Mini Project 2

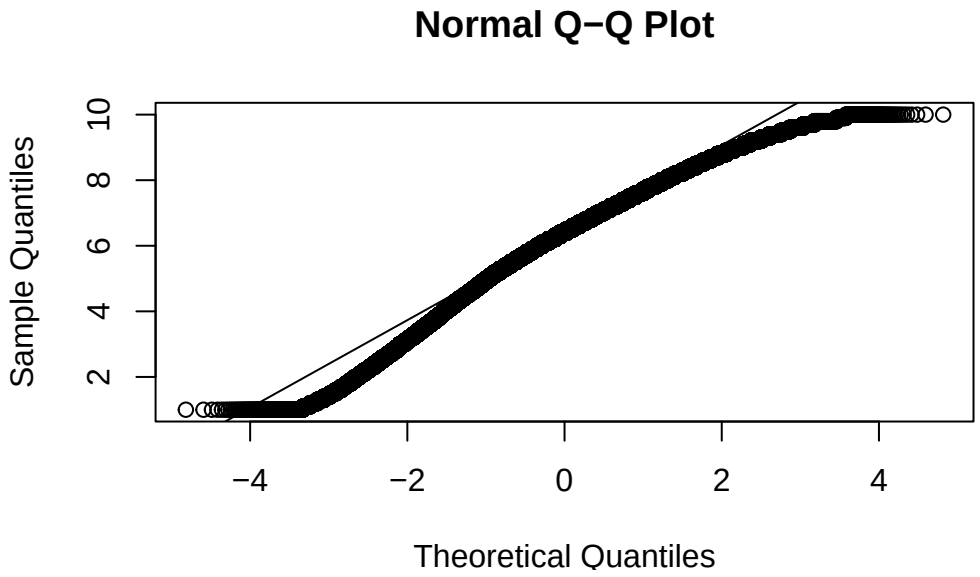
Xiaofeng Cai (5804513)

May 14, 2024

Table of contents

Part 1: Check significant difference between average ratings across different genres

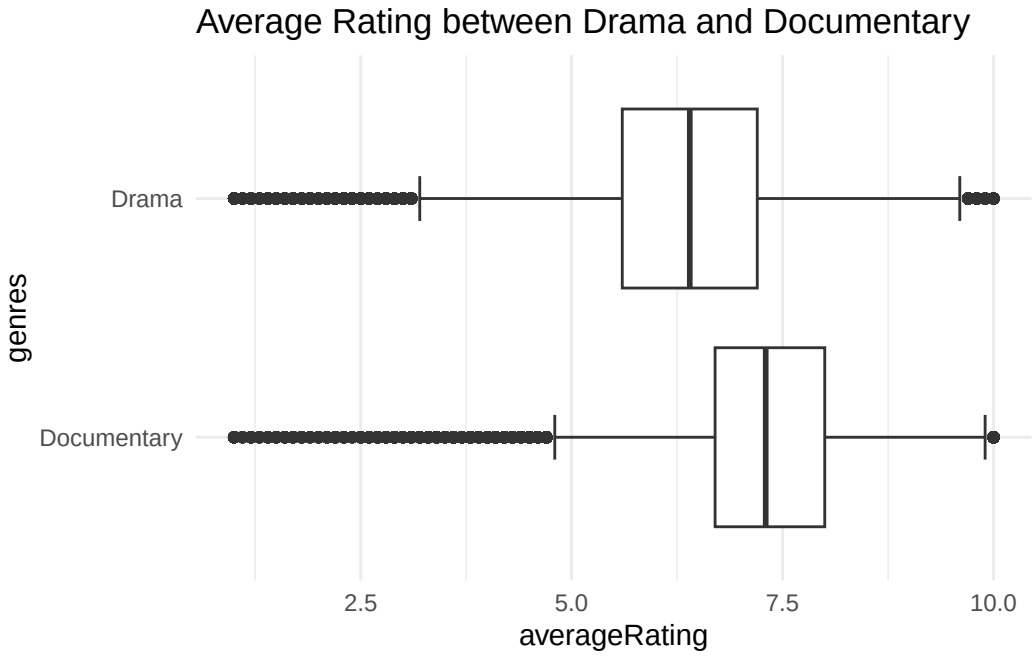
Check normality:



The points on the QQ plot mostly fall along the straight line, so the average ratings across different genres is likely normally distributed.

1. Check the significant difference between average ratings of Drama and Documentary

```
Warning: Removed 233769 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



From the above box plot, we can see that the median of the average ratings between drama and documentary has a significant difference.

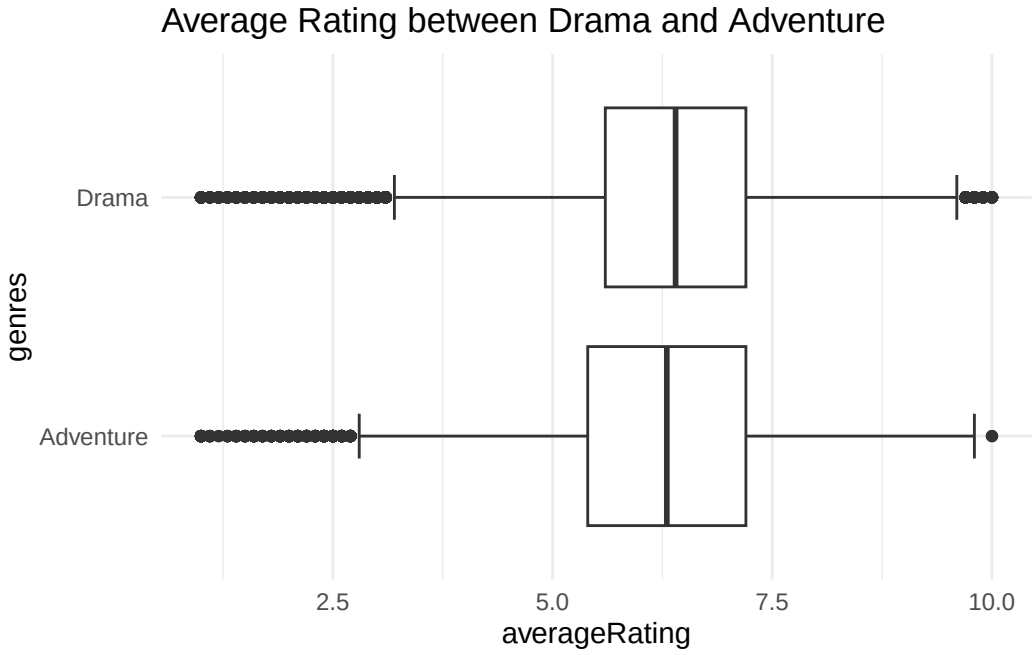
```
    Welch Two Sample t-test

data:  Drama and Documentary
t = -156.35, df = 115068, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8814195 -0.8595948
sample estimates:
mean of x mean of y
 6.364135  7.234642
```

In addition, the p-value from the t-test is less than 0.05, it suggests that there is a statistically significant difference between the average ratings of movies in the Drama genre and those in the Documentary genre.

2. Check the significant difference between average ratings of Drama and Adventure

```
Warning: Removed 144316 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



From the above box plot, it is difficult to distinguish the difference in median average ratings between drama and adventure genres. I will use a t-test to check if there is a statistically difference.

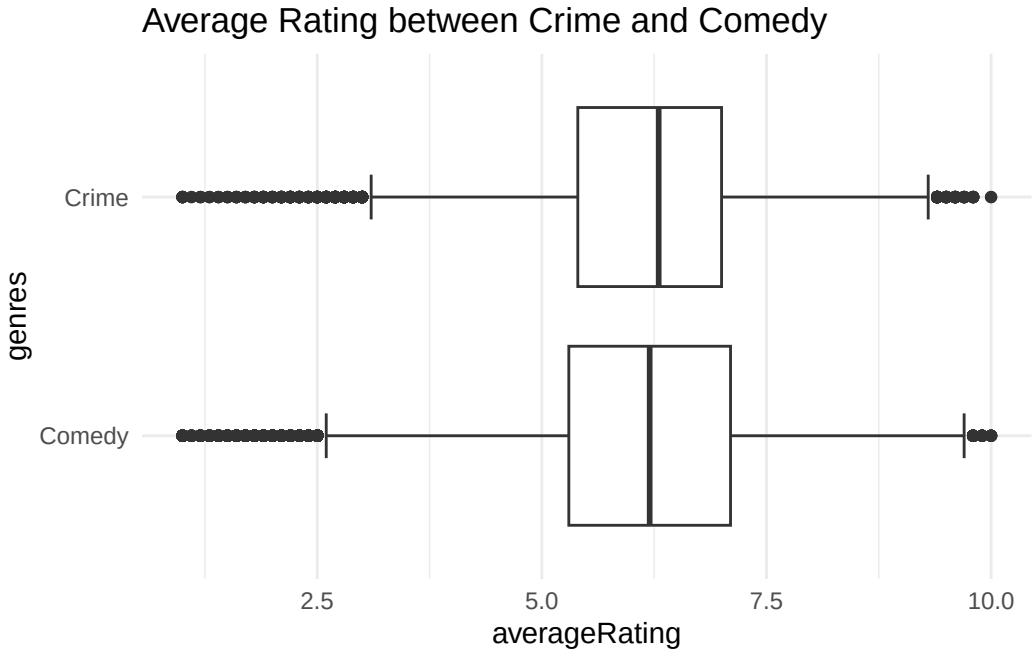
```
    Welch Two Sample t-test

data:  Drama and Adventure
t = 19.066, df = 31895, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1594336 0.1959691
sample estimates:
mean of x mean of y
 6.364135  6.186434
```

The p-value from the t-test is less than 0.05, which indicates there is a statistically significant difference between the average ratings of movies in the Drama genre and those in the Adventure genre.

3. Check the significant difference between average ratings of Crime and Comedy

```
Warning: Removed 85847 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



From the above box plot, it is difficult to distinguish the difference in median average ratings between crime and comedy genres. I will use a t-test to check if there is a statistically difference.

```
    Welch Two Sample t-test

data:  Crime and Comedy
t = 1.5642, df = 61695, p-value = 0.1178
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
 -0.003201194  0.028500618
sample estimates:
mean of x mean of y
 6.187413  6.174763
```

The p-value from the t-test is greater than 0.05, which indicates there is not a statistically significant difference between the average ratings of movies in the Crime genre and those in the Comedy genre.

4. Check the significant difference between average ratings across all different genres

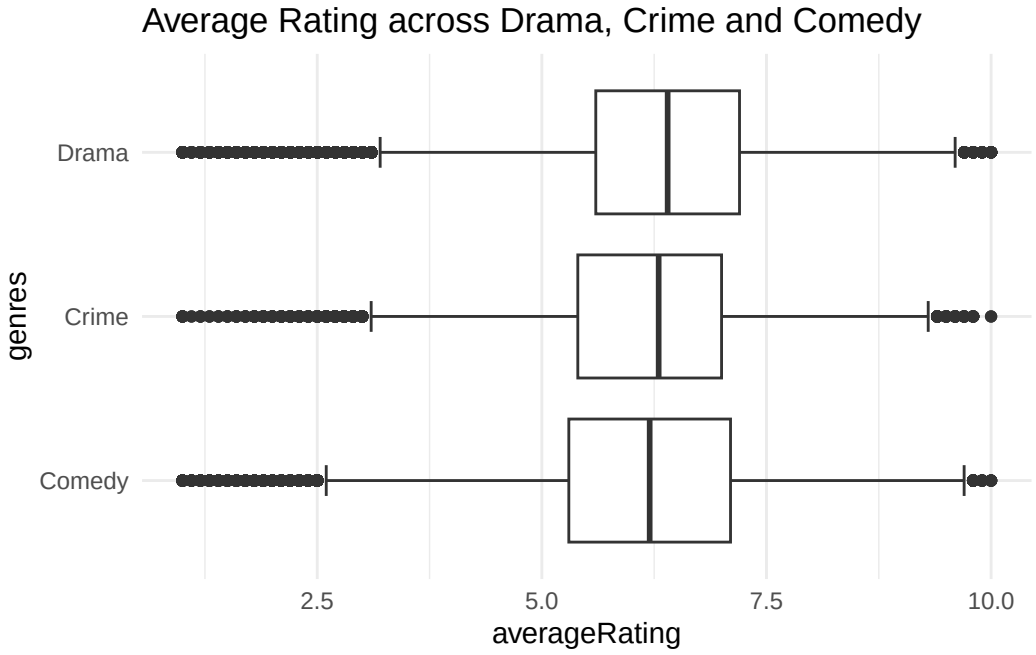
Now, I am going to check if there is a significant difference of average ratings across all different genres.

```
               Df  Sum Sq Mean Sq F value Pr(>F)
genres         28  128126    4576    2592 <2e-16 ***
Residuals   688538 1215546       2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
668310 observations deleted due to missingness
```

From the ANOVA test, the p-value is less than 0.05, so there is a significant difference of average ratings across all different genres.

5. Check the significant difference between average ratings across drama, crime, and comedy

```
Warning: Removed 215661 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



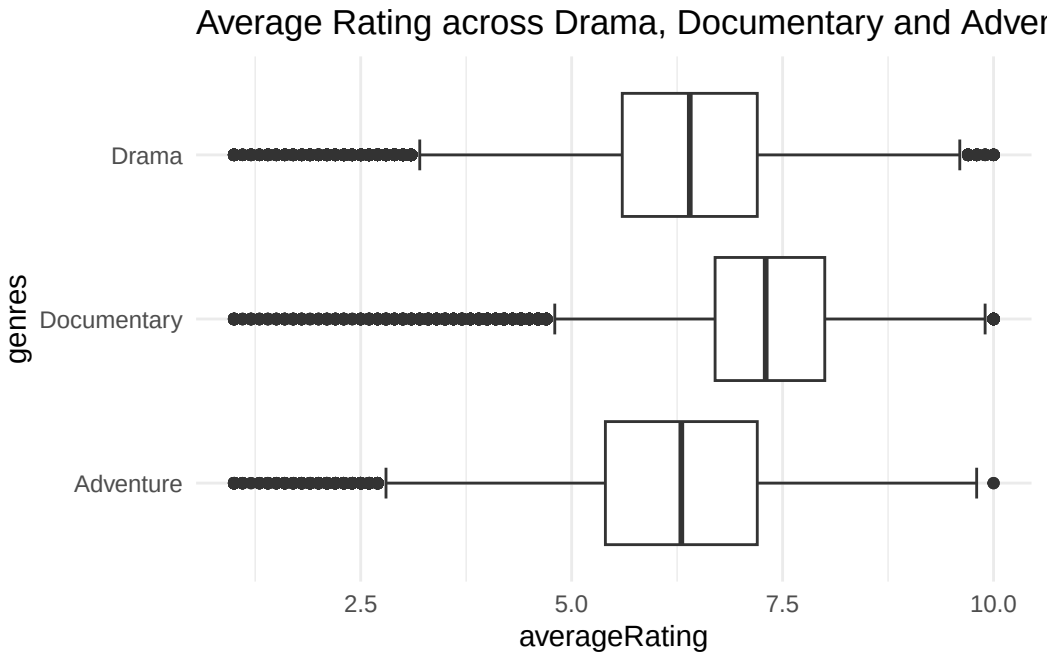
From the above box plot, we can see there are statistically significance of average ratings across these three genres. I will also use ANOVA test to check if my conclusion is correct.

```
                Df Sum Sq Mean Sq F value Pr(>F)
genres           2   2590  1294.8   770.8 <2e-16 ***
Residuals   300254 504326     1.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
215661 observations deleted due to missingness
```

From the ANOVA test, the p-value is less than 0.05, so there is a significant difference of average ratings across these three genres.

6. Check the significant difference between average ratings across drama, documentary, and adventure

```
Warning: Removed 248271 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



From the above box plot, we can see there are statistically significance of average ratings across these three genres. I will also use ANOVA test to check if my conclusion is correct.

```
                Df Sum Sq Mean Sq F value Pr(>F)
genres           2  36780   18390   12049 <2e-16 ***
Residuals   247876 378321       2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
248271 observations deleted due to missingness
```

From the ANOVA test, the p-value is less than 0.05, so there is a significant difference of average ratings across these three genres.

Part 2: Check significant difference between average ratings with genres over year

1. Drama

```
               Df Sum Sq Mean Sq F value Pr(>F)
startYear      120   3089  25.745    16.8 <2e-16 ***
Residuals   163099 249858   1.532
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
129814 observations deleted due to missingness
```

Given that the p-value from the ANOVA test less than 0.05, it suggests a statistically significant difference in average ratings within the drama genre across different years.

2. Crime

```
               Df Sum Sq Mean Sq F value Pr(>F)
startYear     115   1248  10.852   7.103 <2e-16 ***
Residuals   33164  50667   1.528
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
15936 observations deleted due to missingness
```

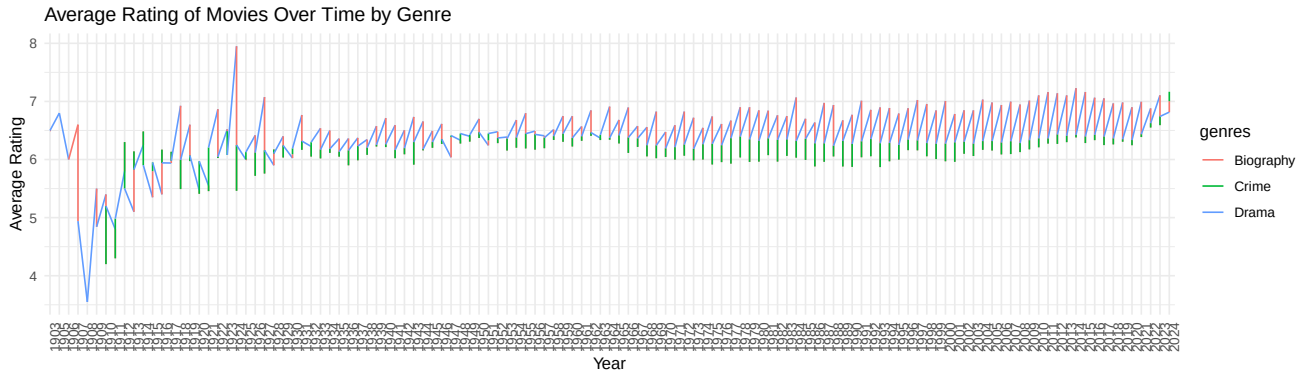
Given that the p-value from the ANOVA test less than 0.05, it suggests a statistically significant difference in average ratings within the crime genre across different years.

3. Biography

```
               Df Sum Sq Mean Sq F value Pr(>F)
startYear     118    454   3.851   3.135 <2e-16 ***
Residuals   10567  12980   1.228
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
9789 observations deleted due to missingness
```

Given that the p-value from the ANOVA test less than 0.05, it suggests a statistically significant difference in average ratings within the crime genre across different years.

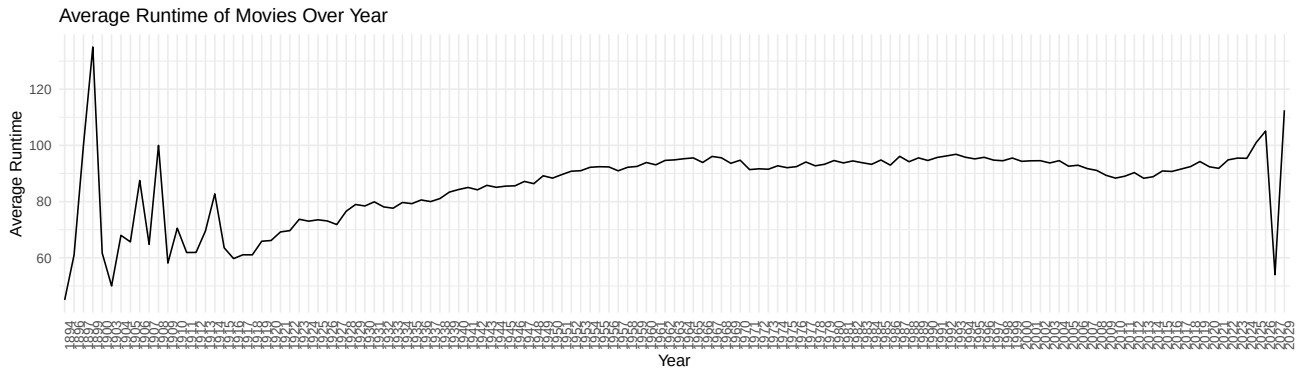
Plot average ratings with these three genres over year



In the line graph, we can observe a significant difference in average ratings among these three genres across different years.

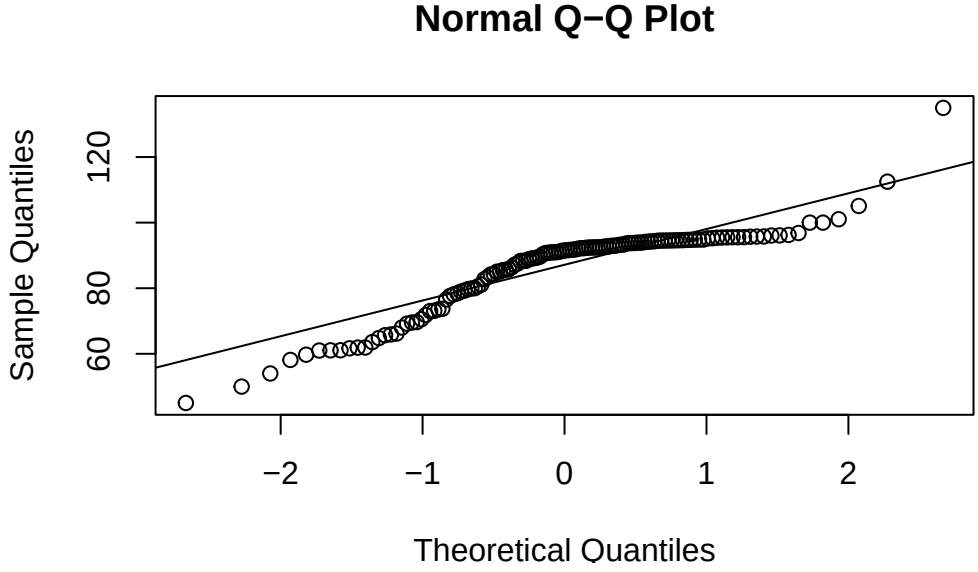
Part 3: The average runtime of movies over year

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



In the line graph, we observe a significant difference in average movie runtime across genres over the years.

I am also going to use ANOVA to check my answer. First, I will check the normality of the data.



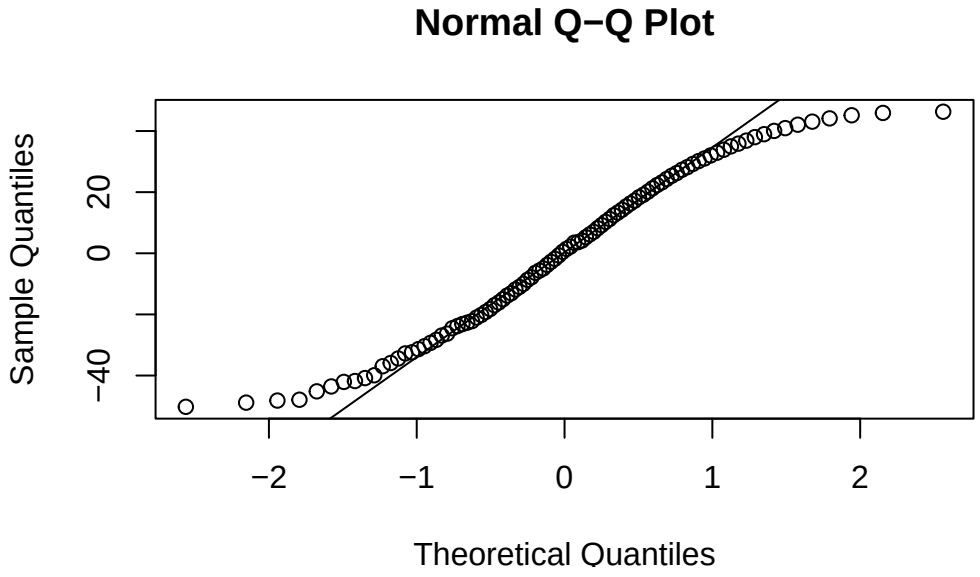
The points on the QQ plot mostly fall along the straight line, so the average runtime of movie is likely normally distributed. Then I can use ANOVA test.

```
                Df Sum Sq Mean Sq F value   Pr(>F)
average_runtime   1  70829   70829   77.14 8.39e-15 ***
Residuals       129 118447     918
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

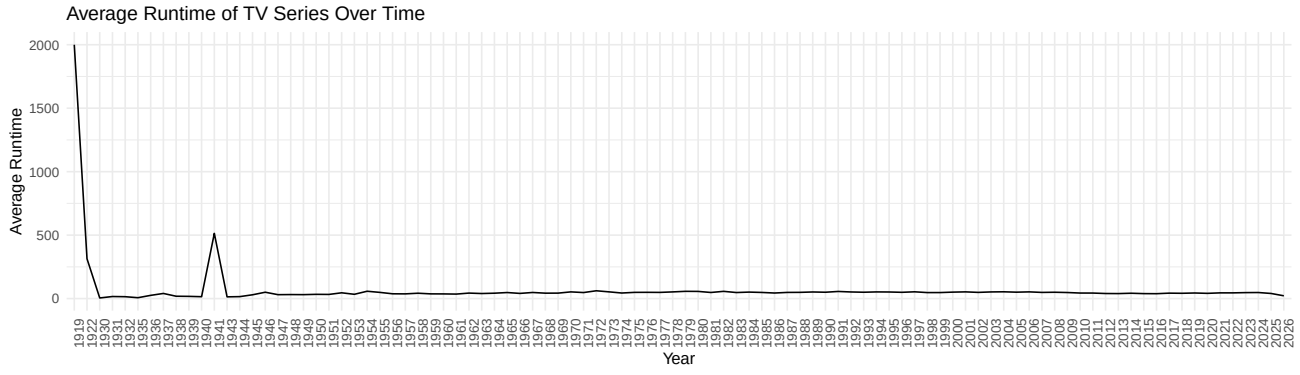
Given that the p-value from the ANOVA test less than 0.05, it also suggests a statistically significant difference in average movie runtime over year.

Part 4: Check if episode lengths (of TV Series) appear to have gotten longer over time

```
                Df Sum Sq Mean Sq F value Pr(>F)
average_runtime  1   4300    4300   5.529 0.0208 *
Residuals       94  73107     778
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

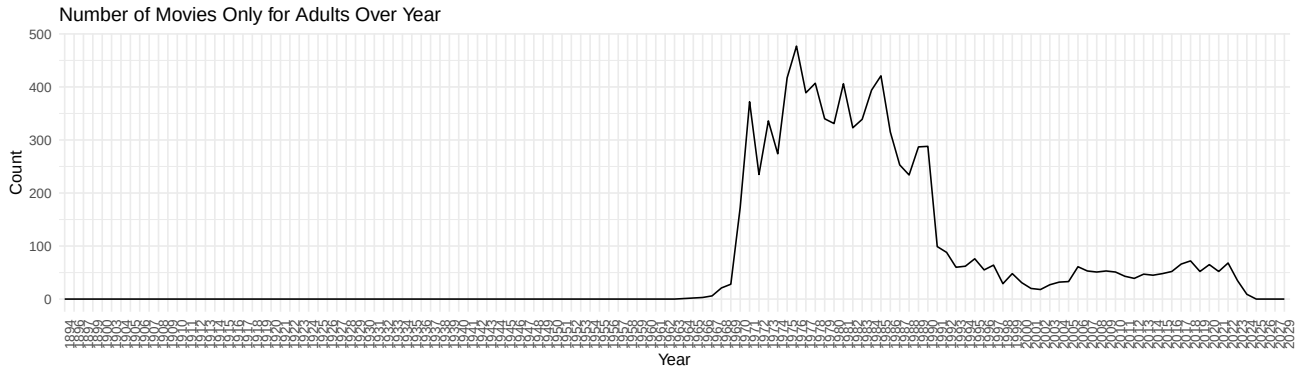


Given that the p-value from the ANOVA test less than 0.05, it suggests a statistically significant difference in episode lengths (of TV Series) over year. The QQ-plot of the residuals also holds the normality assumption. However, it cannot tell us if the episode lengths of TV series have gotten longer over time. Thus, I am going to create a line graph to figure out the pattern.



It is clearly from the line graph that the lengths of TV series episodes experienced a sudden decrease after 1919, followed by an sudden increase in 1941. However, they subsequently declined again and remained relatively constant in the following years. Thus the episode length did not appear to gotten longer over time.

Part 5: Check if the number of movie only for adult increase over year



From the line graph, it is clear that the number of movies only for adults increased suddenly from 1966 to 1991. However, after that time period, the number of movies only for adults has dropped.

Appendix

```
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)

basics <- read_csv("/home/jovyan/100-sp24/Mini_Projects/MP02/data/basics.csv")
ratings <- read_csv("/home/jovyan/100-sp24/Mini_Projects/MP02/data/ratings.csv")

genres_split <- basics %>%
  mutate(genres = strsplit(as.character(genres), ",")) %>%
  unnest(genres)

new_data <- left_join(genres_split, ratings, by = "tconst")


# Part 1: Check significant difference between average ratings across different genres

Check normality:

qqnorm(new_data$averageRating)
qqline(new_data$averageRating)

The points on the QQ plot mostly fall along the straight line, so the average ratings across dif

## 1. Check the significant difference between average ratings of Drama and Documentary


Drama_Doc <- new_data %>%
  filter(genres %in%  c("Drama","Documentary"))
```

```
Drama_Doc_plot <- Drama_Doc %>%
  ggplot(aes(x = averageRating, y =genres)) +
  geom_boxplot(staplewidth = 0.25) +
  theme_minimal() +
  ggtitle("Average Rating between Drama and Documentary")

print(Drama_Doc_plot)
```

From the above box plot, we can see that the median of the average ratings between drama and doc

```
Drama <- filter(Drama_Doc, genres == "Drama")$averageRating
Documentary <- filter(Drama_Doc, genres == "Documentary")$averageRating

t_test_1 <- t.test(Drama, Documentary, alternative = "two.sided")
print(t_test_1)
```

In addition, the p-value from the t-test is less than 0.05, it suggests that there is a statisti

2. Check the significant difference between average ratings of Drama and Adventure

```
Drama_Adv <- new_data %>%
  filter(genres %in%  c("Drama","Adventure"))

Drama_Adv_plot <- Drama_Adv %>%
  ggplot(aes(x = averageRating, y = genres)) +
  geom_boxplot(staplewidth = 0.25) +
  theme_minimal() +
  ggtitle("Average Rating between Drama and Adventure")

print(Drama_Adv_plot)
```

From the above box plot, it is difficult to distinguish the difference in median average ratings

```
Adventure <- filter(Drama_Adv, genres == "Adventure")$averageRating

t_test_2 <- t.test(Drama, Adventure, alternative = "two.sided")
print(t_test_2)
```

The p-value from the t-test is less than 0.05, which indicates there is a statistically signific

3. Check the significant difference between average ratings of Crime and Comedy

```
Cri_Com <- new_data %>%
```

```
  filter(genres %in%  c("Crime","Comedy"))

Cri_Com_plot <- Cri_Com %>%
  ggplot(aes(x = averageRating, y =genres)) +
  geom_boxplot(staplewidth = 0.25) +
  theme_minimal() +
  ggtitle("Average Rating between Crime and Comedy")

print(Cri_Com_plot)
```

From the above box plot, it is difficult to distinguish the difference in median average ratings

```
Crime <- filter(Cri_Com, genres == "Crime")$averageRating
Comedy <- filter(Cri_Com, genres == "Comedy")$averageRating

t_test_3 <- t.test(Crime, Comedy, alternative = "two.sided")
print(t_test_3)
```

The p-value from the t-test is greater than 0.05, which indicates there is not a statistically s

4. Check the significant difference between average ratings across all different genres

Now, I am going to check if there is a significant difference of average ratings across all diff

```
anova_result_1 <- aov(averageRating ~ genres, data = new_data)
summary(anova_result_1)
```

From the ANOVA test, the p-value is less than 0.05, so there is a significant difference of aver

5. Check the significant difference between average ratings across drama, crime, and comedy

```
Dra_Cri_Com <- new_data %>%
  filter(genres %in%  c("Drama","Crime","Comedy"))

Dra_Cri_Com_plot <- Dra_Cri_Com %>%
  ggplot(aes(x = averageRating, y =genres)) +
  geom_boxplot(staplewidth = 0.25) +
  theme_minimal() +
  ggtitle("Average Rating across Drama, Crime and Comedy")

print(Dra_Cri_Com_plot)
```

From the above box plot, we can see there are statistically significance of average ratings acro

```
result_1 <- aov(averageRating ~ genres, data = Dra_Cri_Com)
```

```
summary(result_1)

From the ANOVA test, the p-value is less than 0.05, so there is a significant difference of aver

## 6. Check the significant difference between average ratings across drama, documentary, and ad


Dra_Doc_Adv <- new_data %>%
  filter(genres %in%  c("Drama","Documentary","Adventure"))

Dra_Doc_Adv_plot <- Dra_Doc_Adv %>%
  ggplot(aes(x = averageRating, y =genres)) +
  geom_boxplot(staplewidth = 0.25) +
  theme_minimal() +
  ggtitle("Average Rating across Drama, Documentary and Adventure")

print(Dra_Doc_Adv_plot)

From the above box plot, we can see there are statistically significance of average ratings acro

result_2 <- aov(averageRating ~ genres, data = Dra_Doc_Adv)
summary(result_2)

From the ANOVA test, the p-value is less than 0.05, so there is a significant difference of aver

\newpage

# Part 2: Check significant difference between average ratings with genres over year

## 1. Drama

Drama_data <- filter(Drama_Doc, genres == "Drama")
anova_result_2 <- aov(averageRating ~ startYear, data = Drama_data)
summary(anova_result_2)

Given that the p-value from the ANOVA test less than 0.05, it suggests a statistically significa

## 2. Crime

Crime_data <- filter(new_data, genres == "Crime")
anova_result_3 <- aov(averageRating ~ startYear, data = Crime_data)
summary(anova_result_3)

Given that the p-value from the ANOVA test less than 0.05, it suggests a statistically significa

## 3. Biography

Biography_data <- filter(new_data, genres == "Biography")
```

```
anova_result_4 <- aov(averageRating ~ startYear, data = Biography_data)
summary(anova_result_4)

Given that the p-value from the ANOVA test less than 0.05, it suggests a statistically significa

## Plot average ratings with these three genres over year


new_data$averageRating <- as.numeric(new_data$averageRating)

ratings_num <- new_data %>%
  filter(!is.na(averageRating) & startYear != "\\N" & !is.na(startYear))

ratings_data <- ratings_num %>%
  filter(genres %in% c("Biography", "Crime", "Drama")) %>%
  group_by(startYear, genres) %>%
  summarise(average_rate = mean(averageRating, na.rm = TRUE))

ggplot(ratings_data, aes(x = startYear, y = average_rate, color = genres)) +
  geom_line(group = 1) +
  labs(x = "Year", y = "Average Rating", title = "Average Rating of Movies Over Time by Genre")
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

In the line graph, we can observe a significant difference in average ratings among these three

\newpage
# Part 3: The average runtime of movies over year


new_data$runtimeMinutes <- as.numeric(new_data$runtimeMinutes)

numeric_data <- new_data %>%
  filter(!is.na(runtimeMinutes) & startYear != "\\N" & !is.na(startYear))

run_time_data <- numeric_data %>%
  filter(titleType == "movie") %>%
  group_by(startYear) %>%
  summarise(average_runtime = mean(runtimeMinutes, na.rm = T))

ggplot(run_time_data, aes(x = startYear, y = average_runtime)) +
  geom_line(group = 1) +
  labs(x = "Year", y = "Average Runtime", title = "Average Runtime of Movies Over Year") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))


In the line graph, we observe a significant difference in average movie runtime across genres ov
```

```
I am also going to use ANOVA to check my answer. First, I will check the normality of the data.

qqnorm(run_time_data$average_runtime)
qqline(run_time_data$average_runtime)

The points on the QQ plot mostly fall along the straight line, so the average runtime of movie i


anova_result_5 <- aov(startYear ~ average_runtime, run_time_data)
summary(anova_result_5)

Given that the p-value from the ANOVA test less than 0.05, it also suggests a statistically sign

\newpage
# Part 4: Check if episode lengths (of TV Series) appear to have gotten longer over time


new_data$runtimeMinutes <- as.numeric(new_data$runtimeMinutes)

numeric_data <- new_data %>%
  filter(!is.na(runtimeMinutes) & startYear != "\\N" & !is.na(startYear))

TV <- numeric_data %>%
  filter(titleType == "tvSeries") %>%
  group_by(startYear) %>%
  summarise(average_runtime = mean(runtimeMinutes, na.rm = T))

anova_result_6 <- aov(startYear ~ average_runtime, TV)
summary(anova_result_6)



qqnorm(resid(anova_result_6))
qqline(resid(anova_result_6))


Given that the p-value from the ANOVA test less than 0.05, it suggests a statistically significa


ggplot(TV, aes(x = startYear, y = average_runtime)) +
  geom_line(group = 1) +
  labs(x = "Year", y = "Average Runtime", title = "Average Runtime of TV Series Over Time") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))


It is clearly from the line graph that the lengths of TV series episodes experienced a sudden de
```

```
\newpage
# Part 5: Check if the number of movie only for adult increase over year


mov_adult <- numeric_data %>%
  filter(titleType == "movie") %>%
  group_by(startYear) %>%
  summarise(count = sum(isAdult == 1, na.rm = TRUE))

ggplot(mov_adult, aes(x = startYear, y = count)) +
  geom_line(group= 1) +
  labs(x = "Year", y = "Count", title = "Number of Movies Only for Adults Over Year") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))


From the line graph, it is clear that the number of movies only for adults increased suddenly fr
```