EXCELLENCE SINCE 1817

PRESIDENCY UNIVERSITY

KOLKATA

Data Analysis Project
About all Shows Available on Netflix

Registration No. 18214110028 Roll No. STAT-008

Name:- Suryadeep Ghosh

Department of Statistics

In this project we try to analyze the “Netflix TV Shows and Movies” dataset. This data set was created to list all shows available on Netflix streaming and was acquired in May 2022 containing data available in the United States. **Netflix** is a subscription-based streaming service that allows people to watch TV shows and movies without commercials on an internet-connected device.

This dataset has two files containing the titles (titles.csv) and the cast (credits.csv) for the title. In this analysis we work with the ’titles.csv’ file.This dataset contains 5806 unique titles on Netflix with 15 columns containing their information, including:

1. **id:** The title ID on JustWatch.
2. **title:** The name of the title.
3. **show_type:** TV show or movie.
4. **description:** A brief description.
5. **release_year:** The release year.
6. **age_certification:** The age certification.
7. **runtime:** The length of the episode (SHOW) or movie.
8. **genres:** A list of genres.
9. **production_countries:** A list of countries that produced the title.
10. **seasons:** Number of seasons if it’s a SHOW.
11. **imdb_id:** The title ID on IMDB.
12. **imdb_score:** Score on IMDB.
13. **imdb_votes:** Votes on IMDB.
14. **tmdb_popularity:** Popularity on TMDB.
15. **tmdb_score:** Score on TMDB.

IMDb (an abbreviation of Internet Movie Database) is an online database of information related to films, television series, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews.

TMDB (The Movie Database) is a community built movie and TV database.

In this project we applied various statistical methods and tools to predict the IMDb and TMDB scores of a movie/show. We divided the project in two parts.

In the first part we performed the exploratory analysis where we analyzed the data by drawing various diagrams to depict some dependence of the IMDb Scores and TMDB Scores on different predictors and then made some crucial observations from there. We also observed some interesting insights in this part.

In the second part we make conclusions about the observations we made earlier in the exploratory analysis using statistical methods and tools i.e testing of hypothesis, inference, estimation and various other regression methods.

Importing the Dataset:

First we import the dataset.

We check dimension of the data.

```
[1] 5806   15
```

The names of the coloumns of the dataset are,

```
 [1] "id"                   "title"                "type"
 [4] "description"          "release_year"         "age_certification"
 [7] "runtime"              "genres"               "production_countries"
[10] "seasons"              "imdb_id"              "imdb_score"
[13] "imdb_votes"           "tmdb_popularity"      "tmdb_score"
```

The first few rows of the dataset are,

	id	title	type	description
1	ts300399	Five Came Back: The Reference Films	SHOW	This collection includes 12 World War
2	tm84618	Taxi Driver	MOVIE	A mentally unstable Vietnam War vet
3	tm127384	Monty Python and the Holy Grail	MOVIE	King Arthur, accompanied by his squi
4	tm70993	Life of Brian	MOVIE	Brian Cohen is an average young Jew
5	tm190788	The Exorcist	MOVIE	12-year-old Regan MacNeil begins to
6	ts22164	Monty Python's Flying Circus	SHOW	A British sketch comedy series with t

We see that the **'seasons'** coloumn mostly consists of missing values which will not contribute to our analysis. So we remove that coloumn from the dataset and then remove other rows which has missing values.

Now the dimension of the dataset is,

```
[1] 5041   14
```

We check the variable type of each coloumns of the dataset.

```
'data.frame':	5041 obs. of  14 variables:
 $ id                  : chr  "tm84618" "tm127384" "tm70993" "tm190788" ...
 $ title               : chr  "Taxi Driver" "Monty Python and the Holy Grail" "Life of B
 $ type                : chr  "MOVIE" "MOVIE" "MOVIE" "MOVIE" ...
 $ description         : chr  "A mentally unstable Vietnam War veteran works as a night-
 $ release_year        : int  1976 1975 1979 1973 1969 1971 1964 1980 1967 1966 ...
 $ age_certification   : chr  "R" "PG" "R" "R" ...
 $ runtime             : int  113 91 94 133 30 102 170 104 110 117 ...
 $ genres              : chr  "['crime', 'drama']" "['comedy', 'fantasy']" "['comedy']"
 $ production_countries: chr  "['US']" "['GB']" "['GB']" "['US']" ...
 $ imdb_id             : chr  "tt0075314" "tt0071853" "tt0079470" "tt0070047" ...
 $ imdb_score          : num  8.3 8.2 8 8.1 8.8 7.7 7.8 5.8 7.7 7.3 ...
 $ imdb_votes          : int  795222 530877 392419 391942 72895 153463 94121 69053 11118
 $ tmdb_popularity     : num  27.6 18.2 17.5 95.3 12.9 ...
 $ tmdb_score          : num  8.2 7.8 7.8 7.7 8.3 7.5 7.6 6.2 7.5 7.1 ...
 - attr(*, "na.action")= 'omit' Named int [1:765] 1 33 34 35 92 93 94 97 98 99 ...
  ..- attr(*, "names")= chr [1:765] "1" "33" "34" "35" ...
```

Now we work with this data.

Exploratory Data Analysis:

First we calculate the **summary measures** of the variables in the data.

type:

```
   Length     Class      Mode 
     5041 character character 
```

release_year:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1953    2015    2018    2016    2020    2022 
```

runtime:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   46.00   87.00   79.63  106.00  235.00 
```

production_countires:

```
   Length     Class      Mode 
     5041 character character 
```

imdb_score:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.600   5.800   6.600   6.536   7.400   9.500 
```

imdb_votes:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      5     616    2534   24349   11039 2268288 
```

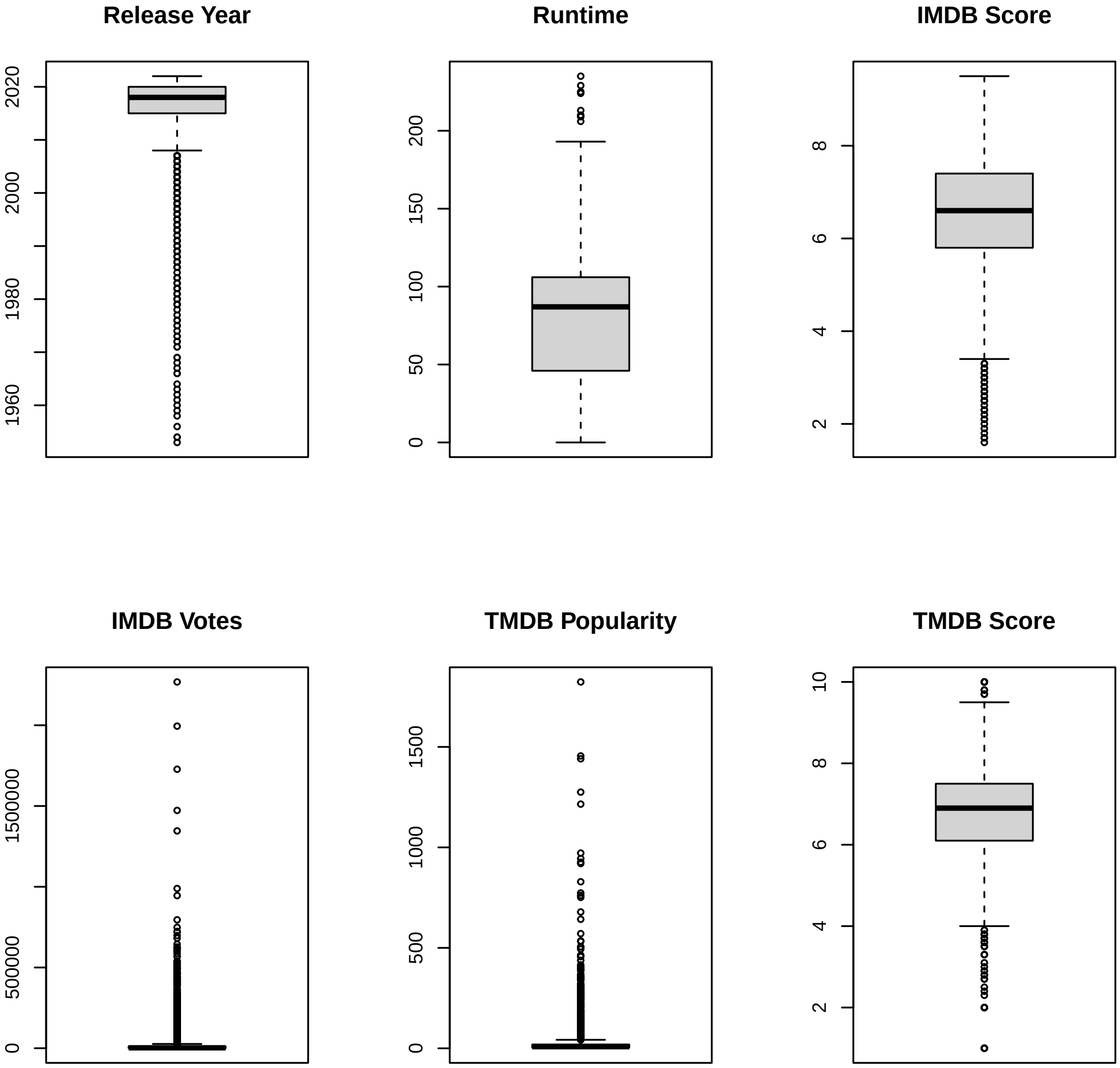
tmdb_popularity:

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
   0.600    3.607    8.311   24.243   19.161 1823.374 
```

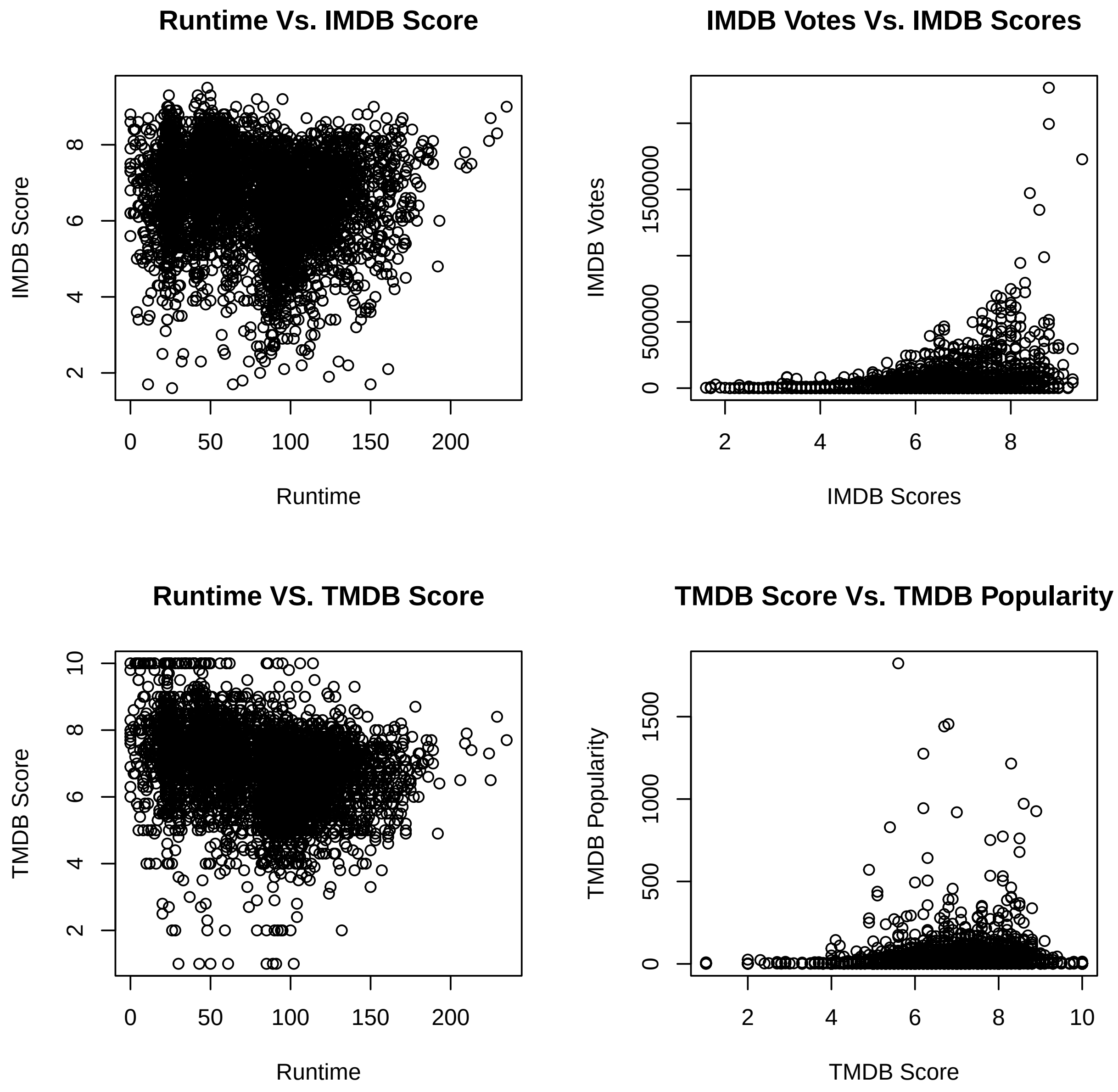
tmdb_score:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   6.100   6.900   6.812   7.500  10.000 
```

Now,
We draw boxplots of the numeric variables of the data.



Then we draw the scatterplot of different combinations of those variables to observe any pattern of relationship among them.



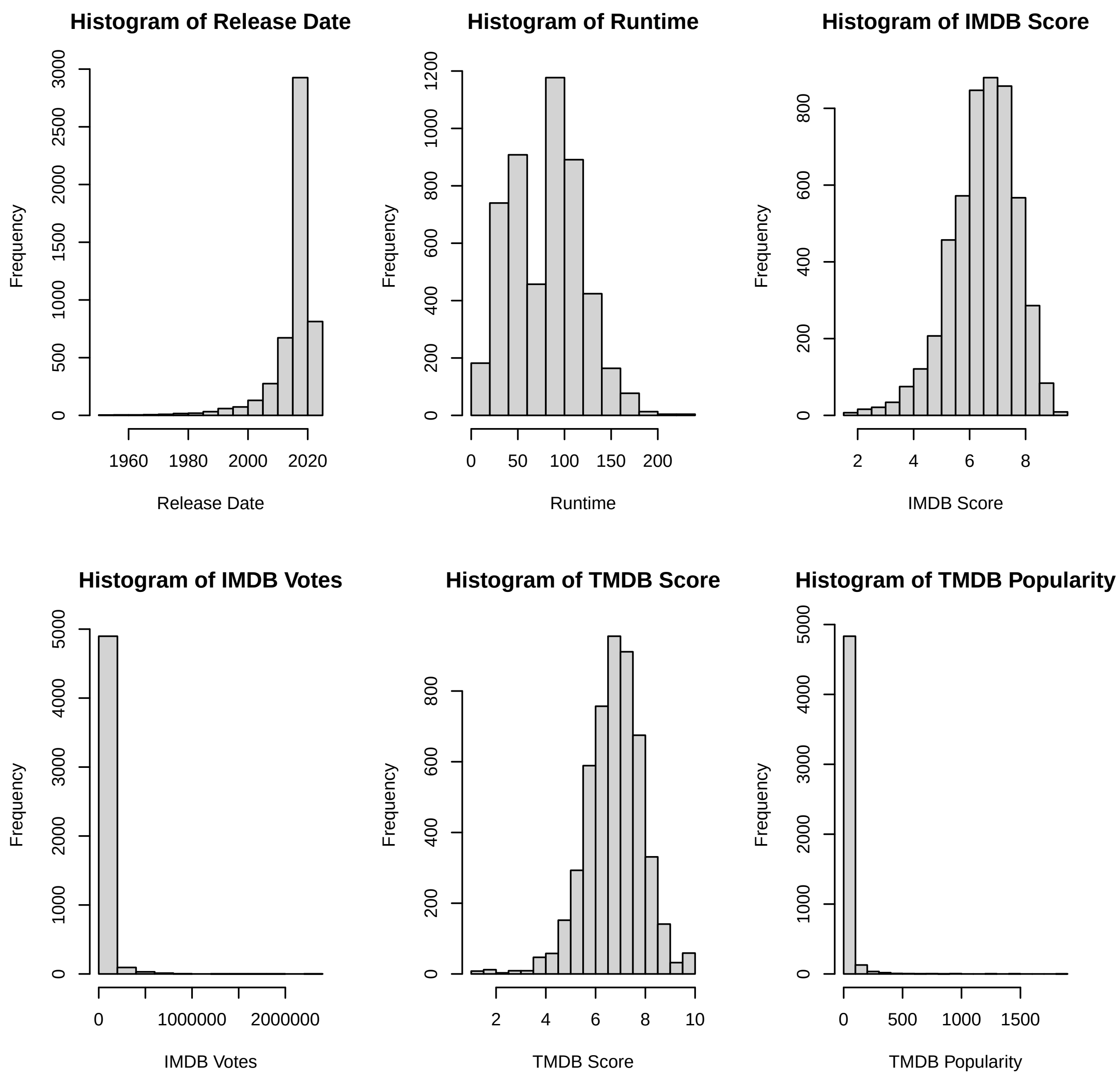
Runtime Vs. IMDB Score

IMDB Votes Vs. IMDB Scores

Runtime VS. TMDB Score

TMDB Score Vs. TMDB Popularity

We also draw the histograms of these variables.



Now we try to see which movies or shows haave the highest and lowest rating on IMDb and TMDB.

IMDb:

Highest Rating:

The highest rating on IMDb is,

```
[1] 9.5
```

The movie/show which has the highest rating on IMDb is,

```
[1] "Breaking Bad"
```

It is a,

```
[1] "SHOW"
```

A short description of the show is,

```
[1] "When Walter White, a New Mexico chemistry teacher, is diagnosed with Stage III canc
```

The release year of this show is,
‘«echo=F,comment=NA»= netflix$release_{y}ear[which.max(netflix$imdb$_{s}core)]$@
The age certification of this show is,

```
[1] "TV-MA"
```

The Runtime of this show in hours is,

```
[1] 48
```

The genre of this show is,

```
[1] "['drama', 'thriller', 'crime']"
```

The production country of this show is,

```
[1] "['US']"
```

The number of votes this show got on IMDb is,

```
[1] 1727694
```

Lowest Rating:

The lowest rating on IMDb is,

```
[1] 1.6
```

The movie/show which has the lowest rating on IMDb is,

```
[1] "He's Expecting"
```

It is a,

```
[1] "SHOW"
```

A short description of the show is,

```
[1] "When a successful ad executive who's got it all figured out becomes pregnant, he's
```

The release year of this show is,

```
[1] 2022
```

The age certification of this show is,

```
[1] "TV-PG"
```

The Runtime of this show in hours is,

```
[1] 26
```

The genre of this show is,

```
[1] "['drama', 'comedy', 'romance']"
```

The production country of this show is,

```
[1] "['JP']"
```

The number of votes this show got on IMDb is,

```
[1] 2735
```

TMDB:

Highest Rating:

The highest rating on TMDB is,

```
[1] 10
```

The movies/shows which have the highest rating on TMDB are,

```
 [1] "Pink Zone"
 [2] "Little Baby Bum"
 [3] "Dharmakshetra"
 [4] "The Haunted House"
 [5] "Rainbow Ruby"
 [6] "Magic Cellphone"
 [7] "Transformers: Rescue Bots Academy"
 [8] "Rainbow Rangers"
 [9] "Unrequited Love"
[10] "The Gift"
[11] "The Unknown Hitman: The Story of El Cholo AdriÃ¡n"
[12] "Three Words to Forever"
[13] "Garth Brooks: The Road I'm On"
[14] "The Possessed"
[15] "Morphle"
[16] "Secreto bien guardado"
```

```
[17] "Singapore Social"
[18] "Felipe Esparza: Bad Decisions"
[19] "Happy Jail"
[20] "The Writer"
[21] "The Queen and the Conqueror"
[22] "How To Ruin Christmas"
[23] "Styling Hollywood"
[24] "Octonauts and the Great Barrier Reef"
[25] "A Queen Is Born"
[26] "No hay tiempo para la vergÃŒenza"
[27] "Legend Quest: Masters of Myth"
[28] "True: Friendship Day"
[29] "Lugar de Mulher"
[30] "Nailed It! Germany"
[31] "Six Windows in the Desert"
[32] "Bangkok Buddies"
[33] "The Charming Stepmom"
[34] "Futmalls.com"
[35] "ç$^{33}Online$”
[36] "Uncle Naji in UAE"
[37] "Fate of Alakada"
[38] "Buddi"
[39] "True Tunes"
[40] "Word Party Songs"
[41] "Mighty Little Bheem: Diwali"
[42] "Selling Tampa"
[43] "Action Pack"
[44] "Marriage or Mortgage"
[45] "The Wedding Coach"
[46] "Thomas & Friends: All Engines Go - Race for the Sodor Cup"
[47] "How to Be A Cowboy"
[48] "The Big Shot Game Show"
[49] "StarBeam: Beaming in the New Year"
[50] "Just in Time"
[51] "Mighty Little Bheem: Kite Festival"
```

We notice that a lot of movies/shows have this highest. So we try to find the most popular one among them.

The movie/show which has the highest popularity among them on TMDB is,

```
[1] "The Haunted House"
```

It is a,

```
[1] "SHOW"
```

A small description of the show is,

```
[1] "With help from a 102-year-old goblin dwelling beneath their haunted apartment build
```

The release year of the show is,

```
[1] 2016
```

The age-certification of the show is,

```
[1] "TV-PG"
```

The runtime of the show in hours is,

```
[1] 25
```

The genre of the show is,

```
[1] "['thriller', 'fantasy', 'horror', 'animation', 'comedy', 'drama', 'family']"
```

The production country of the show is,

```
[1] "['KR']"
```

The TMDB popularity of the show is,

```
[1] 13.649
```

Lowest Rating:

The lowest rating on TMDB is,

```
[1] 1
```

The movie/show which has the lowest rating on TMDB is,

```
[1] "Inborn Pair"
```

It is a,

```
[1] "SHOW"
```

A short description of the show is,

```
[1] "Betrothed while in utero, a resort group president enters into an arranged marriage
```

The release year of this show is,

```
[1] 2011
```

The age certification of this show is,

```
[1] "TV-14"
```

The Runtime of this show in hours is,

```
[1] 43
```

The genre of this show is,

```
[1] "['drama', 'comedy']"
```

The production country of this show is,

```
[1] "['TW']"
```

The TMDB popularity of the show is,

```
[1] 9.281
```

Now,

We want to know the TMDB score of the show ('Breaking Bad') which has the highest IMDb score.

```
[1] 8.8
```

The TMDB score of the show ('He's Expecting') which has the lowest IMDb score is,

```
[1] 4
```

The IMDb score of the show ('The Haunted House') which has the highest TMDB score is,

```
[1] 8.4
```

The IMDb score of the show which has the lowest TMDB score is,

```
[1] 7.1
```

So, we can observe that IMDb scores and TMDB scores are not really agreeing to each other completely. Though they agree on a general pattern that a show that has higher/lower rating on IMDb also has a higher/lower rating on TMDB and vice versa.

Comparison Between IMDb and TMDB scores using Exploratory Analysis:

Comparson between IMDb and TMDB scores of US and Japan based Movies/Shows:

We already know that movies and shows produced in the US are already world famous but in recent years movies and shows produced in Japan are becoming really popular worldwide. So we want to compare the IMDb and TMDB score of US and Japan based movies/shows.

We seperate all the data where production country is only US and only Japan.

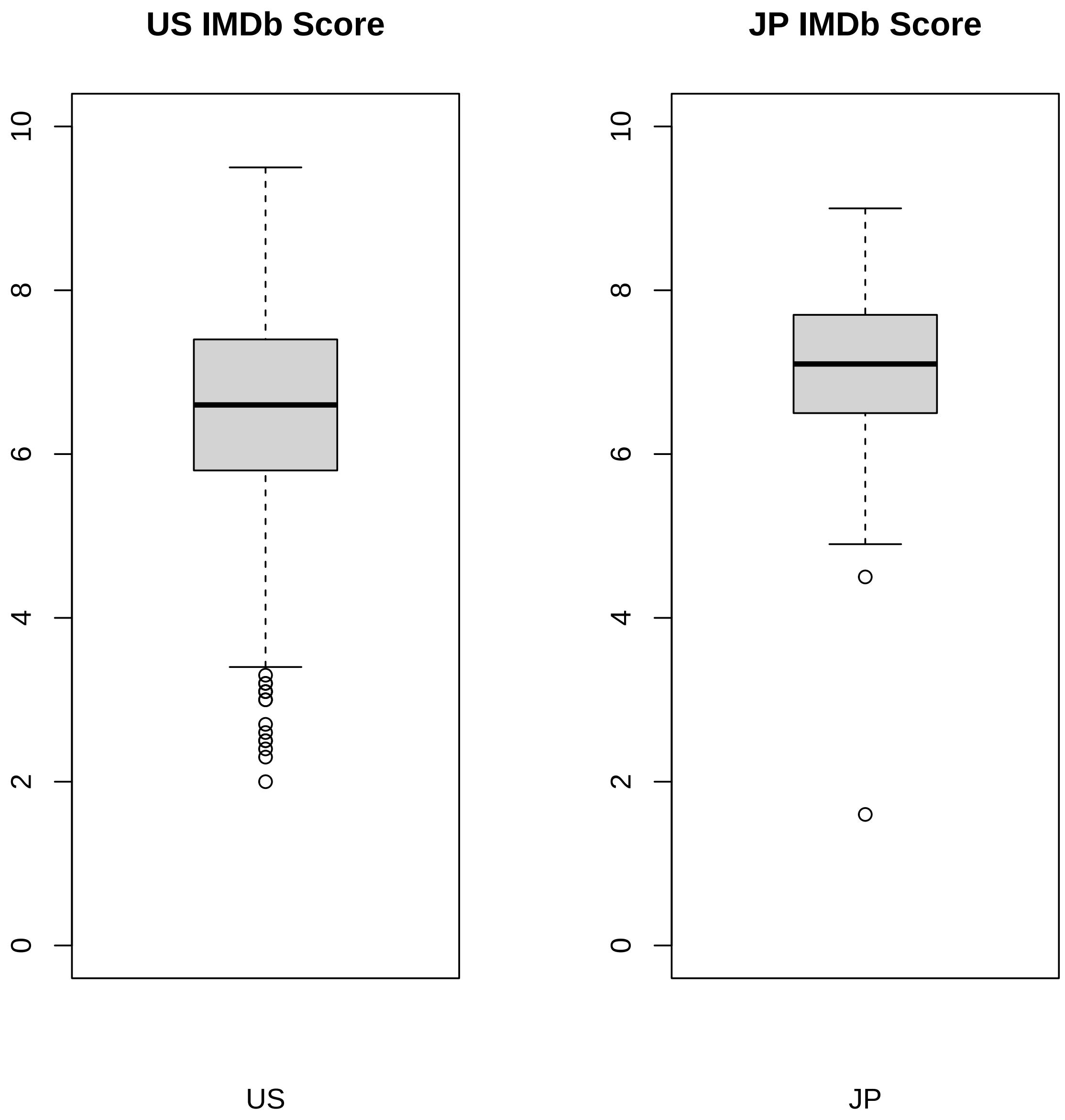
Dimension of data where production country is only US,

```
[1] 1772   14
```

Dimension of data where production country is only Japan,

```
[1] 240  14
```

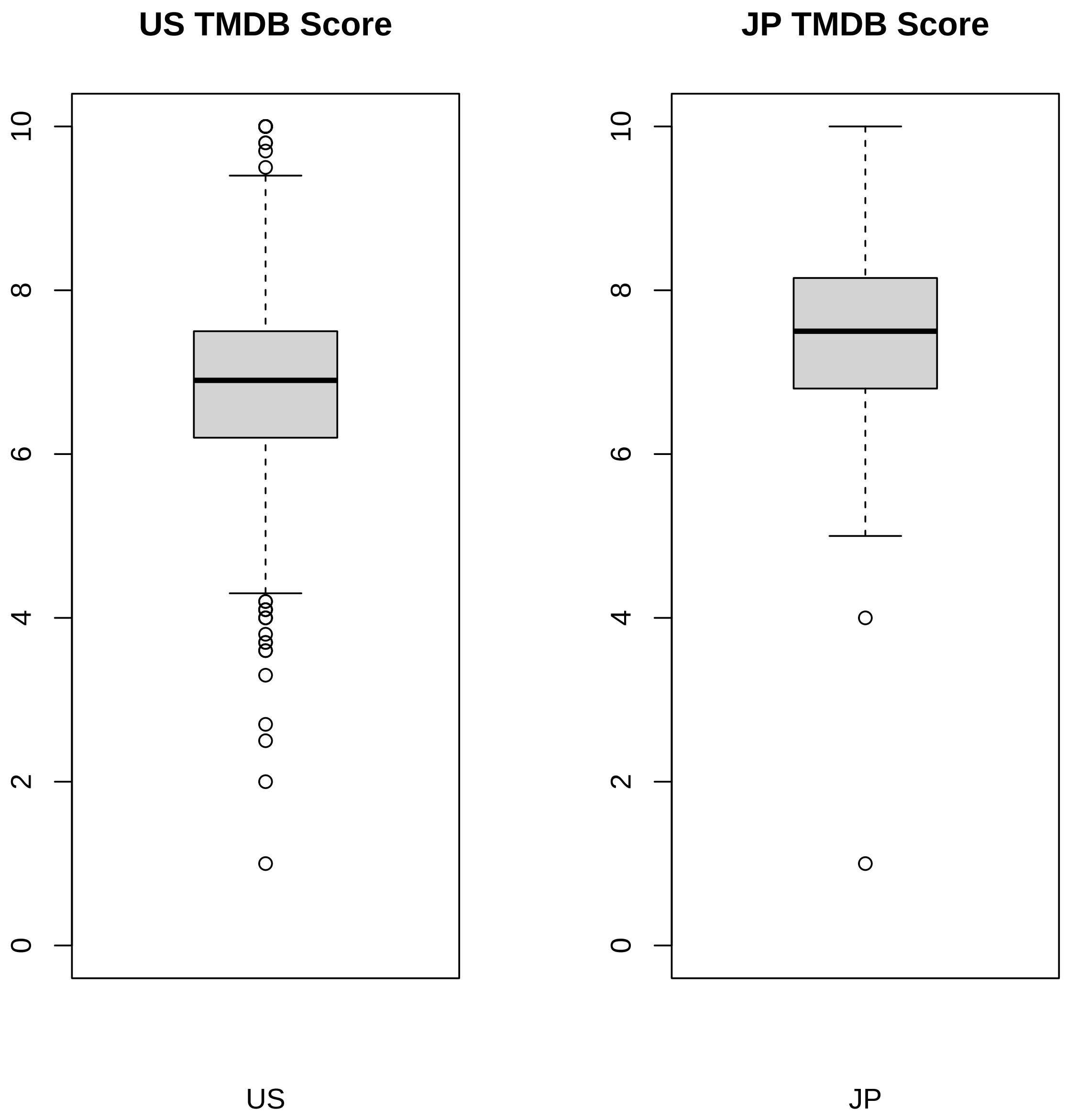
We plot the boxplot of their IMdb Ratings for comparison.



We observe that mean IMDb rating of movies/shows produced in Japan (7.08125) is higher than the mean IMDb rating of movies/shows produced in US (6.552596).

We leave it here for inferential data analysis.

Then we plot the boxplot of their TMDB Ratings for comparison.



We observe that mean TMDB rating of movies/shows produced in Japan (7.433333) is higher than the mean IMDb rating of movies/shows produced in US (6.844582).

We leave it here for inferential data analysis.

Comparison Between IMDb and TMDB Scores of Movies and Shows:

Now we are interested to see that if people enjoy movies or shows more. Our hypothesis based on recent studies is that people like to watch TV-shows more than Movies as TV shows can devote more time to the story than any Movie.

We seperate all the data for movies only and shows only.

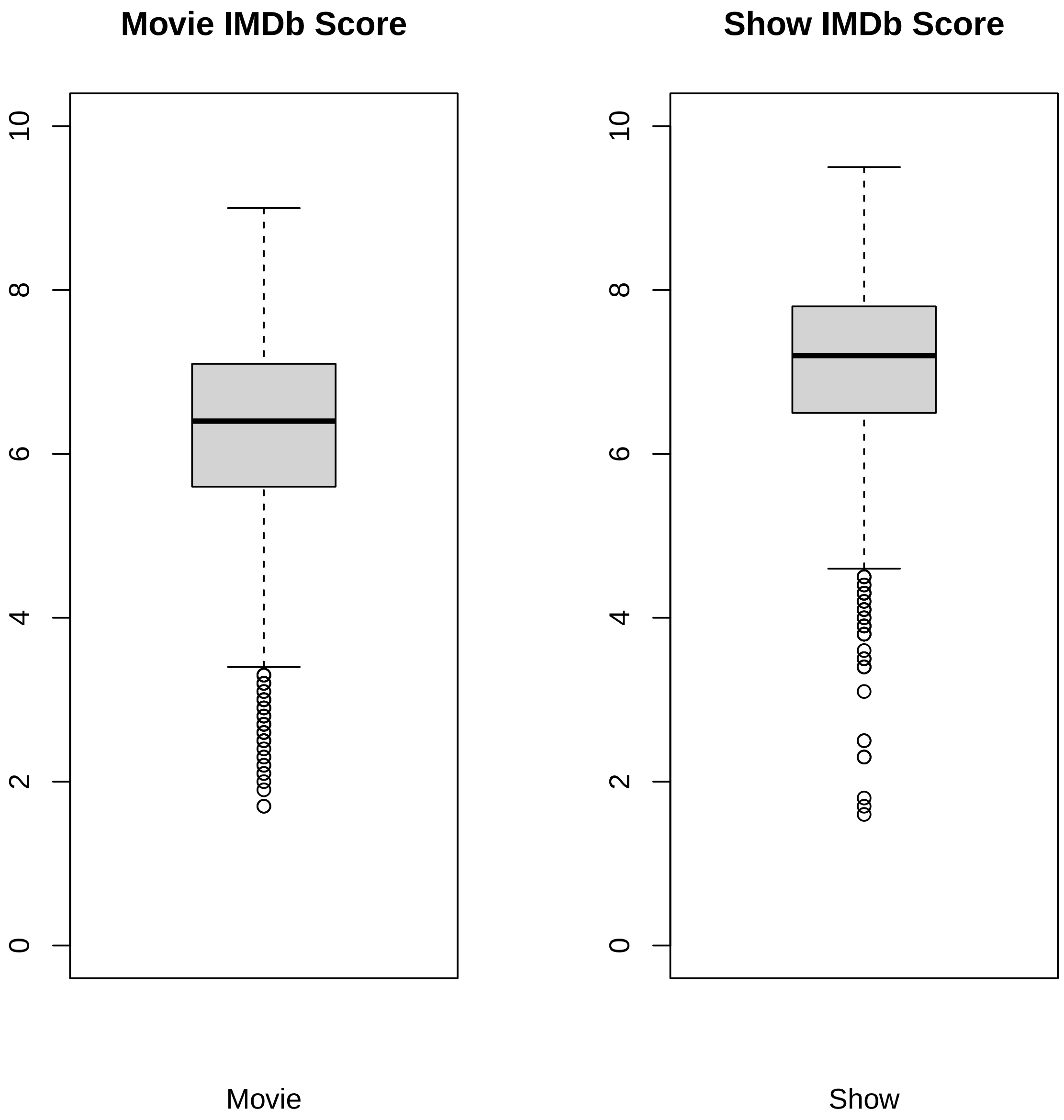
Dimension of the data for movies is,

```
[1] 3255   14
```

Dimension of the data for shows is,

```
[1] 1786   14
```

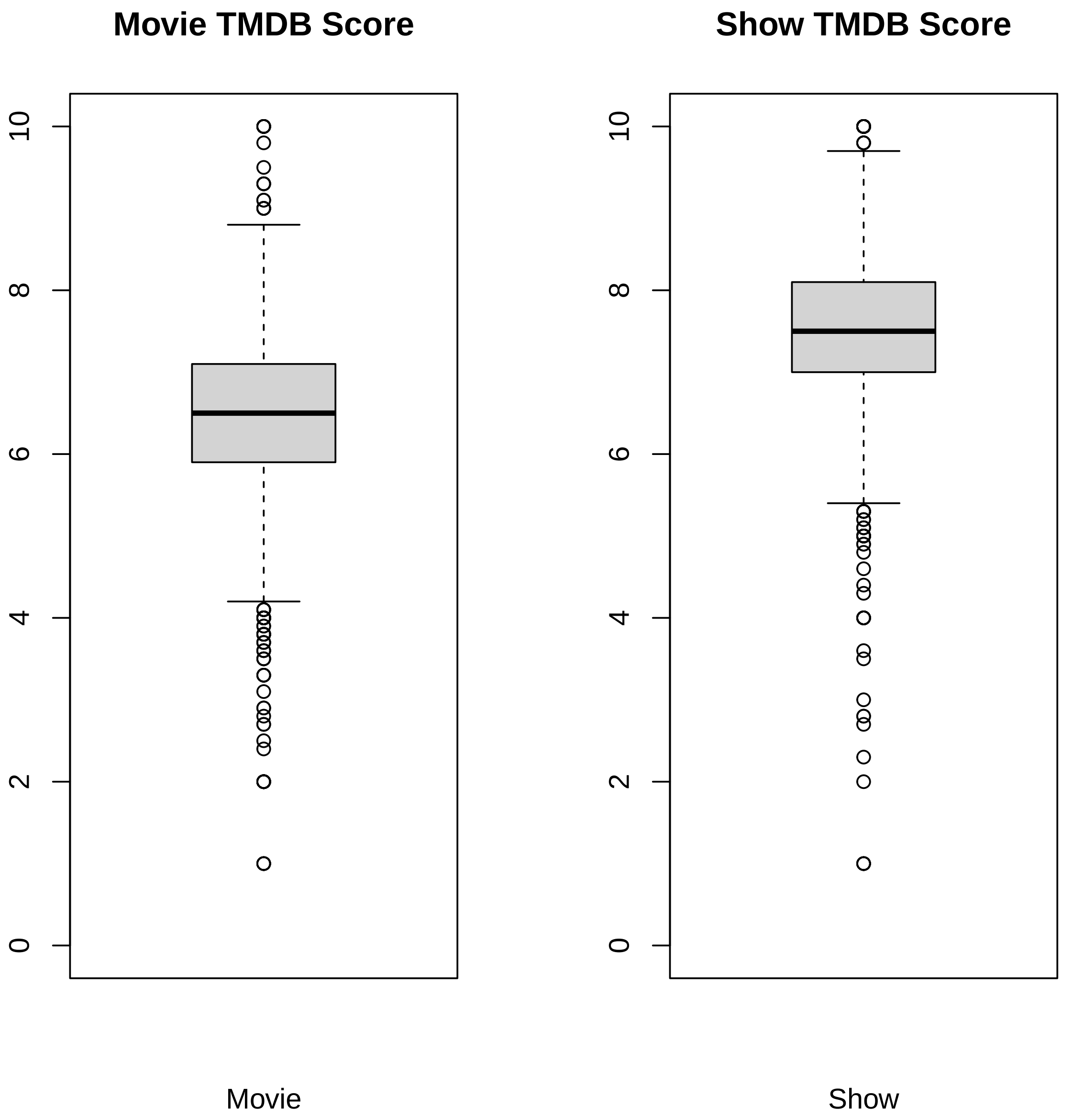
We plot the boxplot of their IMdb Ratings for comparison.



We observe that the mean IMDb score for movies (6.266329) is lower than mean IMDb score of shows (7.026708).

We leave this here for inferenial data analysis.

We plot the boxplot of their TMDB Ratings for comparison.



We observe that the mean TMDB score for movies (6.440399) is lower than mean TMDB score of shows (7.489306).

We leave this here for inferenial data analysis.

Comparison Between the IMDb and TMDB scores of Short and Long Duration movies/shows:

Now we are interested to see whether people like movies/shows of longer time duration or shorter time duration.

We define movies/shows which have runtime ≤ 80 as **short** and which have runtime > 80 as **long** movies/shows.

First we seperate all the data for short movies/shows and for long movies/shows.

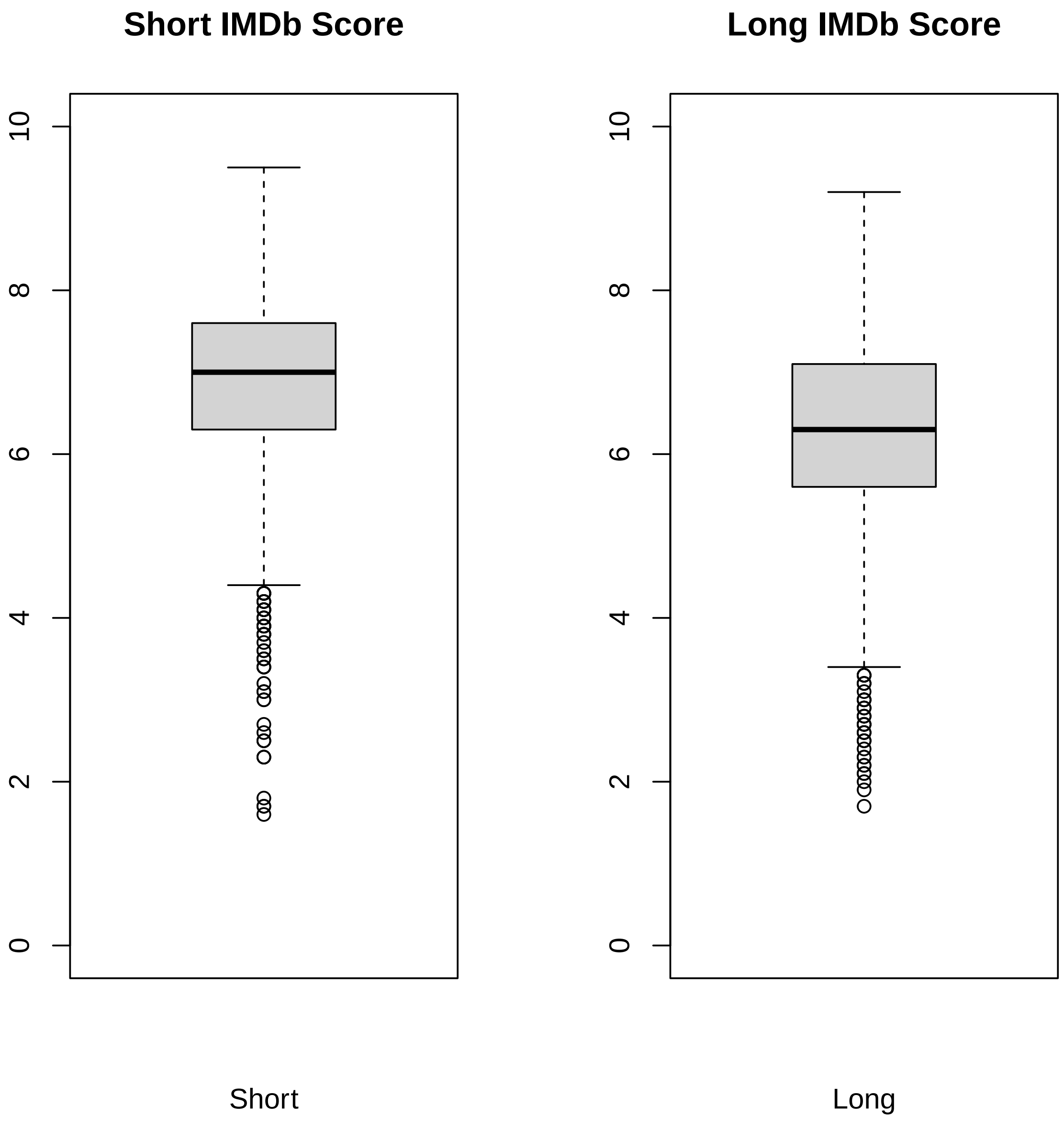
Dimension of the data for short movies/shows,

```
[1] 2287   14
```

Dimension of the data for long movies/shows,

```
[1] 2754   14
```

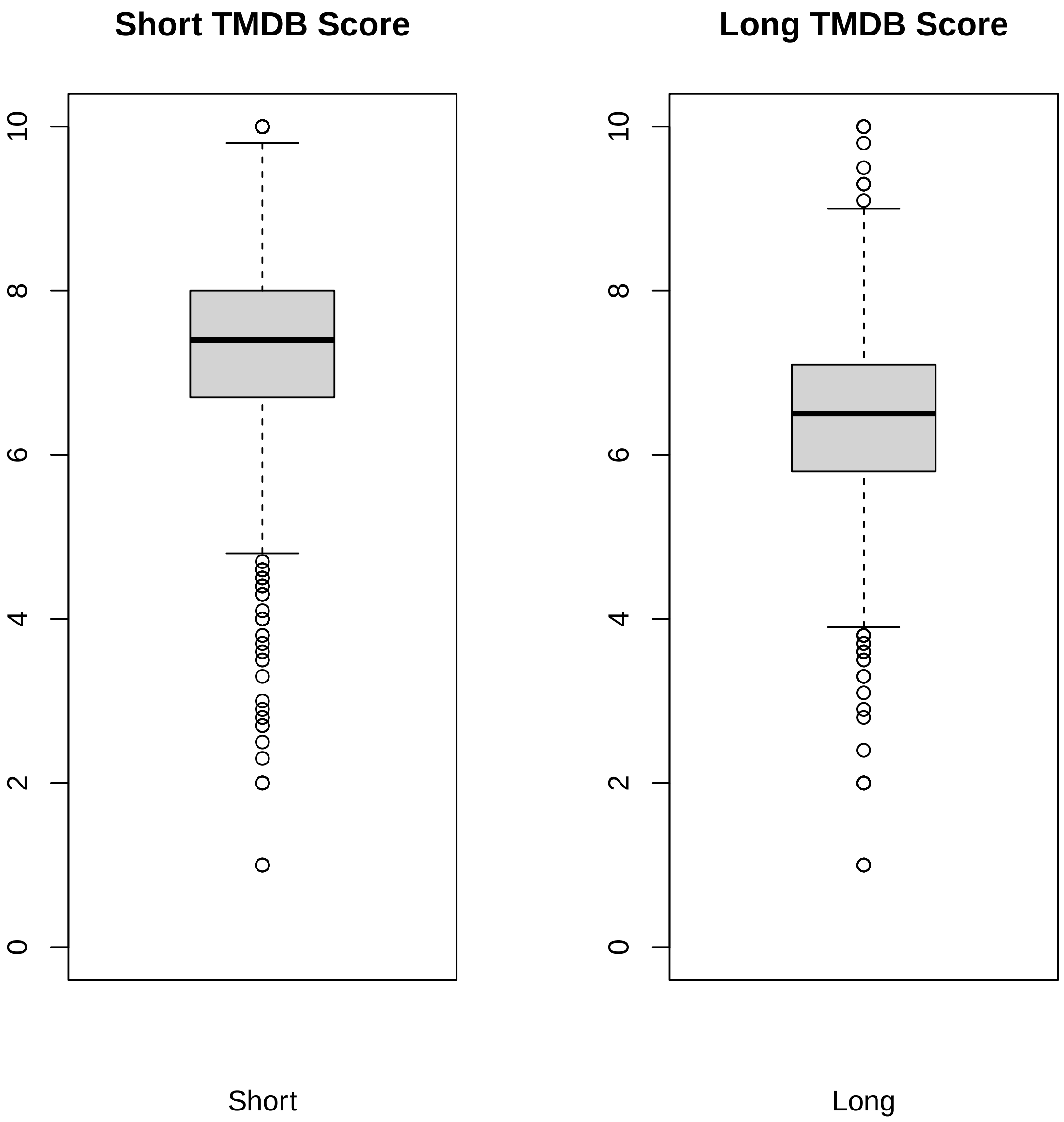
We plot the boxplot of their IMdb Ratings for comparison.



We observe that mean of the IMDb scores for short movies/shows (6.879843) is higher than mean of the IMDb scores for long movies/shows (6.249964).

We leave this here for inferential data analysis.

We plot the boxplot of their TMDB ratings for comparison.



We observe that the mean of the TMDB scores for short movies/shows (7.262352) is higher than the mean of the TMDB scores for long movies/shows (6.438054).

We leave this here for inferential data analysis.

Comparison Between IMDb and TMDB scores of Comedy and Drama Genre movies/shows:

Now we are interested to see whether people like movies/shows of Comedy genre or Drama genre.

First we seperate all the data for Comedy genre and for Drama genre.

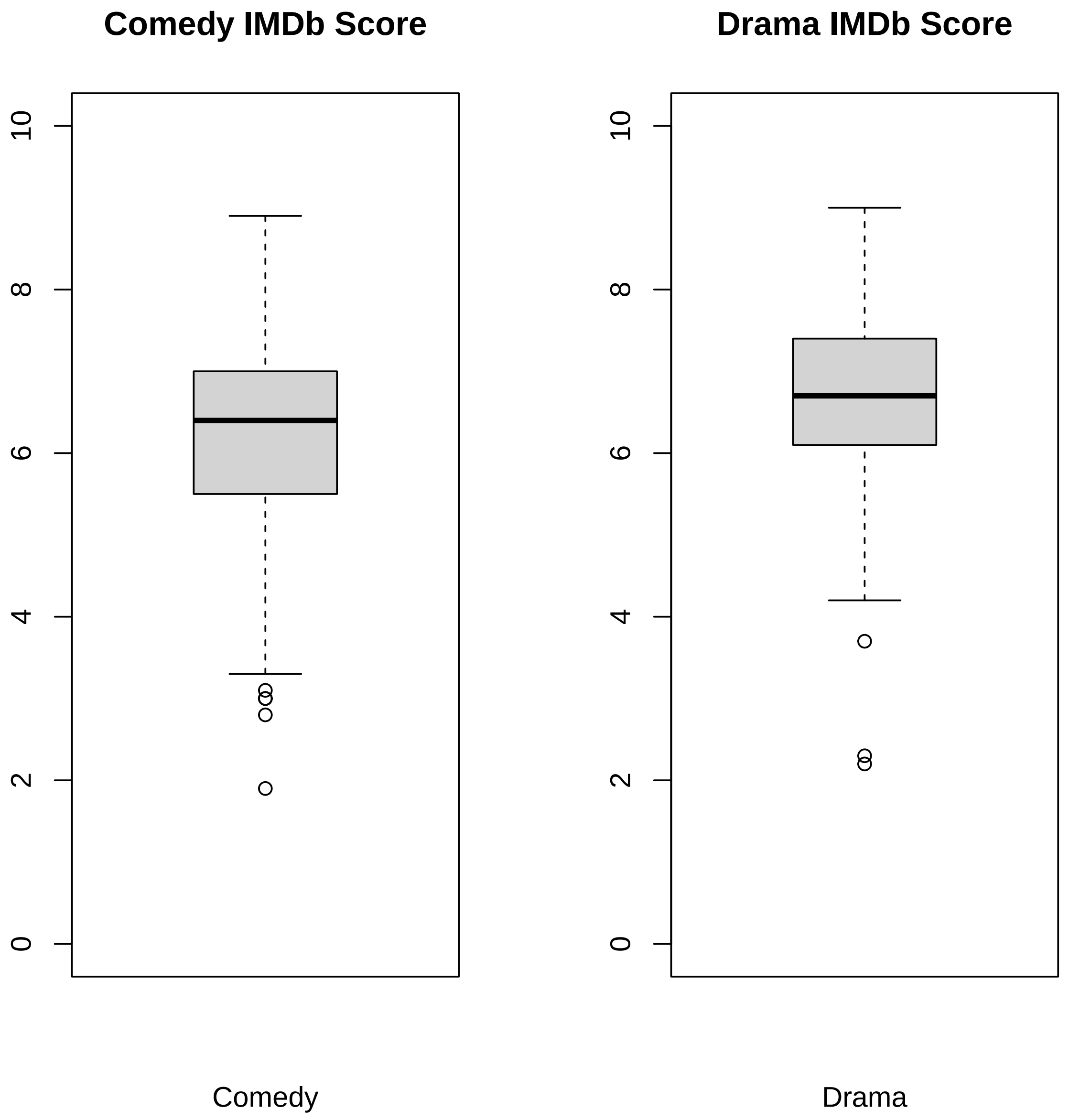
Dimension of the data for comedy genre,

```
[1] 422  14
```

Dimension of the data for Drama genre,

```
[1] 231  14
```

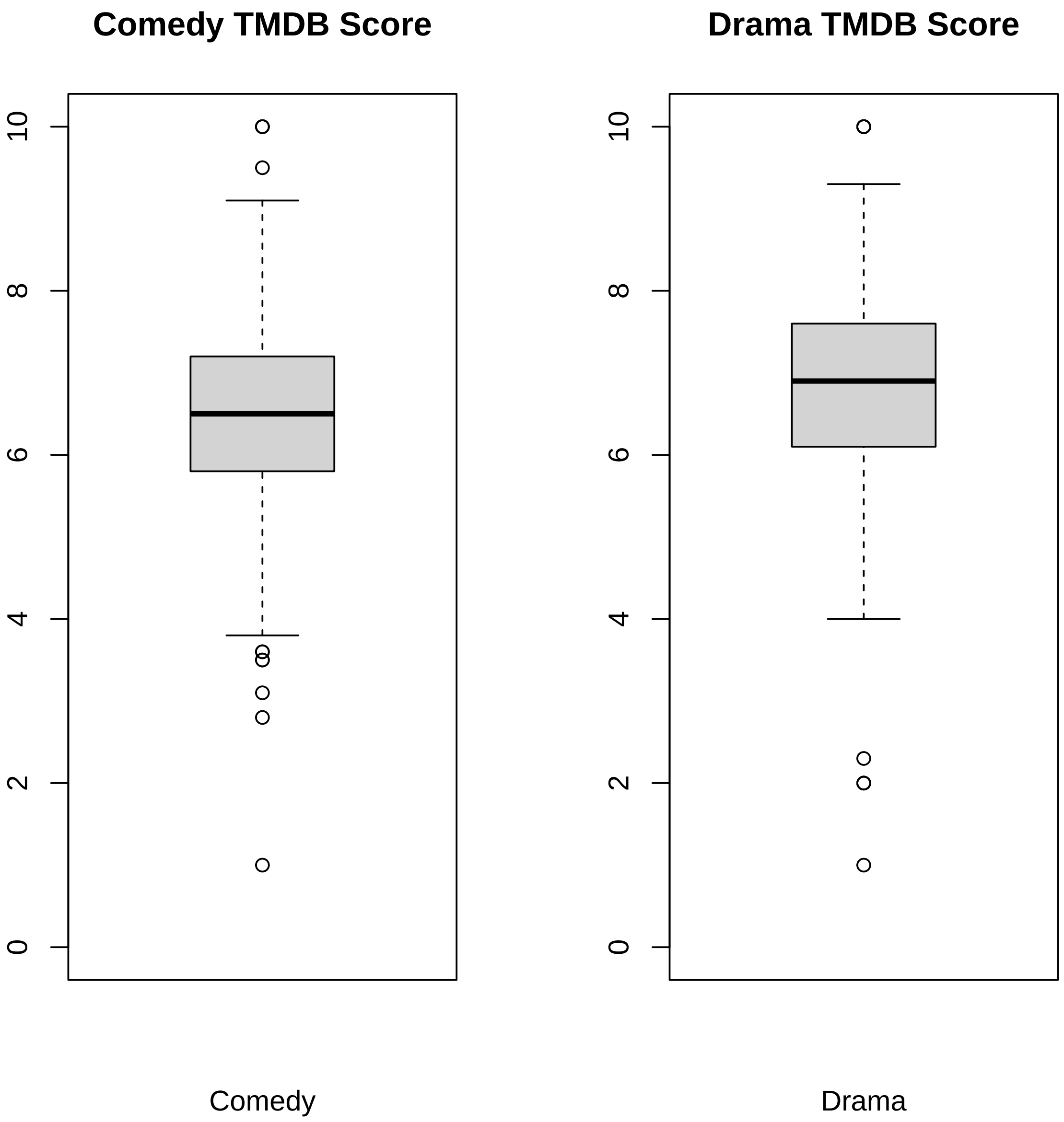
We plot the boxplot of their IMdb Ratings for comparison.



We observe that mean of IMDb scores of movies/shows of Comedy genre (6.262559) is lower than that of movies/shows of Drama genre (6.725108).

We leave it here for inferential data analysis.

We plot the boxplot of their TMDB ratings for comparison.



We observe that mean of TMDB scores of movies/shows of Comedy genre (6.466351) is lower than that of movies/shows of Drama genre (6.8).

We leave it here for inferential data analysis.

Comparison Between IMDb and TMDB scores of Old and New movies/shows:

We define movies/shows which have release year $\leq$ 2016 as **old** movies/shows and which have release year $>$ 2016 as **new** movies/shows.

First we seperate all the data for old movies/shows and for new movies/shows.

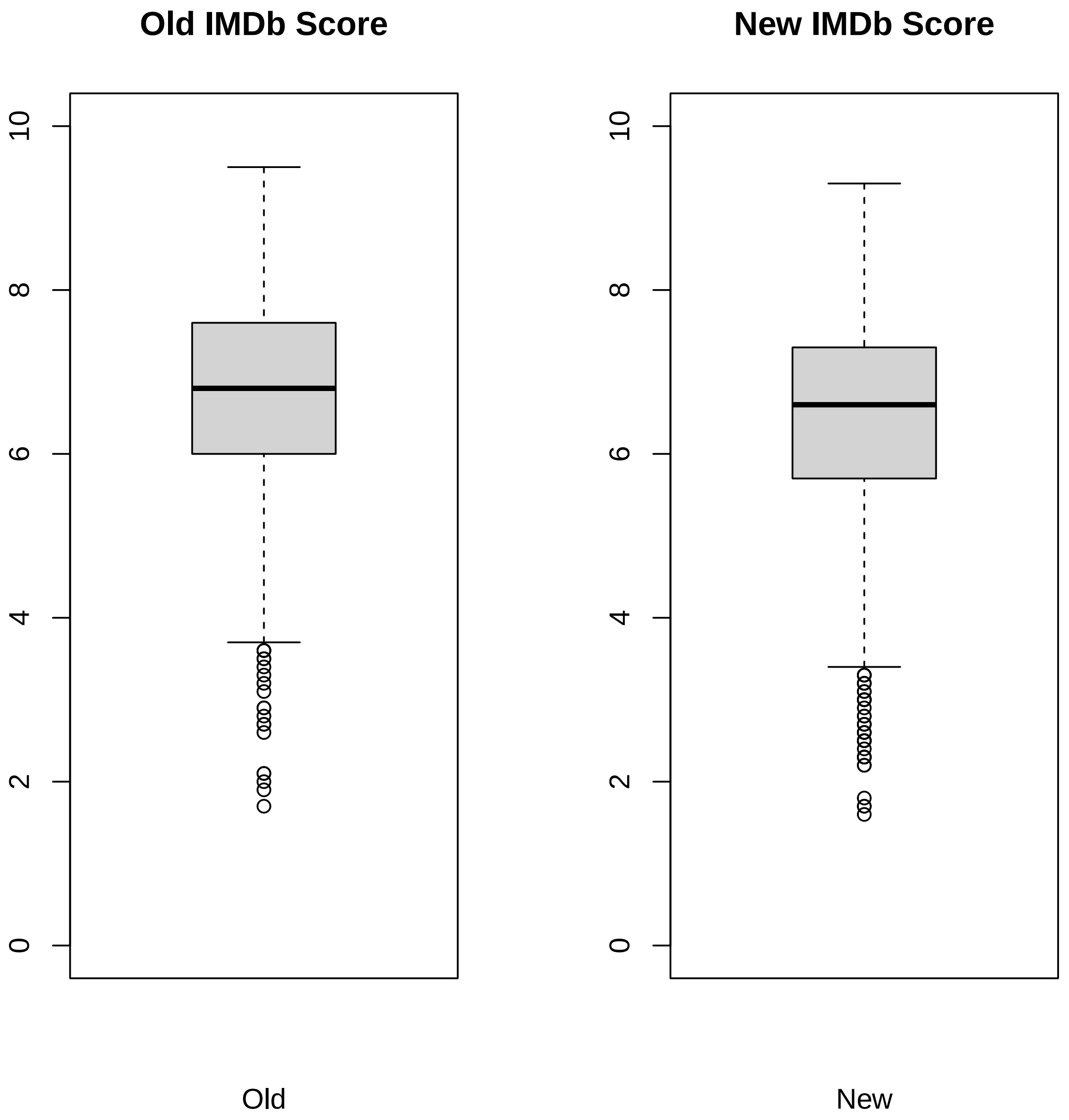
Dimension of the data for old movies/shows,

```
[1] 1643   14
```

Dimension of the data for new movies/shows,

```
[1] 3398   14
```

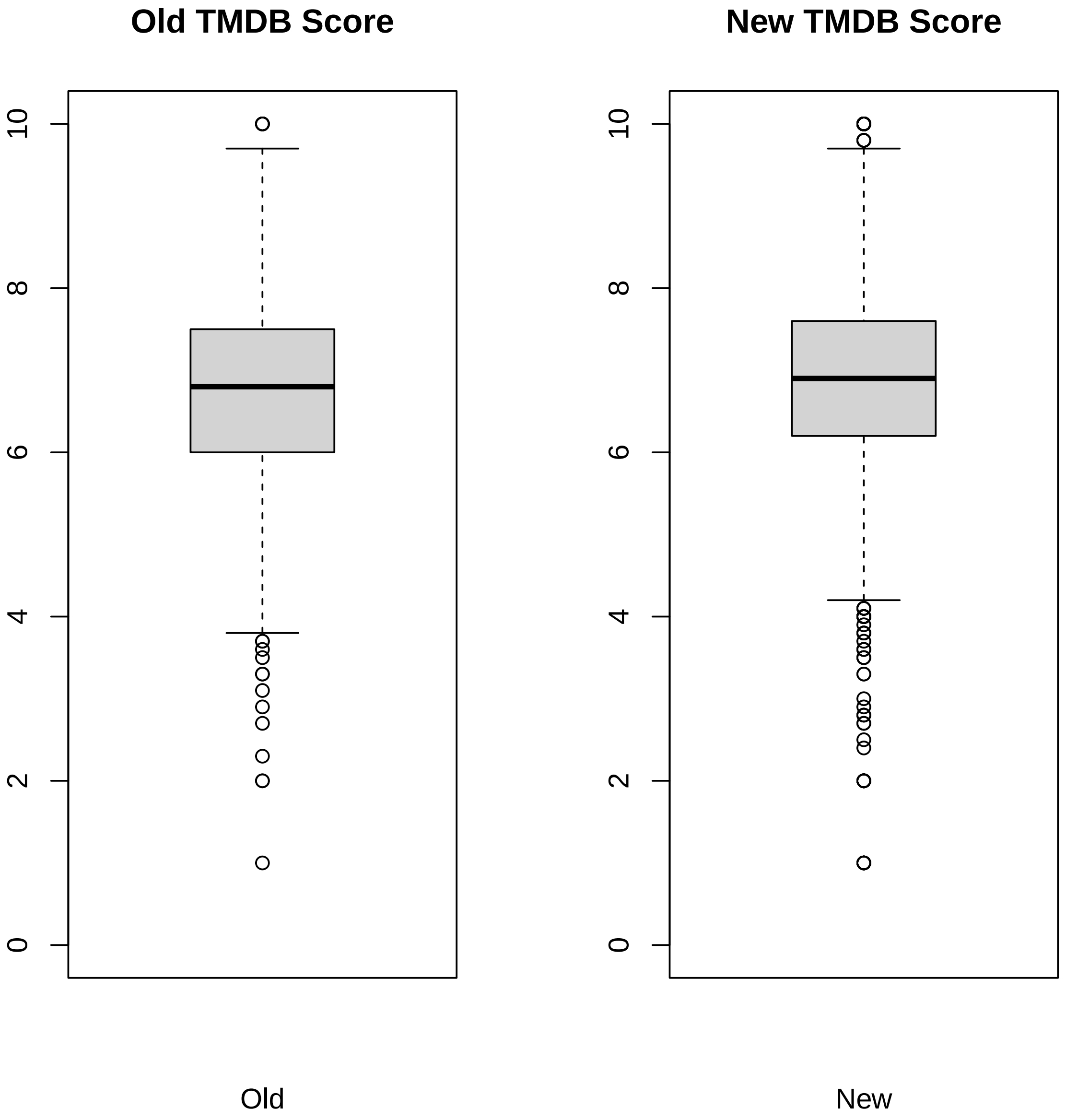
We plot the boxplot of their IMdb Ratings for comparison.



We observe that mean of IMDb scores of old movies/shows (6.72538) is higher than that of new movies/shows (6.444026).

We leave it here for inferential data analysis.

We plot the boxplot of their TMDB ratings for comparison.



We observe that mean of TMDb scores of old movies/shows (6.7056) is lower than that of new movies/shows (6.863479).

We leave it here for inferential data analysis.

Inferential Data Analysis:

Now we shall conduct inferential data analysis by various statistical methods and tools like testing of hypothesis, regrsssion etc.

First we want to test the observations we made in exploratory data analysis.

Testing Theory:

In exploratory data analysis we observed a number of differences of mean values of IMDb and TMDB ratings in different cases. Now we want to test whether the difference in mean is statistically significant or not.

Clearly this is a two sample test problem.

So, first we have to test if the population variances are equal or not. For that we conduct F-test.

- If they are equal, we perform Fisher's t-test to test whether the difference of their mean is statistically ssignificant or not.
- If they are not equal, we encounter Behren-Fisher problem. In that case we conduct Welch's approximate t-test to test whether the difference of their mean is statistically ssignificant or not.

Comparison Between IMDb and TMDB scores using Inferential Analysis:

Comparson between IMDb and TMDB scores of US and Japan based Movies/Shows:

IMDb:

First we want to test if the variance of these two samples are equal.

The null hypothesis is that the population variances are equal and the alternative hypothesis is that they are not equal.

$$H_0 : \sigma_{11}^2 = \sigma_{12}^2 \text{ Vs. } H_1 : \sigma_{11}^2 \neq \sigma_{12}^2$$

«echo=F,comment=NA» var.test(US$imdb_score, JP$imdb$_score, alternative = ”two.sided”$)@

We see that the p-value= 0.0005766 which is < 0.05.

So the null hypotheis is rejected. i.e The population variances are not equal.

So, we have to conduct Welch's approximate t-test to compare their population mean.

The null hypothesis is that the difference between the population means of IMDb ratings of US and Japan based movies is 0

and

The alternative hypotheis is that the difference between the population means of IMDb ratings of US and Japan based movies is less than 0.

$$H_0 : \delta_1 = 0 \text{ Vs. } H_1 : \delta_1 < 0$$

```
Welch Two Sample t-test

data:  US$imdb_score and JP$imdb_score
t = -8.0479, df = 338.39, p-value = 7.207e-15
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.4203099
sample estimates:
mean of x mean of y
 6.552596  7.081250
```

We see that the p-value= 0.0000000000007207 which is < 0.05.

So the null hypothesis is rejected i.e. the difference between population means is less than 0.

So, The population mean rating of movies/shows produced in US is less than the population mean rating of movies/shows produced in Japan.

TMDB:

First we want to test if the variance of these two samples are equal.

The null hypothesis is that the population variances are equal and the alternative hypothesis is that they are not equal.

$$H_0 : \sigma_{21}^2 = \sigma_{22}^2 \text{ Vs. } H_1 : \sigma_{21}^2 \neq \sigma_{22}^2$$

```
F test to compare two variances

data:  US$tmdb_score and JP$tmdb_score
F = 1.046, num df = 1771, denom df = 239, p-value = 0.6644
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8577276 1.2576991
sample estimates:
ratio of variances
          1.045998
```

We see that the p-value= 0.6644 which is > 0.05.

So the null hypotheis is accepted i.e The population variances are equal.

So, we have to conduct Fisher's t-test to compare their population mean.

The null hypothesis is that the difference between the population means of TMDB ratings of US and Japan based movies/shows is 0

and

The alternative hypotheis is that the difference between the population means of TMDB ratings of US and Japan based movies/shows is less than 0.

$$H_0 : \delta_2 = 0 \text{ Vs. } H_1 : \delta_2 < 0$$

```
Two Sample t-test

data:  US$tmdb_score and JP$tmdb_score
t = -8.4431, df = 2010, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.4740002
sample estimates:
mean of x mean of y
 6.844582  7.433333
```

We see that the p-value< 0.00000000000000022 which is < 0.05.

So the null hypothesis is rejected i.e. the difference between population means is less than 0.

So, The population mean rating of movies/shows produced in US is less than the population mean rating of movies/shows produced in Japan.

Comparison Between IMDb and TMDB Scores of Movies and Shows:

IMDb:

First we want to test if the variance of these two samples are equal.

The null hypothesis is that the population variances are equal and the alternative hypothesis is that they are not equal.

$$H_0 : \sigma_{31}^2 = \sigma_{32}^2 \text{ Vs. } H_1 : \sigma_{31}^2 \neq \sigma_{32}^2$$

```
F test to compare two variances

data:  movie$imdb_score and show$imdb_score
F = 1.0697, num df = 3254, denom df = 1785, p-value = 0.1079
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9853141 1.1601272
sample estimates:
ratio of variances
         1.069681
```

We see that the p-value$= 0.1079$ which is > 0.05.

So the null hypotheis is accepted i.e The population variances are equal.

So, we have to conduct Fisher's t-test to compare their population mean.

The null hypothesis is that the difference between the population means of IMDb ratings of movies and shows is 0

and

The alternative hypotheis is that the difference between the population means of IMDb ratings of movies and shows is less than 0.

$$H_0 : \delta_3 = 0 \text{ Vs. } H_1 : \delta_3 < 0$$

```
Two Sample t-test

data:  movie$imdb_score and show$imdb_score
t = -23.629, df = 5039, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.7074388
sample estimates:
mean of x mean of y
 6.266329  7.026708
```

We see that the p-value< 0.00000000000000022 which is < 0.05.

So the null hypothesis is rejected i.e. the difference between population means is less than 0.

So, The population mean rating of movies less than the population mean rating of shows.

TMDB:

First we want to test if the variance of these two samples are equal.

The null hypothesis is that the population variances are equal and the alternative hypothesis is that they are not equal.

$$H_0 : \sigma_{41}^2 = \sigma_{42}^2 \text{ Vs. } H_1 : \sigma_{41}^2 \neq \sigma_{42}^2$$

```
F test to compare two variances

data:  movie$tmdb_score and show$tmdb_score
F = 0.92909, num df = 3254, denom df = 1785, p-value = 0.0758
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8558162 1.0076540
sample estimates:
ratio of variances
         0.9290946
```

We see that the p-value$= 0.0758$ which is > 0.05.

So the null hypotheis is accepted i.e The population variances are equal.

So, we have to conduct Fisher's t-test to compare their population mean.

The null hypothesis is that the difference between the population means of TMDB ratings of movies and shows is 0

and

The alternative hypotheis is that the difference between the population means of TMDB ratings of movies and shows is less than 0.

$$H_0 : \delta_4 = 0 \text{ Vs. } H_1 : \delta_4 < 0$$

```
Two Sample t-test

data:  movie$tmdb_score and show$tmdb_score
t = -34.648, df = 5039, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.9991025
sample estimates:
mean of x mean of y
 6.440399  7.489306
```

We see that the p-value< 0.00000000000000022 which is < 0.05.

So the null hypothesis is rejected i.e. the difference between population means is less than 0.

So, The population mean rating of movies less than the population mean rating of shows.

Comparison Between the IMDb and TMDB scores of Short and Long Duration movies/shows:

IMDb:

First we want to test if the variance of these two samples are equal.

The null hypothesis is that the population variances are equal and the alternative hypothesis is that they are not equal.

$$H_0 : \sigma_{51}^2 = \sigma_{52}^2 \text{ Vs. } H_1 : \sigma_{51}^2 \neq \sigma_{52}^2$$

```
F test to compare two variances

data:  short$imdb_score and long$imdb_score
F = 0.96576, num df = 2286, denom df = 2753, p-value = 0.3849
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8930177 1.0447226
sample estimates:
ratio of variances
         0.9657568
```

We see that the p-value$= 0.3849$ which is > 0.05.

So the null hypotheis is accepted i.e The population variances are equal.

So, we have to conduct Fisher's t-test to compare their population mean.

The null hypothesis is that the difference between the population means of IMDb ratings of short and long duration movies/shows is 0

and

The alternative hypotheis is that the difference between the population means of IMDb ratings of short and long duration movies/shows is greater than 0.

$$H_0 : \delta_5 = 0 \text{ Vs. } H_1 : \delta_5 > 0$$

```
Two Sample t-test

data:  short$imdb_score and long$imdb_score
t = 20.091, df = 5039, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.5782999       Inf
sample estimates:
mean of x mean of y
 6.879843  6.249964
```

We see that the p-value<0.00000000000000022 which is <0.05.

So the null hypothesis is rejected i.e. the difference between population means is greater than 0.

So, The population mean rating of short duration movies/shows is greater than the population mean rating of long duration.

TMDB:

First we want to test if the variance of these two samples are equal.

The null hypothesis is that the population variances are equal and the alternative hypothesis is that they are not equal.

$$H_0 : \sigma^2_{61} = \sigma^2_{62} \text{ Vs. } H_1 : \sigma^2_{61} \neq \sigma^2_{62}$$

```
F test to compare two variances

data:  short$tmdb_score and long$tmdb_score
F = 1.3046, num df = 2286, denom df = 2753, p-value = 2.705e-11
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.206296 1.411220
sample estimates:
ratio of variances
          1.304552
```

We see that the p-value$=0.00000000002705$ which is <0.05.

So the null hypotheis is rejected i.e The population variances are not equal.

So, we have to conduct Welch's approximate t–test to compare their population mean.

The null hypothesis is that the difference between the population means of TMDB ratings of short and long duration movies/shows is 0

and

The alternative hypotheis is that the difference between the population means of TMDB ratings of short and long duration movies/shows is greater than 0.

$$H_0 : \delta_6 = 0 \text{ Vs. } H_1 : \delta_6 > 0$$

```
Welch Two Sample t-test

data:  short$tmdb_score and long$tmdb_score
t = 26.955, df = 4581.2, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.7739871       Inf
sample estimates:
mean of x mean of y
 7.262352  6.438054
```

We see that the p-value< 0.00000000000000022 which is < 0.05.

So the null hypothesis is rejected i.e. the difference between population means is greater than 0.

So, The population mean rating of short duration movies/shows is greater than the population mean rating of long duration.

Comparison Between IMDb and TMDB scores of Comedy and Drama Genre movies/shows:

IMDb:

First we want to test if the variance of these two samples are equal.

The null hypothesis is that the population variances are equal and the alternative hypothesis is that they are not equal.

$$H_0 : \sigma^2_{71} = \sigma^2_{72} \text{ Vs. } H_1 : \sigma^2_{71} \neq \sigma^2_{72}$$

```
F test to compare two variances

data:  comedy$imdb_score and drama$imdb_score
F = 1.2677, num df = 421, denom df = 230, p-value = 0.04475
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.005614 1.585773
sample estimates:
ratio of variances
           1.26768
```

We see that the p-value$= 0.04475$ which is < 0.05.

So the null hypotheis is rejected i.e The population variances are not equal.

So, we have to conduct Welch's approximate t-test to compare their population mean.

The null hypothesis is that the difference between the population means of IMDb ratings of comedy and drama genre movies/shows is 0

and

The alternative hypotheis is that the difference between the population means of IMDb ratings of comedy and drama genre movies/shows is less than 0.

$$H_0 : \delta_7 = 0 \text{ Vs. } H_1 : \delta_7 < 0$$

```
Welch Two Sample t-test

data:  comedy$imdb_score and drama$imdb_score
t = -5.219, df = 522.5, p-value = 1.3e-07
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.3165089
sample estimates:
```

```
mean of x mean of y 
 6.262559  6.725108 
```

We see that the p-value= 0.00000013 which is < 0.05.

So the null hypothesis is rejected i.e. the difference between population means is less than 0.

So, The population mean rating of comedy genre movies/shows is less than the population mean rating of drama genre movies/shows.

TMDB:

First we want to test if the variance of these two samples are equal.

The null hypothesis is that the population variances are equal and the alternative hypothesis is that they are not equal.

$$H_0 : \sigma^2_{81} = \sigma^2_{82} \text{ Vs. } H_1 : \sigma^2_{81} \neq \sigma^2_{82}$$

```
F test to compare two variances

data:  comedy$tmdb_score and drama$tmdb_score
F = 0.72807, num df = 421, denom df = 230, p-value = 0.005377
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5775596 0.9107652
sample estimates:
ratio of variances 
         0.7280734 
```

We see that the p-value= 0.005377 which is < 0.05.

So the null hypotheis is rejected i.e The population variances are not equal.

So, we have to conduct Welch's approximate t-test to compare their population mean.

The null hypothesis is that the difference between the population means of TMDB ratings of comedy and drama genre movies/shows is 0

and

The alternative hypotheis is that the difference between the population means of TMDB ratings of comedy and drama genre movies/shows is less than 0.

$$H_0 : \delta_8 = 0 \text{ Vs. } H_1 : \delta_8 < 0$$

```
Welch Two Sample t-test

data:  comedy$tmdb_score and drama$tmdb_score
t = -3.2464, df = 413.94, p-value = 0.000632
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1642186
sample estimates:
```

```
mean of x mean of y
 6.466351  6.800000
```

We see that the p-value= 0.000632 which is < 0.05.

So the null hypothesis is rejected i.e. the difference between population means is less than 0.

So, The population mean rating of comedy genre movies/shows is less than the population mean rating of drama genre movies/shows.

Comparison Between IMDb and TMDB scores of Old and New movies/shows:

IMDb:

First we want to test if the variance of these two samples are equal.

The null hypothesis is that the population variances are equal and the alternative hypothesis is that they are not equal.

$$H_0 : \sigma^2_{91} = \sigma^2_{92} \text{ Vs. } H_1 : \sigma^2_{91} \neq \sigma^2_{82}$$

```
F test to compare two variances

data:  old$imdb_score and new$imdb_score
F = 1.0291, num df = 1642, denom df = 3397, p-value = 0.4954
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9474364 1.1192655
sample estimates:
ratio of variances
          1.029142
```

We see that the p-value= 0.4954 which is > 0.05.

So the null hypotheis is accepted i.e The population variances are equal.

So, we have to conduct Fisher's t-test to compare their population mean.

The null hypothesis is that the difference between the population means of IMDb ratings of old and new movies/shows is 0

and

The alternative hypotheis is that the difference between the population means of IMDb ratings of old and new movies/shows is greater than 0.

$$H_0 : \delta_9 = 0 \text{ Vs. } H_1 : \delta_9 > 0$$

```
Two Sample t-test

data:  old$imdb_score and new$imdb_score
t = 8.1835, df = 5039, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.2247927       Inf
sample estimates:
mean of x mean of y
 6.725380  6.444026
```

We see that the p-value< 0.00000000000000022 which is < 0.05.

So the null hypothesis is rejected i.e. the difference between population means is greater than 0.

So, The population mean rating of old movies/shows is greater than the population mean rating of new movies/shows.

TMDB:

First we want to test if the variance of these two samples are equal.

The null hypothesis is that the population variances are equal and the alternative hypothesis is that they are not equal.

$$H_0 : \sigma^2_{10,1} = \sigma^2_{10,2} \text{ Vs. } H_1 : \sigma^2_{10,1} \neq \sigma^2_{10,2}$$

```
F test to compare two variances

data:  old$tmdb_score and new$tmdb_score
F = 0.85742, num df = 1642, denom df = 3397, p-value = 0.0003434
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7893454 0.9325028
sample estimates:
ratio of variances
         0.8574177
```

We see that the p-value= 0.0003434 which is < 0.05.

So the null hypotheis is rejected i.e The population variances are not equal.

So, we have to conduct Welch's approximate t-test to compare their population mean.

The null hypothesis is that the difference between the population means of TMDB ratings of old and new movies/shows is 0

and

The alternative hypotheis is that the difference between the population means of TMDB ratings of old and new movies/shows is less than 0.

$$H_0 : \delta_{10} = 0 \text{ Vs. } H_1 : \delta_{10} < 0$$

```
Welch Two Sample t-test

data:  old$tmdb_score and new$tmdb_score
t = -4.7272, df = 3481, p-value = 1.183e-06
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1029294
sample estimates:
mean of x mean of y
 6.705600  6.863479
```

We see that the p-value= 0.000001183 which is < 0.05.

So the null hypothesis is rejected i.e. the difference between population means is less than 0.

So, The population mean rating of old movies/shows is less than the population mean rating of new movies/shows.

Model Fitting:

Now we want to predict IMDb ratings and TMDb ratings of a show using other predictor variables like release year, runtime, imdb votes, type, tmdb popularity. So we fit regression models.

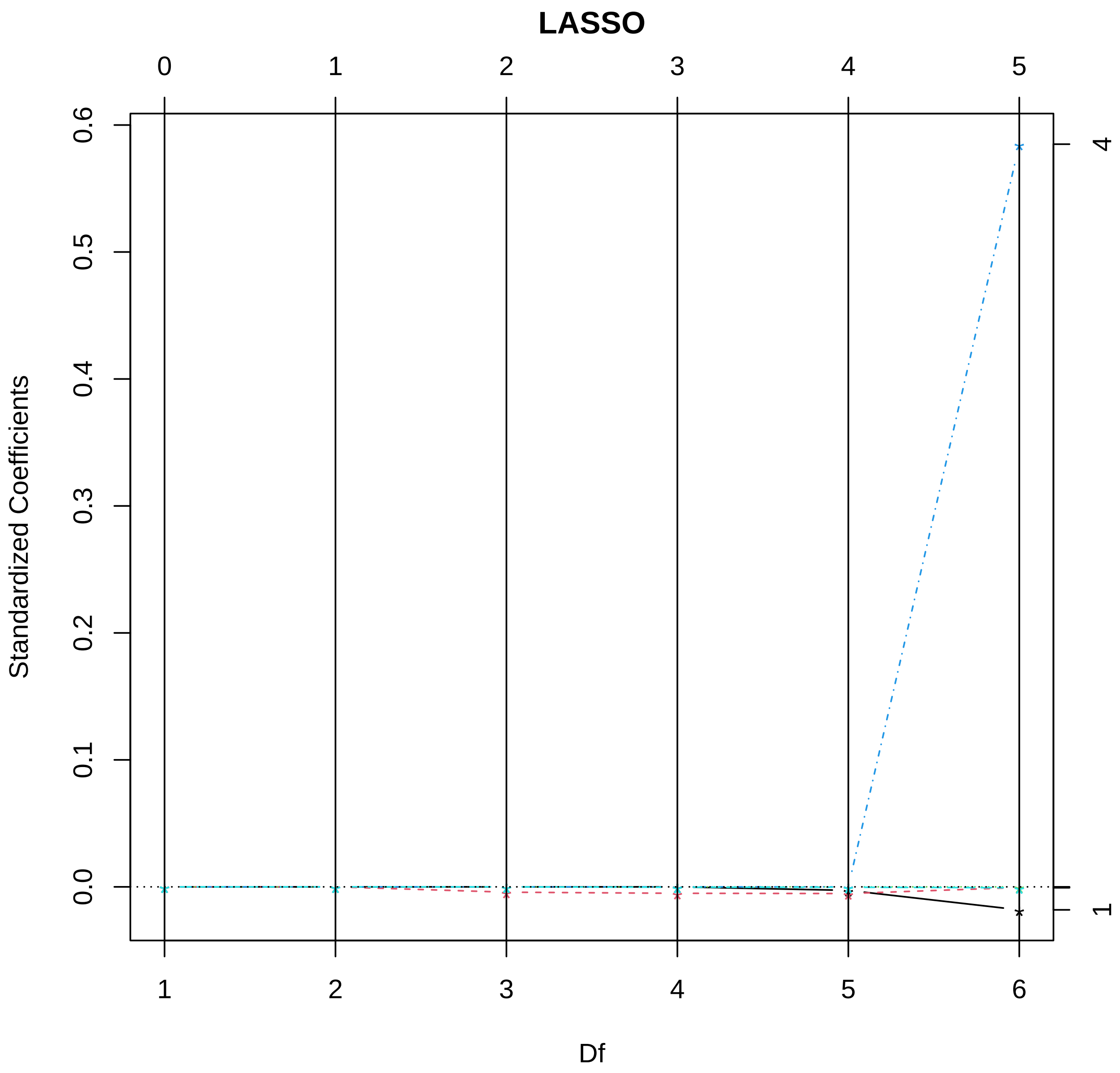
First we conduct variable selection process using LASSO method.

Variable SelctionUsing LASSO Mehtod in case of IMDb Score:

Here we conduct variable selection using LASSO in order to fit a model to predict IMDb score of a movie/show using the predictors release year, runtime, imdb votes, tmdb popularity, tmdb score. We are discarding the 'type' covariate as LASSO method can not deal with factor covariates.

```
Loaded lars 1.3
```

The plot of coefficient paths is,

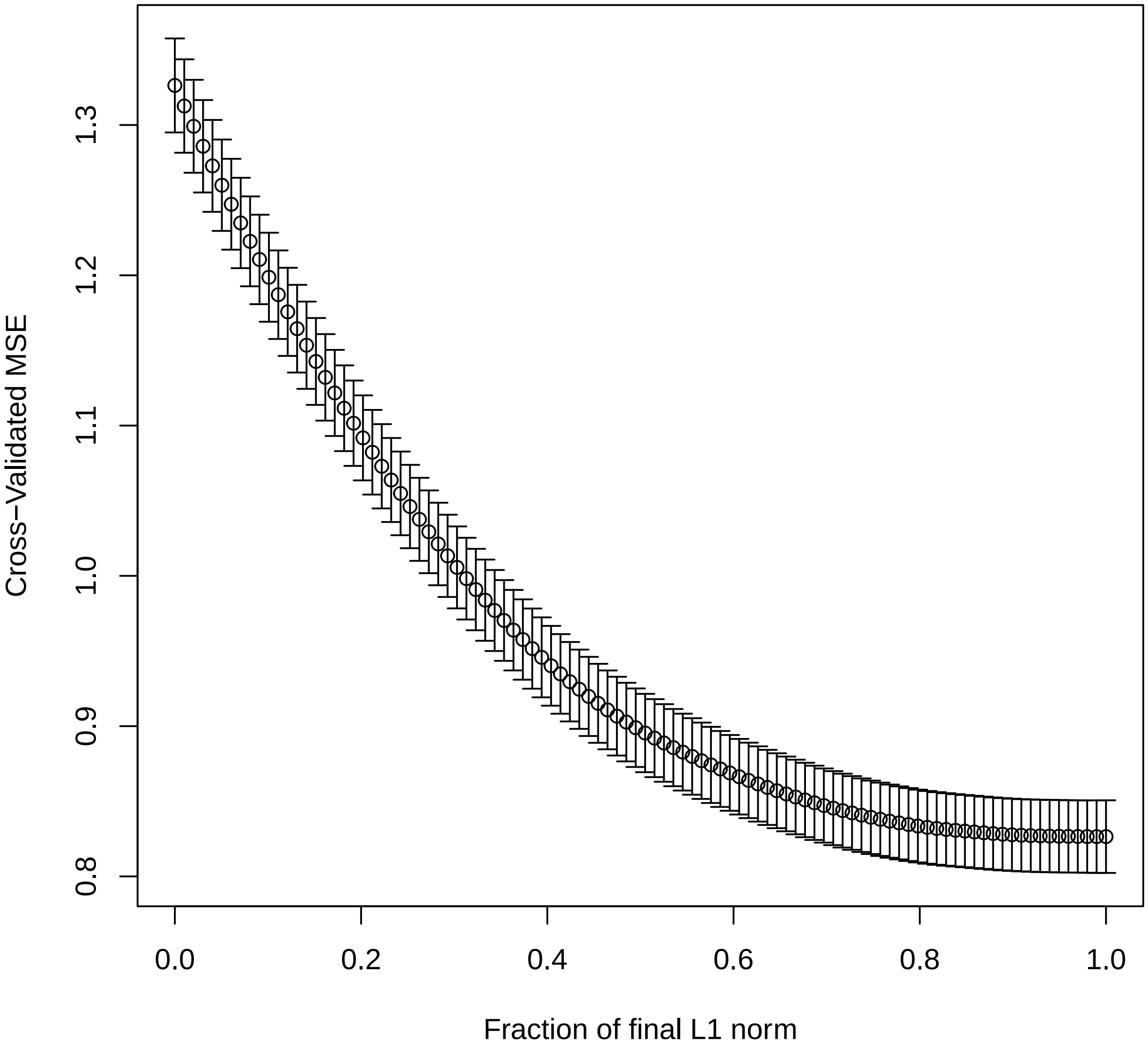


The coefficients for each step in path are,

```
     netflix.release_year netflix.runtime netflix.imdb_votes netflix.tmdb_score
[1,]          0.000000000    0.0000000000       0.000000e+00          0.0000000
[2,]          0.000000000    0.0000000000       2.522213e-06          0.0000000
[3,]          0.000000000   -0.0041858865       2.781194e-06          0.0000000
[4,]          0.000000000   -0.0050711100       2.882032e-06          0.0000000
[5,]         -0.002765165   -0.0052622011       2.850019e-06          0.0000000
[6,]         -0.018108059   -0.0004934913       1.535795e-06          0.5849196
     netflix.tmdb_popularity
[1,]            0.0000000000
[2,]            0.0000000000
[3,]            0.0000000000
[4,]           -0.0002810311
[5,]           -0.0002951927
```

```
[6,]            -0.0006809550
```

Using cross validation to estimate optimal position in path,
The cross validation MSE plot,



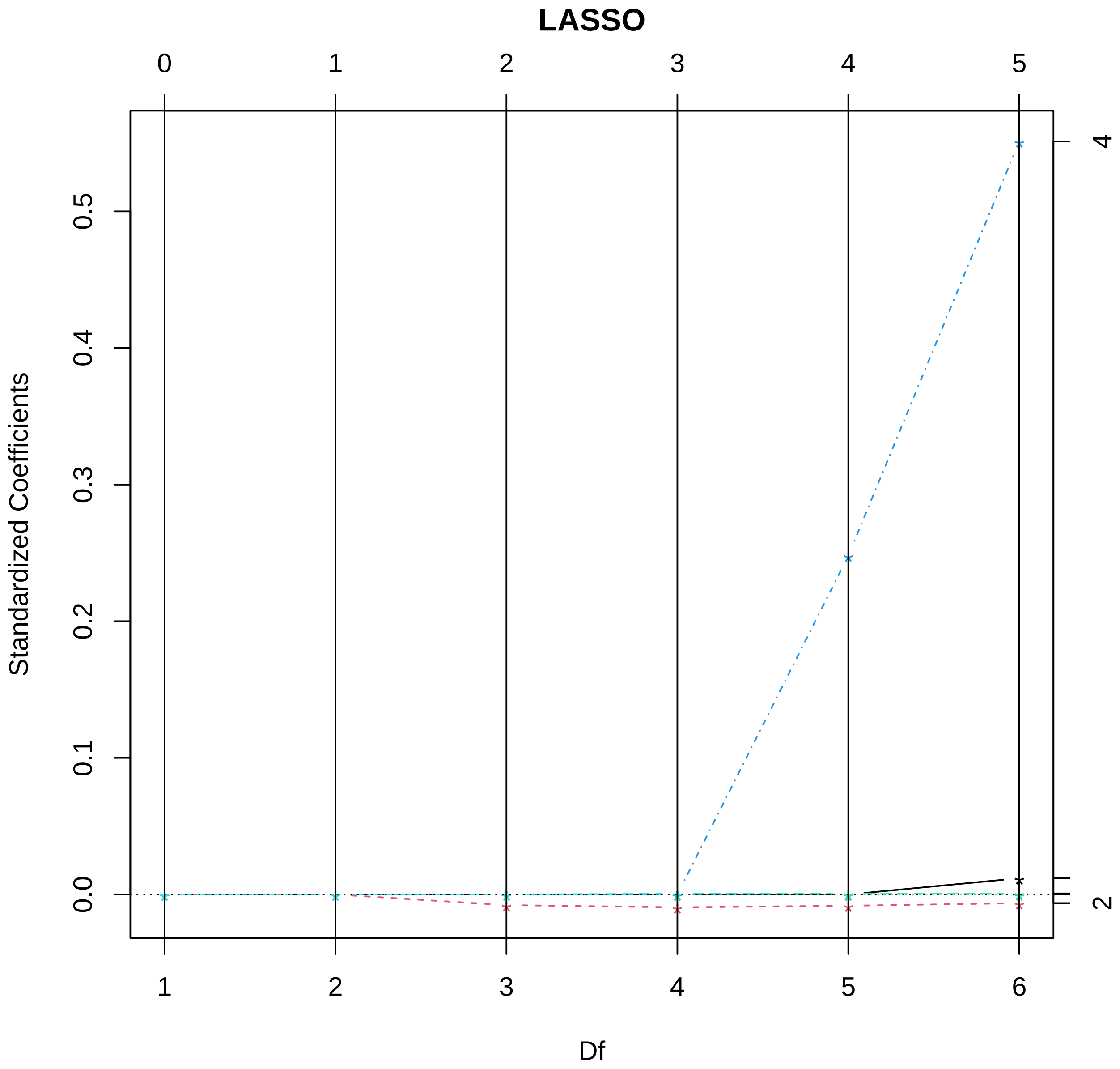
The coefficients are,

```
   netflix.release_year          netflix.runtime       netflix.imdb_votes 
          -1.323659e-02            -2.007587e-03             1.953069e-06 
     netflix.tmdb_score netflix.tmdb_popularity 
           3.992040e-01            -5.584730e-04 
```

Fitting a LASSO Regression Model to Predict TMDB Score:

Here we conduct variable selection using LASSO in order to fit a model to predict TMDB score of a movie/show using the predictors release year, runtime, imdb votes, tmdb popularity, imdb score. We are discarding the 'type' covariate as LASSO method can not deal with factor covariates.

The plot of coefficient paths is,

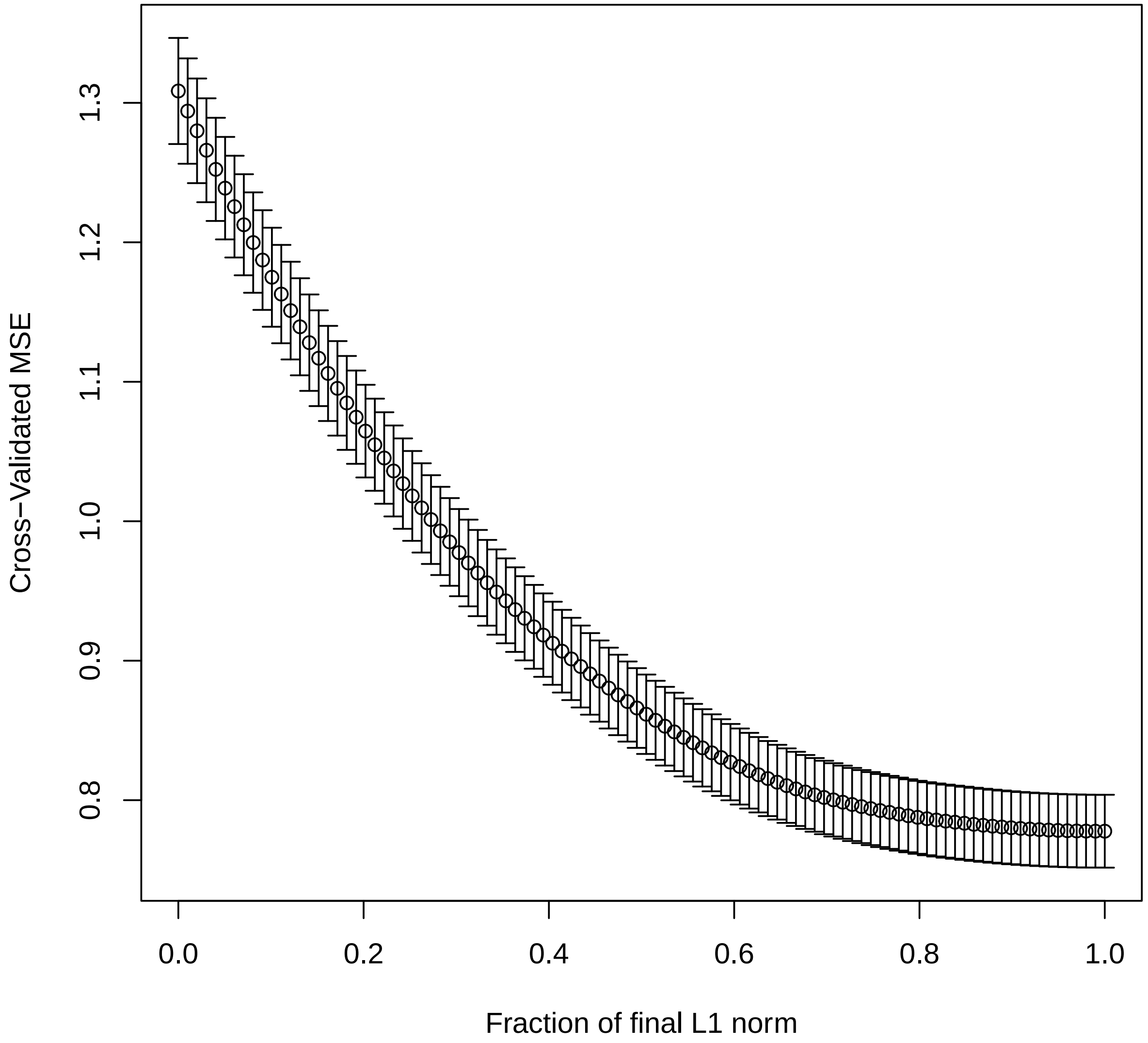


The coefficients for each step in path are,

```
     netflix.release_year netflix.runtime netflix.imdb_votes netflix.imdb_score
[1,]           0.00000000     0.000000000       0.000000e+00          0.0000000
[2,]           0.00000000     0.000000000       1.412695e-06          0.0000000
[3,]           0.00000000    -0.007772044       1.893551e-06          0.0000000
[4,]           0.00000000    -0.009425612       1.923848e-06          0.0000000
[5,]           0.00000000    -0.008227342       1.196642e-06          0.2479855
[6,]           0.01189899    -0.006347887       4.922919e-07          0.5512193
     netflix.tmdb_popularity
[1,]            0.0000000000
[2,]            0.0000000000
```

```
[3,]            0.0000000000
[4,]            0.0004392619
[5,]            0.0006076067
[6,]            0.0007410238
```

Using cross validation to estimate optimal position in path,
The cross validation MSE plot,



The coefficients are,

```
netflix.release_year         netflix.runtime      netflix.imdb_votes
        5.119612e-03           -7.418696e-03            8.935914e-07
  netflix.imdb_score netflix.tmdb_popularity
        3.784536e-01            6.650102e-04
```

We see that in both cases all the variables are significant.

Fitting Multiple Linear Regression Model to Predict IMDb Score:

Here we try to fit a multiple linear regression model to predict IMDb score of a movie/show using the predictors release year, runtime, imdb votes, type, tmdb popularity, tmdb score.

The model will be of the form,

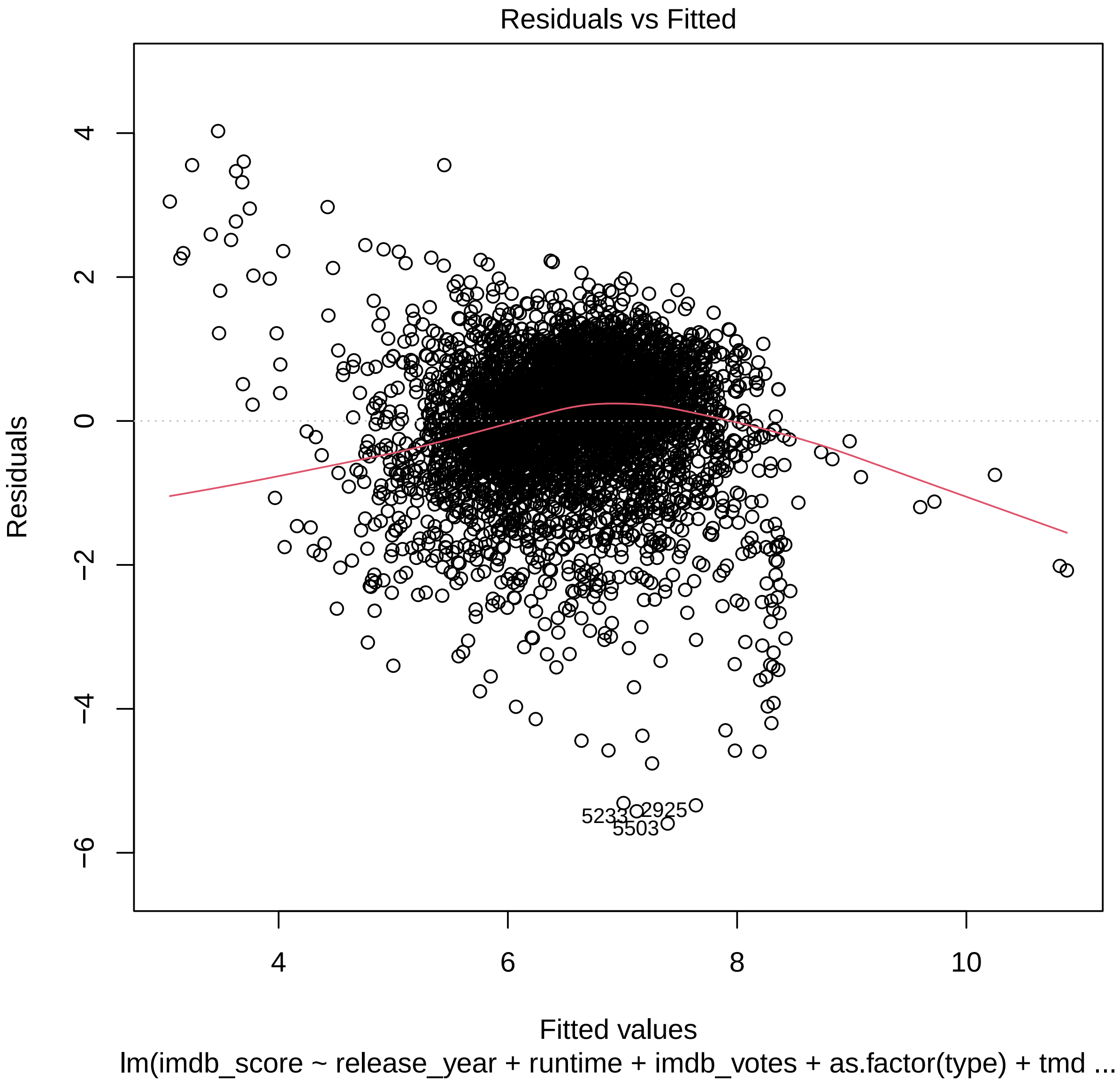
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} \;\forall i = 1(1)5041$$

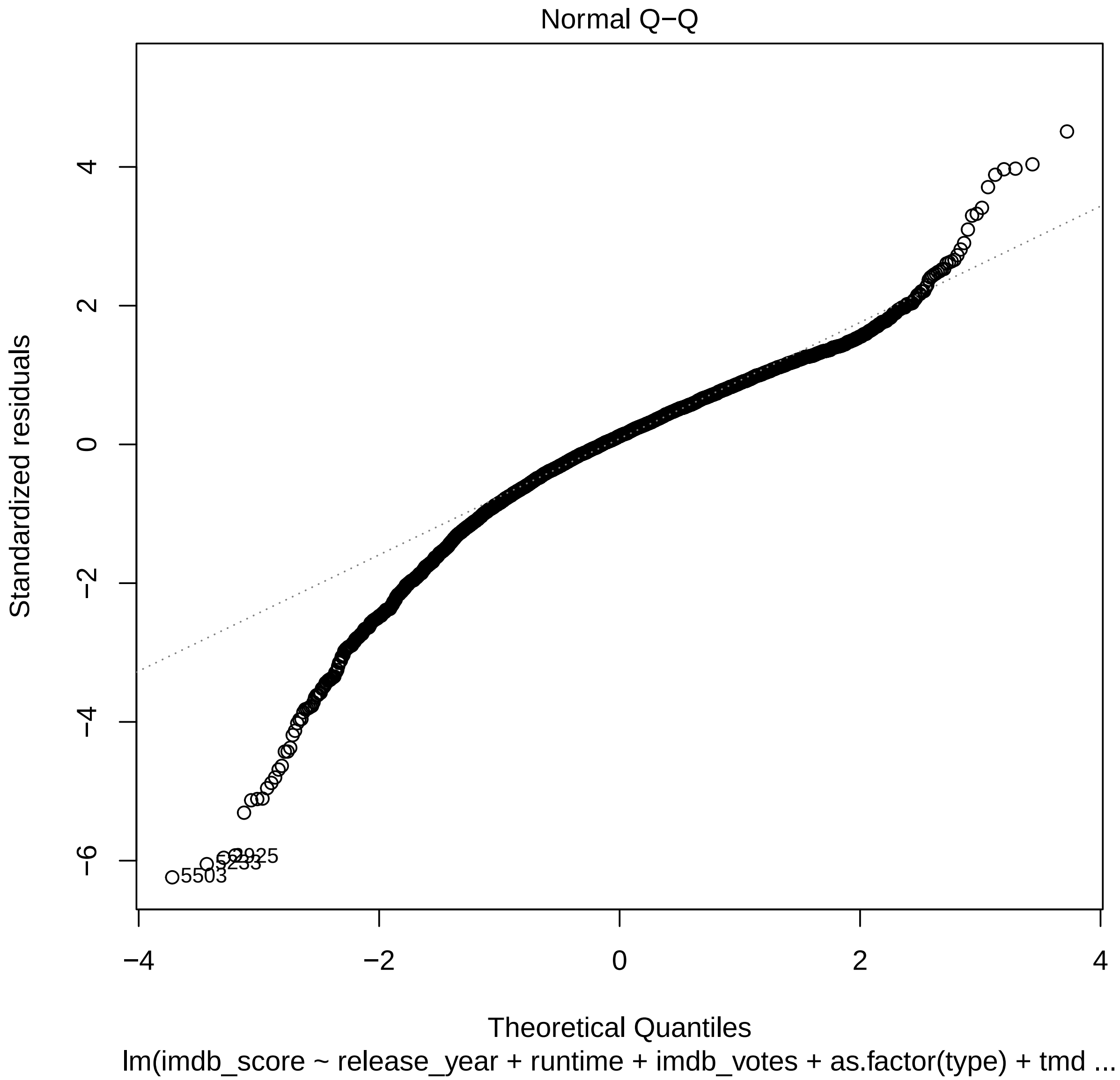
where,

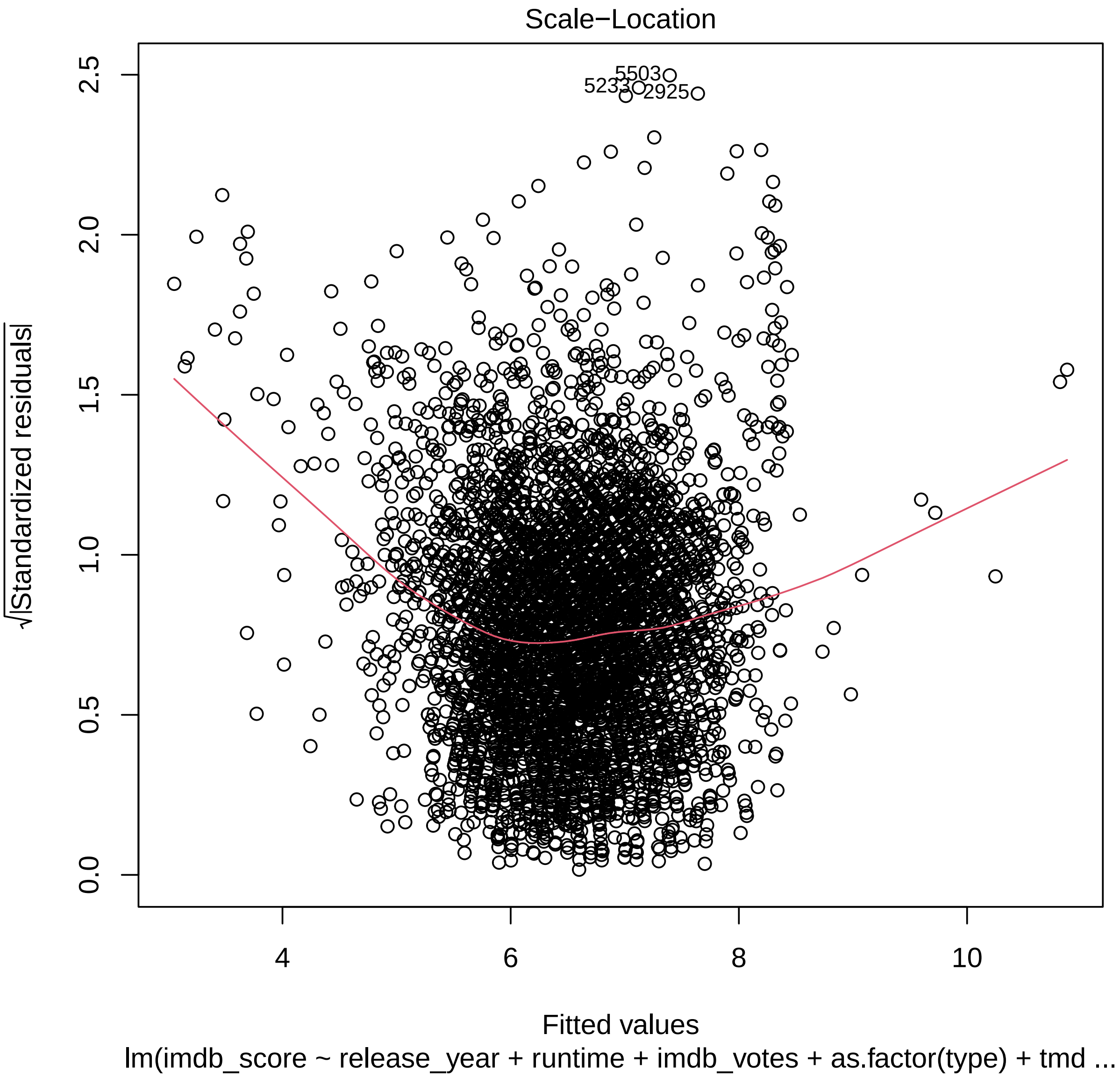
x_{i1} = Release year of the i^{th} movie/show.
x_{i2} = Runtime of the i^{th} movie/show.
x_{i3} = IMDb votes of the i^{th} movie/show.
x_{i4} = level 'show' of the factor covariate Type of the i^{th} movie/show.
x_{i5} = TMDB score of the i^{th} movie/show.
x_{i6} = TMDB popularity of the i^{th} movie/show.

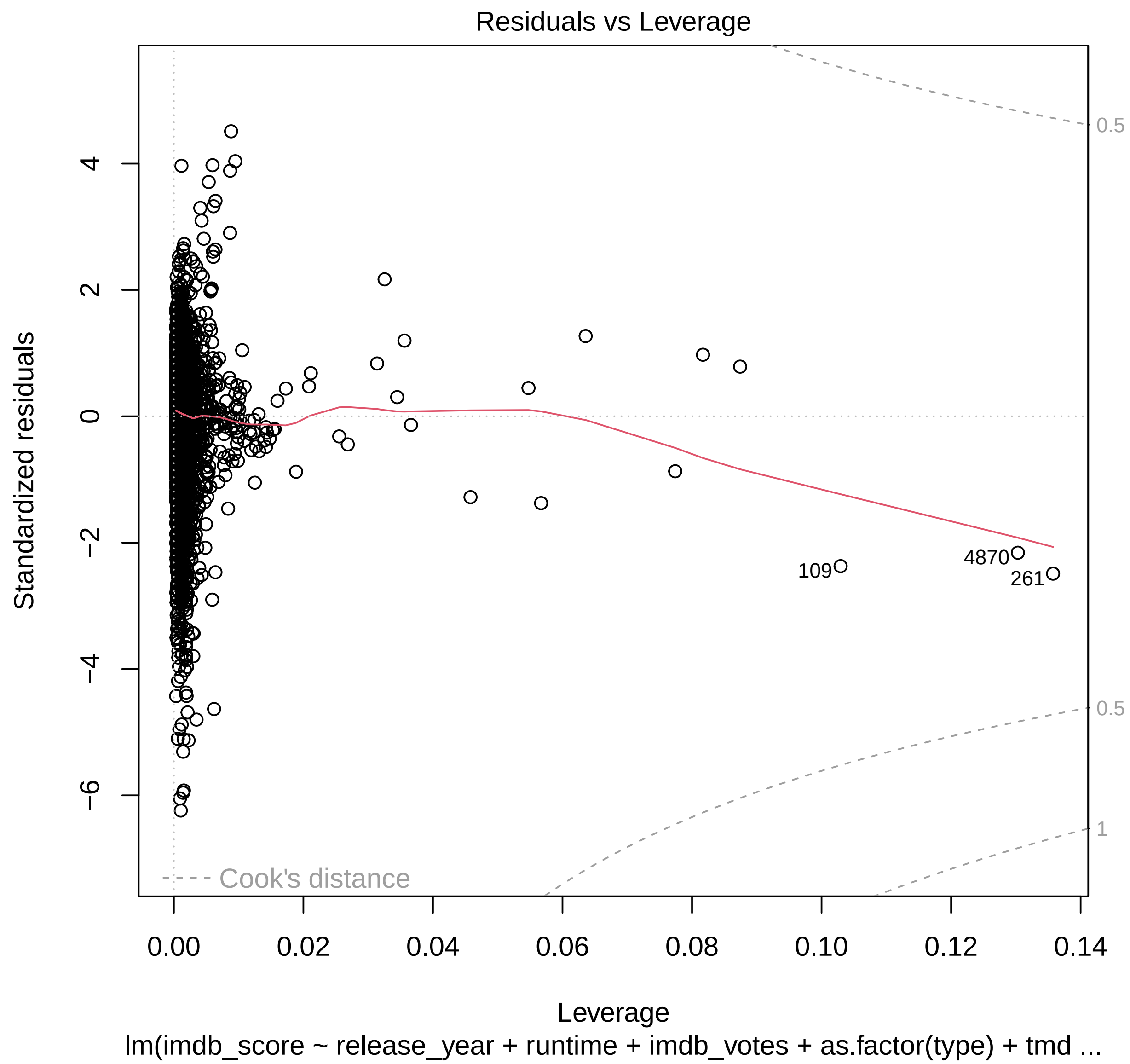
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.2132	3.5876	10.37	0.0000
release_year	-0.0173	0.0018	-9.73	0.0000
runtime	0.0039	0.0005	7.38	0.0000
imdb_votes	0.0000	0.0000	9.87	0.0000
as.factor(type)SHOW	0.4980	0.0445	11.20	0.0000
tmdb_score	0.5410	0.0124	43.49	0.0000
tmdb_popularity	-0.0008	0.0002	-4.32	0.0000

We plot some graphs, Q-Q plot to understand the model better.


Residuals vs Fitted
Residuals
4
2
0
−2
−4
−6
Fitted values
4
6
8
10
5233
2925
5503
lm(imdb_score ~ release_year + runtime + imdb_votes + as.factor(type) + tmd ...


Normal Q−Q
Standardized residuals
4
2
0
−2
−4
−6
−4
−2
0
2
4
Theoretical Quantiles
lm(imdb_score ~ release_year + runtime + imdb_votes + as.factor(type) + tmd ...
5503
5233
2925


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
4
6
8
10
Fitted values
5503
5233
2925
lm(imdb_score ~ release_year + runtime + imdb_votes + as.factor(type) + tmd ...



After fitting the model we see that all the covariates are significant to the model.

Fitting Multiple Linear Regression Model to Predict TMDB Score:

Here we try to fit a multiple linear regression model to predict TMDB score of a movie/show using the predictors release year, runtime, imdb votes, type, tmdb popularity, imdb score.

The model will be of the form,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} \ \forall i = 1(1)5041$$

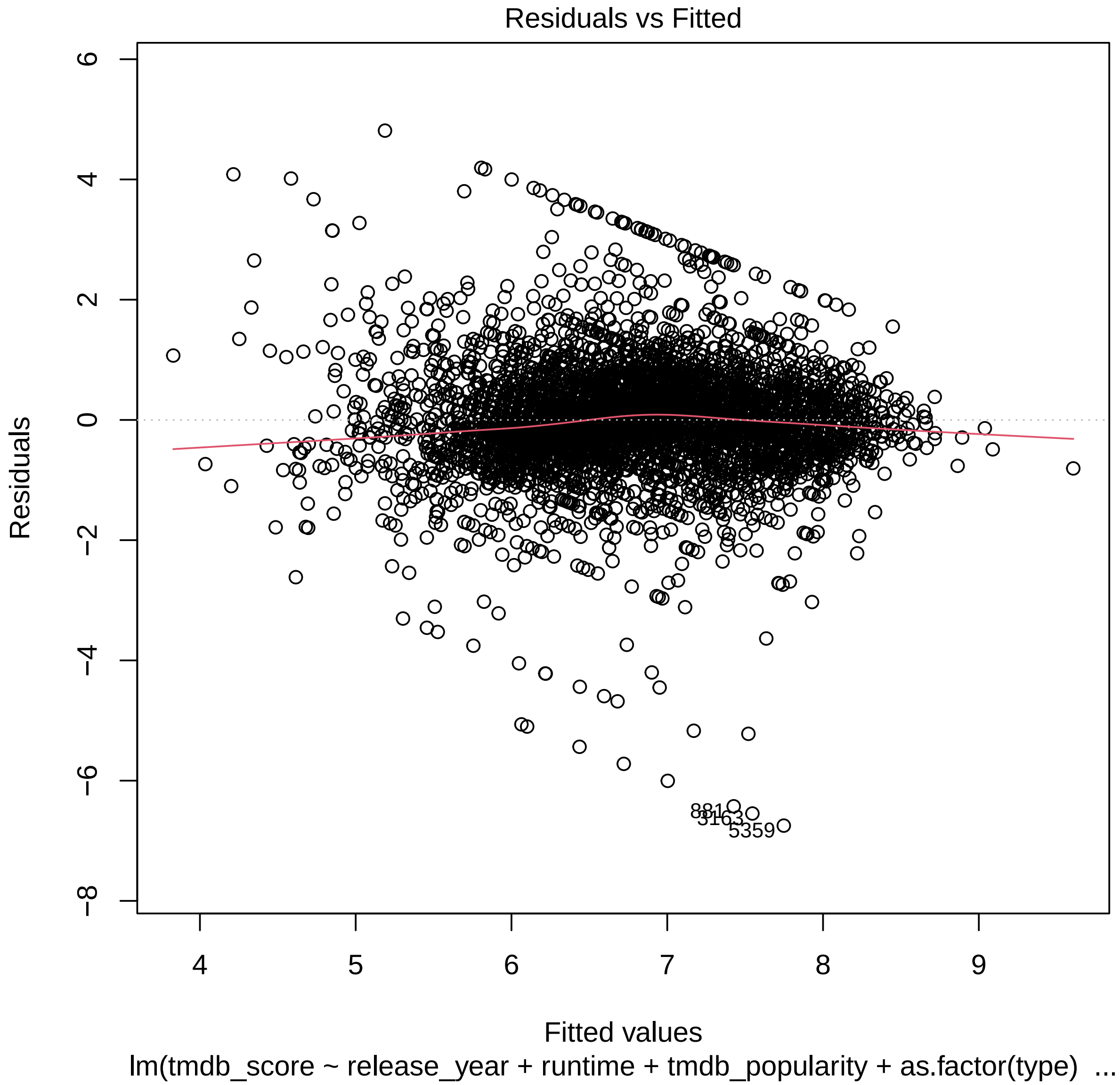
where,

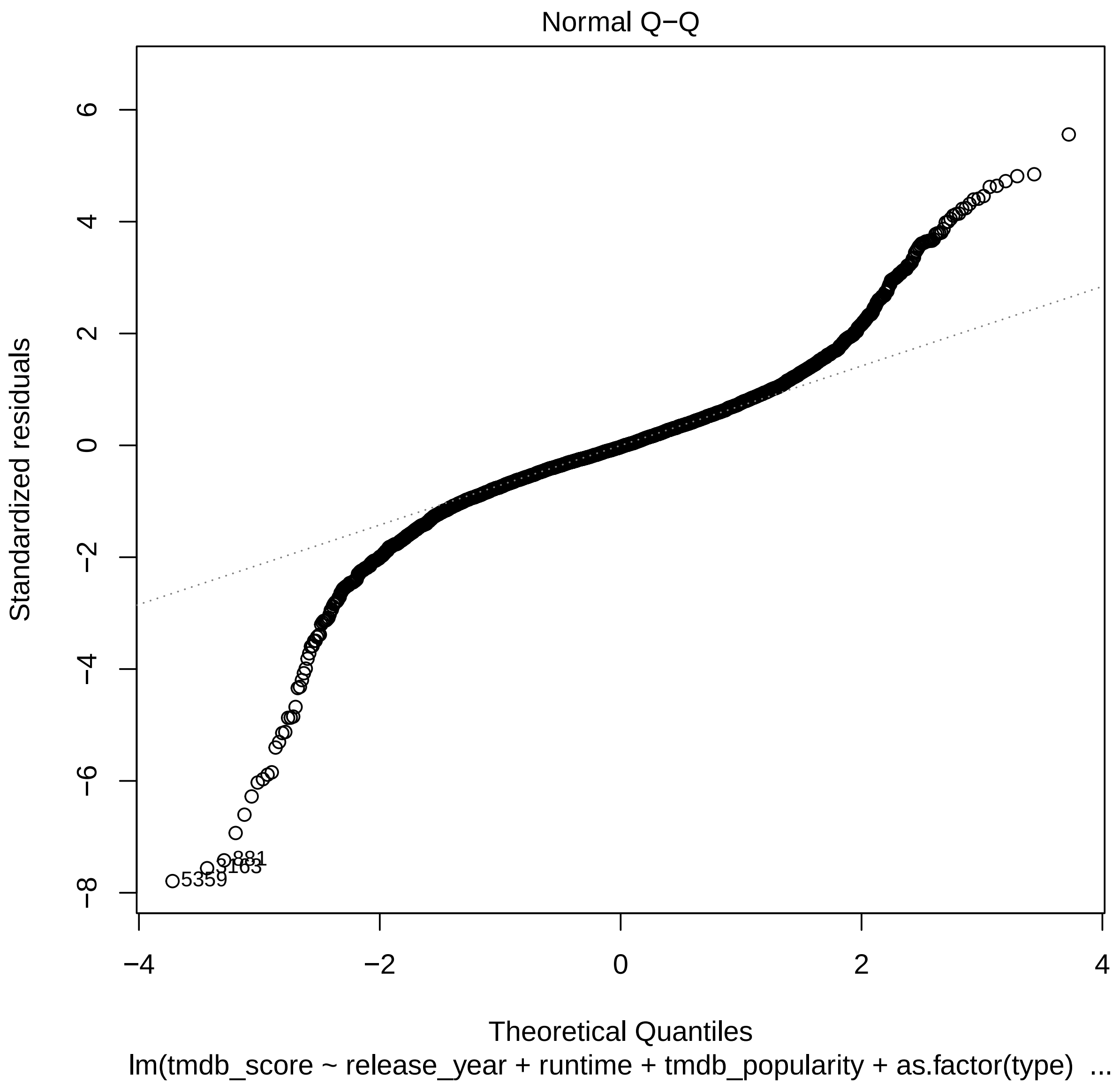
x_{i1} = Release year of the i^{th} movie/show.
x_{i2} = Runtime of the i^{th} movie/show.
x_{i3} = TMDB popularity of the i^{th} movie/show.
x_{i4} = level 'show' of the factor covariate 'Type' of the i^{th} movie/show.
x_{i5} = IMDb score of the i^{th} movie/show.

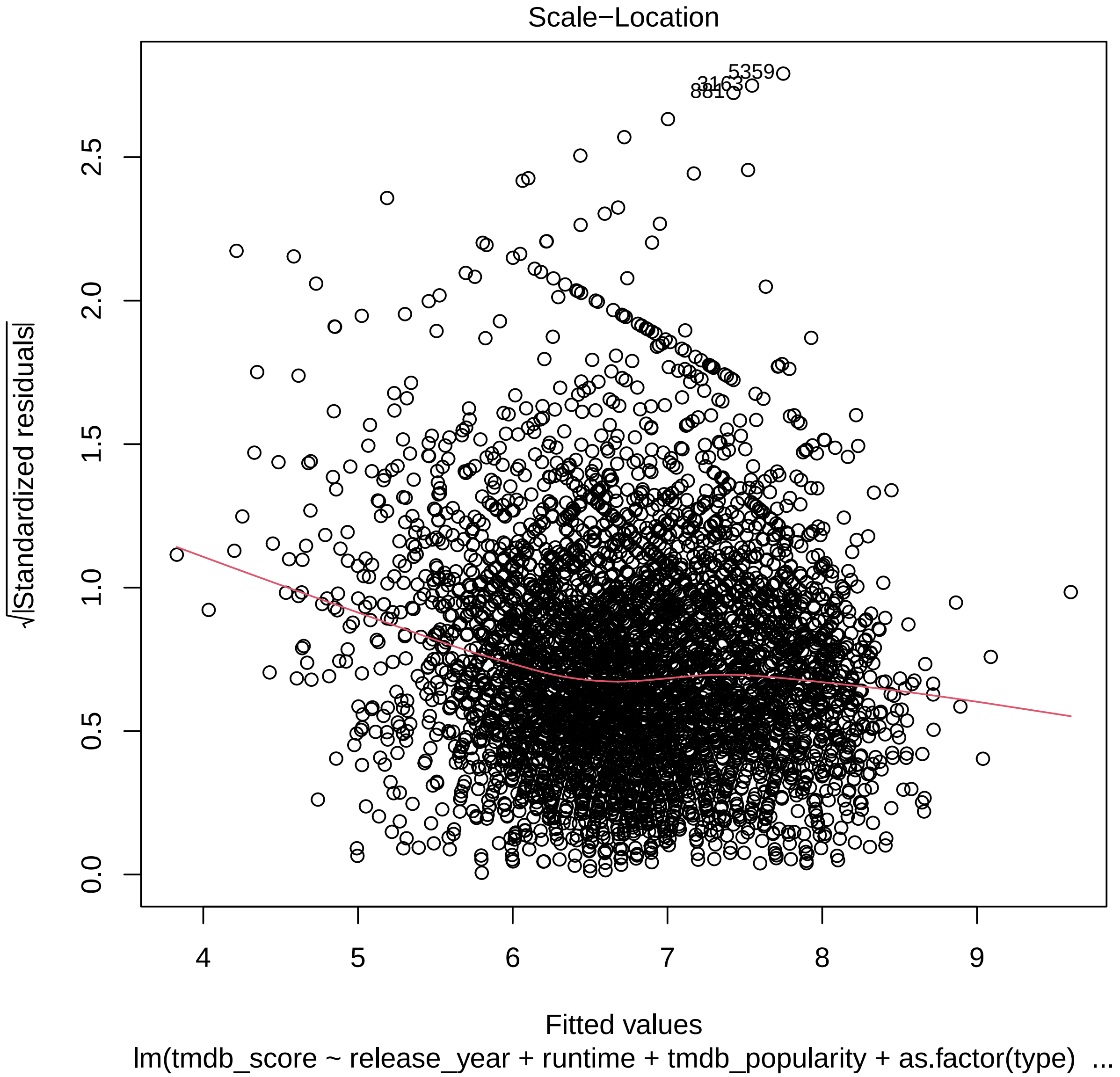
x_{i6} = IMDb votes of the i^{th} movie/show.

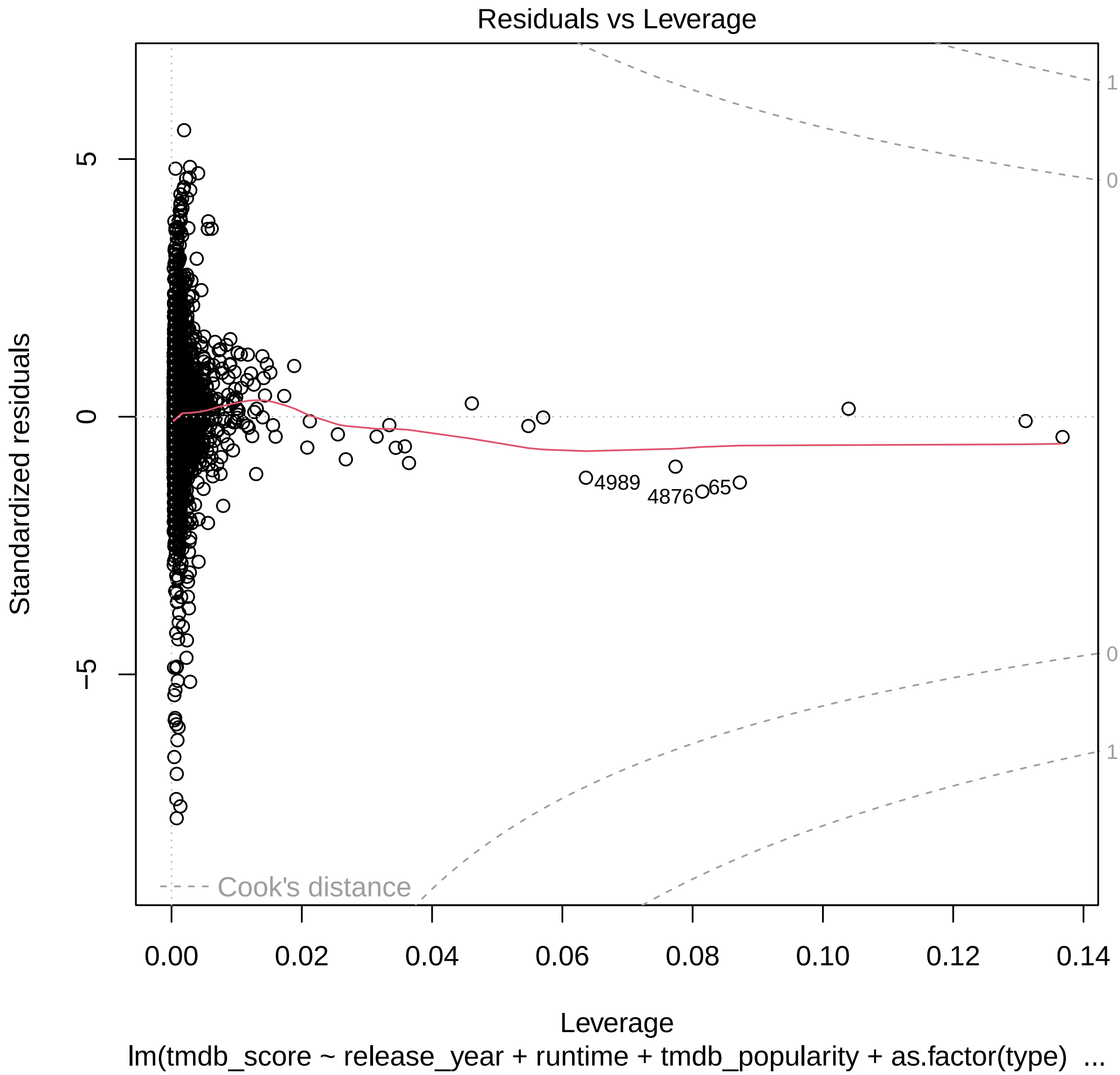
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-20.6509	3.4905	-5.92	0.0000
release_year	0.0119	0.0017	6.91	0.0000
runtime	-0.0011	0.0005	-2.19	0.0284
tmdb_popularity	0.0006	0.0002	3.41	0.0007
as.factor(type)SHOW	0.5674	0.0427	13.28	0.0000
imdb_score	0.5049	0.0116	43.49	0.0000
imdb_votes	0.0000	0.0000	3.20	0.0014

We plot some graphs, Q-Q plot to understand the model better.




Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
Theoretical Quantiles
lm(tmdb_score ~ release_year + runtime + tmdb_popularity + as.factor(type) ...
881
3163
5359


Scale−Location
5359
3163
881
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
4
5
6
7
8
9
Fitted values
lm(tmdb_score ~ release_year + runtime + tmdb_popularity + as.factor(type) ...



After fitting the model we see that all the covariates are significant to the model.

Conclusions and Discussions:

Based on both exploratory and inferential data analysis, in the light of given data we conclude that,

1. People enjoy movies/shows produced in US more than movies/shows produced in Japan. This confirms the recent trend of watching japanese media all around the world.

2. People enjoy watching TV shows more than watching movies.

3. People enjoy watching short duration movies/shows rather than long duration movies/shows.

4. People enjoy watching movies/shows of drama genre rather than of comedy genre.

5. On IMDb, people like older movies/shows more than newer movies/shows. On the contrary on TMDB, people like newer movies/shows more than older movies/shows.

6. The IMDb rating of a movie/show depend on the **Release Year**, **Runtime**, **Type**, **Number of IMDb Votes**, **TMDB score** and **TMDB popularity** of that movie/show.

7. The TMDB rating of a movie/show depend on the **Release Year**, **Runtime**, **Type**, **Number of IMDb votes**, **IMDb score** and **Number of IMDb Votes** of that movie/show.

Appendix:

R Codes:

```r
data=read.csv("C:\\Users\\LENOVO\\Downloads\\netflix.csv")
dim(data)
names(data)
summary(netflix)
str(netflix)
par(mfrow=c(2,3))
boxplot(netflix$release_year,main="Release Year")
boxplot(netflix$runtime,main="Runtime")
boxplot(netflix$imdb_score,main="IMDB Score")
boxplot(netflix$imdb_votes,main="IMDB Votes")
boxplot(netflix$tmdb_popularity,main="TMDB Popularity")
boxplot(netflix$tmdb_score,main="TMDB Score")
par(mfrow=c(2,2))
plot(netflix$runtime,netflix$imdb_score,main="Runtime Vs. IMDB Score",xlab="Runtime",yla
plot(netflix$imdb_score,netflix$imdb_votes,main="IMDB Votes Vs. IMDB Scores",xlab="IMDB
plot(netflix$runtime,netflix$tmdb_score,main="Runtime VS. TMDB Score",xlab="Runtime",yla
plot(netflix$tmdb_score,netflix$tmdb_popularity,main="TMDB Score Vs. TMDB Popularity",xl
par(mfrow=c(2,3))
hist(netflix$release_year,main="Histogram of Release Date",xlab="Release Date")
hist(netflix$runtime,main="Histogram of Runtime",xlab="Runtime")
hist(netflix$imdb_score,main="Histogram of IMDB Score",xlab="IMDB Score")
hist(netflix$imdb_votes,main="Histogram of IMDB Votes",xlab="IMDB Votes")
hist(netflix$tmdb_score,main="Histogram of TMDB Score",xlab="TMDB Score")
hist(netflix$tmdb_popularity,main="Histogram of TMDB Popularity",xlab="TMDB Popularity")
max(netflix$imdb_score)
netflix$title[which.max(netflix$imdb_score)]
netflix$type[which.max(netflix$imdb_score)]
netflix$description[which.max(netflix$imdb_score)]
netflix$release_year[which.max(netflix$imdb_score)]
netflix$age_certification[which.max(netflix$imdb_score)]
netflix$runtime[which.max(netflix$imdb_score)]
netflix$genres[which.max(netflix$imdb_score)]
netflix$production_countries[which.max(netflix$imdb_score)]
netflix$imdb_votes[which.max(netflix$imdb_score)]
```

```r
min(netflix$imdb_score)
netflix$title[which.min(netflix$imdb_score)]
netflix$type[which.min(netflix$imdb_score)]
netflix$description[which.min(netflix$imdb_score)]
netflix$release_year[which.min(netflix$imdb_score)]
netflix$age_certification[which.min(netflix$imdb_score)]
netflix$runtime[which.min(netflix$imdb_score)]
netflix$genres[which.min(netflix$imdb_score)]
netflix$production_countries[which.min(netflix$imdb_score)]
netflix$imdb_votes[which.min(netflix$imdb_score)]
max(netflix$tmdb_score)
netflix$title[which.max(netflix$tmdb_score)]
netflix$type[which.max(netflix$tmdb_score)]
netflix$description[which.max(netflix$tmdb_score)]
netflix$release_year[which.max(netflix$imdb_score)]
netflix$age_certification[which.max(netflix$tmdb_score)]
netflix$runtime[which.max(netflix$tmdb_score)]
netflix$genres[which.max(netflix$tmdb_score)]
netflix$production_countries[which.max(netflix$tmdb_score)]
netflix$tmdb_popularity[which.max(netflix$tmdb_score)]
min(netflix$tmdb_score)
netflix$title[which.min(netflix$tmdb_score)]
netflix$type[which.min(netflix$tmdb_score)]
netflix$description[which.min(netflix$tmdb_score)]
netflix$release_year[which.min(netflix$imdb_score)]
netflix$age_certification[which.min(netflix$tmdb_score)]
netflix$runtime[which.min(netflix$tmdb_score)]
netflix$genres[which.min(netflix$tmdb_score)]
netflix$production_countries[which.min(netflix$tmdb_score)]
netflix$tmdb_popularity[which.min(netflix$tmdb_score)]
netflix$tmdb_score[which(netflix$title=="Breaking Bad")]
netflix$imdb_score[which(netflix$title=="Pink Zone")]
US=subset(netflix,netflix$production_countries=="['US']")
JP=subset(netflix,netflix$production_countries=="['JP']")
dim(US)
dim(JP)
par(mfrow=c(1,2))
boxplot(US$imdb_score,main="US IMDB Score",xlab="US",ylim=c(0,10))
boxplot(JP$imdb_score,main="JP IMDB Score",xlab="JP",ylim=c(0,10))
var.test(US$imdb_score,JP$imdb_score,alternative = "two.sided")
var.test(US$tmdb_score,JP$tmdb_score,alternative = "two.sided")
t.test(US$imdb_score,JP$imdb_score,var.equal=FALSE,alt="less")
t.test(US$tmdb_score,JP$tmdb_score,var.equal=TRUE,alt="less")
movie=subset(netflix,netflix$type=="MOVIE")
show=subset(netflix,netflix$type=="SHOW")
dim(movie)
dim(show)
par(mfrow=c(1,2))
```

```r
boxplot(movie$imdb_score,main="Movie IMDB Score",xlab="Movie",ylim=c(0,10))
boxplot(show$imdb_score,main="Show IMDB Score",xlab="Show",ylim=c(0,10))
var.test(movie$imdb_score,show$imdb_score,alternative = "two.sided")
var.test(movie$tmdb_score,show$tmdb_score,alternative = "two.sided")
t.test(movie$imdb_score,show$imdb_score,var.equal=TRUE,alt="less")
t.test(movie$tmdb_score,show$tmdb_score,var.equal=TRUE,alt="less")
short=subset(netflix,netflix$runtime<=80)
long=subset(netflix,netflix$runtime>80)
dim(short)
dim(long)
par(mfrow=c(1,2))
boxplot(short$imdb_score,main="Short IMDB Score",xlab="Short",ylim=c(0,10))
boxplot(long$imdb_score,main="Long IMDB Score",xlab="Long",ylim=c(0,10))
var.test(short$imdb_score,long$imdb_score,alternative = "two.sided")
var.test(short$tmdb_score,long$tmdb_score,alternative = "two.sided")
t.test(short$imdb_score,long$imdb_score,var.equal=TRUE,alt="greater")
t.test(short$tmdb_score,long$tmdb_score,var.equal=FALSE,alt="greater")
comedy=subset(netflix,netflix$genres=="['comedy']")
drama=subset(netflix,netflix$genres=="['drama']")
dim(comedy)
dim(drama)
par(mfrow=c(1,2))
boxplot(comedy$imdb_score,main="Comedy IMDB Score",xlab="Comedy",ylim=c(0,10))
boxplot(drama$imdb_score,main="Drama IMDB Score",xlab="Drama",ylim=c(0,10))
var.test(comedy$imdb_score,drama$imdb_score,alternative = "two.sided")
var.test(comedy$tmdb_score,drama$tmdb_score,alternative = "two.sided")
t.test(comedy$imdb_score,drama$imdb_score,var.equal=FALSE,alt="less")
t.test(comedy$tmdb_score,drama$tmdb_score,var.equal=FALSE,alt="less")
old=subset(netflix,netflix$release_year<=2016)
new=subset(netflix,netflix$release_year>2016)
dim(old)
dim(new)
par(mfrow=c(1,2))
boxplot(old$imdb_score,main="Old IMDB Score",xlab="Old",ylim=c(0,10))
boxplot(new$imdb_score,main="New IMDB Score",xlab="New",ylim=c(0,10))
var.test(old$imdb_score,new$imdb_score,alternative = "two.sided")
var.test(old$tmdb_score,new$tmdb_score,alternative = "two.sided")
t.test(old$imdb_score,new$imdb_score,var.equal=TRUE,alt="greater")
t.test(old$tmdb_score,new$tmdb_score,var.equal=FALSE,alt="less")
model1=lm(imdb_score~release_year+runtime+imdb_votes+as.factor(type)+tmdb_score+tmdb_pop
summary(model1)
model2=lm(tmdb_score~release_year+runtime+tmdb_popularity+as.factor(type)+imdb_score+imd
summary(model2)
cov1=data.frame(netflix$release_year,netflix$runtime,netflix$imdb_votes,netflix$tmdb_sco
y1=netflix$imdb_score
library(lars)
mod.lasso1=lars(as.matrix(cov1),y1,normalize=F) plot(mod.lasso1,xvar="df") #--plotting-c
coef(mod.lasso1) #--coeffs-for-each-step-in-path
```

```
cv.lasso1=cv.lars(as.matrix(cov1),y1,type="lasso") #--cross-validation
limit1=min(cv.lasso1$cv)+cv.lasso1$cv.error[which.min(cv.lasso1$cv)]
s.cv1=cv.lasso1$index[min(which(cv.lasso1$cv<limit1))]
coef(mod.lasso1,s=s.cv1,mode="fraction")
cov2=data.frame(netflix$release_year,netflix$runtime,netflix$imdb_votes,netflix$imdb_sco
y2=netflix$tmdb_score
library(lars)
mod.lasso2=lars(as.matrix(cov2),y2,normalize=F)
plot(mod.lasso2,xvar="df") #--plotting-coef-paths--
coef(mod.lasso2) #--coeffs-for-each-step-in-path
cv.lasso2=cv.lars(as.matrix(cov2),y2,type="lasso") #--cross-validation
limit2=min(cv.lasso2$cv)+cv.lasso2$cv.error[which.min(cv.lasso2$cv)]
s.cv2=cv.lasso2$index[min(which(cv.lasso2$cv<limit2))]
coef(mod.lasso2,s=s.cv2,mode="fraction")
```

Bibliography:

1. To download the CSV file of the dataset, Click Here

2. Information about Netflix: click here

3. Information about IMDb: click here

4. Information about TMDB: click here