Intro to Data Analytics - Final Project - Do Website Media Ratings and Awards Earned indicate the quality of a show?

John Salmon

2023-12-11

Introduction

Awards and Ratings for TV shows and Movies are Important metrics. When deciding what to watch or whether or not we want to watch something these are often metrics that run through our heads. For this reason its important that these metrics mean something. In this project I am interested in examining website review scores and information on awards received and nominated for to see how they compare with each other and whether they should be used as valid metrics of quality.

```
netflix <- read.csv("netflix.csv")
```

Dataset Information

The netflix.csv file was sourced from Kaggle.com. It contains data from the Netflix website, the Internet Movie Database(IMDb), The user aggregate scoring site Rotten Tomatoes, and scores from MetaCritic. According to the creator many of the data set the movie review website API's were used. I am not sure how the data from netflix was acquired as Netflix does not have an official API.

The dataset can be found here: https://www.kaggle.com/datasets/ashishgup/netflix-rotten-tomatoes-metacritic-imd

```
print(c("There are ", sum(is.na(netflix)), " missing values in the netflix dataset."))
```

```
## [1] "There are "
## [2] "43767"
## [3] " missing values in the netflix dataset."
```

```
print(c("The IMDb column has ", sum(is.na(netflix$Rotten.Tomatoes.Score)), " missing values. Meta Criti
```

```
## [1] "The IMDb column has "              "9098"
## [3] " missing values. Meta Critic has " "11144"
## [5] " missing values."
```

```
netflix <- netflix[!is.na(netflix$Rotten.Tomatoes.Score),]##Adds only columns with data
netflix <- netflix[!is.na(netflix$Metacritic.Score),]##does the same for
netflix <- netflix[!is.na(netflix$IMDb.Score),]
netflix <- netflix[!duplicated(netflix$Title),]
```

```r
##and lets replace the NA's for awards earned with zero's
netflix[is.na(netflix$Awards.Nominated.For),] <- 0
netflix[is.na(netflix$Awards.Received),] <- 0
```

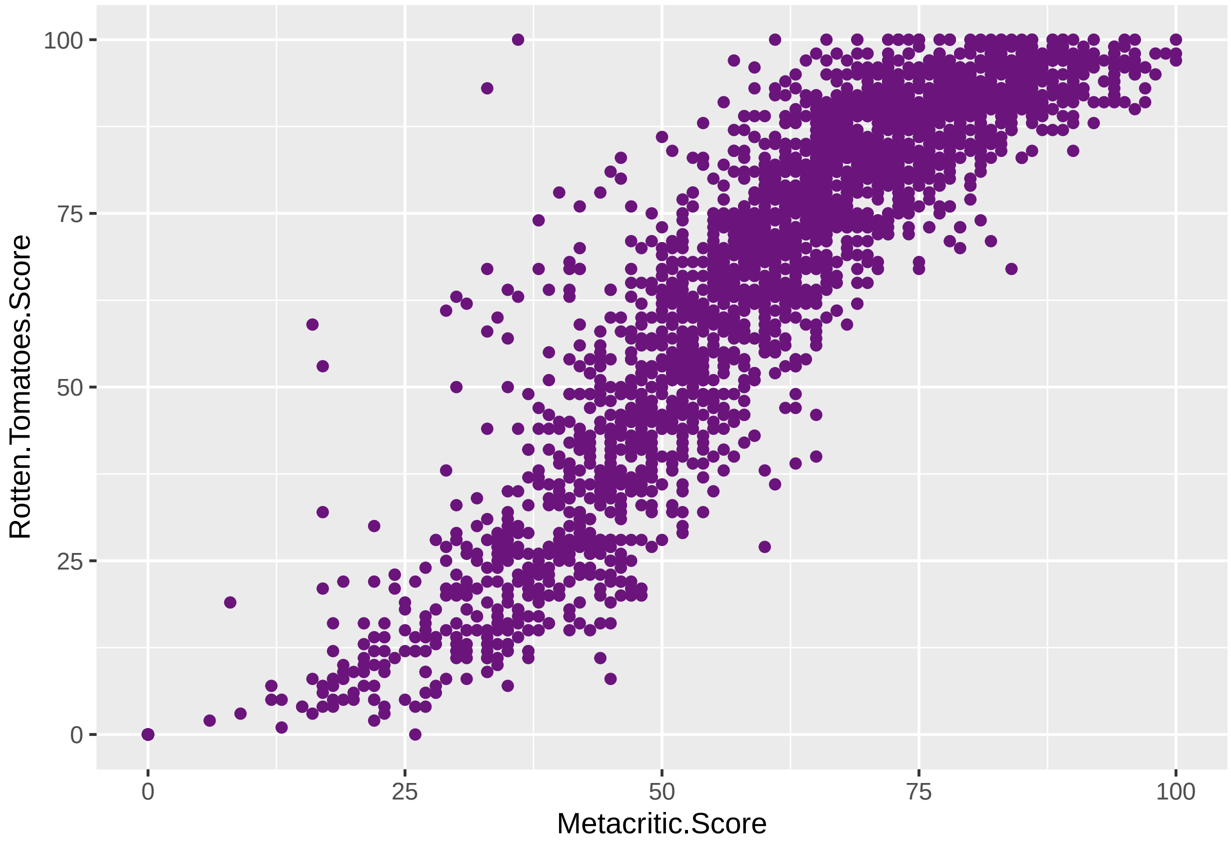
Cleaning The Data

According to the is.na() of the dataset there are 43,767 missing values in the entire dataset. This looks like a scary number but this project is mainly interested in looking at the Rotten Tomatoes, Meta Critic, and IMDb scores, as well as the Awards Information. Which roughly halves the missing values. These missing values need to be dealt with. The missing data is out of scope to collect and patch into the data set at the moment so for this analysis I am choosing to remove rows with missing data. Duplicate data is slightly different, because we are dealing with scores on a scale there will be duplicate numerical values, but movie titles tend to be more unique and can be identified and removed based on title to account for any cases where a movie was accidently entered into the data set twice.

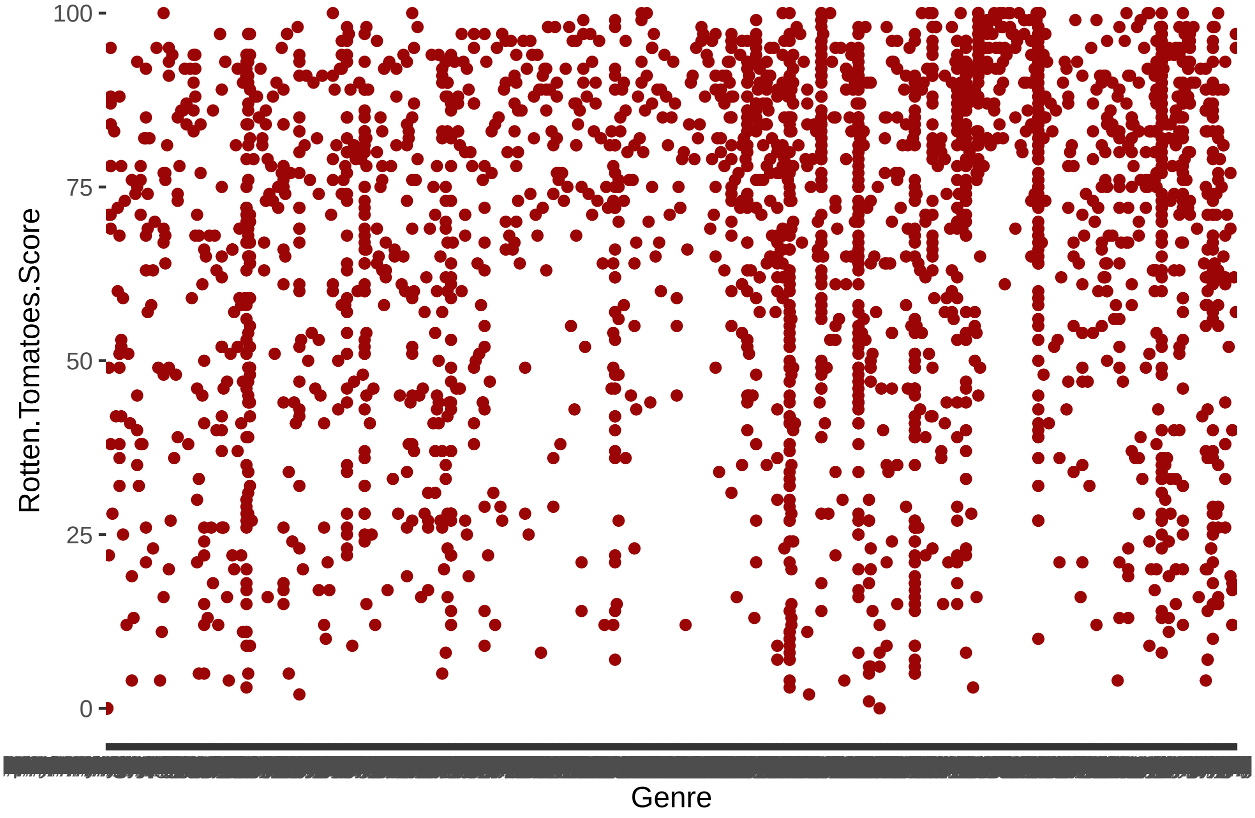
```r
summary(netflix)
```

```
##     Title              Genre               Tags            Languages
##  Length:3834        Length:3834        Length:3834        Length:3834
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Series.or.Movie    Hidden.Gem.Score Country.Availability   Runtime
##  Length:3834        Min.   :0.000    Length:3834          Length:3834
##  Class :character   1st Qu.:0.000    Class :character     Class :character
##  Mode  :character   Median :2.900    Mode  :character     Mode  :character
##                     Mean   :2.537
##                     3rd Qu.:3.900
##                     Max.   :9.200
##    Director            Writer              Actors            View.Rating
##  Length:3834        Length:3834        Length:3834        Length:3834
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    IMDb.Score    Rotten.Tomatoes.Score Metacritic.Score Awards.Received
##  Min.   :0.000   Min.   :  0.00        Min.   :  0.00   Min.   :  0.000
##  1st Qu.:0.000   1st Qu.:  0.00        1st Qu.:  0.00   1st Qu.:  0.000
##  Median :6.300   Median : 52.00        Median : 51.00   Median :  2.000
##  Mean   :4.612   Mean   : 45.87        Mean   : 41.79   Mean   :  8.655
##  3rd Qu.:7.200   3rd Qu.: 82.00        3rd Qu.: 68.75   3rd Qu.:  8.000
##  Max.   :9.300   Max.   :100.00        Max.   :100.00   Max.   :300.000
##  Awards.Nominated.For  Boxoffice          Release.Date
##  Min.   :  0.00       Length:3834        Length:3834
##  1st Qu.:  0.00       Class :character   Class :character
##  Median :  5.00       Mode  :character   Mode  :character
##  Mean   : 16.87
##  3rd Qu.: 16.00
##  Max.   :355.00
##  Netflix.Release.Date Production.House   Netflix.Link        IMDb.Link
##  Length:3834          Length:3834        Length:3834        Length:3834
##  Class :character     Class :character   Class :character   Class :character
##  Mode  :character     Mode  :character   Mode  :character   Mode  :character
##
##
##
##    Summary             IMDb.Votes         Image              Poster
##  Length:3834        Min.   :      0   Length:3834        Length:3834
##  Class :character   1st Qu.:      0   Class :character   Class :character
##  Mode  :character   Median :  24773   Mode  :character   Mode  :character
##                     Mean   : 110643
##                     3rd Qu.: 131814
##                     Max.   :2354197
##  TMDb.Trailer       Trailer.Site
##  Length:3834        Length:3834
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

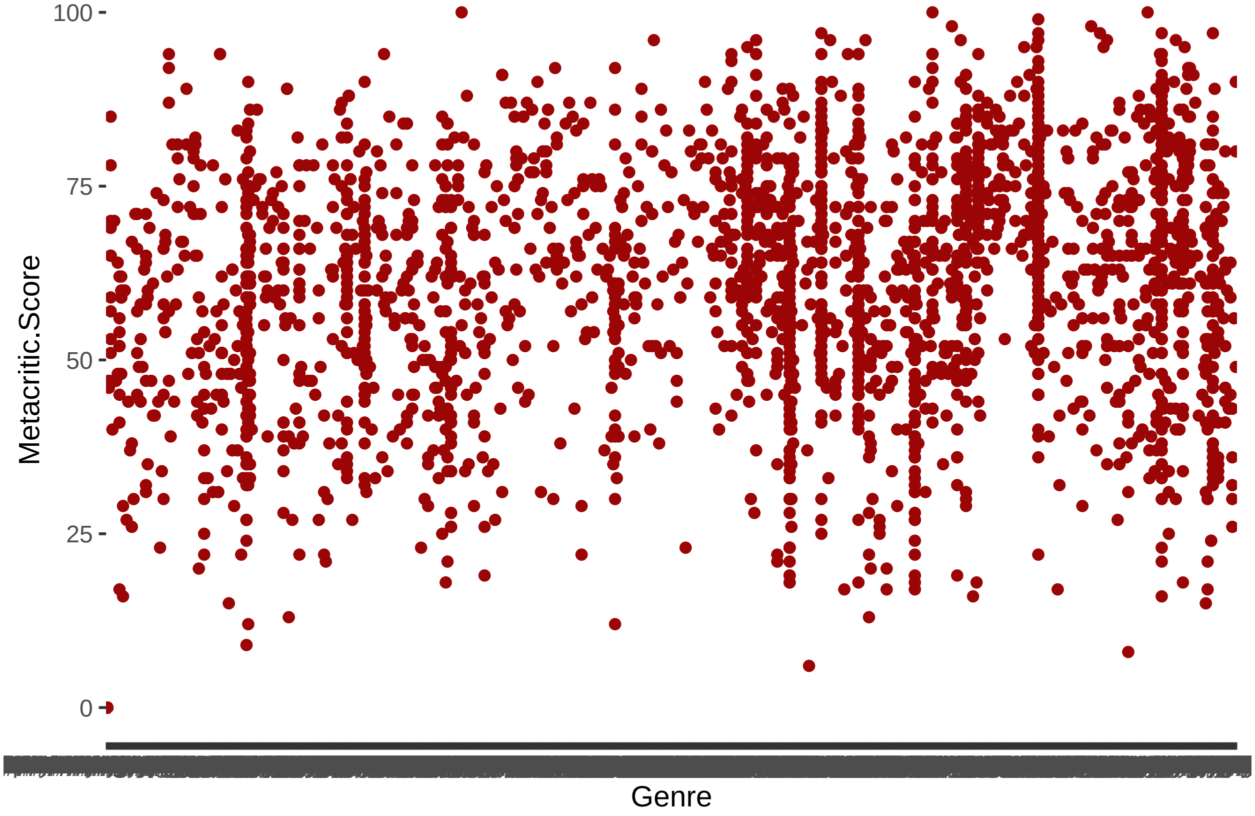
```
ggplot(netflix, mapping = aes(x = Metacritic.Score, y = Rotten.Tomatoes.Score)) + geom_point(color = "#
```



```r
ggplot(netflix, mapping = aes(x = Genre, y = Rotten.Tomatoes.Score)) + geom_point(color = "#9B0505")
```



```r
ggplot(netflix, mapping = aes(x = Genre, y = Metacritic.Score)) + geom_point(color = "#9B0505")
```



## Exploratory Plots.

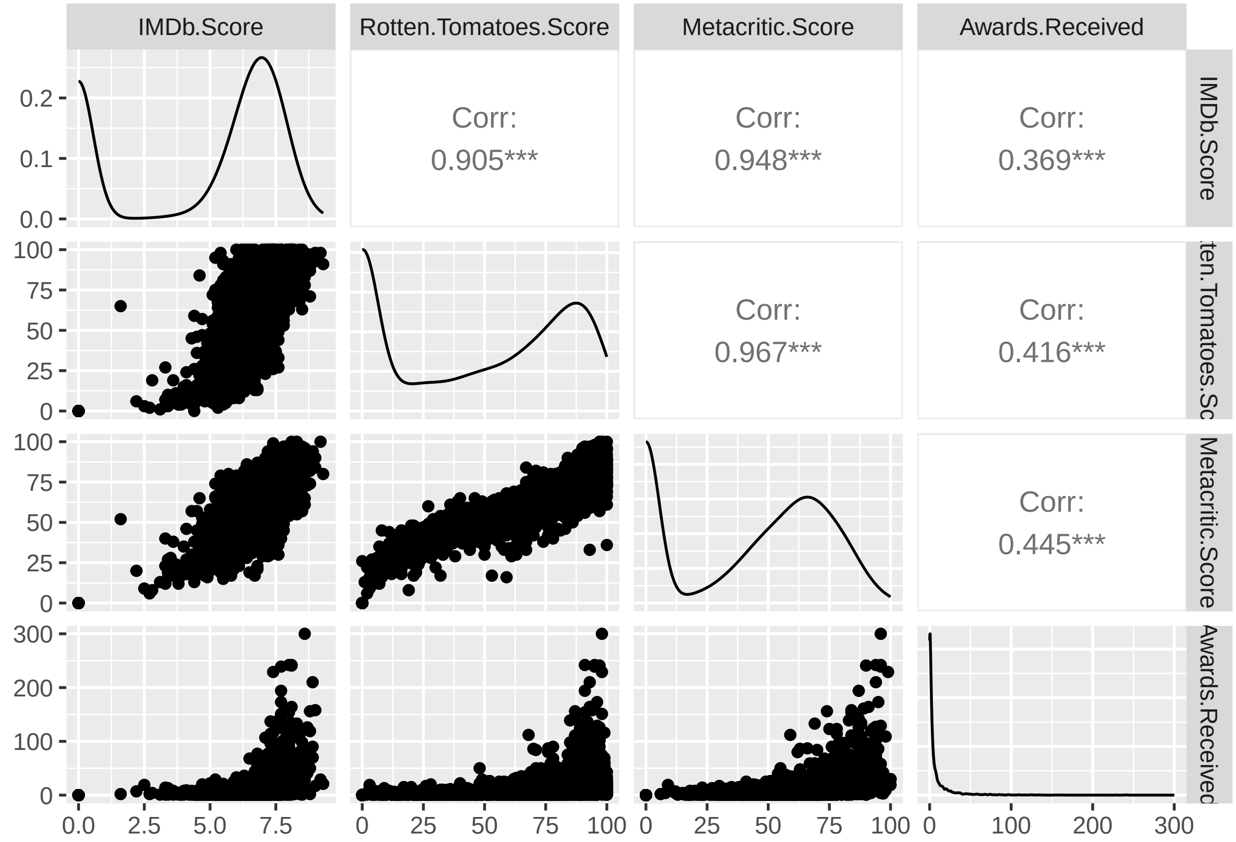
The first plot looks at Meta Critic Scores compared with Rotten Tomatoes Score. For the most part there seems to be a positive line that would fit the data well meaning that for most points the scores match up. But the interesting points are the ones separate from the main line clustering as those are the points where there is a sizable mismatch between the rotten tomato's scores and Meta Critic scores. We can try a regression equation and examine the coefficient and residuals to see how well the scores fit.

The second and third lots are looking at respective website scores and genres. Looking at the plots it is difficult to make specific statements about the data because there are so many data points on the graphs, but we can see that generally the Rotten Tomato's scores seem to be much more spread between their 0 to 100 range where as the MetaCritic scores seem to bunch up much closer around the 25 - 75 range per genre.

```
cor(netflix$Rotten.Tomatoes.Score, netflix$Awards.Nominated.For)
```

```
## [1] 0.455177
```

```
ggpairs(netflix[,c(13, 14, 15, 16)])
```



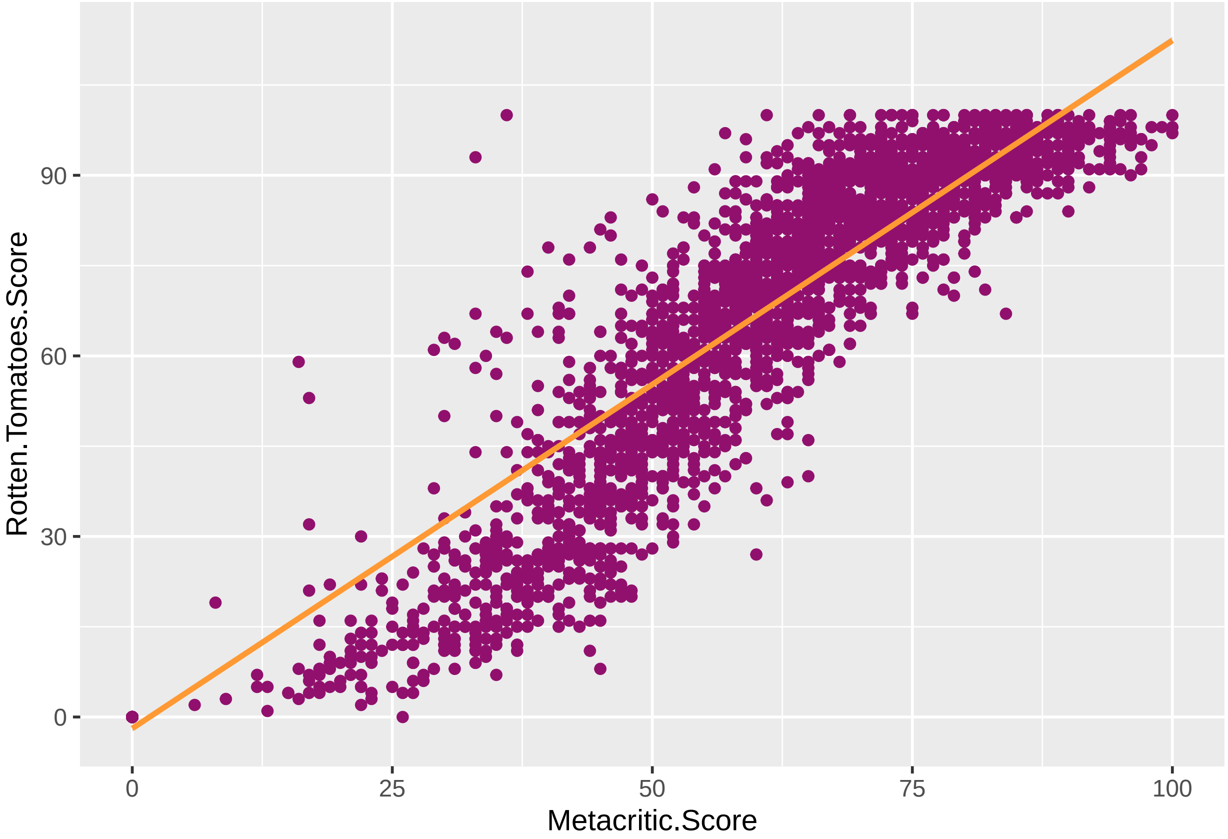
```
Tomatoes_Critic <- lm(Rotten.Tomatoes.Score ~ Metacritic.Score, netflix)
summary(Tomatoes_Critic)
```

```
##
## Call:
## lm(formula = Rotten.Tomatoes.Score ~ Metacritic.Score, data = netflix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.538  -4.113   1.902   3.459  60.750
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.901880   0.254237  -7.481 9.11e-14 ***
## Metacritic.Score  1.143099   0.004833 236.538  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.561 on 3832 degrees of freedom
## Multiple R-squared:  0.9359, Adjusted R-squared:  0.9359
## F-statistic: 5.595e+04 on 1 and 3832 DF,  p-value: < 2.2e-16
```

Regression Analysis

Fitting a line to that first plot we can get some more in depth insight. The formula for our line is $Y = 1.41221X - 19.98424$, the multiple r-squared value at 0.9359 is very close to 1 showing that our line is a pretty good fit for the data.

```
ggplot(netflix, mapping = aes(x = Metacritic.Score, y = Rotten.Tomatoes.Score)) + geom_point(color = "#9
```



This visualization of our line on our chart seems to match our conclusion based on the multiple r-squared value. This shows that for the most part the scores of the rotten tomatoes and metacritic websites seem to match pretty well for the most part. An interesting future project could be to look at the obvious outliers that sit far away from the line and potentially identify what is creating such a mismatch.

```
Tomatoes_IMDb <- lm(Rotten.Tomatoes.Score ~ IMDb.Score, netflix)

IMDb_Critic <- lm(IMDb.Score ~ Metacritic.Score, netflix)

summary(Tomatoes_IMDb)
```

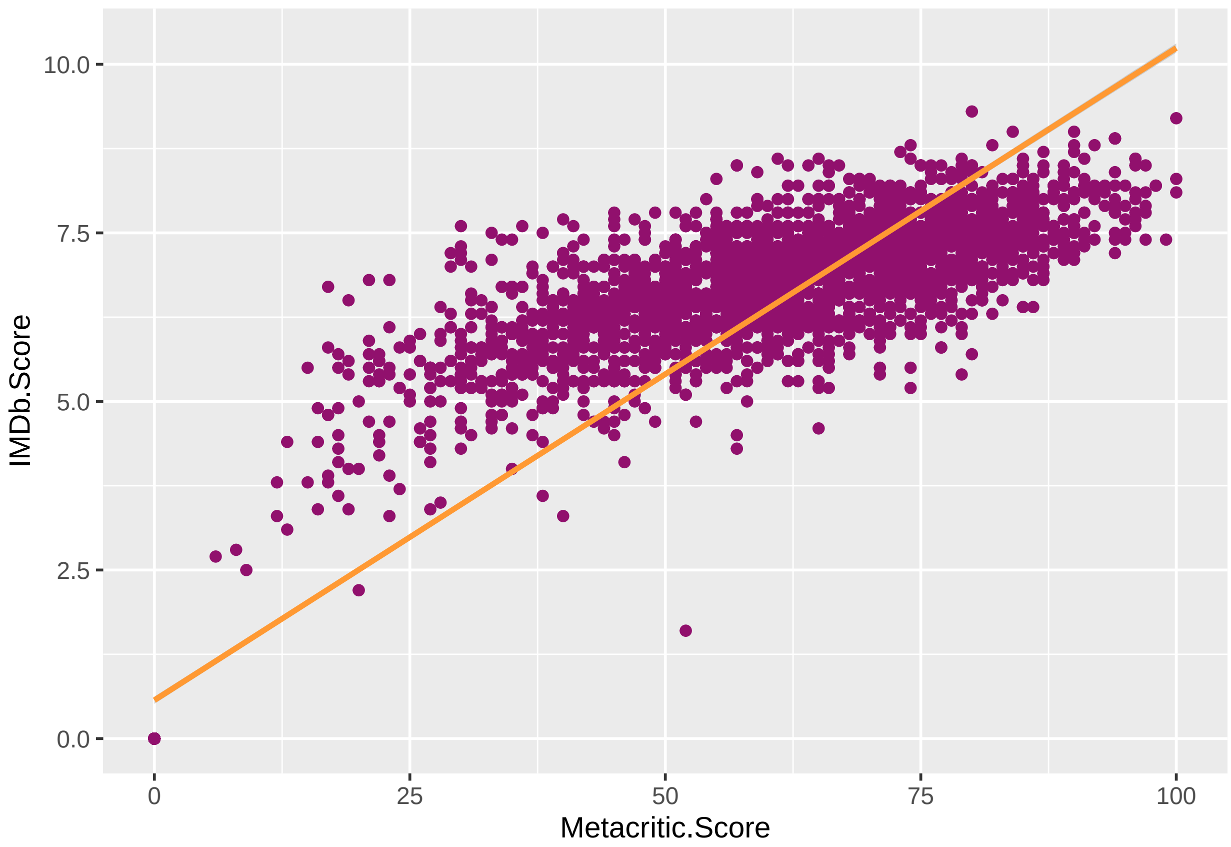
```
##
## Call:
## lm(formula = Rotten.Tomatoes.Score ~ IMDb.Score, data = netflix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.810  -4.473   2.467   9.708  50.696
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.46697    0.44864  -5.499 4.07e-08 ***
## IMDb.Score  10.48194    0.07943 131.962  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 16.04 on 3832 degrees of freedom
## Multiple R-squared:  0.8196, Adjusted R-squared:  0.8196
## F-statistic: 1.741e+04 on 1 and 3832 DF,  p-value: < 2.2e-16
```

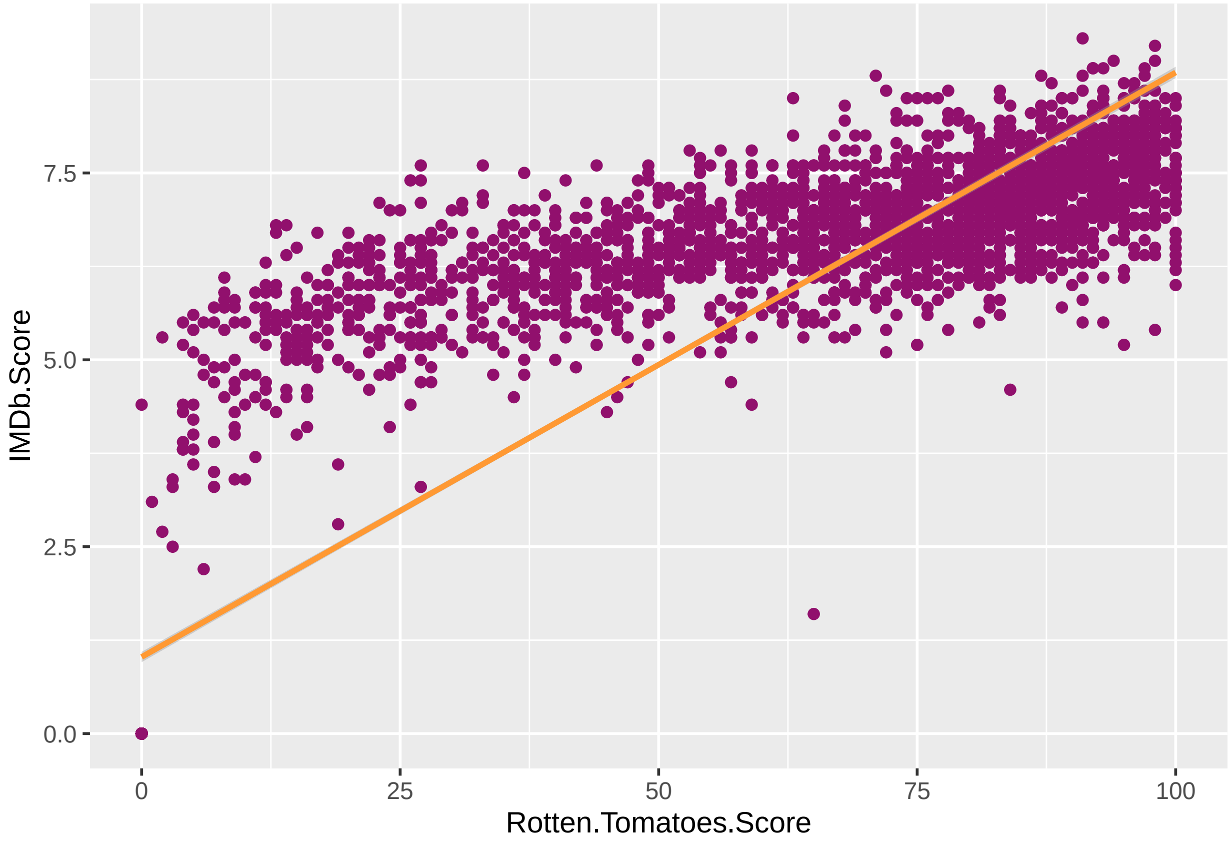
```r
summary(IMDb_Critic)
```

```
##
## Call:
## lm(formula = IMDb.Score ~ Metacritic.Score, data = netflix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9989 -0.5690 -0.5237  0.6305  4.4866
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.5689865  0.0276527   20.58   <2e-16 ***
## Metacritic.Score 0.0967299  0.0005256  184.03   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.04 on 3832 degrees of freedom
## Multiple R-squared:  0.8983, Adjusted R-squared:  0.8983
## F-statistic: 3.387e+04 on 1 and 3832 DF,  p-value: < 2.2e-16
```

```r
ggplot(netflix, mapping = aes(x = Metacritic.Score, y = IMDb.Score)) + geom_point(color = "#91106D") +
```



```r
ggplot(netflix, mapping = aes(x = Rotten.Tomatoes.Score, y = IMDb.Score)) + geom_point(color = "#91106D
```



Examining the relationships between the scores of other aggregate review websites gets a bit more interesting. Based on the charts above, we can see that the relationship between MetaCritic and IMDb and IMDb and Rotten Tomatoes is not quite as strong with multiple r squared values of 0.8983 and 0.8196 respectively. This means that there is a slight mismatch mismatch between the scores of the previous two sites and IMDb's scores in some cases.

```
scores_awardsEarned <- lm(Awards.Received ~ Metacritic.Score+Rotten.Tomatoes.Score+IMDb.Score, netflix)
scores_awardsNominated <- lm(Awards.Nominated.For ~ Metacritic.Score+Rotten.Tomatoes.Score+IMDb.Score, n
```

Awards and Scores

We have established that Rotten Tomatoes, Metacritic, and to a slightly lesser extent IMDb seem to agree on the quality of shows based on the scores shown on their websites. These websites can serve as a window into what the average viewer thinks of a show. To get the opinions of the professional viewers we will examine the number of awards nominated for and rewards received for the shows.

```
par(mfrow=c(1,1))
summary(scores_awardsEarned)
```

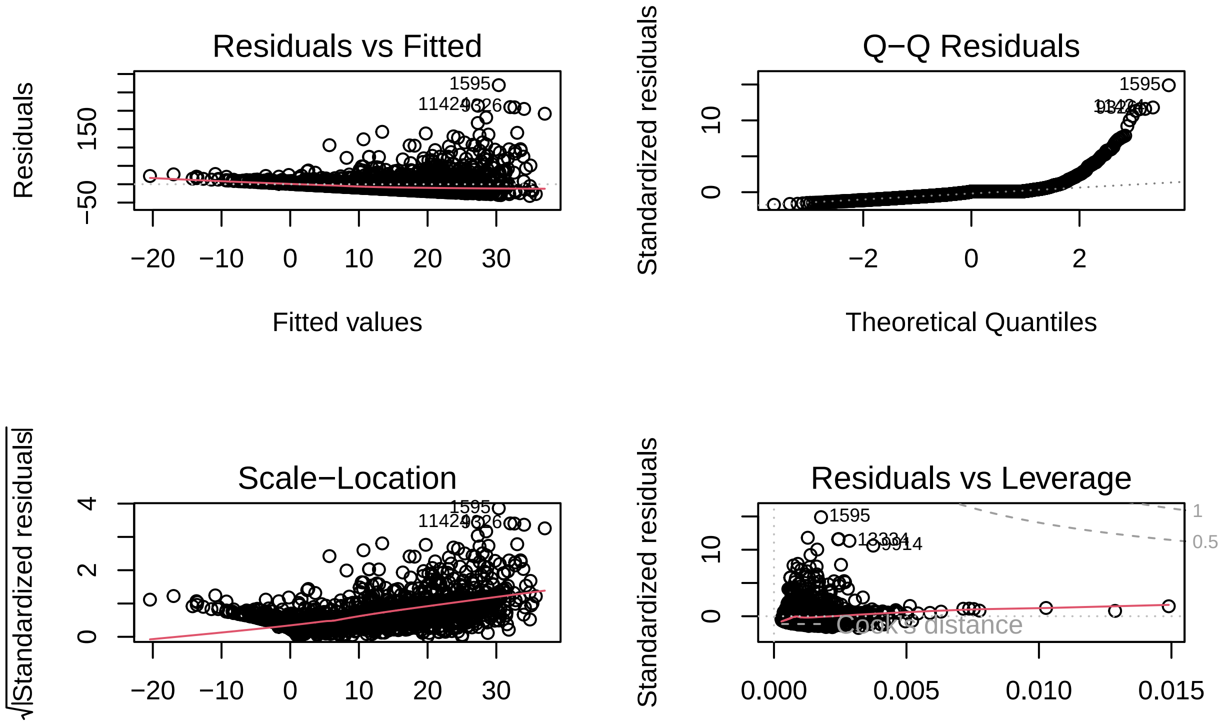
```
##
## Call:
## lm(formula = Awards.Received ~ Metacritic.Score + Rotten.Tomatoes.Score +
##     IMDb.Score, data = netflix)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -31.870  -8.442   1.524   1.693 269.639
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.69279    0.50907  -3.325 0.000892 ***
## Metacritic.Score         0.82662    0.04870  16.974  < 2e-16 ***
## Rotten.Tomatoes.Score -0.17445    0.03094  -5.638 1.84e-08 ***
## IMDb.Score              -3.51234    0.28447 -12.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.12 on 3830 degrees of freedom
## Multiple R-squared:  0.2313, Adjusted R-squared:  0.2307
## F-statistic: 384.2 on 3 and 3830 DF,  p-value: < 2.2e-16
```

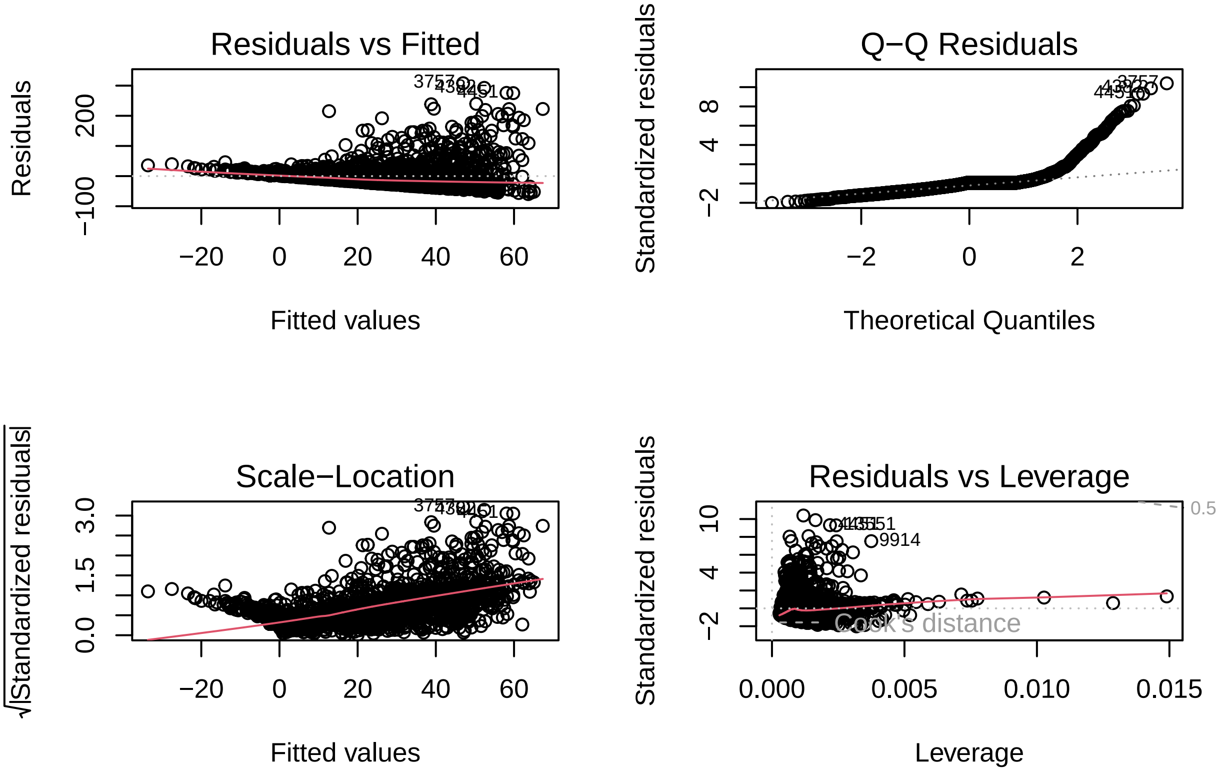
```r
summary(scores_awardsNominated)
```

```
##
## Call:
## lm(formula = Awards.Nominated.For ~ Metacritic.Score + Rotten.Tomatoes.Score +
##     IMDb.Score, data = netflix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.645 -14.472   2.433   2.433 308.025
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -2.43291    0.83280  -2.921  0.00351 **
## Metacritic.Score         1.45209    0.07967  18.227  < 2e-16 ***
## Rotten.Tomatoes.Score -0.31012    0.05062  -6.127 9.87e-10 ***
## IMDb.Score              -5.88889    0.46537 -12.654  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.65 on 3830 degrees of freedom
## Multiple R-squared:  0.2705, Adjusted R-squared:   0.27
## F-statistic: 473.5 on 3 and 3830 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(scores_awardsEarned)
```



```r
plot(scores_awardsNominated)
```



```r
par(mfrow=c(1,1))
```

Based on these R scores: 0.2705 for website ratings and awards received as well as 0.2313 for website ratings and awards nominated for. These mean that website ratings are not accurate predictors for how many awards a show will earn/be nominated for. The four graphs and especially the residuals vs fitted graphs illustrate this pretty well. We can see that at the lower end of the fitted values there aren't that many residuals while at the higher levels the residuals grow. In other words, it seems that there is a disconnect in opinions between internet reviewers and Award Board Members.

##Conclusion

#What does the analysis say? To sum up, IMDb, Rotten Tomatoes, and Meta Critic scores tend to match up pretty well. Meta Critic and Rotten Tomatoes especially, seem to have corollary scores meaning a score on one website will have a likelyhood of sharing a similar score on the other. Factoring in IMDb's scores and there is not as strong of a relation ship, but there still seems to be one present. While there is not enough information to concretely explain the difference it could possibly be attributed to a difference in scales for scoring used on each website. Experimentation and more analysis would be needed to say for certain.

While the websites scores seem to agree in most cases, looking at the website scores and their relationship with the awards a show or movie has been nominated for or received there is a larger mismatch. The aggregate scores from the websites are not very good predictors of whether or not a show or movie will be nominated for or earn any awards. This seems to suggest that the opinions on the shows and movies of the people giving the awards do not match those of the users of the website. It could be that the ammount of awards given is so low that they just cant possibly match the ratings of the websites. Not every highly rated show gets an award simply because there aren't enough unique awards to give. This is also not something that can be concretely concluded with the analytics in this project. However, it certainly would be an interesting project in the future to examine these trends and even the media with awards to see what makes an award winning show.

#Applications and Future Projects User review data and award information are important potential indicators of a show or movies quality. It would be interesting to collect survey data and see how that matches the websites and award data as well. A long term project would be training a model on collected data to recommend shows based on a general public opinion, critic opinion, and other factors. Personalizing this model so that it recommends things tailored to my tastes would be especially interesting. It would also be interesting to examine award data specifically and compare things like genre, production budget, cast, and other factors to determine whether there is any bias in which shows and movies receive awards. This analysis and the future analysis would also have similar applications and maybe similar results in the music field and could be interesting to explore as well.