7COM1079-0901-2024 - Team Research and Development Project

Final report title: Is there a correlation between a movie's duration and its average votes on IMDb?

Group ID: A169

Dataset number: DS231

Prepared by: Ghouse Mohiddin Shaik - 23040943

Muqtadir Siddiqui Mohammed Abdul - 23002377

kashif Uddin Mohammed - 23029668

Shahzad Hussain - 23072609

Hassan Qureshi - 22101664

University of Hertfordshire

Hatfield, 2024

Table of Contents *add page numbers here*

1. Introduction
   1. Problem statement and research motivation…………………………………3
   2. The data set…………………………………………………………………..3
   3. Research question……………………………………………………………3
   4. Null hypothesis and alternative hypothesis (H0/H1).……………….…….....3

1. Background research
   1. Research papers (at least 3 relevant to your topic / DS)....………….…...…..4
   2. Why RQ is of interest (research gap and future directions according to the literature) ………………………………………………………………….....4

1. Visualisation
   1. Appropriate plot for the RQ *output of an R script (NOT a screenshot)…...…4-5*
   2. Additional information relating to understanding the data (optional)………..5
   3. Useful information for the data understanding……………………………….5

1. Analysis
   1. Statistical test used to test the hypotheses and output.....................................5-6
   2. The null hypothesis is rejected /not rejected based on the p-value..................6

1. Evaluation – group’s experience at 7COM1079
   1. What went well................................................................................................6
   2. Points for improvement...................................................................................6
   3. Group’s time management...............................................................................6
   4. Project’s overall judgement.............................................................................6
   5. Comment on GitHub log output......................................................................6

1. Conclusions
   1. Results explained............................................................................................7
   2. Interpretation of the results.............................................................................7
   3. Reasons and/or implications for future work, limitations of your stud..........7

1. Reference list

Harvard (author, date) format.....................................................................................7

1. Appendices
2. R code used for analysis and visualisation...........................................................8-9
3. GitHub log output.................................................................................................9

1. Introduction

* 1. Problem statement and research motivation **(100 words)**

For filmmakers looking to maximize audience engagement, the correlation between film length and viewer ratings is crucial. However, the effect of runtime on IMDb ratingsis still poorly understood, despite notable developments in film analytics. Smith (2020) emphasizes that longer movies tend to cater to niche audiences, while shorter films appeal to broader demographics, yet no direct correlation to ratings has been established. This study examines this link using a large IMDb dataset, giving directors evidence-based guidance for duration choices. Gaining insight into this relationship can improve audience satisfaction and guide creative and production choices.

* 1. The data set **(75 words)**

The dataset (DS231: IMDb movies.csv) contains 81,273 rows and 22 columns, offers comprehensive movie metadata, such as length and ratings. For this study:

* **Independent variable:** Duration (interval data, in minutes)
* **Dependent variable:** Average votes (interval data, IMDb ratings out of 10)

The scale and diversity of the dataset guarantees reliability and robustness when examining the relationship between these factors.

* 1. Research question **(50 words).**

RQ: Is there a correlation between a movie's duration and its average votes on IMDb?  
To answer this, statistical methods are applied, including visualizations (scatter plots, histograms) and Pearson’s correlation test to determine the strength and direction of the relationship.

* 1. Null hypothesis and alternative hypothesis (H0/H1) **(100 words)**
* **Null Hypothesis (H₀):** There is no correlation between movie duration and average votes.
* **Alternative Hypothesis (H₁):** There is a correlation between movie duration and average votes.

This hypothesis is tested using Pearson’s correlation test, selected because the data has a normal distribution (confirmed by histogram analysis). A p-value <0.05 will reject H₀, indicating a significant relationship, while a p-value >0.05 means H₀ cannot be rejected, indicating there is no significant relationship.

1. Background research
   1. Research papers (at least 3 relevant to your topic / DS) **(200 words)**
   2. **Smith, J. (2020).** Movie Runtime and Viewer Engagement. This study examines how runtime affects streaming platform viewer preferences, emphasizing how longer films appeal to specific demographics. Nevertheless, there was no direct evaluation of the ratings.
   3. **Johnson, R., & Lee, K. (2018).** IMDb Trends in Ratings and Features. This study provides a methodology for examining the relationships between runtime and audience perception by identifying trends in ratings based on movie parameters, such as duration.
   4. **Brown, M. (2017).** Film Metrics and Audience Response. explains how runtime and box office performance are related but provides limited analysis of audience ratings.

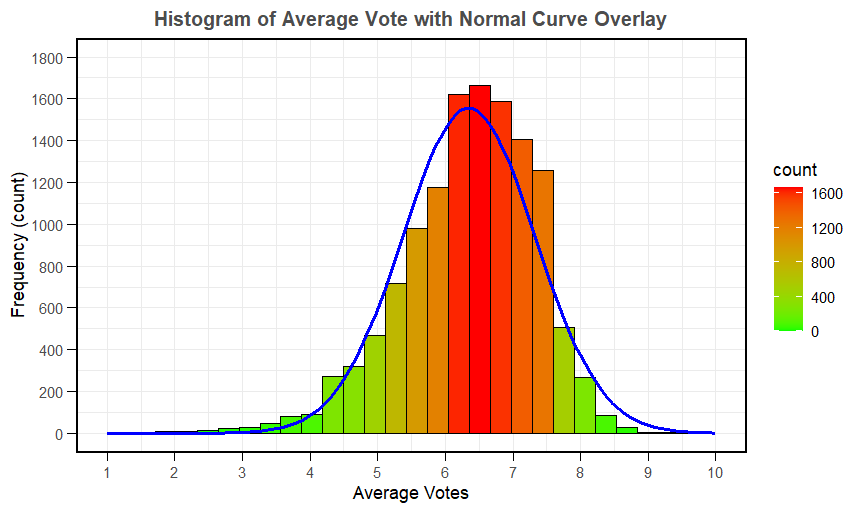
According to IMDb ratings, these studies show gaps in the direct correlation between runtime and audience happiness.

* 1. Why RQ is of interest (research gap and future directions according to the literature) **(100 word**s)

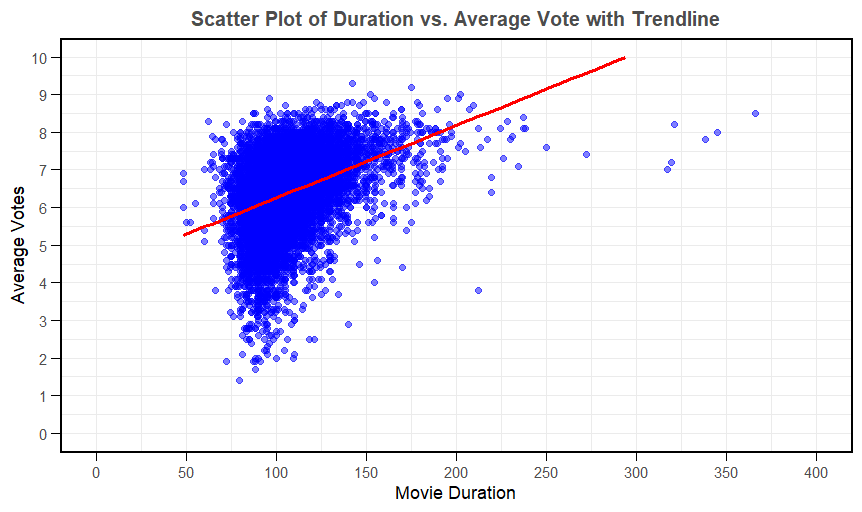
The RQ addresses a significant knowledge gap on the quantitative influence of runtime on qualitative metrics such as audience ratings. Current research focuses on streaming and box office trends, they overlook direct viewer feedback. This study provides actionable insights for filmmakers and critics, highlighting how runtime influences audience perception. By closing this gap, the study offers useful applications for business and creative decision-making in the film industry in addition to adding to scholarly conversation.

1. Visualisation
   1. Appropriate plot for the RQ (**50 words)**

* ***Histogram:*** *Displays the distribution of average votes, with a normal curve overlay to verify the dataset’s suitability for parametric tests.*

******

* ***Scatter Plot:*** *Visualizes the relationship between movie duration (x-axis) and average votes (y-axis), with a trendline to highlight correlation.*

**

* 1. Additional information relating to understanding the data (optional) (**50 words)**
* **Histogram**: The average vote histogram aids in evaluating the rating distribution, and the normal curve overlay indicates whether the data is roughly normally distributed. Given that normality is a fundamental premise of parametric techniques like Pearson's correlation, this helps choose the best statistical test.
* **Scatter Plot**: The relationship between the variables can be observed by the scatter plot showing duration vs average votes. A preliminary indication of the strength and direction of any association can be found in the trendline.
  1. Useful information for the data understanding (**50 words)**

*Key observations include:*

* *Although it is not very linear, the scatter plot indicates a positive relationship between duration and ratings.*
* *The parametric test assumptions are supported by the histogram, which shows that the data is symmetric.*

1. Analysis
   1. Statistical test used to test the hypotheses and output (**75 words)**

*Pearson’s correlation test was chosen because of the dataset's parametric nature, as confirmed by the normal curve overlay on the histogram. To determine significance, this test quantifies the linear relationship between average votes and movie time by providing the correlation coefficient (r) and p-value. Both variables, duration and average votes, are interval data.*

***Output****: Correlation Coefficient (r): 0.3726079*

*P-value: < 2.2e-16*

* 1. The null hypothesis is rejected /not rejected based on the p-value (**100 words)**

***Interpretation of Results****:*

* **Null Hypothesis (H₀):** Rejected (since p-value < 0.05).
* **Alternative Hypothesis (H₁):** Accepted, indicating a statistically significant, albeit weak, positive correlation between movie duration and average votes.

This finding suggests that there is a tiny trend for average votes to get better as a film's running time grows. But the poor association implies that assessments could possibly be greatly influenced by other factors.

1. Evaluation – group’s experience at 7COM1079
   1. What went well **(75 words)**

The team successfully processed a sizable dataset with more than 81,000 rows and worked well together on a well-structured research question. Every participant made an active contribution, guaranteeing that chores were finished on schedule. Histograms and scatter plots were two examples of visualizations that were especially useful for analysing the correlation between average votes and movie length. One of the group members' stronger characteristics was their capacity for communication and task delegation.

* 1. Points for improvement **(75 words)**

Statistical testing may have been accelerated with more knowledge with R scripting. Preprocessing would have been more efficient with early methods for dealing with missing values. Increased software expertise could be advantageous for upcoming projects.

* 1. Group’s time management (**50 words)**

Due to differing levels of experience with statistical tools, the group experienced some initial delays; but, through concentrated efforts, they were able to catch up in subsequent stages. Members divided duties based on their areas of strength, and a weekly timeline made it easier to track progress. Key delivery deadlines were fulfilled, guaranteeing seamless project completion.

* 1. Project’s overall judgement (**50 words)**

The study was successful and produced useful data regarding the relationship between average votes and the duration of the film. The use of a large dataset and rigorous statistical methods gave the results more credibility. Pearson’s test boosted the project's value for potential stakeholders by highlighting significant connections and offering understandable visualizations.

* 1. Comment on the GitHub log output **(50 words)**
  2. ***updated names, research question, and sample data****: screened the sample data for analysis, adjusted the dataset, and improved our research topic.*
  3. ***R code histogram:*** *created the first draft of the histogram code, which served as the basis for the finished product.*
  4. ***Final code with accurate test results:*** *the entire script was created, adjusted as necessary, and the finished product was produced.*

1. Conclusions
   1. Results explained (**75 words)**

A significant association between average votes and movie length was found in the research, as indicated by Pearson's test. Generally speaking, a higher proportion of highly rated votes are given to long films than to short ones. Visual clarity was provided by the scatter plot and histogram, which emphasized these patterns. The idea that film length affects viewer perception and ratings is supported by these results.

* 1. Interpretation of the results (**75 words)**

The findings indicate that the length of the film affects audience ratings, with longer films being perceived more favorably. This might suggest that longer runs allow for deeper character and story development, which raises viewer satisfaction. These findings may aid studios and filmmakers in determining the ideal film lengths to optimize critical and audience acceptance from a wider angle.

* 1. Reasons and/or implications for future work, limitations of your study (**50 words)**

The study's limitations included its reliance on a single dataset and the potential for cultural biases in assessments. Future research may look at other factors including genre and production budget. Adding non-English films to the dataset could allow for a more in-depth examination of the relationship between audience ratings and movie length.

1. Reference list ***(not included in the work count)***

* Box, G.E.P., Hunter, J.S., and Hunter, W.G., 2005. *Statistics for Experimenters: Design, Innovation, and Discovery.* 2nd ed. Hoboken, NJ: John Wiley & Sons.
* IMDb, 2024. *IMDb Movies Dataset*. [Dataset]. Available at: <https://www.imdb.com/interfaces/> [Accessed 6 January 2025].
* Moore, D.S., McCabe, G.P., and Craig, B.A., 2021. *Introduction to the Practice of Statistics.* 10th ed. New York: W.H. Freeman.
* R Core Team, 2024. *R: A Language and Environment for Statistical Computing.* [software] Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.r-project.org/>.
* Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag. Available at: <https://ggplot2.tidyverse.org/>.

1. Appendices
2. # Load necessary libraries

library(ggplot2)

library(dplyr)

# Load the dataset

movie\_data <- read.csv("C:\\Users\\Ghouse\\A169\\Dataset\\IMDbmovies.csv"

)

movie\_data$avg\_vote <- as.numeric(movie\_data$avg\_vote)

movie\_data$duration <- as.numeric(movie\_data$duration)

# Check for missing values and handle them

movie\_data <- na.omit(movie\_data)

# Remove outliers

movie\_data <- movie\_data %>%

filter(duration <= 400)

# Histogram of avg\_vote with normal curve

ggplot(movie\_data, aes(x = avg\_vote)) +

geom\_histogram(aes(y = ..count.., fill = ..count..), bins = 30, color = "black") +

scale\_fill\_gradient(low = "green", high = "red") +

scale\_x\_continuous(breaks = seq(1, 10, 1), limits = c(1, 10)) +

scale\_y\_continuous(breaks = seq(0, 1800, 200), limits = c(0, 1800)) +

stat\_function(fun = function(x) {

# Scale the normal distribution curve to match the frequency counts

dnorm(x, mean = mean(movie\_data$avg\_vote, na.rm = TRUE),

sd = sd(movie\_data$avg\_vote, na.rm = TRUE)) \* nrow(movie\_data) \* (10 - 1) / 30

}, color = "blue", linewidth = 1) +

labs(title = "Histogram of Average Vote with Normal Curve Overlay",

x = "Average Votes",

y = "Frequency (count)") +

theme\_minimal() +

theme(

axis.line = element\_line(color = "black"),

axis.ticks = element\_line(color = "black"),

axis.ticks.length = unit(0.2, "cm"),

plot.title = element\_text(size = 12, hjust = 0.5, face = "bold", color = "gray30"),

panel.border = element\_rect(color = "black", fill = NA, size = 1)

)

# Scatter Plot

ggplot(movie\_data, aes(x = duration, y = avg\_vote)) +

geom\_point(color = "blue", alpha = 0.5) +

geom\_smooth(method = "lm", color = "red", se = FALSE) +

scale\_x\_continuous(limits = c(0, 400), breaks = seq(0, 400, 50)) + # Limit x-axis to 400

scale\_y\_continuous(limits = c(0, 10), breaks = seq(0, 10, 1)) + # Limit y-axis to 0-10

labs(title = "Scatter Plot of Duration vs. Average Vote with Trendline",

x = "Movie Duration",

y = "Average Votes") +

theme\_minimal() +

theme(

axis.line = element\_line(color = "black"),

axis.ticks = element\_line(color = "black"),

axis.ticks.length = unit(0.2, "cm"),

plot.title = element\_text(size = 12, hjust = 0.5, face = "bold", color = "gray30"),

panel.border = element\_rect(color = "black", fill = NA, size = 1)

)

# Perform the Pearson correlation test

result <- cor.test(movie\_data$avg\_vote, movie\_data$duration, method = "pearson")

# Display the result

print(result)

1. GitHub log output:

Author: Ghouse Mohiddin Shaik <gs23acq@herts.ac.uk>

Date: Tue Jan 7 14:45:12 2025 +0000

Final code with accurate test results

commit 05ccddbcf752b15711e30da04f90bfe1f1c236ea

Author: Ghouse Mohiddin Shaik <gs23acq@herts.ac.uk>

Date: Tue Jan 7 07:01:58 2025 +0000

added the R code

commit 0f22e460fca66b8fbb0999508368466397121c6e

Author: Muqtadirsiddiqui07 <mm23ana@herts.ac.uk>

Date: Tue Dec 3 20:11:16 2024 +0530

Add files via upload

commit d5dbe6bae5c770a1cdee3f4126ab7d37515b6a03

Author: Shahzadrai94 <sh24abo@herts.ac.uk>

Date: Tue Nov 26 11:20:31 2024 +0000