Modeling and prediction for movies

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(corrplot)
library(GGally)
library(gridExtra)
```

Load data

```
load('movies.Rdata')
```

Part 1: Data

The dataset is comprised of a random sample of 651 movies produced and released before 2016. It includes information from two sources: Rotten Tomatoes and IMDB. Since the dataset is a random sample, the results of this study can be generalized to the population of all movies produced and released before 2016. However, we are limited to movies that are documented on Rotten Tomatoes and IMDB. We cannot use this data to establish causal relationships, as the data are observational, rather than being from a randomized experiment.

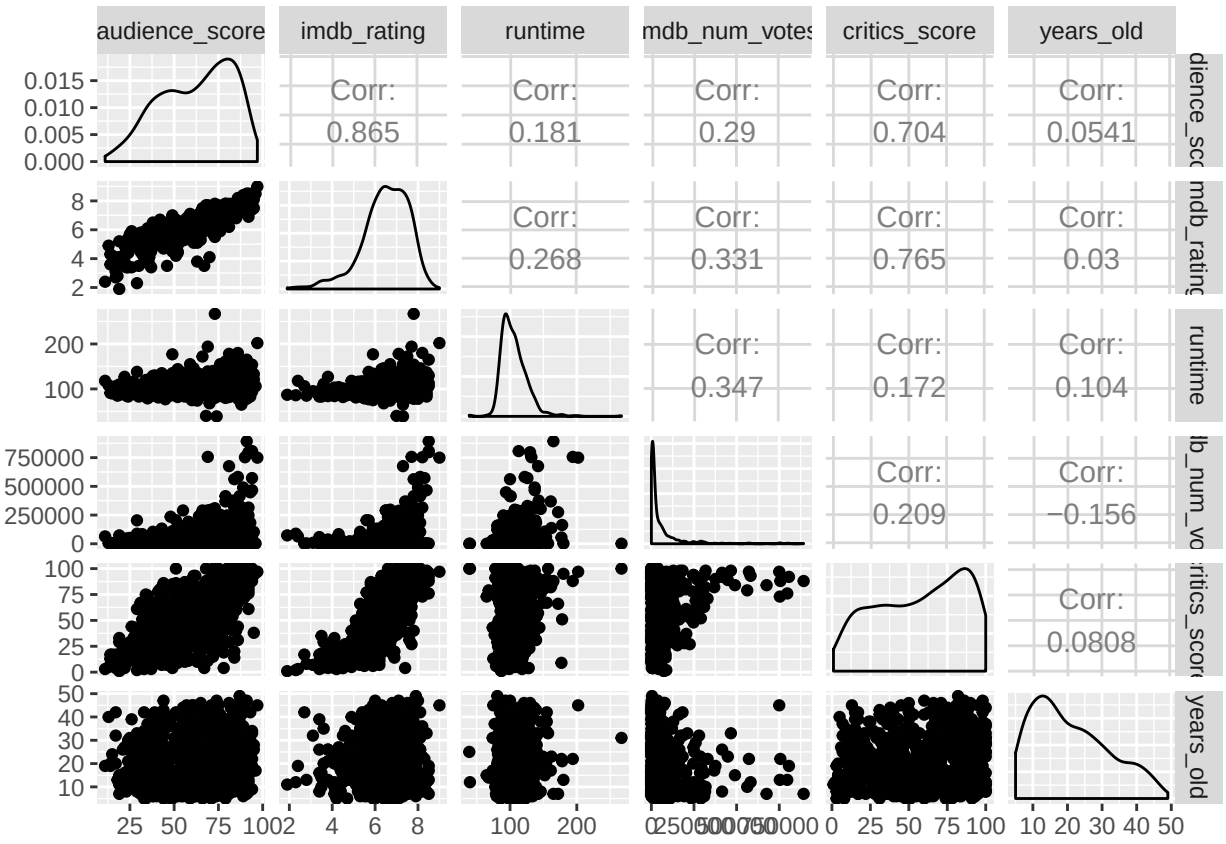
Part 2: Research question

The movie *Avengers: Endgame* was released in April, 2019, and went on to break several box office records, such as the highest grossing opening weekend, bringing in $1,223,641,414 worldwide. It currently has an audience score of 90% on Rotten Tomatoes. With so many variables contributing to the description of a movie (genre, runtime, release month, etc.), what attributes make a movie popular?

Part 3: Exploratory data analysis

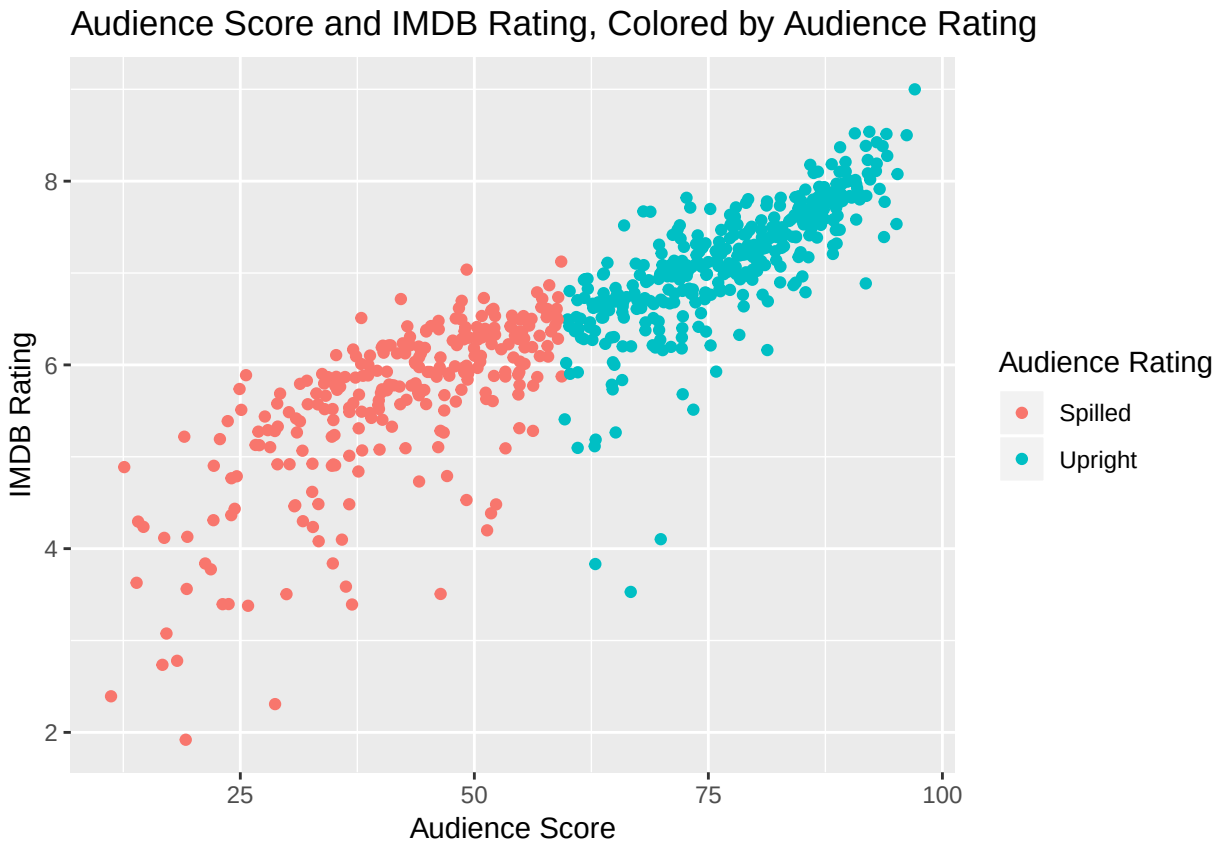
First, I created a new variable, **years_old**, which quantifies how many years old the movie is based on it's theater release year. I then visualized the relationships among the numeric variables, along with their individual distributions, in the following chart.

```
attach(movies)
movies <- mutate(movies, years_old = 2019 - thtr_rel_year)
movies$runtime <- as.numeric(movies$runtime)
num <- select(movies, audience_score, imdb_rating, runtime, imdb_num_votes,
              critics_score, years_old)
ggpairs(num)
```



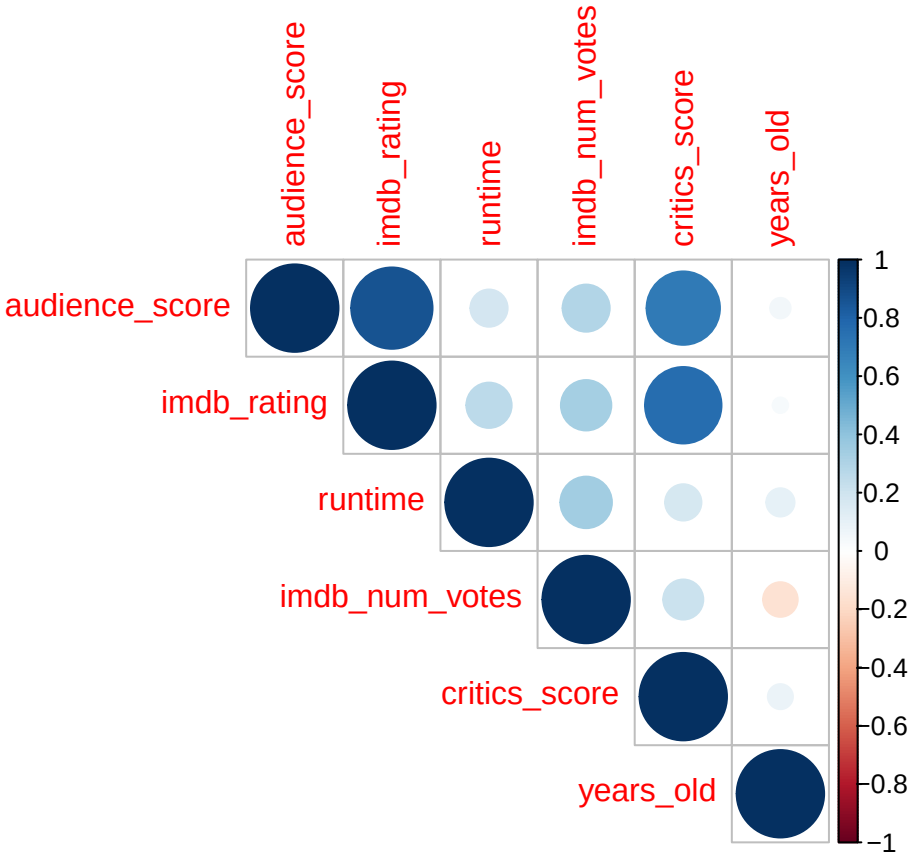
We see a strong, positive correlation between our response variable, **audience_score**, and the predictor variable **imdb_rating**. The below chart visualizes this relationship further, and colors each movie according to its **audience_rating**, either "Spilled" or "Upright".

```
ggplot(movies, aes(x=audience_score, y=imdb_rating, color=audience_rating)) +
  geom_jitter() + xlab('Audience Score') + ylab('IMDB Rating') +
  ggtitle('Audience Score and IMDB Rating, Colored by Audience Rating') +
  labs(color='Audience Rating')
```



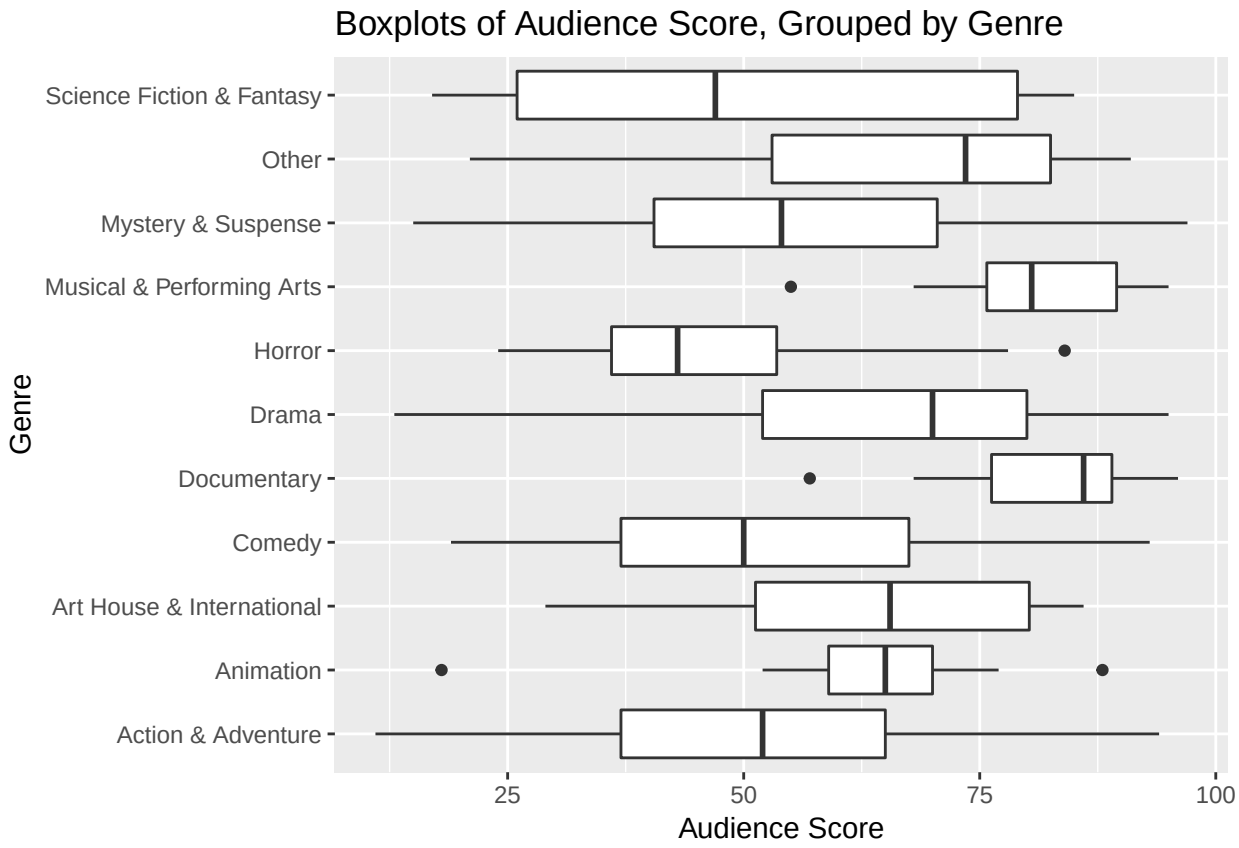
I then visualized the relationships among numeric predictors further in the correlation plot below. In addition to the strong, positive correlations with **audience_score** and both **imdb_rating** and **critics_score**, there is a strong, positive correlation between **critics_score** and **imdb_rating**.

```
cors <- cor(num, method='pearson', use='na.or.complete')
corrplot(cors, type='upper')
```



In the grouped box plot below, we see that the **genre** of "Documentary" has the highest median **audience_score**, followed by "Musical & Performing Arts", while "Horror" has the lowest. "Action & Adventure" has the widest range in **audience_score**.

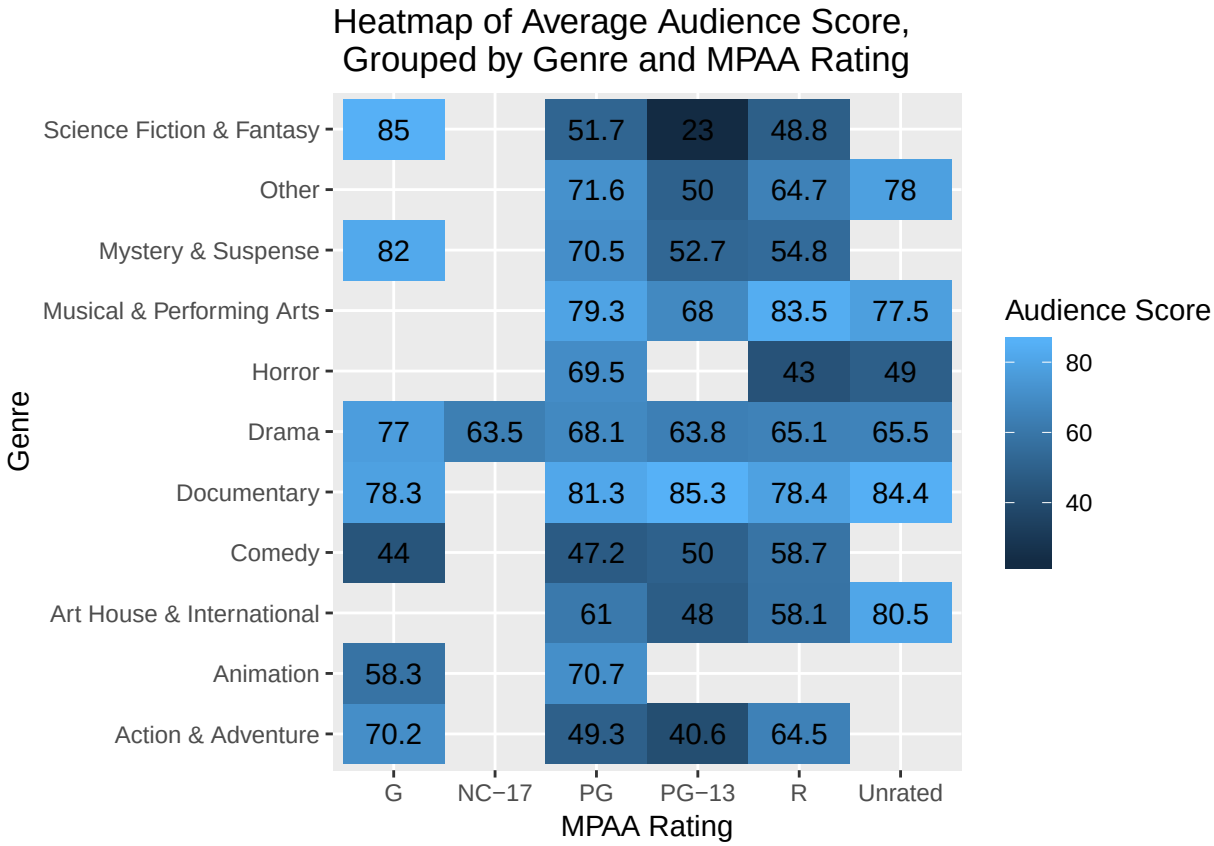
```
ggplot(movies, aes(x=genre, y=audience_score)) + geom_boxplot() + coord_flip() +
  xlab('Genre') + ylab('Audience Score') +
  ggtitle('Boxplots of Audience Score, Grouped by Genre')
```



Continuing our exploration of the **genre** and **audience_score**, we consider these two variables in relation to the **mpaa_rating**, and find that PG-13 Documentaries have the highest score, followed by R Musical & Performing Arts. The lowest scoring combination is PG-13 Science Fiction & Fantasy.

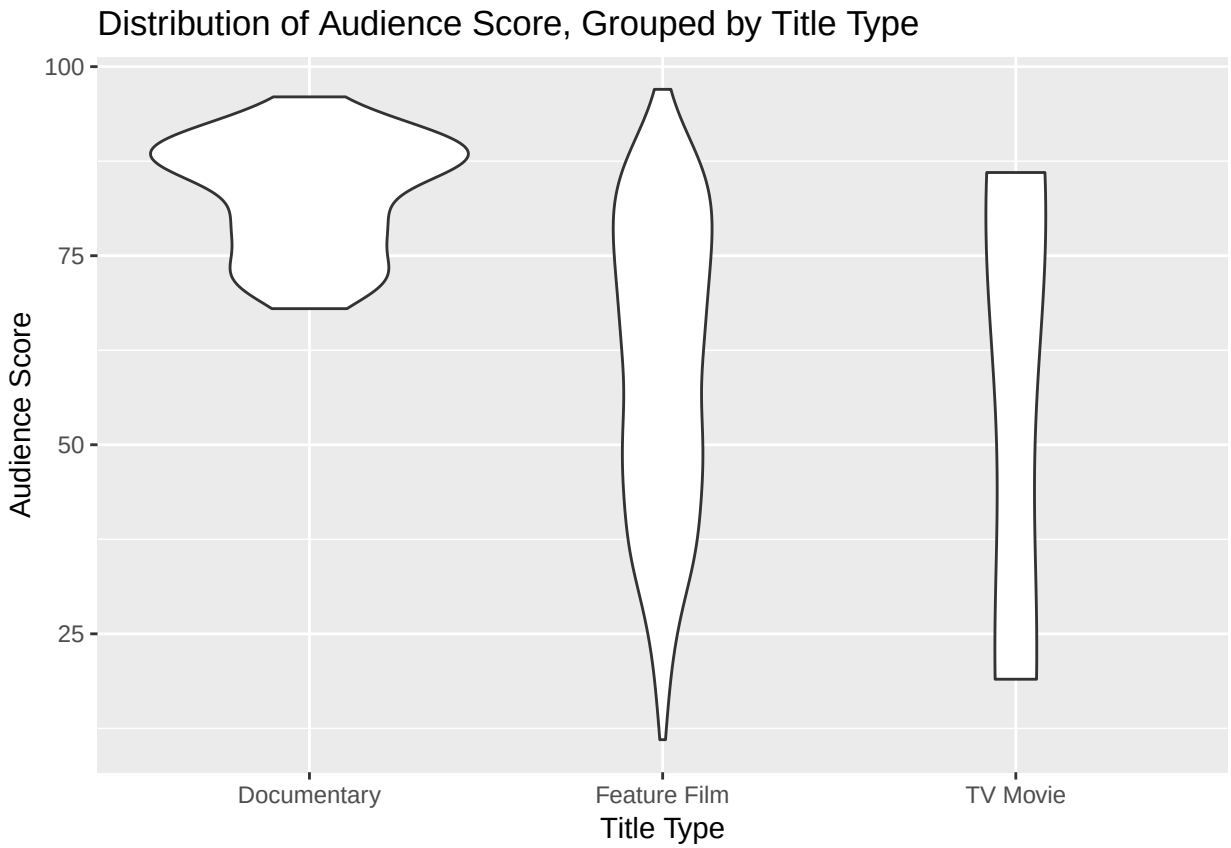
```
data <- movies %>% group_by(mpaa_rating, genre) %>% summarise(avg=mean(audience_score))

ggplot(data, aes(data$mpaa_rating, data$genre)) + geom_tile(aes(fill=data$avg)) +
  xlab('MPAA Rating') + ylab('Genre') +
  ggtitle('Heatmap of Average Audience Score, \n Grouped by Genre and MPAA Rating') +
  geom_text(aes(label = round(data$avg, 1))) +
  labs(fill='Audience Score')
```



This violin chart for the three movie types confirms the popularity of Documentaries that has been displayed in the other visualizations. TV Movies are not as popular as Feature Films and Documentaries.

```
ggplot(movies, aes(title_type, audience_score)) + geom_violin() +
  xlab('Title Type') + ylab('Audience Score') +
  ggtitle('Distribution of Audience Score, Grouped by Title Type')
```



Part 4: Modeling

To select a model, I used the backward stepwise p-value approach, since my question is focused more on model interpretation than predictability. The full model includes 19 of the 33 available variables (including the **years_old** one that I created). It leaves out the title, studio, director, and actor variables as these factors had too many levels for this analysis; it leaves out the URLs as they don't provide any predictive or explanatory power; and it leaves out the years of theater and DVD releases, which is captured in **years_old**. I removed missing data before fitting the model, which reduced the sample size from 651 to 619.

```
movies <- movies %>% na.omit()

movies$thtr_rel_month <- as.character(movies$thtr_rel_month)
movies$thtr_rel_day <- as.character(movies$thtr_rel_day)
movies$dvd_rel_month <- as.character(movies$dvd_rel_month)
movies$dvd_rel_day <- as.character(movies$dvd_rel_day)

model <- lm(audience_score ~ genre + runtime + mpaa_rating + years_old + thtr_rel_month +
              thtr_rel_day + dvd_rel_month + dvd_rel_day + imdb_rating + imdb_num_votes +
              critics_rating + critics_score + audience_rating + best_pic_nom +
              best_pic_win + best_actor_win + best_actress_win + best_dir_win + top200_box,
            data = movies)
summary(model)
```

```
##
## Call:
## lm(formula = audience_score ~ genre + runtime + mpaa_rating +
```

```
##     years_old + thtr_rel_month + thtr_rel_day + dvd_rel_month +
##     dvd_rel_day + imdb_rating + imdb_num_votes + critics_rating +
##     critics_score + audience_rating + best_pic_nom + best_pic_win +
##     best_actor_win + best_actress_win + best_dir_win + top200_box,
##     data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.5452  -4.1679   0.2606   3.8573  21.1121
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -4.364e+00  4.458e+00  -0.979   0.3282
## genreAnimation                  4.819e+00  3.091e+00   1.559   0.1196
## genreArt House & International -2.150e+00  2.442e+00  -0.880   0.3791
## genreComedy                     2.157e+00  1.257e+00   1.716   0.0868 .
## genreDocumentary                1.657e+00  1.863e+00   0.889   0.3742
## genreDrama                      1.664e-02  1.127e+00   0.015   0.9882
## genreHorror                    -1.363e+00  1.906e+00  -0.715   0.4748
## genreMusical & Performing Arts  2.441e+00  2.461e+00   0.992   0.3218
## genreMystery & Suspense        -2.442e+00  1.446e+00  -1.689   0.0919 .
## genreOther                      3.843e-01  2.173e+00   0.177   0.8597
## genreScience Fiction & Fantasy  1.335e-01  2.771e+00   0.048   0.9616
## runtime                        -3.258e-02  1.965e-02  -1.658   0.0979 .
## mpaa_ratingNC-17               -1.426e+01  7.957e+00  -1.792   0.0737 .
## mpaa_ratingPG                  -1.009e+00  2.172e+00  -0.464   0.6426
## mpaa_ratingPG-13               -1.624e+00  2.256e+00  -0.720   0.4719
## mpaa_ratingR                   -1.725e+00  2.156e+00  -0.800   0.4240
## mpaa_ratingUnrated             -4.635e-02  2.578e+00  -0.018   0.9857
## years_old                       4.427e-02  3.431e-02   1.290   0.1976
## thtr_rel_month10               -1.640e+00  1.413e+00  -1.161   0.2463
## thtr_rel_month11               -7.830e-01  1.557e+00  -0.503   0.6152
## thtr_rel_month12               -4.810e-01  1.429e+00  -0.337   0.7366
## thtr_rel_month2                -5.793e-01  1.690e+00  -0.343   0.7319
## thtr_rel_month3                -4.593e-01  1.520e+00  -0.302   0.7626
## thtr_rel_month4                -3.536e-01  1.567e+00  -0.226   0.8215
## thtr_rel_month5                 6.457e-01  1.546e+00   0.418   0.6763
## thtr_rel_month6                -2.248e-01  1.351e+00  -0.166   0.8679
## thtr_rel_month7                -1.406e+00  1.532e+00  -0.918   0.3590
## thtr_rel_month8                -2.090e+00  1.605e+00  -1.302   0.1934
## thtr_rel_month9                 4.884e-01  1.542e+00   0.317   0.7516
## thtr_rel_day10                 -4.040e-01  1.846e+00  -0.219   0.8269
## thtr_rel_day11                 -1.332e-01  1.816e+00  -0.073   0.9416
## thtr_rel_day12                  3.180e-02  1.937e+00   0.016   0.9869
## thtr_rel_day13                 -1.774e+00  2.096e+00  -0.846   0.3978
## thtr_rel_day14                  2.708e+00  2.256e+00   1.200   0.2306
## thtr_rel_day15                  1.785e+00  1.776e+00   1.005   0.3153
## thtr_rel_day16                  2.751e+00  1.960e+00   1.403   0.1611
## thtr_rel_day17                 -1.076e+00  1.881e+00  -0.572   0.5675
## thtr_rel_day18                  2.262e+00  2.034e+00   1.112   0.2666
## thtr_rel_day19                 -1.421e+00  1.785e+00  -0.796   0.4263
## thtr_rel_day2                  -1.008e+00  2.315e+00  -0.435   0.6635
## thtr_rel_day20                  1.811e+00  1.859e+00   0.974   0.3303
## thtr_rel_day21                  1.601e+00  1.840e+00   0.870   0.3848
```

```
## thtr_rel_day22                6.968e-01  1.838e+00   0.379   0.7049
## thtr_rel_day23                2.466e+00  2.125e+00   1.161   0.2463
## thtr_rel_day24                2.442e+00  2.392e+00   1.021   0.3078
## thtr_rel_day25               -1.147e+00  1.973e+00  -0.581   0.5613
## thtr_rel_day26                6.991e-01  2.212e+00   0.316   0.7521
## thtr_rel_day27                6.223e-01  2.022e+00   0.308   0.7584
## thtr_rel_day28               -2.764e+00  2.194e+00  -1.260   0.2082
## thtr_rel_day29                2.575e+00  2.022e+00   1.274   0.2034
## thtr_rel_day3                -7.886e-02  2.287e+00  -0.034   0.9725
## thtr_rel_day30               -1.267e+00  2.146e+00  -0.590   0.5551
## thtr_rel_day31                4.295e+00  3.155e+00   1.361   0.1740
## thtr_rel_day4                 1.667e+00  2.429e+00   0.687   0.4927
## thtr_rel_day5                 1.297e+00  1.971e+00   0.658   0.5109
## thtr_rel_day6                 3.245e+00  1.979e+00   1.640   0.1017
## thtr_rel_day7                -4.280e-01  1.857e+00  -0.230   0.8179
## thtr_rel_day8                 1.499e+00  1.877e+00   0.799   0.4249
## thtr_rel_day9                 4.493e-01  2.022e+00   0.222   0.8242
## dvd_rel_month10              -2.202e+00  1.558e+00  -1.413   0.1581
## dvd_rel_month11              -1.883e+00  1.655e+00  -1.138   0.2557
## dvd_rel_month12               2.003e-01  1.579e+00   0.127   0.8991
## dvd_rel_month2               -1.894e+00  1.563e+00  -1.212   0.2261
## dvd_rel_month3               -1.268e+00  1.495e+00  -0.848   0.3969
## dvd_rel_month4               -2.517e+00  1.597e+00  -1.576   0.1156
## dvd_rel_month5               -1.817e+00  1.521e+00  -1.195   0.2327
## dvd_rel_month6                4.678e-01  1.548e+00   0.302   0.7627
## dvd_rel_month7               -4.523e-01  1.627e+00  -0.278   0.7811
## dvd_rel_month8               -1.107e+00  1.620e+00  -0.683   0.4948
## dvd_rel_month9                5.297e-01  1.570e+00   0.337   0.7360
## dvd_rel_day10                -5.089e-02  2.173e+00  -0.023   0.9813
## dvd_rel_day11                -1.604e+00  2.157e+00  -0.744   0.4575
## dvd_rel_day12                 1.202e+00  2.307e+00   0.521   0.6027
## dvd_rel_day13                -3.757e+00  2.157e+00  -1.742   0.0821 .
## dvd_rel_day14                -3.972e+00  2.623e+00  -1.514   0.1305
## dvd_rel_day15                -1.320e+00  1.986e+00  -0.665   0.5065
## dvd_rel_day16                -1.892e+00  1.859e+00  -1.017   0.3095
## dvd_rel_day17                 3.798e+00  2.154e+00   1.763   0.0785 .
## dvd_rel_day18                -1.548e-02  1.991e+00  -0.008   0.9938
## dvd_rel_day19                -3.457e+00  2.241e+00  -1.543   0.1235
## dvd_rel_day2                  1.370e+00  2.160e+00   0.634   0.5262
## dvd_rel_day20                 2.364e-01  2.143e+00   0.110   0.9122
## dvd_rel_day21                -1.801e+00  2.243e+00  -0.803   0.4225
## dvd_rel_day22                 1.583e+00  2.281e+00   0.694   0.4880
## dvd_rel_day23                -2.271e+00  2.192e+00  -1.036   0.3006
## dvd_rel_day24                -9.268e-01  2.114e+00  -0.438   0.6612
## dvd_rel_day25                -4.169e+00  2.647e+00  -1.575   0.1160
## dvd_rel_day26                -3.058e+00  2.039e+00  -1.499   0.1344
## dvd_rel_day27                -8.067e-01  2.027e+00  -0.398   0.6909
## dvd_rel_day28                -1.582e+00  2.357e+00  -0.671   0.5024
## dvd_rel_day29                 2.424e-01  2.191e+00   0.111   0.9119
## dvd_rel_day3                 -1.197e-01  2.200e+00  -0.054   0.9566
## dvd_rel_day30                -2.290e+00  2.131e+00  -1.074   0.2831
## dvd_rel_day31                 4.933e-01  2.794e+00   0.177   0.8599
## dvd_rel_day4                 -2.497e-01  2.094e+00  -0.119   0.9051
## dvd_rel_day5                 -5.454e-01  2.217e+00  -0.246   0.8057
```

```
## dvd_rel_day6                  -6.631e-01  1.904e+00  -0.348   0.7277
## dvd_rel_day7                  -8.043e-01  1.988e+00  -0.405   0.6859
## dvd_rel_day8                  -1.558e-01  2.149e+00  -0.073   0.9422
## dvd_rel_day9                   4.660e-01  2.222e+00   0.210   0.8340
## imdb_rating                    9.352e+00  5.404e-01  17.305   <2e-16 ***
## imdb_num_votes                 2.408e-06  3.582e-06   0.672   0.5017
## critics_ratingFresh           -4.815e-02  9.920e-01  -0.049   0.9613
## critics_ratingRotten          -1.044e+00  1.557e+00  -0.670   0.5031
## critics_score                 -2.728e-03  2.776e-02  -0.098   0.9218
## audience_ratingUpright         2.072e+01  8.776e-01  23.608   <2e-16 ***
## best_pic_nomyes                4.968e+00  1.997e+00   2.488   0.0132 *
## best_pic_winyes               -4.086e+00  3.520e+00  -1.161   0.2463
## best_actor_winyes              3.459e-02  9.046e-01   0.038   0.9695
## best_actress_winyes           -1.064e+00  9.853e-01  -1.080   0.2808
## best_dir_winyes                1.082e+00  1.295e+00   0.835   0.4040
## top200_boxyes                 -1.906e+00  2.079e+00  -0.917   0.3597
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.888 on 507 degrees of freedom
## Multiple R-squared:  0.9044, Adjusted R-squared:  0.8835
## F-statistic: 43.22 on 111 and 507 DF,  p-value: < 2.2e-16
```

Starting with the full model above, I removed the variable with the highest p-value and refit the model. I repeated this process until all of the remaining predictors had a significant p-value. This approach resulted in a model with an adjusted R^2 of 0.8847 that is comprised of 6 predictors: **genre**, **runtime**, **thtr_rel_month**, **imdb_rating**, **audience_rating**, **best_pic_nom**. These attributes are the most significant in determining a movie's popularity, and we can interpret them as follows, based on the output of the regression:

- All else held constant, when the **genre** is Mystery & Suspense, the audience score decreases by -3.19626 on average.
- All else held constant, for every 1 minute increase in **runtime**, the audience score decreases by -0.03448 on average.
- All else held constant, when the **thtr_rel_month** is 8 (August), the audience score decreases by -2.91225 on average.
- All else held constant, for every 1 point increase in **imdb_rating**, the audience score increases by 9.77714 on average.
- All else held constant, when the **audience_rating** is Upright, the audience score increases by 20.42074 on average.
- All else held constant, when the **best_pic_nom** is "yes", the audience score increases by 3.45914 on average.

The intercept is negative, which doesn't provide any insight in this case, as the audience score cannot be below 0.

```
model <- lm(audience_score ~ genre + runtime + thtr_rel_month + imdb_rating +
              audience_rating + best_pic_nom,
            data = movies)
summary(model)
```

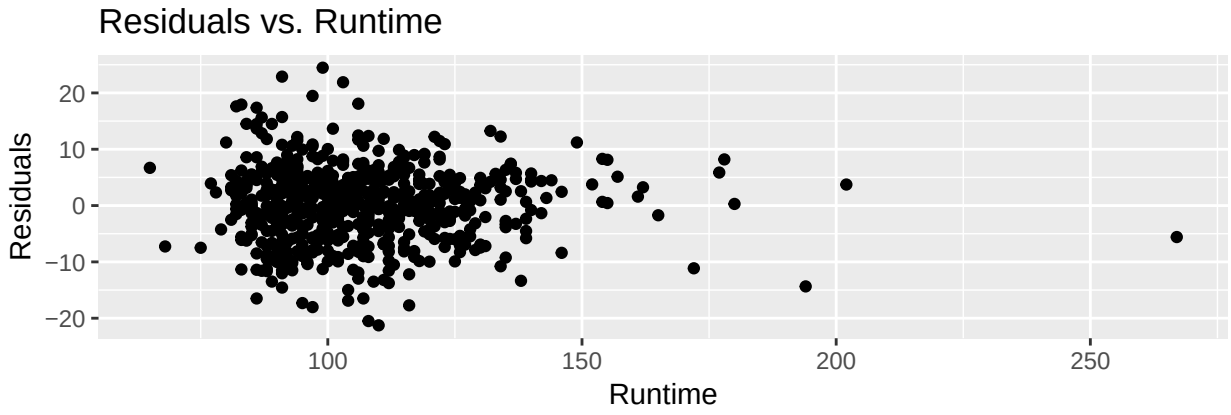
```
##
## Call:
## lm(formula = audience_score ~ genre + runtime + thtr_rel_month +
##     imdb_rating + audience_rating + best_pic_nom, data = movies)
##
```

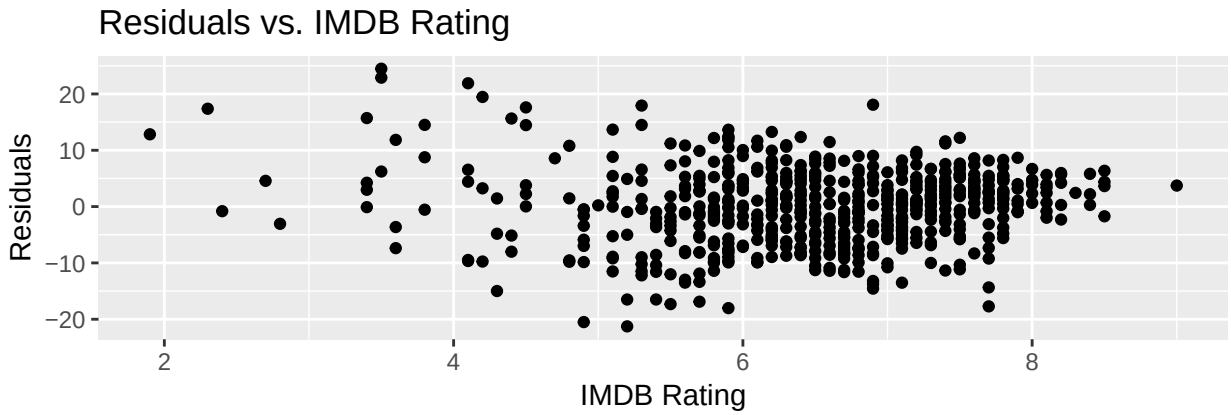
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2670  -4.5200   0.4443   4.4107  24.4784
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -7.89853    2.76999  -2.851   0.0045 **
## genreAnimation                   4.78980    2.63555   1.817   0.0697 .
## genreArt House & International -2.31003    2.18743  -1.056   0.2914
## genreComedy                      1.39935    1.15468   1.212   0.2260
## genreDocumentary                 1.33304    1.49457   0.892   0.3728
## genreDrama                      -0.67155    0.98555  -0.681   0.4959
## genreHorror                     -1.77912    1.72946  -1.029   0.3040
## genreMusical & Performing Arts   2.79975    2.21251   1.265   0.2062
## genreMystery & Suspense         -3.19626    1.29118  -2.475   0.0136 *
## genreOther                      -0.12908    2.01507  -0.064   0.9489
## genreScience Fiction & Fantasy   0.41181    2.59015   0.159   0.8737
## runtime                         -0.03448    0.01666  -2.069   0.0390 *
## thtr_rel_month10                -2.11300    1.21597  -1.738   0.0828 .
## thtr_rel_month11                -2.32335    1.33683  -1.738   0.0827 .
## thtr_rel_month12                -0.54975    1.26078  -0.436   0.6630
## thtr_rel_month2                 -0.28196    1.49463  -0.189   0.8504
## thtr_rel_month3                 -1.68735    1.31036  -1.288   0.1984
## thtr_rel_month4                 -1.06642    1.36174  -0.783   0.4339
## thtr_rel_month5                  0.31313    1.36337   0.230   0.8184
## thtr_rel_month6                 -0.71461    1.19034  -0.600   0.5485
## thtr_rel_month7                 -1.74541    1.34545  -1.297   0.1950
## thtr_rel_month8                 -2.91225    1.37370  -2.120   0.0344 *
## thtr_rel_month9                 -0.89404    1.29523  -0.690   0.4903
## imdb_rating                      9.77714    0.39372  24.832   <2e-16 ***
## audience_ratingUpright          20.42074    0.79947  25.543   <2e-16 ***
## best_pic_nomyes                  3.45914    1.61739   2.139   0.0329 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.853 on 593 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8847
## F-statistic: 190.6 on 25 and 593 DF,  p-value: < 2.2e-16
```

Next, we perform diagnostics for our MLR model to check if they support the model assumptions.

I) The plots below show a random scatter around 0, indicating that there is no linear relationship between the numeric predictors and audience score.

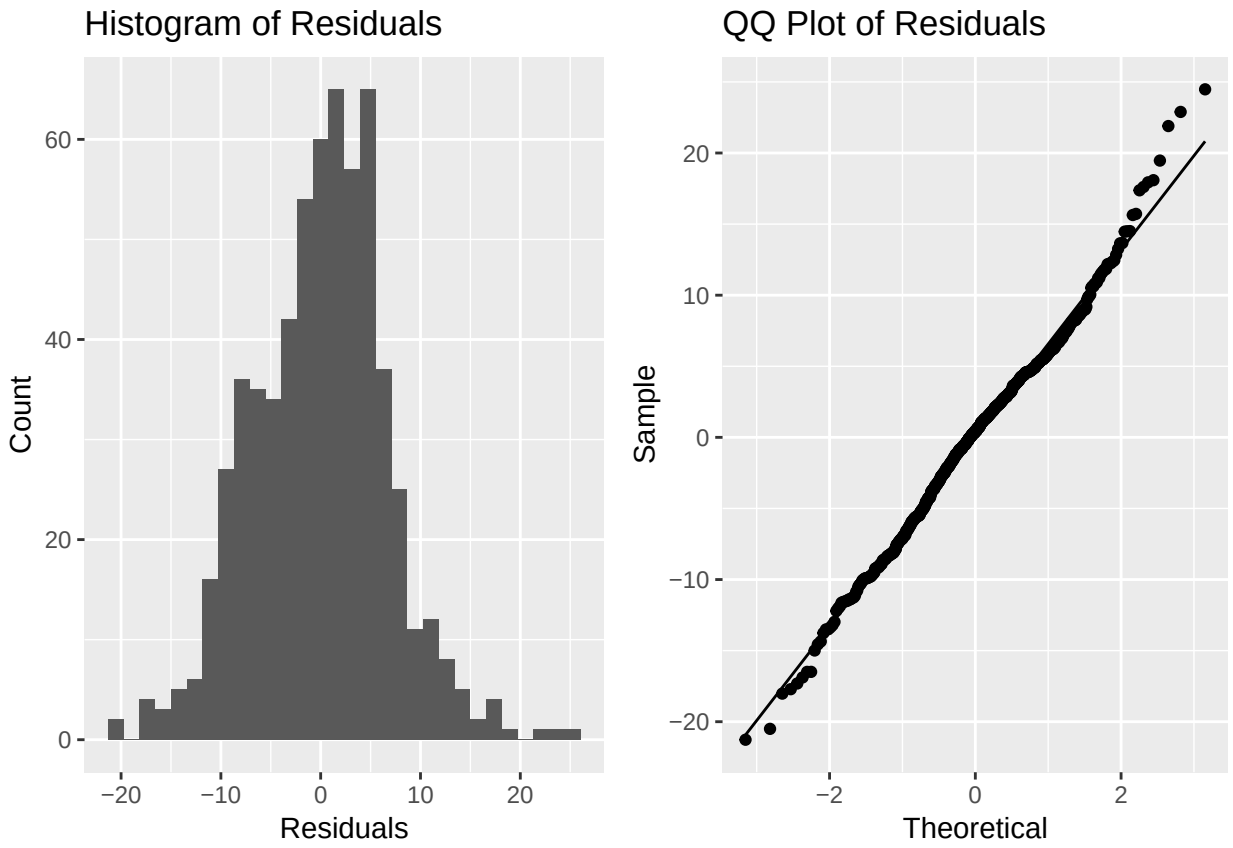
```
resid <- data.frame(Residuals=model$residuals, Runtime=movies$runtime,
                    IMDB=movies$imdb_rating, Fitted=model$fitted)
plot1 <- ggplot(resid, aes(Runtime, Residuals)) + geom_point() +
  ggtitle('Residuals vs. Runtime')
plot2 <- ggplot(resid, aes(IMDB, Residuals)) + geom_point() +
  ggtitle('Residuals vs. IMDB Rating') + xlab('IMDB Rating')
grid.arrange(plot1, plot2, ncol=1)
```





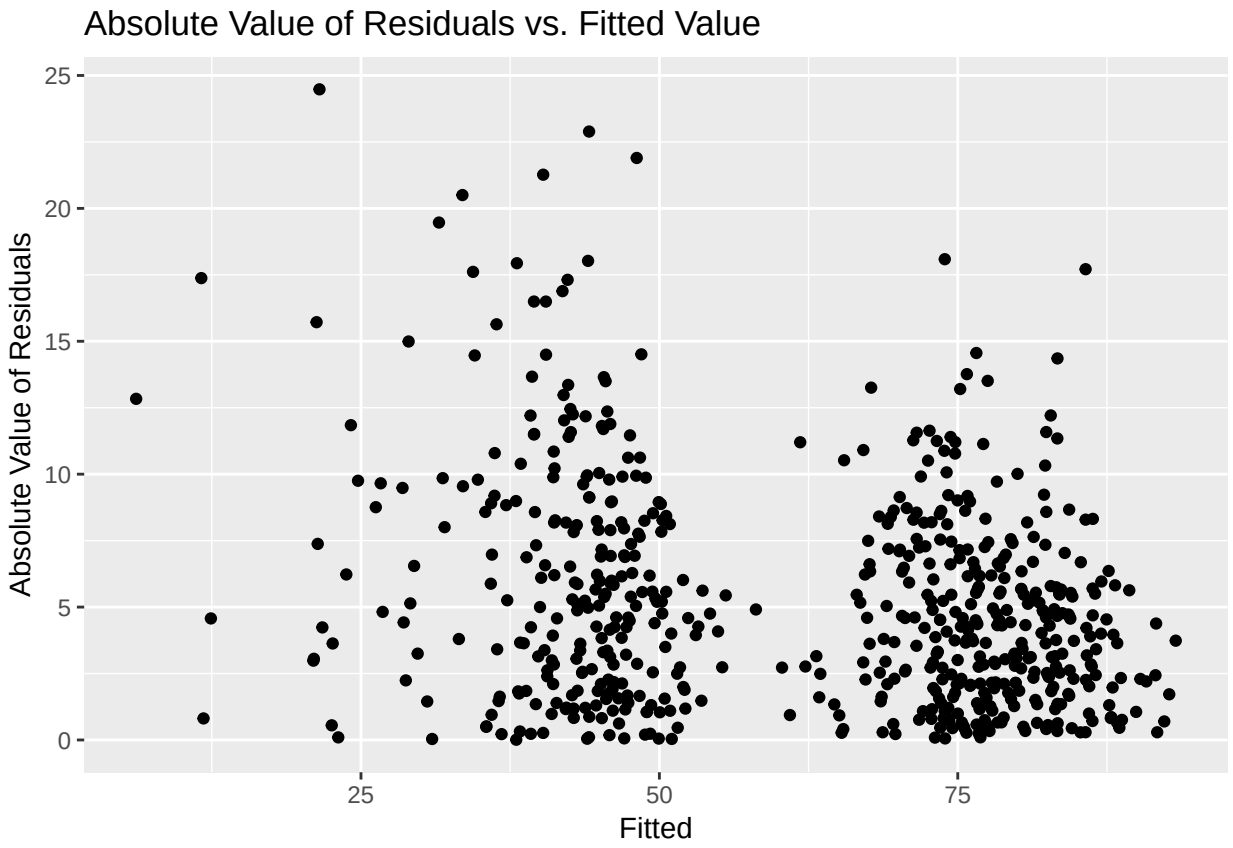
II) The histogram and QQ plot of residuals below show that the residuals are nearly normal with mean 0. There are some minor irregularities, but nothing to be cause for concern.

```
plot1 <- ggplot(resid, aes(Residuals)) + geom_histogram() +
  ggtitle('Histogram of Residuals') + ylab('Count')
plot2 <- ggplot(resid, aes(sample=Residuals))+ stat_qq() + stat_qq_line() +
  xlab('Theoretical') + ylab('Sample') + ggtitle('QQ Plot of Residuals')
grid.arrange(plot1, plot2, ncol=2)
```



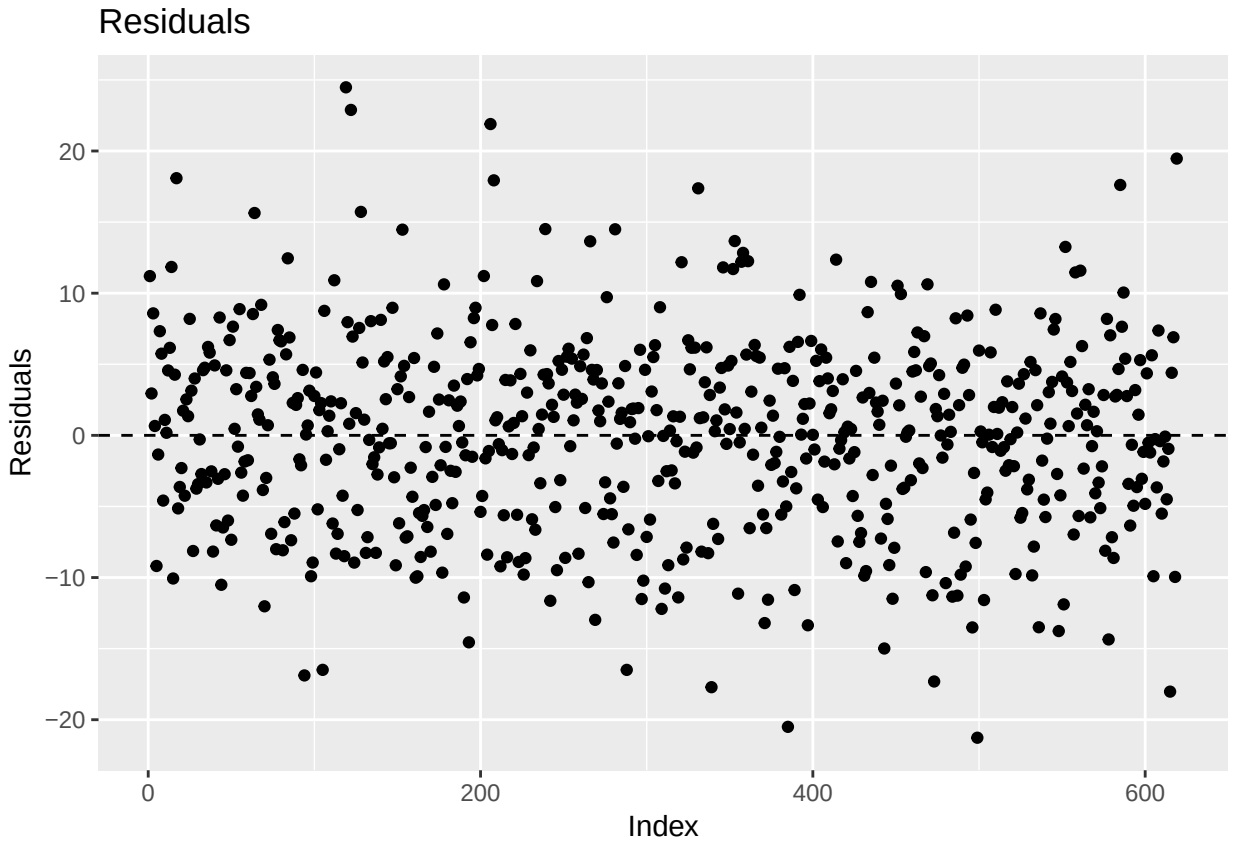
III) The constant variability of residuals is shown in the below plot, which displays a random scatter of constant width around 0.

```
ggplot(resid, aes(Fitted, abs(Residuals))) + geom_point() +
  ylab('Absolute Value of Residuals') +
  ggtitle('Absolute Value of Residuals vs. Fitted Value')
```



IV) The below chart indicates the residuals are independent, which means the observations are independent.

```
ggplot(resid, aes(1:length(Residuals), Residuals)) + geom_point() + xlab('Index') +
  ggtitle('Residuals') + geom_hline(yintercept=0, linetype='dashed')
```



Part 5: Prediction

In keeping with the Marvel movie theme, I made a prediction for *Captain America: Civil War* which was released in May 2016. The true audience score of this movie on Rotten Tomatoes is 89%. The data comes from Rotten Tomatoes and IMDB.

```
newmovie <- data.frame(genre='Action & Adventure',
                       runtime=146,
                       thtr_rel_month='5',
                       imdb_rating=7.8,
                       audience_rating='Upright',
                       best_pic_nom='no')

predict(model, newmovie, interval = "prediction", level = 0.95)
```

```
##        fit     lwr      upr
## 1 84.06268 70.2525 97.87286
```

The predicted audience score is 84%, which is a slight underestimate of the true audience score, 89%. However, the true score is captured in the 95% prediction interval of [71, 97].

Part 6: Conclusion

The exploratory data analysis revealed the variation of popularity among the different genres, ratings, and title types. For instance, we noticed documentaries were more highly rated than other genres and title types. It

also showed correlations between different predictors and the response, audience score. The regression model quantified what attributes make a movie popular: **genre**, **runtime**, **thtr_rel_month**, **imdb_rating**, **audience_rating**, **best_pic_nom**. One shortcoming is that a leading predictor, **audience_rating**, is closely related to **audience_score** by definition. With an R^2 of 0.8847, there is room for improvement in the model, which could be remedied by increasing our sample size of 651. In addition to collecting more data, future studies could try nonlinear techniques and make predictions for more movies to get a better sense of predictive accuracy.