# Final Project_Individual_BI

ANUSHREE RAIPAT

2024-03-03

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ lubridate 1.9.3     ✔ tibble    3.2.1
## ✔ purrr     1.0.2     ✔ tidyr     1.3.1
## ✔ readr     2.1.5
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
d <- read.csv('cleaned_dataset_specific_columns.csv')
```

```
d$Released_Year <- as.numeric(d$Released_Year)
```

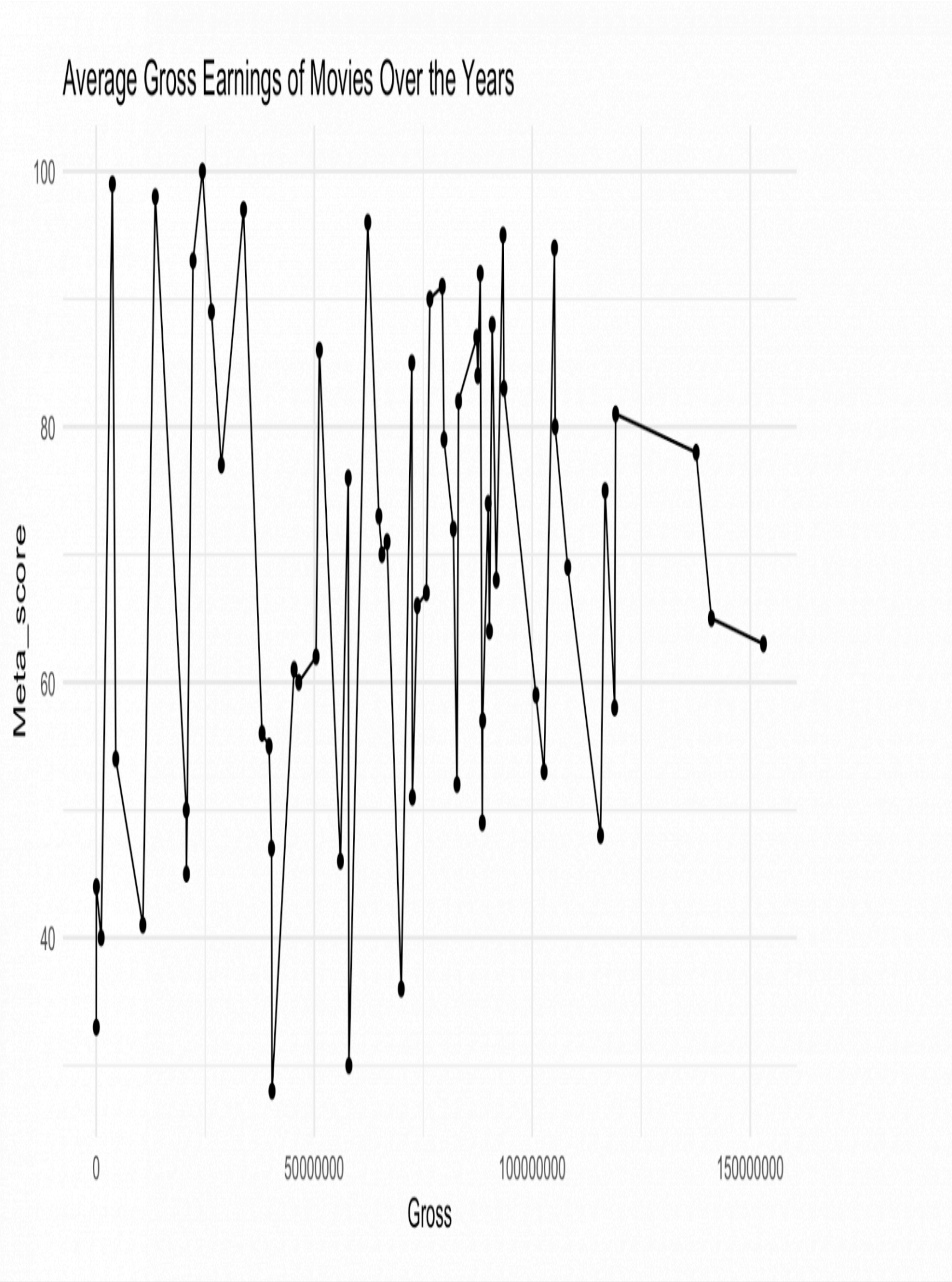
```
## Warning: NAs introduced by coercion
```

```
d$Gross <- as.numeric(d$Gross)
```

```
data <- na.omit(d)
```

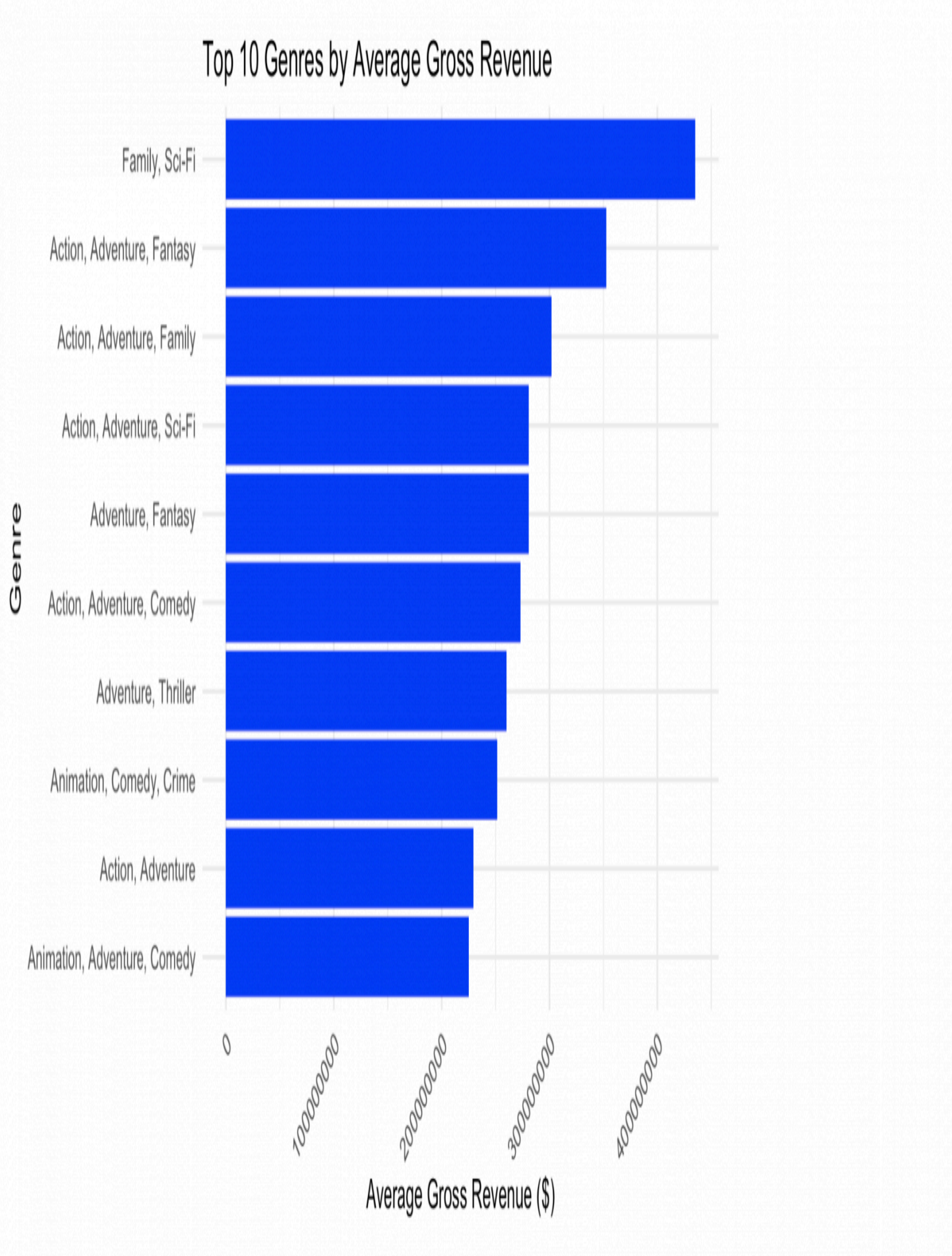
```
# Aggregate data to calculate mean gross per year
average_gross_per_year <- aggregate(Gross ~ Meta_score, data = data, FUN = mean)
```

```
# Plotting the average gross per year
ggplot(average_gross_per_year, aes(x = Gross, y = Meta_score)) +
  geom_line() +
  geom_point() +
  theme_minimal() +
  labs(title = "Average Gross Earnings of Movies Over the Years",
       x = "Gross",
       y = "Meta_score")
```




```
 "Average Gross Earnings of Movies Over the Years."

#1. Title and Axes:
   -#The plot represents the average gross earnings of movies over time.
    - #The x-axis is labeled "Gross," which likely represents the monetary value (in
```

```
dollars).
   - #The y-axis is not explicitly labeled, but it appears to represent some form of
rating or score (values range from 40 to 100).

#2. Trend and Variability:
   - #The line graph shows a fluctuating trend with significant variability in
earnings.
   - #Over the years, the average gross earnings of movies have experienced ups and
downs.
   - #Notably, there is a decline in average gross earnings toward the right end of
the plot.

#3. Interpretation:
   - #The declining trend suggests that recent movies may be earning less on average
compared to earlier years.
   - #Factors such as changing audience preferences, market saturation, or economic
conditions could contribute to this trend.
   - #Filmmakers and studios might need to adapt their strategies to address this
decline.


Genre_gross <- d %>%
  group_by(Genre) %>%
  summarise(Genre_gross  = mean(Gross, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(desc(Genre_gross))

top_genre <- head(Genre_gross, 10)

options(scipen = 999)

ggplot(top_genre , aes(x = reorder(Genre, Genre_gross), y = Genre_gross)) +
  geom_bar(stat = "identity", fill = "blue") +
  coord_flip() + # Flip the coordinates to make the plot horizontal
  labs(title = "Top 10 Genres by Average Gross Revenue", x = "Genre", y = "Average
Gross Revenue ($)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

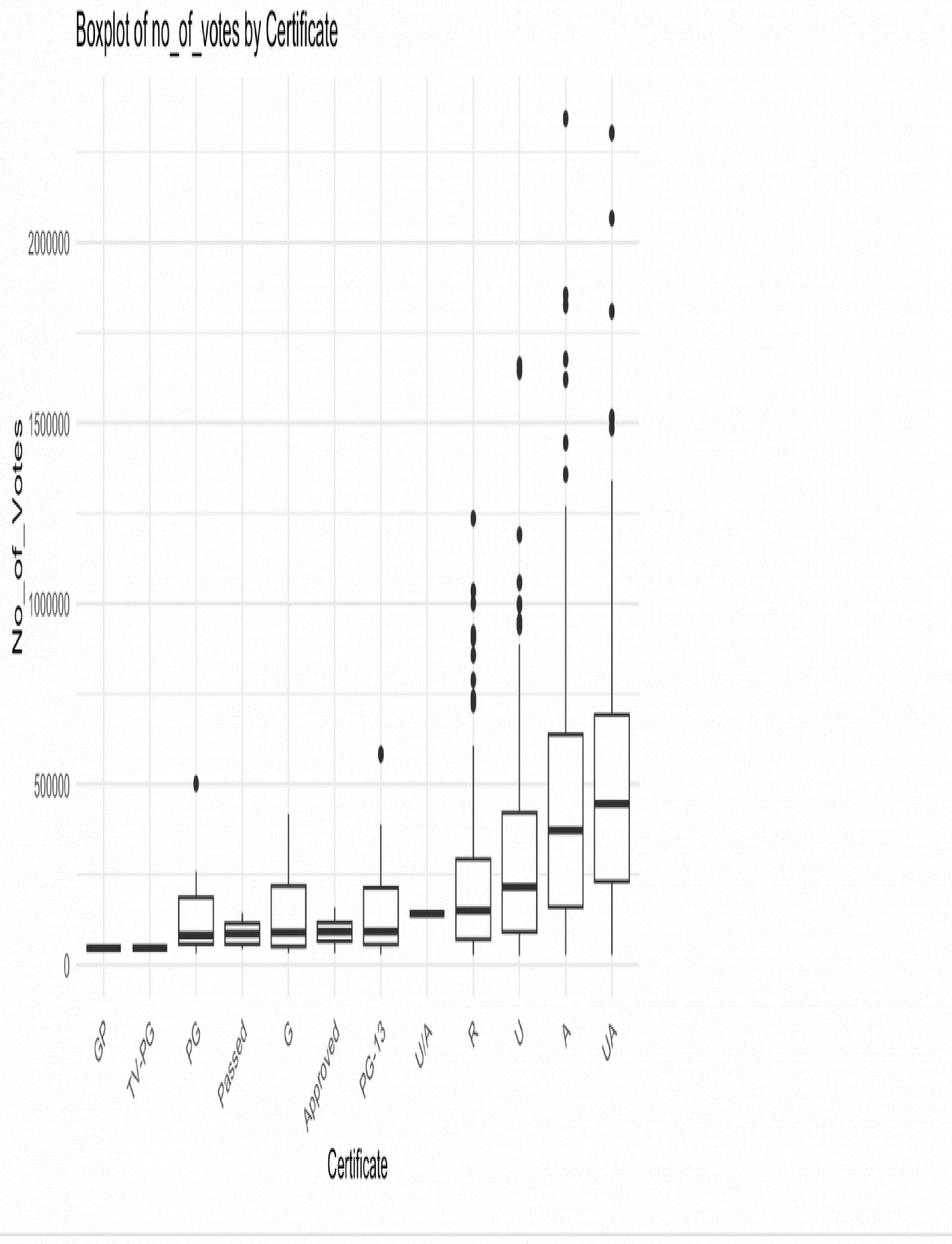



#Title and Axes:

- The plot represents the **average gross revenue** of movies for various genres.
- The x-axis represents the average gross revenue in dollars (ranging from 0 to over $400,000,000).
- The y-axis lists the top 10 movie genres.

```
2. #Genre Rankings:
   - The genres are listed in descending order of average gross revenue.
   - The top 10 genres are as follows:
     - Family, Sci-Fi: This combination genre has the highest average gross revenue
among all listed genres.
     - Action, Adventure, Fantasy: A mix of action, adventure, and fantasy genres.
     - Action, Adventure, Family: Combining action, adventure, and family elements.
     - Action, Adventure, Sci-Fi: A blend of action, adventure, and science fiction.
     - Adventure, Fantasy: Focusing on adventure and fantasy themes.
     - Action, Adventure, Comedy: A mix of action, adventure, and comedy.
     - Adventure, Thriller: Combining adventure and thriller elements.
     - Animation, Comedy, Crime: A genre mix involving animation, comedy, and crime.
     - Action, Adventure: Pure action and adventure.
     - Animation, Adventure, Comedy: Animated movies with adventure and comedy.

3. Insights:
   - The dominance of family-oriented genres(such as Family, Sci-Fi) suggests that
movies appealing to a broad audience tend to generate higher revenue.
   - Action and adventuregenres consistently appear in the top rankings.
   - Sci-Fi and fantasy elements contribute significantly to revenue.
   - Comedy and animation also play a role in revenue generation.
```

```r
d$No_of_Votes <- as.numeric(gsub(" min", "", d$No_of_Votes))

median_No_of_Votes <- d %>%
  group_by(Certificate) %>%
  summarise(Median_No_of_Votes = median(No_of_Votes, na.rm = TRUE)) %>%
  arrange(Median_No_of_Votes)

d$Certificate <- factor(d$Certificate, levels = median_No_of_Votes$Certificate)

# Create the boxplot
ggplot(d, aes(x = Certificate, y = No_of_Votes)) +
  geom_boxplot() +
  labs(title = "Boxplot of no_of_votes by Certificate", x = "Certificate", y =
"No_of_Votes") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




1. Title and Axes:
   - The plot shows the distribution of the number of votes received by movies.
   - The x-axis represents different movie certificates, including GP, TV-PG, PG,

```
Passed, G, Approved, PG-13, U/A, R, and U.
   - The y-axis represents the number of votes, ranging from 0 to 2,000,000.

2. Interpretation:
   - Each box represents the distribution of votes for movies with a specific
certificate.
   - Key observations:
     - U/A" Certificate: This category has the widest range of votes, indicating
variability. The median number of votes is relatively high.



# Load necessary libraries
d$IMDB_Rating <- as.numeric(d$IMDB_Rating)
d$Runtime <- as.numeric(d$Runtime)
d$Gross <- as.numeric(d$Gross)

data_clean <- na.omit(d[, c("Gross", "IMDB_Rating", "Runtime")])

# Linear regression model predicting Gross based on IMDB_Rating, Meta_score,
No_of_Votes, and Runtime
model <- lm(Gross ~ IMDB_Rating + Runtime, d = data_clean)

# Summary of the model
summary(model)




Call:
lm(formula = Gross ~ IMDB_Rating + Runtime, data = data_clean)

Residuals:
       Min         1Q     Median         3Q        Max
-155216849  -64203863  -38928111   25091929  850165892

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -276426860  115076994  -2.402 0.016557 *
IMDB_Rating   34606775   14944159   2.316 0.020856 *
Runtime         648766     169301   3.832 0.000138 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 113100000 on 711 degrees of freedom
Multiple R-squared:  0.03579,   Adjusted R-squared:  0.03308
F-statistic: 13.19 on 2 and 711 DF,  p-value: 0.000002362

#The model suggests that
a 1. Model Summary:
   - The model aims to predict the gross revenue of movies based on two predictors:
```

IMDB rating and runtime.
   - The model's performance is summarized by the coefficients, p-values, and R-squared values.

2. Coefficients:
   - Intercept: The estimated gross revenue when both IMDB rating and runtime are zero is approximately -$276,426,860 (negative value).
   - IMDB Rating: For every one-unit increase in IMDB rating, the average gross revenue increases by approximately $34,606,775.
   - Runtime: For every one-minute increase in runtime, the average gross revenue increases by approximately $648,766.

3. Significance:
   - The p-values associated with IMDB rating and runtime are both less than 0.05 (the typical significance level).
   - This indicates that both predictors are statistically significant in predicting gross revenue.

4. Adjusted R-squared:
   - The adjusted R-squared value is 0.03308.
   - It represents the proportion of variability in gross revenue explained by the model.
   - In this case, only about 3.31% of the variability in gross revenue can be explained by IMDB rating and runtime.

Overall Interpretation:
   - The model suggests that higher IMDB ratings and longer runtimes are associated with higher average gross revenue for movies.
   - However, the overall predictive power of the model is relatively low, as indicated by the low R-squared value.