Modeling and prediction for movies

Setup

Load packages

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
library(statsr)
```

```
## Warning: package 'statsr' was built under R version 4.0.5
```

```
## Warning: package 'BayesFactor' was built under R version 4.0.5
```

```
## Warning: package 'coda' was built under R version 4.0.5
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.5
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.0.5
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.0.5
```

Load data

Make sure your data and R Markdown files are in the same directory. When loaded your data file will be called `movies` . Delete this note when before you submit your work.

```
load("D:/Statistics with R/Data Analysis Project/Course 3/movies.Rdata")
movies <- na.omit(movies) # Remove the NA values from the dataset
```

Part 1: Data

The dataset consists of movie information from Rotten Tomatoes and IMBD.The data set is comprised of 651 randomly sampled movies produced and released before 2016.We have 32 variables recorded for each movie.It is possible to do an observational study with this dataset, which can be generalized to movies produced and released before 2016.

Part 2: Research question

As a data scientist for the Paramount Pictures, I am interested in learning what attributes make a movie popular and also the relationship between genre, ratings and critic scores for a movie? Do the ratings of critics and audience coincide?

Part 3: Exploratory data analysis

Let’s do EDA on the variables we want to consider.

```
str(movies %>%
      select(genre,imdb_rating,critics_score,audience_score))
```

```
## tibble [619 x 4] (S3: tbl_df/tbl/data.frame)
##  $ genre         : Factor w/ 11 levels "Action & Adventure",..: 6 6 4 6 7 6 6 5 6 1 ...
##  $ imdb_rating   : num [1:619] 5.5 7.3 7.6 7.2 5.1 7.2 5.5 7.5 6.6 6.8 ...
##  $ critics_score : num [1:619] 45 96 91 80 33 57 17 90 83 89 ...
##  $ audience_score: num [1:619] 73 81 91 76 27 76 47 89 66 75 ...
##  - attr(*, "na.action")= 'omit' Named int [1:32] 6 25 94 100 131 172 175 184 198 207 ...
##   ..- attr(*, "names")= chr [1:32] "6" "25" "94" "100" ...
```

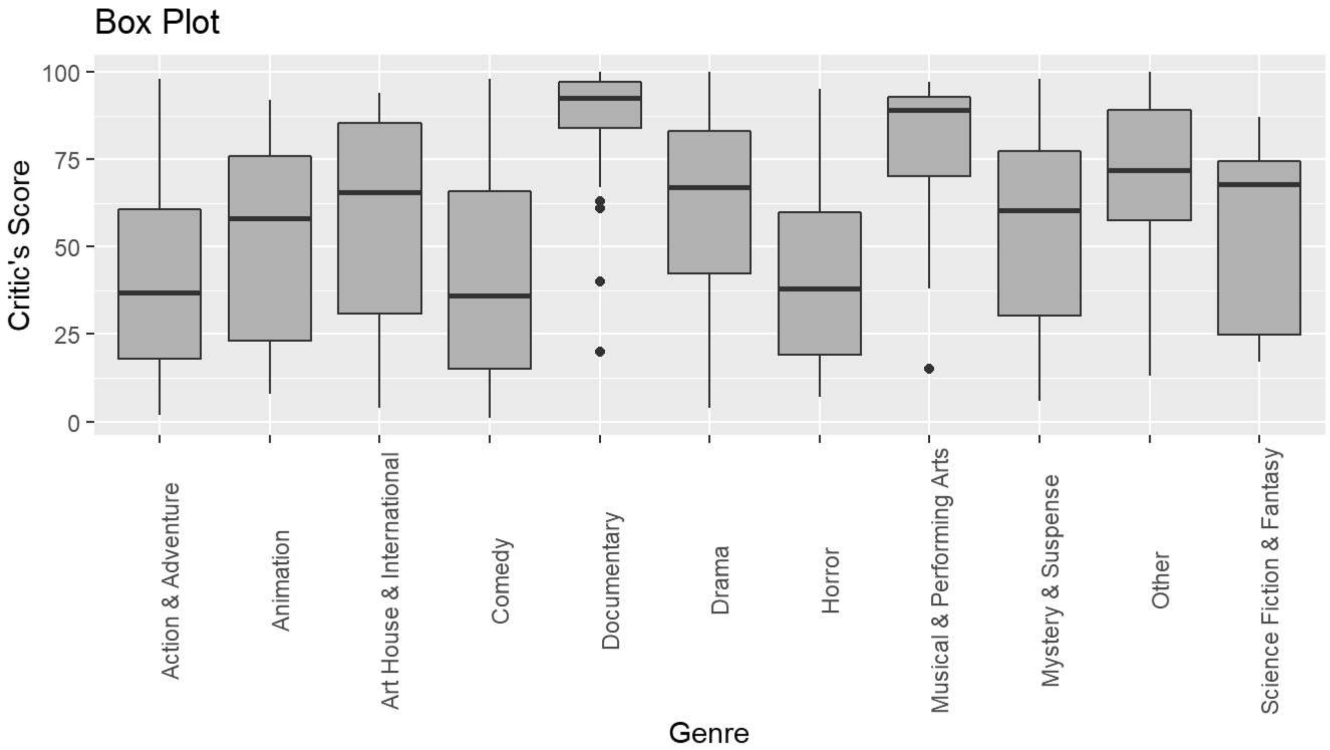
```
summary(movies %>%
          select(genre,imdb_rating,critics_score,audience_score))
```

```
##                  genre      imdb_rating    critics_score    audience_score
##  Drama             :298   Min.   :1.900   Min.   :  1.00   Min.   :11.00
##  Comedy            : 86   1st Qu.:5.900   1st Qu.: 33.00   1st Qu.:46.00
##  Action & Adventure: 62   Median :6.600   Median : 61.00   Median :65.00
##  Mystery & Suspense: 56   Mean   :6.486   Mean   : 57.43   Mean   :62.21
##  Documentary       : 40   3rd Qu.:7.300   3rd Qu.: 82.50   3rd Qu.:80.00
##  Horror            : 22   Max.   :9.000   Max.   :100.00   Max.   :97.00
##  (Other)           : 55
```

From the above data summary we can see that among all the genre’s the most common one is Drama with 305 movies out of 651 being of this genre. Comedy comes in second with 87 movies,which is much lower in number than that of Drama. The summary tells us that the average score is generally between 55% to 65%. But the lowest IMDB ratings and audience scores is around the 10% mark of the scoring system, the critic scores indicate a low rating of around 1% of the scoring system. This is a clear indication of critics scores being different from audience views. While the highest scores still range between 90% - 100% .

Now lets create a visualization between genres and the relevent scores and ratings. First lets see the Critic’s scores:

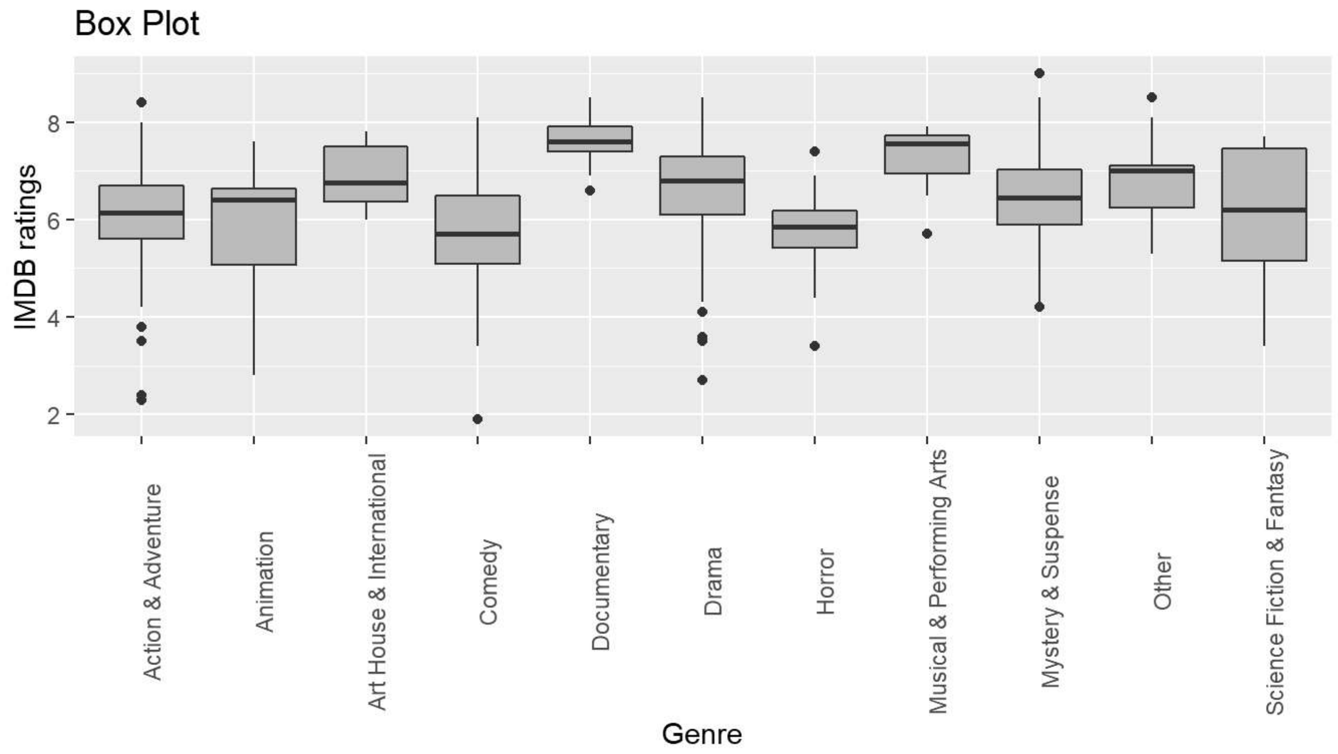
```
ggplot(data = movies, aes(x = genre, y = critics_score, fill = "genre"))+
  theme(axis.text.x = element_text(angle = 90)) +
  geom_boxplot(fill="cyan")+
  labs(title="Box Plot",
      x="Genre",
      y="Critic's Score")
```



From this plot we can see the co-relation between the Critic’s scores and the Genre of the movies. This shows that the genre with the highest number of scores are Documentaries and Musical and Performing arts. The rest of the genre shows an evenly distributed number of scores, with most scores evenly distributed between the lowest and the highest.

In the case of the IMDB scores:

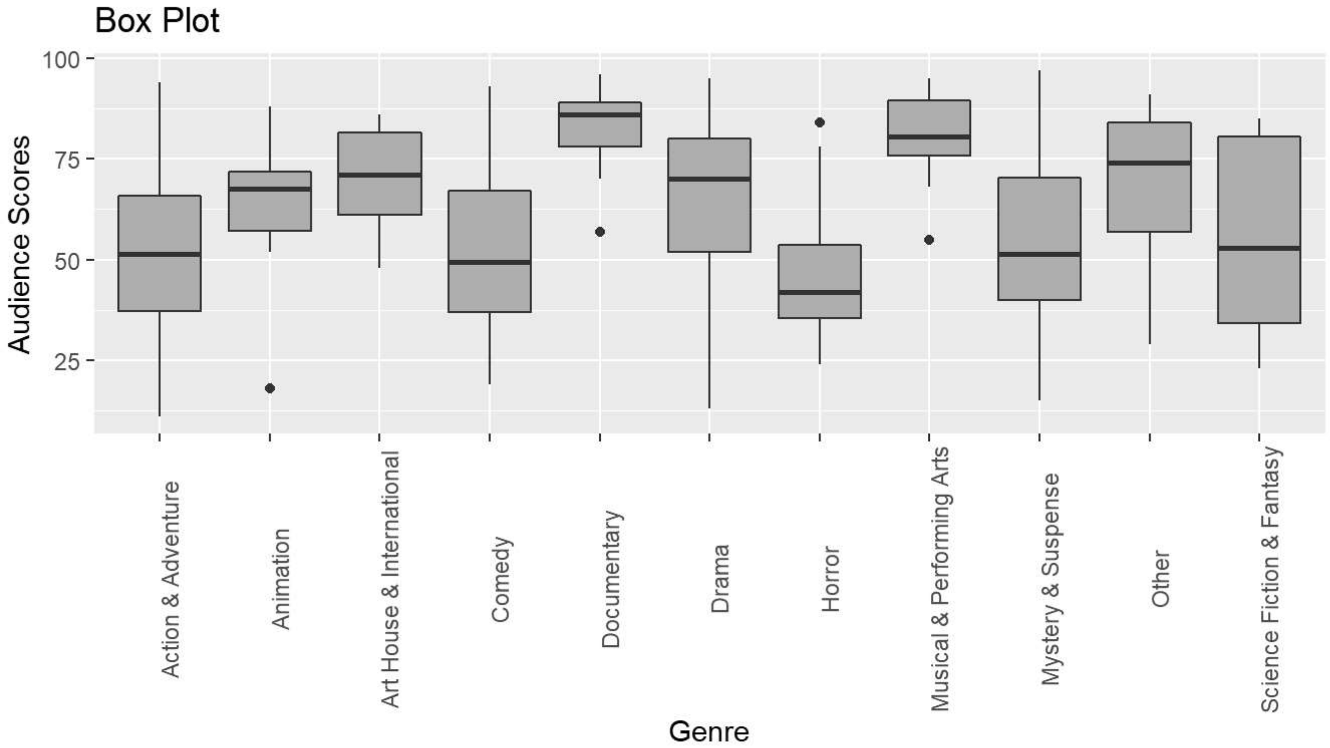
```
ggplot(data = movies, aes(x = genre, y = imdb_rating, fill = "genre"))+
 theme(axis.text.x = element_text(angle = 90)) +
 geom_boxplot(fill="sky blue") +
 labs(title="Box Plot",
       x="Genre",
       y="IMDB ratings")
```



From this plot we can see that most of the genres except Science fiction & Fantasy shows a similar trend. All of these show a higher number of above average scores, with the remaining scores being distributed evenly in the lower 50% of the scoring system. This aligns with the scoring seen among the Critic's scores.

Finally we will look at the audience scores:

```
ggplot(data = movies, aes(x = genre, y = audience_score, fill = "genre"))+
  theme(axis.text.x = element_text(angle = 90)) +
  geom_boxplot(fill="violet") +
   labs(title="Box Plot",
        x="Genre",
        y="Audience Scores")
```



Documentaries and Musical and Performing arts shows a similar trend as the other two plots. But the other genres are different when compared to other plots.

Now let’s see the co-relation between these,

```
# The co-relation between IMDB ratings and Critic's scores
movies %>%
  summarise(cor(imdb_rating,critics_score))
```

```
## # A tibble: 1 x 1
##   `cor(imdb_rating, critics_score)`
##                               <dbl>
## 1                             0.762
```

```
#Co-relation between IMDB rating and audience scores
movies %>%
  summarise(cor(imdb_rating,audience_score))
```

```
## # A tibble: 1 x 1
##   `cor(imdb_rating, audience_score)`
##                                <dbl>
## 1                              0.861
```

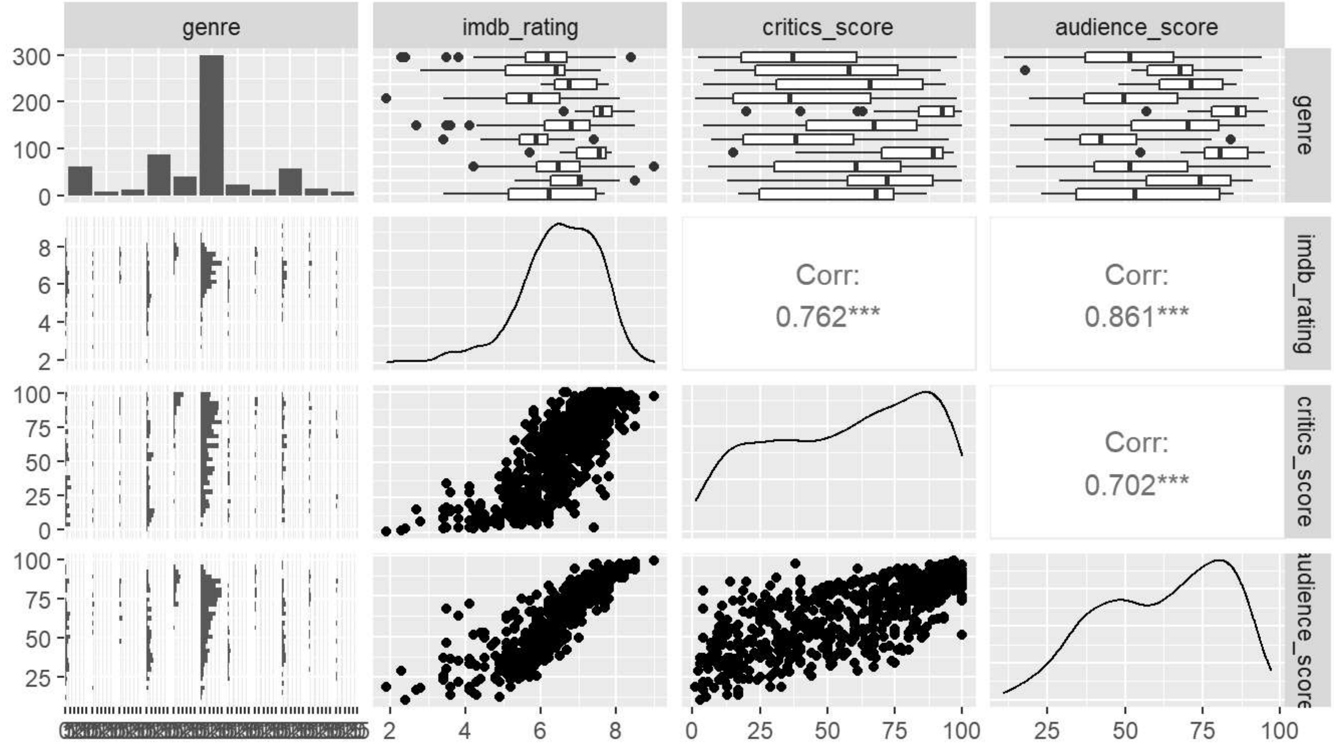
```
#Co-relation between Critic's scores and audience scores
movies %>%
  summarise(cor(critics_score,audience_score))
```

```
## # A tibble: 1 x 1
##   `cor(critics_score, audience_score)`
##                                  <dbl>
## 1                                0.702
```

Now we will also plot a co-relation chart divided by genre,

```
ggpairs(movies , columns = c(3,13,16,18))
```

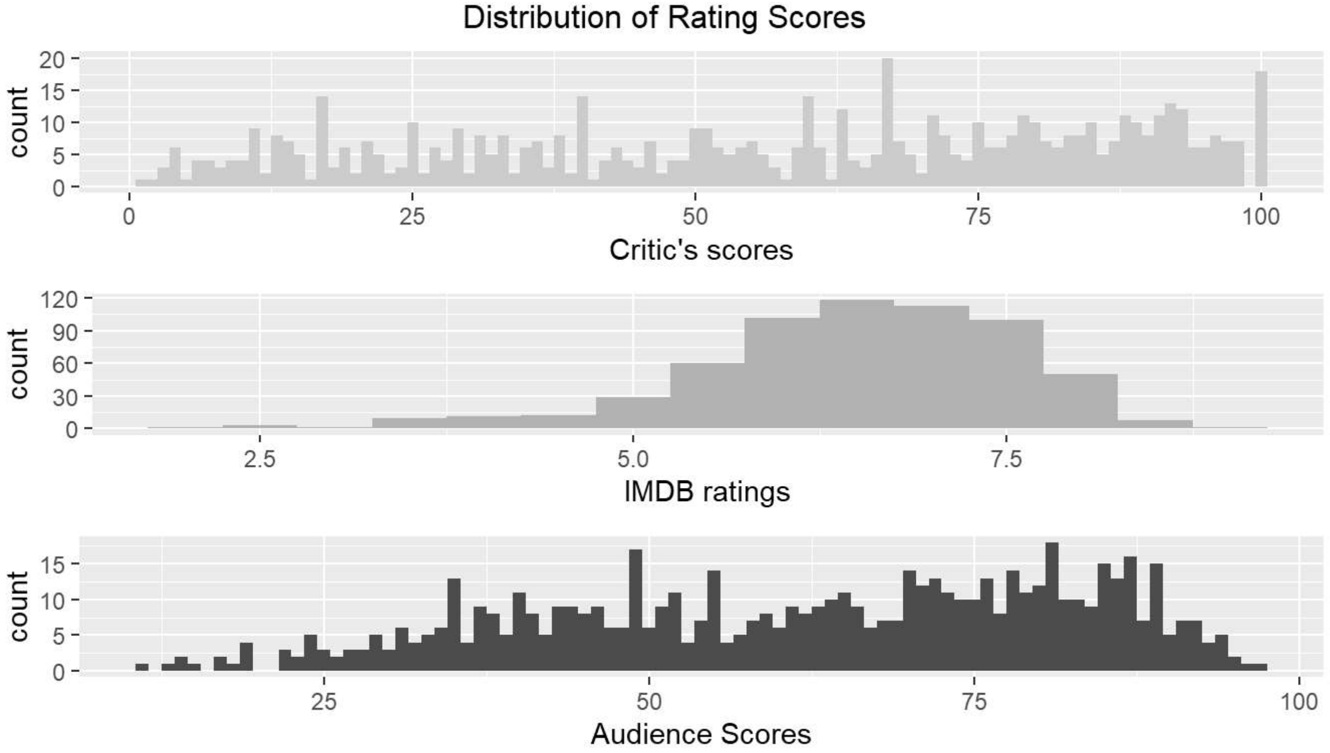
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can see a strong correlation between the scores and ratings. This suggests that the audience on Rotten tomatoes as well as IMDB generally agree with the Critic scores.

Now let’s plot a histogram of these distribution:

```
plot1 <- ggplot(data=movies, aes(x=critics_score)) +
  geom_histogram(binwidth=1, fill="light blue") +
  xlab("Critic's scores")
plot2 <- ggplot(data=movies, aes(x=imdb_rating)) +
  geom_histogram(binwidth=0.5, fill="cyan") +
  xlab("IMDB ratings")
plot3 <- ggplot(data=movies, aes(x=audience_score)) +
  geom_histogram(binwidth=1, fill="red") +
  xlab("Audience Scores")
grid.arrange(plot1, plot2, plot3,
             top="Distribution of Rating Scores")
```



We can see that the IMDB ratings show a nearly normal distribution, while both the Rotten Tomatoes audience scores and the Critic's scores show a left-skewed distribution. Hence IMDB ratings might be the best target response variable to use for the prediction model.

Part 4: Modeling

Now we will do modeling for the target response variable, which in our case is the movie ratings in terms of scores given.From the previous plot we see a clear normal distribution in the IMDB rating, and the fact that it has a high correlation factor with both of the other rating systems, we will use IMDB ratings as the target response variable.

In this model, we will use the backward elimination mode i.e. start with all of the variables which can affect the ratings and remove certain variables to create a parsimonious model.

The initial variables will be: 1. Genre (genre) 2. Movie run time in minutes (runtime) 3. MPAA rating of movie (mpaa_rating) 4. Whether or not the movie won a best picture Oscar(best_pic_win) 5. Have any of the main actors ever won an Oscar (best_actor_win) 6. Have any of the main actresses ever won an Oscar (best_actress_win) 7. Has the director ever won an Oscar (best_dir_win) 8. Critics score on Rotten Tomatoes(critics_score) 9.Audience score on Rotten Tomatoes(audience_score)

```
initial_model <-lm(imdb_rating ~ genre+runtime+mpaa_rating+
                      best_pic_win+best_actor_win+best_actress_win+
                      best_dir_win+critics_score+audience_score,
                      data=movies)
summary(initial_model)
```

```
##
## Call:
## lm(formula = imdb_rating ~ genre + runtime + mpaa_rating + best_pic_win +
##     best_actor_win + best_actress_win + best_dir_win + critics_score +
##     audience_score, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.37972 -0.18656  0.04029  0.27695  1.19382
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.302247   0.182810  18.064  < 2e-16 ***
## genreAnimation                 -0.526633   0.191897  -2.744  0.00625 **
## genreArt House & International  0.245957   0.152986   1.608  0.10843
## genreComedy                    -0.156834   0.079830  -1.965  0.04992 *
## genreDocumentary                0.280981   0.113054   2.485  0.01321 *
## genreDrama                      0.029459   0.069675   0.423  0.67259
## genreHorror                     0.073774   0.120042   0.615  0.53907
## genreMusical & Performing Arts  0.034069   0.152167   0.224  0.82292
## genreMystery & Suspense         0.213357   0.090214   2.365  0.01835 *
## genreOther                     -0.002352   0.137549  -0.017  0.98636
## genreScience Fiction & Fantasy -0.086459   0.176947  -0.489  0.62530
## runtime                         0.005006   0.001153   4.343 1.65e-05 ***
## mpaa_ratingNC-17               -0.016896   0.489030  -0.035  0.97245
## mpaa_ratingPG                  -0.132053   0.138768  -0.952  0.34168
## mpaa_ratingPG-13               -0.073818   0.141707  -0.521  0.60261
## mpaa_ratingR                   -0.042566   0.137311  -0.310  0.75667
## mpaa_ratingUnrated             -0.180387   0.160275  -1.125  0.26084
## best_pic_winyes                 0.052203   0.193247   0.270  0.78715
## best_actor_winyes               0.026232   0.056309   0.466  0.64149
## best_actress_winyes             0.066843   0.062008   1.078  0.28148
## best_dir_winyes                 0.050262   0.081093   0.620  0.53562
## critics_score                   0.010444   0.000985  10.603  < 2e-16 ***
## audience_score                  0.033401   0.001368  24.407  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.469 on 596 degrees of freedom
## Multiple R-squared:  0.8163, Adjusted R-squared:  0.8096
## F-statistic: 120.4 on 22 and 596 DF,  p-value: < 2.2e-16
```

Now lets perform backward elimination method:

```
final_model <- step(initial_model, direction = "backward",trace = FALSE)
summary(final_model)
```

```
##
## Call:
## lm(formula = imdb_rating ~ genre + runtime + critics_score +
##     audience_score, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34317 -0.19822  0.04111  0.26913  1.16952
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.182655   0.130365  24.413  < 2e-16 ***
## genreAnimation                  -0.490449   0.177001  -2.771  0.00576 **
## genreArt House & International  0.219695   0.148553   1.479  0.13969
## genreComedy                     -0.148901   0.078375  -1.900  0.05793 .
## genreDocumentary                 0.216866   0.101665   2.133  0.03331 *
## genreDrama                       0.047788   0.067161   0.712  0.47702
## genreHorror                      0.087379   0.117252   0.745  0.45642
## genreMusical & Performing Arts  0.011288   0.150541   0.075  0.94025
## genreMystery & Suspense          0.251075   0.087187   2.880  0.00412 **
## genreOther                      -0.019440   0.136217  -0.143  0.88657
## genreScience Fiction & Fantasy -0.074232   0.176285  -0.421  0.67384
## runtime                          0.005435   0.001060   5.127 3.97e-07 ***
## critics_score                    0.010438   0.000962  10.850  < 2e-16 ***
## audience_score                   0.033525   0.001360  24.641  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4681 on 605 degrees of freedom
## Multiple R-squared:  0.8143, Adjusted R-squared:  0.8103
## F-statistic:   204 on 13 and 605 DF,  p-value: < 2.2e-16
```

The model now has 4 variables Genre, Run time, Critic Score, and the Audience Score. The model also has a slightly larger adjusted r-squared value of 0.8103, hence obtaining almost the same variability with fewer variables from the dataset.

Now lets generate ANOVA for the final model:

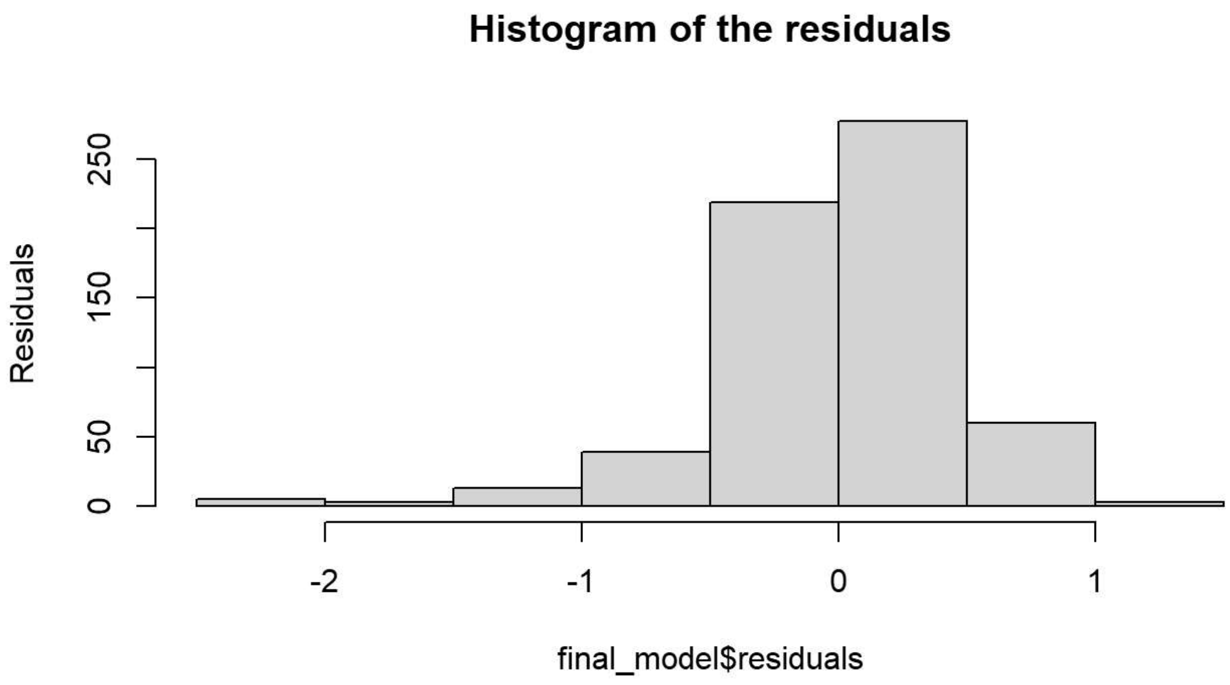
```
anova(final_model)
```

```
## Analysis of Variance Table
##
## Response: imdb_rating
##                 Df  Sum Sq Mean Sq F value    Pr(>F)
## genre           10 155.021  15.502   70.74 < 2.2e-16 ***
## runtime          1  36.246  36.246  165.40 < 2.2e-16 ***
## critics_score    1 256.893 256.893 1172.26 < 2.2e-16 ***
## audience_score   1 133.060 133.060  607.18 < 2.2e-16 ***
## Residuals      605 132.582   0.219
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the above tables that the variables in the final model are all statistically significant, hence we will reject the null hypothesis in favor of the alternative hypothesis that the coefficient of at least one of the variables is not 0. From the summary of final model we can see that only 3 out of the 10 genres show statistical significance (p-value less than 0.05).Of these genres, Documentaries and Mystery and Suspense have the highest coefficients of all the genres. The remaining significant variables are similar to genre, all have a positive effect on movies ratings.

Now lets have a look at histogram of the residual plots for the variables:

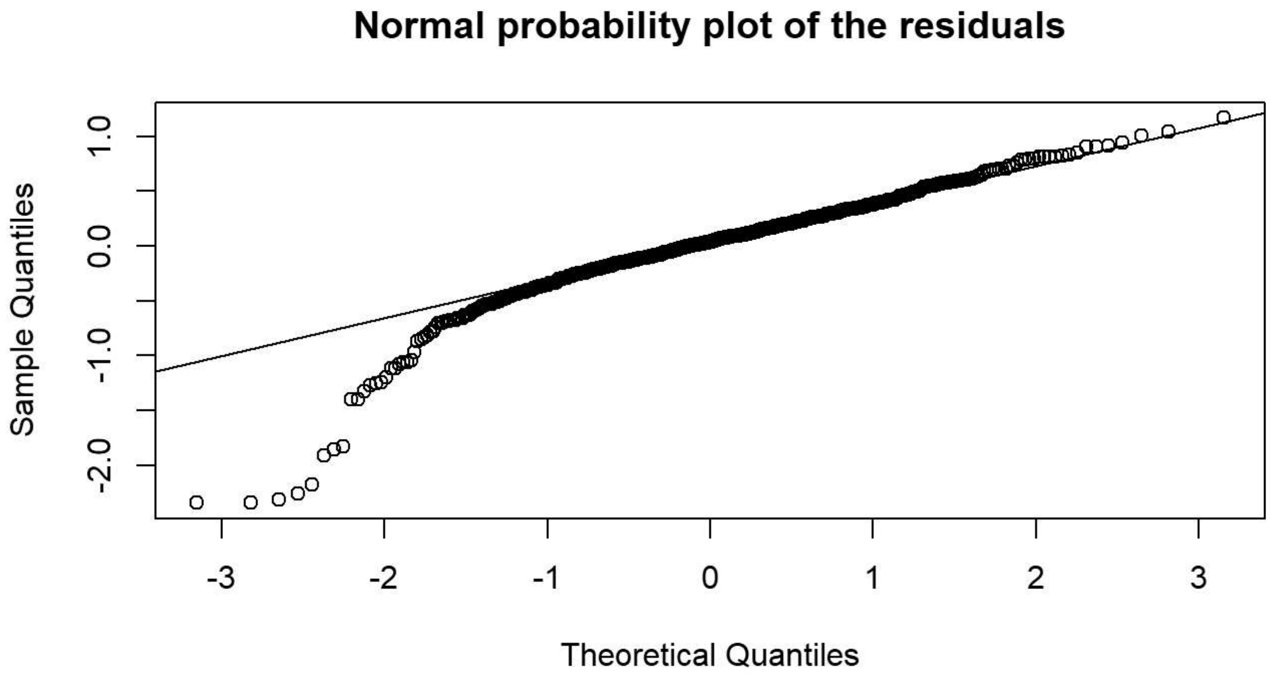
```
hist(final_model$residuals,ylab="Residuals",
      main="Histogram of the residuals")
```



The histogram is left-skewed.

Now lets have a look at the Normal probability plot of the residuals:

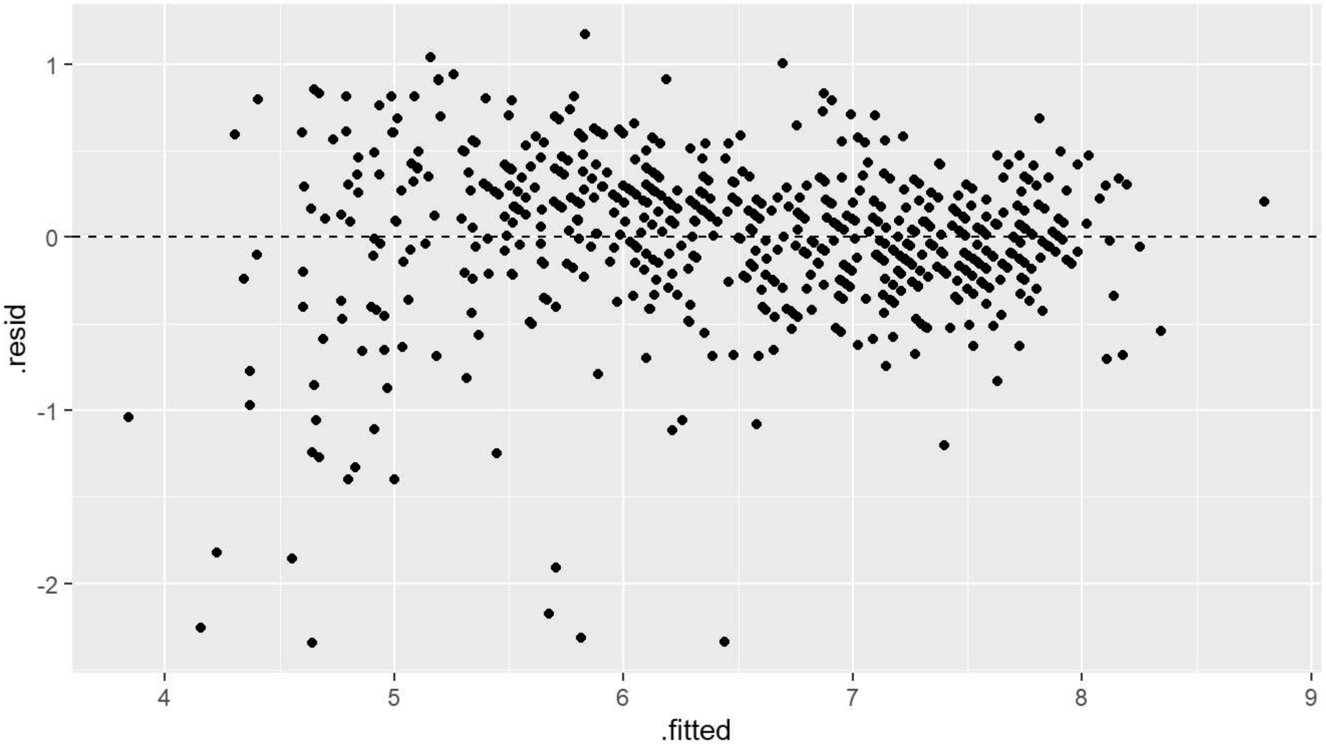
```
qqnorm(final_model$residuals,
        main="Normal probability plot of the residuals")
qqline(final_model$residuals)
```



This plot shows that the skewness is towards the tail.

Now lets look at the residuals vs fitted values plot to check the non-linearity, outliers, variances and any unequal errors:

```
ggplot(data = final_model, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed")
```



From the plot we can see that the variance of the residual is constant. But there are outliers, mostly towards the lower left portion of the plot.

Part 5: Prediction

Now lets test the model to get a prediction for the IMDB rating score for the film, “The Five Heartbeats” staring Michael Wright, Tico Wells and Leon. All variables obtained from IMDB and Rotten Tomatoes.

```
movie_2016 <- data.frame(genre = "Musical & Performing Arts",
                         runtime = 121, mpaa_rating = "R",
                         best_dir_win = "no" ,critics_score = 38,
                         audience_score = 95)
predict(final_model,movie_2016)
```

```
##        1
## 7.433087
```

The predicted rating is 7.43, while the actual rating is 7.5 which is nearly close.

The prediction interval for this movie is:

```
predict(final_model,movie_2016, interval = "predict")
```

```
##        fit     lwr      upr
## 1 7.433087 6.47055 8.395624
```

So the predicted rating score could range from 6.47 to 8.4, which is inclusive of the actual IMDB rating of 7.43 .

Lets look at another movie. “Alice in Wonderland” is a Action and Adventure film starring Johnny Depp and Anne Hathaway. Again, all the variables have been taken from IMDB and Rotten Tomatoes.

```
movie_2016 <- data.frame(genre = "Action & Adventure",
                         runtime = 108, mpaa_rating = "PG",
                         best_dir_win = "No" ,critics_score = 52,
                         audience_score = 55)
predict(final_model,movie_2016)
```

```
##        1
## 6.156282
```

From the model, the predicted IMDB rating score is 6.15, while the actual is 6.5, a discrepancy of about 0.35 points.

Now lets look at the prediction interval:

```
predict(final_model,movie_2016, interval = "predict")
```

```
##        fit      lwr     upr
## 1 6.156282 5.229373 7.08319
```

This interval shows that the predictions could range from 5.22 to 7.08, a range of almost 1.84 points on the rating system. The actual IMDB rating falls within this interval.

The upper and lower values in each predicted interval indicate the 95% Confidence interval (CI). In both cases the predicted and actual value fall within the 95% CI. Hence we can be 95% confident in the accuracy of the predicted IMDB rating.

Part 6: Conclusion

From the data we can conclude that the genre of a movie greatly effects the expectations of the movie and hence the Critic’s and Audience scores and ratings as well. Also Critic’s scores and Audience scores generally do seem to coincide with up to about a 10% difference. This difference may be due to personal preference or the way in which Critics analyze the movie compared to common audiences. The generation of the various models in this project provided the tools to come up with a prediction model for the popularity of a movie to some extent. This prediction used IMDB rating scores as the target variable. As seen above, there are discrepancies between the predicted and actual IMDB scores of the 2 movies tested. In order to improve the prediction capability of the models, several steps can be taken: 1. Movies can be stratified by genre and the IMDB ratings, Critic’s scores and audience scores for each genre can be accurately analyzed to obtain the true statistical variability in each genre. Separate models for each genre can also be created and used in combination with a more general model to obtain more accurate predicted rating scores. 2. A larger dataset can be used to improve on variability of data.