


Figure 7.5. Best Model Result

Our best model without clustering in DataRobot is Nystroem Kernel SVM Classifier. From the figure above we can see that the F1 score is 0.8271 and the sensitivity is 0.9054, both are higher than the manual model we made. It is because the model in DataRobot included more features like Genre which improve the accuracy of the prediction.

7.6. Manual Classification with Clustering Interpretation

We partitioned the data with a 60% training set and a 40% testing set. Then, we selected variables bo_year_rank, worldwidegross, runtimeMinutes, numVotes, and interaction term bo_and num, which is the multiple of numVotes and bo_year_rank to run the regression with the training dataset. We added the interaction term bo_and_num since we think these two variables have a combined effect on our target variable. The clusters are then turned into dummy variables

labeled kc1 to kc7 for the 8 k-means clusters, and Hc1 to Hc7 for the 8 hierarchical clusters; kc8 and Hc8 are the base clusters.

```
                      Estimate         Std. Error z value             Pr(>|z|)
(Intercept)       1.8925565134298    0.9743439612481   1.942              0.05209 .
bo_year_rank     -0.0036171775095    0.0017011091640  -2.126              0.03347 *
worldwidegross    0.0000000005251    0.0000000020980   0.250              0.80235
runtimeMinutes    0.0074951947261    0.0081985129982   0.914              0.36060
numVotes          0.0000014391645    0.0000024897589   0.578              0.56324
bo_and_num        0.0000000937138    0.0000000347605   2.696              0.00702 **
kc11             -3.9965603367915    0.7415334216983  -5.390       0.000000070619 ***
kc21             -2.4240265647585    0.5505605309918  -4.403       0.000010684597 ***
kc31            -20.9719738078867  510.5322688576541  -0.041              0.96723
kc41             -4.7755334143612    0.4181068569250 -11.422 < 0.0000000000000002 ***
kc51            -23.0189784793185 4605.5316029126598  -0.005              0.99601
kc61             13.4455020305319 2911.9930465535022   0.005              0.99632
kc71             -4.1530700373475    0.6611740100008  -6.281       0.000000000336 ***
Hc11             19.7095333862099 4605.5318796870160   0.004              0.99659
Hc21              1.0728039543145    0.8461226258202   1.268              0.20483
Hc31              1.0617579820146    0.5891407521811   1.802              0.07151 .
Hc41              0.6402398842516    0.4587355546517   1.396              0.16282
Hc51             -2.1219443548177    0.4915825378727  -4.317       0.000015848132 ***
Hc61              1.3906496018233    0.4956409915998   2.806              0.00502 **
Hc71              0.9238631122109    0.6123264328726   1.509              0.13136

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2043.22  on 1473  degrees of freedom
Residual deviance:  823.79  on 1454  degrees of freedom
AIC: 863.79

Number of Fisher Scoring iterations: 17

  bo_year_rank worldwidegross runtimeMinutes       numVotes     bo_and_num            kc1
      2.562005      10.458888       3.314271       6.554176       3.413863       3.838174
             kc2            kc3            kc4            kc5            kc6            kc7
      6.448899       2.568529       4.311886       2.702608       1.240076       5.235448
             Hc1            Hc2            Hc3            Hc4            Hc5            Hc6
      2.439962       3.141902       5.229091       2.154493       4.387739       5.060259
             Hc7
      2.893723
```

Figure 7.6: Best Cluster Model Result

As shown in the figure, there are a few variables that are insignificant. Those variables are worldwidegross, kc3, kc5, kc6, Hc1, Hc2, Hc4, and Hc7. The VIF of worldwidegross, numVotes, kc2, kc7, Hc3, and Hc6 are above 5 with worldwidegross being the only variable above 10. Hence, other variables do not have collinearity issues except for worldwidegross.

7.7. Cluster Model Confusion Matrix Interpretation

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 644  64
         1 101 665

               Accuracy : 0.8881
                 95% CI : (0.8708, 0.9037)
    No Information Rate : 0.5054
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7762

 Mcnemar's Test P-Value : 0.005069

            Sensitivity : 0.8644
            Specificity : 0.9122
         Pos Pred Value : 0.9096
         Neg Pred Value : 0.8681
             Prevalence : 0.5054
         Detection Rate : 0.4369
   Detection Prevalence : 0.4803
      Balanced Accuracy : 0.8883

       'Positive' Class : 0
```

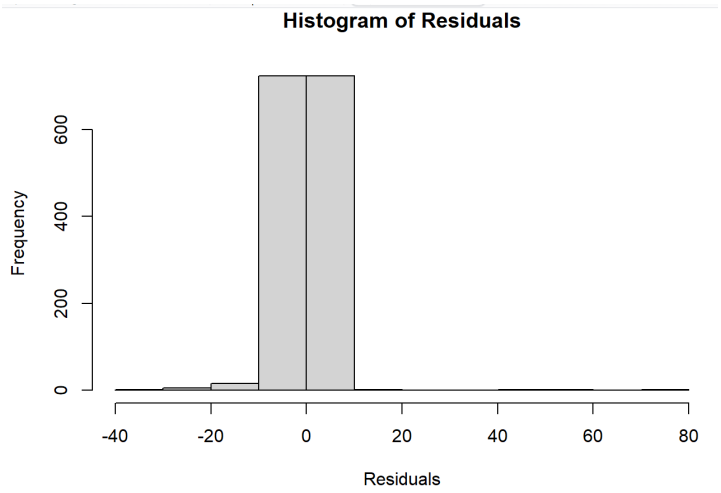
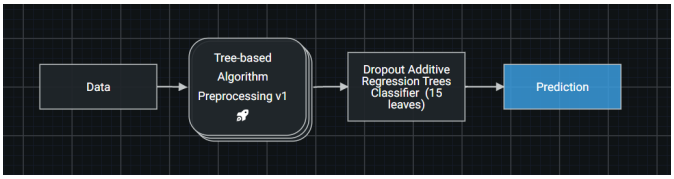


Figure 7.7.1. Confusion Matrix and Histogram of Residuals

From the result above, we can see that the accuracy, 0.8881 is significantly higher than that of the previous model. Sensitivity is 0.8644 and specificity is 0.9122. This tells us the model can make correct predictions 88.81% of the time. Sensitivity tells us how correctly the model can identify a movie that has an average rating higher than 6.37 when its rating is higher than 6.37. Thus, logistic regression can correctly predict 86.44% of popular movies as popular movies. Specificity measures how correctly the model can identify a movie that does not have an average rating higher than 6.37 when its rating is not higher than 6.37. Thus, logistic regression can predict 91.22% of the wrong values correctly. The overall accuracy is 0.8881 which is an improvement from the model without clustering (0.6798). The F1 score for this regression is 0.88. F1 score is the weighted average of sensitivity and specificity values with 1 being the best and 0 being the worst. We think the F1 score of 0.88 is very good, but the model can be further improved. Furthermore, from the histogram above, we can see the residual is skewed to the right.

7.8. DataRobot Cluster Model Interpretation



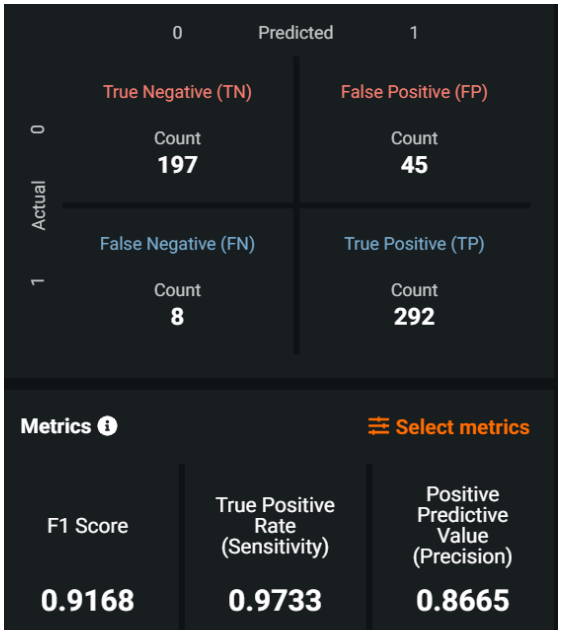


Figure 7.8.1. Confusion Matrix

From this, we can calculate the Accuracy as being 0.9022. The F1 score is 0.9168. The model's ability to correctly predict a popular movie as being popular is 0.9733, this is a high number indicating the model can strongly predict popular movies as popular. The Specificity/Precision is 0.8665 which tells us the model is less accurate in predicting non-popular movies correctly.

7.9. Compare and Contrast Between Manual and DataRobot Models

Classification	Manual	DataRobot
Without Clusters	Accuracy: 0.7186 Sensitivity: 0.6786 Specificity: 0.7534 F1 score: 0.7258	Accuracy: 0.7934 Sensitivity: 0.9054 Specificity: 0.6585 F1 score: 0.8272
Cluster	Accuracy: 0.8881 Sensitivity: 0.8644 Specificity: 0.9122 F1 score: 0.88	Accuracy: 0.9022 Sensitivity: 0.9733 Specificity: 0.8665 F1 score: 0.9168

Table 7.9.1: Model Accuracy Comparison

To ensure all our models are consistent, we use the same features as well as clusters when searching for the models on DataRobot. The difference in our Manual vs DataRobot data is the way the cluster variables are presented. For the Manual model, we created dummy variables for both cluster types. The 8 k-means clusters were divided into 7 dummy variables; the 8 hierarchical clusters were also divided into 7 dummy variables. We did not use the data set with the dummy variables in DataRobot because we believe DataRobot would be able to identify these items as categorical unlike R. From the results, we see the DataRobot cluster model (Dropout Additive Regression Trees Classifier (15 leaves)) did very well. It has a higher Accuracy, Sensitivity, and F1 score. This indicates the model can better predict popular movies as popular relative to the manual model. However, the manual model shows it can better predict non-popular movies correctly.

Classification	Manual	DataRobot
Without Clusters	Bo_and_num Number of votes	Genre split1 Bo_and_num
Cluster	K-mean clusters 4, 7, 1, 2 Hierarchical clusters 5 and 6 Bo_year_rank Bo_and_num	K-mean clusters 8, 2, 5 Hierarchical clusters 1, 2, 3 Bo_and_num

Table 7.9.2: Important Features Comparison

For clustering, the Manual and DataRobot model both showed the k-mean clusters are relevant and useful in the predictions. Bo_and_num interaction term also shows significance in both models.

8. Conclusion

8.1. Regression Analysis Business Insights

To succeed in the movie industry, movie producers should create movies that are within average runtimes such as 1 hour to 1 hour and 30 minutes which are the movies in the k-means cluster of 6 and 7. It might be intuitive to research movies that have high average ratings to produce movies that can appeal to mainstream audiences. However, our regression analysis suggests that average rating is impacting worldwide gross negatively. Thus, producers should not focus on movies with high average ratings when researching which types of movies to produce. This is because if only a few people rate the movie and the ratings are high, the average rating will be high. Hence, producers should try to research movies that have a large number of votes because having a higher number of votes means the movie is popular. So, movie producers can

research the movies that have a high number of votes to understand which genre, actors, or plots current popular movies have. For instance, if Marvel superhero movies are popular, a movie producer can try to recreate superhero movies with a different backstory. Therefore, if production companies want to succeed in producing a movie that will receive a high worldwide gross, they should focus on researching the types of movies that their target customer segment wants to watch since this will be the most important factor for earning high worldwide gross.

8.2. Classification Business Insights

We can observe that the number of votes and runtime minutes are widely used in the first three levels of the classification tree and ensemble models. Thus, we suggest businesses in the movie industry can pay more attention to these dimensions when they plan to produce new movies. Firstly, they can produce movies with popular topics to trigger more audience votes and assure the number of votes is greater than 166,834 on the IMDb website. Secondly, businesses can also produce more movies with longer runtime minutes than 105 minutes.

On the other hand, the business can refer to the km clusters 2, 5, 8 to produce the movie, like km cluster 5 which is highest in worldwide gross, domestic gross, oversea gross, high in average rating, runtime minutes, and numbers of votes, and km cluster 8 which with low worldwide gross, low domestic gross, low overseas gross, low runtime minutes, medium average rating, and low number of votes. They can produce new movies like those in cluster 5 to make them popular and also to avoid producing new movies like those in cluster 8.

9. Reflection About the Journey

After experiencing the whole process of modeling - data visualization, data wrangling, setting target variables and features, and partitioning training and validation sets - we realized the importance of each step to receive the model with the best performance. In the visualization part, we need to plot the data to see the relationship between each feature to understand the data structure. We also can use some visualization tools to exclude the outliers to assure the data quality. To select the target variable and features, we can add some interaction terms and higher-order variables to improve the prediction accuracy of the model. In the data training part, some of the steps are iterative. Thus, we have to patiently tune the parameters until we receive the ideal results. Instead of doing it manually, there are some useful tools to help us, like grid search. Moreover, we can fit the data into several models to test out which one can best capture the data structure and return us the best test score. We learned a lot during this journey from data preprocessing to building models and interpreting meaningful results. Even though we faced several challenges such as errors in Rstudio and models not being implemented correctly, we persevere and produce the best results. Overall, it was a journey that was worthwhile for all of us in the Lucky Tiger Team.