dPrep IMBd project

Team 3 dPrep

2025-02-13

Research Motivation

Research question and method

This study aims to investigate whether directors with longer careers tend to produce higher-rated movies on IMDb. The research question, "Do directors with a long career have better-rated movies?", explores the relationship between career longevity and film quality. This question is relevant as it can provide insights into whether experience plays a significant role in filmmaking success, informing both aspiring and established directors, film producers, and scholars studying the film industry.

By analyzing IMDb datasets, we will measure a director's career length based on the span between their first and last directed movie and compare this to the average IMDb ratings of their films. Additionally, the number of productions attributed to each director will be taken into account, allowing us to explore whether a higher volume of work influences overall ratings. The findings could provide insight into whether experience and productivity correlate with higher audience appreciation.

To answer this question, a multiple regression analysis is chosen as the primary research method. Regression is well-suited for this study because it allows us to examine the relationship between a director's career length and the average IMDb rating of their movies while also considering the number of productions they have directed. By applying a multiple linear regression model, we can determine whether a longer career and/or a higher number of directed movies are associated with better ratings, while controlling for potential variability. Additionally, a boxplot analysis is included to categorize directors into career-length groups, and a scatter plot will be used to explore the relationship between the number of movies directed and IMDb ratings. This combination of methods ensures a comprehensive and statistically sound approach to answering the research question.

Way of deployment

To effectively communicate the findings, the results of the analysis will be presented through a combination of a structured PDF report and an interactive dashboard. The PDF report will provide a clear and concise summary of the research process, statistical findings, and key conclusions, ensuring accessibility for academic and industry professionals. Meanwhile, the interactive dashboard will allow users to explore the data dynamically, offering visual representations such as scatter plots and boxplots to illustrate trends in film ratings over a director's career. This dual approach ensures that the study's conclusions are both comprehensible and actionable for a broad audience, including researchers, filmmakers, and industry stakeholders.

The potential use of this workflow

An automated and reproducible workflow ensures transparency. This allows researchers to replicate and extend findings. The workflow facilitates comparative studies, and it saves time on the other hand. It also promotes open science by enabling collaboration. The workflow can be adapted for various film-related analyses and can be used as a valuable teaching tool for data-driven research. Students benefit from its structured approach and also simplify their projects. The reproducible workflow also enhances and ensures

scientific integrity. Its adaptability encourages broader applications, and fosters knowledge-sharing in the scientific community.

Data Exploration

Install and load required packages

```
library(tidyverse)
library(dplyr)
library(readr)
library(tinytex)
```

Step 1: Downloading and loading the IMBd data

```
data_title_crew <- read_tsv("https://datasets.imdbws.com/title.crew.tsv.gz")
data_title_basics <- read_tsv("https://datasets.imdbws.com/title.basics.tsv.gz")
data_title_ratings <- read_tsv("https://datasets.imdbws.com/title.ratings.tsv.gz")
data_name_basics <- read_tsv("https://datasets.imdbws.com/name.basics.tsv.gz")
```

Step 2

2.1 Filter for movies

```
movies <- data_title_basics %>%
  filter(titleType == "movie") %>%
  select(tconst, primaryTitle, startYear)
```

2.2a Extract Director information

```
directors <- data_title_crew %>%
  select(tconst, directors) %>%
  filter(!is.na(directors)) %>%
  separate_rows(directors, sep = ",")
```

2.2b Replace "\\N" in director dataset with actual NA's

```
directors <- directors %>%
  mutate(directors = na_if(directors, "\\N"))
```

2.3a Merge director IDs with name.basics.tsv.gz to get birthYear and calculate career span.

inner_join is used so only matching rows are selected

```
merged_director_data <- directors %>%
  inner_join(data_name_basics, by = c("directors" = "nconst"))
```

2.3b Select relevant columns in merged director data set

```
merged_director_data <- merged_director_data %>%
  select(directors, primaryName, birthYear, deathYear, tconst)
```

2.4a Merge director dataset with movies dataset

```
data_director_career <- merged_director_data %>%
  inner_join(movies, by = "tconst") %>%
  group_by(directors)
```

2.4b Replace "\\N" in dataset with actual NA's

```
data_director_career$birthYear[data_director_career$birthYear == "\\N"] <- NA
data_director_career$deathYear[data_director_career$deathYear == "\\N"] <- NA
data_director_career$startYear[data_director_career$startYear == "\\N"] <- NA
```

2.4c Turn character values into numeric values for the year variables

```
data_director_career$birthYear <- as.numeric(data_director_career$birthYear)
data_director_career$deathYear <- as.numeric(data_director_career$deathYear)
data_director_career$startYear <- as.numeric(data_director_career$startYear)
```

2.5 Compute career length in merged director/movie dataset

```
data_director_career <- data_director_career %>%
summarise(
    career_start = min(startYear),
    career_end = max(startYear),
    num_movies = n(),  # Count movies directed
    .groups = "drop"
  ) %>%
  mutate(career_length = career_end - career_start)
```

2.6 Merge director career data with movie ratings data

```
data_director_ratings <- directors %>%
  inner_join(data_title_ratings, by = "tconst") %>%
  group_by(directors) %>%
  summarise(
    avg_rating = mean(averageRating, na.rm = TRUE),
    .groups = "drop"
)
```

2.7 Final data (merge director career data with director ratings data)

```
final_data <- data_director_career %>%
  left_join(data_director_ratings, by = "directors") %>%
  filter(!is.na(avg_rating))  # Remove missing rating
```

Step 3: Statistical Analysis and Visualization

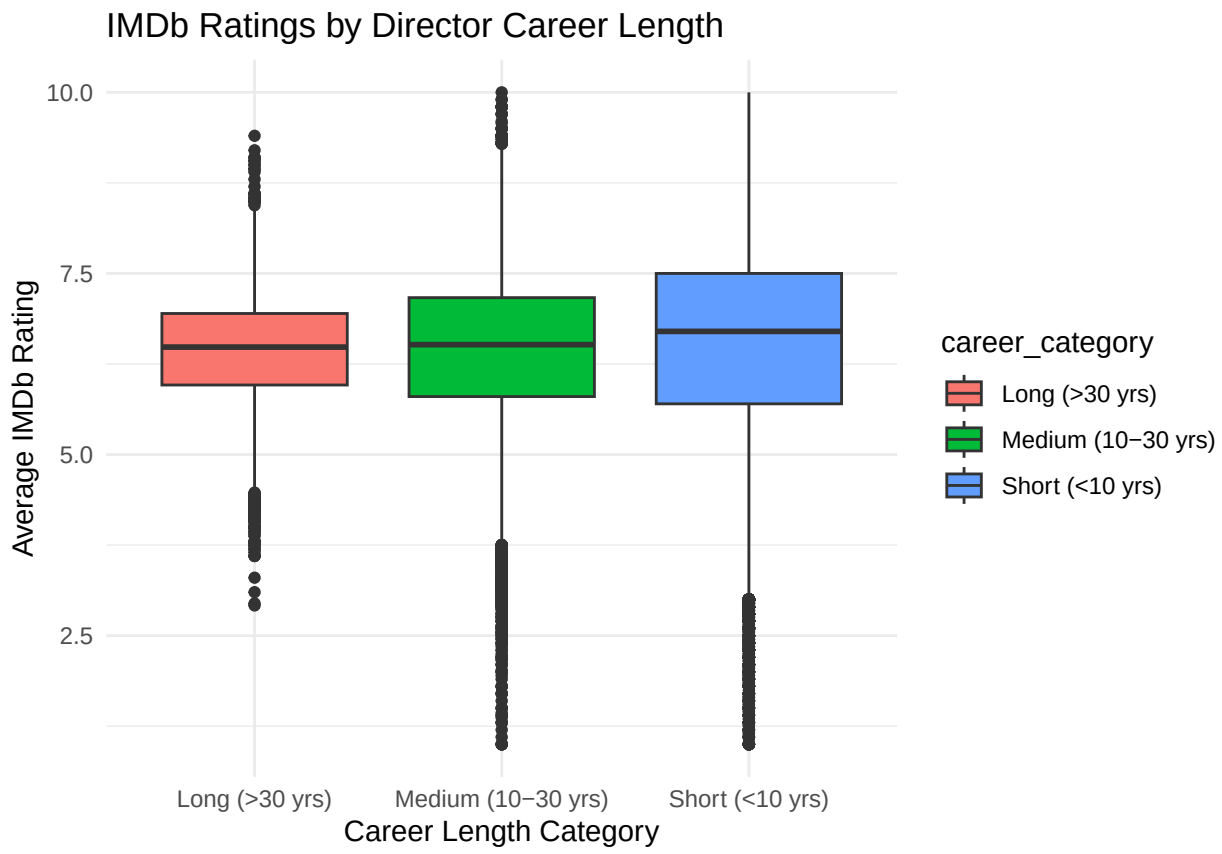
3.1 Multiple Linear Regression Analysis

```
model_1 <- lm(avg_rating ~ career_length + num_movies, data = final_data)
summary(model_1)
```

```
##
## Call:
## lm(formula = avg_rating ~ career_length + num_movies, data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5530 -0.7530  0.1470  0.8978  3.6678
##
## Coefficients:
##                 Estimate Std. Error  t value Pr(>|t|)
## (Intercept)    6.5620623  0.0042001 1562.374  < 2e-16 ***
## career_length -0.0015627  0.0004572   -3.418 0.000632 ***
## num_movies    -0.0091024  0.0005640  -16.138  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.34 on 138152 degrees of freedom
##   (23714 observations deleted due to missingness)
## Multiple R-squared:  0.003307,   Adjusted R-squared:  0.003292
## F-statistic: 229.2 on 2 and 138152 DF,  p-value: < 2.2e-16
```

3.2 Boxplot Analysis

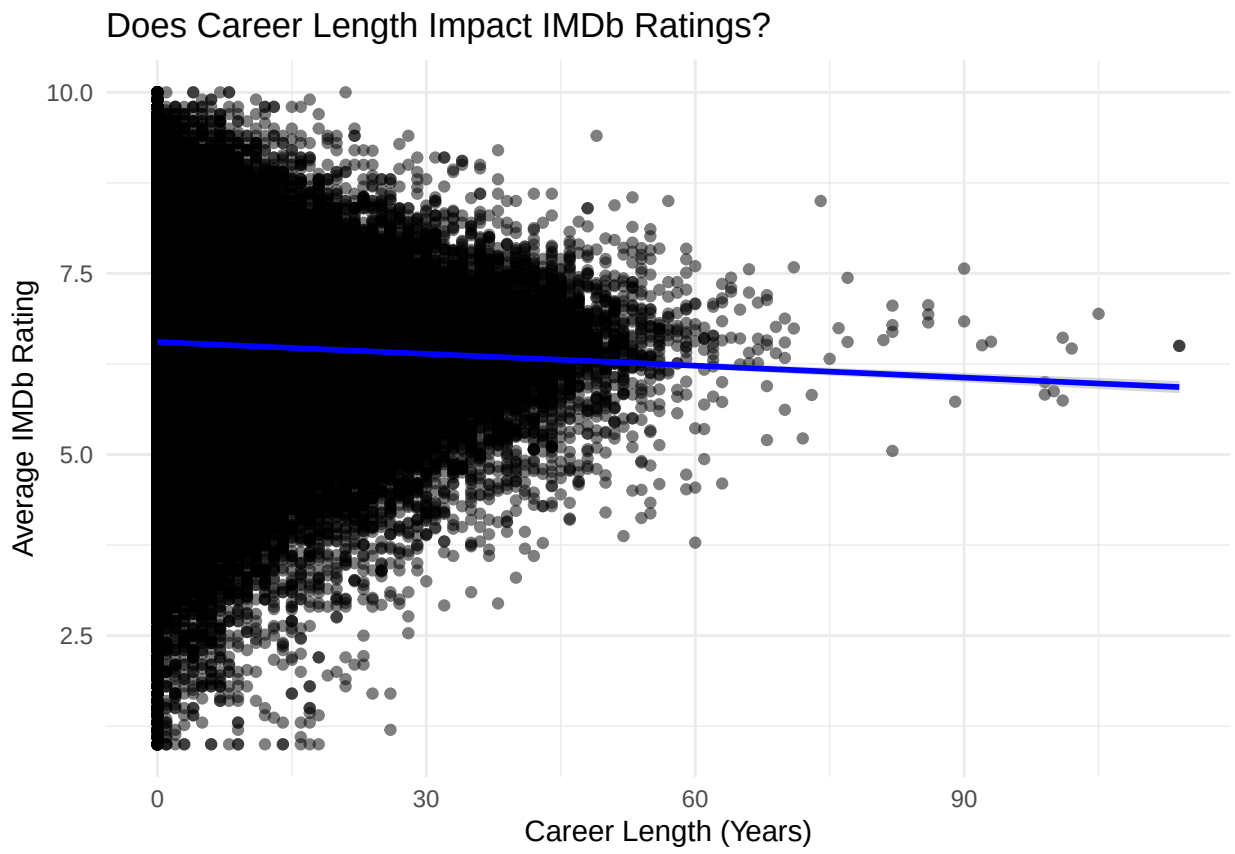
```r
final_data <- final_data %>%
  mutate(career_category = case_when(
    career_length < 10 ~ "Short (<10 yrs)",
    career_length >= 10 & career_length <= 30 ~ "Medium (10-30 yrs)",
    career_length > 30 ~ "Long (>30 yrs)"
  ))

ggplot(na.omit(final_data), aes(x = career_category, y = avg_rating, fill = career_category)) +
  geom_boxplot() +
  labs(title = "IMDb Ratings by Director Career Length",
       x = "Career Length Category",
       y = "Average IMDb Rating") +
  theme_minimal()
```



3.3 Scatter Plots: Career Length & Productivity vs Ratings

```
ggplot(final_data, aes(x = career_length, y = avg_rating)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(title = "Does Career Length Impact IMDb Ratings?",
       x = "Career Length (Years)",
       y = "Average IMDb Rating") +
  theme_minimal()
```

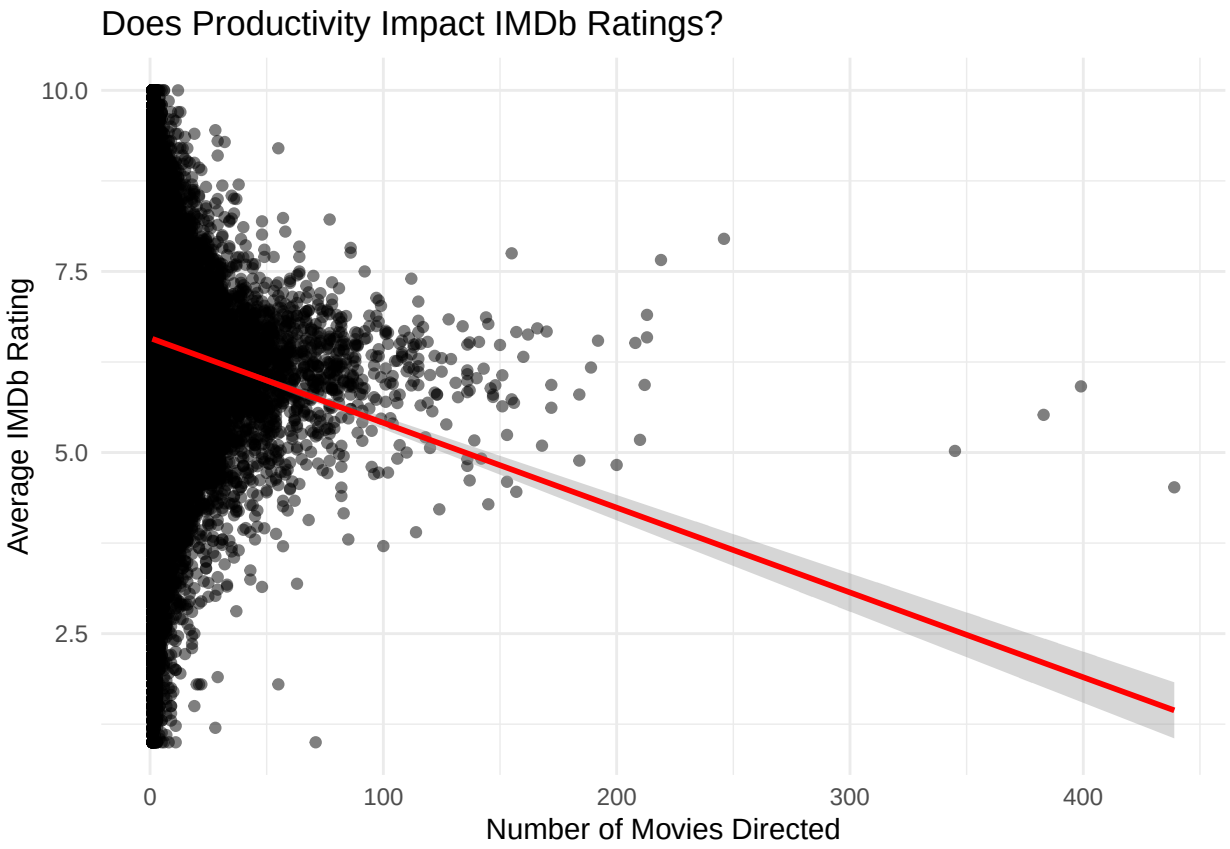
```
## `geom_smooth()` using formula = 'y ~ x'
```



3.4 Number of Movies Directed vs IMDb Rating

```
ggplot(final_data, aes(x = num_movies, y = avg_rating)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Does Productivity Impact IMDb Ratings?",
       x = "Number of Movies Directed",
       y = "Average IMDb Rating") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


Does Productivity Impact IMDb Ratings?
Average IMDb Rating
10.0
7.5
5.0
2.5
0
100
200
300
400
Number of Movies Directed