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

### Loading the file

```
df_imdb <- read.csv("~/IMDb/Movies.csv", encoding = "UTF-8")
```

### 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

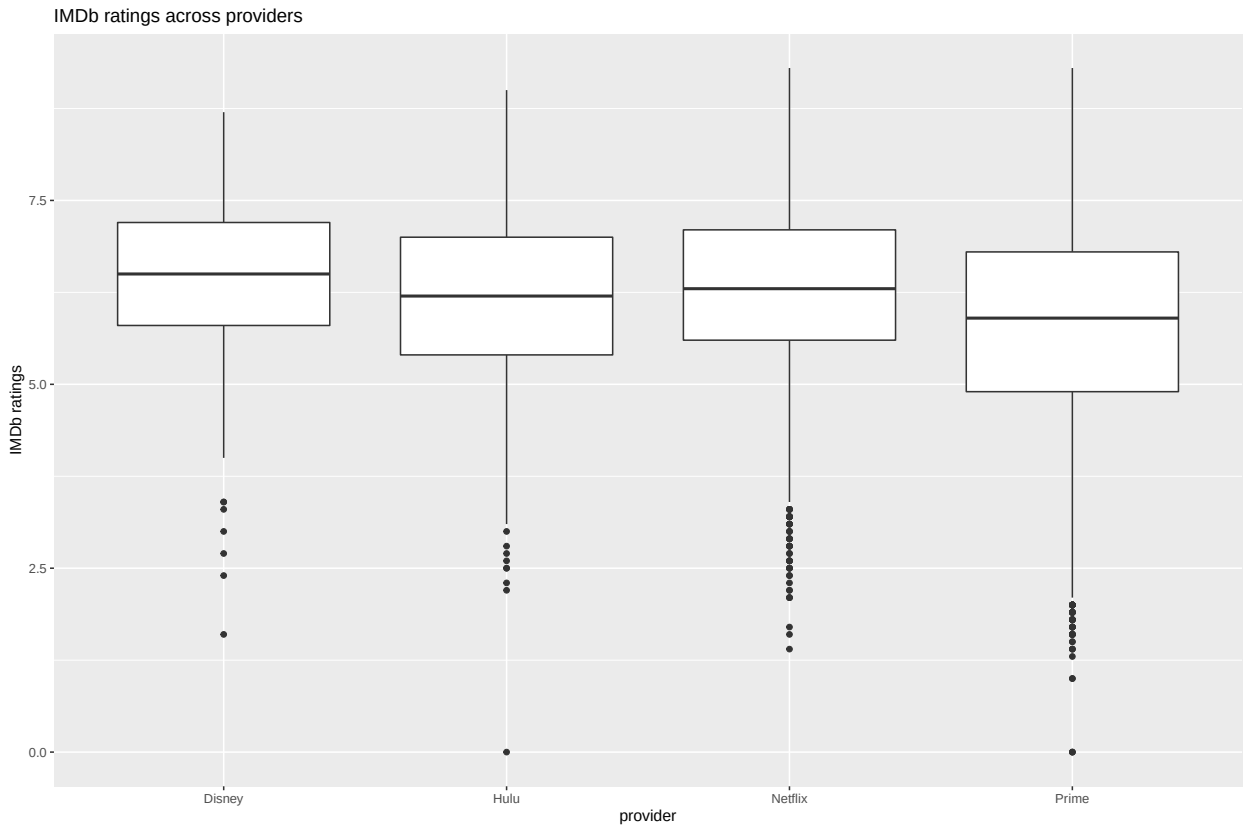
```
df_imdb1 <- df_imdb1 %>%
  filter(has_movie == 1)
nrow(df_imdb1)
```

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

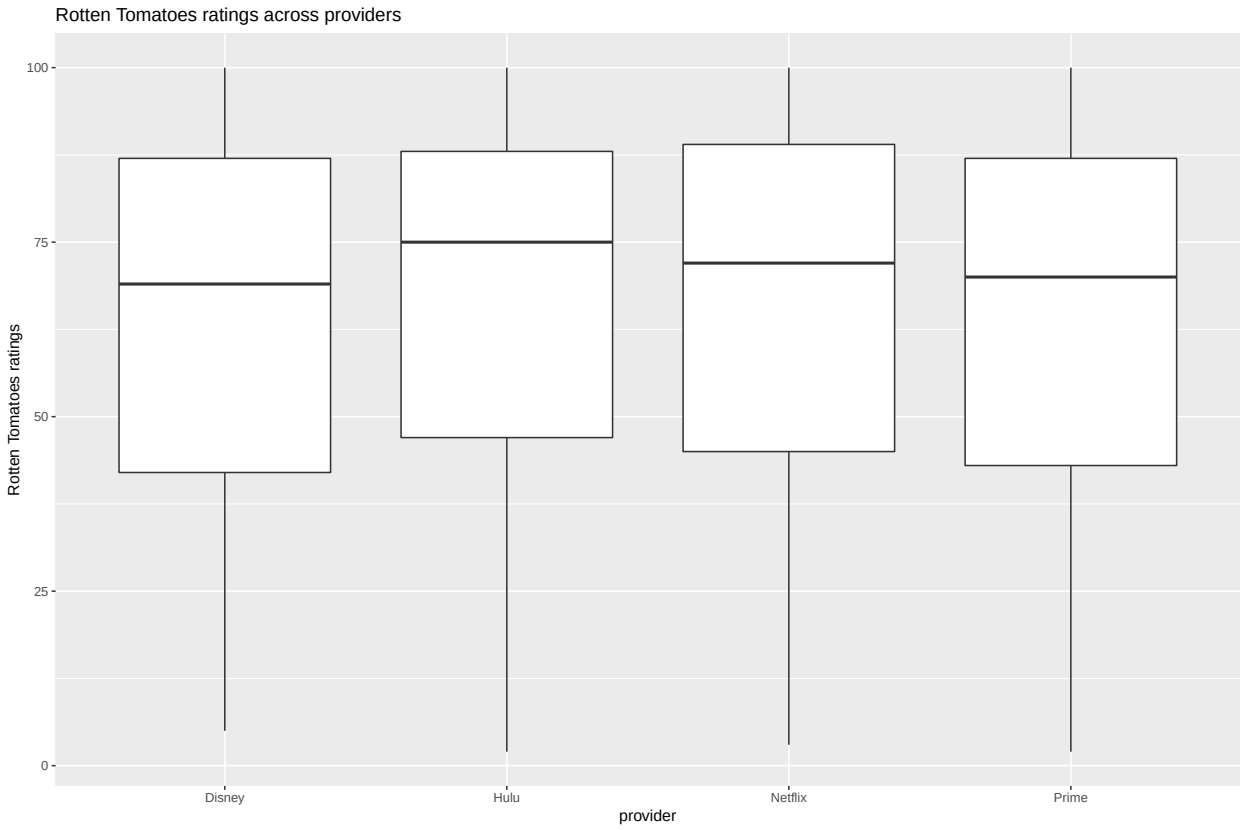
```
write.csv(df_imdb1, "~/IMDb/Movies2.csv")
```

### 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")
```

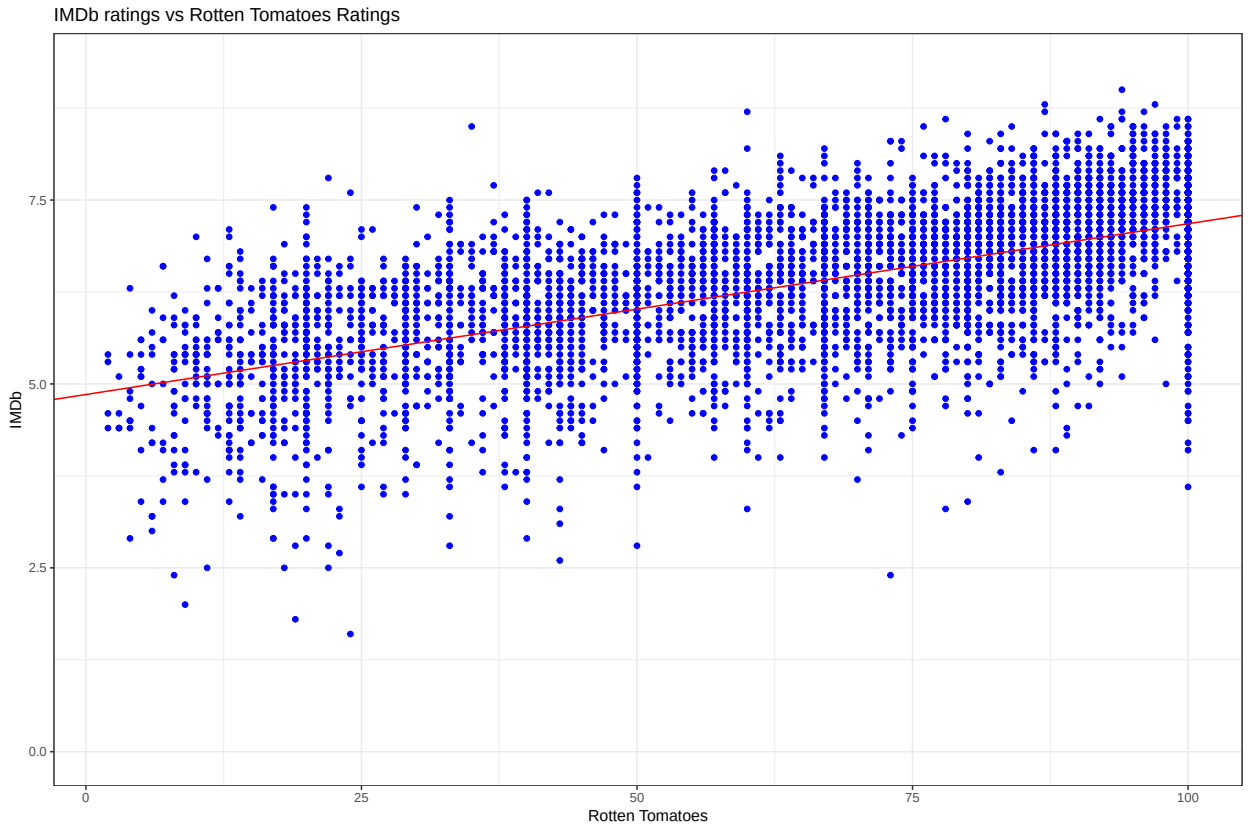



```
df_imdb1 %>%
  ggplot() +
  geom_boxplot(aes(x = provider, y = Rotten.Tomatoes), na.rm = TRUE) +
  labs(x = "provider", y = "Rotten Tomatoes ratings") +
  ggtitle("Rotten Tomatoes ratings across providers")
```



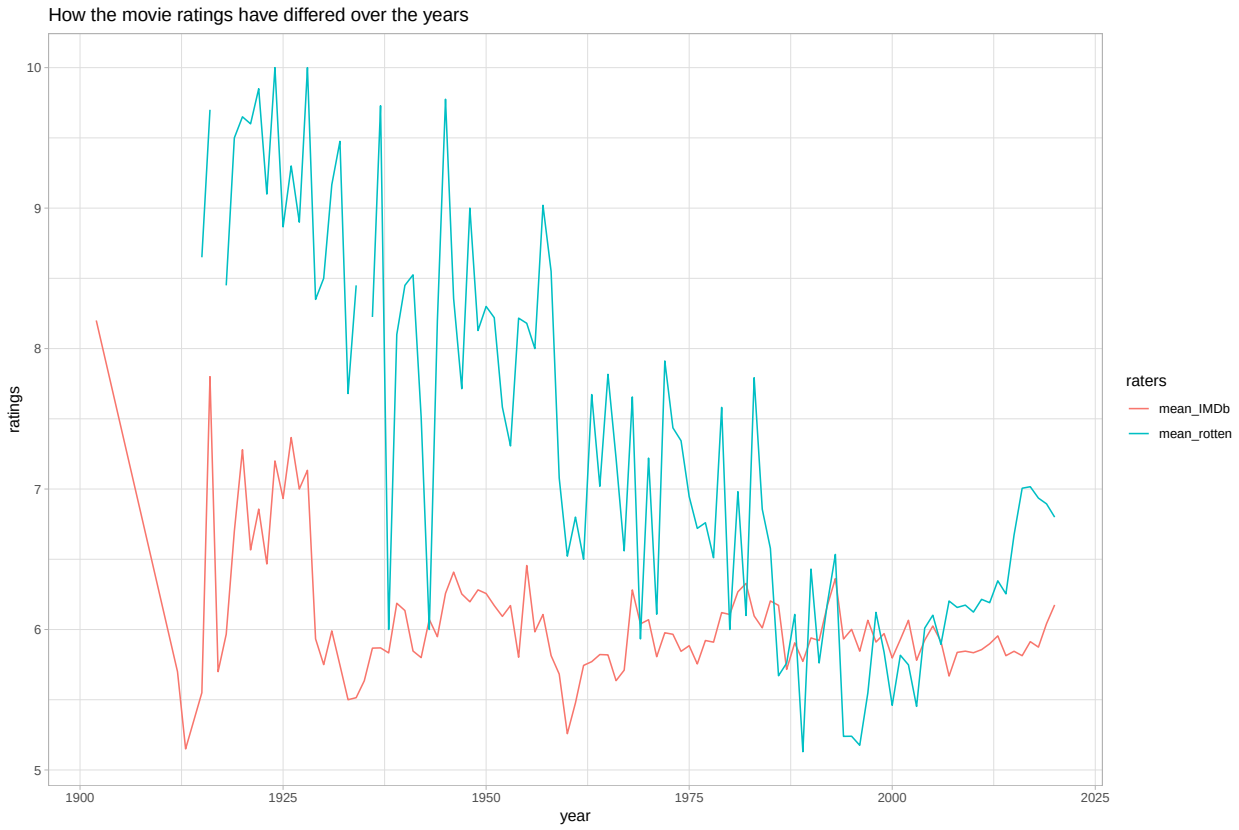

### Graph of IMDb vs Rotten Tomatoes

```
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")
```

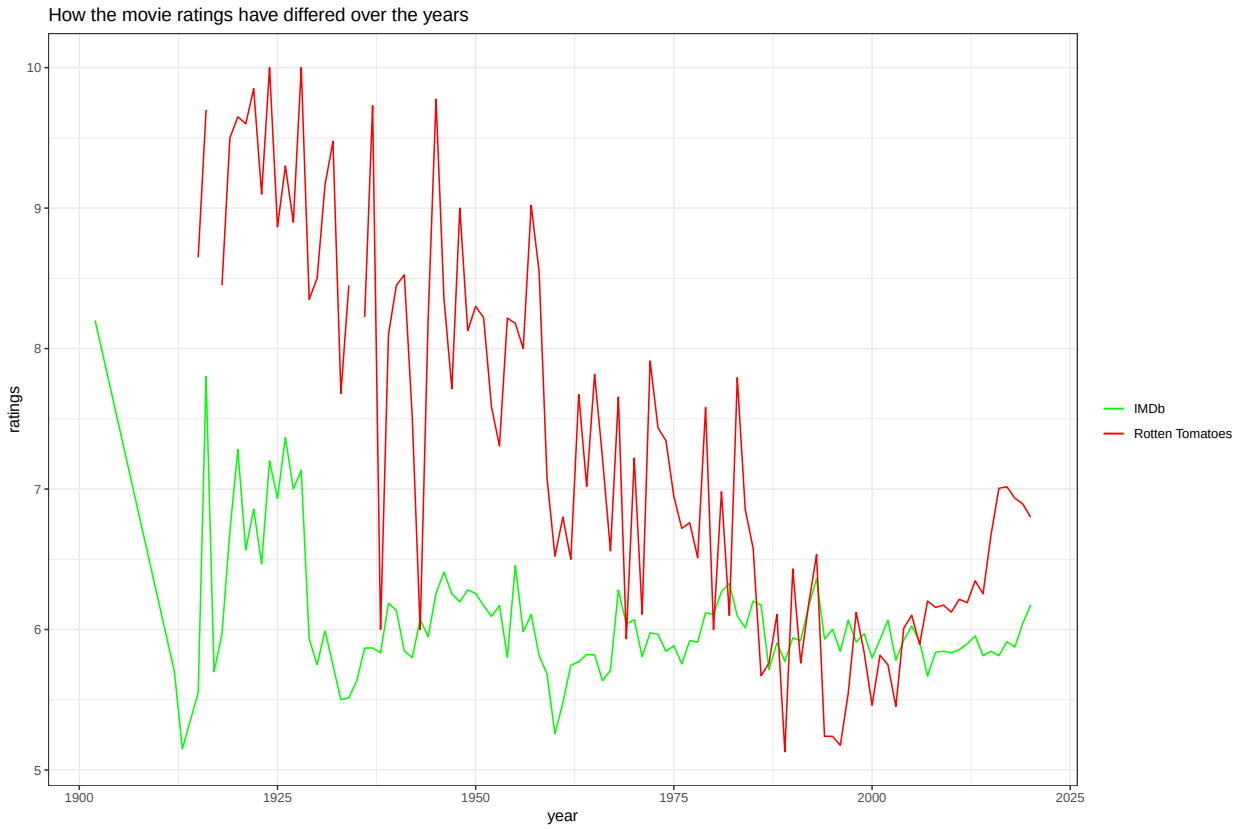



```
df_imdb2 <- df_imdb %>%
  mutate(new_rotten = as.double(Rotten.Tomatoes/10)) %>%
  group_by(Year) %>%
  summarize(mean_IMDb = mean(IMDb, na.rm = TRUE), mean_rotten = mean(new_rotten, na.rm=TRUE))
```

```
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()
```




## The Director with the best ratings

```
df_imdb3 <- df_imdb %>%
  select(Title, Directors, IMDb, Rotten.Tomatoes) %>%
  separate(Directors, into = c("Dir1", "Dir2", "Dir3"), sep = ",") %>%
  gather(Dir1:Dir3, key = "dir", value = "directors")
```

```
## Warning: Expected 3 pieces. Additional pieces discarded in 87 rows [365, 670,
## 1061, 1258, 1276, 1727, 1801, 1889, 1913, 2255, 2289, 2404, 2561, 2976, 3048,
## 3051, 3193, 3292, 3809, 4023, ...].
```

```
## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 16507 rows [1, 2,
## 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, ...].
```

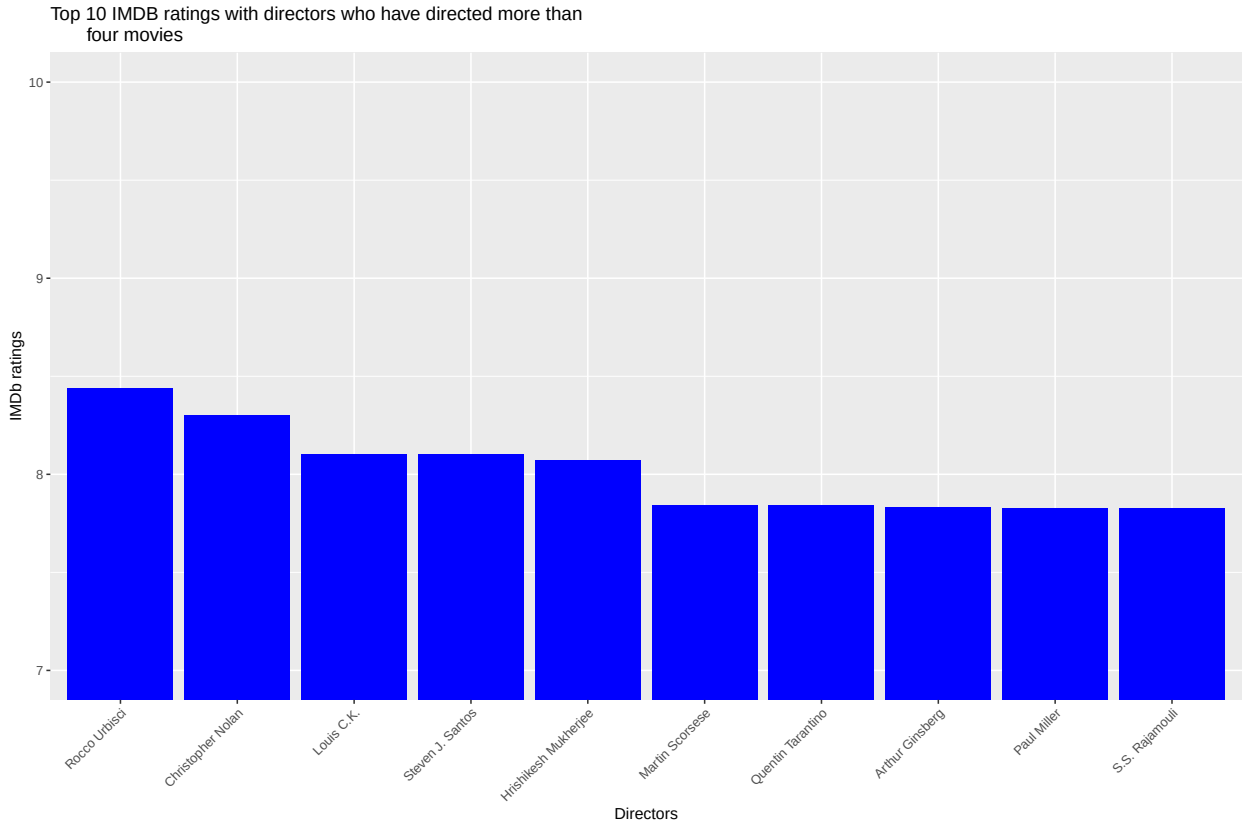
### Checking out new dataset

```
df_imdb4 <- df_imdb3 %>%
  filter(directors != "") %>%
  group_by(directors) %>%
  summarize(imdb = mean(IMDb, na.rm = TRUE), rotten = mean(Rotten.Tomatoes, na.rm = TRUE), count = n())
nrow(df_imdb4)
```

```
## [1] 12242
```

```
df_imdb4 <- df_imdb4 %>%
  mutate(rank = rank(desc(count), ties.method = "first")) %>%
  arrange(rank)
```

```
df_imdb4 %>%
  filter(count >= 4) %>%
  mutate(new_rank = rank(desc(imdb))) %>%
  arrange(new_rank) %>%
  filter(new_rank <= 10) %>%
  ggplot() +
  geom_bar(aes(x = reorder(directors, -imdb), y = imdb), fill = "blue", stat = "identity") +
  coord_cartesian(ylim = c(7, 10)) +
  labs(x = "Directors", y = "IMDb ratings", title = "Top 10 IMDB ratings with directors who have directe
       four movies") +
  theme(axis.text.x=element_text(angle=45, hjust=1))
```




```
df_imdb5 <- df_imdb %>%
  separate(Genres, into = c("gen1", "gen2", "gen3", "gen4", "gen5","gen6", "gen7"), sep = ",") %>%
  gather(gen1:gen7, key = "gen_num", value = "genre") %>%
  filter(genre != "NA") %>%
  select(Title,genre, IMDb, Rotten.Tomatoes)
nrow(df_imdb5)
```

```
## [1] 39354
```

```r
df_imdb6 <- df_imdb5 %>%
  group_by(genre) %>%
  summarize(imdb = mean(IMDb, na.rm = TRUE),rotten = mean(Rotten.Tomatoes, na.rm = TRUE), count = n()) %
  mutate(rank = rank(desc(count), ties.method = "first")) %>%
  arrange(rank)
```

```r
df_imdb6 %>%
  mutate(new_rank = rank(desc(imdb))) %>%
  filter(new_rank <= 10) %>%
  ggplot() +
  geom_bar(aes(x = reorder(genre, -imdb), y = imdb), fill = "purple", stat = "identity") +
  coord_cartesian(ylim = c(5, 9)) +
  labs(x = "Genres", y = "IMDb ratings", title = "Top 10 Genres with the highest IMDb ratings") +
  theme(axis.text.x=element_text(angle=45, hjust=1))
```