Mini Project 2

Albert Chui (albertchui) Joseph Zaki (josephzaki)

2024-05-08

Table of contents

1 Abstract

In this paper we explore data from the official IMdB Documentation detailing 942007 movies. We find that the average rating between genres varies between 5 and 7 points. For the ratings of selected genres of news, reality-tv, horror, documentary we see that documentaries are relativly stable, horror has been decreasing, reality-tv has been increasing, and news has been extremly variable throughout the years. In addition, runtime of movies has been increasing over the years, to a current 2024 average of 120 minutes. The runtime of episodes in tv series had been increasing until 24 years ago, when it has suddenly started decreasing until it hit 40 minutes, then bounced back to 50 minutes.

2 Dataset Examination

Let us first look at all of our unique genres in this dataset:

genres	avg_rating	num_movies
Documentary	7.235	163250
Short	7.126	4243
Biography	6.975	20475
History	6.927	20115
Sport	6.767	14821
Animation	6.754	25459
Music	6.726	23800
Family	6.503	34319
War	6.47	10493
Film-Noir	6.467	882
Reality-TV	6.385	24960
Drama	6.364	293034
News	6.351	10100
Talk-Show	6.269	22822
Musical	6.247	12137
Romance	6.238	60294
Crime	6.187	49216
Adventure	6.186	39866
Fantasy	6.176	21801
Comedy	6.175	173668
Game-Show	6.128	8983
Mystery	6.094	23625
Western	5.954	8729
Action	5.945	66914
Thriller	5.729	51903

genres	avg_rating	num_movies
Sci-Fi	5.671	19062
Adult	5.561	11519
Horror	5.134	44503

3 Testing for Differences in Mean Rating Across Genres

In this section we will perform several hypothesis tests on different genres to determine if there is a difference in the mean rating based on genre.

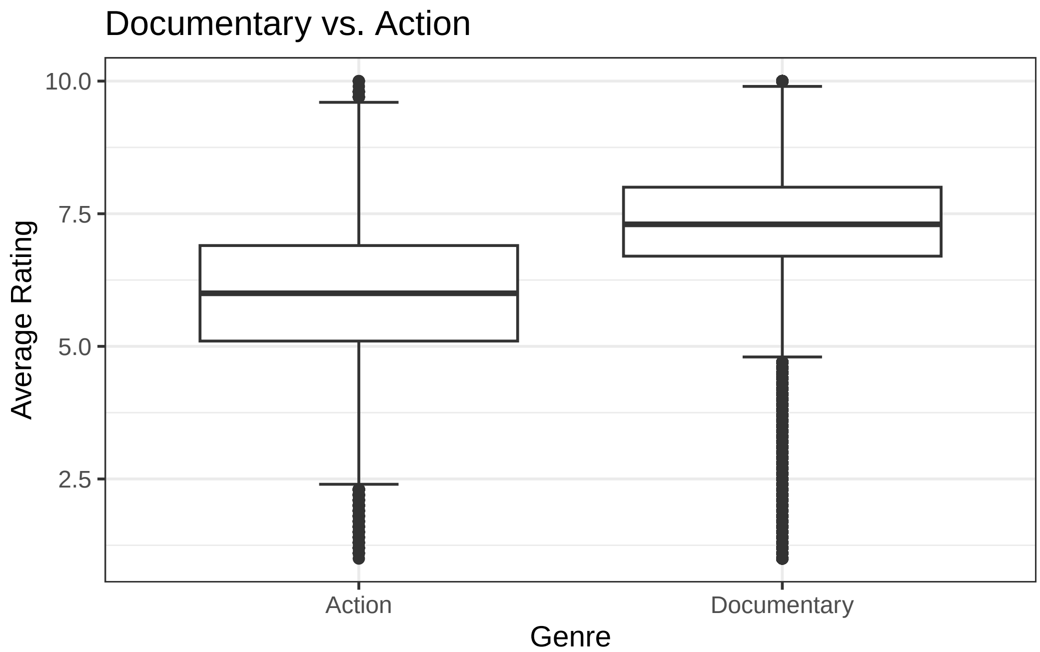
3.1 Documentary VS Action

Here, we specifically test for a difference in mean rating between documentaries and action programs. To do this we test the following hypotheses:

$$\begin{cases} H_0 : \mu_{documentary} = \mu_{action} \\ H_a : \mu_{documentary} \neq \mu_{action} \end{cases}$$

```
    Welch Two Sample t-test

data:  movies %>% filter(genres == "Documentary") %>% pull(averageRating) and movies %>%
t = 149.26, df = 68285, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.272481 1.306345
sample estimates:
mean of x mean of y
 7.234642  5.945229
```



From the results of our two-sample t-test we can see that our p-value is smaller than 0.05, which means there is a significant difference in mean ratings across Documentaries and Action movies/series. From the boxplot, we can see that the mean rating for both categories seems to clearly differ.

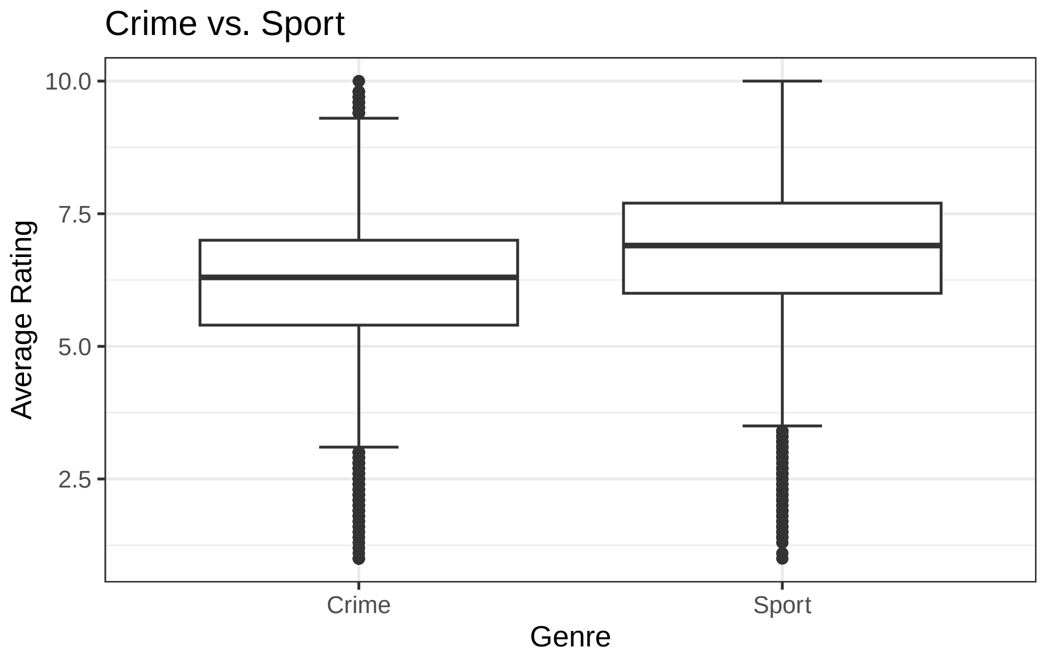
3.2 Crime vs. Sports

Here, we specifically test for a difference in mean rating between crime and sports programs. To do this we test the following hypotheses:

$$\begin{cases} H_0 : \mu_{crime} = \mu_{sports} \\ H_a : \mu_{crime} \neq \mu_{sports} \end{cases}$$

```
    Welch Two Sample t-test

data:  movies %>% filter(genres == "Crime") %>% pull(averageRating) and movies %>% filte
t = -31.313, df = 7909.9, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6153849 -0.5428749
sample estimates:
mean of x mean of y
 6.187413  6.766543
```



From the results of our two-sample t-test we can see that our p-value is smaller than 0.05, which means there is a significant difference in mean ratings across Crime and Sports movies/series. From the boxplot, we can see that the mean rating for both categories differs slightly, but only through our t-test can we tell that this is a significant differnece.

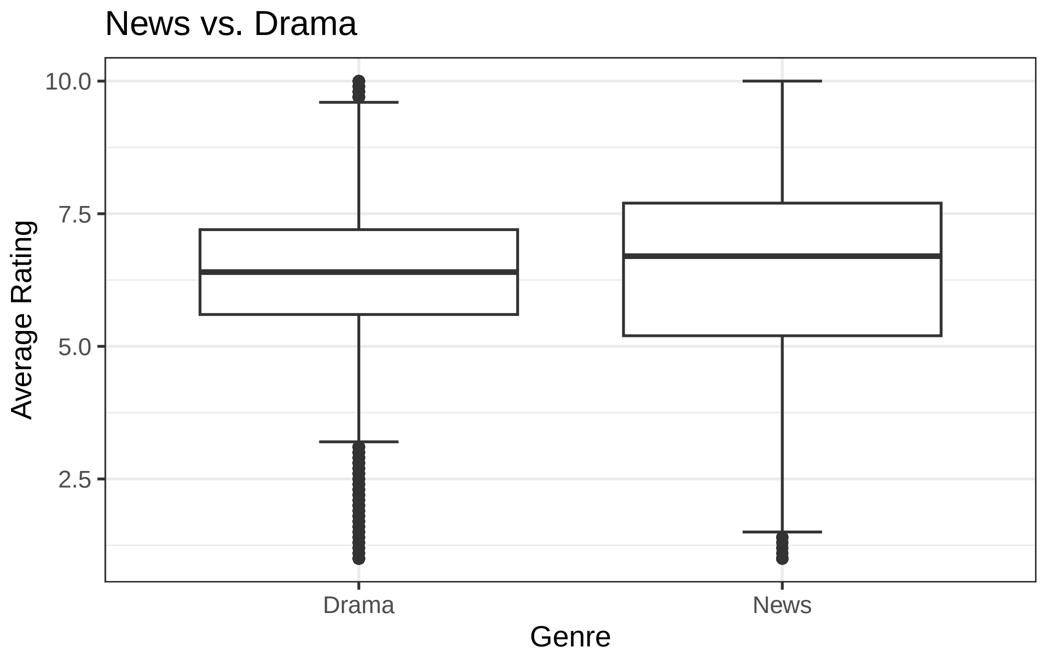
3.3 News vs Drama

Here, we specifically test for a difference in mean rating between News and Drama programs. To do this we test the following hypotheses:

$$\begin{cases} H_0 : \mu_{news} = \mu_{drama} \\ H_a : \mu_{news} \neq \mu_{drama} \end{cases}$$

```
    Welch Two Sample t-test

data:  movies %>% filter(genres == "News") %>% pull(averageRating) and movies %>% filter
t = -0.37348, df = 2709.3, p-value = 0.7088
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.08026006  0.05457742
sample estimates:
mean of x mean of y
 6.351294  6.364135
```



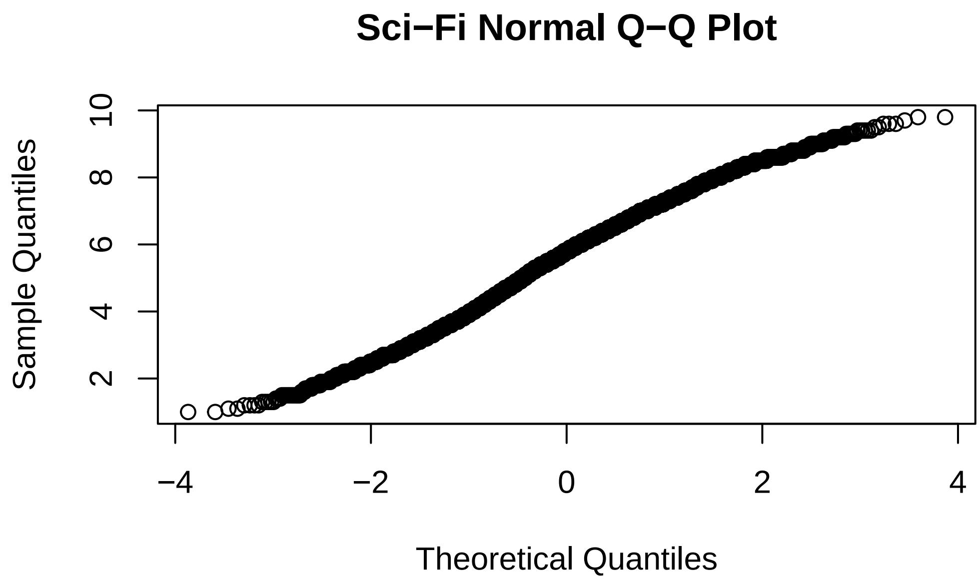
From the results of our two-sample t-test we can see that our p-value is larger than 0.05, which means there is not a significant difference in mean ratings across Crime and Sports movies/series. From the boxplot, we can see that the mean rating for both categories is very similar, appearing almost identical.

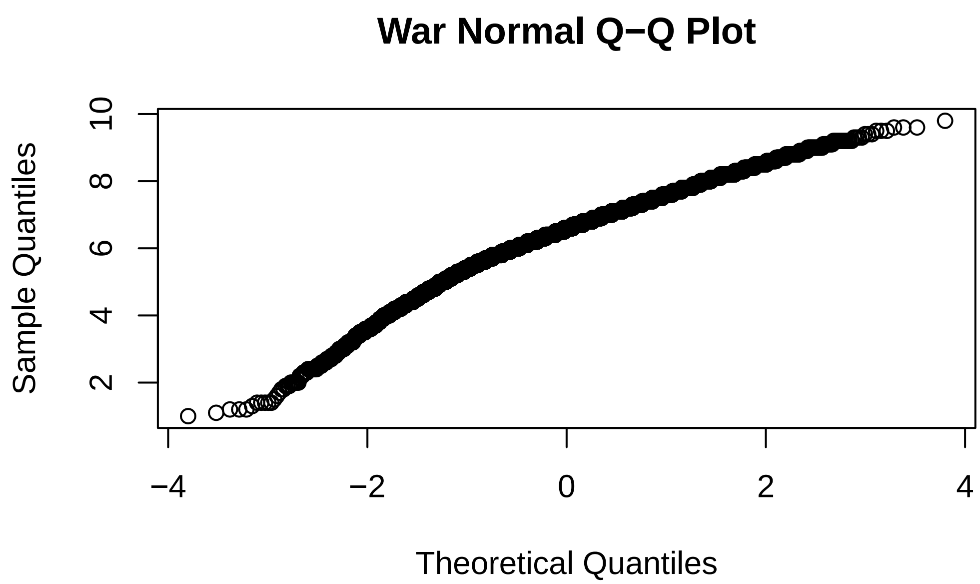
3.4 Romance vs Sci-Fi vs War vs Western

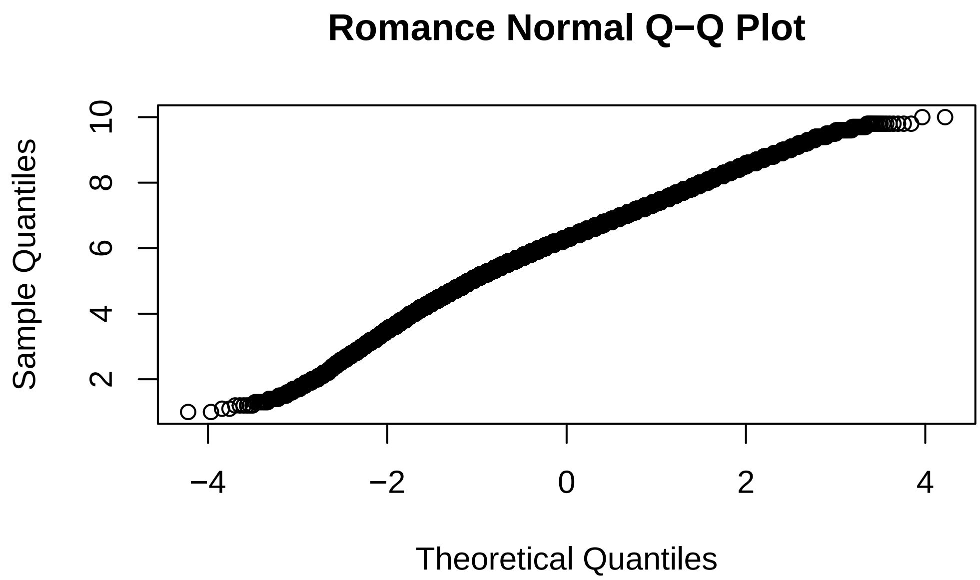
In this section we test for differences in mean ratings across groups larger than 2 genres, meaning we need to perform an Analysis of Variance test.

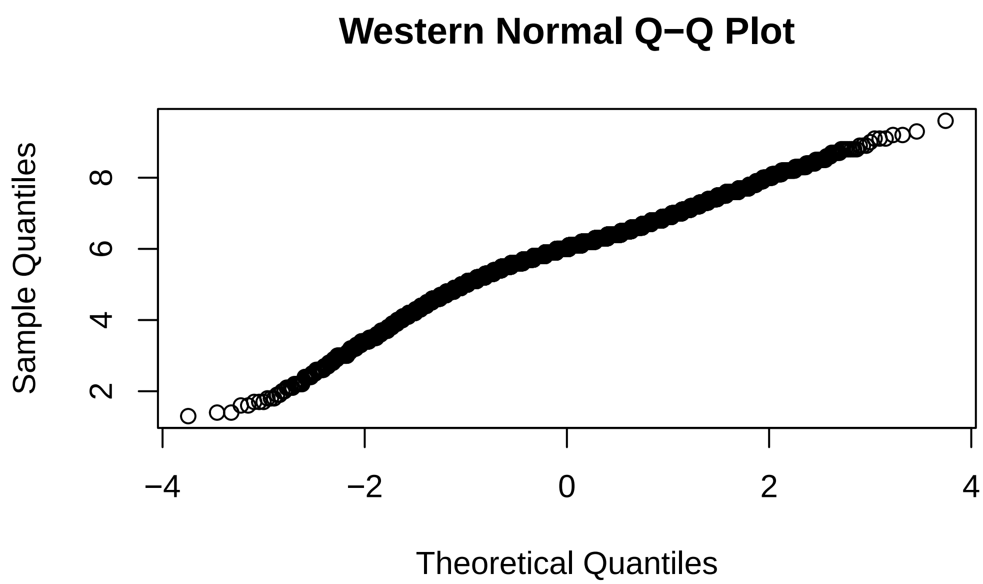
3.4.1 Check Assumptions

1) **Normality**


Sci−Fi Normal Q−Q Plot
Sample Quantiles
2
4
6
8
10
−4
−2
0
2
4
Theoretical Quantiles

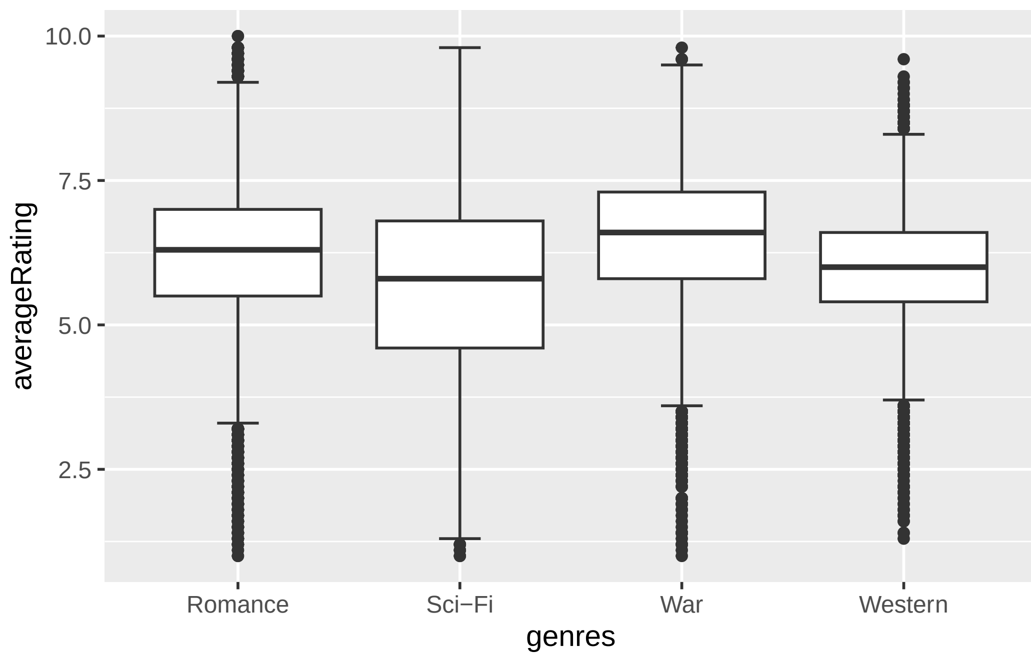

War Normal Q−Q Plot
Sample Quantiles
2
4
6
8
10
−4
−2
0
2
4
Theoretical Quantiles





From the above Q-Q Plots, we can see that our normality assumption holds for all four genres.

2) Homoscedasticity (Constant Variance)



From the above boxplots, we can see that the assumption of Homoscedasticity holds for this dataset as all four genres seem to have approximately the same variance.

3) Independence

Although we cannot test for independence, it is unlikely that the ratings for one movie or series strongly effects the rating for a different movie or series, meaning our assumption of Independence is most likely not violated.

3.5 Conduct ANOVA Test

$$\begin{cases} H_0 : \mu_{war} = \mu_{romance} = \mu_{western} = \mu_{sci-fi} \\ H_a : \text{At least one mean differs from the others} \end{cases}$$

```
               Df Sum Sq Mean Sq F value Pr(>F)
genres          3   3328  1109.2   696.8 <2e-16 ***
Residuals   62743  99875     1.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

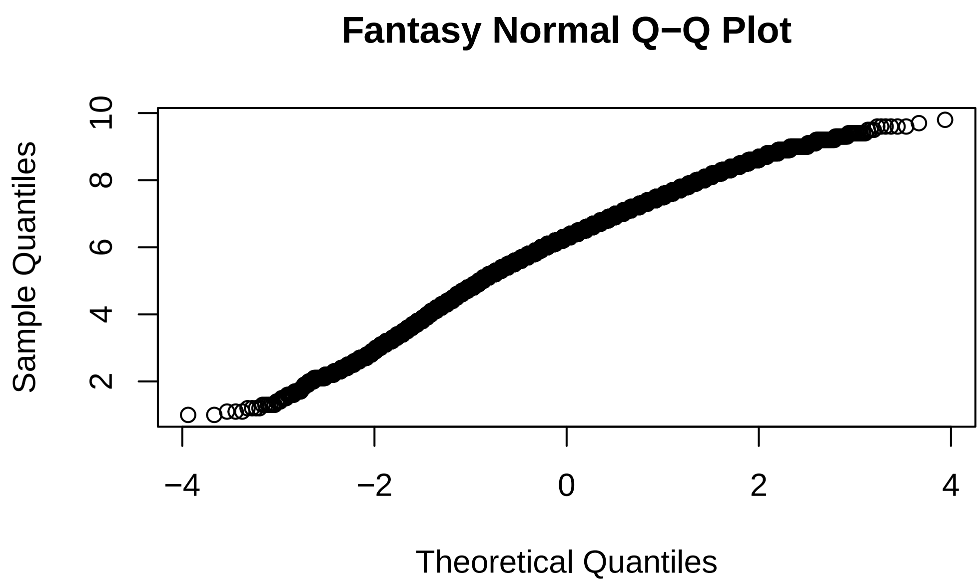
From the results of our ANOVA test above, we can see that our p-value is very small, much smaller than 0.05, meaning we can reject $H_0 : \mu_{war} = \mu_{romance} = \mu_{western} = \mu_{sci-fi}$, indicating there is a significant difference in the mean ratings for at least one of the tested genres (Romance, War, Western, Sci-Fi).

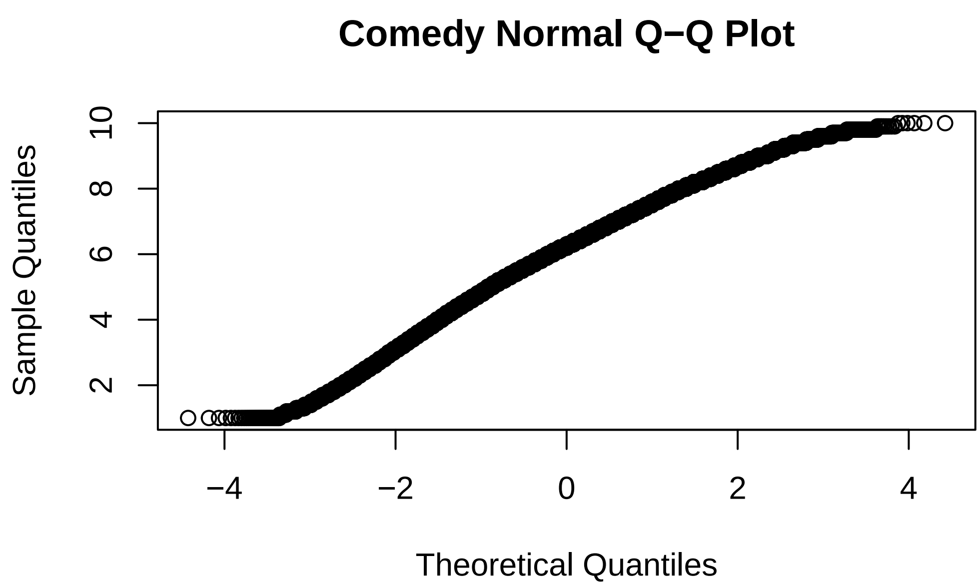
3.6 Fantasy vs Comedy vs Horror

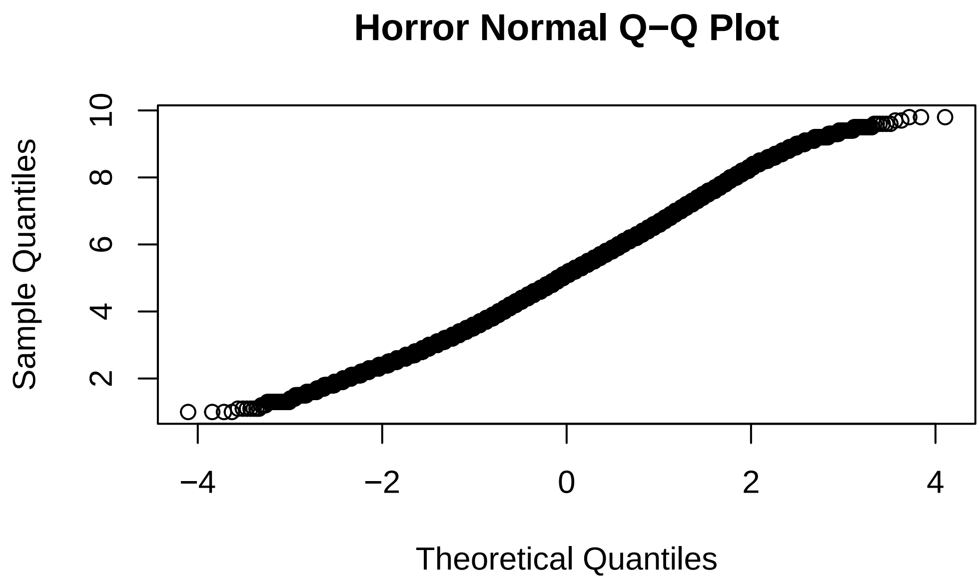
In this section we test for differences in mean ratings across groups larger than 2 genres, meaning we need to perform an Analysis of Variance test.

3.6.1 Check Assumptions

1) **Normality**

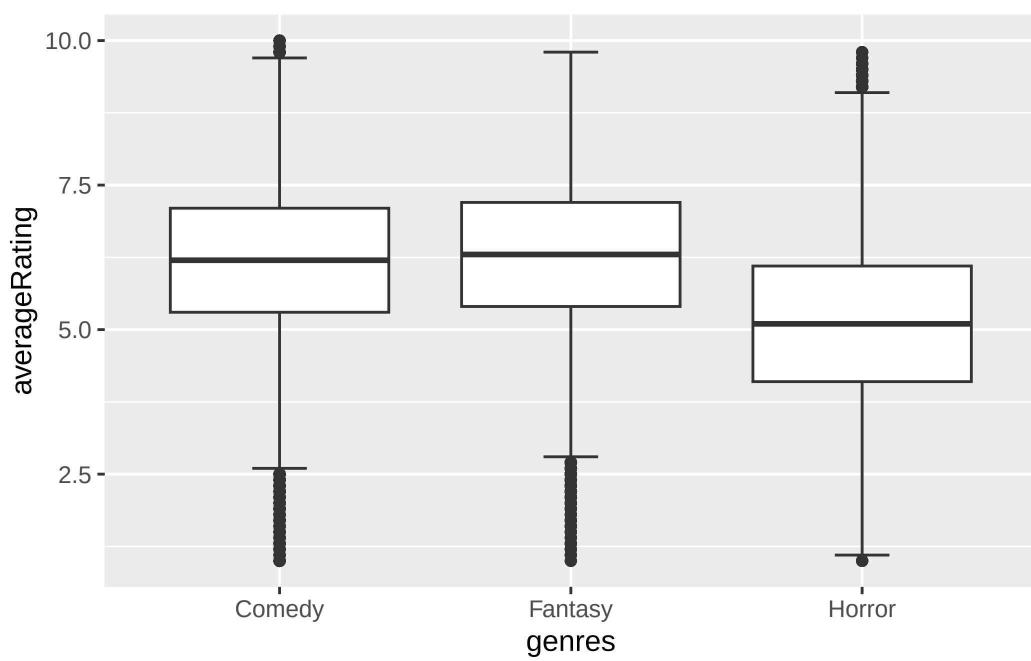






From the above Q-Q Plots, we can see that our normality assumption holds for all three genres.

2) Homoscedasticity (Constant Variance)



From the above boxplots, we can see that the assumption of Homoscedasticity holds for this dataset as all three genres seem to have approximately the same variance.

3) Independence

Although we cannot test for independence, it is unlikely that the ratings for one movie or series strongly effects the rating for a different movie or series, meaning our assumption of Independence is most likely not violated.

3.7 Conduct ANOVA Test

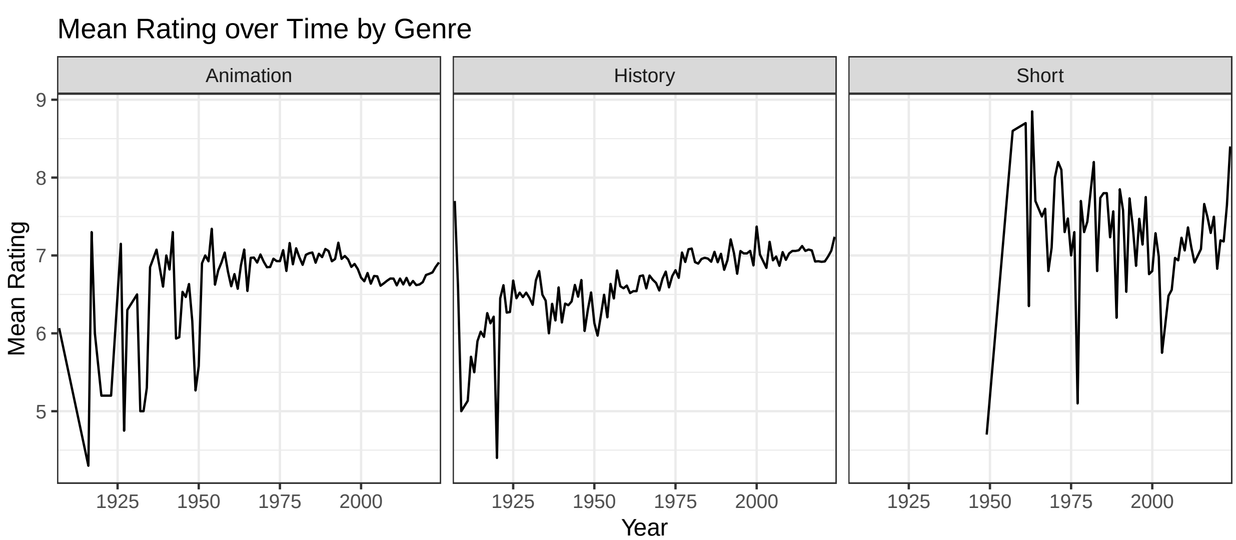
$$\begin{cases} H_0 : \mu_{fantasy} = \mu_{comedy} = \mu_{horror} \\ H_a : \text{At least one mean differs from the others} \end{cases}$$

```
                Df Sum Sq Mean Sq F value Pr(>F)
genres           2  22058   11029    5566 <2e-16 ***
Residuals   140661 278717       2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results of our ANOVA test above, we can see that our p-value is very small, much smaller than 0.05, meaning we can reject $H_0 : \mu_{fantasy} = \mu_{comedy} = \mu_{horror}$, indicating there is a significant difference in the mean ratings for at least one of the tested genres (Fantasy, Comedy, Horror).

4 Testing for Differences in Mean Ratings Within Genres Across Years

In this section we investigate if there are significant changes to the mean ratings of several genres across different years. We focus on the genres of History, Short Films, and Animation.



From the above graph displaying mean ratings in each year by genre, we can already see that the mean rating within each genre seems to vary greatly each year and tends to oscillate especially before 1950. In order to statistically test for difference in mean rating across years, we conducted an ANOVA test for each genre.

4.1 Conduct ANOVA Test for Shorts, History, and Animation Mean Ratings Across Years

4.1.1 Check Assumptions

Since we have tested the assumption of Normality, Homoscedasticity, and Independence for other genres and found that they hold, it is reasonable to assume that they hold for these genres as well since all the data is coming from the same source.

4.2 Conduct ANOVA Test

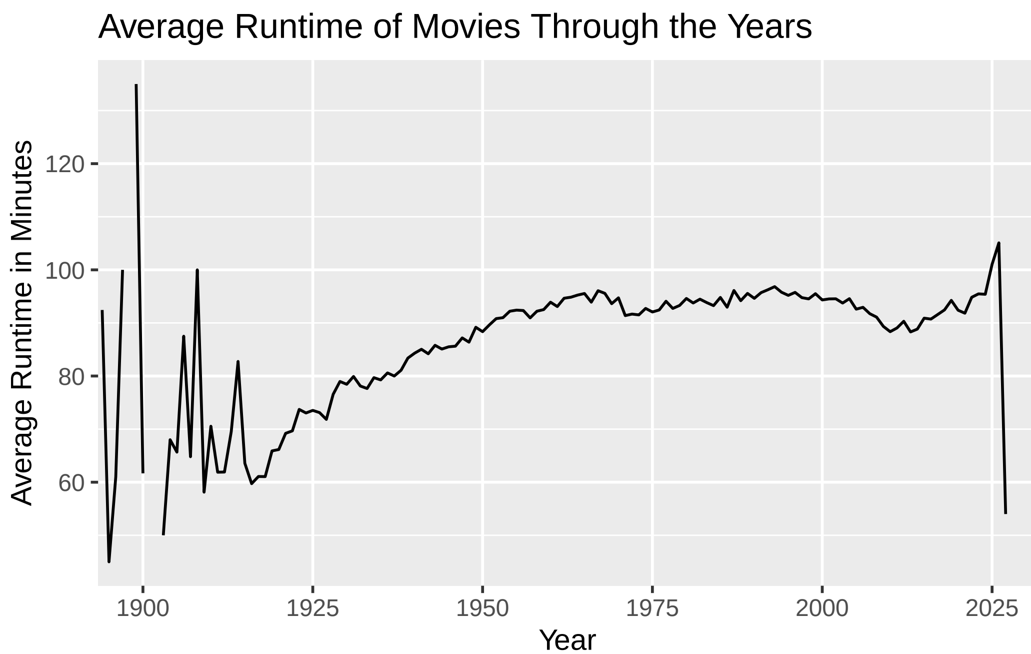
$$\begin{cases} H_0 : \mu_i = \mu_j ; \text{for } i, j \in \text{Years and } i \neq j \\ H_a : \text{At least one mean differs from the others} \end{cases}$$

```
               Df Sum Sq Mean Sq F value Pr(>F)
startYear     117    551   4.711   3.199 <2e-16 ***
Residuals   27061  39856   1.473
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results of our ANOVA test above, we can see that our p-value is very small, much smaller than 0.05, meaning we can reject $H_0: *{i}* = {j} ; for i,j Years and i j $, indicating there is a significant difference in the mean ratings within genres across years for the tested genres (Shorts, Animation, and History).

5 Testing for differences between Runtime and Years for Movies and TV series

5.1 Movies



Visually, it seems that average movie length have not varied much year to year after to 1950. Prior to 1950, mean length varied greatly year to year. They have remained constant after 1950 at run times of 90 to 100 minutes. We can see that the mean run time in 2024 appears very low, this could be caused by the fact that the year is not over and many movies have not yet been released.

5.1.1 Conduct ANOVA Test For Mean Runtime of Movies

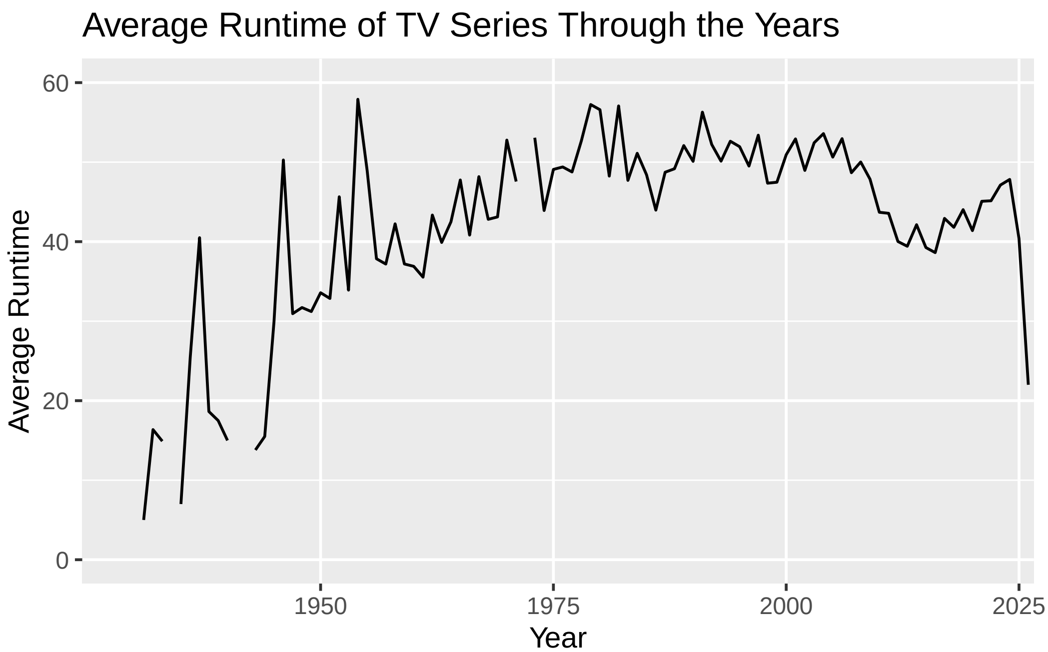
$$\begin{cases} H_0 : \mu_i = \mu_j ; \text{for } i, j \in \text{Years and } i \neq j \\ H_a : \text{At least one mean differs from the others} \end{cases}$$

```
                Df    Sum Sq Mean Sq F value Pr(>F)
startYear      131 1.645e+07  125557   13.31 <2e-16 ***
Residuals   687772 6.486e+09    9431
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA Test results, we find a p-value much lower than 0.05, meaning we can reject $H_0 : \mu_i = \mu_j ; \text{for } i, j \in \text{Years and } i \neq j$ indicating that the runtimes of movies do change significantly year to year.

5.2 TV Series

```
ggplot(series_timed_runlength)+
  geom_line(aes(x=`startYear`, y = `avg_runtime`), group = 1) +
  labs(x = "Year", y = "Average Runtime", title = "Average Runtime of TV Series Throug
  scale_x_discrete(breaks = seq(0, max(series_timed_runlength$startYear), by = 25)) +
  ylim(0, 60)
```



From the plot, we can see that prior to 1950, mean episode runtime varied greatly year to year. They have stabilized after 1950 at runtimes of around 50 minutes. We can see that the mean run time in 2024 appears very low, this could be caused by the fact that the year is not over and many TV Series have not yet been released. Overall, it does appear that the average runtime of TV Series has increased since 1920-1975 up to about 50 minutes where runtimes have stabilized.

5.3 Conduct ANOVA Test for Mean Runtime of TV Series

$$\begin{cases} H_0 : \mu_{\text{before 1975}} = \mu_{\text{after 1975}} \\ H_a : \mu_{\text{before 1975}} < \mu_{\text{after 1975}} \end{cases}$$

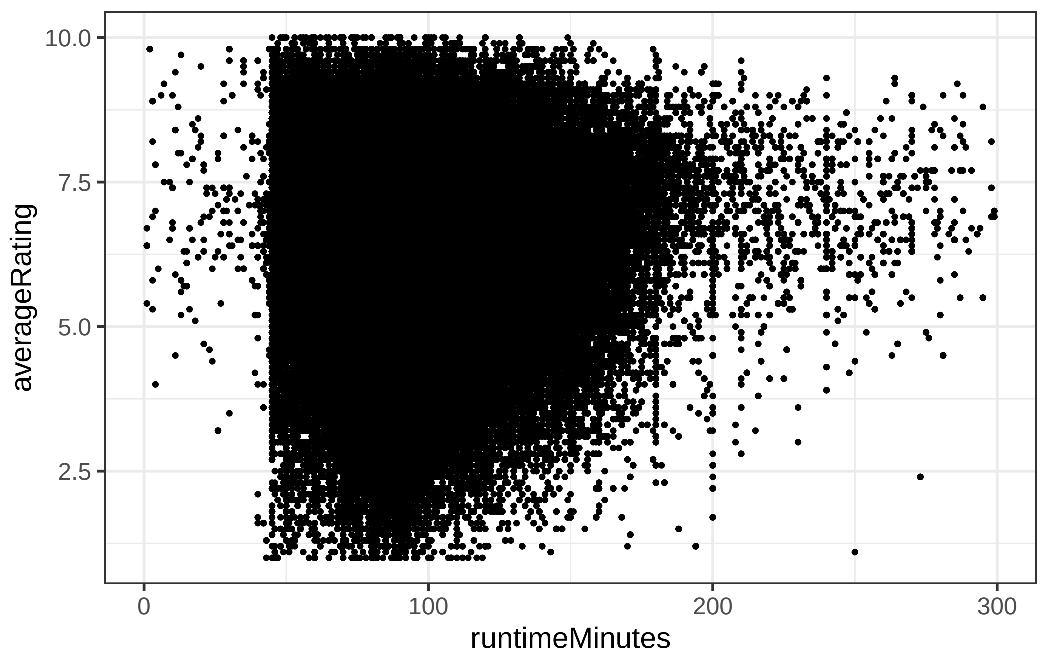
```
    Welch Two Sample t-test

data:  before_1975_runtimes and after_1975_runtimes
```

```
t = 1.0731, df = 45.034, p-value = 0.8555
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 121.1835
sample estimates:
mean of x mean of y
 95.00250  47.75681
```

By performing a two-sample t-test on the average runtimes of TV Series before 1975 vs after 1975, we find a p-value of 0.8555. This means we fail to reject $H_0 : \mu_{\text{before 1975}} = \mu_{\text{after 1975}}$, meaning we did not find a significant difference in mean runtime of TV Series in the years before 1975 and after 1975. This means TV Series runtimes have not gotten significantly longer over time.

6 Does the Runtime of a Movie affect it's Rating? (Runtime over 2 hours vs less than 2 hours)



From the above scatterplot, we can see that beyond 150 minutes of runtime, an increase in runtime is associated with an increase in the average rating of a film.

To test this we can perform a two sample t-test on the following hypotheses to test if there is a significant difference in mean rating between movies with runtimes less than 150 minutes and movies with runtimes of more than 150 minutes.

$$\begin{cases} H_0 : \mu_{\text{less than 150 minutes}} = \mu_{\text{greater than 150 minutes}} \\ H_a : \mu_{\text{less than 150 minutes}} < \mu_{\text{greater than 150 minutes}} \end{cases}$$

```
    Welch Two Sample t-test

data:  runtime_ratings_lessthan2 and runtime_ratings_morethan2
t = -47.851, df = 14867, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.521758
sample estimates:
mean of x mean of y
 6.107989  6.648321
```

Through our hypothesis test, we find a p-value smaller than 0.05, meaning we can reject $H_0 : \mu_{\text{less than 150 minutes}} = \mu_{\text{greater than 150 minutes}}$, indicating that the average rating of films longer than 150 minutes is significantly greater than the average rating of films less than 150 minutes.