DESAUTELS | McGill

MGSC 661 Multivariate Statistics Analysis

THE IMDB PREDICTION CHALLENGE

Abinav Ramesh Sundararaman	(260872704)
Amanda Zeng	(260531550)
Wei Zeng	(260683975)
Junkang Zhang	(260867086)
Chris Zong	(260872947)

November 6, 2018

1 Introduction

IMDb is an online database which stores information about movies and TV shows, and offers a rating scale, 1 to 10, that allows users to rate films. Its rating is one of the most trusted metrics used worldwide to determine the quality and fame of a movie. This website stores tremendous amount of information on a movie, such as duration minutes, director name, actor names, genres, color, content rating, etc. In our project, data are extracted not only from IMDb but also from some external sources, such as Facebook, to make movies rating predictions.

1.1 Summary of the project

The main aim of this project is to predict IMDb ratings of 12 movies that are about to be released in the month of November 2018. To achieve this objective, this project analyzes 3132 movies' information of which are released between 1916 to 2018. In this project, we want to determine the most statistically important variables and build a model to predict IMDb scores. There are four main phases in the project: exploring distribution and relationships between variables, testing non-linearity, building a multiple regression model, and refining the model.

In the first phase, by looking at a scatter plot between Y and each X_i, we examine the distribution of the independent and dependent variables, especially that of quantitative independent variables. For some crucial categorical variables, boxplots are generated to show the dispersion in the data. Then, some useless variables are removed, and this process allows us to size down the data to focus and model easier. Moreover, "genres" column is split to add some missing dummy variables.

In the second phase, relationships among all variables are explored by looking at the correlation matrix and evaluating their collinearity. Linear regression for each predictor with the dependent variable, IMDb score, is performed to evaluate each model's r-squared. Non-constant variance and heteroskedasticity test are performed to flag potential heteroskedastic predictors. We also correct those linear regression models if there is a presence of heteroskedasticity. Based on the results of correlation matrix and linear regression, we keep the predictors with high r-squared and remove their highly correlated predictors with lower r-squared.

In the third phase, the nonlinearity test (Tukey test) is used to see which variables violate the linearity assumption. We isolate these variables and run several polynomial functions to identify the optimal degree. We used the K-fold test and ANOVA test to select the best degree.

In the last phase, we test our model with out of the sample and try to select the best model with the smallest MSE.

2 Data Description

Distribution of the dependent variable: the distribution of IMDb ratings is left-skewed, and more than 90% movies have IMDb score between 5 and 8. (Appendix 1)

Distribution of the independent variables: independent variables such as title, movie id, movie_IMDB_link, plot_keywords, plot_summary are removed because they do not show any predictive power. Several categorical predictors, such as a distributor, director, actor_1_name, actor_1_known_for, actor_2_name, and actor_1_known_for, are also removed because they are only composed of name and text, and they are not useful to build predictive models. In fact, actors' name and directors' names do not repeat often enough for us to create dummy variable. Moreover, budget and movie budget are removed because some movies do not provide budget information even after movies are released. Plus, budget's currency is different because movies may be produced by different countries. Therefore, it is hard to use these variables as good predictors.
Based on scatter plot, histogram and boxplot, the following distribution is observed for each independent variable: (Appendix 2-74)

Release_month: release month is uniformly distributed among all months. January and October have slightly more released movies.

Release_day: release day is uniformly distributed among all days in the month.

Release_year: most of the movies are released between 2000 and 2016, and very few movies are released before 1980. In fact, this variable's distribution is left-skewed.

Duration_mins: this variable is highly concentrated with duration length around 100 minutes, and the score ranges from 2 to 8.5. But most of movies concentrated from with score of 4 to 8. The distribution of this variable is right-skewed.

Language: about 97% of movies have English as movie language. In this case, it makes no sense to consider language as a strong predictor since most of the movies are in English.

Country: the top ten countries are selected based on the number of movies in the dataset. They are Australia, Canada, France, Germany, India, Italy, New Zealand, Spain, UK, and the USA, and we group other countries as "Others". According to boxplot, USA, UK, and other countries seem to be normally distributed. The movies with the lowest score are from the USA. Moreover, few outliers are present in the USA movies distribution. New Zealand is the country has the highest overall movie score, which concentrates from 7.2 to 8.

Content_rating: there are in total eight different categories. According to boxplot, "approved" movies have an overall very high movie score. For "R", "PG", and "PG-13" movies, scores seem normally distributed. For "TV-14" and "Not rated" productions, scores are widely distributed, meaning that for the same type of rating, the film's score varies immensely.

Number_news_articles: most of the data are concentrated between 0 and 5000. Based on the dis-

tribution, for those that have more number of news articles, the IMDb score tends to be higher than average than those do not. The distribution of this variable is right-skewed.

Director_facebook_likes: most of the data are concentrated between 0 and 1000 likes with score varying from 5 to 8. Some directors are extremely well-known with Facebook likes greater than 10,000. The score of their movies is between 6 to 9. The distribution of this variable is right-skewed.

Actor_1_facebook_likes: there are some actors with an extremely high number of Facebook likes. About half number of actors have Facebook likes between 0 and 1000, and the other half is between 2000 to 50000 likes. Most of the data are concentrated with a score of 5 to 8. The distribution of this variable is right-skewed.

Actor_1_star_meter: about 90% of the movies have an actor 1-star meter smaller than 5000. It seems that as actor 1-star meter increase, movie score tends to go down. The distribution of this variable is right-skewed.

Actor_2_facebook_likes: about 85% movies have actor 2 Facebook like smaller than 1,000. About 300 movies have actor 2 Facebook likes range from 6,000 to 27,000. There is one actor 2 has extremely high Facebook likes, which is 137,000. In fact, the distribution of this variable is right-skewed.

Actor_2_star_meter: most the movies have actor 2 meters smaller than 10,000. Most of movies' actor 2 meters are concentrated with meter smaller than 5,000. Their movies score are most likely to be concentrated between 6 to 7. The distribution of this variable is right-skewed.

Actor_3_facebook_likes: most the movies have actor 3 Facebook likes smaller than 2,000. Movies scores are well distributed, and a large number their movies have a rating around 6.5. The distribution of this variable is right-skewed.

Actor_3_star_meter: a large number of movies have actor 3 meter smaller than 10,000. Most of movies' actor 3 meter are concentrated with meter smaller than 6,000. Their movies score are most likely to be concentrated near 6.5. The distribution of this variable is right-skewed.

Color: according to boxplot, black and white movies tend to have much higher IMDb score than that of the color movies. The median score of black and white movies is 7.5 and that of color is about 6.8. The range of black and white movies' score is from 5.5 to 9. On the other hand, the range of color movies' score varies from 4 to 9.

critic_reviews_number: number of critic review is mostly concentrated between 0 and 200 with a movie's score varying from 6 to 7.5. The distribution of this variable is right-skewed.

Genres: this variable will be split in a larger phase to get more genres' dummy variables.

User_votes_number: as the number of user votes increases, the movies score increases. However, most of the data are concentrated between 10,000 and 200,000. The distribution of this variable is right-skewed.

Cast_total_facebook_likes: about 97% of movies have cast total Facebook likes varying from 0 to 50,000 likes. $\frac{2}{3}$ of these movies have likes between 0 and 10,000. The distribution of this variable is

right-skewed.

Number_of_faces_in_movie_poster: about 93% of movies have the number of faces in the poster fewer than 4. There is a movie have 31 faces in the poster, which shifts its distribution. In fact, the distribution of this variable is right-skewed.

User_reviews_number: based on the scatter plot (Appendix 46-47), it seems that as the number of user reviews increases, the movies score increases. But most of the data are concentrated between 10,000 and 200,000 user reviews. The distribution of this variable is right-skewed.

Movie_facebook_likes: about 70% of movies have movie Facebook likes fewer than 1,000. The score of these movies seems to be well distributed between 4 and 7.5. The distribution of this variable is right-skewed.

Sum_total_likes: about 97% of movies are concentrated with sum total likes between 0 and 50,000. $\frac{2}{3}$ these movies are between 0 to 8,000. Movie score is mostly concentrated between 5.5 and 7.5. As the number of total likes increase, movie score would increase. The distribution of this variable is right-skewed.

Ratio_movie_cast_likes: most of the data are concentrated near 0 and 5. The distribution of this variable is right-skewed.

movie_meter_IMDB_pro: about half number of movies are concentrated around a value of 5,000. As movie meter decreases, movie score tends to increase. The distribution of this variable is right-skewed.

Number_of_votes: number of votes are concentrated around 10,000. As the number of votes increases, movie score tends to increase. The distribution of this variable is right-skewed.

Dummy variables: boxplots for dummy variables are also created to detect if movie score can vary if a specific genre is present. We evaluate variables such as action, adventure, sci-fi, thriller, musical, romance, western, sport, horror, drama, war, animation, and crime. Based on the boxplots, action, sci-fi, and thriller movies have a slightly negative effect on IMDb score comparing to other types of movies while holding other variables constant. Horror movies is 1 score lower than non-horror movies. Adventure, musical, romance, or crime movies may not have a strong impact on the IMDb score. Western, sport, drama, war, or animation movies have a positive effect on the IMDb score.

3 Model Selection

To build the desired model, we first clean the entire dataset, correct the wrong data entries, and remove all the movies with N/A values because of the error thrown for the regression models. We also remove some unnecessary categorical predictors from the dataset, such as plot summary and IMDb website, so that we can focus on the more useful ones. The primary three categorical predictors we choose are country, content rating, and color because we believe that all three categories have different target groups with different sizes

and different group of people may rate movies differently. We use ten most frequently appeared value in these predictors to create dummy variables and categorize everything else as "others" for each of these categorical variables if it is the case. We then create a complete list of "genres" dummy variables to see each one's predictive power.

Since our primary concern of the model is multicollinearity, we run a correlation test to remove predictors that have either a positive or negative correlation greater than |0.4|, and the dataset ends up with 35 predictors, which includes 21 genre predictors. To narrow down the number of predictors, we perform a simple linear regression for each predictor against IMDB_score to evaluate which predictors have the highest r-squared. From the scatter plots (Appendix 22-57), we also notice that there are many predictors with extremely high value on the X-axis. Therefore, we decide to remove them from the dataset to test whether this step can improve the r-squared. Since heteroskedasticity can also influence the results generated by linear regression, residual plots are performed to detect a such issue. After removing all the extreme values by analyzing cook's Distance and performing a linear regression to correct heteroskedastic errors, we have obtained the predictors that have significant predictive power for the model building.

Other than linear regression, polynomial regression model is also very worthwhile to test. We perform an ANOVA test to see which degree is the most optimal for each predictor and K-fold test to examine which degree has the minimum MSE. Although for some predictors, the test results indicate that the predictor should use degree 4 or 7, we only choose degree 2 for all the significant polynomial predictors. The reason for this procedure is that degree 2 is very significant based on the ANOVA tests' result and is robust when extreme values are present. The model is well built by using all the predictors that have a relatively high r-squared value and a correlation coefficient below |0.4|. During the predictors selection process, we cannot use some "good" predictors, such as critic review number, user votes, etc., due to missing information on the new 12 movies, and therefore we decide to remove them.

During the training process, we realize that all the predictors involved with Facebook are very sensitive to the data collection time. Hence, Facebook likes predictors are extremely volatile and their coefficients may be impossible to predict accurately. As a result, we decided to use actor meter and movie meter's value in the model because they are more stable and robust.

After deciding on the predictors, we also test the interaction with movie meter with every genres variable. However, none of the interactions seems significant and the adjusted R-squared decreases, meaning that the extra variables in the model are unnecessary.

After all the above procedures, the model for predicting movie's IMDb socre is finalized. The equation for the regression model is as followed.

$$IMDB\_score = \beta_0 + \beta_1 \times duration\_mins + \beta_2 \times duration\_mins^2 + \beta_3 \times movie\_meter\_IMDB\_pro$$
$$+\beta_4 \times number\_news\_articles + \beta_5 \times number\_news\_articles^2 + \beta_6 \times actor\_1\_star\_meter + \beta_7 \times actor\_2\_star\_meter$$
$$+\beta_8 \times number\_of\_faces\_in\_movie\_poster + \beta_9 \times Documentary + \beta_{10} \times Biography + \beta_{11} \times Drama + \beta_{12} \times Crime$$
$$+ \beta_{13} \times Romance + \beta_{14} \times Thriller + \beta_{15} \times Music + \beta_{16} \times Fantasy + \beta_{17} \times Action + \beta_{18} \times Family$$
$$+ \beta_{19} \times Comedy + \beta_{20} \times Horror + \beta_{21} \times Scifi$$

4 Results

To assess the predictive power of the model, we decide to use a "modified" version of the validation test set; size of test data is $\frac{1}{6}$of the entire dataset. We believe that the validation test set is very easy to use. In this case, we can lessen the method's drawback by increasing the size of the training dataset. By running the model more than 20 times and recording the MSE values, we conclude that our model's MSE values generally varies from 0.6 to 0.8.

Table 1 is the summary of the regression model. Although not all predictors are significant, the model may generate inferior prediction without those predictors for out of sample tests.

Table 1: Regression Model Summary

	Dependent variable:
	imdb_score
$duration\_mins$	14.981^{***} (1.050)
$duration\_mins^2$	-2.820^{***} (0.920)
$movie\_meter\_IMDB\_pro$	-0.00000^{***} (0.00000)
$number\_news\_articles$	12.867^{***} (0.951)
$number\_news\_articles^2$	-7.677^{***} (0.890)
$actor\_1\_star\_meter$	0.00000 (0.00000)
$actor\_2\_star\_meter$	0.00000 (0.00000)
$number\_of\_faces\_in\_movie\_poster$	-0.027^{***} (0.009)
$Documentary$	1.285^{***} (0.159)
$Biography$	0.119^{*} (0.072)
$Drama$	0.367^{***} (0.041)

Crime	0.147*** (0.047)
Romance	−0.131*** (0.041)
Thriller	−0.146*** (0.045)
Music	−0.222*** (0.083)
Fantasy	−0.013 (0.054)
Action	−0.359*** (0.045)
Family	−0.251*** (0.065)
Comedy	−0.111** (0.043)
Horror	−0.331*** (0.062)
Scifi	−0.033 (0.052)
Animation	0.799*** (0.103)
Sport	0.019 (0.083)
War	0.189** (0.089)
Western	0.161 (0.125)
History	−0.017 (0.093)
Musical	0.225* (0.129)
Mystery	0.047 (0.059)
Constant	6.573*** (0.053)
Observations	3,010
R^2	0.328
Adjusted R^2	0.322
Residual Std. Error	0.877 (df = 2981)
F Statistic	51.973*** (df = 28; 2981)
Note:	*p<0.1; **p<0.05; ***p<0.01

Interpretation of numerical data

The duration minutes has a convex and downward sloping relationship with the IMDb score. When the duration minutes of a movie increase, it will have a positive impact on the IMDb scores. However, when the movie is too long, the score may be potentially lowered. This result matches our intuition since a short movie cannot convey its idea clearly while a long movie may be tiresome for the viewers.

The movie meter parameter has a slightly negative relationship with the IMDb score. ~~IMDb measures the popularity of the movies in an ascending order. Therefore~~, the more popular a movie is, the lower its movie meter is. This relationship explains the negativity of the coefficient. Because there are millions of movies on

This sentence is not true..movie meter just how much people talk about a movie.. its not really a measure of bad movies.I will send a link regarding movie meter

This parameter plays an important role when a movie is heavily talked about. Its a very good indicator of hype for a movie

the website, this parameter may play a role when the movie is not popular.

The number of news articles has a similar relationship with the IMDb score as duration minutes. The more news article a movie has, the higher IMDb it will have, and diminishing return appears after a certain point, meaning the score declines.

Actor 1-star meter and actor 2-star meter both have a slightly positive relationship with the IMDb score. The more famous actor stars are, the lower actor meter they will get. We assume that some good movies with high ratings usually are performed by non-famous actors who have good acting skills.

Actor star meter is a good indicator of the perception of the actor among the masses.

Interpretation of categorical data

All the categorical data in the models are associated with movie genres. In general, if a movie belongs to the following genres, documentary, biography, drama, crime, animation, sport, war, western, musical, and mystery, the movie will receive extra scores given all the other factors constant. Contradictorily, romance, thriller, music, fantasy, action, family, comedy, horror, sci-fi, and history movies will have lower scores with all other factors being constant.

Given the number of variables in the model, the overfitting issue may also emerge. We decide to collect five out of sample movies' data to test if the MSE falls in the same range. The results have shown that our model generate more or less the same results for those five movies. The model is thus valid for predicting a general IMDb score. However, we discover the model generally produces a conservative prediction, meaning that the predictions is often lower than the actual result. We find that one of the limitations of our model is that we are not taking into consideration the variations of predictors over time. In other words, we assume that the distribution of our predictors remain constant over time. For instance, users may read other people's reviews and scores, and get influenced by other users' comments. Also, actor meters are highly dependent on actor's popularity. The popularity can fluctuate greatly. Therefore, the prediction may not be 100% accurate.

Please remove this. it underappreciates our model too much.

Given our knowledge, spline line test and time series regression may improve the prediction power of a model. However, by studying scatter plots, we do not find any visual clue of the spline line. Moreover, polynomial has already created acceptable results, and we do not have to run spline to further minimize the MSE. The reason that we have not performed a time series regression is that we believe all users can rate the movies at any time. The time of the data collection may influence the values we have in the training model. Therefore, finding patterns without knowing the data collection time may not further improve the model. To predict the new 12 movies' IMDb score, all the information are mainly collected from the IMDb website and input into the model. However, there are still some missing value for the new movies. The movie, Second Act, does not provide any information about the duration minutes publicly, and the variable is a critical predictor for the model. Therefore, we use the median duration minutes from the train dataset to be the duration minutes for that movie, which is 100 minutes. We believe median is a more rigorous value than average for

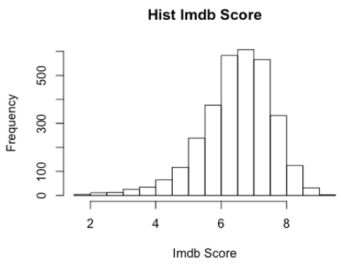
We dont really have to mention this.. its unnecessary. Please remove this. We dont have to be utterly truthful :)

missing values because average is easily influenceable by extreme values. The predicted score are in table 2.

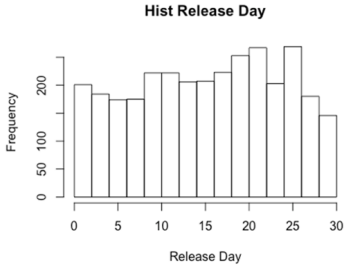
Table 2: Prediction of the 12 new movies' score

Movie	Score Prediction
The Grinch	6.50
Postcards from London	6.27
The Long Dumb Road	5.93
Crimes of Grindelwald	6.31
Instant Family	6.26
The Cloverhitch Killer	6.34
Robin Hood	6.03
Creed II	6.84
The Women of Marwen	6.54
Ralph Breaks the Internet	7.13
Second Act	6.01
Becoming Astrid	7.00

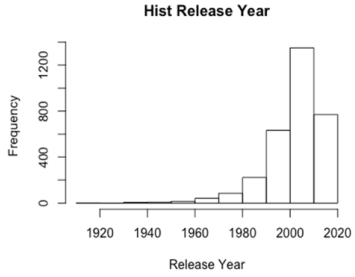
A Appendix



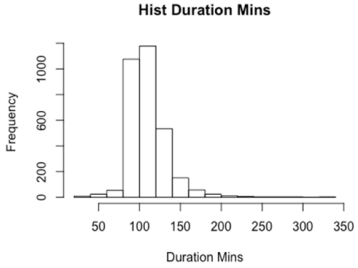
(1) Histogram



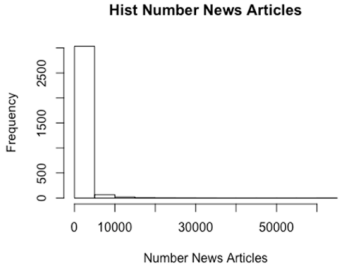
(2) Histogram



(3) Histogram

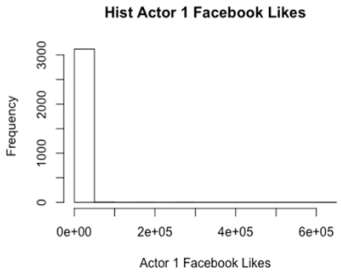


(4) Histogram



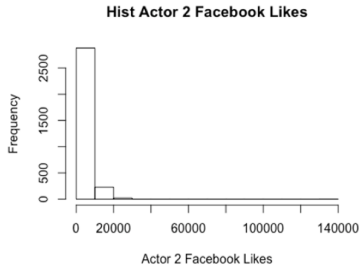
(5) Histogram

(6) Histogram

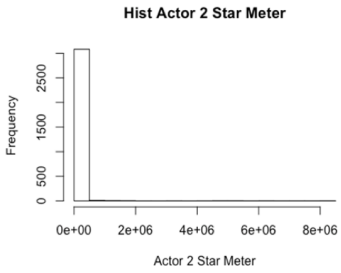


(7) Histogram

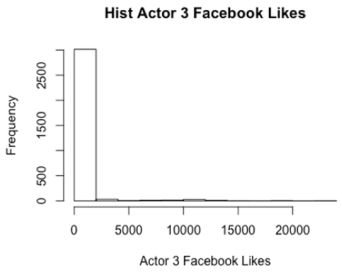
(8) Histogram



(9) Histogram

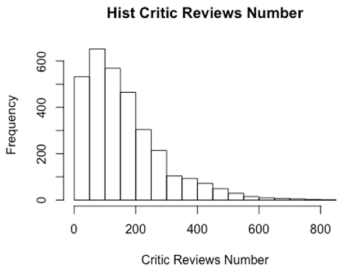


(10) Histogram

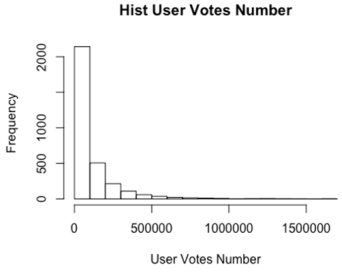


(11) Histogram

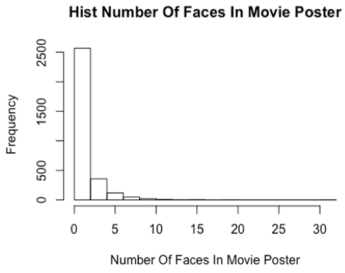
(12) Histogram



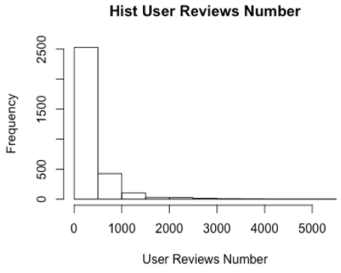
(13) Histogram



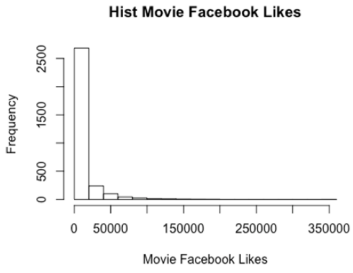
(14) Histogram



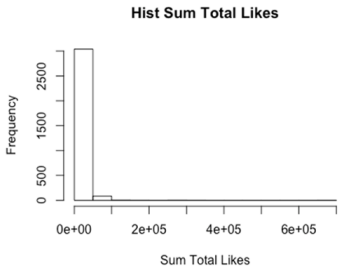
(15) Histogram



(16) Histogram

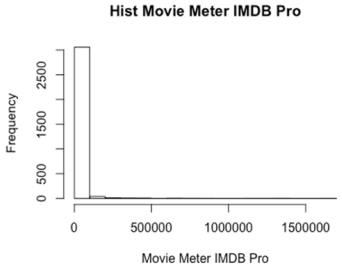


(17) Histogram

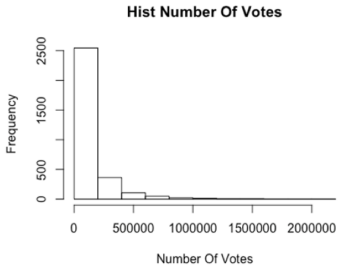


(18) Histogram

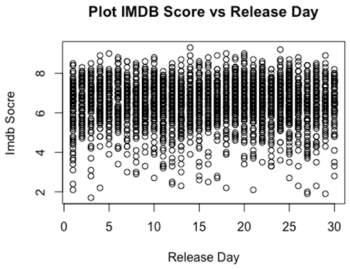
(19) Histogram



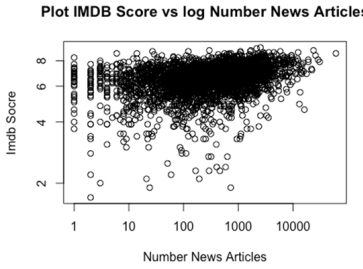
(20) Histogram



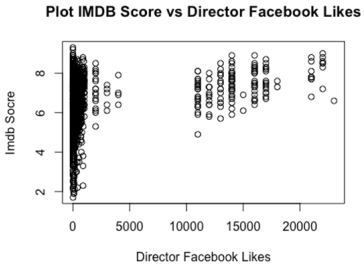
(21) Histogram



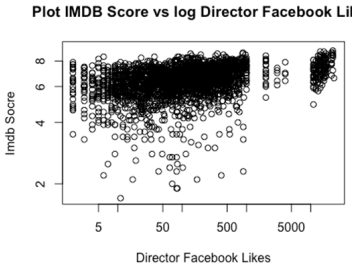
(22) Histogram



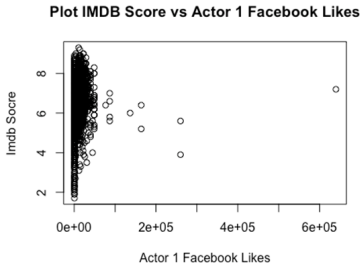
(23) Scatter Plot



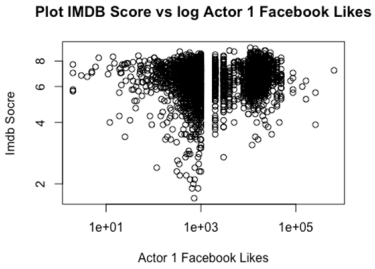
(24) Scatter Plot



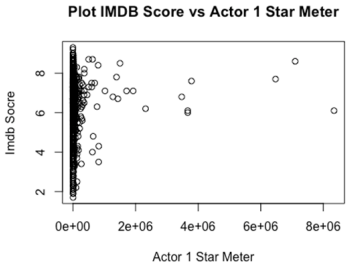
(25) Scatter Plot



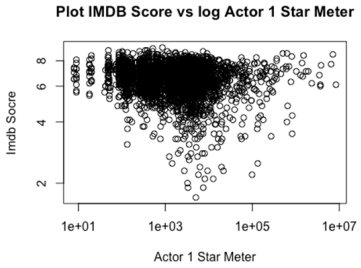
(26) Scatter Plot



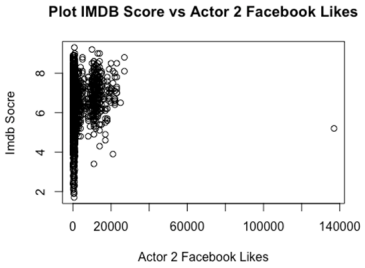
(27) Scatter Plot



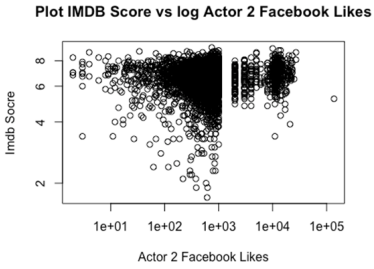
(28) Scatter Plot



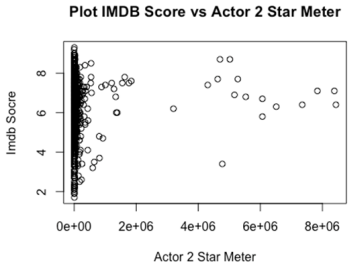
(29) Scatter Plot



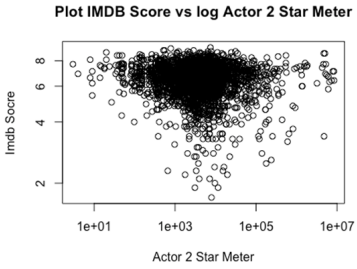
(30) Scatter Plot



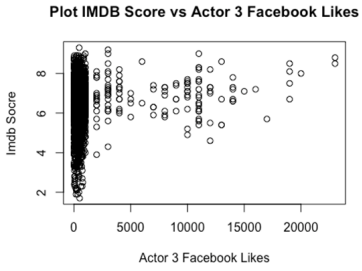
(31) Scatter Plot



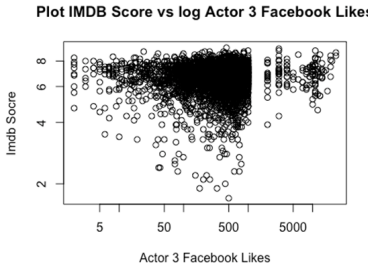
(32) Scatter Plot



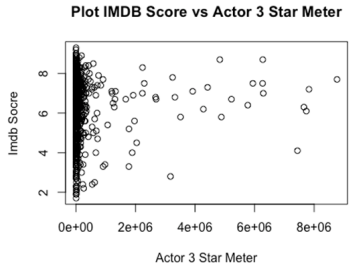
(33) Scatter Plot



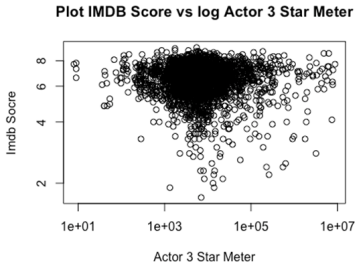
(34) Scatter Plot



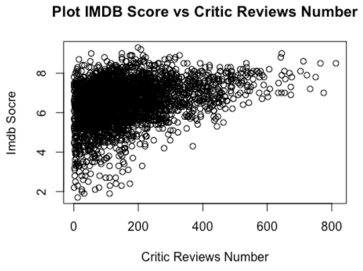
(35) Scatter Plot



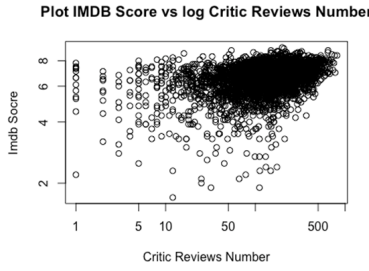
(36) Scatter Plot



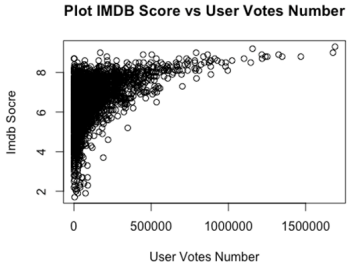
(37) Scatter Plot



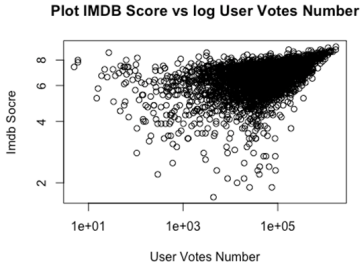
(38) Scatter Plot



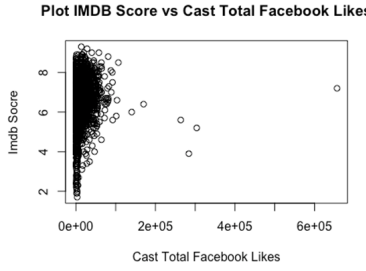
(39) Scatter Plot



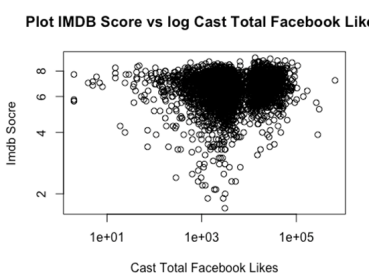
(40) Scatter Plot



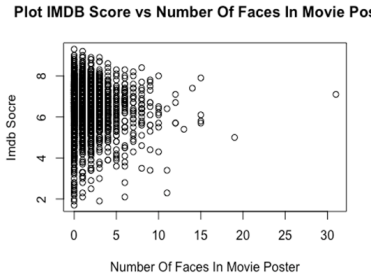
(41) Scatter Plot



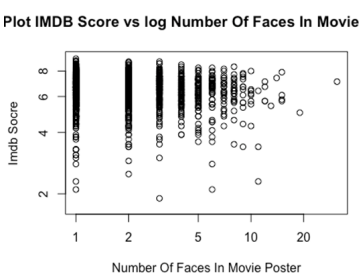
(42) Scatter Plot



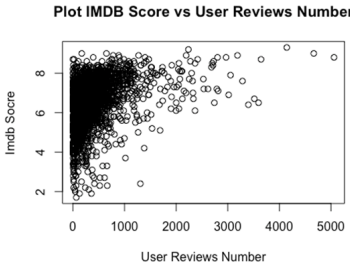
(43) Scatter Plot



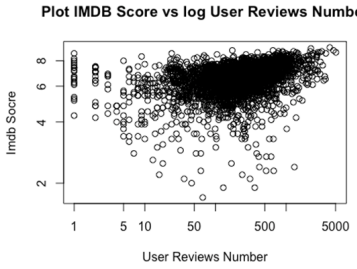
(44) Scatter Plot



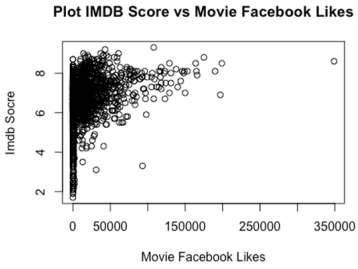
(45) Scatter Plot



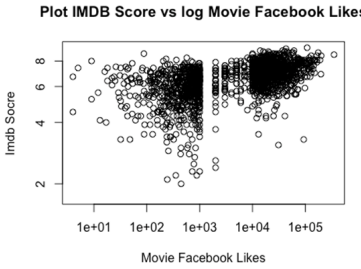
(46) Scatter Plot



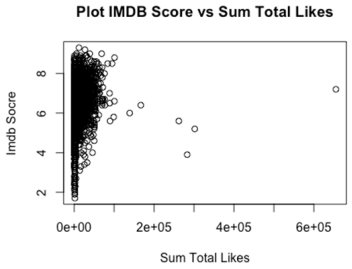
(47) Scatter Plot



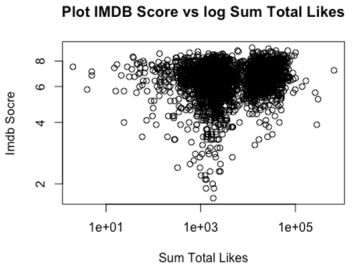
(48) Scatter Plot



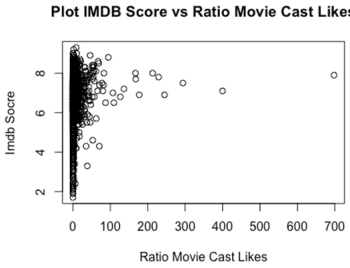
(49) Scatter Plot



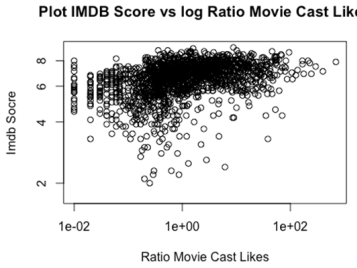
(50) Scatter Plot



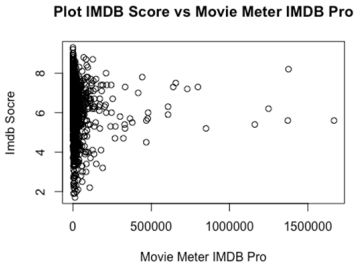
(51) Scatter Plot



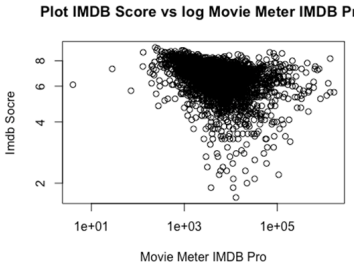
(52) Scatter Plot



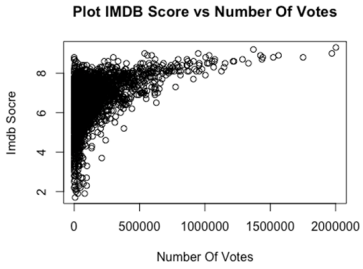
(53) Scatter Plot



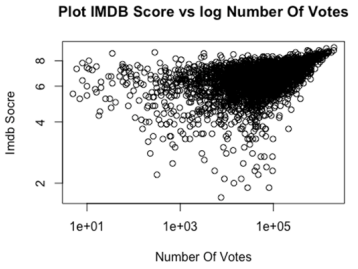
(54) Scatter Plot



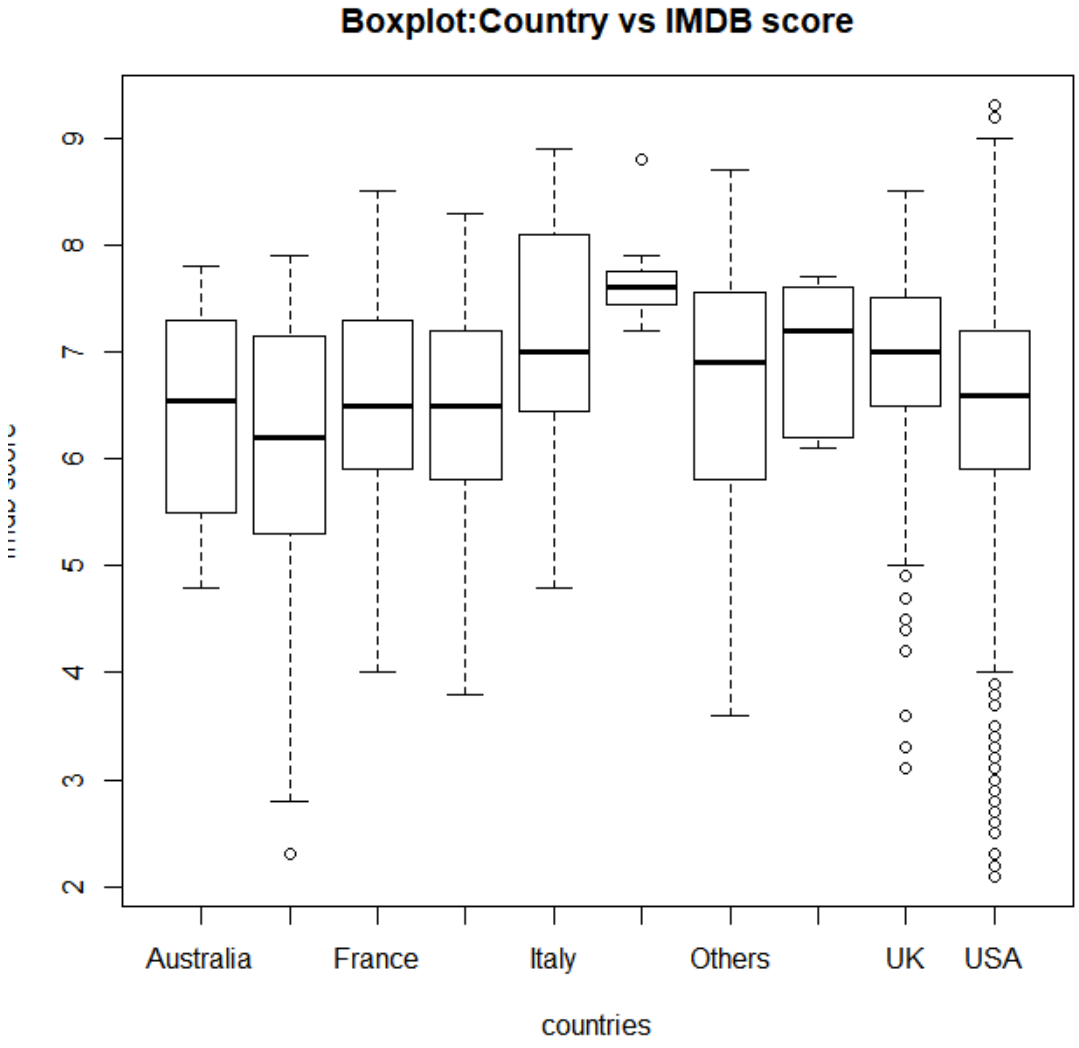
(55) Scatter Plot



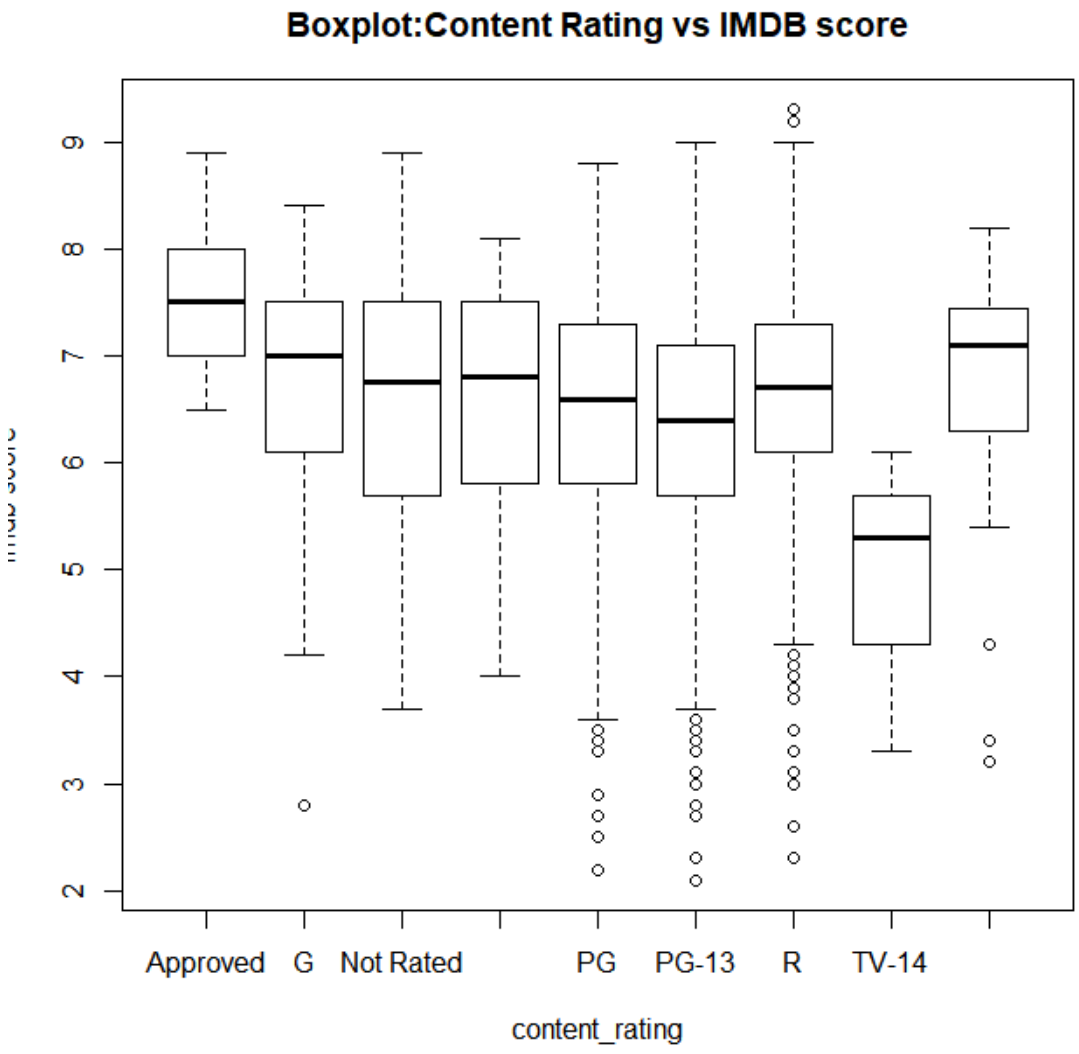
(56) Scatter Plot



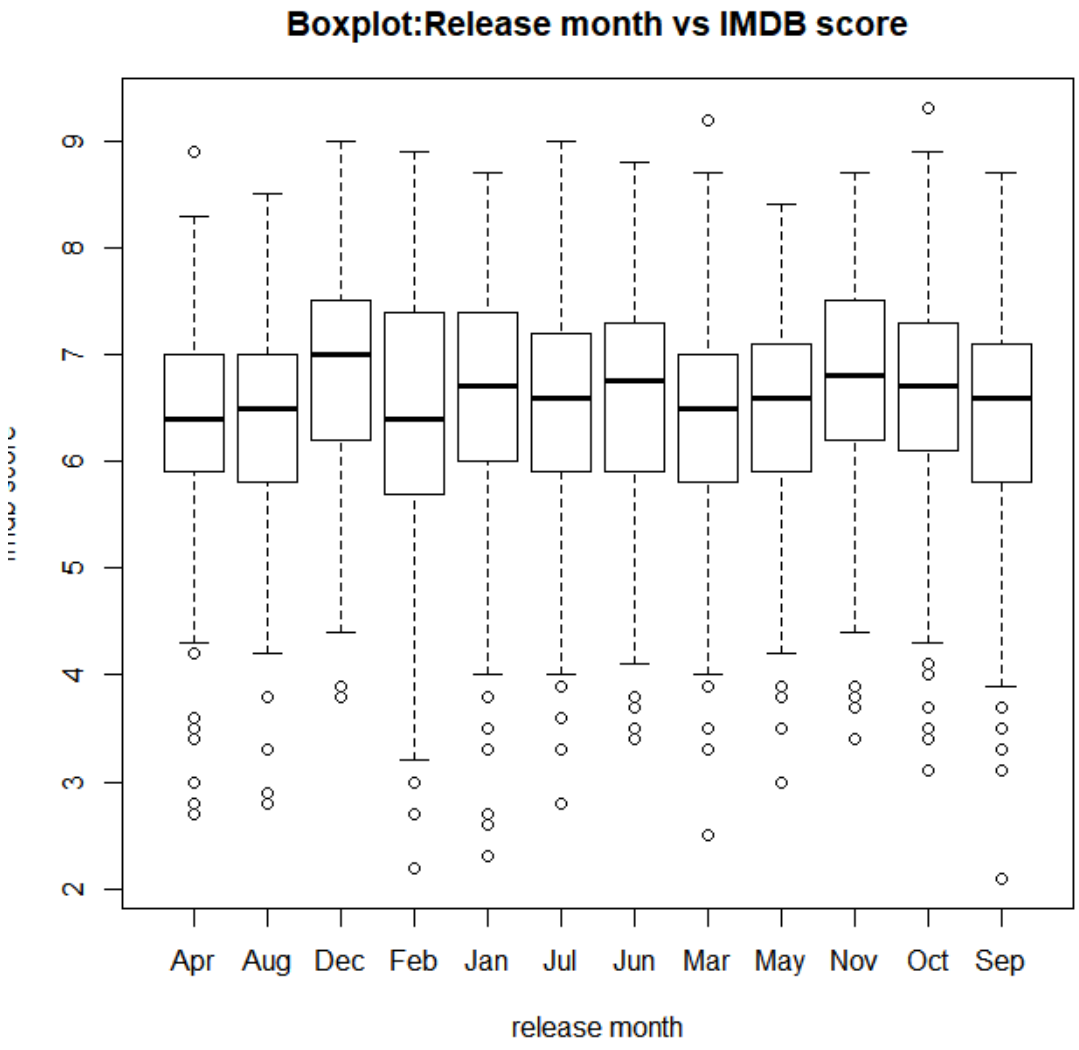
(57) Scatter Plot


Boxplot:Country vs IMDB score
2
3
4
5
6
7
8
9
Australia
France
Italy
Others
UK
USA
countries

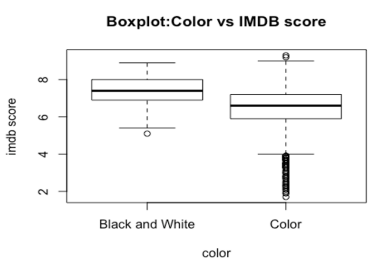
(58) Boxplot


Boxplot:Content Rating vs IMDB score
2
3
4
5
6
7
8
9
Approved
G
Not Rated
PG
PG-13
R
TV-14
content_rating

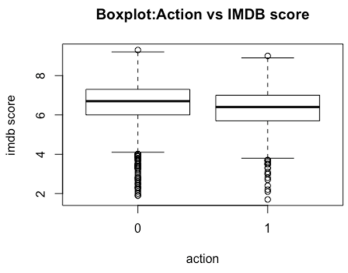
(59) Boxplot



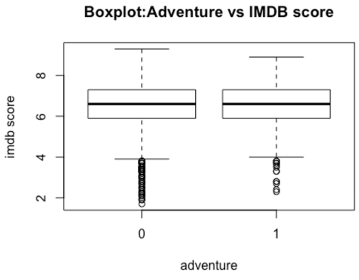
(60) Boxplot



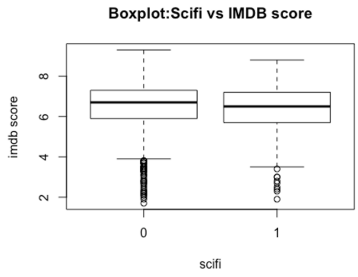
(61) Boxplot



(62) Boxplot

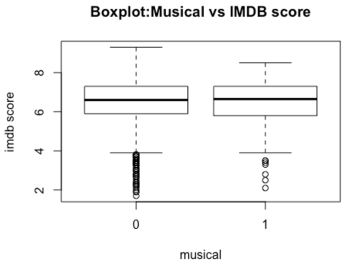


(63) Boxplot

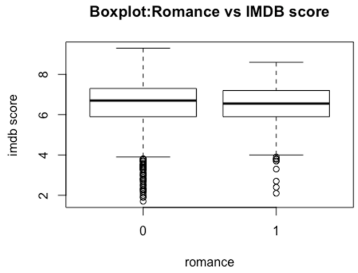


(64) Boxplot

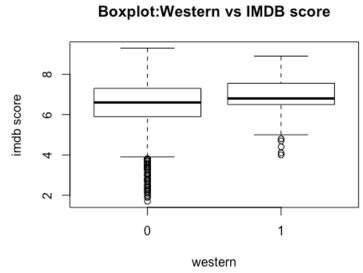
(65) Boxplot



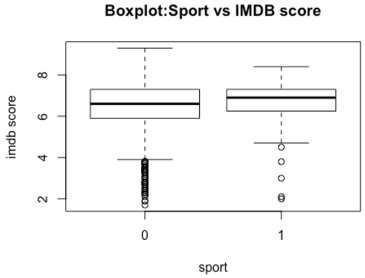
(66) Boxplot



(67) Boxplot

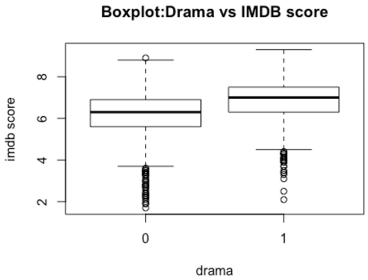


(68) Boxplot

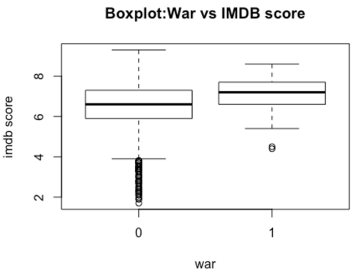


(69) Boxplot

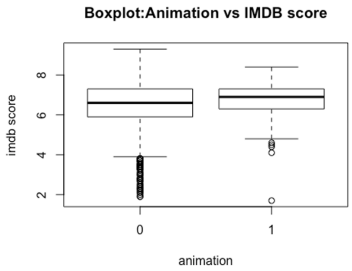
(70) Boxplot



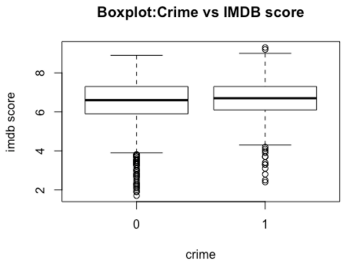
(71) Boxplot



(72) Boxplot



(73) Boxplot



(74) Boxplot