Runtime and Ratings: Are Longer Horror Movies Better (or Scarier)?

2024-10-15

Project Purpose

Horror Movies can be a great way to escape reality and immerse yourself, and give an outlet to relieve stress in day to day life. I wanted to know if it's worth it to watch a long movie or if saving time on a movie won't affect the viewing experience. Because of this idea, I decided to investigate how movie run time relates to the perceived quality of the movie by viewers. I want to know if short movies are rated as highly as long movies, and decide if I need to set more time aside to watch.

The population of the dataset is **all Horror Movies that have been listed on IMDb**. Also known as the International Movies Database, IMDb is a popular online movie database and user reviewing website. This website provides a repository for hundreds of thousands of movies, tv shows, and other forms of entertainment programs, with detailed information on hundreds of variables ranging from the individuals involved in the making of the content, aspects of the content itself, and one of the critical variables, user ratings (IMDB | Help, n.d.-b). IMDB claims that, as of September 2024, they have 693,133 individual Movies within their database, and 255,604 entertainment programs within the Horror genre (Press Room - IMDB, n.d.). Using their advanced title search, I have found that the population of all horror movies on the IMDB platform is 42,758 (Advanced Search, n.d.). Therefore, the population I have derived my sample from are the 42,758 Horror Movies that are listed on IMDB.

Since IMDB collects information on hundreds if not thousands of variables, I've decided to choose one of the most relevant qualities about the movie as the variable of interest, the **Average IMDb movie user rating**. IMDB offers users with an account the ability to rate a movie on an ordinal scale from 1 to 10, with 10 being high and 1 being low, indicating their vote on the quality of the movie (IMDB | Help, n.d.). These ratings are then aggregated and averaged, creating the Average IMDB movie user rating score for each individual movie. Therefore, while each rating is based on an ordinal scale, the average of these ordinal scales is, in practice, a continuous variable. This figure can inform potential movie watchers on the potential quality of a movie before choosing to watch the movie or not. This variable can be leveraged as a decision factor to an individual choosing which movie to watch.

The factor that I will be leveraging within the parameter test with my variable of interest is the Runtime of each individual movie. Using this as my factor, the parameter that I am interested in testing is **whether there is a significant difference between the average IMDb ratings of long horror movies and short horror movies**. This will inform if I am able to enjoy quality, short movies, or if passionate horror movie goers are required to sit through longer films.

Data Collection

The dataset **"IMDb Horror: Chilling Movie Dataset"** has been found and sourced from Kaggle (IMDB Horror: Chilling Movie Dataset, 2023). This dataset was collected directly from the IMDb website, and looks at a sample of 837 movies categorized as horror. The sample is a collection of horror movies that have more than 25,000 reviews on the website, ensuring that a sufficient rating sample size has been reached. Some of the key variable columns in the dataset include release date, genre, director, runtime, rating, and gross profits. The data is in a .csv format and the gross profits column contains some NaN values. All other columns contain complete data and do not require further cleaning. The data is licensed under CC0, which is a public domain, and has no copyright.

Statistical Analysis

First, I loaded in the dataset. I downloaded in here all relevant libraries (`mosaic` and `ggplot2`) and looked over each variable using a head function:

```r
data <- read.csv("Horror Movies IMDb.csv")
library(mosaic)
library(ggplot2)
head(data, 5)
```

```
##   Movie.Title Movie.Year Runtime                     Genre Rating
## 1       Alien       1979     117            Horror, Sci-Fi    8.5
## 2      Psycho       1960     109 Horror, Mystery, Thriller    8.5
## 3 The Shining       1980     146             Drama, Horror    8.4
## 4   The Thing       1982     109   Horror, Mystery, Sci-Fi    8.2
## 5     Tumbbad       2018     104    Drama, Fantasy, Horror    8.2
##           Director     Votes   Gross
## 1     Ridley Scott  9,05,275 $78.90M
## 2 Alfred Hitchcock  6,89,068 $32.00M
## 3  Stanley Kubrick 10,51,582 $44.02M
## 4   John Carpenter  4,39,793 $13.78M
## 5  Rahi Anil Barve    53,297
```

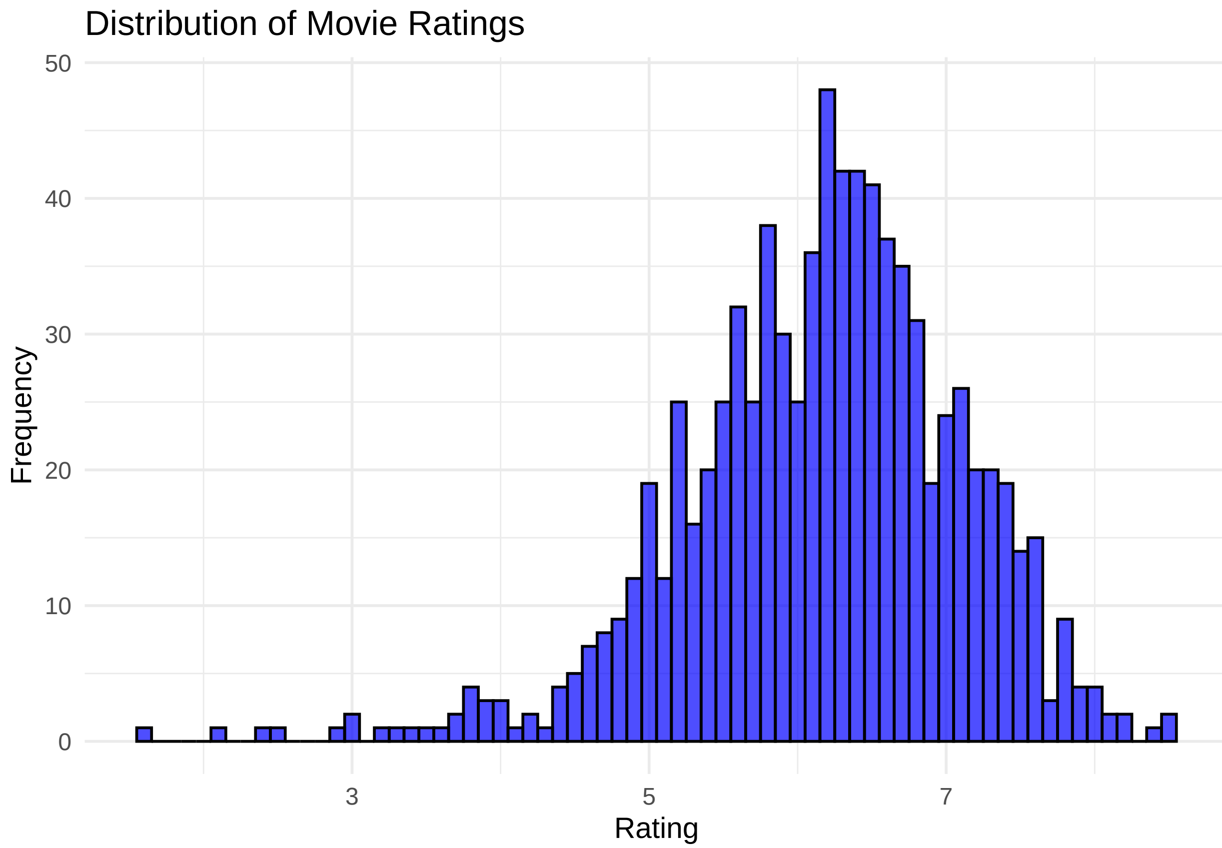
Exploratory Data Analysis (EDA)

To start my analysis, I conducted an introductory EDA to gain a better understanding of the data. The main variables of interest include Rating and Runtime. One key definition I hoped to reach based on my EDA was appropriate constant runtime values to define "long" and "short" horror movies.

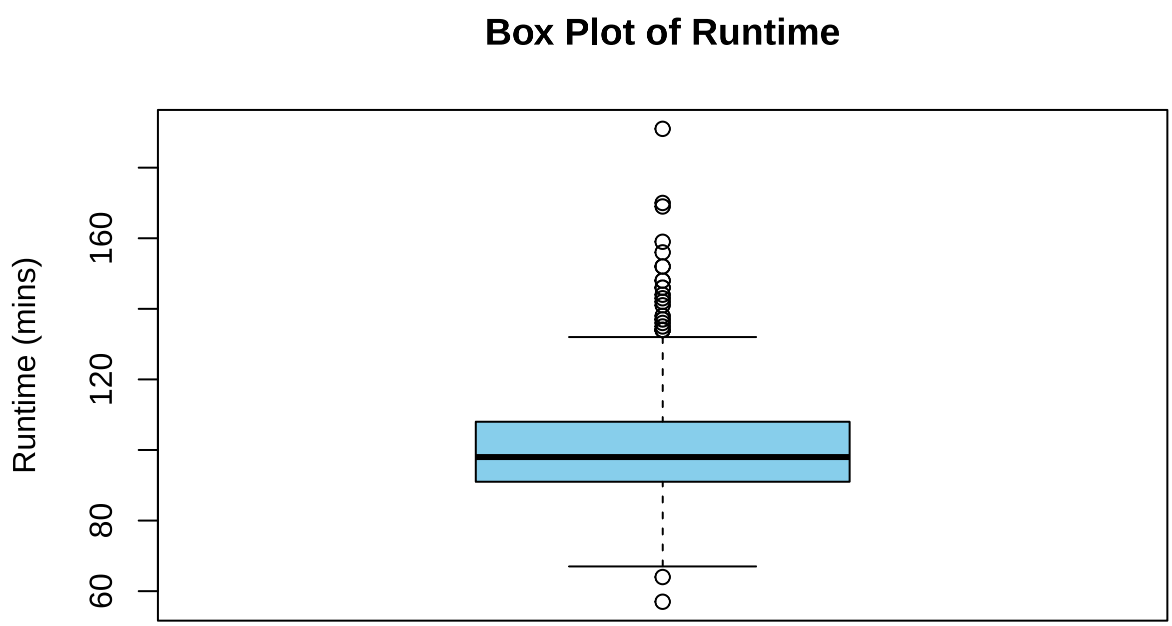
```r
favstats(data$Runtime)
```

```
##  min Q1 median  Q3 max     mean       sd   n missing
##   57 91     98 108 191 100.7679 14.48456 836       0
```

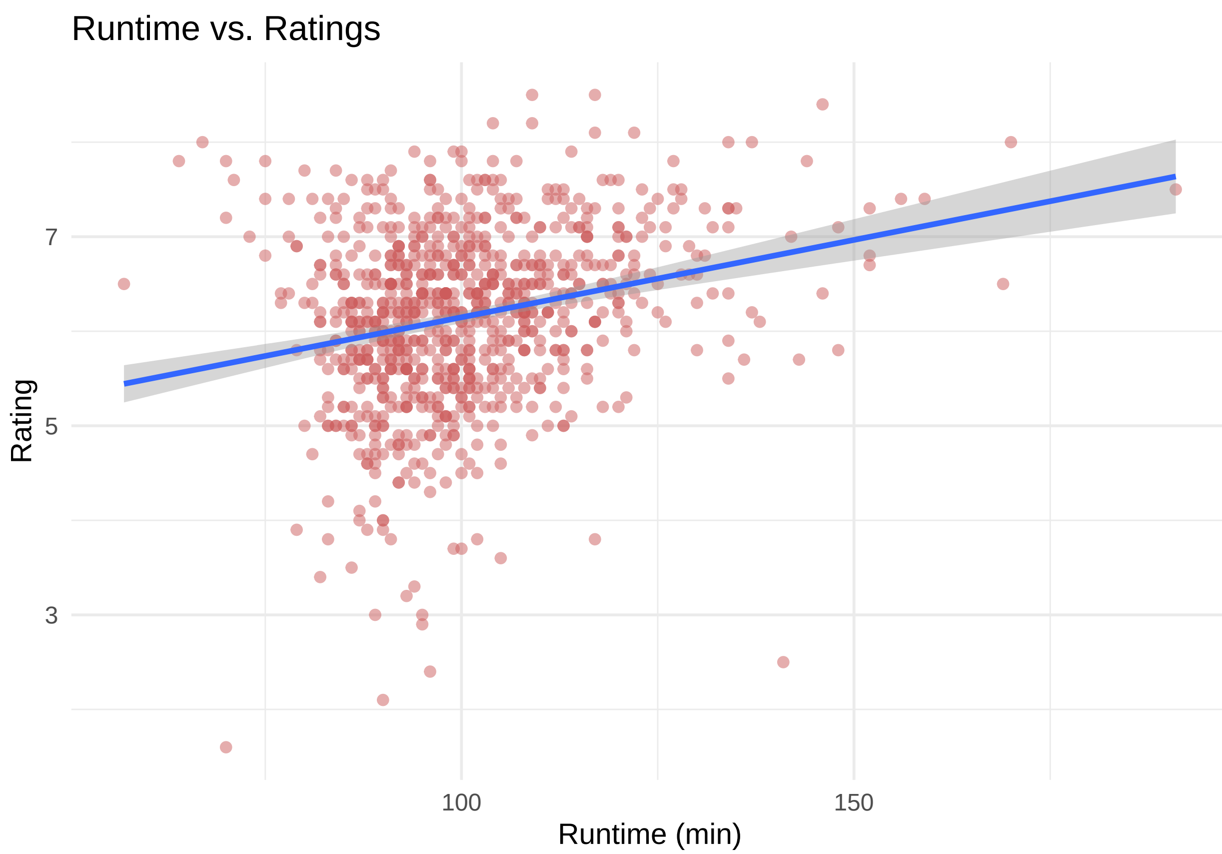
```r
ggplot(data, aes(x = Rating)) +
  geom_histogram(binwidth = 0.1, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Movie Ratings", x = "Rating", y = "Frequency") +
  theme_minimal()
```




```
boxplot(data$Runtime,main="Box Plot of Runtime",ylab="Runtime (mins)",
        col="skyblue",border = "black")
```




```
ggplot(data, aes(x = Runtime, y = Rating)) +
  geom_point(alpha = 0.5, color = "indianred") +
  labs(title = "Runtime vs. Ratings", x = "Runtime (min)", y = "Rating") +
  stat_smooth(method = "lm", formula = y ~ x, geom ="smooth") +
  theme_minimal()
```



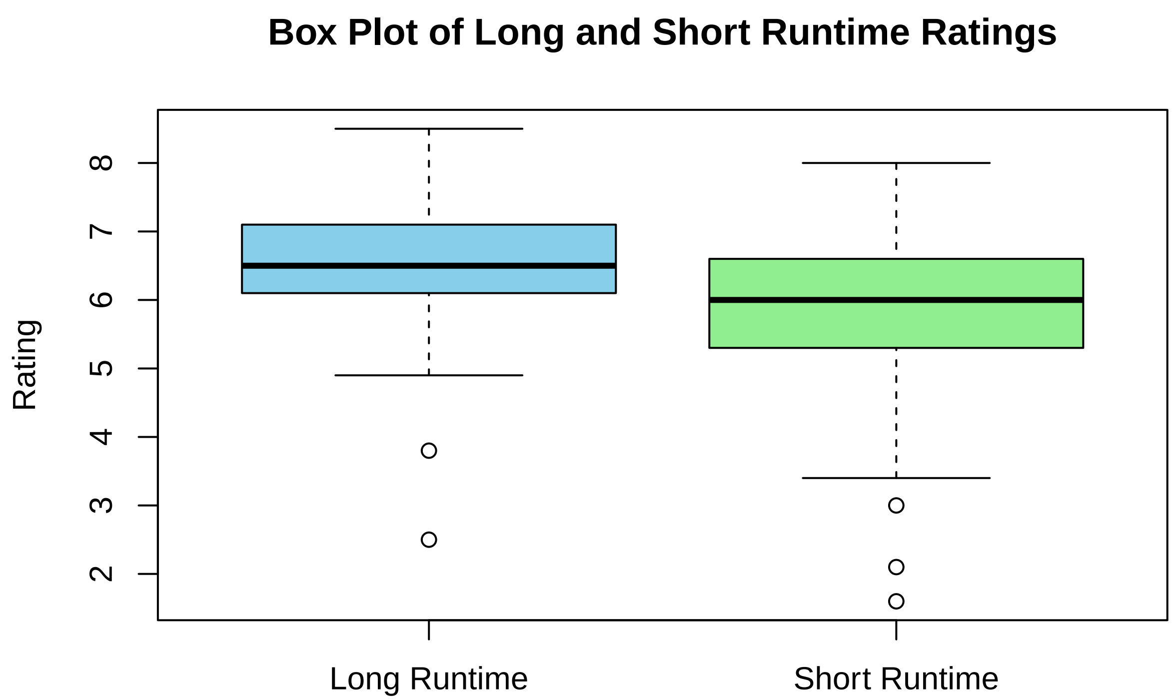
Based off my initial EDA analysis, I gained insight on distributions of ratings and how to best define my test parameters. I found the distribution of movie ratings to be left skewed, and that ratings follow what appears be a relatively normal distribution. I chose to define a "long" movie as any movie that is found to be longer than or equal to Q3 and a "short" movie as any movie that is found to be shorter than or equal to Q1:

long: $\geq Q3$, short: $\leq Q1$

Further, I created a scatter plot to test if there is any visual relationship between runtime and ratings using a line of best fit. I see that there is a slightly positive relationship, which would lead us to believe that longer movies may have higher ratings than shorter movies.

One final EDA visualization I completed was a side-by-side box plot of long and short horror movie ratings. I define long and short horror movie runtimes here that will be used in further analysis. I wanted to better visualize any differences between long and short horror movie ratings to gain a clearer picture before conducting statistical inferences and conclusion:

```
longQ3=filter(data,Runtime>=108)
shortQ1=filter(data,Runtime<=91)
longQ3R=longQ3$Rating
shortQ1R=shortQ1$Rating
boxplot(longQ3R, shortQ1R,
        main="Box Plot of Long and Short Runtime Ratings",
        names=c("Long Runtime", "Short Runtime"),
        ylab="Rating",
        col=c("skyblue", "lightgreen"),
        border="black")
```



By looking at the box plots, I can definitely see there is some level of difference in ratings for movies that have a long runtime relative to a short runtime. To test this, I conducted hypothesis testing analysis on the mean difference between long and short horror movies to see if there is a statistical difference at a 95% confidence level:

Test 1: Hypothesis Testing

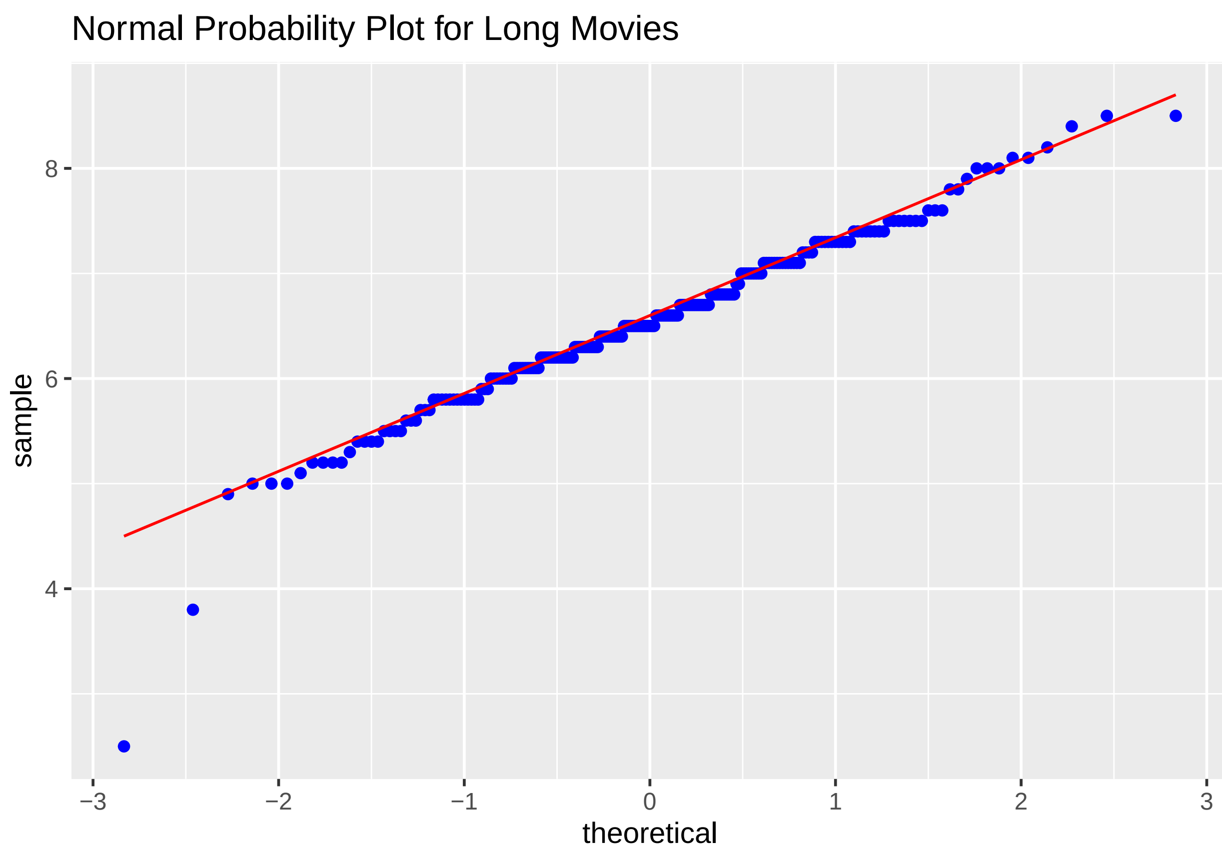
My null hypothesis is that there is no difference between mean long horror movie ratings and mean short horror movie ratings.
My alternative hypothesis is that there is a difference between mean long horror movie ratings and mean short horror movie ratings.

$H_0 : \mu_{Long} = \mu_{Short}$
$H_A : \mu_{Long} \neq \mu_{Short}$

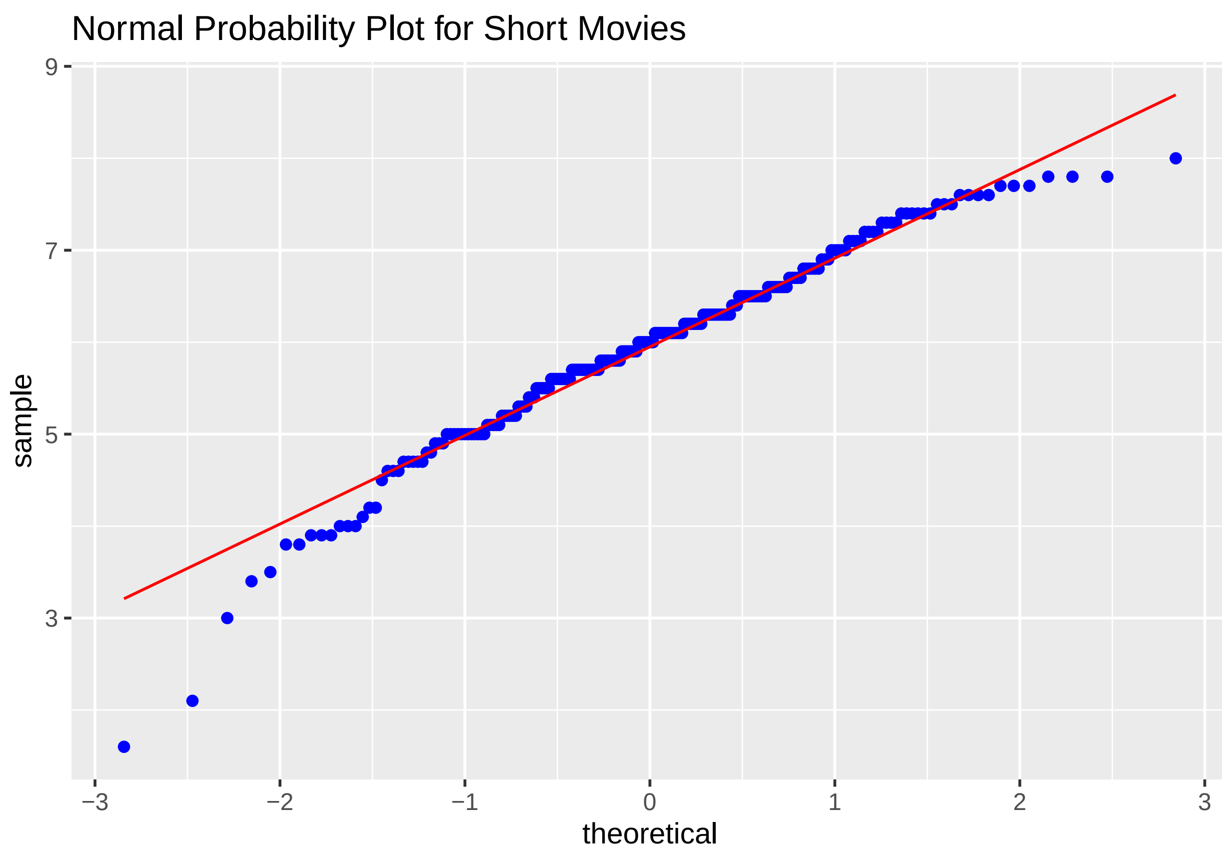
$\alpha = 0.05$

For the test statistic, I chose to use a t-test statistic and compare the two to come to a more accurate conclusion. In t-tests, one of the key assumptions that must be met is normality. In order to check for normality, I created a Normal Probability Plot using stat_qq as a function of qqplot.

```
ggplot(data=longQ3,aes(sample=longQ3R))+
  stat_qq(col="blue")+stat_qqline(col="red")+
  ggtitle("Normal Probability Plot for Long Movies")
```




```
ggplot(data=shortQ1,aes(sample=shortQ1R))+
  stat_qq(col="blue")+stat_qqline(col="red")+
  ggtitle("Normal Probability Plot for Short Movies")
```



Since the produced plots of long horror movies and short horror movies follows roughly a straight line through the middle of the data points, these data can be determined to conform to normal probability models.

Another important factor when conducting a t-test on two samples is the sample variance. I must find out if the samples have equal variance or not to conclude if I should complete an independent sample t-test or a Welch t-test.

```
favstats(longQ3R)
```

```
##  min  Q1 median  Q3 max     mean        sd   n missing
##  2.5 6.1    6.5 7.1 8.5 6.540553 0.7921581 217       0
```

```
favstats(shortQ1R)
```

```
##  min  Q1 median  Q3 max     mean       sd   n missing
##  1.6 5.3      6 6.6   8 5.935714 1.040684 224       0
```

```
cat("Standard diviation of two populations are unequal, therefore, their variances are unequal.")
```

```
## Standard diviation of two populations are unequal, therefore, their variances are unequal.
```

Because both ratings of long and short movies are considered to follow normal probability models, and the variances in the two populations' data sets are unequal, I will use Welch's t-test to compute the test statistic and p-value.

```
t.test(longQ3R,shortQ1R,alternative="two.sided",conf.level=0.95,val.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  longQ3R and shortQ1R
## t = 6.8809, df = 415.92, p-value = 2.197e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4320519 0.7776256
## sample estimates:
## mean of x mean of y
##  6.540553  5.935714
```

p-value: 2.197e-11 < 0.05 (α)

At a significance level of 0.05, I reject the null hypothesis that there is no difference between long horror movie ratings and short horror movie ratings in favour of the alternative hypothesis that there is a difference between long horror movie ratings and short horror movie ratings.

Test 2: Confidence Interval Testing

The confidence interval test is useful to get the range of values within the confidence interval. To conduct this test, I will again define our null and alternative hypotheses. I will test to see if long horror movies have a different mean rating than short horror movies:

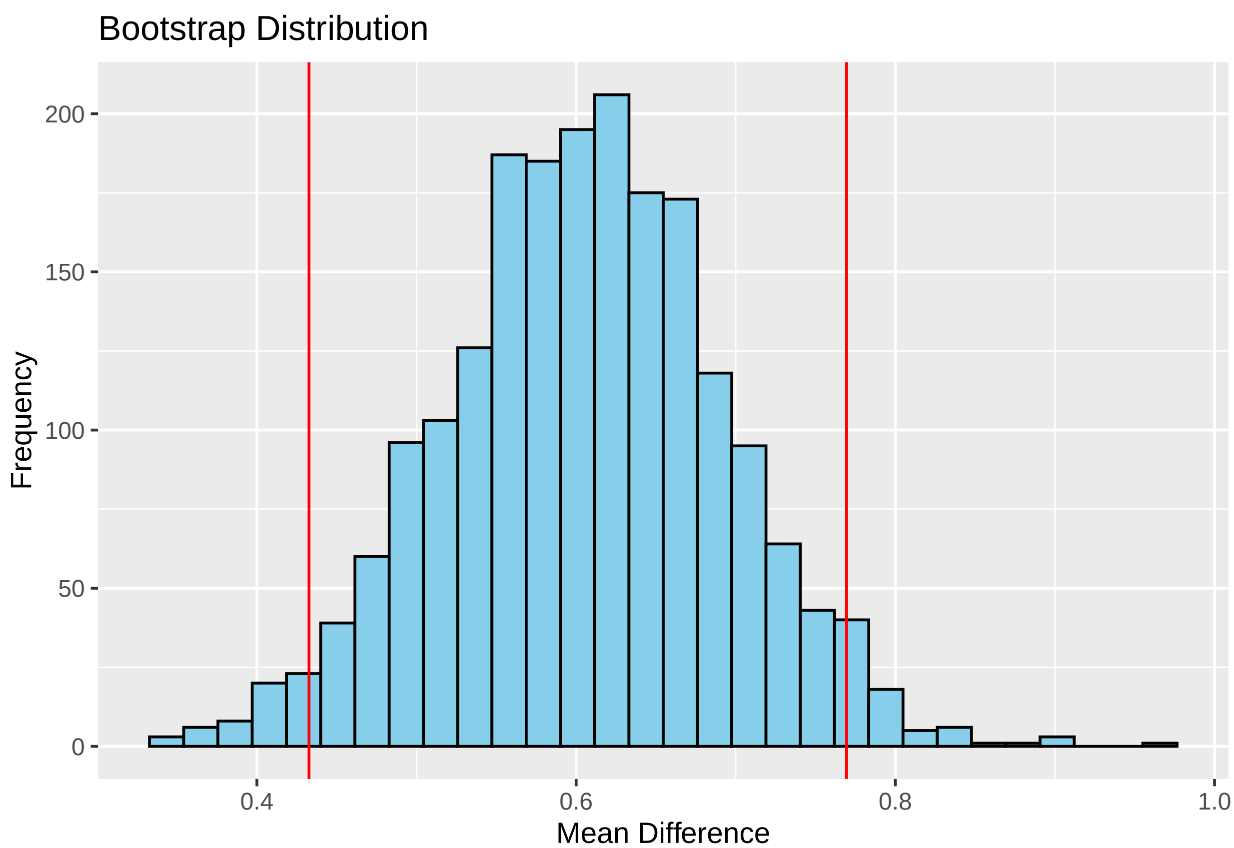
$H_0 : \mu_{Long} = \mu_{Short}$ $H_A : \mu_{Long} \neq \mu_{Short}$

$\alpha = 0.05$

To evaluate the confidence interval, I will create a bootstrap distribution to find our bootstrap confidence interval.

With bootstrapping, I generated many mean differences based on the difference between resampled ratings of long and short horror movies. In my bootstrapping, 2000 repeats of the mean difference between long movie ratings and short movie ratings were simulated to compute the distribution of the mean difference. I then computed the 95% confidence interval based on the simulated distribution of mean difference and visualized the distribution.

```
B=do(2000)*(mean(resample(longQ3R))-mean(resample(shortQ1R)))
LL = quantile(B$result, 0.025)
UL = quantile(B$result, 0.975)
ggplot(B,aes(x=result))+
  geom_histogram(fill="skyblue",col="black")+
  ylab("Frequency")+
  xlab("Mean Difference")+
  ggtitle("Bootstrap Distribution")+
  geom_vline(xintercept = UL, color = "red")+
  geom_vline(xintercept = LL, color = "red")
```




```r
interval=quantile(B$result,c(0.025,0.975))
interval
```

```
##      2.5%     97.5%
## 0.4326404 0.7694628
```

With 95% confidence, the true mean difference between long movie ratings and short movie ratings is captured in the interval of (0.4258151, 0.7766885).

Since the confidence interval does not capture 0, at a 0.05 significance level, I can reject the null hypothesis that the mean rating of long horror movies is equal to the mean rating of short horror movies in favour of the alternative hypothesis that there is a difference between long horror movies ratings and short horror movies ratings."

Conclusion

Through my completed analysis, I have found a statistically significant difference in IMDb movie ratings between short and long horror movies. Based on the results from my hypothesis test and bootstrapping confidence interval, the longest 25% of horror movies tend to receive higher average IMDb user ratings than the shortest 25% of horror movies. As a group of passionate horror fans, I now have the knowledge that informs us that, to have a better likelihood of enjoying a horror movie, a relatively longer movie choice may result in a better viewing experience.

In progressing through this project and identifying my key findings, I learned a great deal about the practical implementation of hypothesis testing and confidence intervals in real-world datasets and real-world scenarios with little guidance. Additionally, I gained strong experience identifying datasets and utilizing decision criteria to accept or reject potential datasets. This project has improved my ability to progress through the stages of a data science project, conceptualize relevant questions and problems, and use my newfound skills to develop tangible insights and solutions.

References

Chen, G. (2024). EDA numerical summaries PDF document. University of Calgary, Central Authentication Service. https://d2l.ucalgary.ca/d2l/le/content/618086/viewContent/6613182/View

Chen, G. (2024). Estimation: Point and interval estimation. University of Calgary, Central Authentication Service. https://d2l.ucalgary.ca/d2l/le/content/618086/viewContent/6613184/View

Chen, G. (2024). Introduction to data visualization. University of Calgary, Central Authentication Service. https://d2l.ucalgary.ca/d2l/le/content/618086/viewContent/6613181/View

Chen, G. (2024). Statistical tests of hypothesis. University of Calgary, Central Authentication Service. https://d2l.ucalgary.ca/d2l/le/content/618086/viewContent/6613187/View

Comprehensive R Archive Network (CRAN). (n.d.). CRAN: Package latexpdf. https://cran.r-project.org/web/packages/latexpdf/index.html

IMDb. (n.d.). Advanced search. https://www.imdb.com/search/title/?title_type=feature&genres=horror&sort=user_rating,desc

IMDb. (n.d.). Press Room - IMDb. https://www.imdb.com/pressroom/stats/

IMDb. (n.d.). What is IMDb? https://help.imdb.com/article/imdb/general-information/what-is-imdb/G836CY29Z4SGNMK5?ref_=helpart_nav_1#

IMDB Horror: Chilling Movie Dataset. (2023, June). Kaggle. https://www.kaggle.com/datasets/shreyanshverma27/imdb-horror-chilling-movie-dataset?resource=download

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A grammar of data manipulation (Version 1.1.4). R package. https://dplyr.tidyverse.org

Xie, Y. (2021). knitr: A general-purpose package for dynamic report generation in R. https://yihui.org/knitr/