Final Project: IMDb Analysis

Charlotte Matthews, Jasmine Jordan-Lake

2024-05-06

Abstract

Within the film industry's dynamic landscape, where audience preferences and trends continually shift, understanding the relationship between critical acclaim, such as IMDb ratings, and commercial success, as measured by box office earnings, and how viewer inclinations influence these is vital. In this study, we have presented findings derived from a comprehensive analysis of IMDb data, showing key trends in the film industry. "IMDb ('Internet Movie Database') is the world's most popular and authoritative source for movie, TV and celebrity content. Find ratings and reviews for the newest movie and TV shows (IMDb 2024)." Over the years, we observed a significant decrease in average IMDb ratings, accompanied by a convergence towards a narrower range of scores. Additionally, average film runtimes have also converged. Our analysis further reveals a general positive relationship between IMDb ratings and box office earnings, as well as between film runtime and IMDb ratings. Overall, this may suggest that evolving audience preferences and changes in filmmaking practices may be influencing these trends. These findings offer valuable insights for stakeholders, suggesting optimized runtime decisions aligned with audience preferences, effective marketing strategies leveraging IMDb ratings and box office correlations, and further exploration of industry dynamics and trends.

Introduction

Dataset Overview:

During this analysis, we utilized an IMDb Dataset from data.world to examine and visualize patterns, correlations, and potential trends. This dataset consists of 5273 observations and 27 features. We were particularly interested in this dataset because of our interest in film industry and the dataset's coverage of IMDb ratings, box office earnings, film runtimes, and classifications, allowing us to explore various areas of film reception and performance. The main question we hope to discuss is: "How have audience preferences and film industry norms changed over time? What impacts have these shifts had, notably on IMDb ratings and Box Office earnings?" Our comprehensive analysis explored these objectives to help promote our discussion:

Objectives:

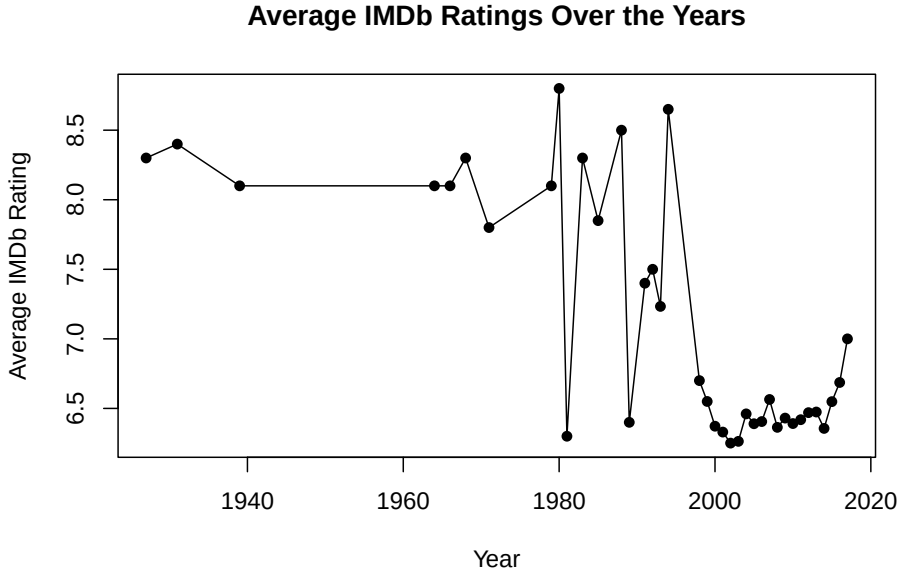
1. **IMDb Rating Trends Over Time and Average Runtime:**
    - Investigate trends in IMDb ratings over different years to assess if IMDb tends to view movies from one generation more favorably compared to others. Additionally, analyze the average runtime of movies over time. This section aims to provide insights into how movie ratings and runtime have evolved over the years and their potential impact on audience perception.
2. **IMDb Rating Frequency and Distribution:**
    - Analyze the frequency and distribution of IMDb ratings across movies in the dataset. This objective aims to understand the distribution of IMDb scores and identify any patterns or trends in audience ratings.

3. **IMDb Rating vs. Box Office Correlation:**
    - Investigate whether a correlation exists between IMDb ratings and Box Office earnings. This exploration aims to analyze the relationship between critical reception and financial success.
4. **Runtime vs. IMDb Rating**
    - Look at the relationship between film runtime and IMDb ratings. By examining this correlation, we may uncover patterns and trends that focus on the interplay between film duration and audience perception, offering valuable insights into how the length of a film impacts its audience reception.
5. **Rated Films vs. Runtime and IMDb Rating**
    - Analyze the variations in IMDb ratings and runtime across subpopulations defined by a film's rating. Look at how different audience demographics perceive films of varying classifications and how the runtimes may vary between these groups.

Data Cleaning

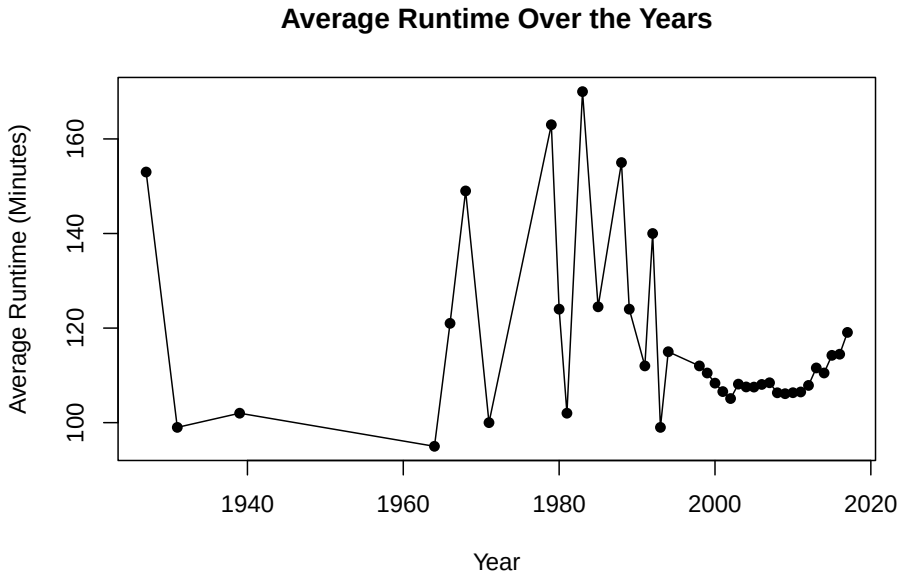
In the data cleaning process, we first loaded the dataset and examined its dimensions, identifying and summing up missing values. Since there are several unecessary columns, we decided that we can keep only the fundamental ones for this analysis: title, year, rated, runtime, imdbrating, boxoffice, and type (keeping only movies). Following this, we cleaned the dataset by removing rows with missing values, resulting in a refined dataset ready for analysis. Additionally, we transformed relevant variables, such as converting runtime to numeric values and adjusting box office earnings for consistency, utilizing regular expressions and 'gsub'. Instead of removing outliers, we retained them in the dataset for further analysis, recognizing their potential to provide insights into the dynamics of the film industry.

Exploratory Data Analysis

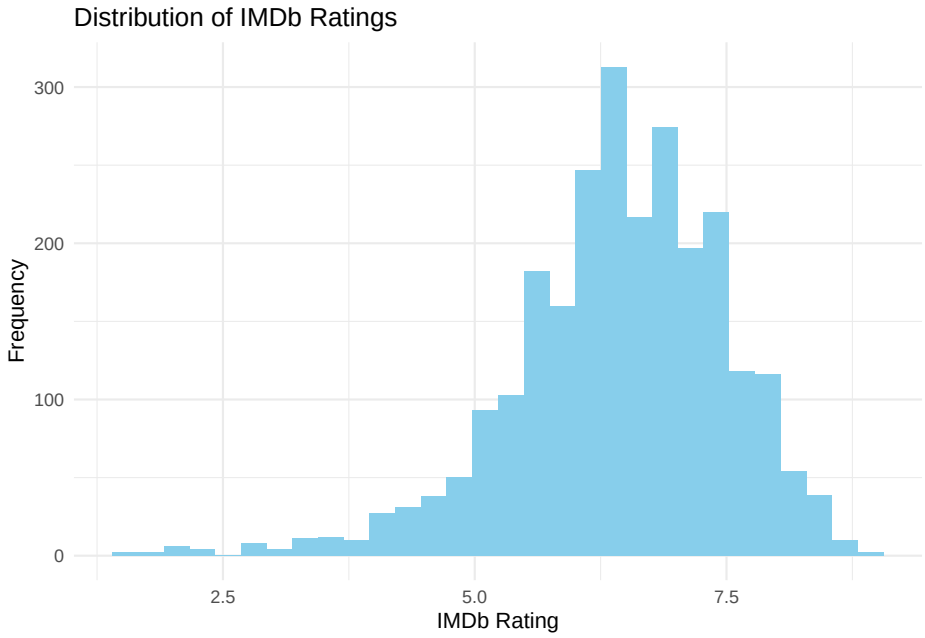


Looking at the average IMDb ratings over the years, we can see that from around the years 1920-1960, there was a significantly high average IMDb rating (~8.0). Within the years from 1960-2000, the variability increases, with a range

of ratings from ~6.25 to ~8.8. Lastly, the 2000s is marked by much less variability, implying a period of convergence. The early years of significantly high IMDb ratings from around 1920-1970s coincide with the well-known notion that this period is the "cinematic golden age." This era was popularized by the "glitz and glamour" of Hollywood, "a period in American filmmaking in which the five major studios, MGM, Paramount, Fox, Warner Bros., and RKO, dominated the production of major motion pictures (Heckman 2021)." While the high variability represented in the middle time period may represent a time of diversification in filmmaking, we eventually come back to a period where the IMDb ratings converge to ~6.5 in the 2000s. This convergence might reflect a homogenization of filmmaking styles, influenced by audience perception and the standardization of production techniques.

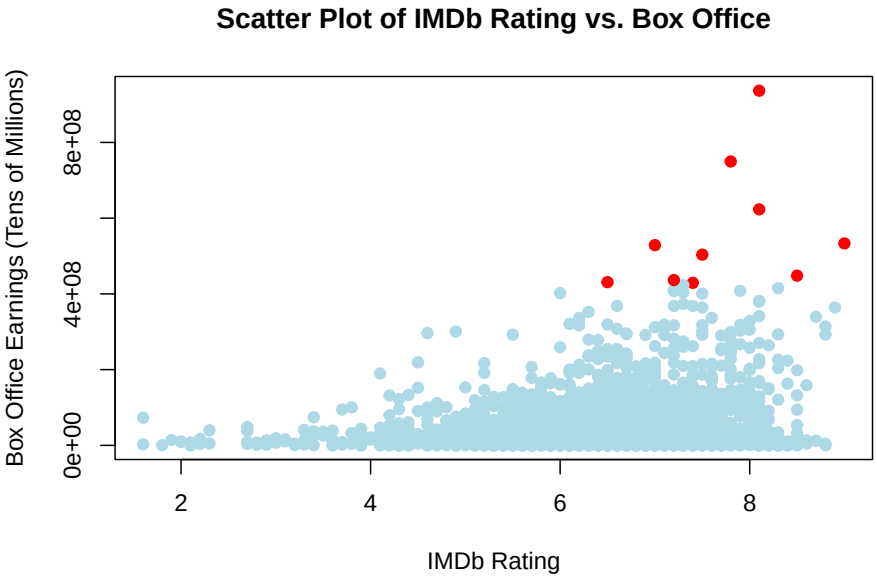


Presented by this second plot, we have the average runtime of a film (in minutes) over the years, revealing a large variability ranging from around 90 to 170 minutes. This analysis tracks the evolving preferences and practices in filmmaking, revealing shifts in narrative styles and audience expectations. However, around the 21st century, there's a notable convergence observed, marked by a significant reduction in runtime variability and a trend towards shorter film durations. This may reflect changing audience viewing habits influenced by the rise of digital platforms and the need for more concise films in an era of abundant media choices. Additionally, standardization in theatrical release formats and the growing significance of streaming services likely contribute to this trend, highlighting the dynamic nature of the film industry as it continues to adapt to modern consumption patterns.



In the third visualization, the distribution of IMDb Ratings is depicted through a histogram. This representation showcases the frequency of ratings within various score intervals, offering a comprehensive view of the rating distribution. With the median rating approximately at 6.6, it suggests that a significant portion of the ratings gravitate towards this score. The majority of the ratings span from ~5.8 to ~7.3, but it is noted that outliers may exist this range. Here, we can gain insights into the distribution pattern of IMDb ratings, highlighting both the common trends and the presence of extreme values.

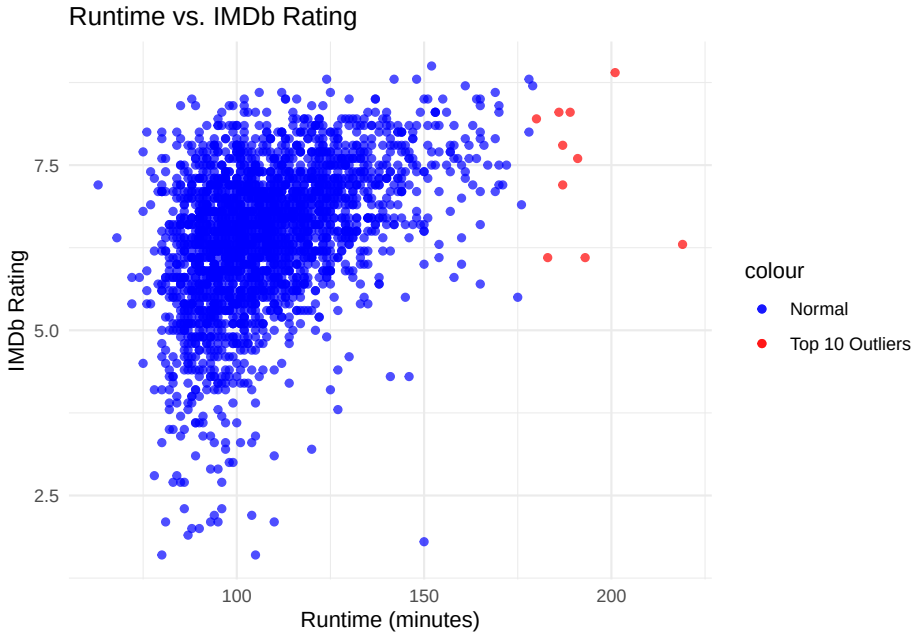
Further Analysis




```
##                                 Title Year Rated Runtime imdbRating
```

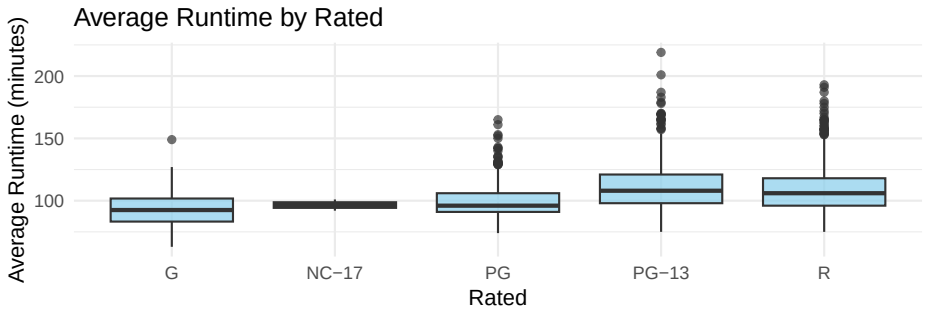
```
## 270                Star Wars: The Force Awakens 2015 PG-13     136        8.1
## 657                                      Avatar 2009 PG-13     162        7.8
## 320                                The Avengers 2012 PG-13     143        8.1
## 104                             The Dark Knight 2008 PG-13     152        9.0
## 678                              Jurassic World 2015 PG-13     124        7.0
## 2660                       Beauty and the Beast 2017    PG     129        7.5
## 152                       The Dark Knight Rises 2012 PG-13     164        8.5
## 1196                                    Shrek 2 2004    PG      93        7.2
## 879  Star Wars: Episode I - The Phantom Menace 1999    PG     136        6.5
## 664                     Avengers: Age of Ultron 2015 PG-13     141        7.4
##       Type BoxOffice
## 270  movie 936658640
## 657  movie 749700000
## 320  movie 623279547
## 104  movie 533316061
## 678  movie 528757749
## 2660 movie 503685062
## 152  movie 448130642
## 1196 movie 436471036
## 879  movie 431000000
## 664  movie 429113729
```

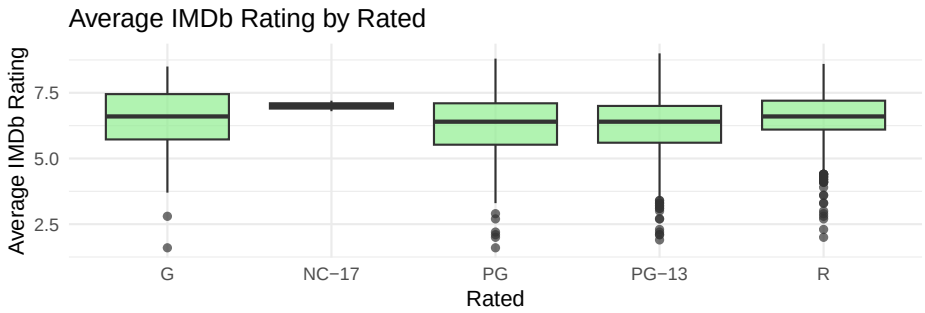
In this further analysis section, we have decided to dive deeper, past just the surface-level observations. This scatterplot depicts the relationships between IMDb Rating and Box Office Earnings (tens of millions of USD). Here, we see a small, positive relationship between the two values, noting that higher IMDb ratings may correlate with higehr Box Office Earnings. Wanting to examine the outliers further, we have found the 10 largest outliers, computing their distance from the median Box Office value. These films stand out because they have exceptionally high box office earnings, but their IMDb ratings aren't necessarily the highest. Take "Star Wars: Episode I - The Phantom Menace" for example, which raked in a whopping 430,000,000 USD despite having an IMDb rating of only 6.5. This discovery suggests that while there was a subtle relationship, a film's commercial success can't always be predicted solely by its critical acclaim, adding layers to our understanding of how movies resonate with audiences.



```
##                                                Title Year     Rated Runtime
## 1474                                Gods and Generals 2003     PG-13     219
## 107  The Lord of the Rings: The Return of the King 2003     PG-13     201
## 3493                          Kabhi Alvida Naa Kehna 2006         R     193
## 1497                                      Grindhouse 2007         R     191
## 380                                          Swades 2004 NOT RATED     189
## 1306                              The Hateful Eight 2015         R     187
## 686                                       King Kong 2005     PG-13     187
## 383                              Bhaag Milkha Bhaag 2013 NOT RATED     186
## 757                                   Pearl Harbor 2001     PG-13     183
## 216                          The Wolf of Wall Street 2013         R     180
##      imdbRating  Type BoxOffice
## 1474        6.3 movie  12900000
## 107         8.9 movie 364000000
## 3493        6.1 movie   3160978
## 1497        7.6 movie  24928753
## 380         8.3 movie   1174643
## 1306        7.8 movie  54116191
## 686         7.2 movie 218100000
## 383         8.3 movie   1626289
## 757         6.1 movie 197761540
## 216         8.2 movie  91330760
```

Analyzing the relationship between runtime and IMDb rating reveals a fairly evident positive correlation, with films boasting longer runtimes typically receiving higher IMDb ratings. However, examining outliers based on the farthest distance from the median runtime and IMDb rating exhibit notable exceptions. For instance, "Gods and Generals" with a lengthy runtime of 219 minutes garnered a modest IMDb rating of 6.3, while "The Wolf of Wall Street," with a shorter runtime compared to "Gods and Generals," received a significantly higher IMDb rating of 8.2. From this analysis, we can conclude that audience and critic perception generally favors lengthier runtimes, acknowledging exceptions as well.





Here, we have divided the movies into the main subpopulations, representing films that have been rated "G, NC-17, PG, PG-13, and R" as respective boxplots. Although there are two additional categories in our data, 'UNRATED' and 'NOT RATED,' we wanted to focus specifically on the main ratings commonly used for classification. When

examining the average runtime of the subpopulations, we see that G and PG films are almost identical in their range that is generally lower, reflecting their suitability for family audiences. Conversely, PG-13 and R-rated films tend to have slightly longer average runtimes, indicative of their appeal to older audiences and exploration of more mature themes. NC-17 films exhibit minimal variation, likely due to their niche audience and fewer entries displayed in the dataset. These trends underscore how different rating categories cater to their distinct target audiences, taking in their preferences.

Looking at the average IMDb rating, suprisingly, G and R rated films have the slightly-higher edge over the PG and PG-13 ones (with NC-17 again having a low representatoin). The difference in IMDb ratings across different rating categories highlights the diverse tastes and preferences of viewers, indicating that film quality transcends content classification.

Conclusion

Key Findings:

1. IMDb Rating Trends Over Time and Average Runtime: In analyzing IMDb rating trends over time and average runtime, we observed a significant evolution in the film industry. Initially, there was high variability in movie runtimes, which gradually decreased over the years, leading to a moderate convergence of average runtimes. We also found that IMDb ratings generally declined in the more recent years compared to previous years. This trend suggests a potential standardization or evolution in movie lengths and IMDb ratings, reflecting changes in audience preferences and industry practices.

2. IMDb Rating Frequency and Distribution: Exploring the frequency and distribution of IMDb ratings revealed valuable insights into the distribution of movie ratings within the dataset. We observed a diverse range of ratings, with the majority clustered around the middle ratings. This distribution pattern indicates a balanced mix of both high and low-rated movies, highlighting the diversity of ratings.

3. IMDb Rating vs. Box Office Correlation: Examining the correlation between IMDb ratings and box office earnings sheds light on the relationship between movie quality and financial success. While there was a positive correlation between IMDb ratings and box office earnings, the strength of this correlation varied across different movies, especially after analyzing the outliers. This finding suggests that while high IMDb ratings may indicate higher box office earnings on average, other factors such likely play significant roles in determining a movie's financial success.

4. Runtime vs. IMDb Rating Our analysis revealed a positive correlation between film runtime and IMDb rating, indicating that longer films tend to receive higher ratings. However, we also identified notable exceptions, emphasizing that runtime alone does not dictate audience and critic perception, underscoring the complexity of factors influencing film ratings.

5. Rated Films vs. Runtime and IMDb Rating Our analysis of films categorized by ratings uncovered trends in both average runtime and IMDb ratings. While G and R-rated films display slightly higher IMDb ratings compared to PG and PG-13 films, the variations in runtime reflect the diverse target audiences in each classification, with films geared towards family audiences having generally shorter runtimes compared with those catering to more mature preferences.

Implications and Actionable Insights:

These findings have several implications for various stakeholders within the industry. Producers and filmmakers can use the insights gained from the analysis to make informed decisions about movie production, such as optimizing runtime to align with audience preferences and prioritizing aspects that contribute to higher audience perceptions. Marketers can leverage the correlation between IMDb ratings and box office earnings to develop effective marketing strategies and plans for maximizing movie revenues and box office performance. Additionally, analysts can further explore the underlying factors driving the observed trends to gain deeper insights into the dynamics of the film industry.

Limitations

While our analysis provides valuable insights into the dynamics of the film industry, several limitations should be acknowledged. Firstly, the dataset's coverage does not fully represent the entire range of films released over the years, possibly changing the results of the proposed trends and relationships we explored. Additionally, it's important to note that IMDb ratings are inherently subjective. While IMDb provides a platform for users to rate and review movies, these ratings are influenced by various factors, including personal biases, cultural backgrounds, and viewing experiences. Thus, while IMDb ratings serve as a significant metric for gauging audience reception, they should be interpreted with caution. Lastly, while we have identified trends and relationships between certain variables, causation cannot be fully proven, and further research is needed to explore the underlying mechanisms driving these relationships.

Acknowledgements

Jasmine did the commentary for the introduction and conclusion, while Charlotte did the commentary for the analysis of the visualizations. We met up several times during the semester to work on the code and presentation together.

Bibliography

Heckmann, Chris. "When Was the Golden Age of Hollywood - and Why Did It End?" StudioBinder, 20 Dec. 2021, www.studiobinder.com/blog/when-was-the-golden-age-of-hollywood/.

"IMDB Top 250 Lists and 5000 plus IMDB Records - Dataset by Studentoflife." Data.World, 10 Feb. 2020, data.world/studentoflife/imdb-top-250-lists-and-5000-or-so-data-records.

Kosourova, Elena. "A Guide to R Regular Expressions with Examples." DataCamp, DataCamp, 14 Oct. 2022, www.datacamp.com/tutorial/regex-r-regular-expressions-guide.

"Ratings, Reviews, and Where to Watch the Best Movies & TV Shows." IMDb, IMDb.com, www.imdb.com/. Accessed 2 May 2024.

Appendix

```r
# Read CSV File And Only Keep Movies
df <- read.csv("IMDBdata_MainData.csv", na.strings = c("", "NA", "N/A", "NULL"))
df <- df[df$Type == "movie", ]

# Keep Only Specified Columns
df <- df[, c("Title", "Year", "Rated", "Runtime", "imdbRating", "Type", "BoxOffice")]

# Check Dimensions and Missing Values
dim(df)
```

```
## [1] 5270    7
```

```r
total_missing <- sum(is.na(df))
total_missing
```

```
## [1] 3104
```

```
# Remove Rows With Missing Values
df_clean <- na.omit(df)
new_total <- sum(is.na(df_clean))
new_total
```

```
## [1] 0
```

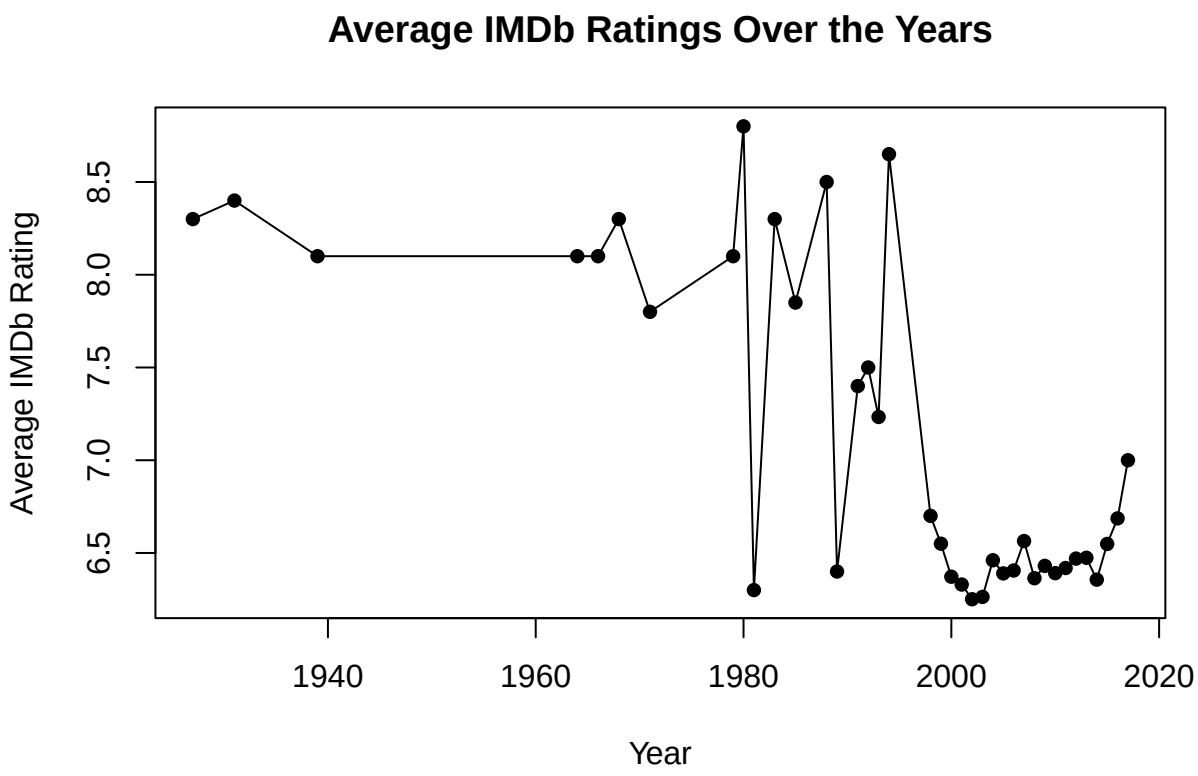
```
dim(df_clean)
```

```
## [1] 2550    7
```

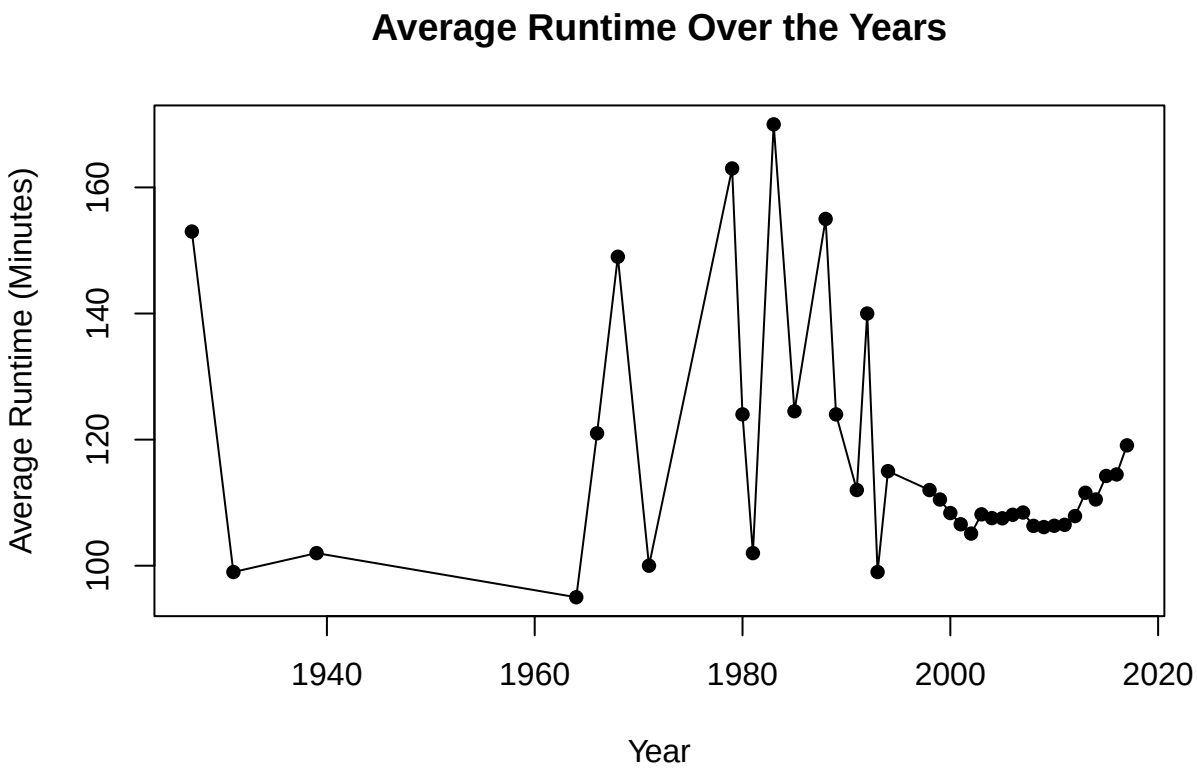
```
head(df_clean)
```

```
##                           Title Year     Rated Runtime imdbRating  Type
## 2              Saving Christmas 2014        PG  80 min        1.6 movie
## 3  Superbabies: Baby Geniuses 2 2004        PG  88 min        2.0 movie
## 12               Disaster Movie 2008     PG-13  87 min        1.9 movie
## 16         From Justin to Kelly 2003        PG  81 min        2.1 movie
## 25                   Himmatwala 2013 NOT RATED 150 min        1.8 movie
## 28            House of the Dead 2003         R  90 min        2.0 movie
##      BoxOffice
## 2   $2,778,297
## 3   $9,016,422
## 12 $14,174,654
## 16  $4,584,577
## 25    $270,880
## 28 $10,199,354
```

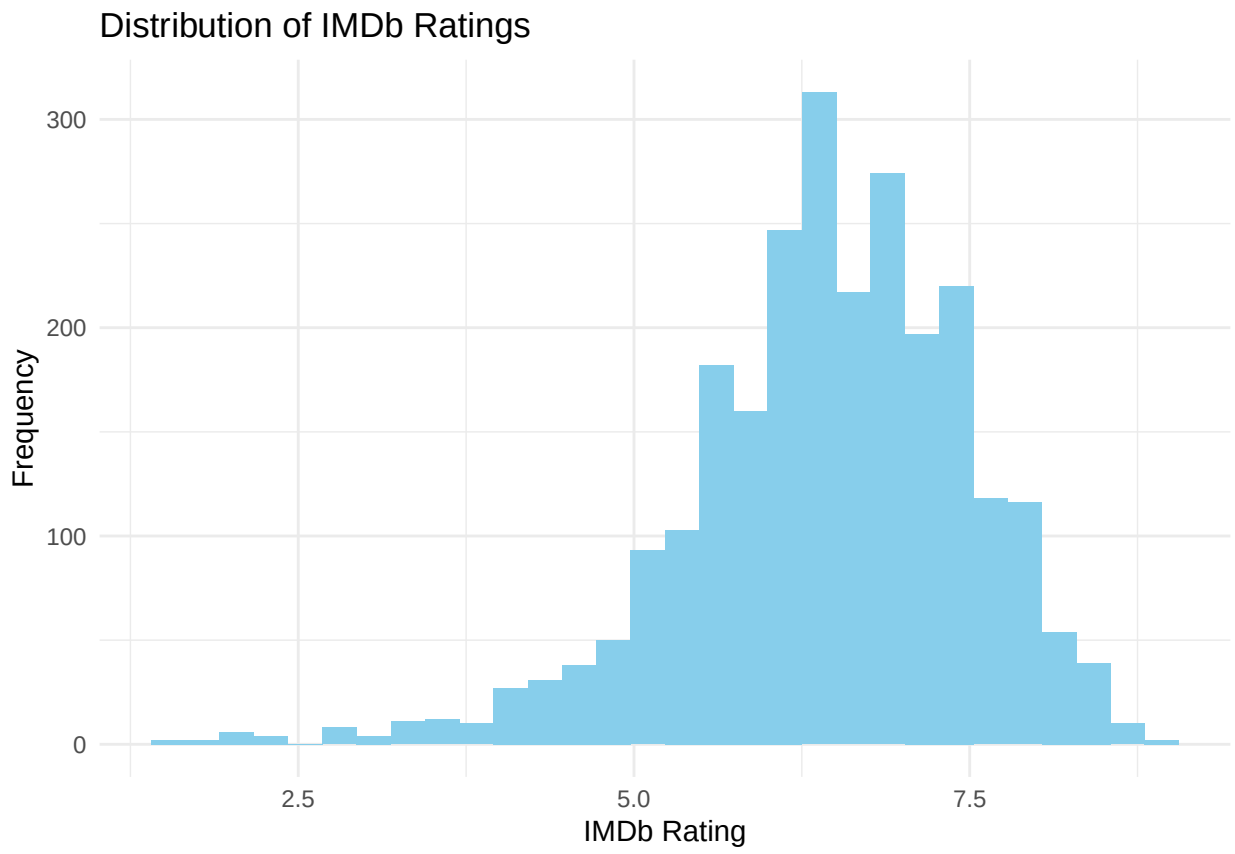
```
# Group by 'Year' And Calculate The Average imdbRating For Each Year
average_rating_by_year <- aggregate(imdbRating ~ Year, data = df_clean, FUN = mean, na.rm = TRUE)

# Plot
plot(average_rating_by_year$Year, average_rating_by_year$imdbRating,
     type = "o",
     pch = 16,
     xlab = "Year",
     ylab = "Average IMDb Rating",
     main = "Average IMDb Ratings Over the Years")
```



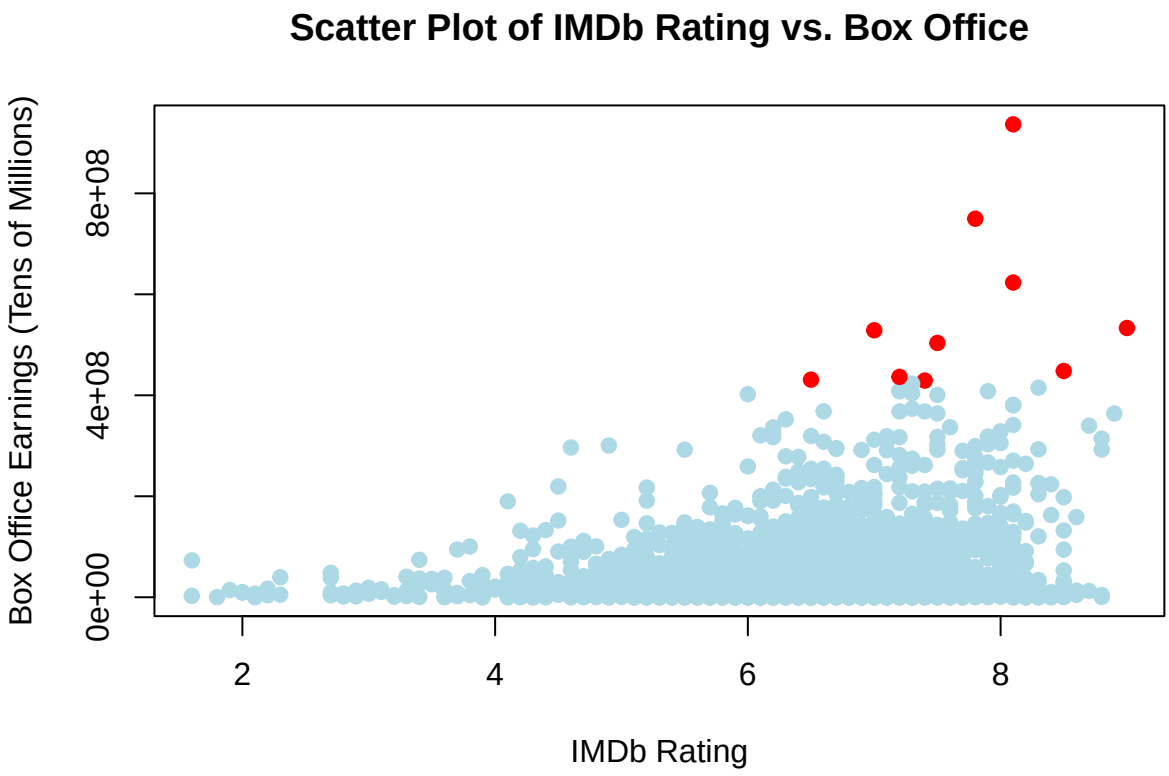

```r
# Use Regular Expressions to Change Runtime Into Numeric Value
library(stringr)
df_clean$Runtime <- as.numeric(str_extract(df_clean$Runtime, "\\d+"))

# Group by 'Year' And Calculate The Average Runtime For Each Year
average_runtime_by_year <- aggregate(Runtime ~ Year, data = df_clean, FUN = mean, na.rm = TRUE)

# Plot
plot(average_runtime_by_year$Year, average_runtime_by_year$Runtime,
     type = "o",
     pch = 16,
     xlab = "Year",
     ylab = "Average Runtime (Minutes)",
     main = "Average Runtime Over the Years")
```




```r
# Plot Distribution of IMDb Ratings
library(ggplot2)
ggplot(df_clean, aes(x = imdbRating)) +
  geom_histogram(fill = "skyblue", bins = 30) +
  ggtitle("Distribution of IMDb Ratings") +
  xlab("IMDb Rating") +
  ylab("Frequency") +
  theme_minimal()
```



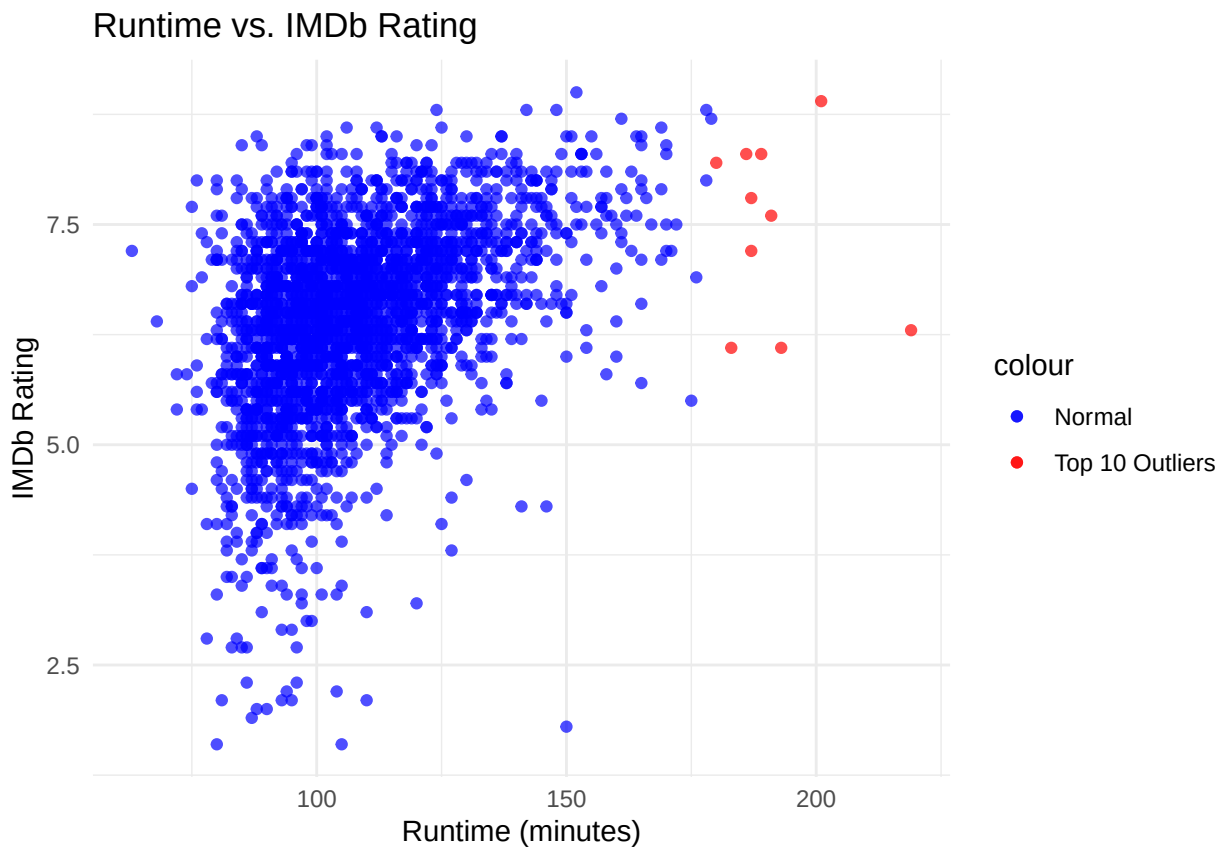

```r
# Remove '$' And ',' Characters And Convert To Numeric
df_clean$BoxOffice <- as.numeric(gsub("[\\$,]", "", df_clean$BoxOffice))

# Select Top 10 Outliers Based On Their Distances
median_boxoffice <- median(df_clean$BoxOffice)
distances_boxoffice <- abs(df_clean$BoxOffice - median_boxoffice)
threshold_boxoffice <- quantile(distances_boxoffice, 0.95)
outliers_boxoffice <- distances_boxoffice > threshold_boxoffice
outliers_df_boxoffice <- df_clean[outliers_boxoffice, ]
top_outliers_df_boxoffice <- head(outliers_df_boxoffice[order(distances_boxoffice[outliers_boxoffice], 

# Plot Top 10 Outliers In Red
plot(df_clean$imdbRating, df_clean$BoxOffice,
     xlab = "IMDb Rating",
     ylab = "Box Office Earnings (Tens of Millions)",
     main = "Scatter Plot of IMDb Rating vs. Box Office",
     pch = 19,
     col = ifelse(row.names(df_clean) %in% row.names(top_outliers_df_boxoffice), "red", "lightblue"),
     cex = 1)
```




```
# Display
top_outliers_df_boxoffice
```

```
##                                         Title Year Rated Runtime imdbRating
## 270                Star Wars: The Force Awakens 2015 PG-13     136        8.1
## 657                                      Avatar 2009 PG-13     162        7.8
## 320                                The Avengers 2012 PG-13     143        8.1
## 104                             The Dark Knight 2008 PG-13     152        9.0
## 678                              Jurassic World 2015 PG-13     124        7.0
## 2660                       Beauty and the Beast 2017    PG     129        7.5
## 152                      The Dark Knight Rises 2012 PG-13     164        8.5
## 1196                                    Shrek 2 2004    PG      93        7.2
## 879  Star Wars: Episode I - The Phantom Menace 1999    PG     136        6.5
## 664                     Avengers: Age of Ultron 2015 PG-13     141        7.4
##       Type BoxOffice
## 270  movie 936658640
## 657  movie 749700000
## 320  movie 623279547
## 104  movie 533316061
## 678  movie 528757749
## 2660 movie 503685062
## 152  movie 448130642
## 1196 movie 436471036
## 879  movie 431000000
## 664  movie 429113729
```

```r
# Calculate Outliers
Q1_x <- quantile(df_clean$Runtime, 0.25)
Q3_x <- quantile(df_clean$Runtime, 0.75)
IQR_x <- Q3_x - Q1_x
lower_cutoff_x <- Q1_x - 1.5 * IQR_x
upper_cutoff_x <- Q3_x + 1.5 * IQR_x

Q1_y <- quantile(df_clean$imdbRating, 0.25)
Q3_y <- quantile(df_clean$imdbRating, 0.75)
IQR_y <- Q3_y - Q1_y
lower_cutoff_y <- Q1_y - 1.5 * IQR_y
upper_cutoff_y <- Q3_y + 1.5 * IQR_y

# Calculate Median Of Both Axes
median_x <- median(df_clean$Runtime)
median_y <- median(df_clean$imdbRating)

# Calculate Distances From Median
distances <- sqrt((df_clean$Runtime - median_x)^2 + (df_clean$imdbRating - median_y)^2)

# Get The Indices
top_outliers_indices <- order(distances, decreasing = TRUE)[1:10]

# Extract Top 10 Biggest Outliers
top_outliers_df <- df_clean[top_outliers_indices, ]

# Create Scatter Plot
library(ggplot2)
ggplot(df_clean, aes(x = Runtime, y = imdbRating)) +
  geom_point(data = top_outliers_df, aes(color = "Top 10 Outliers"), alpha = 0.7) +
  geom_point(data = df_clean[-top_outliers_indices, ], aes(color = "Normal"), alpha = 0.7) +
  scale_color_manual(values = c("blue", "red"), labels = c("Normal", "Top 10 Outliers")) +
  labs(x = "Runtime (minutes)", y = "IMDb Rating", title = "Runtime vs. IMDb Rating") +
  theme_minimal()
```




```
top_outliers_df
```

```
##                                                Title Year     Rated Runtime
## 1474                             Gods and Generals 2003     PG-13     219
## 107  The Lord of the Rings: The Return of the King 2003     PG-13     201
## 3493                          Kabhi Alvida Naa Kehna 2006         R     193
## 1497                                    Grindhouse 2007         R     191
## 380                                        Swades 2004 NOT RATED     189
## 1306                           The Hateful Eight 2015         R     187
## 686                                    King Kong 2005     PG-13     187
## 383                              Bhaag Milkha Bhaag 2013 NOT RATED     186
## 757                                   Pearl Harbor 2001     PG-13     183
## 216                           The Wolf of Wall Street 2013         R     180
##      imdbRating  Type BoxOffice
## 1474        6.3 movie  12900000
## 107         8.9 movie 364000000
## 3493        6.1 movie   3160978
## 1497        7.6 movie  24928753
## 380         8.3 movie   1174643
## 1306        7.8 movie  54116191
## 686         7.2 movie 218100000
## 383         8.3 movie   1626289
## 757         6.1 movie 197761540
## 216         8.2 movie  91330760
```

```r
# Boxplot To See Variability
library(ggplot2)

# Filter Out 'Not Rated' And' Unrated'
df_clean_filtered <- df_clean[!df_clean$Rated %in% c('NOT RATED', 'UNRATED'), ]

# Group By 'Rated'
average_ratings_by_rated <- aggregate(cbind(Runtime, imdbRating) ~ Rated, data = df_clean_filtered, FUN

# Plot Boxplot For Runtime
p1 <- ggplot(df_clean_filtered, aes(x = Rated, y = Runtime)) +
  geom_boxplot(fill = "skyblue", alpha = 0.7) +
  labs(x = "Rated", y = "Average Runtime (minutes)", title = "Average Runtime by Rated") +
  theme_minimal()

# Plot Boxplot For Rating
p2 <- ggplot(df_clean_filtered, aes(x = Rated, y = imdbRating)) +
  geom_boxplot(fill = "lightgreen", alpha = 0.7) +
  labs(x = "Rated", y = "Average IMDb Rating", title = "Average IMDb Rating by Rated") +
  theme_minimal()

# Display
library(gridExtra)
grid.arrange(p1, p2, ncol = 1)
```