Movie analysis - statistical project

Radomska Małgorzata

2024-06-22

Project Description

The aim of the project is to conduct a statistical analysis of data concerning American movies released between 1915 and 2023. The data is publicly available at: https://www.kaggle.com/datasets/willianoliveiragibin/10000-data-about-movies-1915-2023/data. The dataset comprises 10,000 rows.

From the entire dataset, three features were selected for analysis: **Year of Release, Duration in Minutes and Rating on a Scale of 1-10**. Individual movies constitute the elements in the sample.

During the data analysis, basic statistical parameters were computed, data was visually presented through graphs, and statistical hypotheses were formulated and tested.

The R programming language was employed for data analysis purposes.

Data Preparation for Analysis

From the entire dataset, three columns were extracted for statistical analysis: **Year of Release, Duration in Minutes and Rating on a Scale of 1-10**. The column names were translated to Polish.

```
movies <- read.csv("~/GitHub/Movies-analysis/data/movies.csv", row.names=1)

data <- movies[,2:4]
head(data)
```

```
##   Year.of.Release Run.Time.in.minutes Movie.Rating
## 0            1994                 142          9.3
## 1            1972                 175          9.2
## 2            1993                 135          9.2
## 3            1975                  87          9.2
## 4            2008                 152          9.0
## 5            1993                 195          9.0
```

Statistical Analysis

For each feature, the following descriptive statistics were computed:

- Arithmetic mean,
- Harmonic mean,
- Variance,
- Standard deviation,
- Minimum value,
- Maximum value,
- Median,
- Quartiles,
- Range,
- Mode,
- Coefficient of variation,
- Skewness coefficient.

The number of records was assigned to the variable n.

```
(n <- length(data$Year.of.Release))
```

```
## [1] 10000
```

Wyniki analizy dla cechy Rok wydania

Analysis Results for Year of Release Feature

After performing calculations for the **Year of Release** feature, the following results were obtained:

- Arithmetic mean: 2001.414
- Harmonic mean: 0.0004996906
- Variance: 345.8783
- Standard deviation: 18.5978
- Minimum value: 1915
- Maximum value: 2023
- Median: 2007
- Quartiles: 0%: 1915, 25%: 1994, 50%: 2007, 75%: 2015, 100%: 2023
- Range: 108
- Mode: 2018
- Coefficient of variation: 0.9292334
- Skewness coefficient: -0.8918472

```
year <- data$Year.of.Release

(mean_year <- mean(year))
```

```
## [1] 2001.414
```

```r
(hmean_year <- sum(1/year) / n)
```

```
## [1] 0.0004996906
```

```r
(sd_year <- sd(year))
```

```
## [1] 18.5978
```

```r
(min_year <- min(year))
```

```
## [1] 1915
```

```r
(max_year <- max(year))
```

```
## [1] 2023
```

```r
(median_year <- median(year))
```

```
## [1] 2007
```

```r
(quan_year <- quantile(year))
```

```
##   0%  25%  50%  75% 100%
## 1915 1994 2007 2015 2023
```

```r
(range_year <- max_year - min_year)
```

```
## [1] 108
```

```r
(mode_year <- as.numeric(names(sort(table(year), decreasing = T)[1])))
```

```
## [1] 2018
```

```r
(cv_year <- sd_year / mean_year * 100)
```

```
## [1] 0.9292334
```

```r
(var_year <- var(year))
```

```
## [1] 345.8783
```

Analysis Results for Duration in Minutes Feature

After performing calculations for the **Duration in Minutes** feature, the following results were obtained:

- Arithmetic mean: 110.725
- Harmonic mean: 0.009339529
- Variance: 486.338
- Standard deviation: 22.05307
- Minimum value: 45
- Maximum value: 439
- Median: 107
- Quartiles: 0%: 45, 25%: 96, 50%: 107, 75%: 121, 100%: 439
- Range: 394
- Mode: 95
- Coefficient of variation: 19.91698
- Skewness coefficient: 0.7130526

```
duration <- data$Run.Time.in.minutes

(mean_duration <- mean(duration))
```

```
## [1] 110.725
```

```
(hmean_duration <- sum(1/duration) / length(duration))
```

```
## [1] 0.009339529
```

```
(sd_duration <- sd(duration))
```

```
## [1] 22.05307
```

```
(min_duration <- min(duration))
```

```
## [1] 45
```

```
(max_duration <- max(duration))
```

```
## [1] 439
```

```
(median_duration <- median(duration))
```

```
## [1] 107
```

```
(quan_duration <- quantile(duration))
```

```
##   0%  25%  50%  75% 100%
##   45   96  107  121  439
```

```r
(range_duration <- max_duration - min_duration)
```

```
## [1] 394
```

```r
(mode_duration <- as.numeric(names(sort(table(duration), decreasing = TRUE)[1])))
```

```
## [1] 95
```

```r
(cv_duration <- sd_duration / mean_duration * 100)
```

```
## [1] 19.91698
```

```r
(var_duration <- var(duration))
```

```
## [1] 486.338
```

Analysis Results for Rating Feature

After performing calculations for the **Rating** feature, the following results were obtained:

- Arithmetic mean: 6.72702
- Harmonic mean: 0.1509641
- Variance: 0.6744754
- Standard deviation: 0.8212645
- Minimum value: 4.9
- Maximum value: 9.3
- Median: 6.7
- Quartiles: 0%: 4.9, 25%: 6.1, 50%: 6.7, 75%: 7.3, 100%: 9.3
- Range: 4.4
- Mode: 6.7
- Coefficient of variation: 12.20844
- Skewness coefficient: 0.03290048

```r
rate <- data$Movie.Rating

(mean_rate <- mean(rate))
```

```
## [1] 6.72702
```

```r
(hmean_rate <- sum(1/rate) / n)
```

```
## [1] 0.1509641
```

```r
(sd_rate <- sd(rate))
```

```
## [1] 0.8212645
```

```r
(min_rate <- min(rate))
```

```
## [1] 4.9
```

```r
(max_rate <- max(rate))
```

```
## [1] 9.3
```

```r
(median_rate <- median(rate))
```

```
## [1] 6.7
```

```r
(quan_rate <- quantile(rate))
```

```
##   0%  25%  50%  75% 100%
##  4.9  6.1  6.7  7.3  9.3
```

```r
(range_rate <- max_rate - min_rate)
```

```
## [1] 4.4
```

```r
(mode_rate <- as.numeric(names(sort(table(rate), decreasing = TRUE)[1])))
```

```
## [1] 6.7
```

```r
(cv_rate <- sd_rate / mean_rate * 100)
```

```
## [1] 12.20844
```

```r
(var_rate <- var(rate))
```

```
## [1] 0.6744754
```

Visual Presentation of Data

In this part of the project, graphs were generated to present the data:

- histograms,
- density plots,
- empirical cumulative distribution function (ECDF) plots,
- box plots,
- trend line plots.

Histograms

Histograms were created for each feature. Histograms depict the distributions of data for the year of film release, duration in minutes, and ratings. They allow us to observe the distribution of film releases over the years, the spread of durations, and the frequency of ratings on a scale from 1 to 10. Conclusions drawn from the analysis of the following graphs:

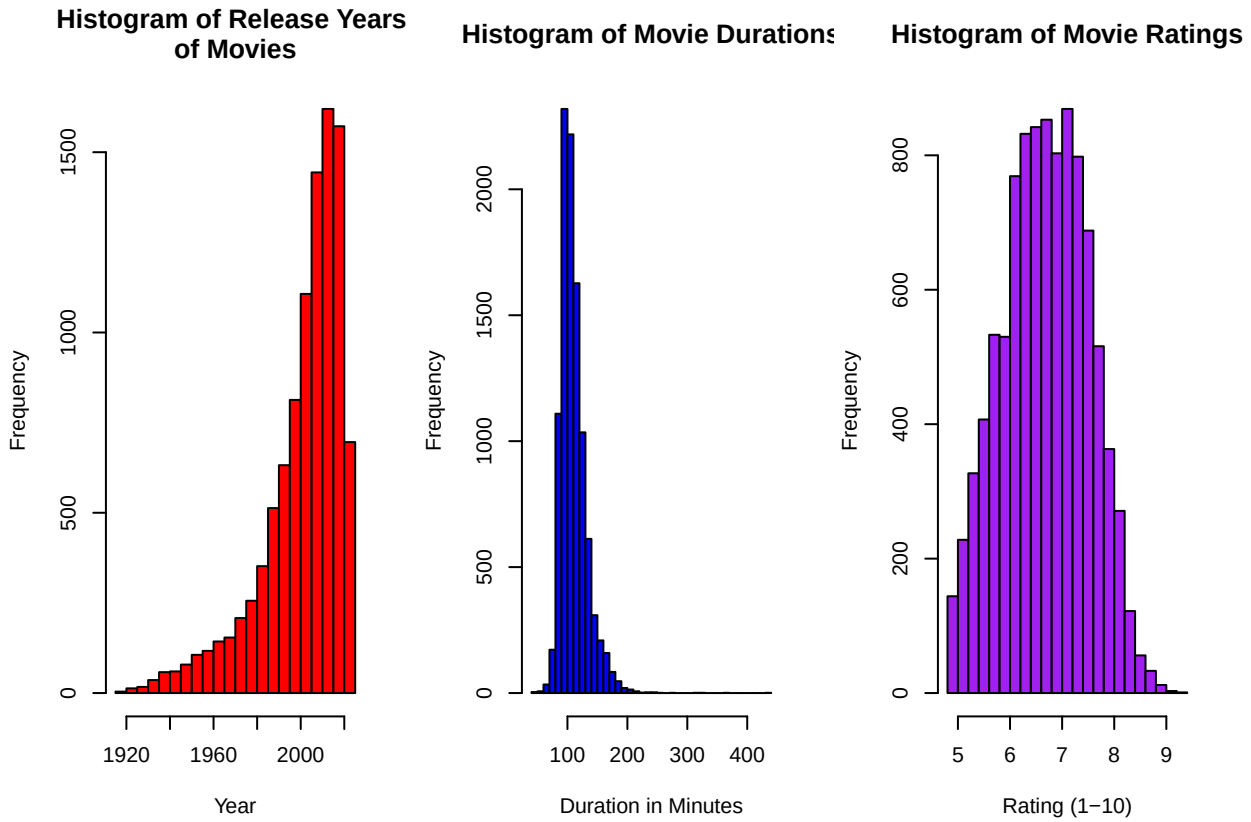
- The number of film releases shows an upward trend over the years with a noticeable decline during the Covid-19 pandemic period.
- Only a few films have durations exceeding 200 minutes.
- The most common ratings fall within the range of 6.0-7.5.

```
par(mfcol = c(1,3))

hist(year, main = "Histogram of Release Years \nof Movies", xlab = "Year",
     ylab = "Frequency", col = "red", breaks = 30)

hist(duration, main = "Histogram of Movie Durations",
     xlab = "Duration in Minutes", ylab = "Frequency", col = "blue",
     breaks = 30)

hist(rate, main = "Histogram of Movie Ratings", xlab = "Rating (1-10)",
     ylab = "Frequency", col = "purple", breaks = 30)
```



Density Plots

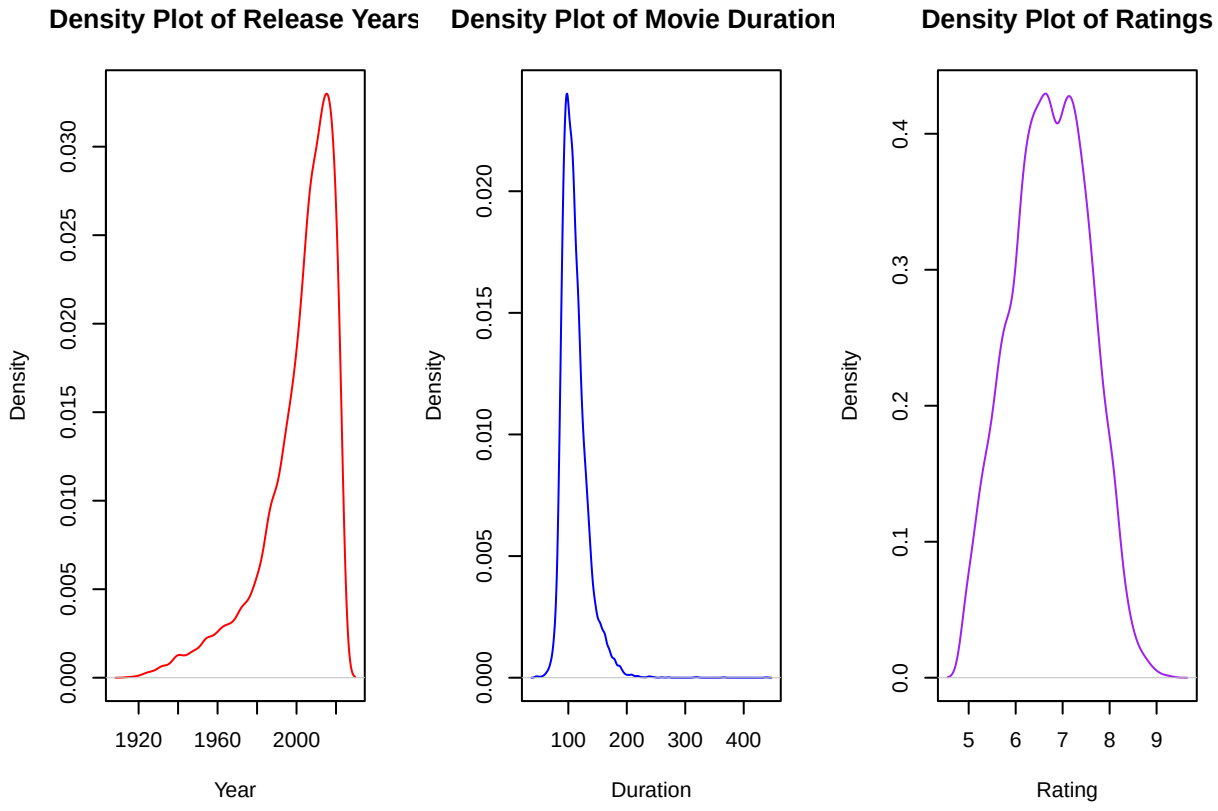
Density plots were generated for each feature. On density plots, we analyze the distributions of the number of films based on release year, duration, and rating. The values on the vertical axis indicate how many films were released in a given year, with a specific duration or rating. Higher density values indicate more films in that particular time frame, release year, or rating. It can be concluded that the shapes of the density plots align with those of histograms.

```
par(mfcol = c(1,3))

plot(density(year), main = "Density Plot of Release Years", xlab = "Year",
     ylab = "Density", col = 'red')

plot(density(duration), main = "Density Plot of Movie Durations",
     xlab = "Duration", ylab = "Density", col = 'blue')

plot(density(rate), main = "Density Plot of Ratings", xlab =  "Rating",
     ylab = "Density", col = 'purple')
```



Empirical CDF Plots

Empirical cumulative distribution function (ECDF) plots were generated for each feature. On ECDF plots, we track how films are distributed based on release year, duration, and rating. The values on the vertical axis indicate the proportion of the dataset represented by films released before a certain year, with a shorter duration, or rated equal to or lower than a specified value. Higher ECDF values indicate a larger percentage of films meeting the criterion. Conclusions drawn from these plots include:

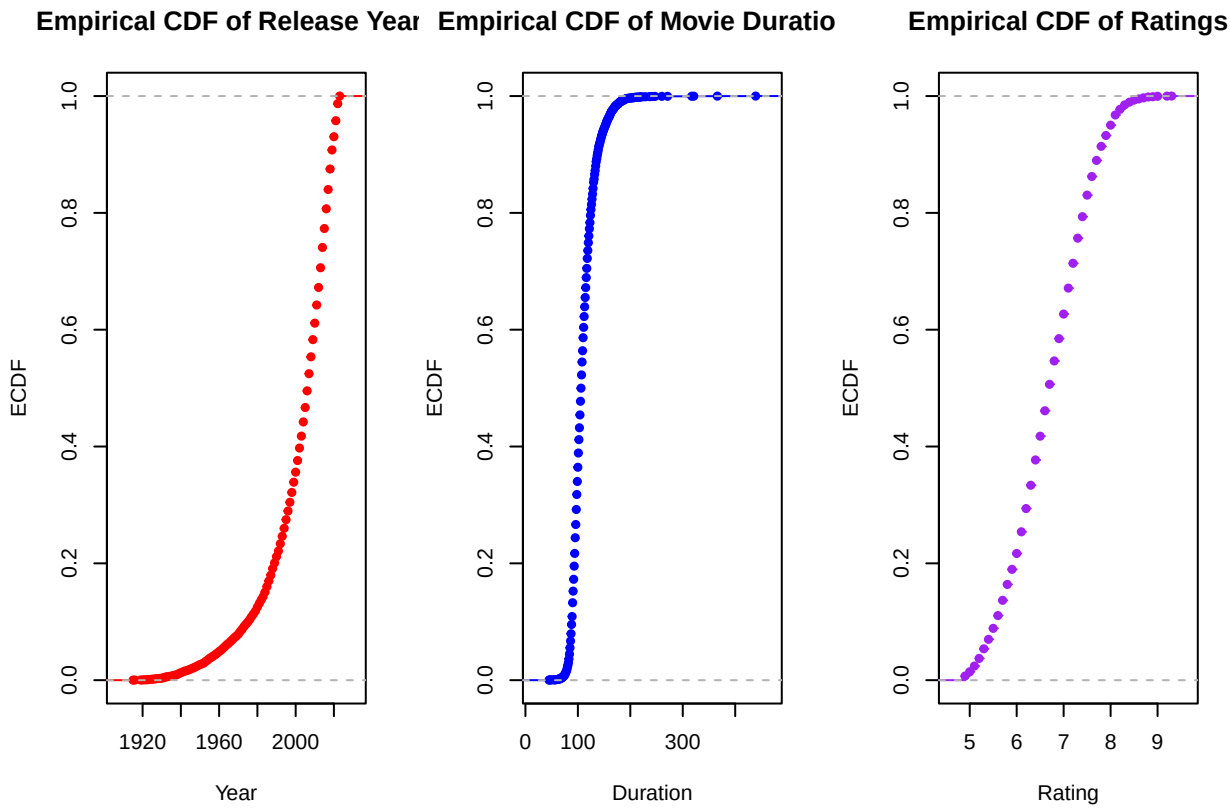
- Approximately 40% of films were released before the beginning of the 21st century.
- Nearly 100% of films have a duration of 200 minutes or less.
- Only 20% of films received a rating of 6.0 or lower.

```
par(mfcol = c(1,3))

plot(ecdf(year), main = "Empirical CDF of Release Years", xlab = "Year",
     ylab = "ECDF", col = 'red', cex = 0.75)

plot(ecdf(duration), main = "Empirical CDF of Movie Durations", xlab =
       "Duration", ylab = "ECDF", col = 'blue', cex = 0.75)

plot(ecdf(rate), main = "Empirical CDF of Ratings", xlab = "Rating",
     ylab = "ECDF", col = 'purple', cex = 0.75)
```



Box Plots

Box plots were generated for each feature. They show 50% of observations, which are between the first and third quartiles. The bold line represents the median. Lines extending from the box inform us about the distribution of data - if the data distribution is more concentrated, the "whiskers" will be narrower, whereas if the data distribution is more dispersed, the whiskers will be wider. Empty points denote outliers. Conclusions that can be drawn from the following plots include:

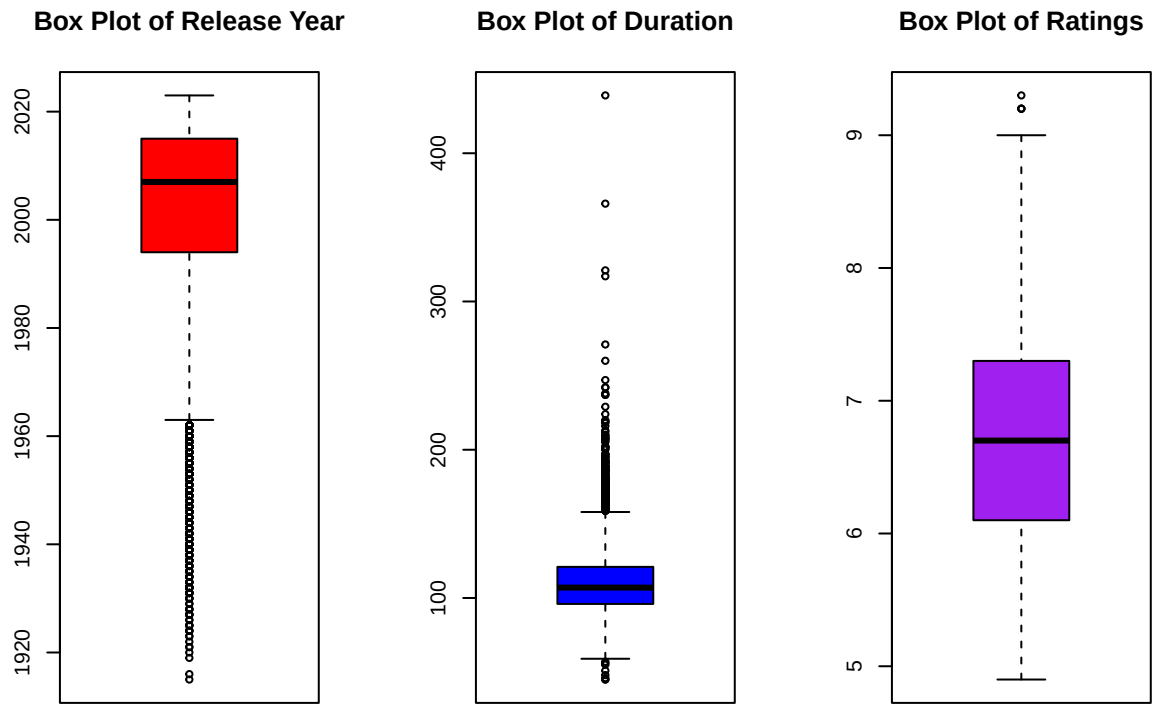
- 50% of films were released within approximately 20 years.
- The distribution of film durations is more concentrated compared to the distributions of other features.
- The median rating is approximately 6.7.

```
par(mfcol = c(1,3))

boxplot(year, col = 'red', main = "Box Plot of Release Year")

boxplot(duration, col = 'blue', main = "Box Plot of Duration")

boxplot(rate, col = 'purple', main = "Box Plot of Ratings")
```



Scatter Plots with Trend Line

Scatter plots with trend lines were generated to show relationships between the following features: release year and rating, release year and film duration, and film duration and rating. On the plots, individual observations are marked with points, and the black line represents the trend. Conclusions that can be drawn from the following plots include:

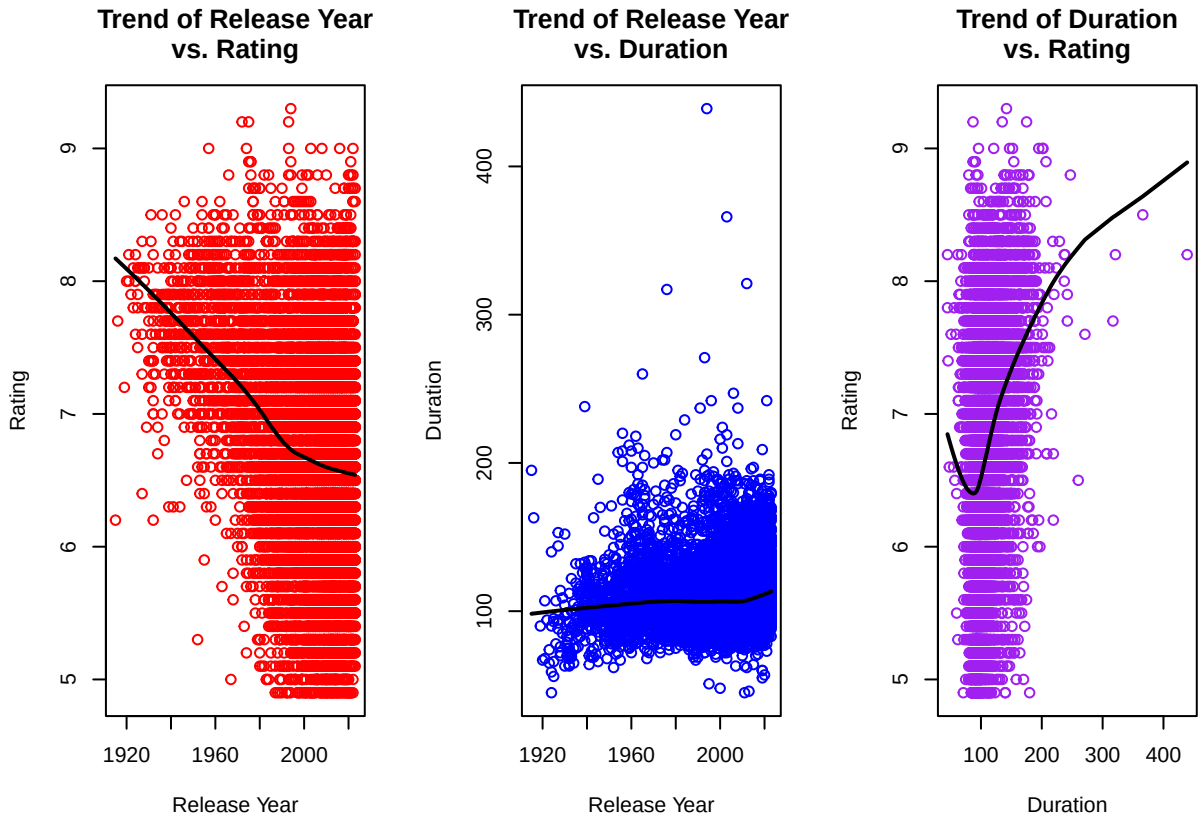
- A decreasing trend in ratings of films over the years can be observed.
- The release year does not significantly affect the duration of films.
- The lowest ratings are given to films with a duration of approximately 100 minutes.

```
par(mfcol = c(1,3))

plot(year, rate, main = "Trend of Release Year\nvs. Rating",
     xlab = "Release Year", ylab = "Rating", col = "red")
lines(lowess(year, rate), col = "black", lwd = 2)

plot(year, duration, main = "Trend of Release Year\nvs. Duration",
     xlab = "Release Year", ylab = "Duration", col = "blue")
lines(lowess(year, duration), col = "black", lwd = 2)

plot(duration, rate, main = "Trend of Duration\nvs. Rating",
     xlab = "Duration", ylab = "Rating", col = "purple")
lines(lowess(duration, rate), col = "black", lwd = 2)
```



Statistical Hypothesis Testing

Three statistical hypotheses were tested. They concern:

- the mean rating of films,
- the correlation between the release year of a film and its duration,
- the ratio of the variance of film duration to the variance of the release year.

Hypothesis regarding the mean rating of films

Null Hypothesis (H0: $\mu = 7$): The mean rating of films is equal to 7.
Alternative Hypothesis (H1: $\mu > 7$): The mean rating of films is greater than 7.
Significance Level: $\alpha = 0.1$.

Student's t-test was used to test the hypothesis.

```
t.test(rate, mu = 7, alternative = "greater", conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  rate
## t = -33.239, df = 9999, p-value = 1
## alternative hypothesis: true mean is greater than 7
## 90 percent confidence interval:
##  6.716494      Inf
## sample estimates:
## mean of x
##   6.72702
```

- The t-statistic value was -33.329.
- Degrees of freedom were 9,999.
- The p-value was 1.
- At a 90% confidence level, the confidence interval was [6.716494; ∞).
- The mean rating of films was 6.72702.

Due to the fact that the p-value ($p > \alpha$), there is no basis to reject the null hypothesis.

Hypothesis regarding the correlation between the release year of the film and its duration

Null hypothesis (H0: $\rho = 0$): There is no significant correlation between the release year of the film and its duration.
Alternative hypothesis (H1: $\rho \neq 0$): There is a significant correlation between the release year of the film and its duration.
Significance level: $\alpha = 0.01$.

Pearson correlation statistics were used to test the hypothesis.

```
cor.test(year, duration, conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  year and duration
## t = 5.416, df = 9998, p-value = 6.236e-08
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  0.02836929 0.07973162
## sample estimates:
##       cor
## 0.05408623
```

- The t-statistic value was 5.416.
- Degrees of freedom were 9,998.
- The p-value was 6.236×10^{-8}.
- At a 99% confidence level, the confidence interval was [0.02836929; 0.07973162).
- The correlation between the release year of the film and its duration was 0.05408623.

Due to the fact that the p-value is less than alpha, the null hypothesis should be rejected in favor of the alternative hypothesis.

Hypothesis regarding the ratio of variance of duration to variance of year

Null hypothesis (H0: $\frac{\sigma^2_{\text{duration}}}{\sigma^2_{\text{year}}} = 2$**):** The ratio of variance of duration to variance of year is at least 2.
Alternative hypothesis (H1: $\frac{\sigma^2_{\text{duration}}}{\sigma^2_{\text{year}}} < 2$**):** The ratio of variance of duration to variance of year is less than 2.
Significance level: $\alpha = 0.05$.

An F-test was used to verify the hypotheses comparing two variances.

```
var.test(duration, year, ratio=2, alternative = "less")
```

```
##
##  F test to compare two variances
##
## data:  duration and year
## F = 0.70305, num df = 9999, denom df = 9999, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is less than 2
## 95 percent confidence interval:
##  0.000000 1.453126
## sample estimates:
## ratio of variances
##           1.406096
```

- The F-statistic value was 0.70305.
- Degrees of freedom were 9,999.
- The p-value was 2.2×10^{-16}.
- At a 95% confidence level, the confidence interval was [0; 1.453126).
- The ratio of variance of duration to variance of year was 1.406096.

Due to the fact that the p-value is less than alpha, we reject the null hypothesis in favor of the alternative hypothesis.

Description of R Environment Functions Used

In the project, the following functions from various R packages were utilized:

- **utils package**:
  - *read.csv()*: reading data from CSV files into the R environment,
  - *head()*: displaying the first few rows of a data frame to understand its structure and contents;
- **base package**:
  - *length()*: returning the number of elements in an object,
  - *mean()*: calculating the arithmetic mean of values in a data vector,
  - *sum()*: calculating the sum of values in a data vector,
  - *min()*: finding the minimum value in a data vector,
  - *max()*: finding the maximum value in a data vector,
  - *table()*: creating a frequency table of values in a data vector,
  - *sort()*: sorting a data vector in ascending or descending order,
  - *names()*: returning the names of elements in an object,
  - *as.numeric()*: converting objects to numeric type;
- **stats package**:
  - *sd()*: calculating the standard deviation of values in a data vector,
  - *median()*: calculating the median of values in a data vector,
  - *quantile()*: calculating quantiles of values in a data vector,
  - *var()*: calculating the variance of values in a data vector,
  - *density()*: computing and plotting probability density function based on data,
  - *ecdf()*: computing empirical cumulative distribution function,
  - *lowess()*: fitting locally weighted regression to data points, enabling visualization of data trends,
  - *t.test()*: comparing means of two data samples,
  - *cor.test()*: performing correlation test between two variables,
  - *var.test()*: performing test of equality of variances between two data samples;
- **graphics package**:
  - *par()*: setting and retrieving graphical options in the current session,
  - *hist()*: creating histograms of data,
  - *plot()*: creating various types of plots,
  - *boxplot()*: creating boxplots of data,
  - *lines()*: adding lines to an existing plot.