Final Project

#Executive Summary

Each row represents a movie available on FilmTV.it, with the original title, year, genre, duration, country, director, actors, average vote and votes. The fie in the English version contains 37,711 movies and 19 attributes.

(we should add what each of the different variables mean)

EDA

Data Cleaning

The first step in our data cleaning is to deal with missing values. First, there werre 88 movies that had no genre. So we used google to fill in the missing genres to ensure we could use those data points for our analysis. We also were able to find one movie that had no title, however, from the description we were able to determine the movie name as well as the genre.

Here we are making film into a factor as it contains categorical variables that would be better analyzed as factors.

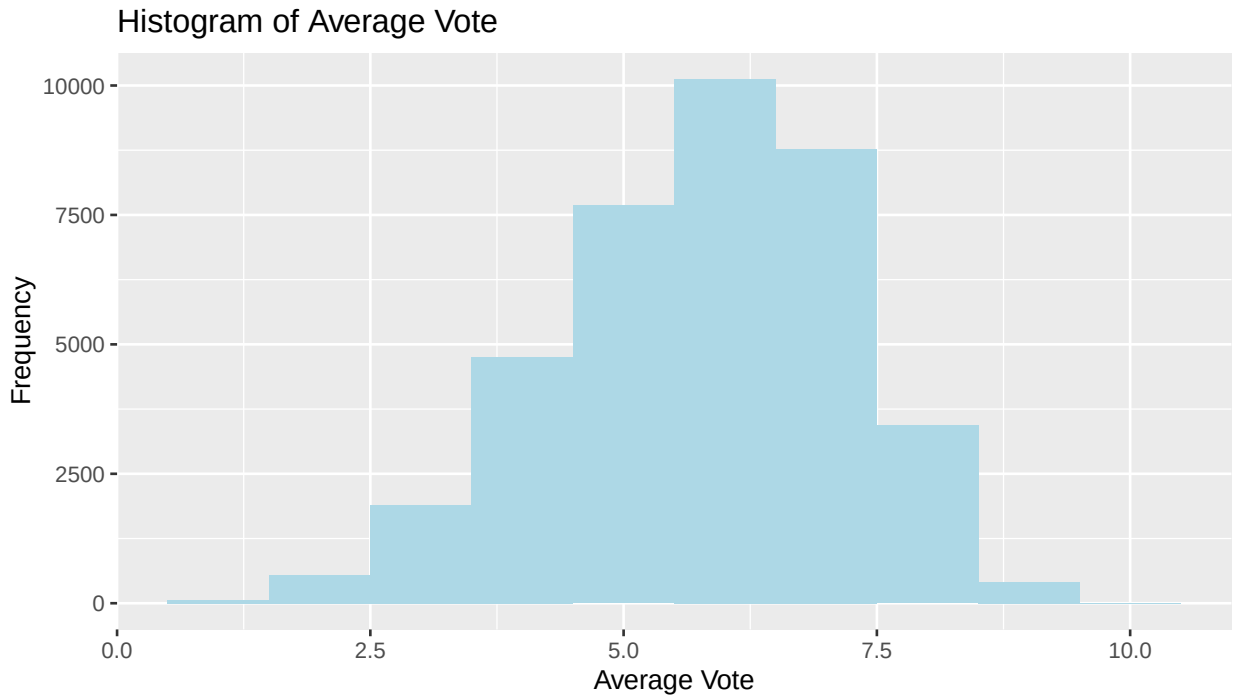
```
##  [1] "Action"      "Adventure"   "Animation"   "Biblical"    "Biography"
##  [6] "Comedy"      "Crime"       "Documentary" "Drama"       "Fantasy"
## [11] "Gangster"    "Grotesque"   "History"     "Horror"      "Mélo"
## [16] "Musical"     "Mythology"   "Noir"        "Romance"     "Romantic"
## [21] "Short Movie" "Sperimental" "Sport"       "Spy"         "Super-hero"
## [26] "Thriller"    "War"         "Western"
```

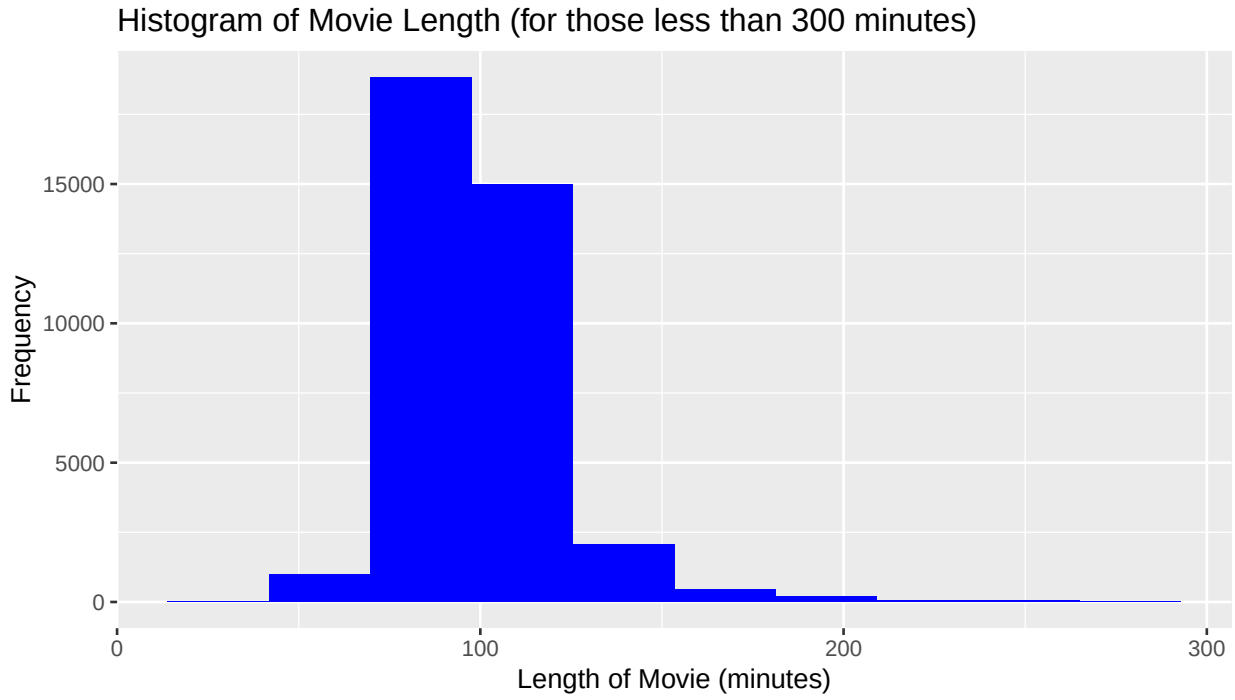
We will also make the year variable into different subsections including different years that will be used as a factor. We are going to make the years in sections of 10 years from 1897 to 2021, that will give us 14 different levels of year modulos to work with

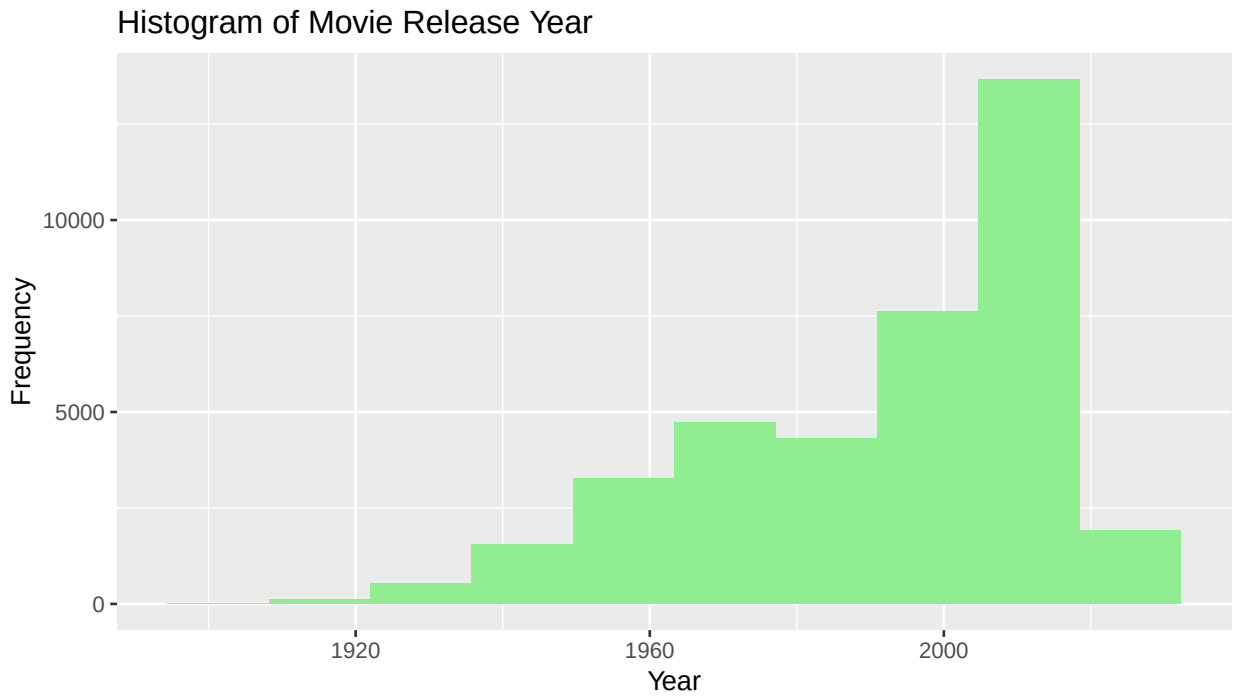
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1897    1975    2000    1992    2012    2021
```

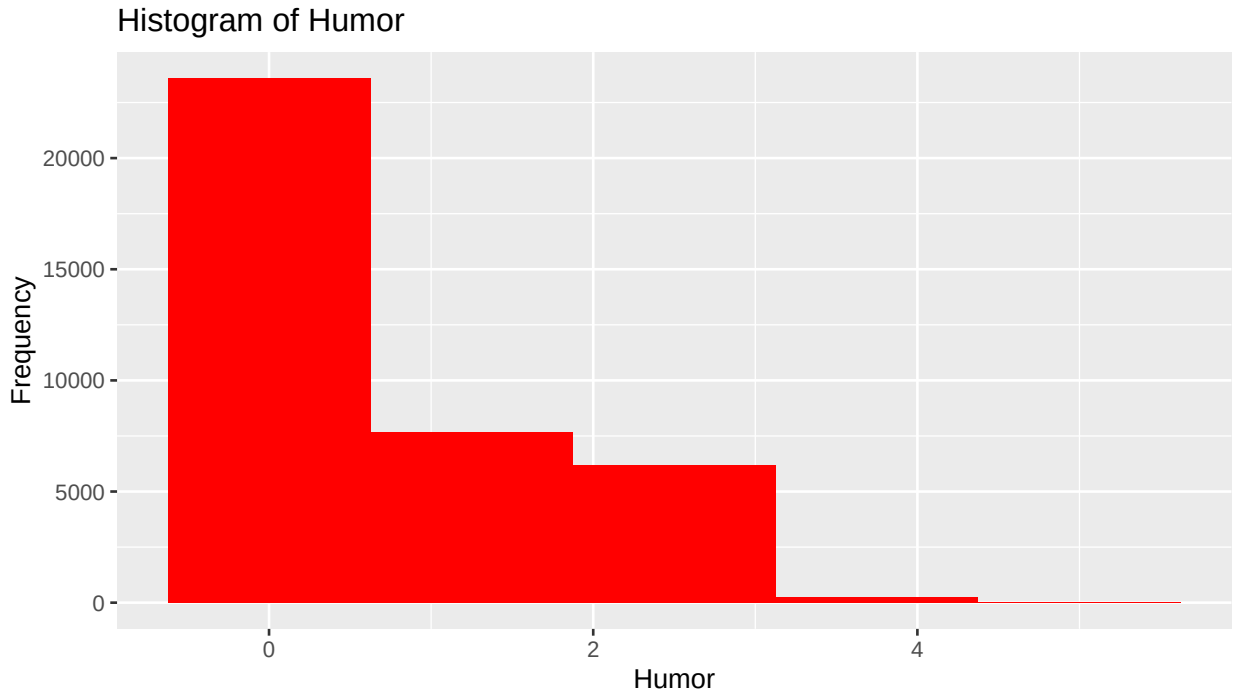
```
## [1890,1900) [1900,1910) [1910,1920) [1920,1930) [1930,1940) [1940,1950)
##           1           1          65         256         689        1190
## [1950,1960) [1960,1970) [1970,1980) [1980,1990) [1990,2000) [2000,2010)
##        2265        2980        3323        3363        4722        7358
## [2010,2020) [2020,2030)
##       10614         884
```

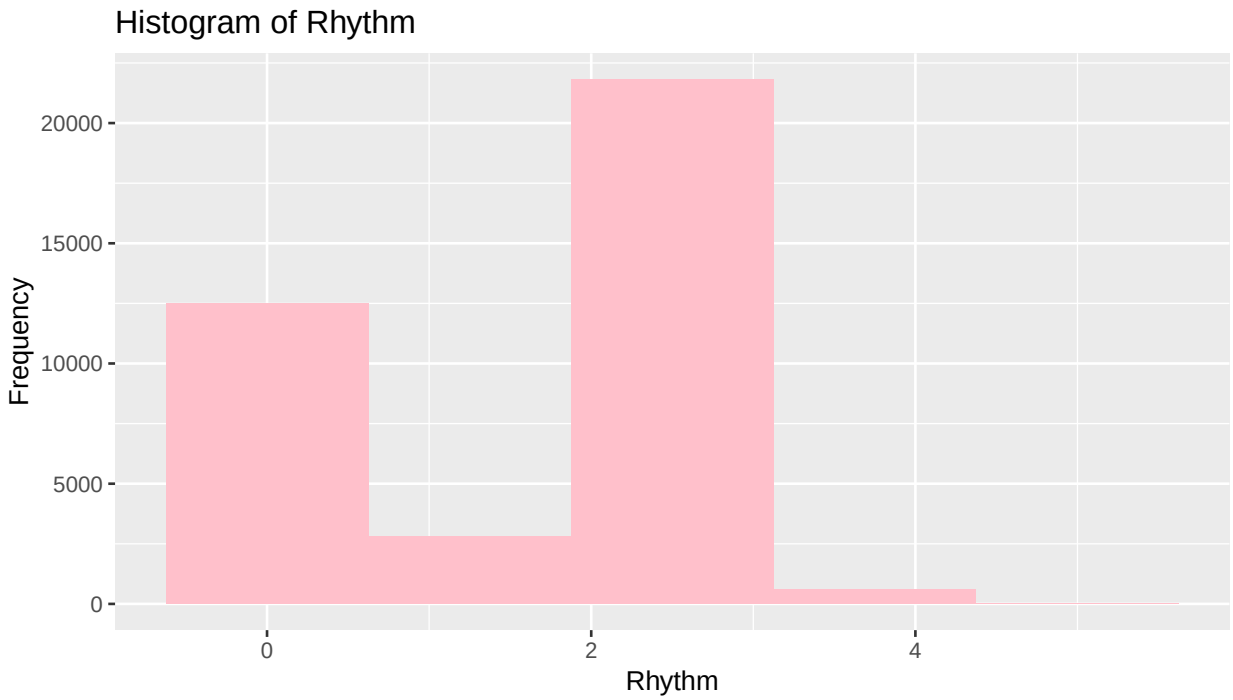
Plots

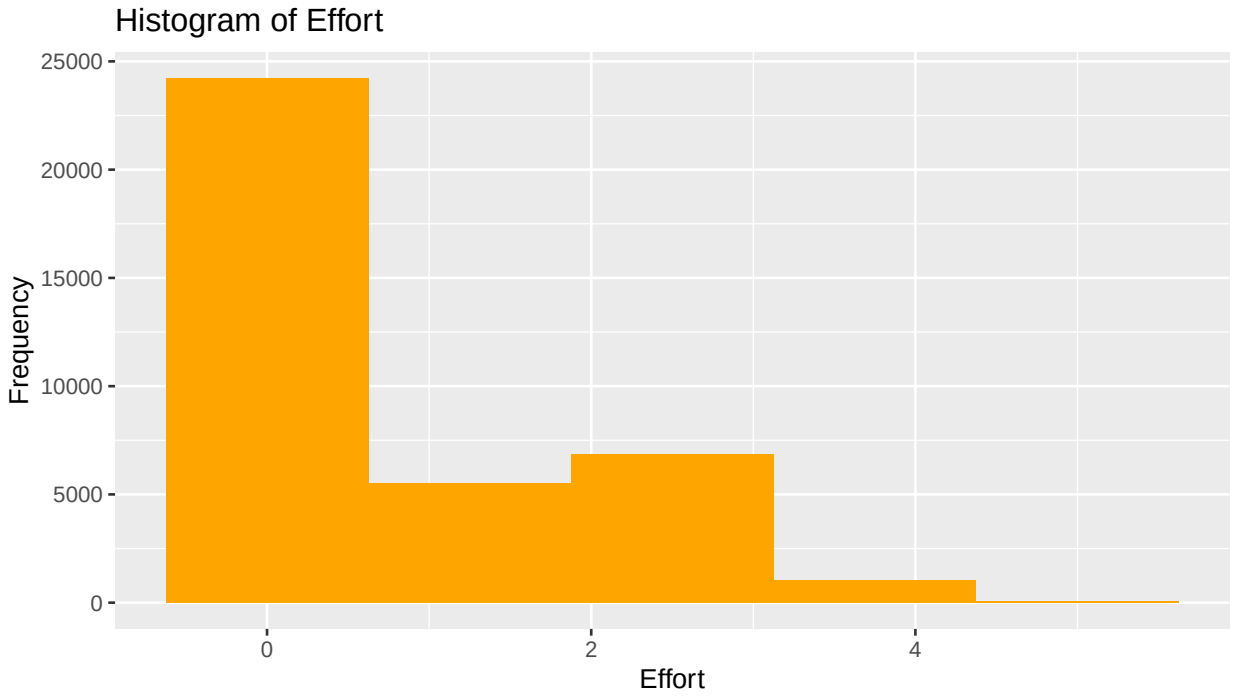

Histogram of Average Vote
Frequency
10000
7500
5000
2500
0
0.0
2.5
5.0
7.5
10.0
Average Vote

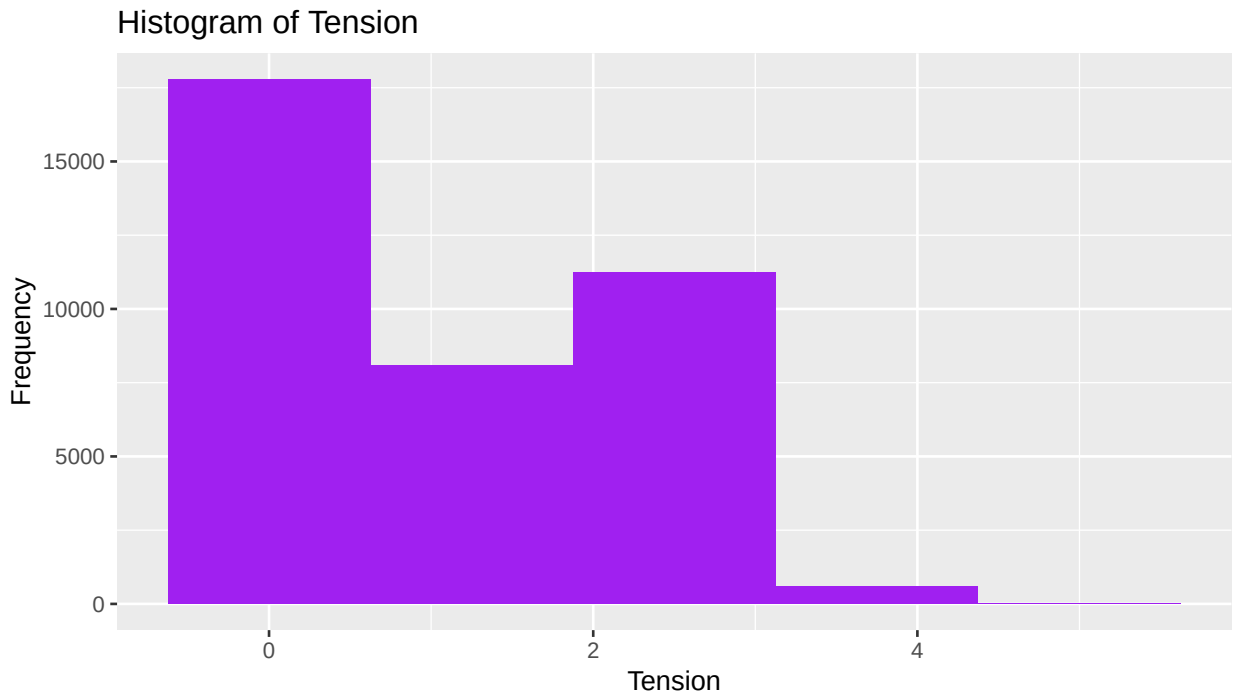

Histogram of Movie Length (for those less than 300 minutes)
Frequency
15000
10000
5000
0
0
100
200
300
Length of Movie (minutes)

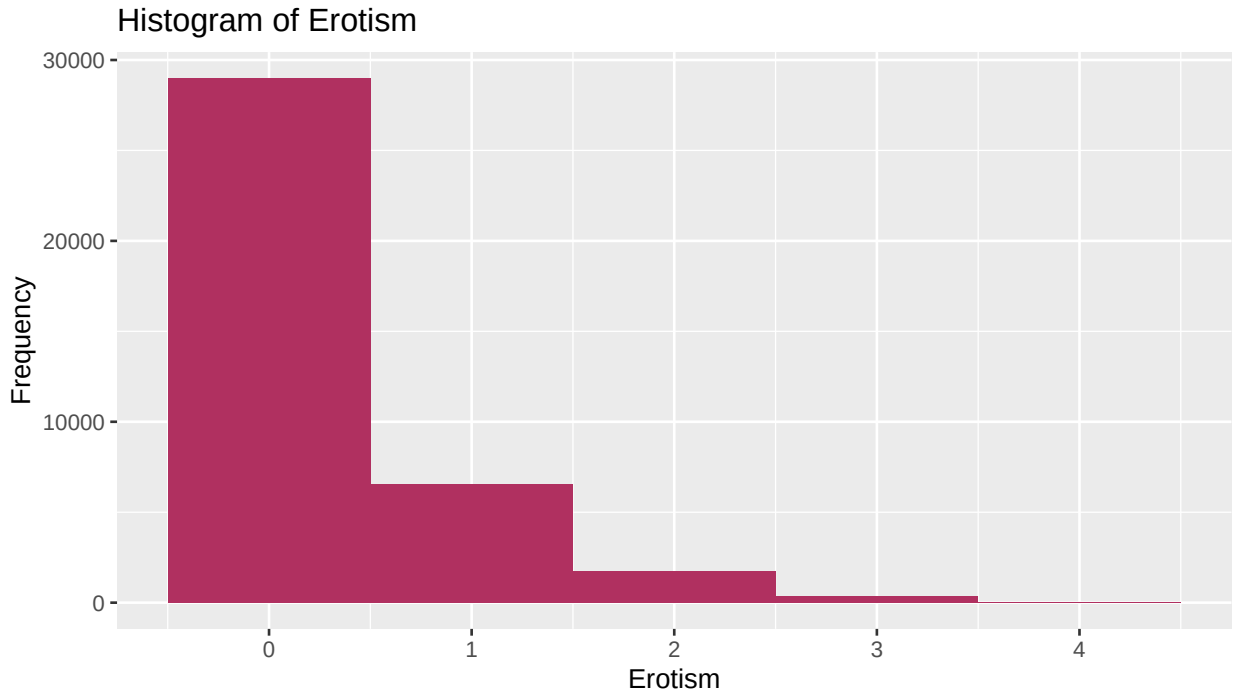

Histogram of Movie Release Year
Frequency
10000
5000
0
1920
1960
2000
Year


Histogram of Humor
Frequency
20000
15000
10000
5000
0
0
2
4
Humor


Histogram of Rhythm
Frequency
20000
15000
10000
5000
0
0
2
4
Rhythm


Histogram of Effort
Frequency
25000
20000
15000
10000
5000
0
0
2
4
Effort



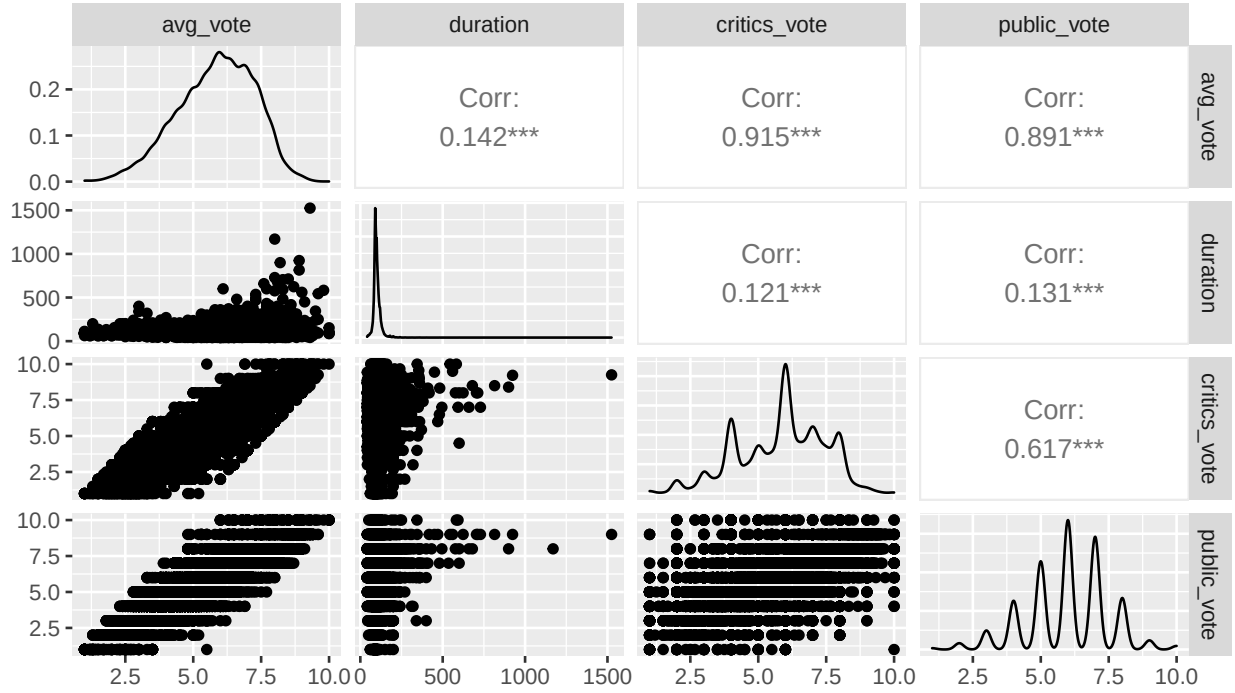


Let's look at the correlation between some of the numerical variables and average vote.

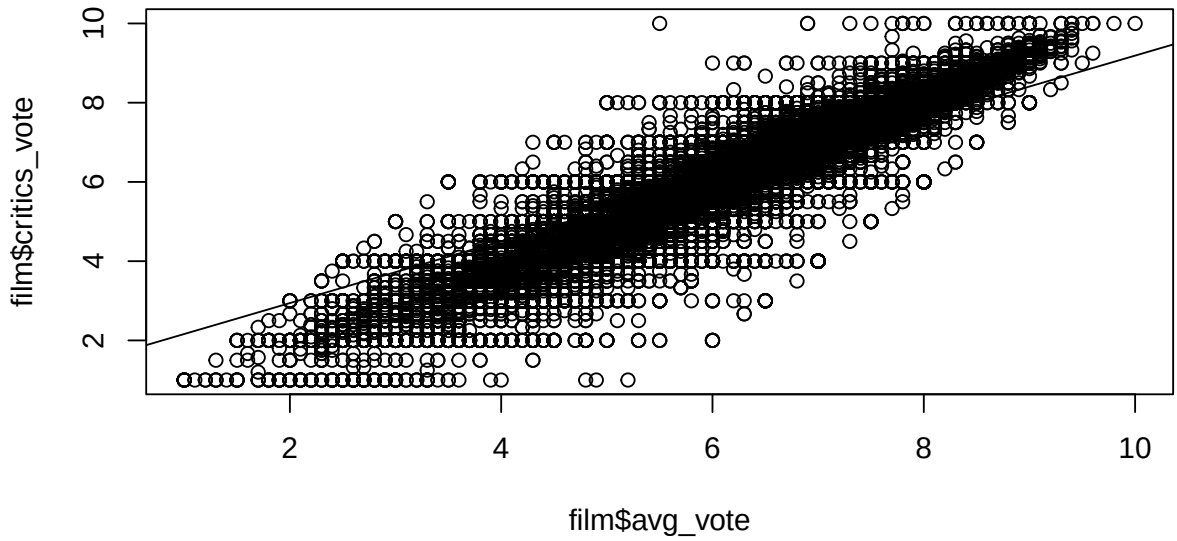
```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 4085 rows containing missing values
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 219 rows containing missing values
```

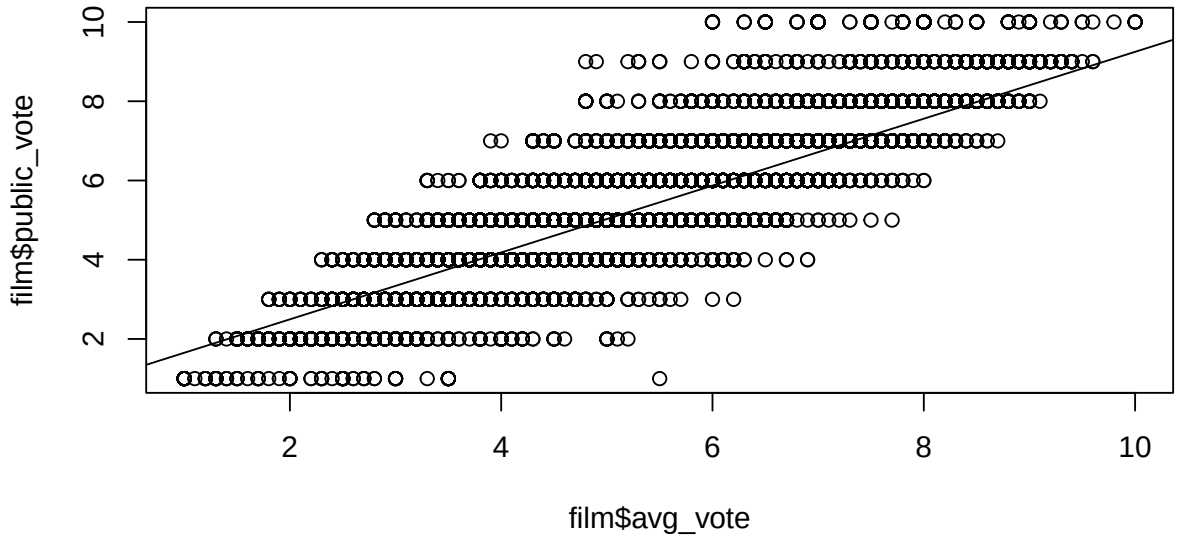
```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 4085 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 219 rows containing missing values

## Warning: Removed 4085 rows containing missing values (geom_point).

## Warning: Removed 4085 rows containing missing values (geom_point).

## Warning: Removed 4085 rows containing non-finite values (stat_density).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 4304 rows containing missing values

## Warning: Removed 219 rows containing missing values (geom_point).

## Warning: Removed 219 rows containing missing values (geom_point).

## Warning: Removed 4304 rows containing missing values (geom_point).

## Warning: Removed 219 rows containing non-finite values (stat_density).
```



In our data we noticed that there are three different voting variables. Our ultimate goal is to predict the average vote, so we want to determine if the public vote and critics vote have a linear relationship or high correlation to the average vote. If that is the case, we will remove those variables from the regression as they have a somewhat linear dependency on one another.



We can see that the scatter plot shows a high correlation between the average vote and the critics vote with a correlation of 0.915. Thus there is almost a linear relationship between the variables so we will remove critics vote for the sake of prediction and regression.



We can see that this scatter plot also shows a high correlation between the average vote and the public vote with a correlation of 0.891. Thus there is almost a linear relationship between the variables so we will remove public vote for the sake of prediction and regression as well.

PCA

Let's create PCA scores for our data set and see if they tell us anything important.

Linear Regression

Let's try to predict the average vote using some of the other variables in the data set. First, we will drop some of the variables in the data set that are not useful for the regression, such as `filmtv_id`,`year`, `title`,`country`, `directors`, `actors`, `description`, and `notes`. We will also remove `critics_vote` and `public_vote` as indicated previously.

##Multiple Regression Let's run a linear model using all of the variables

```
## Anova Table (Type II tests)
##
## Response: avg_vote
##              Sum Sq    Df F value    Pr(>F)
## genre          5366    27  148.69 < 2.2e-16 ***
## duration        556     1  416.15 < 2.2e-16 ***
## total_votes    1741     1 1302.29 < 2.2e-16 ***
## humor           799     1  598.12 < 2.2e-16 ***
## rhythm          913     1  683.07 < 2.2e-16 ***
## effort         1458     1 1090.67 < 2.2e-16 ***
## tension        1071     1  801.62 < 2.2e-16 ***
## erotism         219     1  163.82 < 2.2e-16 ***
## year_mod       4735    13  272.55 < 2.2e-16 ***
## Residuals     50337 37663
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the ANOVA for this test, we can ese that all of the variables used the the regression are significant.

LASSO

Logistic Regression