

IMDb Data Analysis

This document is designed to use data analysis to understand the movie ratings and how different service providers differ in the quality of movies on their platforms

Loading the Packages

```
library(tidyverse)
```

```
## -- Attaching packages -----
```

```
## v ggplot2 3.3.0      v purrr  0.3.4
## v tibble  3.0.1      v dplyr  0.8.5
## v tidyr   1.0.3      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.5.0
```

```
## -- Conflicts -----
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

Loading the file

```
df_imdb <- read.csv("~/IMDb/Movies.csv")
```

Checking out the loaded file

```
head(df_imdb,5)
```

```
##           Title Year Age IMDb Rotten.Tomatoes Netflix Hulu
## 1      Inception 2010 13+  8.8           87%         1     0
## 2      The Matrix 1999 18+  8.7           87%         1     0
## 3 Avengers: Infinity War 2018 13+  8.5           84%         1     0
## 4      Back to the Future 1985  7+  8.5           96%         1     0
## 5 The Good, the Bad and the Ugly 1966 18+  8.8           97%         1     0
## Prime.Video Disney. Type                      Directors
## 1           0         0     0           Christopher Nolan
## 2           0         0     0 Lana Wachowski,Lilly Wachowski
## 3           0         0     0           Anthony Russo,Joe Russo
```

```
## 4      0      0      0      Robert Zemeckis
## 5      1      0      0      Sergio Leone
##                               Genres                               Country
## 1 Action,Adventure,Sci-Fi,Thriller United States,United Kingdom
## 2                               Action,Sci-Fi                               United States
## 3      Action,Adventure,Sci-Fi                               United States
## 4      Adventure,Comedy,Sci-Fi                               United States
## 5                               Western      Italy,Spain,West Germany
##                               Language Runtime
## 1 English,Japanese,French      148
## 2                               English      136
## 3                               English      149
## 4                               English      116
## 5                               Italian      161
```

```
colnames(df_imdb)
```

```
## [1] "Title"      "Year"      "Age"      "IMDb"
## [5] "Rotten.Tomatoes" "Netflix"   "Hulu"     "Prime.Video"
## [9] "Disney."     "Type"     "Directors" "Genres"
## [13] "Country"    "Language"  "Runtime"
```

Refining the dataset

```
df_imdb <- df_imdb %>%
  rename(c("Prime" = "Prime.Video", "Disney" = "Disney."))
df_imdb$Rotten.Tomatoes <- str_replace_all(df_imdb$Rotten.Tomatoes, "%", "")
df_imdb$Rotten.Tomatoes <- as.integer(df_imdb$Rotten.Tomatoes)
sum(is.na(df_imdb$Rotten.Tomatoes))
```

```
## [1] 11586
```

Rechecking dataframe

```
head(df_imdb, 5)
```

```
##                               Title Year Age IMDb Rotten.Tomatoes Netflix Hulu
## 1              Inception 2010 13+ 8.8      87      1      0
## 2              The Matrix 1999 18+ 8.7      87      1      0
## 3  Avengers: Infinity War 2018 13+ 8.5      84      1      0
## 4      Back to the Future 1985 7+ 8.5      96      1      0
## 5 The Good, the Bad and the Ugly 1966 18+ 8.8      97      1      0
##   Prime Disney Type                               Directors
## 1      0      0      0      Christopher Nolan
## 2      0      0      0 Lana Wachowski,Lilly Wachowski
## 3      0      0      0      Anthony Russo,Joe Russo
## 4      0      0      0      Robert Zemeckis
## 5      1      0      0      Sergio Leone
##                               Genres                               Country
```

```
## 1 Action,Adventure,Sci-Fi,Thriller United States,United Kingdom
## 2          Action,Sci-Fi          United States
## 3          Action,Adventure,Sci-Fi          United States
## 4          Adventure,Comedy,Sci-Fi          United States
## 5          Western          Italy,Spain,West Germany
##          Language Runtime
## 1 English,Japanese,French      148
## 2          English      136
## 3          English      149
## 4          English      116
## 5          Italian      161
```

Creating new data frame for data exploration

```
df_imdb1 <- df_imdb %>%
  select(Title, Year, IMDb, Rotten.Tomatoes, Netflix, Hulu, Prime, Disney) %>%
  gather(Netflix:Disney, key = "provider", value = "has_movie")
```

Checking new dataframe

```
head(df_imdb1,5)
```

```
##          Title Year IMDb Rotten.Tomatoes provider has_movie
## 1          Inception 2010 8.8          87 Netflix      1
## 2          The Matrix 1999 8.7          87 Netflix      1
## 3    Avengers: Infinity War 2018 8.5          84 Netflix      1
## 4      Back to the Future 1985 8.5          96 Netflix      1
## 5 The Good, the Bad and the Ugly 1966 8.8          97 Netflix      1
```

```
sprintf("Regular dataset has: %d", nrow(df_imdb))
```

```
## [1] "Regular dataset has: 16744"
```

```
sprintf("Gathered dataset has: %d", nrow(df_imdb1))
```

```
## [1] "Gathered dataset has: 66976"
```

Understading the data better

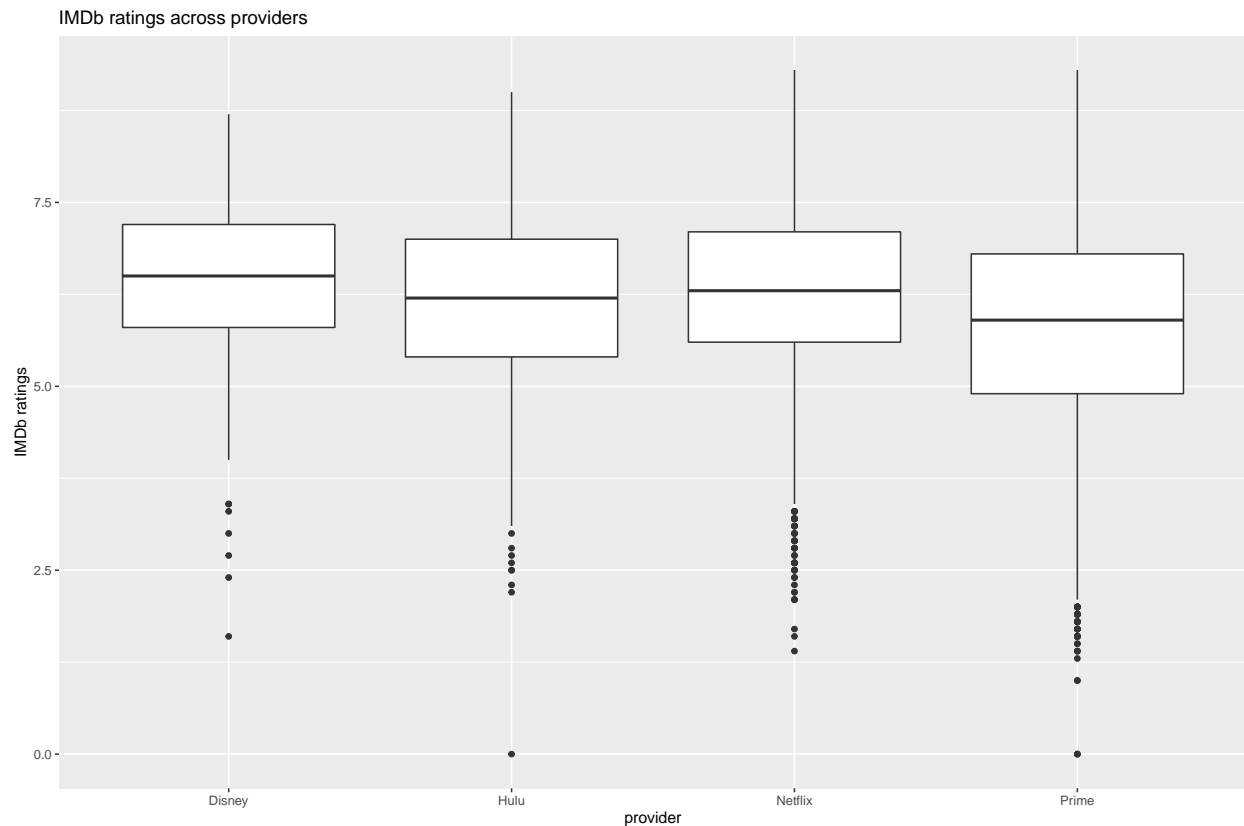
```
df_imdb1 <- df_imdb1 %>%
  filter(has_movie == 1)
nrow(df_imdb1)
```

```
## [1] 17381
```

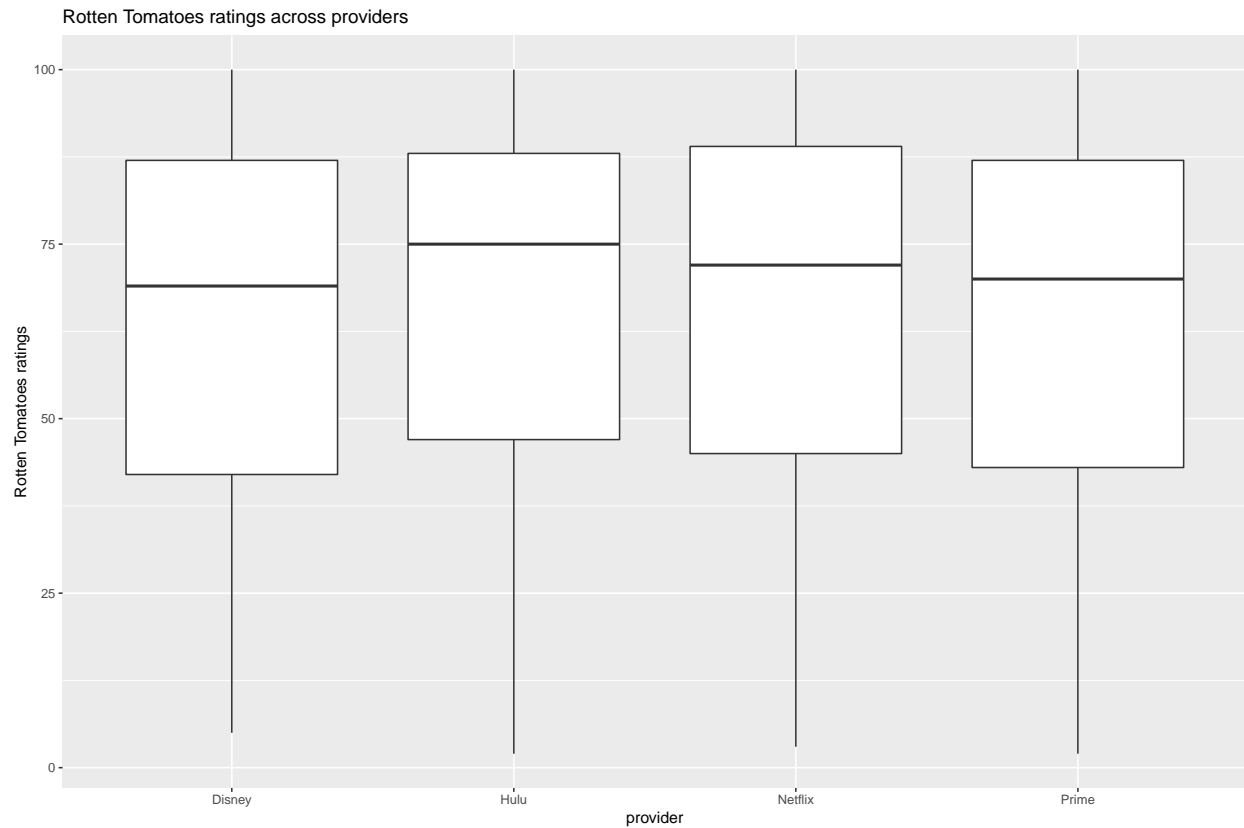
```
write.csv(df_imdb1, "~/IMDb/Movies2.csv")
```

Boxplot graph

```
df_imdb1 %>%  
  ggplot() + geom_boxplot(aes(x = provider, y = IMDb), na.rm = TRUE) +  
  labs(x = "provider", y = "IMDb ratings", title = "IMDb ratings across providers")
```

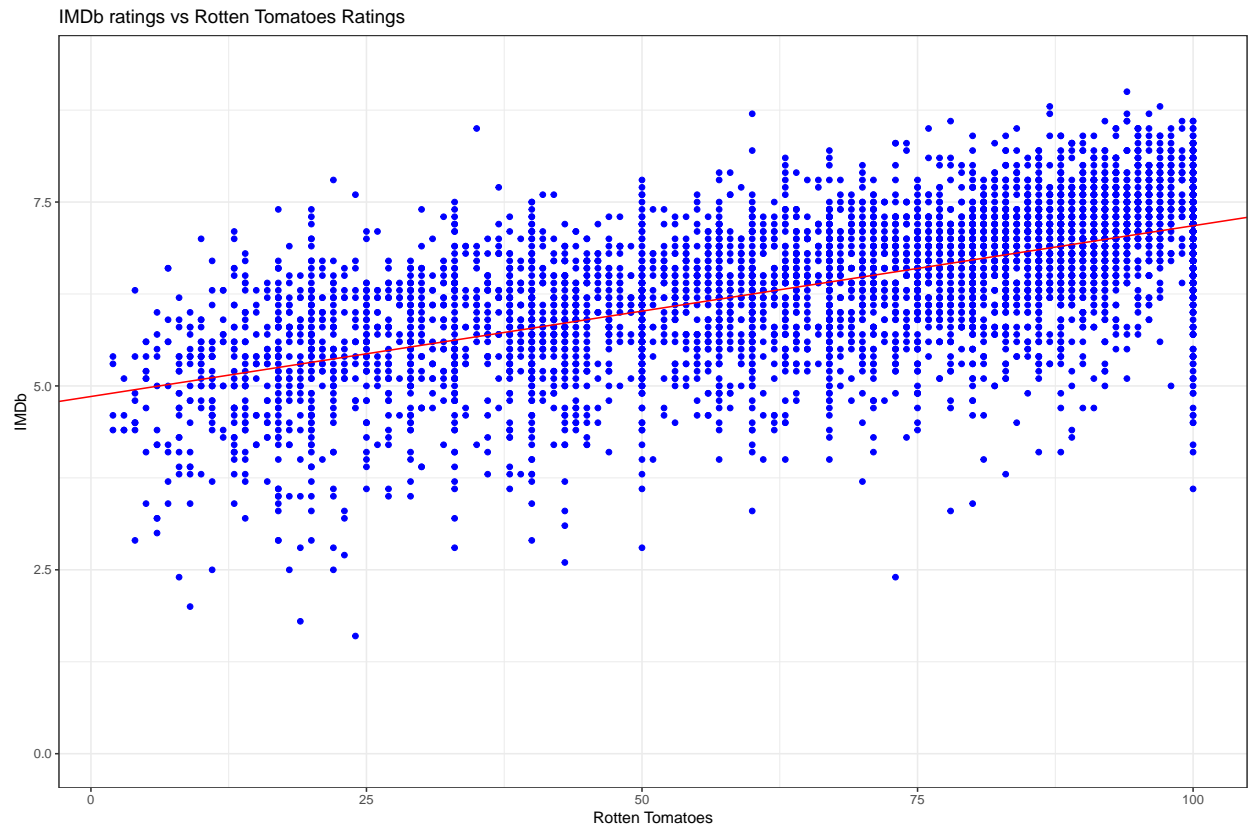


```
df_imdb1 %>%  
  ggplot() +  
  geom_boxplot(aes(x = provider, y = Rotten.Tomatoes), na.rm = TRUE) +  
  labs(x = "provider", y = "Rotten Tomatoes ratings") +  
  ggtitle("Rotten Tomatoes ratings across providers")
```



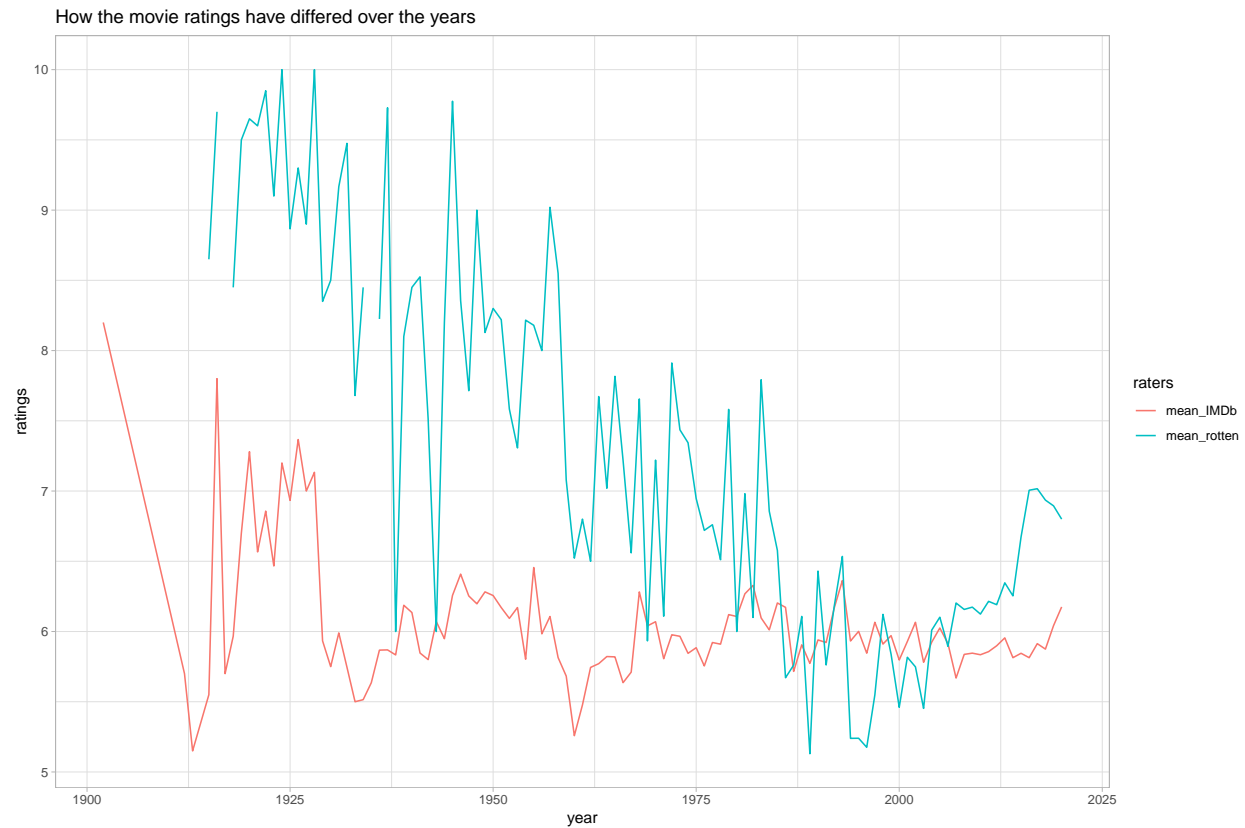
Graph of IMDb vs Rotten Tomatoes

```
model <- lm(IMDb ~ Rotten.Tomatoes, data = df_imdb)
coeff <- coef(model)
ggplot(df_imdb) +
  geom_point(aes(x = Rotten.Tomatoes, y = IMDb), color = "blue", na.rm = TRUE) +
  geom_abline(intercept = coeff[1], slope = coeff[2], color = "red") +
  theme_bw() +
  labs(y = "IMDb", x = "Rotten Tomatoes", title = "IMDb ratings vs Rotten Tomatoes Ratings")
```



```
df_imdb2 <- df_imdb %>%
  mutate(new_rotten = as.double(Rotten.Tomatoes/10)) %>%
  group_by(Year) %>%
  summarize(mean_IMDb = mean(IMDb, na.rm = TRUE), mean_rotten = mean(new_rotten, na.rm=TRUE))

df_imdb2 %>%
  gather(mean_IMDb:mean_rotten, key = "raters", value = "ratings") %>%
  ggplot() +
  geom_line(aes(x = Year, y = ratings, color = raters)) +
  labs(x = "year", y = "ratings", title = "How the movie ratings have differed over the years") +
  theme_light()
```



```
ggplot(df_imdb2) +
  geom_line(aes(x = Year, y = mean_IMDb, color = "IMDb")) +
  geom_line(aes(x = Year, y = mean_rotten, color = "Rotten Tomatoes")) +
  labs(x = "year", y = "ratings", title = "How the movie ratings have differed over the years") +
  scale_colour_manual("", breaks = c("IMDb", "Rotten Tomatoes"), values = c("IMDb"="green",
                                                                              "Rotten Tomatoes"="red"))+
  theme_bw()
```

