FINAL PROJECT

OPIM 5604 – Fall, 2016

TEAM # **6**

CLASS SECTION – **Afternoon**

TEAM MEMBERS:

Anita Sivakumar

Jinsheng Wang

Neha Telagubanajigara Jagannath

Niharika Srivastava

Yuhang Wu

“The work contained and presented here is our work and our work alone.”

Table of Contents

[1. Executive Summary 2](#_Toc468222260)

[1.1 Problem Statement 2](#_Toc468222261)

[1.2 Approach 3](#_Toc468222262)

[1.3 Results 3](#_Toc468222263)

[2. Introduction 3](#_Toc468222264)

[3. Data Understanding & Initial Preparation 3](#_Toc468222265)

[4. Possible Modeling Outcome & Target Variables 5](#_Toc468222266)

[5. Data Exploration 6](#_Toc468222267)

[5.1 Data Recoding 6](#_Toc468222268)

[5.2 Data Visualization 6](#_Toc468222269)

[5.2.1 Data Summarization 6](#_Toc468222270)

[5.2.2 Data Visualization & Mathematical Exploration 9](#_Toc468222271)

[6. Other Variable Treatment 19](#_Toc468222272)

[7. Outlier Detection 19](#_Toc468222273)

[7.1 Variable Transformation 19](#_Toc468222274)

[8. Missing Data Handling 20](#_Toc468222275)

[8.1 Missing Data Pattern 20](#_Toc468222276)

[9. Data Reduction 21](#_Toc468222277)

[9.1 Highly Sparse Variables 21](#_Toc468222278)

[9.2 Highly Correlated Columns 21](#_Toc468222279)

[10. Standardization 22](#_Toc468222280)

[11. Clustering 22](#_Toc468222281)

[12. Principal Component Analysis 22](#_Toc468222282)

[13. Final Outlier Treatment 23](#_Toc468222283)

[14. Modeling 23](#_Toc468222284)

[14.1. Classification Models 23](#_Toc468222285)

[14.2 Regression Models 31](#_Toc468222286)

[15. Conclusion 37](#_Toc468222287)

[15.1 Classification 37](#_Toc468222288)

[15.2 Regression 38](#_Toc468222289)

[16. Next Steps 39](#_Toc468222290)

[17. Appendix 40](#_Toc468222291)

[17.1 Data Dictionary 40](#_Toc468222292)

[17.1 Using Discriminant Analysis for understanding the top variables 40](#_Toc468222293)

# Executive Summary

## Problem Statement

Using the IMDB scraped data build models that can accurately predict whether a movie is bad, average or good and the IMDB score. This model would especially be helpful in cases where the critics reviews are not immediately available. With thousands of movies releasing every year, this might be a better way to tell the greatness of a movie without relying on critics or our own instincts. To answer this question, we selected 5000+ movies from IMDB and analyzed them.

## Approach

For performing this analysis, we started off with the initial data exploration and cleaning. Post preparing the final dataset, we performed regression and classification analysis on the same. Also, we looked deeper into whether outliers and the year the movie was released have any effect on the accuracy of the prediction. Our motivation behind the latter being that the advent of Facebook might have affected the factors affecting the overall perception of the movie.

## Results

On running the two types of models, we observe that for classification, the neural net without outliers and on the whole dataset seems to be performing better while for regression, the dataset with ensemble model without outliers and irrespective of whether the data is segmented based on year title or not performs better.

Since the regression model is preforming better on the whole dataset when compared to classification on part of the data set also because regression might be a better fit for business as it will give user the liberty of choosing the ratings that work for them. Business might prefer using regression over classification

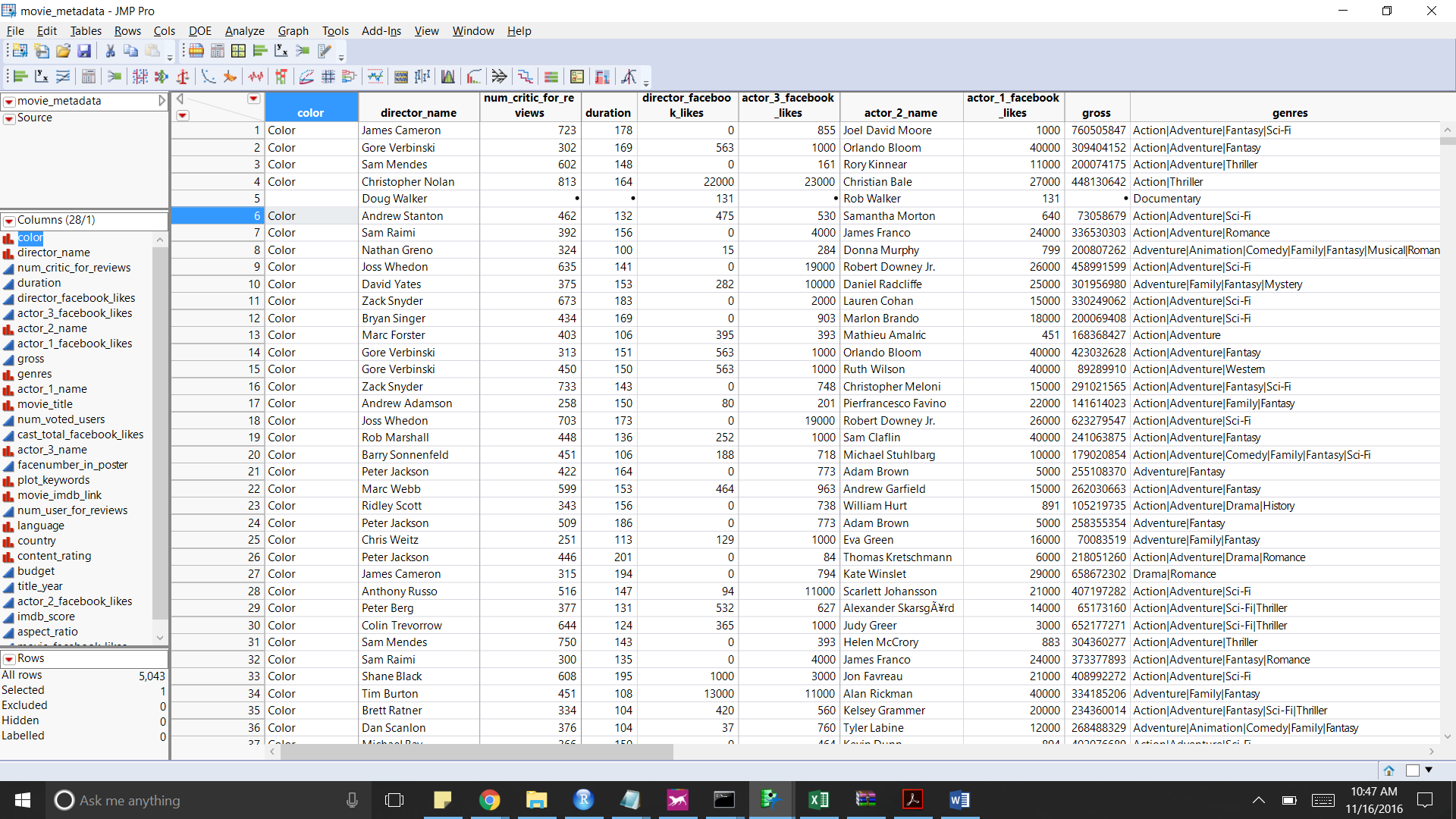
# Introduction

In this project, for the movies that have released in the last 50 years, we attempt to predict their IMBD scores and classify them as good, average or bad. To better understand the data, we started off with initial data exploration, cleaning and pre-processing. On this cleaned and pre-processed data, we applied various prediction and classification techniques to come up with the predicted scores and classification respectively. Furthermore, we have also provided an appendix towards the end listing all the techniques that did not make sense or could not be applied to the given dataset along with the plausible reasons behind the same.

For all our analysis in this project we have used SAS JMP Pro.

# Data Understanding & Initial Preparation

For this analysis, we used the IMDB dataset available on [Kaggle](https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset). The data for this dataset has been scraped by the dataset [author](https://www.kaggle.com/deepmatrix) using Python from [moviecatcher](http://moviecatcher.net/) and [IMDB](http://www.imdb.com/). The data set contains 5043 rows and 28 columns. Below is the screenshot of the same.



**Figure 3.1**

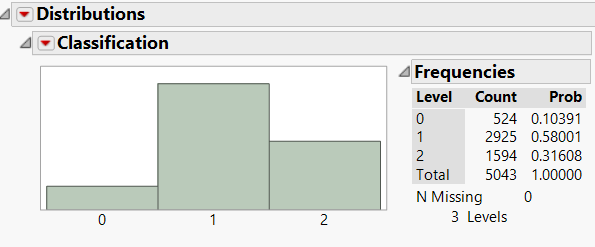
Since the dataset size is not huge, for this analysis we would be using the complete dataset. Also, for classification analysis we created on extra column called ‘movie\_goodness’ based on the following formula:

**IF imdb\_score > 5 then 2**

**ELSE ( IF imdb\_score <= 5 then 0**

**ELSE 1)**

where 2 signifies good, 1 as average and 0 as bad. Below is the final dataset on which we performed complete analysis. Below is the distribution of the same in the data.



**Figure 3.2**



We started off our analysis by checking the datatypes of each of the variables and reassigned the datatypes wherever required. Below is the table with the corrected data type for each of the variable.

**Table 2.1**

|  |  |
| --- | --- |
| **Column Name** | **Corrected Variable Type** |
| color | Nominal Variable |
| director\_name | Nominal Variable |
| num\_critic\_for\_reviews | Continuous Variable |
| duration | Continuous Variable |
| director\_facebook\_likes | Continuous Variable |
| actor\_3\_facebook\_likes | Continuous Variable |
| actor\_2\_name | Nominal Variable |
| actor\_1\_facebook\_likes | Continuous Variable |
| gross | Continuous Variable |
| genres | Nominal Variable |
| actor\_1\_name | Nominal Variable |
| movie\_title | Nominal Variable |
| num\_voted\_users | Continuous Variable |
| cast\_total\_facebook\_likes | Continuous Variable |
| actor\_3\_name | Nominal Variable |
| facenumber\_in\_poster | Continuous Variable |
| plot\_keywords | Nominal Variable |
| movie\_imdb\_link | Nominal Variable |
| num\_user\_for\_reviews | Continuous Variable |
| language | Nominal Variable |
| country | Nominal Variable |
| content\_rating | Nominal Variable |
| budget | Continuous Variable |
| title\_year\* | Nominal Variable |
| actor\_2\_facebook\_likes | Continuous Variable |
| imdb\_score | Continuous Variable |
| aspect\_ratio | Continuous Variable |
| movie\_facebook\_likes | Continuous Variable |

\*Initially we treated the **title\_year** as a continuous variable but on further analysis and building the model we observed better results on treating this variable as a nominal variable.

# Possible Modeling Outcome & Target Variables

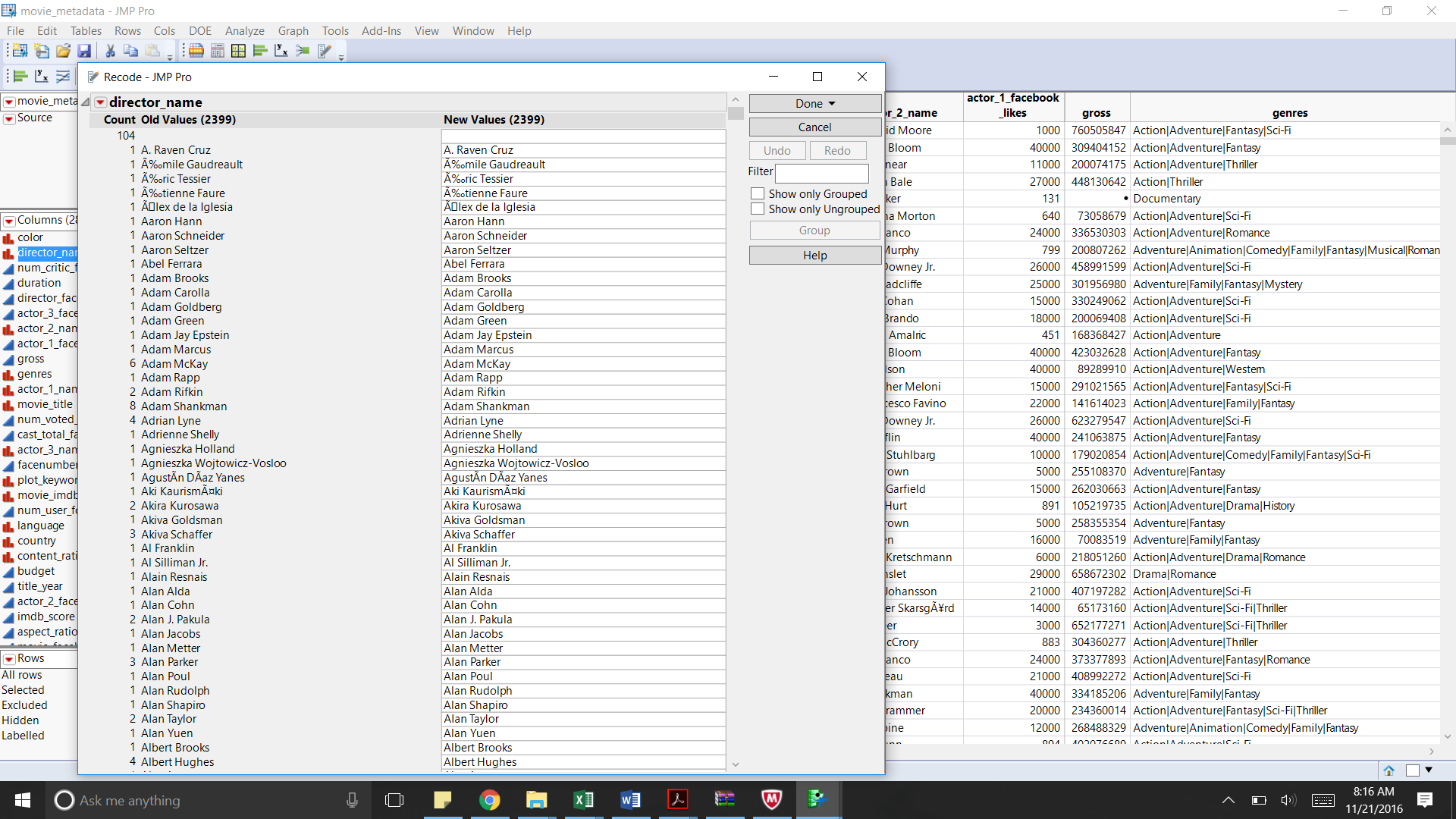
The variables that have been used as target variables are **imdb\_score** for prediction problem and **classification** for the classification problem. These models can be used in a tool that would help understand the possible goodness of the movie based on the given movie’s characteristics, without looking at the critic reviews. Also in the second part of the analysis, we analyze if factors governing the user ratings for movies before the advent of Facebook (2004) are different from the ones after Facebook was launched.

# Data Exploration

Once we had the working sample of 5043, we started with data exploration to better understand our data. In this process, we followed 3 steps. Firstly, we checked the columns with nominal data types for inconsistencies, especially for cases where the same level is represented in two different ways. We corrected such inconsistencies using the Recode function in JMP. Secondly, we performed a mathematical exploration of the data, trying to understand the central tendencies and measure of spread for each of the variables. Finally, we performed visual exploration of the data using various visualization techniques available in JMP to unravel possible pattern in the variables. This section talks about the above 3 processes in the above-mentioned order.

## Data Recoding

While analyzing the various columns for data inconsistencies, we noticed that “**director\_name**” has lots inconsistencies, especially in terms of special characters being used in various names. On cross-checking these values with the source website, we realized that the problem is present in the source website itself. A similar problem is seen in the column “**actor\_1\_name**”. The column “**movie\_title**” has all the instances ending in “Â”. The reason for this can be attributed to the code used for scraping the data. But this issue won’t affect data analysis as the movie\_title column serves the same purpose as Row Id and hence can be removed from the data.



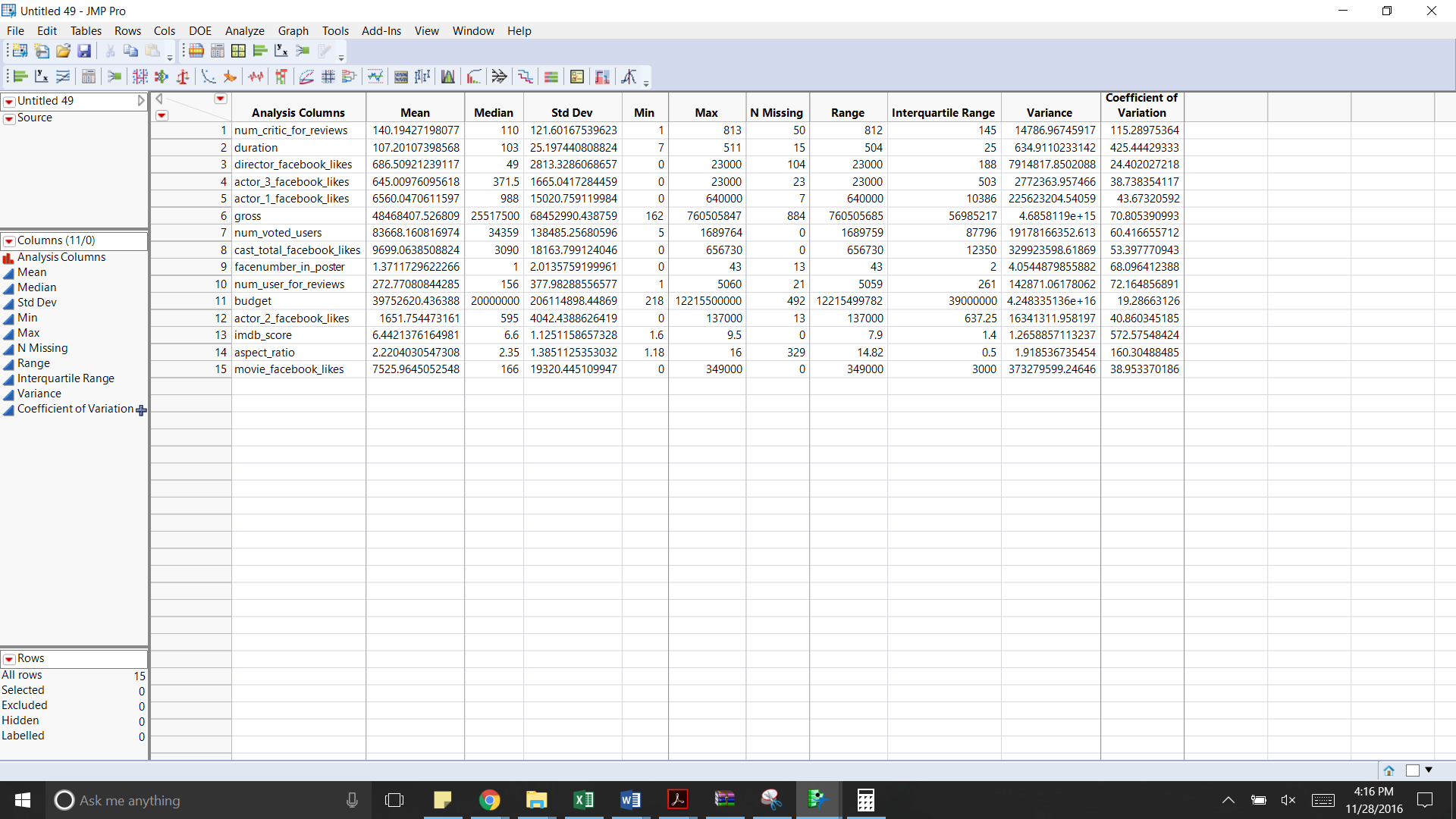
**Figure 5.1**

## Data Visualization

Under this section we consider various summarization and visualization techniques that we used to better understand our data and the relation between variables. The first section starts off with Summarization, followed by a section on Univariate visualization for each of the variable. Finally, we consider bivariate visualization to understand the relationship between various variables. We also tried using Tree maps to get a better sense on the data.

### Data Summarization

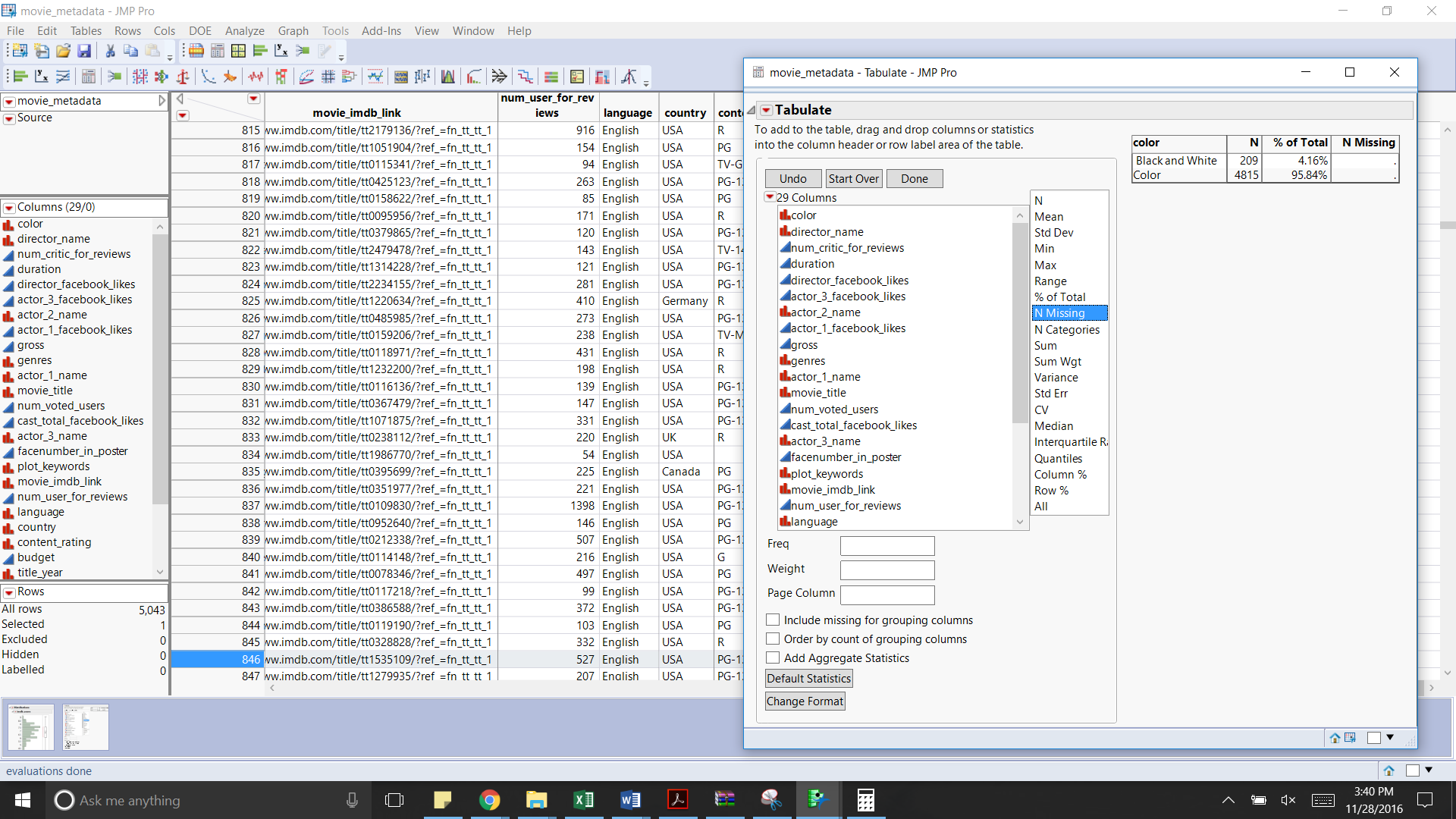
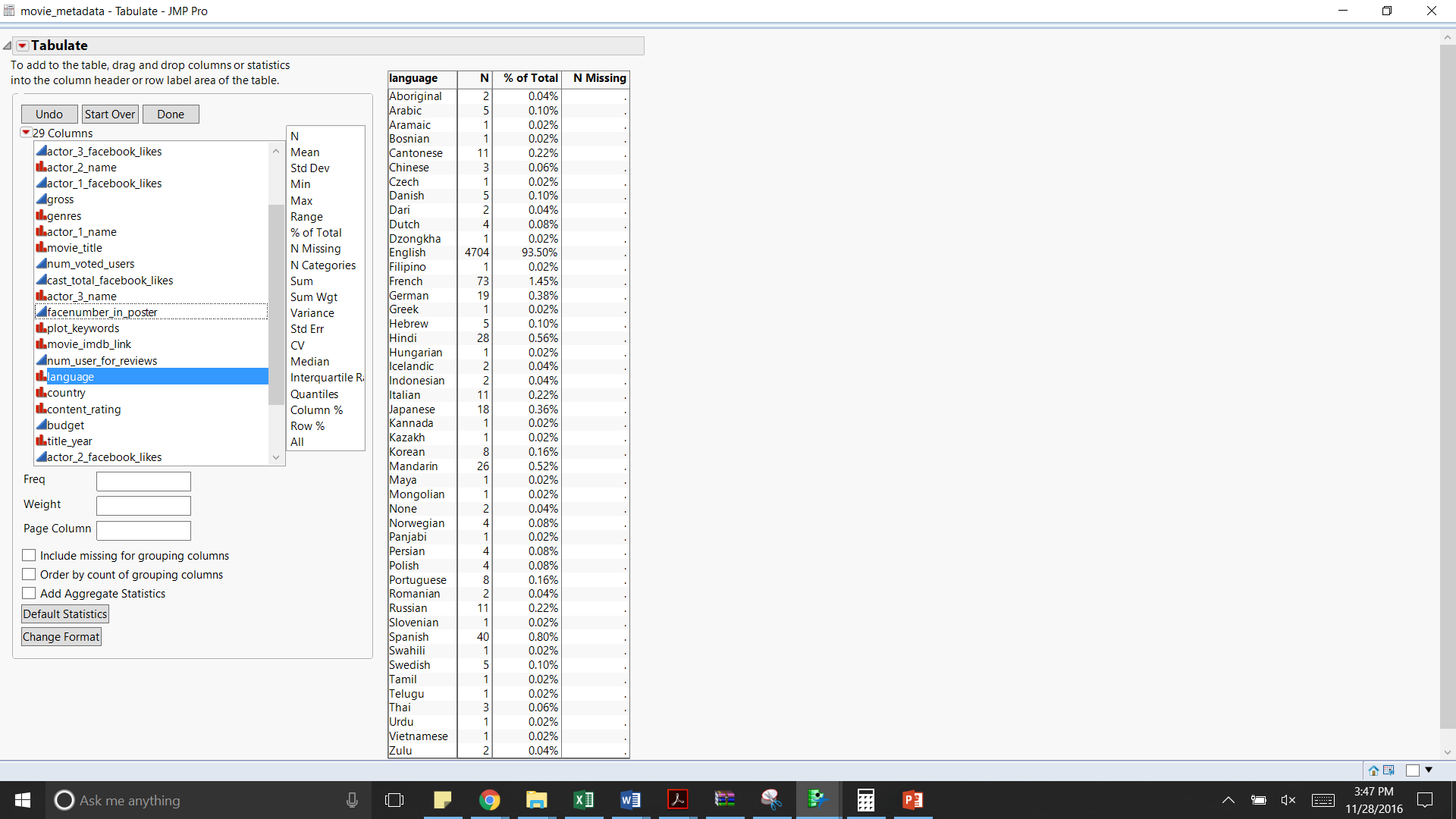
The table below tabulates various summary statistics.



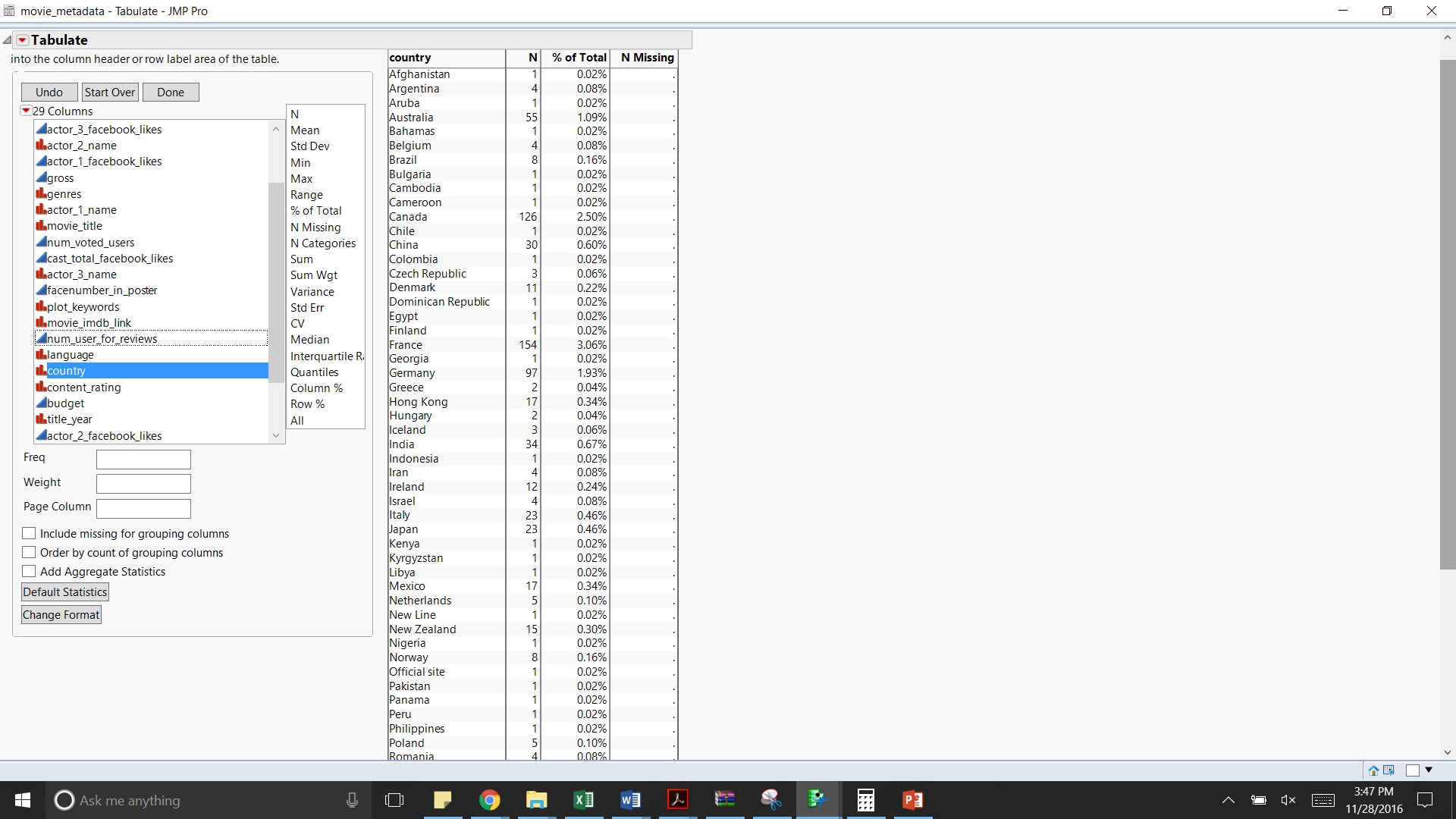
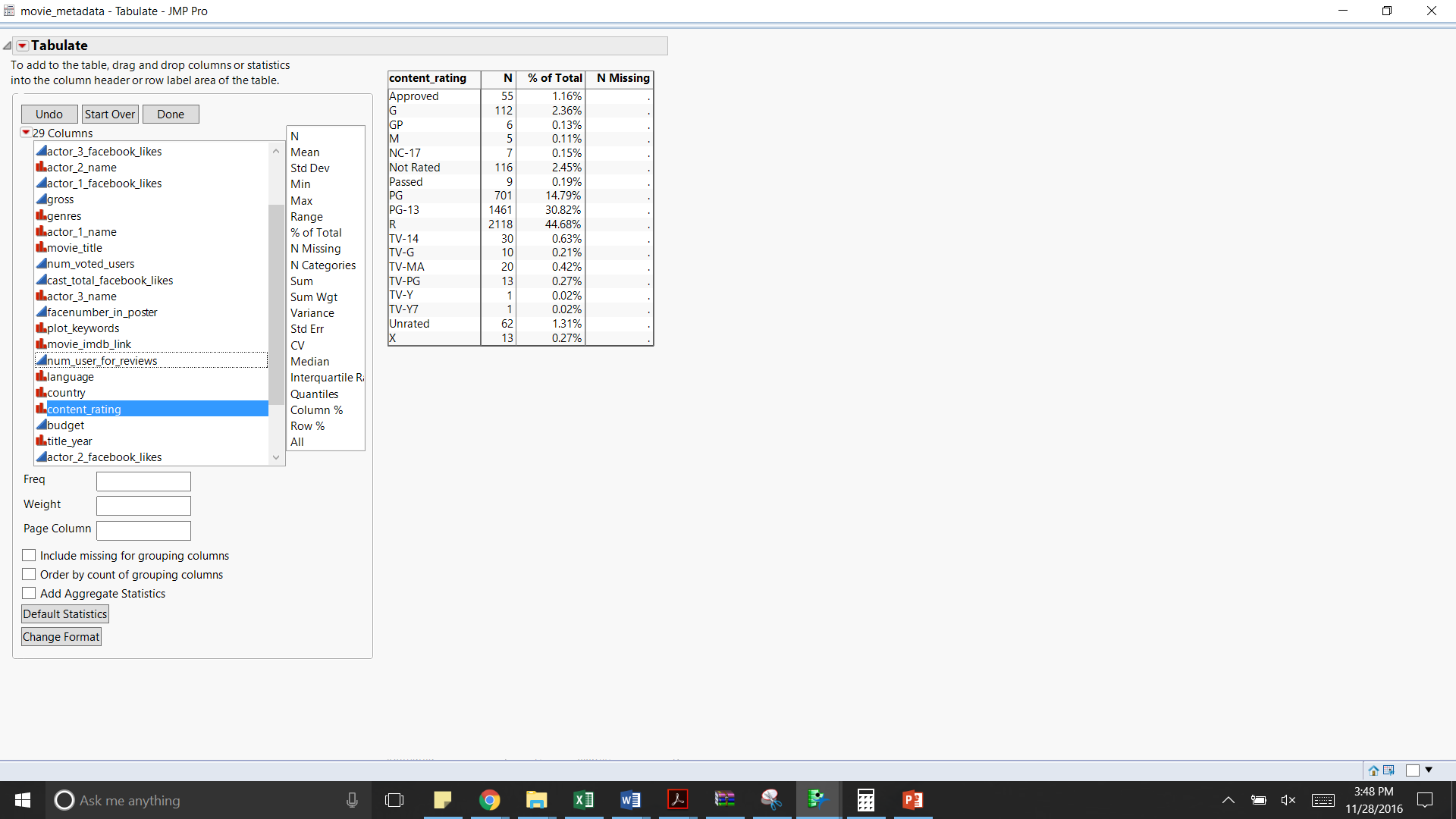
**Figure 5.2**

Below are the inferences from studying the table

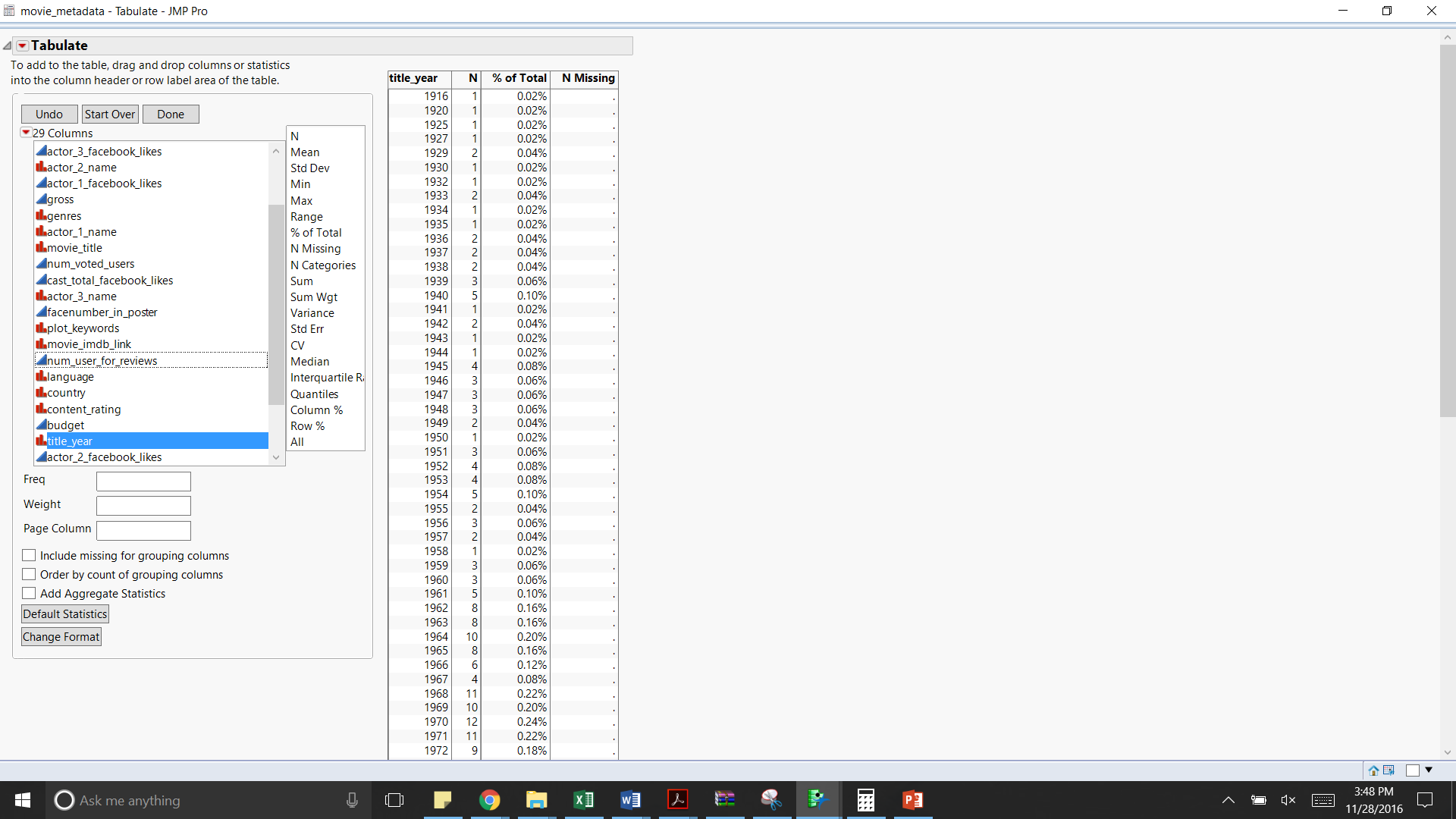
* There are no columns with all values missing, with the target variable present for all the instances
* The columns budget and gross revenue have the maximum number of missing values with around 16% and 9% of the values missing respectively.
* The columns duration and imdb\_score show the maximum variation.
* The variation in the column duration can be attributed to various types of movies ranging from short film to mini- film series.

**Figure 5.3 Figure 5.4**

**Figure 5.5 Figure 5.6**



**Figure 5.7**

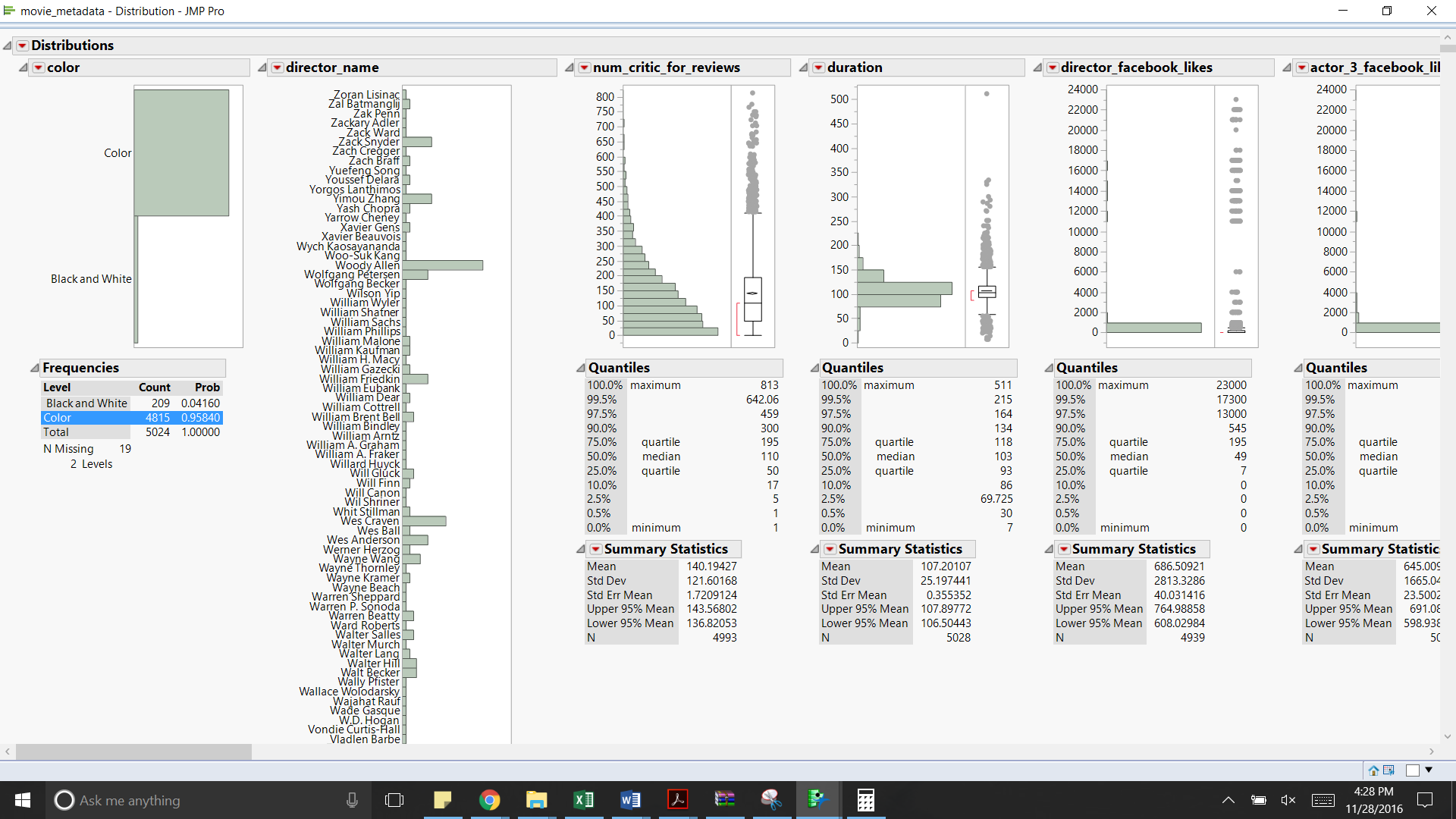
Based on the above pivot tables, we can make the following inferences

* The maximum number of movies reviewed are in Color (~96%).
* Around 94% of the movies are in English, followed by French (~1.4%)
* Most of the movies are from the USA (76%) with UK being the distant second (~9%)
* Around 1.5% of the data relates to TV series instead of movies based on the content rating values.
* Across the years the number of movies being rated is continually increasing uptil 2004 and then setting on a constant with a mean of around 250 movies per year

### Data Visualization & Mathematical Exploration

In this section, we start off with univariate analysis using histogram and bar charts for continuous and nominal variables respectively. Followed by bivariate analysis using scatter plot and mosaic plot and finally multivariate analysis using Tree maps.

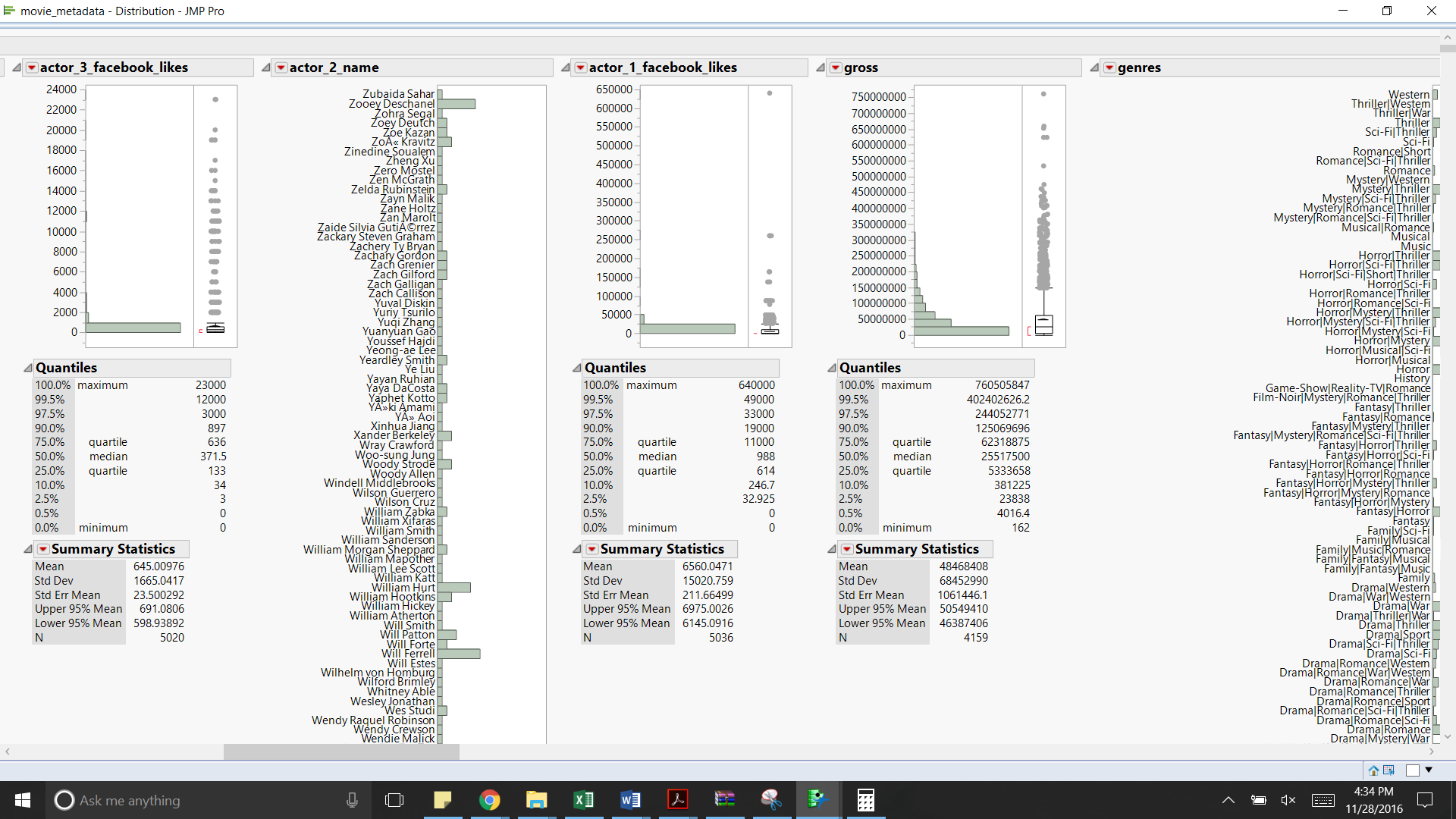
Below is the analysis based on the univariate analysis of the various variables



**Figure 5.8**

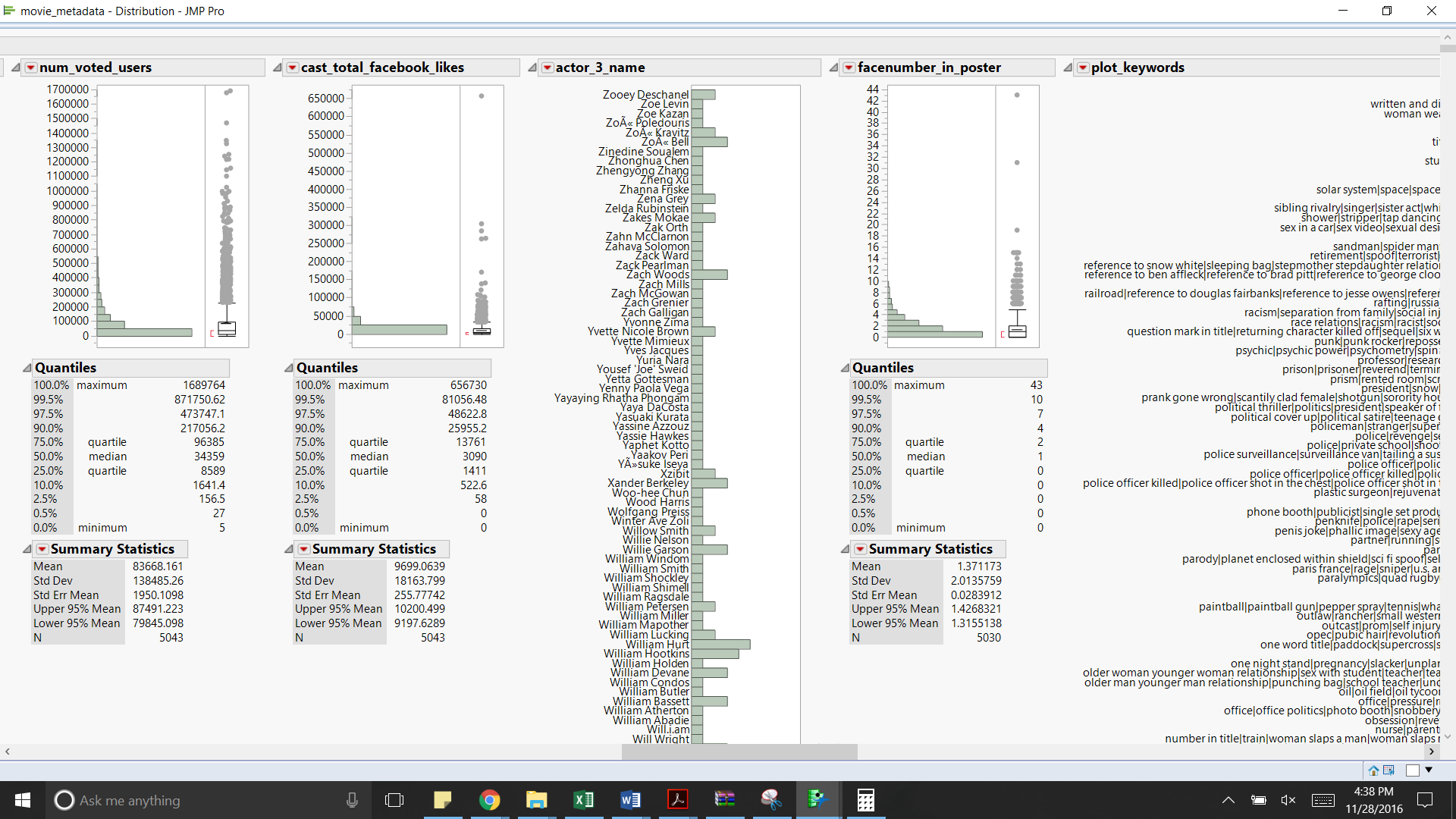
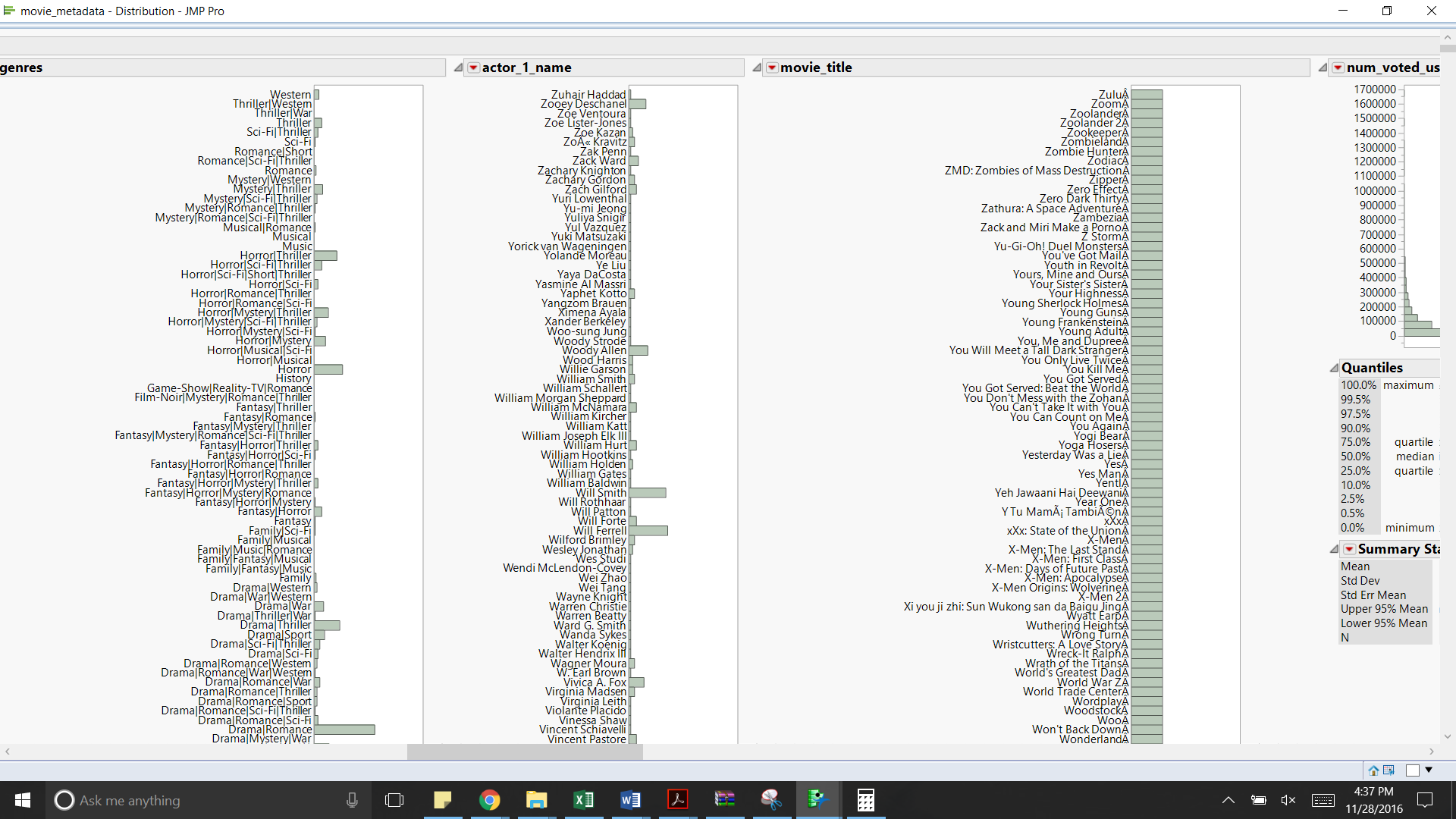
**Table 4.2.2.1**

|  |  |
| --- | --- |
| **Name of the variable** | **Observations** |
| Color | There are two categories of film colorizations: Color (4815 observations) and Black and White (209 observations). The dataset contains most of the movie which belong to color films |
| director\_name | The director names associated with the movies is mostly unique with few names occurring more than once. It could act as a potential explanatory variable to indicate if a movie directed by a director has a trend of having a good/bad/average ranking |
| num\_critic\_for\_reviews | The number of critic’s reviews for a movie ranges between 1 - 813. These variables could be used to determine if a critic’s review can have an influence on the target variable |
| duration | The minimum duration of a movie/TV show is 7 minutes and the maximum duration is 511 minutes. Most of the movies have a 90-minute duration. This can be used as an explanatory variable to see how a duration of a movie/TV show can influence the target variable |
| director\_facebook\_likes | The number of likes for a director on Facebook ranges between 0 - 2300, with Joseph Gordon-Levitt getting the maximum likes. This can be used as an explanatory variable to determine how the popularity of a director can contribute to the success of a movie. |



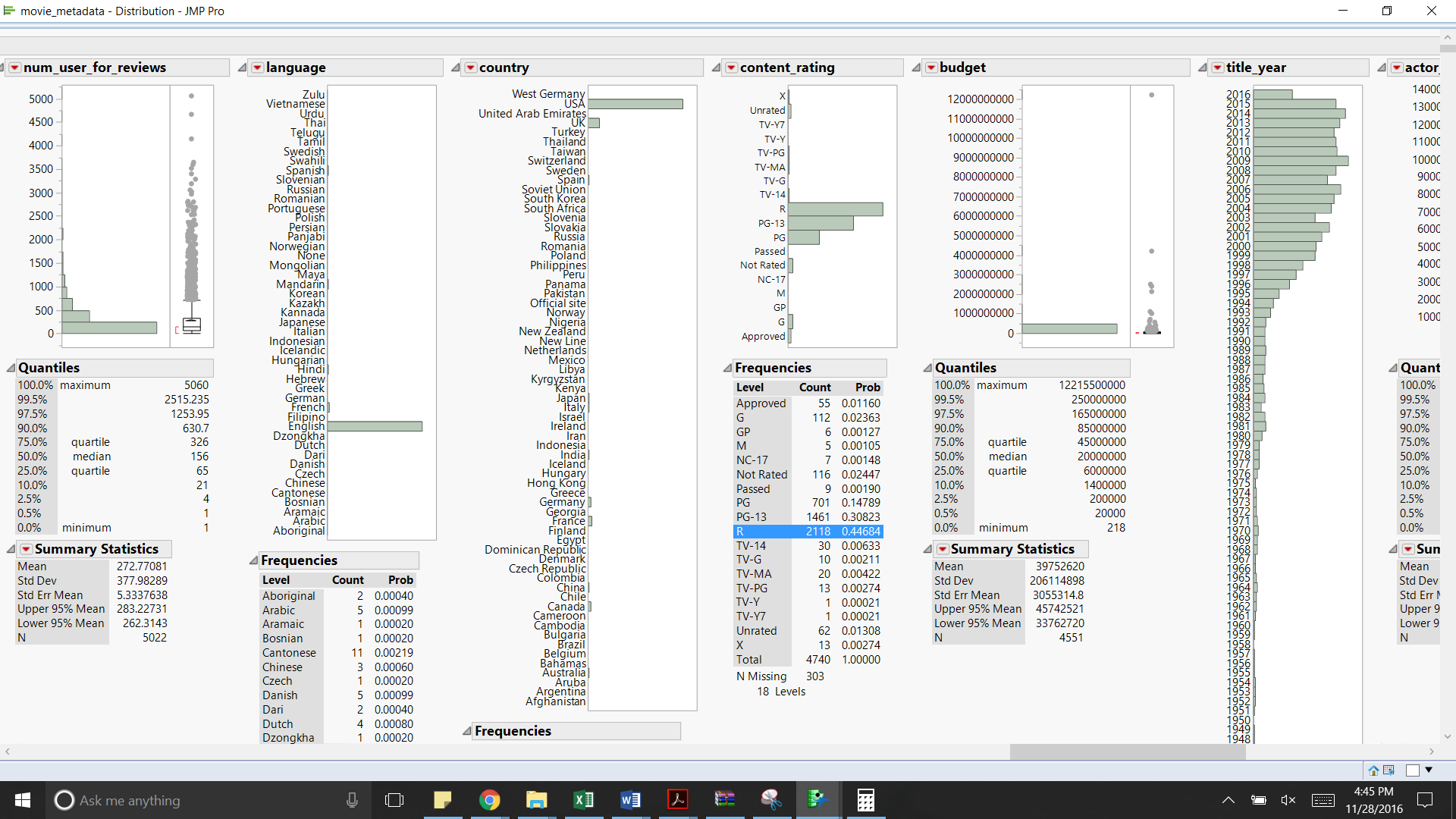
**Figure 5.9**

|  |  |
| --- | --- |
| actor3\_facebook\_likes | The Facebook likes for the 3rd actor in a movie ranges between 0 - 23000, with Joseph Gordon-Levitt getting the maximum likes. This can be used as an explanatory variable to determine how the popularity of an actor can contribute to the success of a movie. |
| actor\_2\_name | The actors listed as 2nd actor in a movie have been listed as 2nd actors in a minimum of 1 movie and a maximum of 20 movies. The actor’s names can be used as an explanatory variable to see if there is a relation between the success of a movie and the cast. |
| actor\_1\_facebook\_likes | The Facebook likes for the 1st actor in a movie ranges between 0 - 640000, with Darcy Donavan getting the maximum likes. This can be used as an explanatory variable to determine how the popularity of an actor can contribute to the success of a movie. |
| gross | The earnings the movies made at the box office range between 162 - 760505847, with most movies making an income of 34964818. The highest grossing movie is Avatar. It can be used as an explanatory variable to see how the target variable might be related to it (if a highest grossing movie associated with a movie which has a high score / is a good movie) |



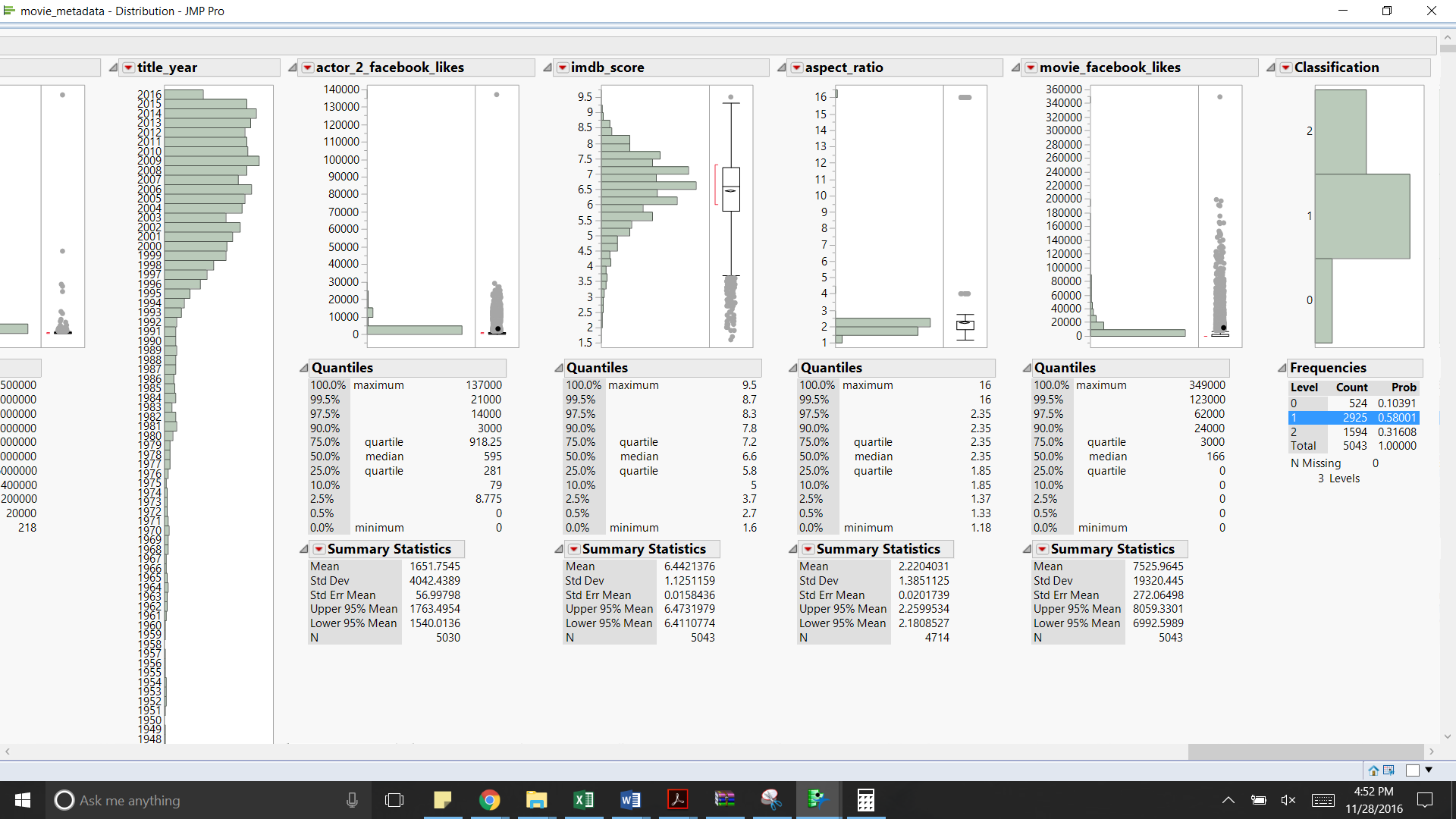
**Figure 5.10**

|  |  |
| --- | --- |
| genres | The genres range between 1 - 236. Most of the movies belong to the category “Drama”. This could be an explanatory variable to determine how a genre of a movie can influence a movie’s success |
| actor\_1\_name | The actors listed as 1st actor in a movie have been listed as 1st actors in a minimum of 1 movie and a maximum of 49 movies. Robert De Niro has been listed as the 1st actor in most movies. The actor’s names can be used as an explanatory variable to see if there is a relation between the success of a movie and the cast. |
| movie\_title | The movie names are unique. Hence they can be eliminated from being considered as an explanatory variable |
| num\_voted\_users | The number of votes from users ranges between 5 - 1689764. The Shawshank Redemption got the maximum votes from users. This variable can be used as a predictor to determine its influence on the target variable. |
| cast\_total\_facebook\_likes | The total Facebook likes for the movie cast ranges between 0 - 656730. This variable can be used as a substitute explanatory variable to the variables actor\_1\_facebook\_likes, actor\_2\_facebook\_likes, actor\_3\_facebook\_likes and director\_facebook\_likes to see how the combination has an influence on the target variable |
| actor\_3\_name | Most of the actors have acted in minimum of 1 movie to a maximum of 5 movies. There are around 3000 actors. |
| facenumber\_in\_poster | The number of face number in posters ranges from 0 to 44 and a mean of 1.37. |
| plot\_keywords | The plot keywords are almost unique to every movie and it varies with genre |
| movie\_imdb\_link | The IMDB link associated with every movie is also unique. Hence it does not add much value to the data. |



**Figure 5.11**

|  |  |
| --- | --- |
| num\_user\_for\_reviews | The number of reviews for a movie varies from 1 to 5060. This variable might add some value to predict the target variable. |
| language | Most of the number of movies listed in this database are English movies. |
| country | Most of the movies in this dataset are American movies followed by the ones from UK, Canada, France and Germany. |
| content\_rating | Content Rating for most of the movies is ‘R’ followed by PG - 13 and PG. Very few movies have their content ratings approved. |
| budget | The budget allocated for a movie varies from 218 to 12215500000. The mean of the budget comes to be 39752620. So, this dataset contains low budget as well as high budget movies. This variable can also have an impact on the IMDB score. |
| title\_year | This dataset contains movies which has released from 1916 to 2015. Among them, most of the movies have been released in the year 2010 and 2005. There are chances of the old movies’ IMDB score being calculated in a different way i.e. the variables considered during that time might have been different compared to the recent ones. |



**Figure 5.12**

|  |  |
| --- | --- |
| actor\_2\_facebook\_likes | Some actors have got 137000 Facebook likes and some have not got any. This might also be used as one of the contributing explanatory variable for determining the imdb score. |
| imdb\_score | The IMDB Score varies from 1.6 to a maximum of 9.5. The average IMDB score comes out to be 6.4. |
| aspect\_ratio | The aspect\_ratio of the movie ranges from 1.18 to 16. But the mean is only 2.2. Hence there are a few outliers which needs to be tackled. |
| movie\_facebook\_likes | The Facebook likes for a movie varies from 0 to 349000. The average number of Facebook likes is around 7500. |
| classification | Average rated movies constitute the biggest chunk of the whole date (~58%) followed by good and bad (~32% and ~10% respectively) |

Below are some of the insights based on the bivariate analysis between variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Y Axis** | **X Axis** | **Chart** | **Observations** |
| imdb\_score | title\_year |  | The IMDB scores for the almost all the movies released before the year 1968 is above 5. The movies which have been released after 1968 seems to have got scores less than 5 too.  There are chances of the old movies being evaluated with different parameters and that could have been the reason for their high scores. |
| movie\_facebook\_likes | title\_year |  | The number of Facebook likes seems to increase with the title\_year. The reason for this might be due to the year in which Facebook was launched. |
| facenumber\_in\_poster | title\_year |  | The face number in posters also seems to increase with title\_year. This might be attributed to more multi-starrer movies being made across years or an increase in number of popular actors every year |
| mean(imdb\_score) | language |  | Except for a few exceptions, the average imdb\_score doesn’t seem to be showing much variation across languages, even though some of the languages have just one movie. And since we have a very few movies in languages other than English, for the current dataset it might mean that language is not a major differentiator for the imdb\_score |
| mean(imdb\_score) | country |  | Except for a few exceptions, the average imdb\_score doesn’t seem to be showing much variation across countries, even though some of the countries have just one movie. And since we have only few movies from countries other than the USA, for the current dataset it might mean that country is not a major differentiator for the imdb\_score |
| Mean(movies\_facebook\_likes) | Imdb\_score |  | Except for one one major exception ( a Justin Bieber movie – a celebrity who has a strongly polarized fan base). For most of the movies as the number of Facebook likes increase so does the imdb\_score |

|  |  |  |  |
| --- | --- | --- | --- |
| **X axis** | **Y axis** | **Graph** | **Insights/Observations** |
| language | color |  | The dominated category of variable “language” is English, which contributes to 93.5% of “language”. Most of movies, no matter language, are made in color technique. Only one kind of language, Filipino, belongs to black & white movie totally which only has one movie. It is mainly because of the time when data represents. |
| country | color |  | Most of areas prefer to produce color movies. Only Kenya, Libya produce only black & white movies. But each of them only has one instance. It is mainly because of the time when data represents. |
| content\_rating | color |  | Most of categories of content is rated for color movies. It is mainly because of the time when data represents. |
| country | director\_name |  | The dominated country where directors come from is USA. Canada and European countries also have lots of directors included in this dataset. |
| language | actor\_2\_name |  | Most of actors participate in English movies. The variables of “actor\_1\_name”, “actor\_2\_name” have same property, the duplicated works will not be included in this table. |
| country | actor\_2\_name |  | The dominated country where actors come from is USA. Canada and European countries also have lots of actors included in this dataset. The variables of “actor\_1\_name”, “actor\_2\_name” have same property, the duplicated works will not be included in this table. |
| language | genres |  | Most of movies’ language are English. Also, the prevalent genre of movies is combination of comedy, Drama, Romance. |
| content\_rating | genres |  | Most of movies are categorized as PG, PG-13 and R. So they are not suitable for children audience. Also, the prevalent genre of movies are combination of comedy, Drama, Romance. |
| language | plot\_keywords |  | Most of plots are shown in the English movies |
| country | plot\_keywords |  | Most of plots are shown in the USA, UK or other European movies. |
| content\_rating | plot\_keywords |  | Most of plots are shown in the movies categorized as PG, PG-13 and R. So most of movies are not suitable for children audience. |
| content\_rating | lauguage |  | Most of movies are made in English and categorized as PG, PG-13 and R. So they are not suitable for children audience. |

Below are some of the observations based on the multivariate analysis.

**Table 4.2.2.9**

|  |  |
| --- | --- |
| **Chart** | **Observations** |
|  | Steven Spielberg, Woody Allen, Martin Scorsese and Clint Eastwood seem to be the most popular directors |
|  | Movies directed by Steven Spielberg seem to have the highest imdb scores |
|  | Movies directed by Christopher Nolan seem to be more popular |
|  | Anne Hathway and Scarlett Johnson are among the actors categoriised as “actor 3” to have the most facebook likes |
|  | Johnny Depp seems to be the most popular among the actors categorised as “actor 1” |
|  | Morgan Freeman seems to be the most popular among the actors categorised as “actor 2” |
|  | Steven Spielberg, Peter Jackson and Christopher Nolan are among the directors who have directed highest grossing movies |
|  | Action|Adventure|Sci-Fi movies are among the most liked genres |
|  | Movies belonging to the genre of Action|Adventure|Sci-Fi, Comedy, Comedy|Romance have been the highest grossing |
|  | The face number for most of the actors ranges from 1 to 4.The number of face numbers in posters ranges from 1 to 43. |
|  | Most of the movies have content\_rating as ‘R’ and there are significant number of movies which are unrated, not rated and approved and G. The correlation between content rating and IMDB score can give us some insight about the relation between content rating and good movies. |
|  | Movies which have got the IMDB score as 7 have invested the maximum. Some low budget movies also have the IMDB score as 8.7. So budget might not actually be the deciding criteria for a good IMDB score. |

Post these analysis, three of the variables- director\_name, actor\_1\_name, actor\_2\_name, actor\_3\_name were removed from the data analysis since most of the values in these columns were unique and with the dataset being really small, these won’t add any value to analysis. To include these variables in the analysis we might have to look at a bigger set of data. Also since the columns genre and plot\_keywords are textual in nature, they cannot be used in our current analysis but some text mining techniques can be used to derive useful variables from the same

# Other Variable Treatment

Since for this analysis we are comparing the movies starting from 1920 until 2016, one of the major factor that needs to be normalized is the budget and the gross revenue, whose current value might be very different because of inflation. To tackle this problem, based on [this](http://www.usinflationcalculator.com/inflation/historical-inflation-rates/) link we normalized all the budget and revenue values for the year 2016.

In doing so, to simplify the analysis, we have assumed that the gross revenue for a given movie is from the year it was released.

# Outlier Detection

Based on the analysis in the previous section we identified the columns that showed large number of outliers. We have used majorly 2 techniques to solve this concern – Variable Transformation and Mahalanobis Distance. Detecting outliers using correlation didn’t provide us with a robust technique to separate out the outliers throughout. In the first step, we performed variable transformation for the variable with large number of outliers. And towards the end of data cleaning process we performed the Mahalanobis analysis to identify, separate and treat the outliers differently.

## Variable Transformation

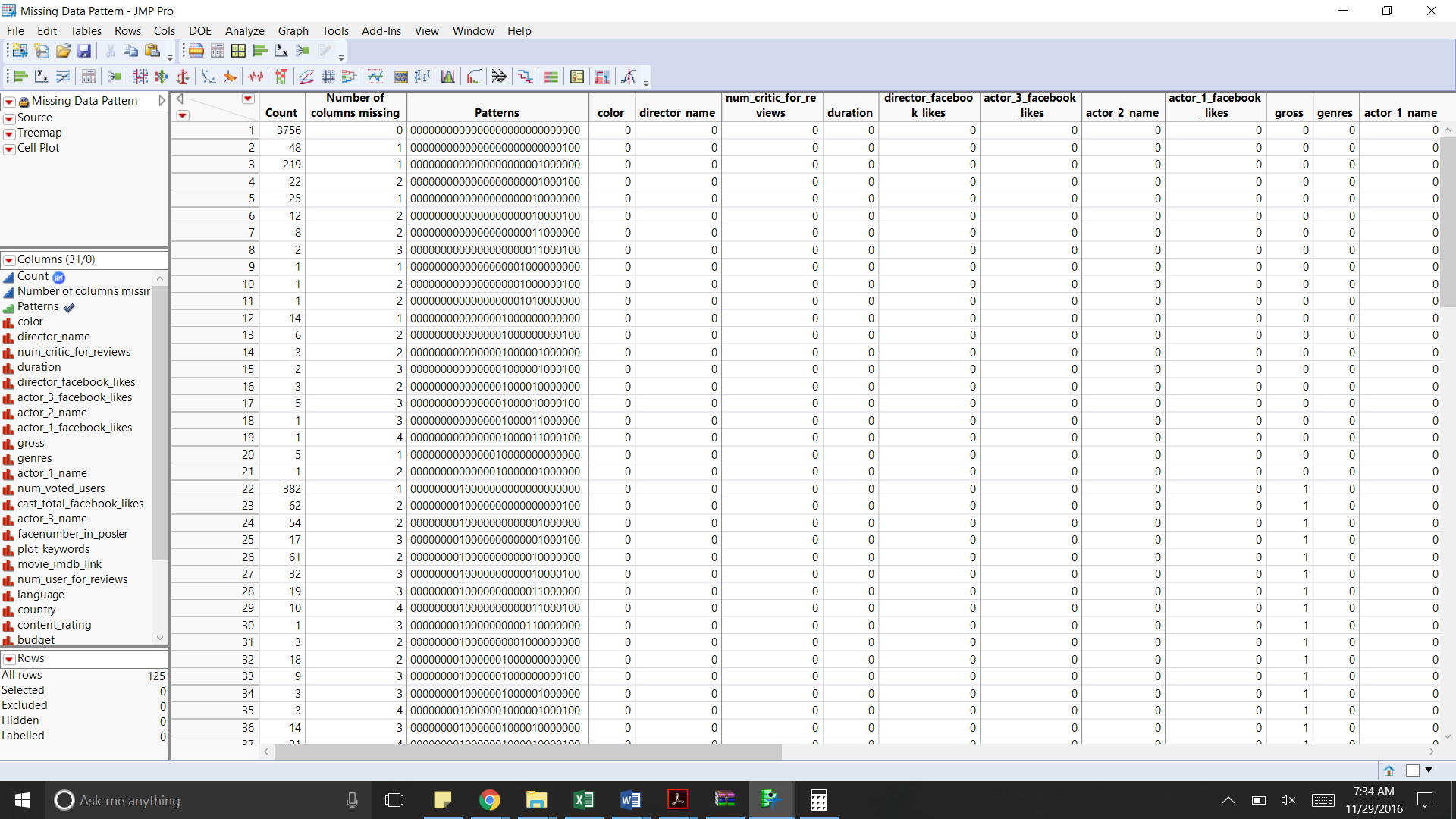
Inferring from the univariate analysis performed in the previous section, we performed the variable transformation using the Continuous Fit option available in the analyze options which automatically chose the best fit for the given variable. The table below shows the type of transformation chosen for each of the variable and the histogram before and after the transformation. The transformations below aim to reduce the number of outliers for each of the distributions. But whether these outliers work the best for the model, can only be checked by comparing models with and without variable transformations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Distribution (Before) | Transformation | Transformation Formula | Distribution (After) |
| num\_critic\_for\_reviews |  | Weibull Density | Weibull Density (:num\_critic\_for\_reviews, 1.1287372058788, 146.344395947755, 0) |  |
| cast\_total\_facebook\_likes |  | Johnson SI | (Log ((: cast\_total\_facebook\_likes - (-79.7229205148896)) / 1) \* 0.671418381540576  + (-5.51799080705756)) \* 1 |  |
| Num\_user\_for\_reviews |  | Weibull Density | Weibull Density (:num\_user\_for\_reviews, 0.849139065705588, 248.688180293987, 0) |  |

# Missing Data Handling

## Missing Data Pattern

Using the missing data pattern option in JMP we tried identifying the rows and columns with high missing values. Below is the screenshot and excel file with the data for the same.



**Figure 8.1**

* Around 75% of the data has no values missing
* The columns gross (~18%) and budget (~10%) account for the maximum rows with missing values with around 4% of the data having both the values missing.
* Since the dataset has only 5043 rows with 3 levels of classification and since only 75% of values are missing, we would be keeping all 5043 rows for further analysis.

# Data Reduction

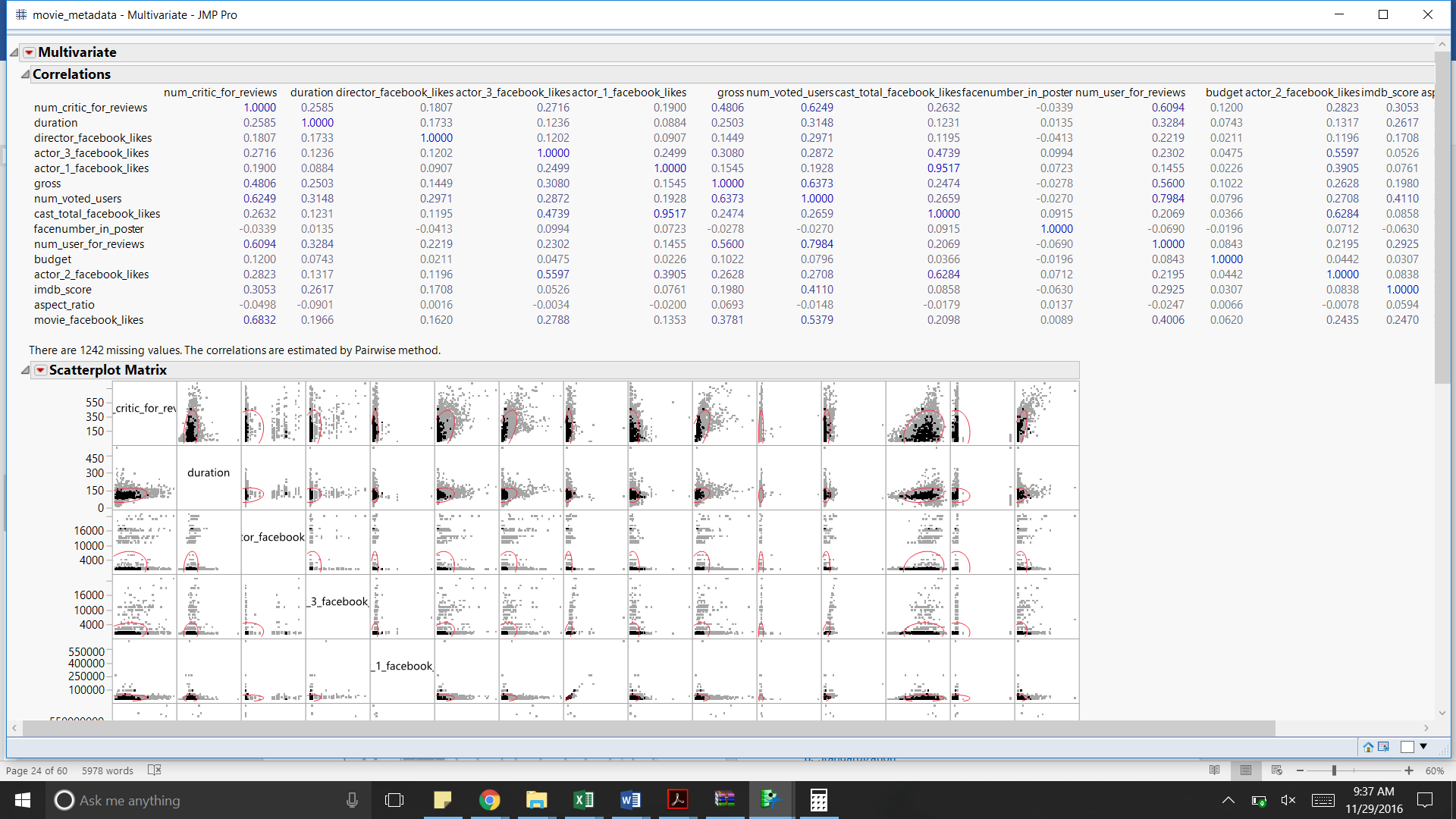
In this section, using various methods like checking for sparsity, high correlation and principal component analysis, we aim to reduce the complexity of the model by reducing the number of fields that won’t be useful in the model.

## Highly Sparse Variables

Based on our analysis there were no variables identified which were sparse.

## Highly Correlated Columns

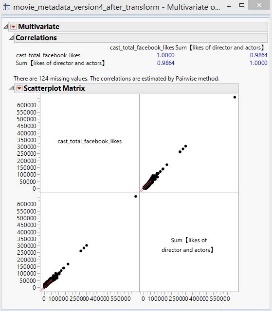
We performed the correlation analysis by using the multivariate option in the analyze menu. Below are the results of the correlation analysis.



**Figure 9.1**

Based on the above analysis we observe that the following columns show high correlation

* Since actor\_1\_facebook\_likes are highly correlated with cast\_total\_facebook\_likes (~0.95), we would be dropping this column
* Also on combing Facebook likes of all the cast and directors, this new metrics is coming to be highly correlated to the Facebook likes for the director (~0.99), for this reason we have dropped the column director\_facebook\_likes.



**Figure 9.2**

* Also, the columns pairs num\_voted\_users & num\_user\_for\_review and movie\_facebook\_likes & num\_critic\_for reviews have a good correlation of ~0.80 and ~0.68 respectively. Based on this analysis while running the models we will keep an eye on these columns so to check if the presence of both is having any effect on the model.
* One interesting correlation that we noticed was that for the movies released before 1961 the imdb score is highly correlated to the number critic reviews (~0.7267)

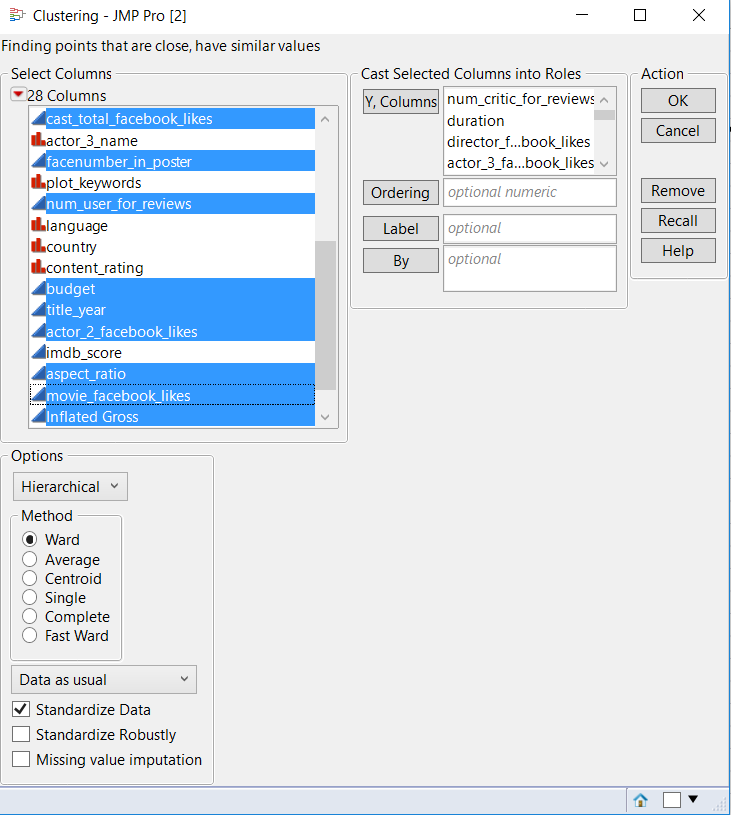
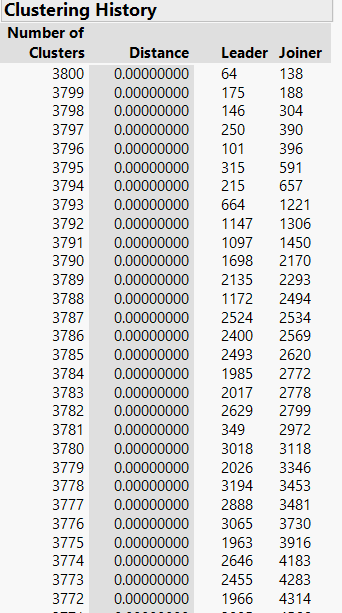
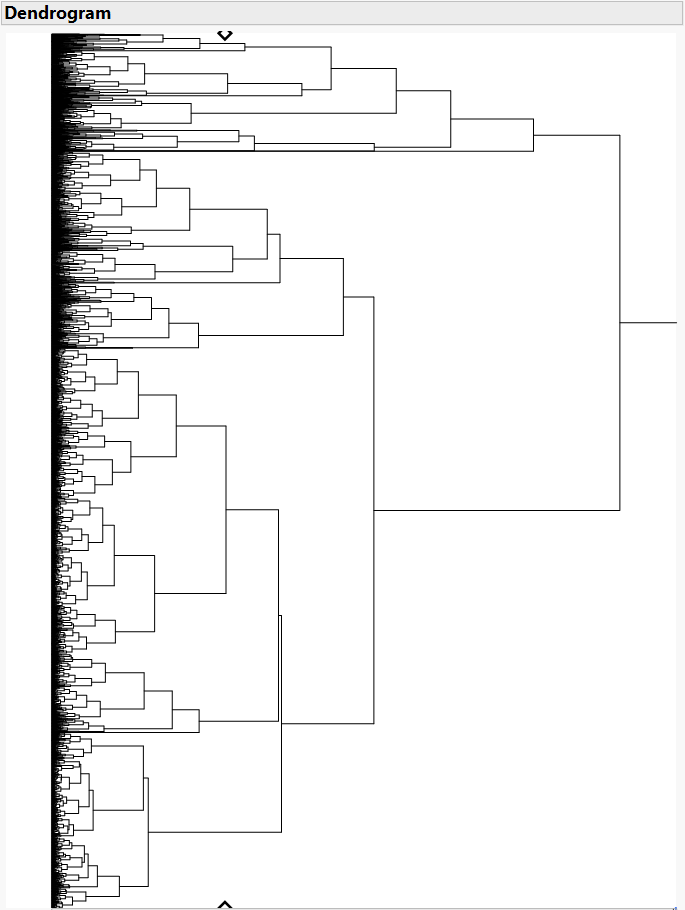
# 10. Standardization

Since our current data contains few variables in the range of 0 and 1 and other set of variables with values going as high as 200,000, to perform various analysis techniques like clustering and principal component analysis, we need to standardize our data and get all the variables to the same scale. We have used the following formula for standardization: (Xi- Xmin)/(Xmax-Xmin).

Using the above mentioned formula we have standardized the following continuous variables : num\_crirtic\_reviews, duration, Inflated gross, cast\_total\_facebook\_likes, num\_voted\_users, facenumber\_in\_posters, num\_user\_for\_reviews, Inflated budget, aspect\_ratio, movie\_facebook\_likes.

# 11. Clustering

Using the clustering option under Multivariate options under Analyze we performed clustering. Below is the dendrogram for the same.Based on this we could possibly divide our data into 3 clusters but due to a low count of rows in the data, we won’t be proceeding with this approach

**Figure 11.1 Figure 11.2 Figure 11.3.3**

# 12. Principal Component Analysis

For performing principal component analysis, we tried understanding the possible natural grouping in the whole dataset and for data separately based on the advent of Facebook (2004). Focusing on the continuous variables, initially we considered the following parameters - num\_critic\_for\_reviews, director\_facebook\_likes actor\_3\_facebook\_likes, actor\_1\_facebook\_likes, num\_voted\_users, cast\_total\_facebook\_likes facenumber\_in\_poster, num\_user\_for\_reviews, actor\_2\_facebook\_likes, movie\_facebook\_likes.

For all the trials, none of the analysis could capture more than 32% of the variation because of which we have not used PCA in our analysis.

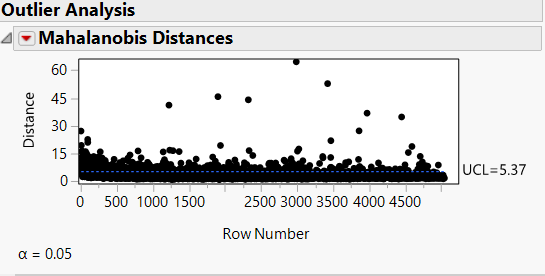
# 13. Final Outlier Treatment

For the final outlier treatment, we calculated the Mahalanobis distances for each of the row. Based on the UCL limit α=0.05 and how far off a row is from the UCL, we created 2 datasets

● Dataset 1: Includes all the rows within the UCL= 5.37

● Dataset 2: Includes all the rows

While building the model, we will build it in on the 2 datasets, based on its performance on the validation set we will select the best model.



**Figure 13.1**

# 14. Modeling

Using the dataset created in the previous steps, we built 2 sets of models

1. Classification Models – These models classify the movies into three different categories – poor, average and good
2. Regression Models – These models predict the IMDB score using the selected predictor variables

## 14.1. Classification Models

The classification models were built on three datasets divided based on different time periods -

1. The entire dataset (5043 rows)
2. Movies with title\_year <2004
3. Movies with title\_year >= 2004

We also built the models with the outliers removed (based on Mahalanobis distance (α=0.05 and UCL=5.37)) to check if they have any impact on the accuracy of the model

Also, since majority of the predictor variables are based on Facebook likes and since Facebook was launched in the year 2004, we wanted to check if the advent of Facebook has had any effects on the IMDB scores associated with the movies.

Before we start to build a model, we need to determine the training, validation and test dataset.

Since the size of the dataset is 5043 only, we decided to divide the data as training and test only.

Also, as the dataset is unbalanced, we have used stratified sampling to create a balanced train and unbalanced test datasets.

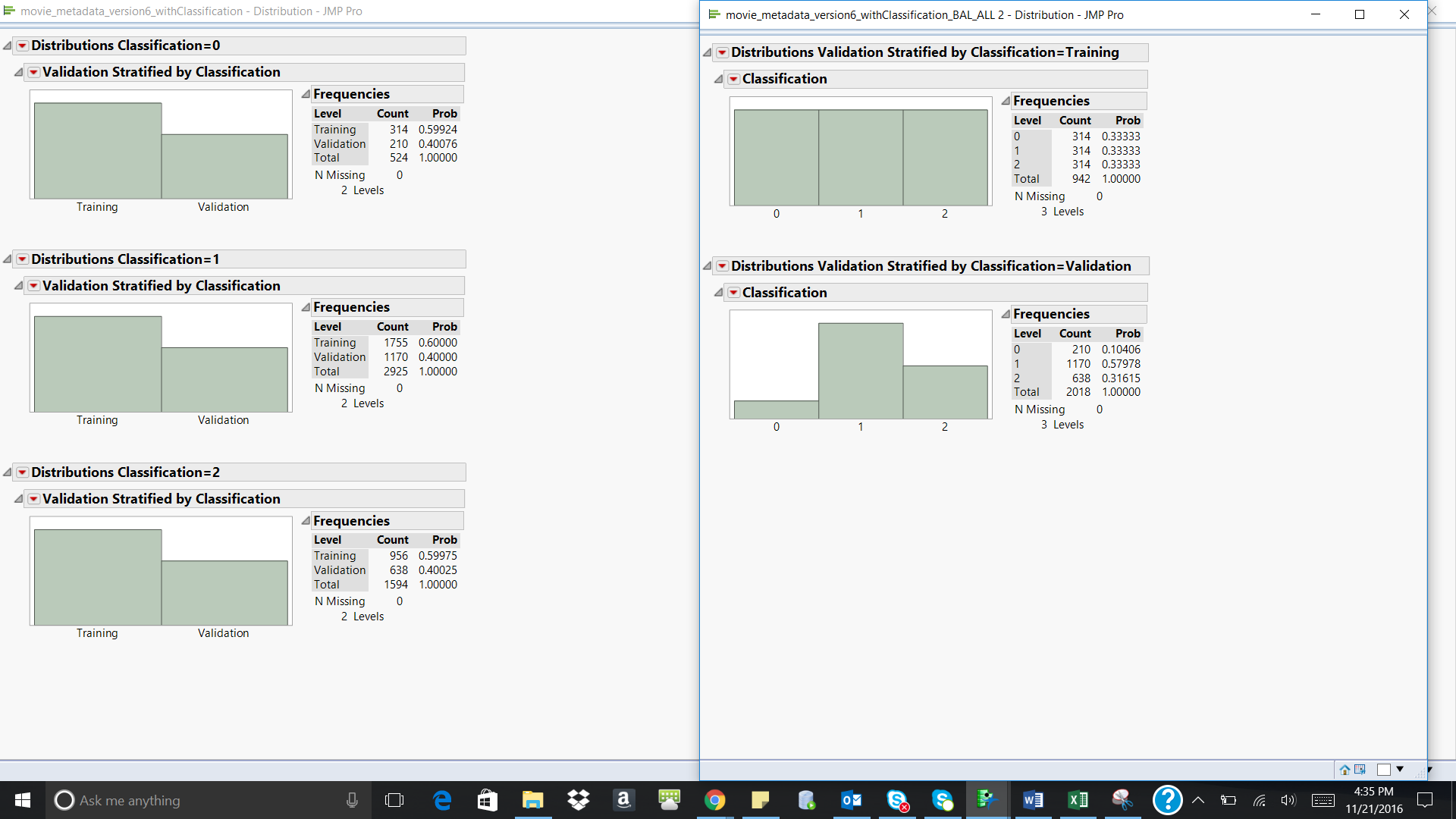
Below is the analysis for the **whole dataset with all the outliers** included.

First, we built the models taking all 5043 rows included.

We created a stratified sample based on the classification of movies and considered ‘0’ as the focal group (since this level has minimum number of instances) with the proportions set as following.

1. Training Set: 0.6
2. Validation Set: 0.4

Since the number of 0s are less as compared to 1’s and 2’s we selected that as the focal group and created a balanced training sample and unbalanced validation sample.

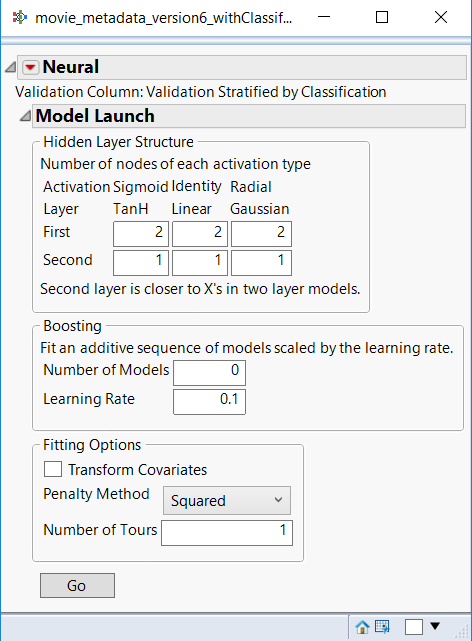
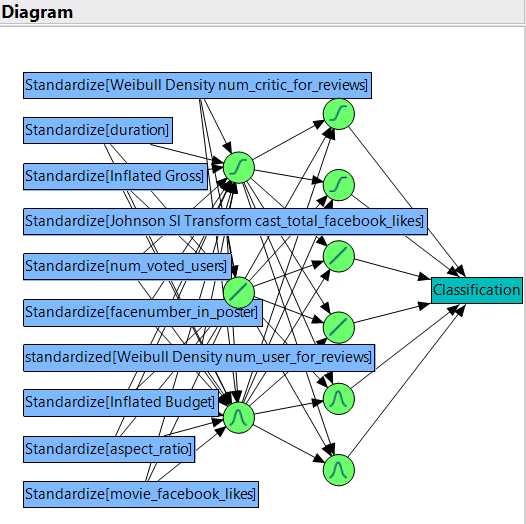


**Figure 14.3**

We have used three different classification techniques-

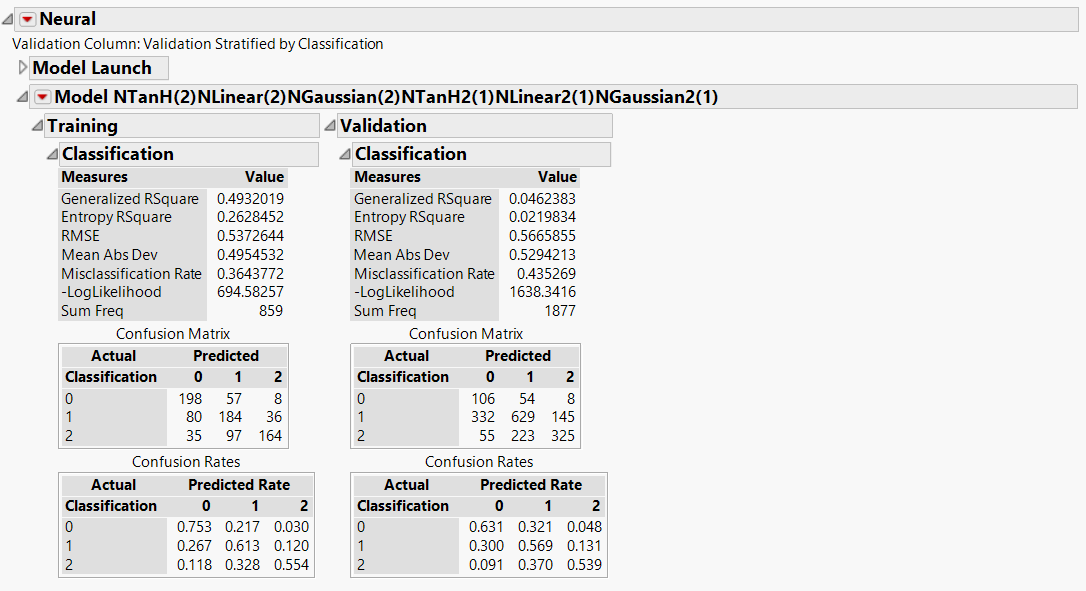
1. Neural networks
2. Discriminant analysis
3. Decision trees

For **neural nets** the following approach was used. Using the Neural options under Analyze-> Model, we ran the model for ‘classification’ using all the predictor variables. Below is the screenshot of the neural diagram

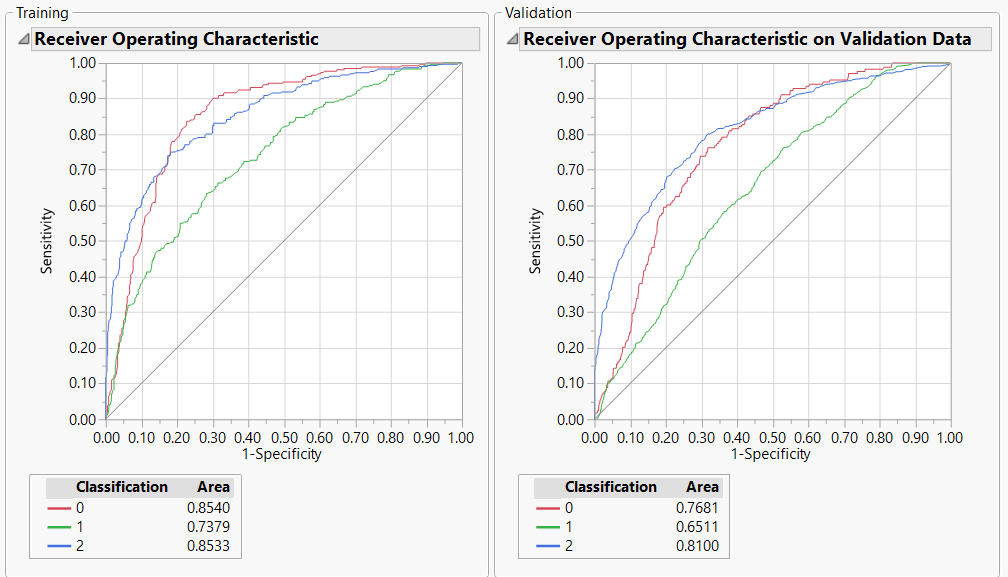
**Figure 14.4**

We ran the model for higher level of complexity but it performed at the same level as the model above. Based on this, the above model was the final selected model. The model consists of 4 layers – 1 input layer, 2 hidden layers and 1 output layer.

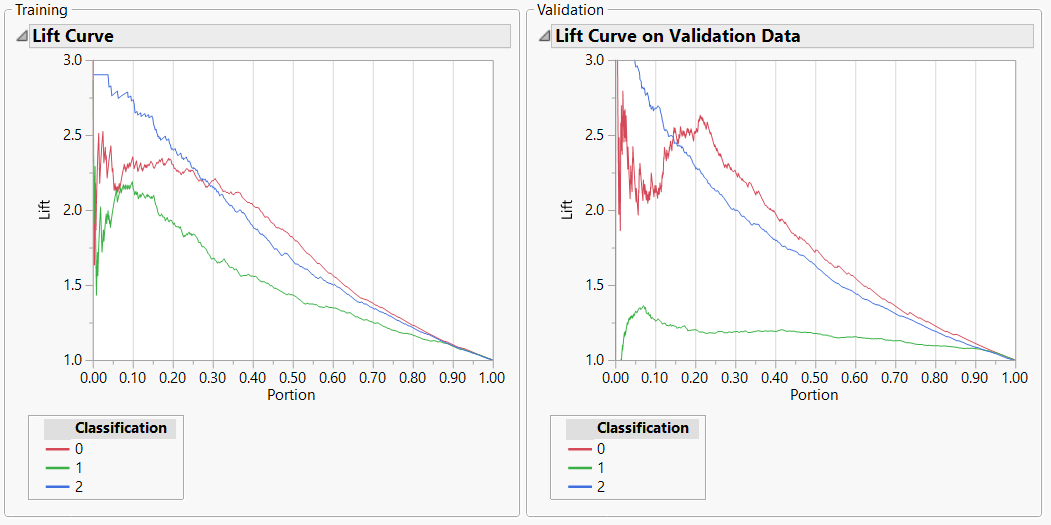


**Figure 14.5**

The model performs almost equally good on the training and validation data with an accuracy of 56.4% and the misclassification rate for bad movies being 30%, average being 46% and good being 36%. Below are the ROC and the Lift curves for the selected model



**Figure 14.6**

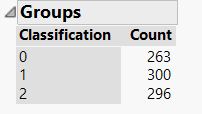
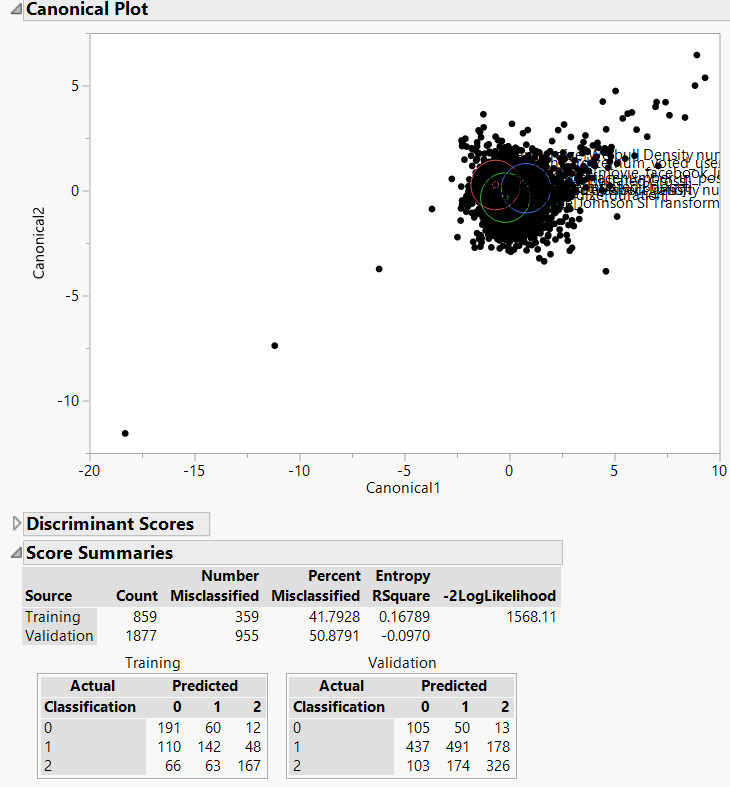


**Figure 14.7**

From the lift curve above the following observations were made:

|  |  |  |
| --- | --- | --- |
| Classification | Training data | Validation data |
| 0 | The lift value is around 2.25 on an average across the first three deciles and then constantly decreases across the other deciles | The lift value is constant for the second and the third decile and then constantly decreases across the other deciles.It varies from 2 to 2.75 in the first decile. |
| 1 | The lift value ranges from 1 to 2.1 in the first decile and then constantly decreases across all the other deciles. | Even though the lift value is low when compared to 0s and 2s, it is consistent till the seventh decile and then slowly reduces |
| 2 | The lift value is very high in the first decile but reduces consistently across all the other deciles | The lift value is more than 3 in the first decile but reduces consistently in all the other deciles |

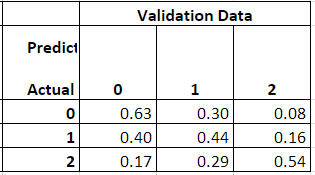
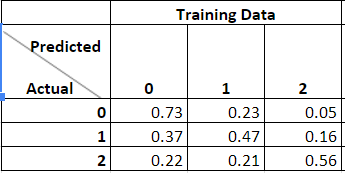
For **discriminant analysis,** we followed the following steps. Using the Discriminant option under Analyze->Multivariate, we ran the discriminant analysis for the variable ‘classification’ using only the continuous predictors (covariates). Below is the canonical plot for the analysis



**Figure 14.8**

Below are the ROC curves for the 2 models.

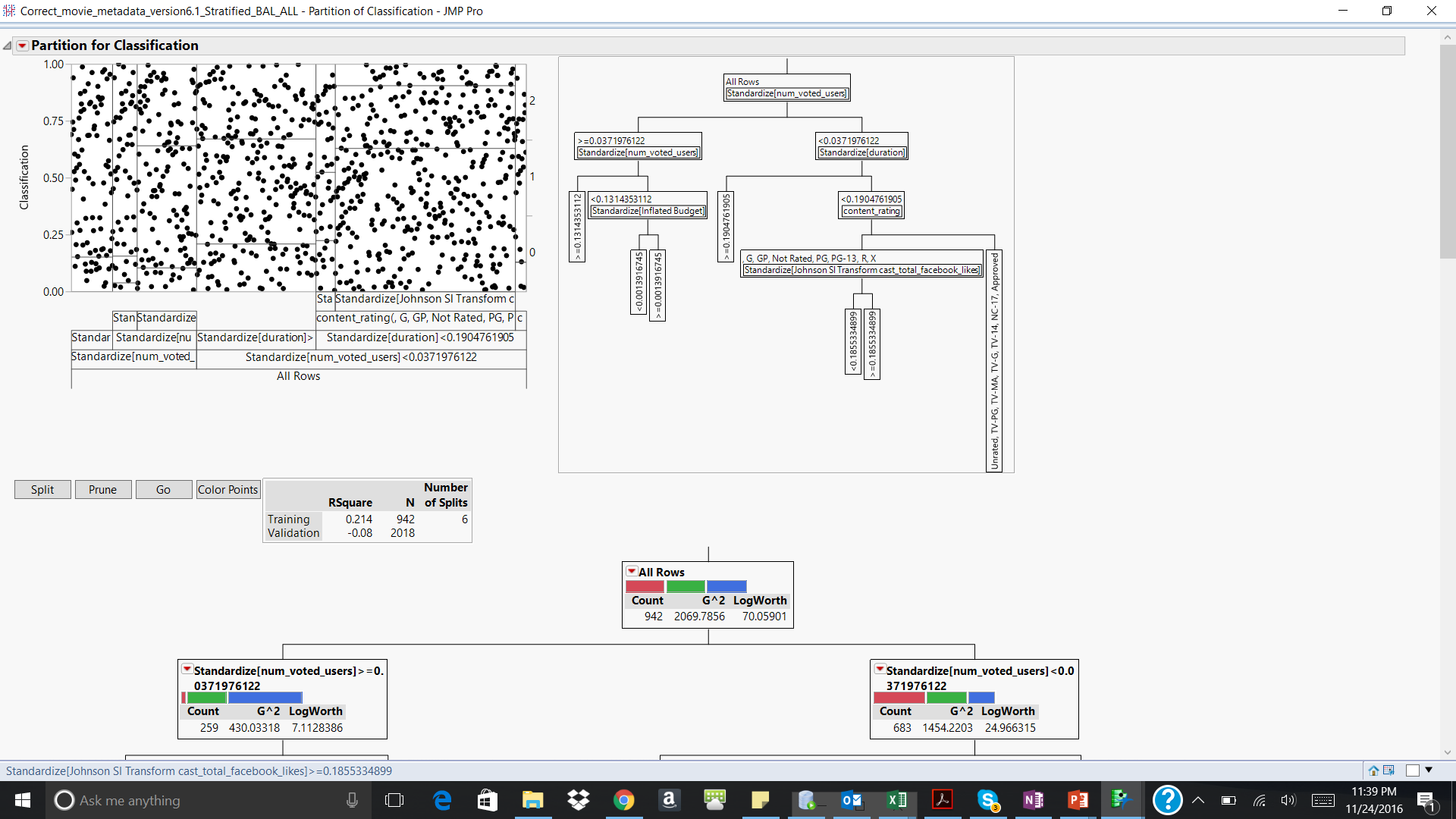
Confusion matrix is as follows



From the observations above, the model seems to perform equally well in both the Training and the validation dataset.

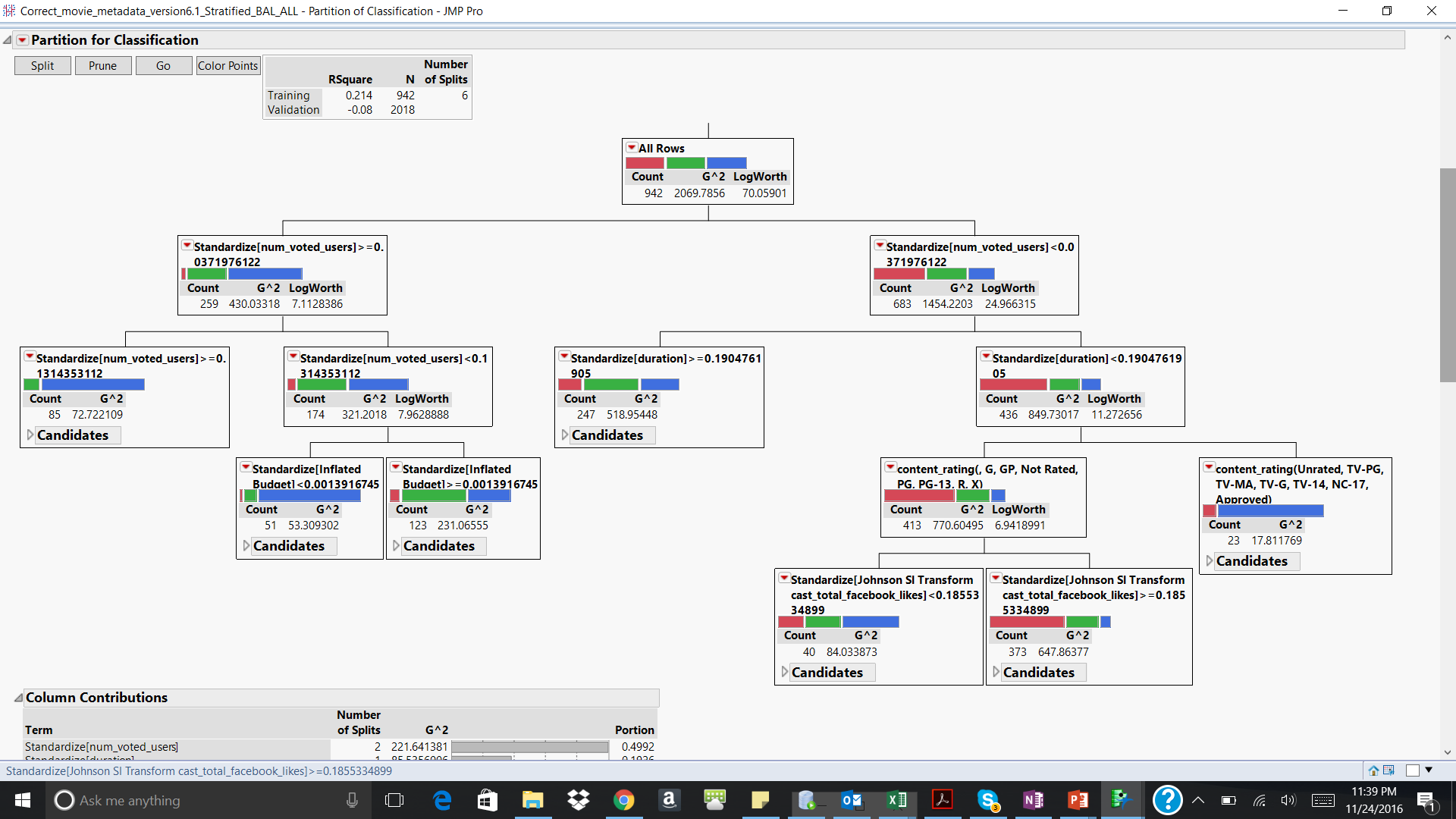
The accuracy rate in the training data set is 58.20% and the validation data set is 49.12% and the misclassification rate (validation) for bad movies being 30%, average being 46% and good being 36%. Since the performance of the model across the training and validation datasets are almost the same, the model doesn’t seem to be overfitting on the data.

For **decision tree,** we followed the following steps. Using the Partition option under Analyse -> Modeling, we ran the decision tree model for the given data set. Since the validation dataset was not being recognized by JMP, we manually went about splitting the tree.

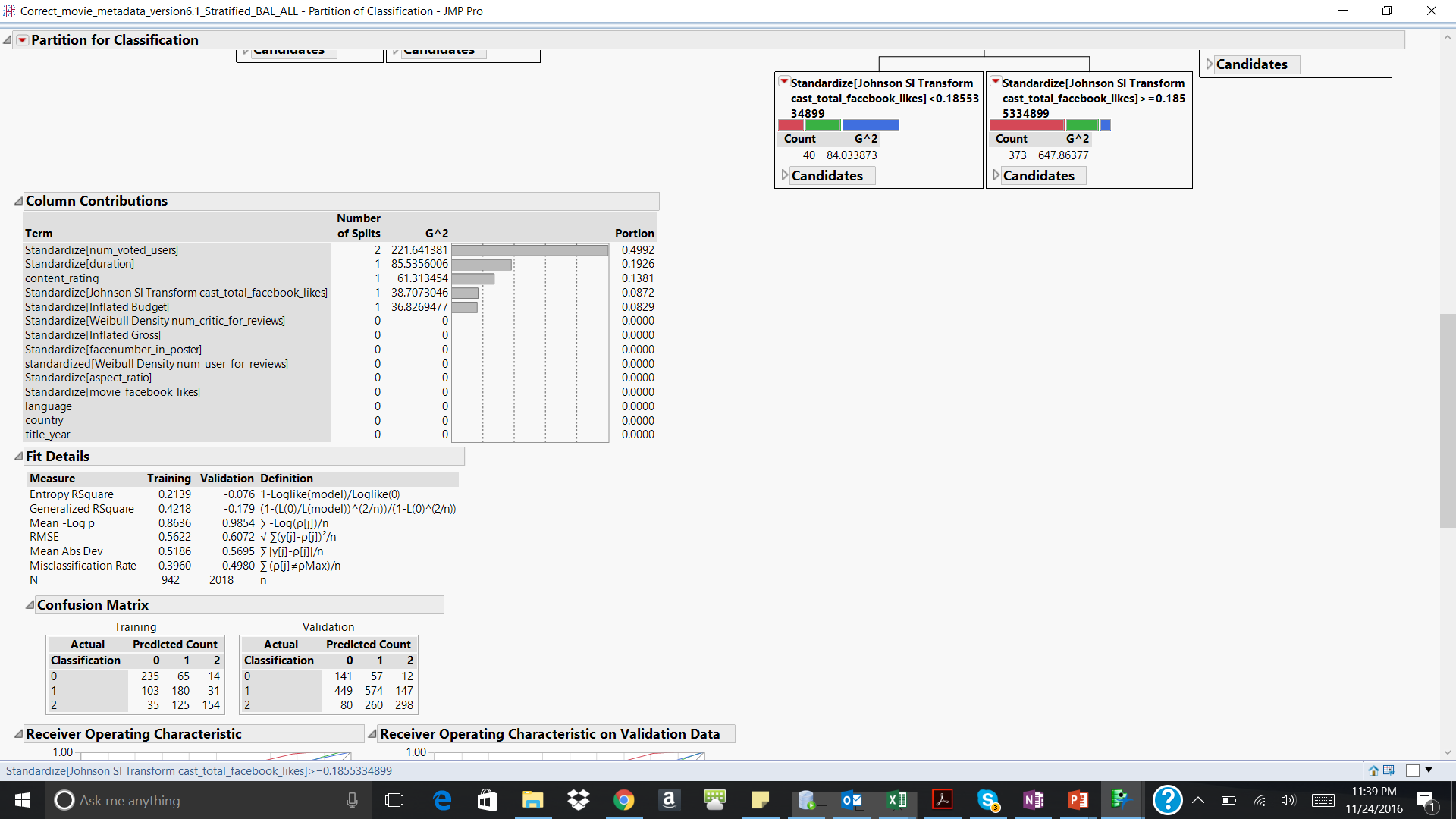


**Figure 14.9**

For the given decicion tree, the first level of split is done using num\_voted\_users, followed by duration and content rating. After the split on the variable budget there is no more contribution from the other variables. After this point the misclassification rate in the validation data also increases. Hence, there are no more splits that are executed.

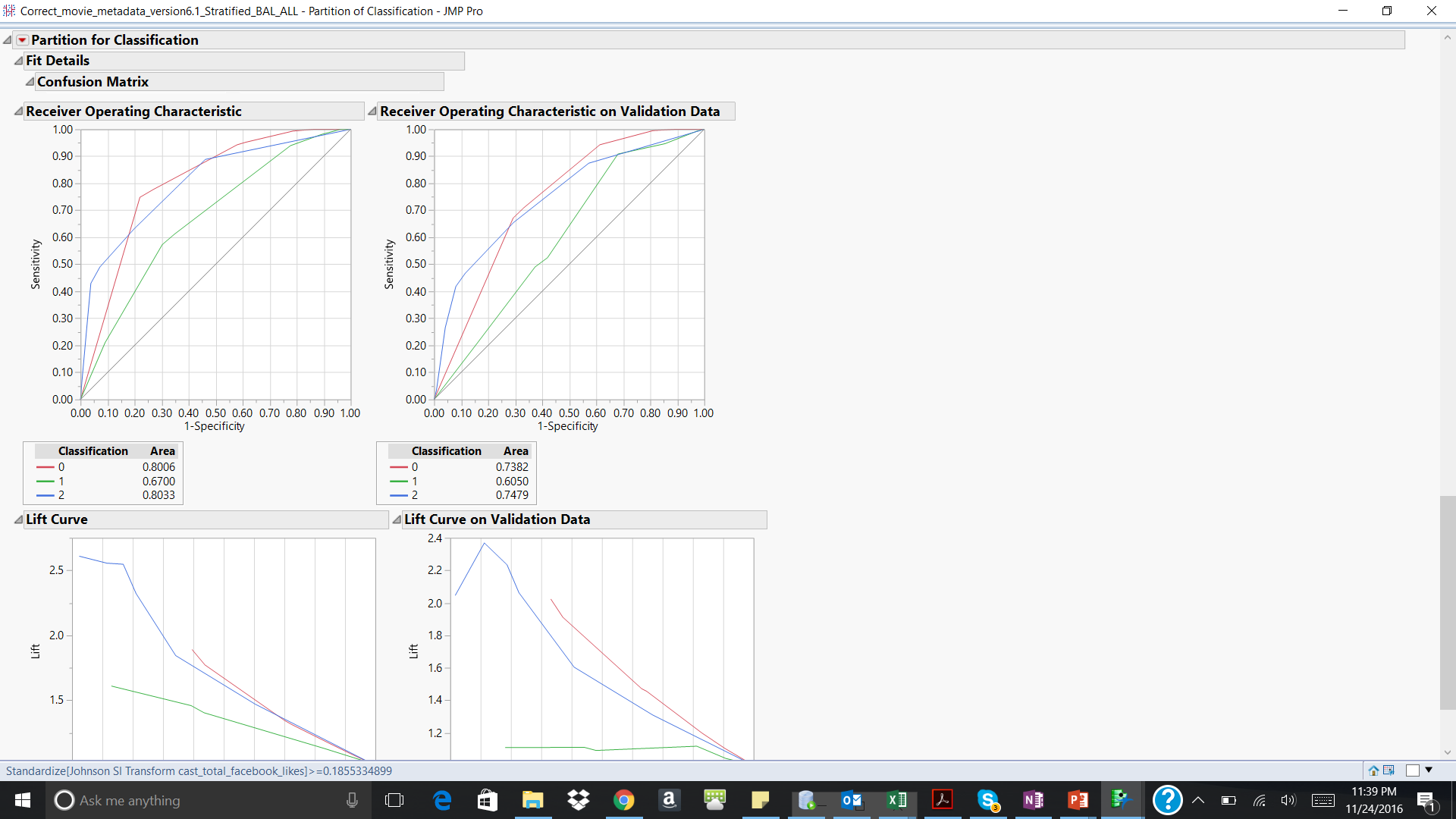


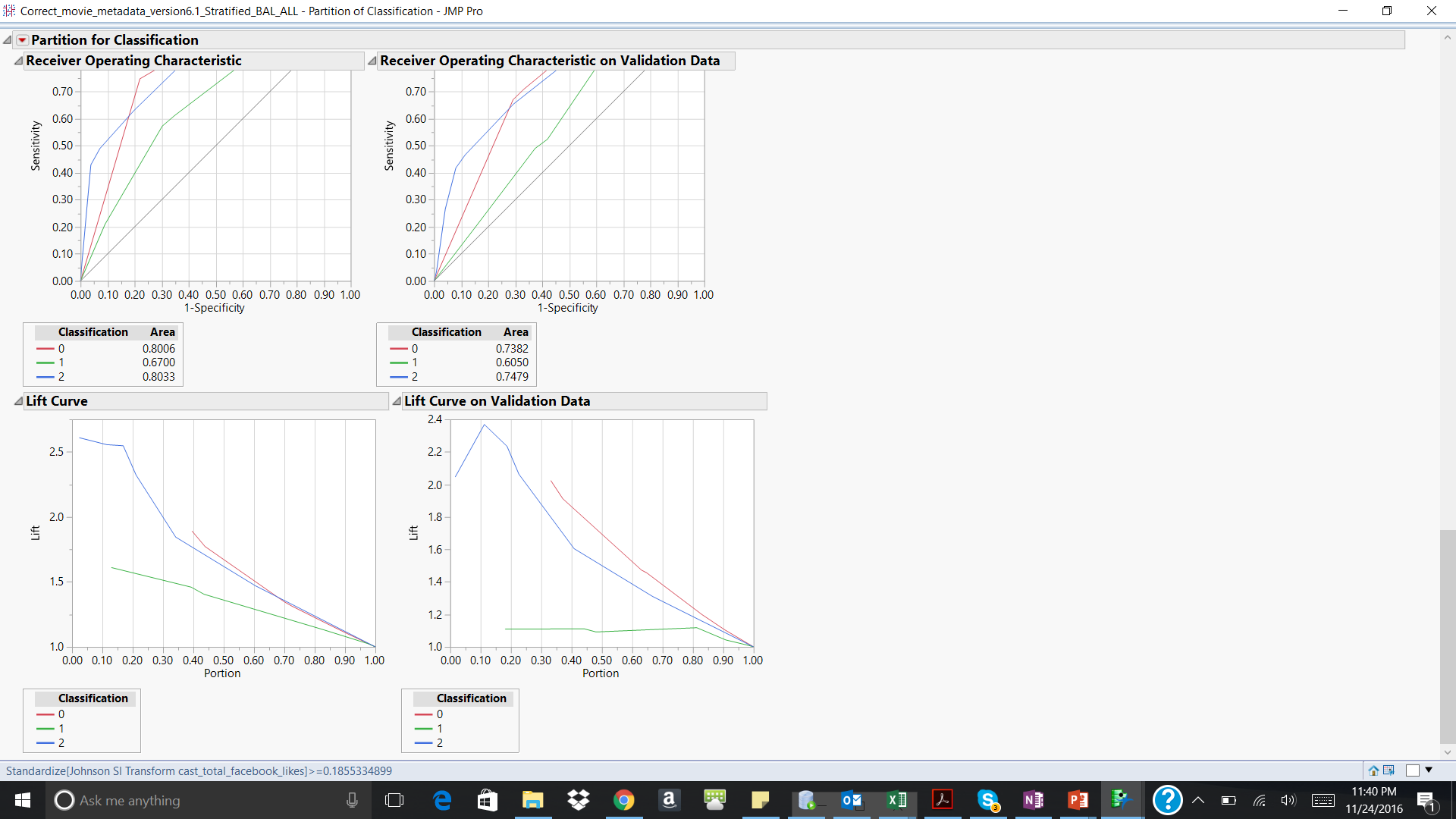
**Figure 14.10**



**Figure 14.11**

Below are the ROC and the lift curves for training and validation. In the validation data, the AUC for the Good(2), Average(1) and Poor(0) movies is almost comparable. Hence, the model seems to predict each of these categories at a comparable accuracy rate. Also the misclassification rates for validation are as follows 29%, 57%, 35% for bad, average and good respectively.

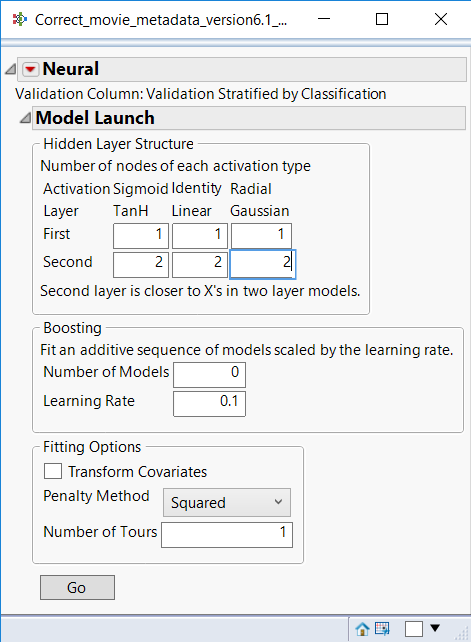
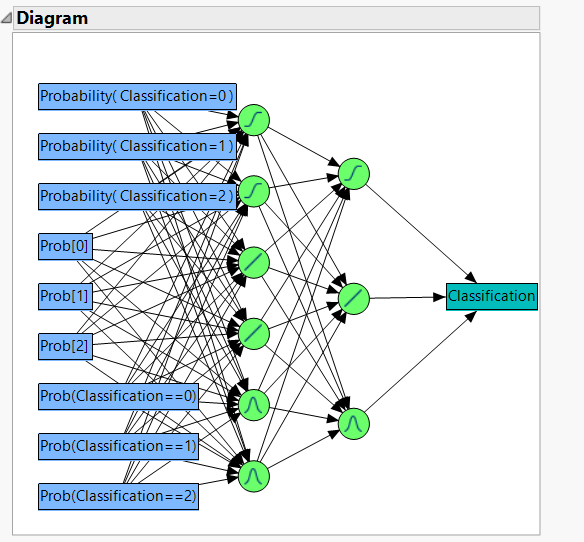




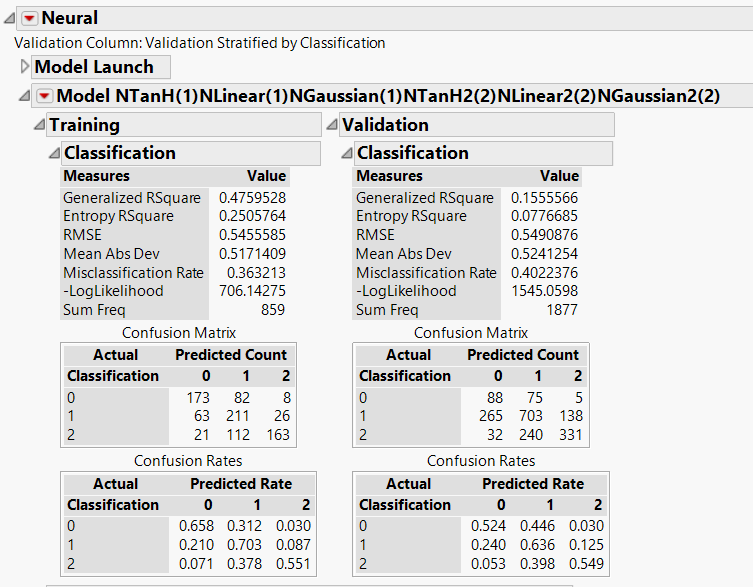
**Figure 14.12**

In the validation data, the model seems to perform well for 2s in the first decile and then the performance constantly decreases in the other deciles. The lift value for 1s is negligible in the 1st two deciles, after which there is a lift of around 1.1 which remains constant thereafter for the next two deciles and decreases. The lift value for 0s is negligible in the 1st three deciles, after which there is a lift of around 2 which decreases thereafter.

For **ensemble model**, we used neural networks on the outputs of the models built above. Choosing the Neural network to build the ensemble model, below is the screenshot of the same.

**Figure 14.13**



**Figure 14.14**

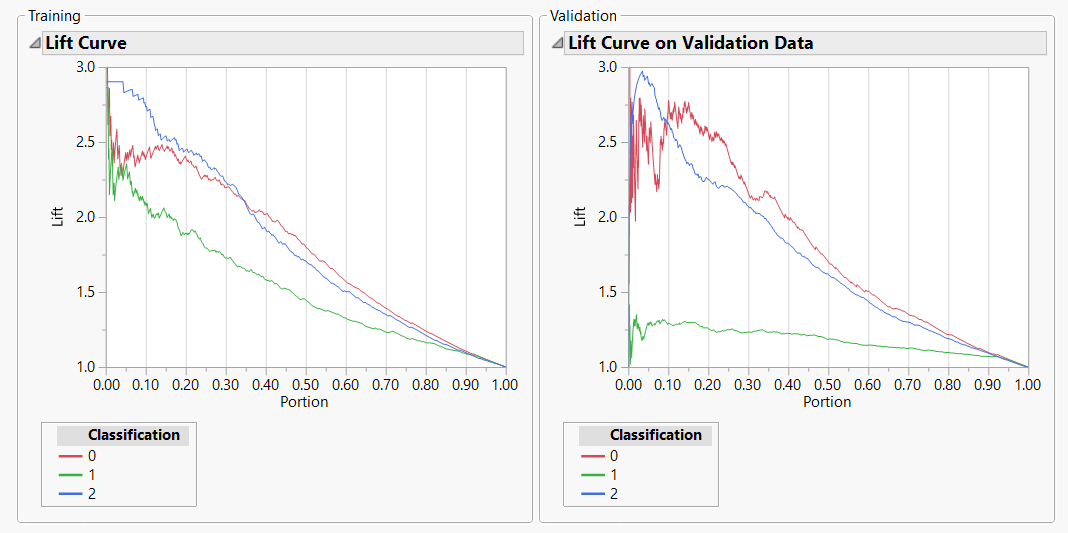
The accuracy of the model on the validation data is 60% which is better when compared to discriminant analysis, neural networks and decision tree individually. But misclassification rate (for validation) which seems to high when compared to the neural net is as follows 30%, 53%, 29% for bad, average and good movies respectively.

Below are the ROC and the lift curves for the model



**Figure 14.15**

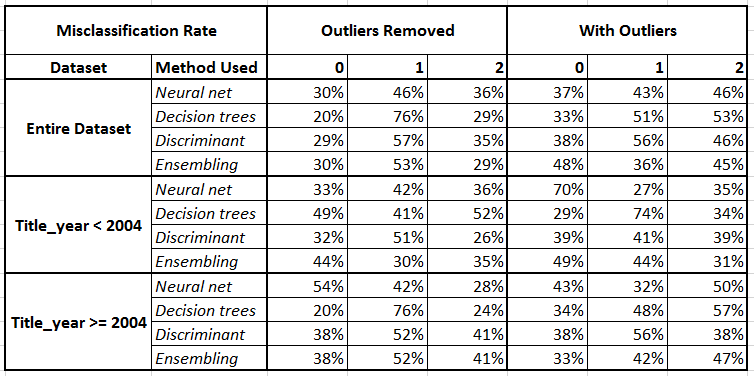
In the validation data, the AUC is higher for the 2s (‘Good) when compared to 1s(‘Average’) and 0s (‘Poor’) . Hence, the model seems to predict the Good movies with more accuracy when compared to ‘Average’ and ‘Poor’ ones.



**Figure 14.16**

In the validation data, the model seems to perform well for 0s and 1s in the first decile and then the performance constantly decreases in the other deciles. The lift value for 1 seems to increase in the 1st decile after which it remains constant.

Below are the misclassification rates for all the models run under classification

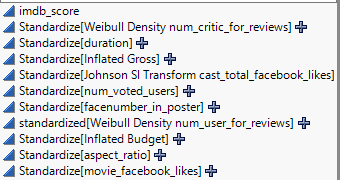


Post analyzing these models, the **neural nets** seem to be working the best for each set in the matrix. Post this we took all the neural nets and compared the results for with outliers vs without outliers and entire dataset vs the segmented dataset (based on title year)

**Note**: The details of all the models can be found in the Model Appendix.docx. Also since the classification dataset has 3 levels, in the appendix we have taken the IMDB dataset and converted into classification problem with 2 levels to test logistic regression on the same data.

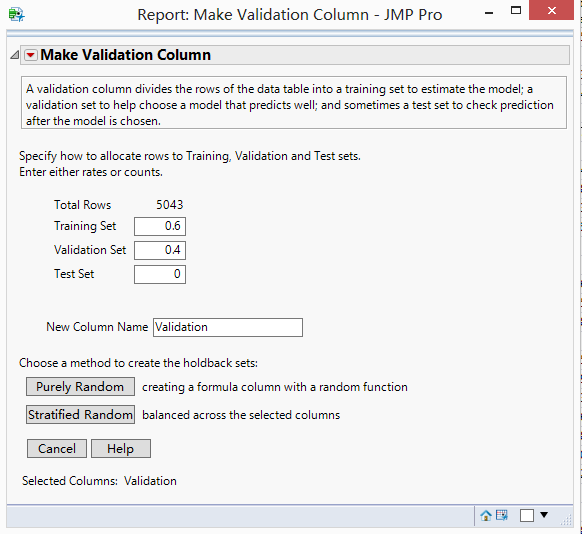
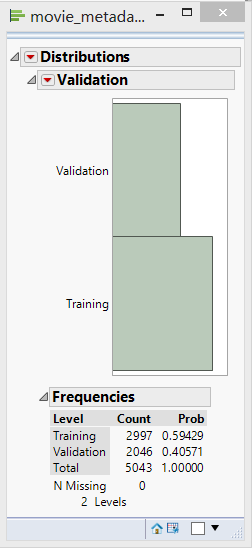
## 14.2 Regression Models

Using the dataset created in the exploratory analysis, the following were selected as the explanatory variables.

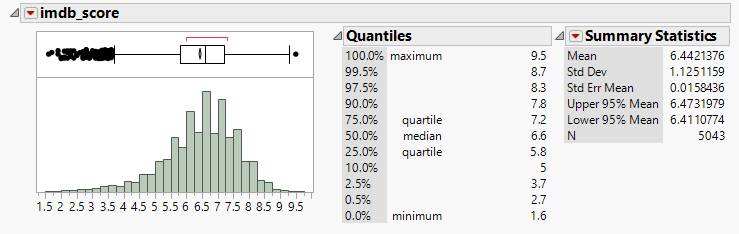


**Figure 14.17**

Firstly, we split dataset into two groups, Training and Testing and would adopt K-folder testing in stepwise to decrease the possibility of overfitting.

**Figure 14.18**



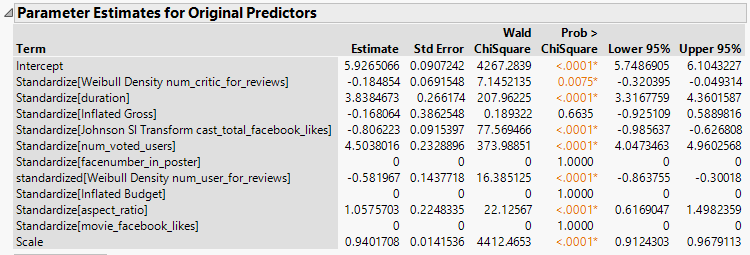
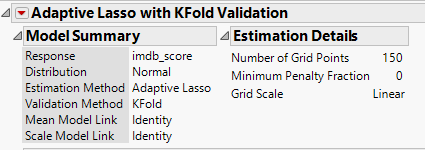
**Figure 14.19**

Given the distribution of imdb\_score, the average mean is almost equal to median of population and upper 95% mean is similarly same as lower 95% mean. So, we can assume that the target variable follows normal distribution.

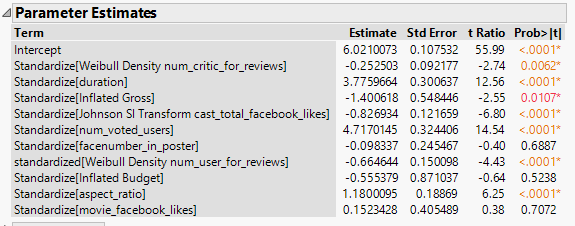
Based on that observation, we mainly check the three models for training and validation data,

1. Generalized Regression for Normal distribution target variable.(GR)
2. Standard Least Squares with minimal report. (SLS)
3. Neural Net (NN)

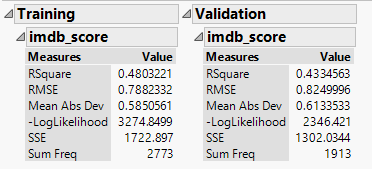
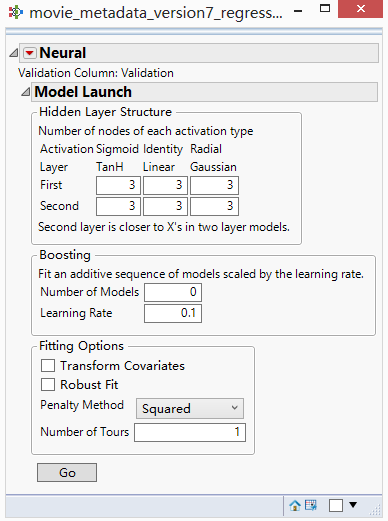
To increase accuracy of model, we use stepwise with k-folder cross validation to determine the variables which contribute most in model. Based on this analysis, all the variables were used for model building exercise. Using the **whole dataset with ouliers**, below are the summaries for the 3 models.



**Figure 14.20: Generalized Regression (3)**

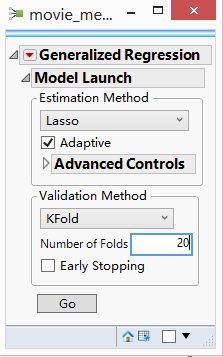
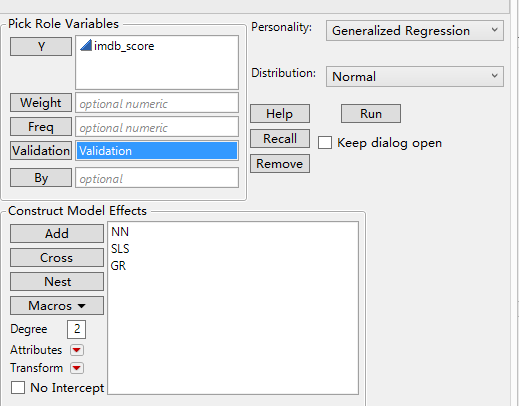


**Figure 14.21: Standard Least Squares (3):**



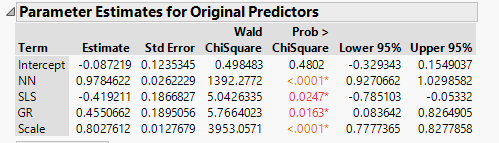
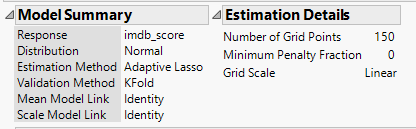
**Figure 14.22: Neural Net (3):**

**Figure 4.10 : Ensemble Generalized Regression (1):**



**Figure 14.23**

**Figure 4.10 : Ensemble Generalized Regression (2):**



Below are the performance metrics of each of the model run:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RSquare Calculation** | **GR** | **SLS** | **NN** | **Ensemble GR** |
| **Training** | 0.348972 | 0.352135 | 0.4954444 | 0.545071 |
| **Rank in Training** | 4 | 3 | 2 | 1 |
| **RSquare**  **Validation** | 0.341445 | 0.341058 | 0.43913 | 0.499421 |
| **Rank in Validation** | 3 | 4 | 2 | 1 |

We can see result based on RSquare metric. The Ensemble Generalized Regression model is best model for this dataset.

Below are the performance metrics of each of the models run for movies released **before 2004 and on data having outliers:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RSquare Calculation** | **GR** | **SLS** | **NN** | **Ensemble GR** |
| **Training** | 0.294801 | 0.301529 | 0.604691 | 0.642202 |
| **Rank in Training** | 4 | 3 | 2 | 1 |
| **RSquare**  **Validation** | 0.278873 | 0.276995 | 0.4856404 | 0.522066 |
| **Rank in Validation** | 4 | 3 | 2 | 1 |

We can see result based on RSquare metric. The Ensemble Generalized Regression model is best model for this dataset.

Below are the performance metrics of each of the models run for movies released **after 2004 and on data having outliers:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RSquare Calculation** | **GR** | **SLS** | **NN** | **Ensemble GR** |
| **Training** | 0.303529 | 0.304471 | 0.442976 | 0.523765 |
| **Rank in Training** | 4 | 3 | 2 | 1 |
| **RSquare**  **Validation** | 0.327148 | 0.321649 | 0.4128989 | 0.535395 |
| **Rank in Validation** | 3 | 4 | 2 | 1 |

We can see result based on RSquare metric. The Ensemble Generalized Regression model is best model for this dataset.

Below are the performance metrics of each of the models run for movies **on data without outliers:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RSquare Calculation** | **GR** | **SLS** | **NN** | **Ensemble GR** |
| **Training** | 0.310591 | 0.312118 | 0.453157 | 0.454684 |
| **Rank in Training** | 4 | 3 | 2 | 1 |
| **RSquare**  **Validation** | 0.305628 | 0.300393 | 0.465969 | 0.466623 |
| **Rank in Validation** | 4 | 3 | 2 | 1 |

We can see result based on RSquare metric. The Ensemble Generalized Regression model is best model for this dataset.

Below are the performance metrics of each of the models run for movies released **before 2004 and on data without outliers:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RSquare Calculation** | **GR** | **SLS** | **NN** | **Ensemble GR** |
| **Training** | 0.352083 | 0.354167 | 0.6263621 | 0.588542 |
| **Rank in Training** | 4 | 3 | 1 | 2 |
| **RSquare**  **Validation** | 0.360947 | 0.357988 | 0.4424563 | 0.460059 |
| **Rank in Validation** | 3 | 4 | 2 | 1 |

We can see result based on RSquare metric. The Ensemble Generalized Regression model is best model for this dataset.

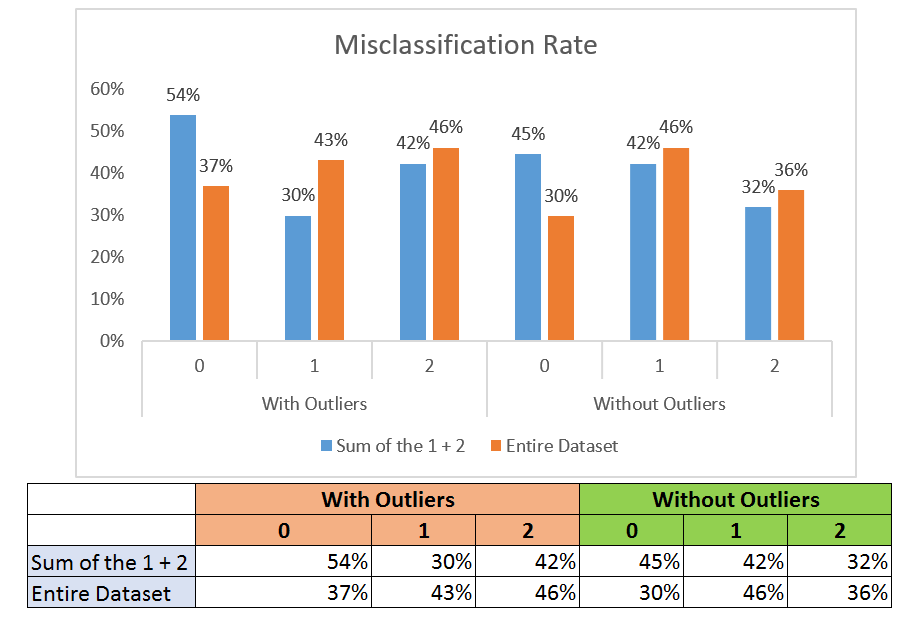
Below are the performance metrics of each of the models run for movies released **after 2004 and on data without outliers:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| RSquare Calculation | GR | SLS | NN | Ensemble GR |
| Training | 0.319962 | 0.319962 | 0.499522 | 0.505253 |
| Rank in Training | 4 | 3 | 2 | 1 |
| RSquare  Validation | 0.348578 | 0.348578 | 0.551298 | 0.55377 |
| Rank in Validation | 4 | 3 | 2 | 1 |

We can see result based on RSquare metric. The Ensemble Generalized Regression model is best model for this dataset.

# 15. Conclusion

# 15.1 Classification



**Figure 15.1**

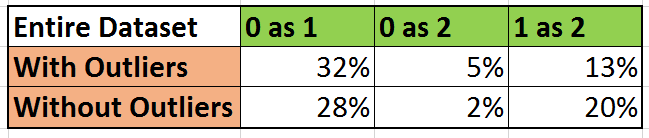
**With Outliers vs Without Outliers**

* Misclassification rates seems to increase for the 'Average' movies when the outliers are excluded.
* Misclassification rates seems to decrease for the 'Good' and 'Poor' movies when the outliers are excluded.
* Outliers seem to have an effect in varying magnitudes on 'poor', 'average' and 'good' movies

**Whole Dataset vs Dataset segmented based on title year**

* Misclassification Rates seems to increase for 'Poor' rated movies when we do separate analysis for movies released before and after 2004(Year of launch for Facebook)
* Misclassification Rates seems to decrease for 'Average' and 'Good' rated movies when we do separate analysis for movies released before and after 2004(Year of launch for Facebook)

We can conclude by saying that when misclassification occurs for 'Bad' rated movies, it might be an annoyance the viewers. Whereas if a 'good' movie is classified as 'poor’ rated or 'average' rated movie, and if a 'average' rated movie is classified as 'Good' or 'Bad', might not be high (unless a movie is at the brink of average and bad classification)



Also on analyzing the misclassification rate for the selected models, we observe that ‘bad’ movies being classified as 'Average and Good' movies is comparatively lesser in the model without outliers. Average movies being classified as 'Good' movies is comparatively lesser in the model with outliers. But since Poor movies being classified as 'Average and Good' movies increases the annoyance factor when compared to Average movies being classified as 'Good' movies, the model without outliers seems to perform better than the one with outliers included.

Based on the above analysis it is better to better to do the analysis for the entire dataset rather than dividing the dataset with respect to time period. But we need to perform a better in-depth understanding of the misclassification rate to get a better understanding of the models in absence of a cost factor.

# 15.2 Regression

Through regression analysis, we can see no matter dataset’s property, the Ensemble Model Always has the best result. So we list table here to compare the result:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Data After 2004 | Data Before 2004 | Whole Data |
| RSquare of Ensembel Gr in Training | 0.523765 | 0.642202 | 0.545071 |
| RSquare of Ensembel Gr in Validation | 0.535395 | 0.522066 | 0.499421 |

Comparison between data without outlier:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Data After 2004 | Data Before 2004 | Whole Data |
| RSquare of Ensembel Gr in Training | 0.505253 | 0.588542 | 0.454684 |
| RSquare of Ensembel Gr in Validation | 0.55377 | 0.460059 | 0.466623 |

We can conclude following issues:

1. The chosen outliers contain meaningful information in modeling, that is the reason why models of data with outliers perform better than one of data without outliers
2. Also even though splitting data based on 2004 performs better, at an overall level the datasets with or without split based on the year perform almost the same.

# 16. Next Steps

Beyond the analysis that we have done on this project. Some of the things that we can further look at are

* Analyze how the IMBD score affects the profits the movie makes to understand how a misclassification of a newly released movie might affect the business
* For the current analysis we have used all the variables in the analysis, work towards reducing the number of variables like Facebook based attributes, country and language to build simpler models
* Since the model with outliers are performing better in classifying the average movies and the model without outliers are performing better in classifying the bad and good movie. Explore building an ensemble of the two types of analysis
* Perform an analysis like the outlier analysis by including and excluding the variables which have correlations in the range of 0.5-0.8
* Analyze the effect of the variable transformation done for handling outliers by comparing models with transformed variables and models without transformed variables
* Use text mining techniques for the fields genre and plot\_keywords to derive new meaning full variables from these fields
* Based on using the cut-off of year 2004, the whole dataset seems to be performing better when compared to the analysis performed post breaking the dataset. Analyze the validity of this claim by choosing various cut offs between 2000 and 2008 as since the effects of Facebook might even be noticed on the movies released before the advent of Facebook and also because Facebook did not become popular before the years 2008-2009
* Based on the current analysis, the variable language and country don’t seem to be the top factors in deciding the IMDB score of the movies. Analyze a bigger and more diverse set of data to check the validity of this claim.
* Using a bigger dataset, perform an analysis on the outliers to check if a different model might work better for them
* Look at websites beyond IMDB (like Rotten Tomatoes) to make the analysis more robust

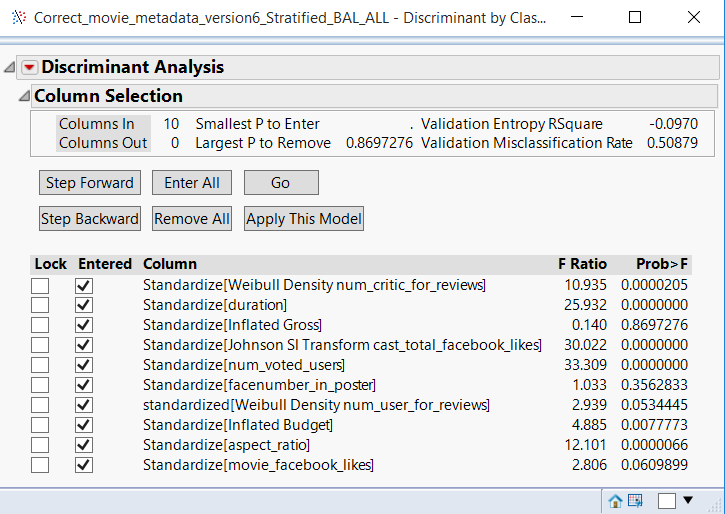
# 17. Appendix

## 17.1 Data Dictionary

The data dictionary below summarizes all the columns present in the data.

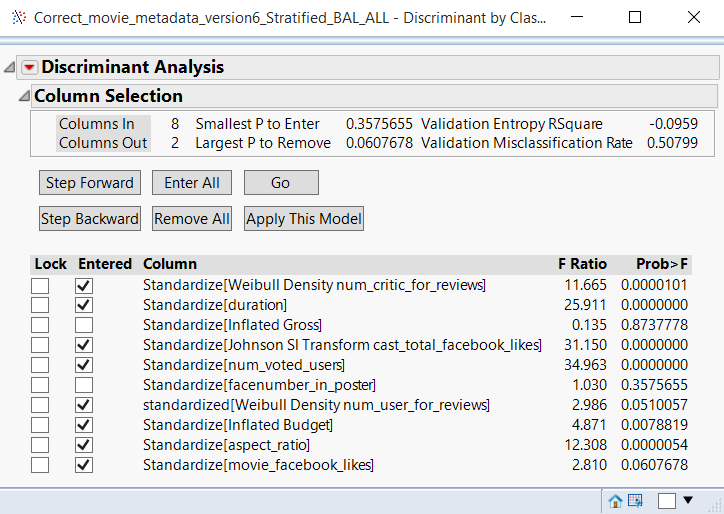


## 17.1 Using Discriminant Analysis for understanding the top variables



The misclassification rate with all predictors included was 0.54309.

On excluding the predictors which had a smaller F Ratio (Gross and Budget) the classification rate reduced by a negligible margin.



Hence we decided to make use of all the predictor variables.

The confusion matrix for the model with the excluded predictors is as below. It can be seen that there is not much difference in the accuracy / error rate of the model due to which these predictor variables can still be included in the data

