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

Here we are making film into a factor as it contains categorical variables that would be better analyzed as factors. We will also remove the films that contain no genre.

```
##  [1] 14754 15203 15249 15301 15309 15525 15643 15871 15896 16554 16780 17126
## [13] 17214 17522 17571 17597 17598 17639 17836 18043 18291 18387 20006 20156
## [25] 20503 20776 20820 20878 20976 21061 21256 21439 21598 21716 22116 22118
## [37] 22258 22380 22381 22821 22837 22885 22978 23056 23181 23277 23477 23553
## [49] 23972 24004 24048 24272 24292 24567 25630 25706 25720 25936 25937 25938
## [61] 26106 26334 26421 26457 26515 26553 26631 26704 26709 26724 26776 26936
## [73] 27088 27572 27577 27728 28549 29300 29310 30000 30171 30191 30437 33546
## [85] 35885 35987 36126 36732
```

```
##  [1] filmtv_id    title        year         genre        duration
##  [6] country      directors    actors       avg_vote     critics_vote
## [11] public_vote  total_votes  description  notes        humor
## [16] rhythm       effort       tension      erotism
## <0 rows> (or 0-length row.names)
```

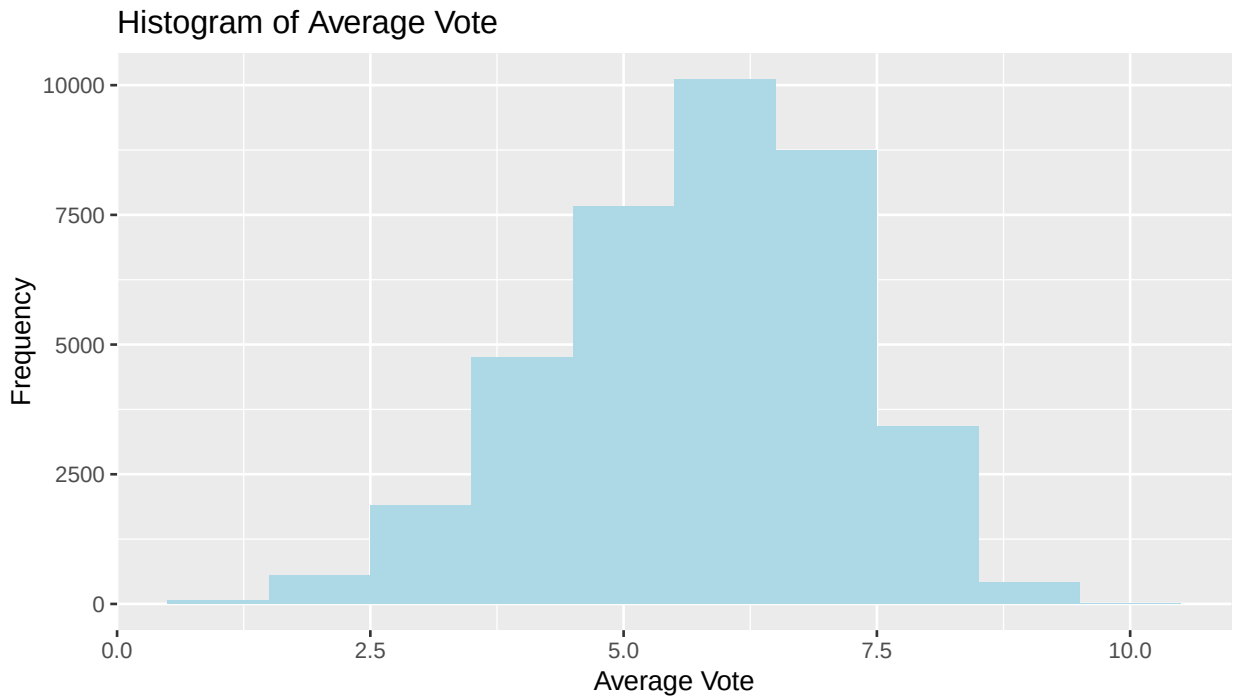
```
##  [1] "Action"      "Adventure"   "Animation"   "Biblical"    "Biography"
##  [6] "Comedy"      "Crime"       "Documentary" "Drama"       "Fantasy"
## [11] "Gangster"    "Grotesque"   "History"     "Horror"      "Mélo"
## [16] "Musical"     "Mythology"   "Noir"        "Romantic"    "Short Movie"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

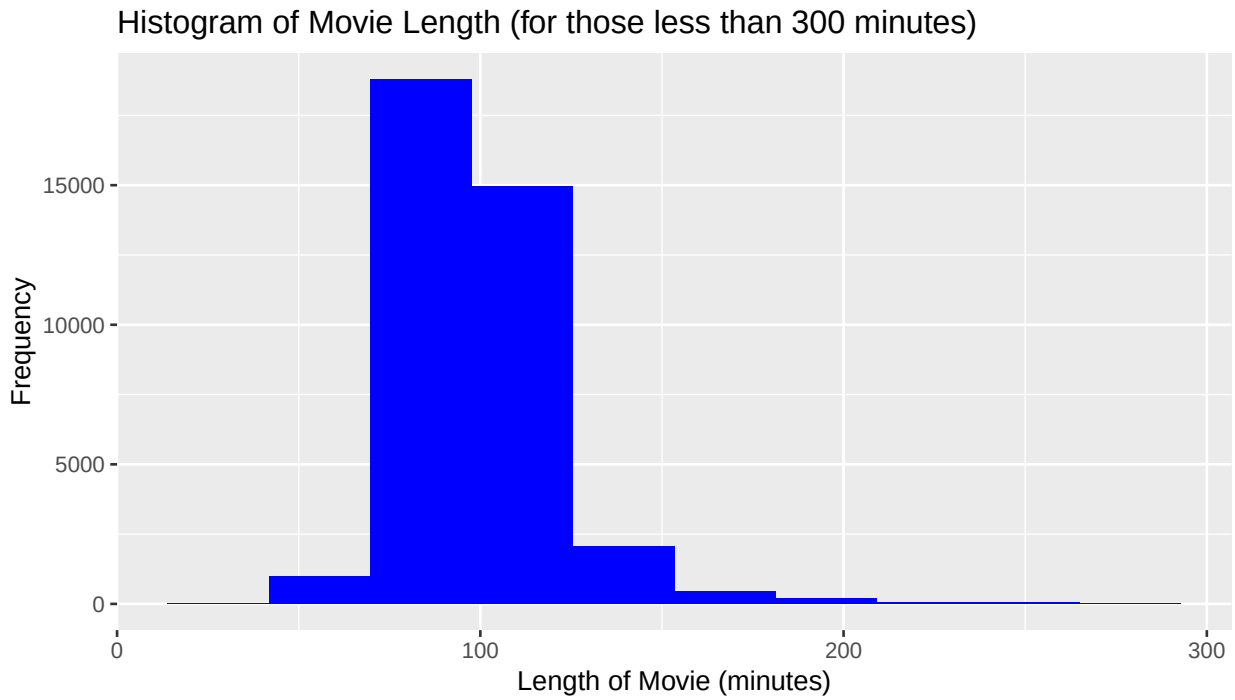
We will also make the year variable into different subsections including different years that will be used as a factor. We are going to make the years in sections of 10 years from 1897 to 2021, that will give us 14 different levels of year modulos to work with

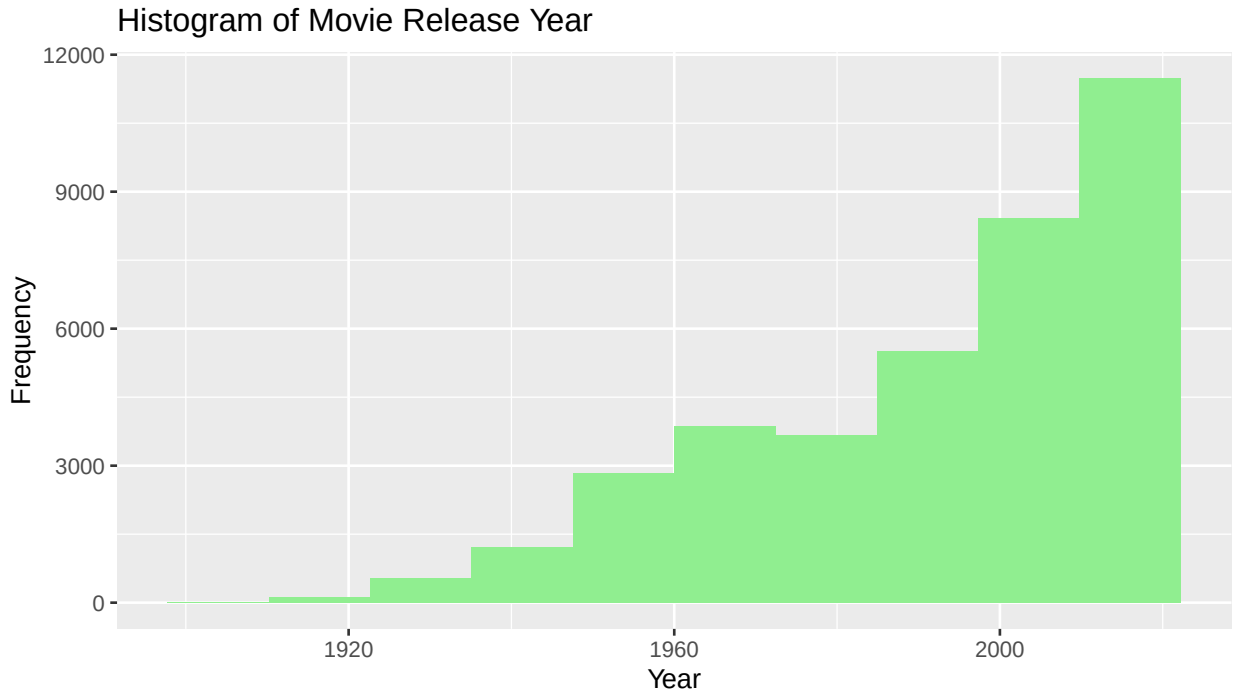
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1909    1975    1999    1992    2012    2021
```

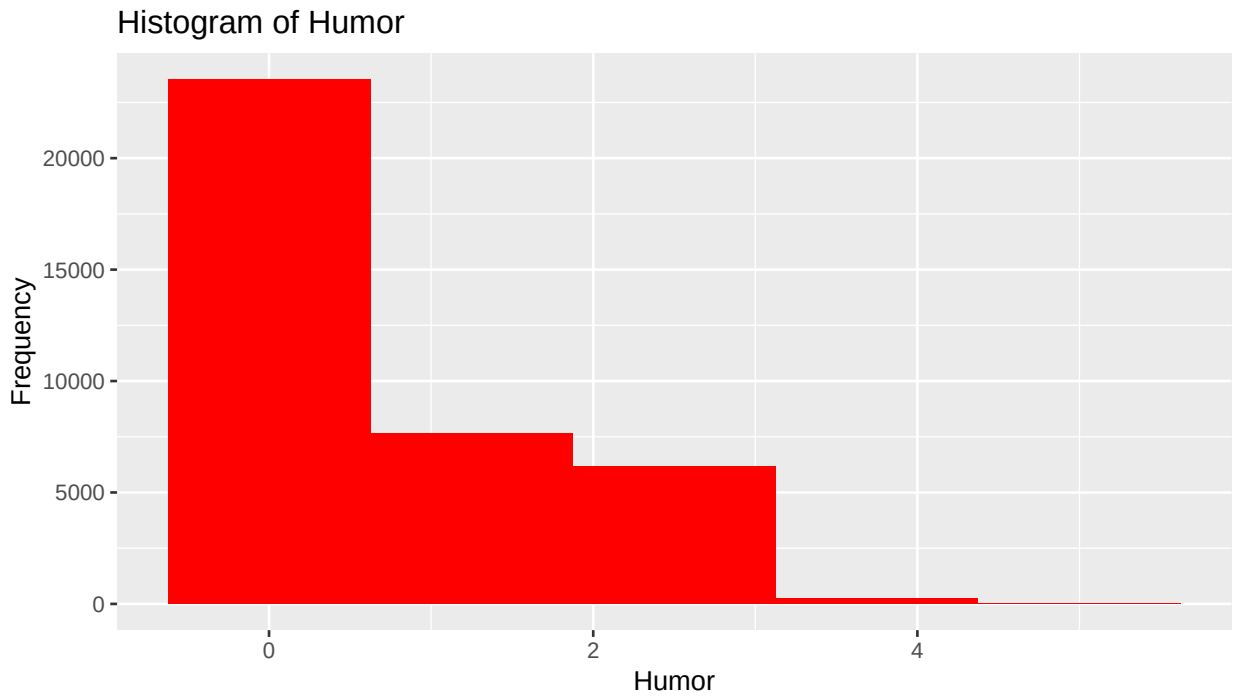
```
## [1900,1910) [1910,1920) [1920,1930) [1930,1940) [1940,1950) [1950,1960)
##           1          65         256         688        1187        2264
## [1960,1970) [1970,1980) [1980,1990) [1990,2000) [2000,2010) [2010,2020)
##        2971        3314        3355        4713        7326       10599
## [2020,2030)
##         884
```

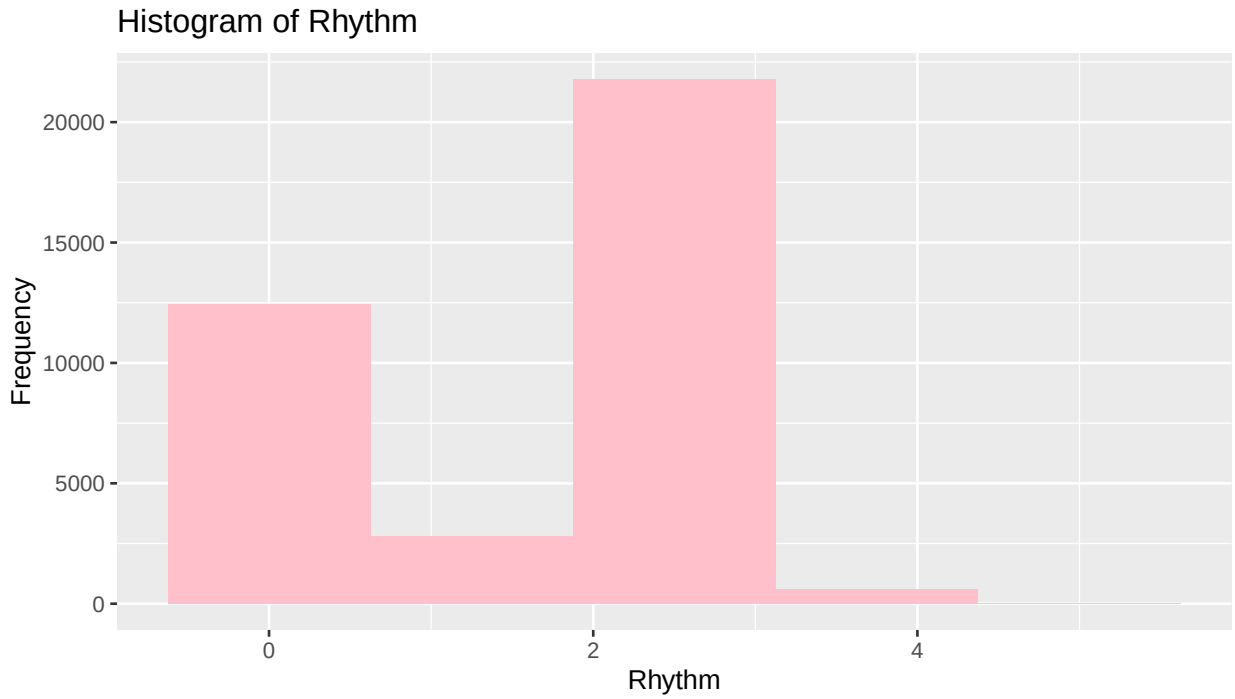
Plots

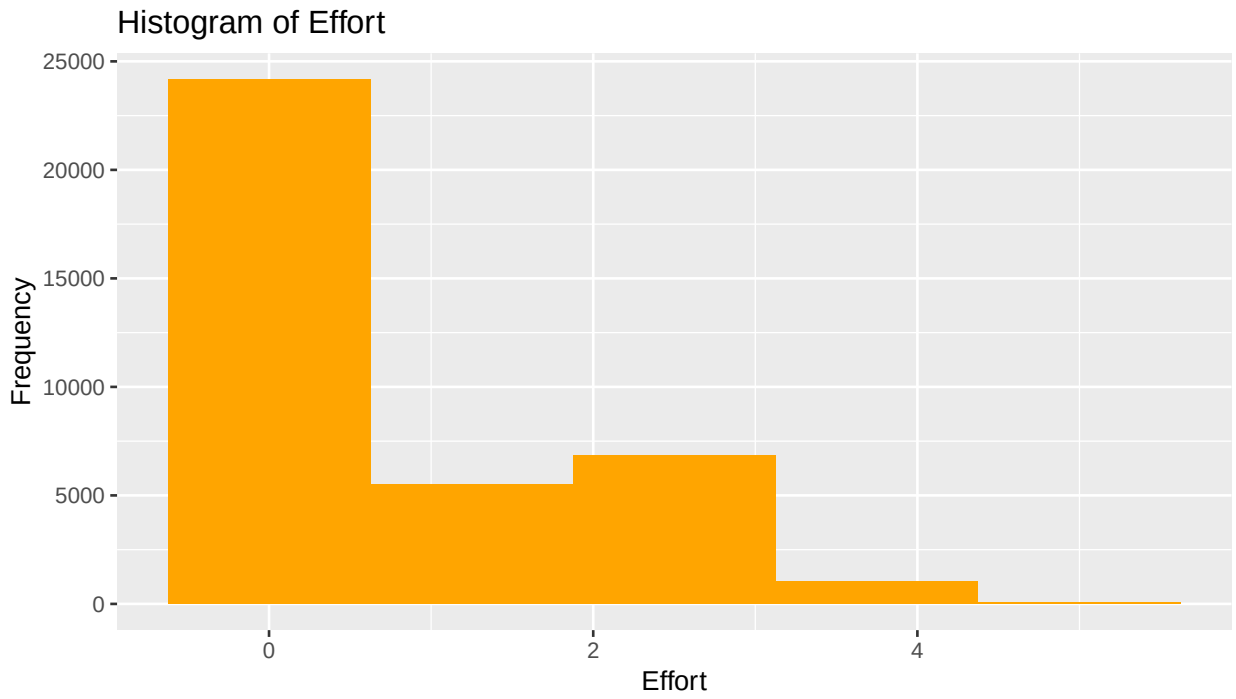


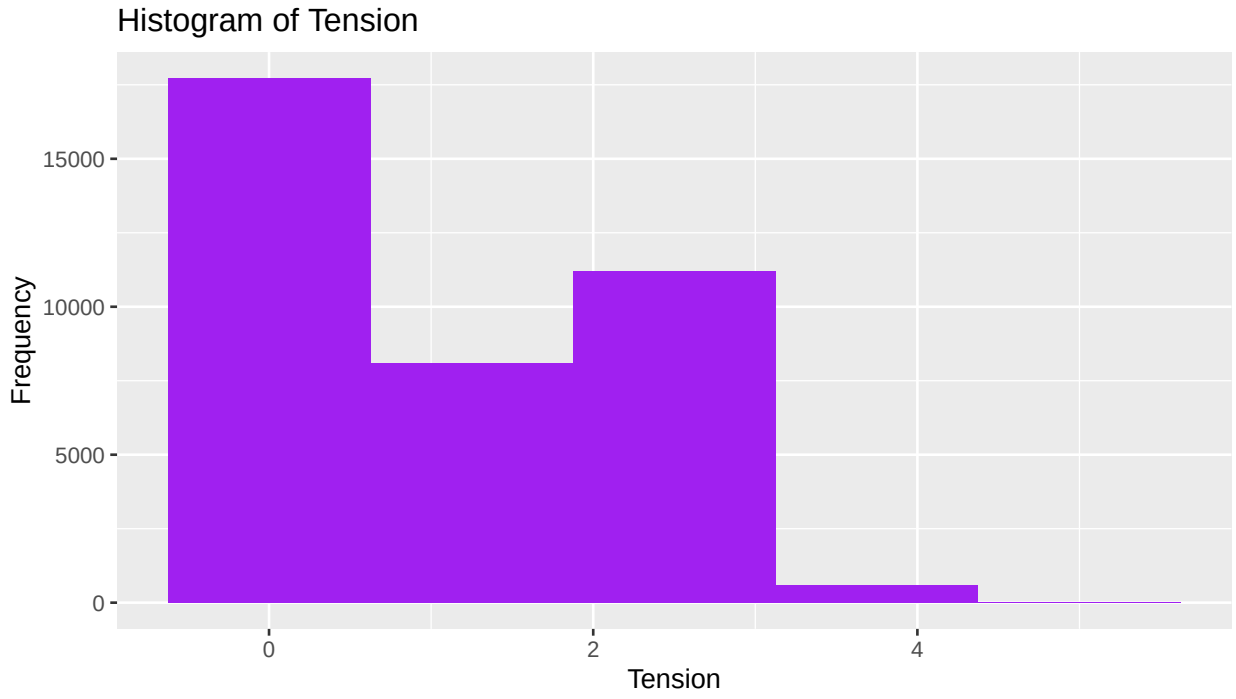

Histogram of Movie Length (for those less than 300 minutes)
Frequency
15000
10000
5000
0
0
100
200
300
Length of Movie (minutes)

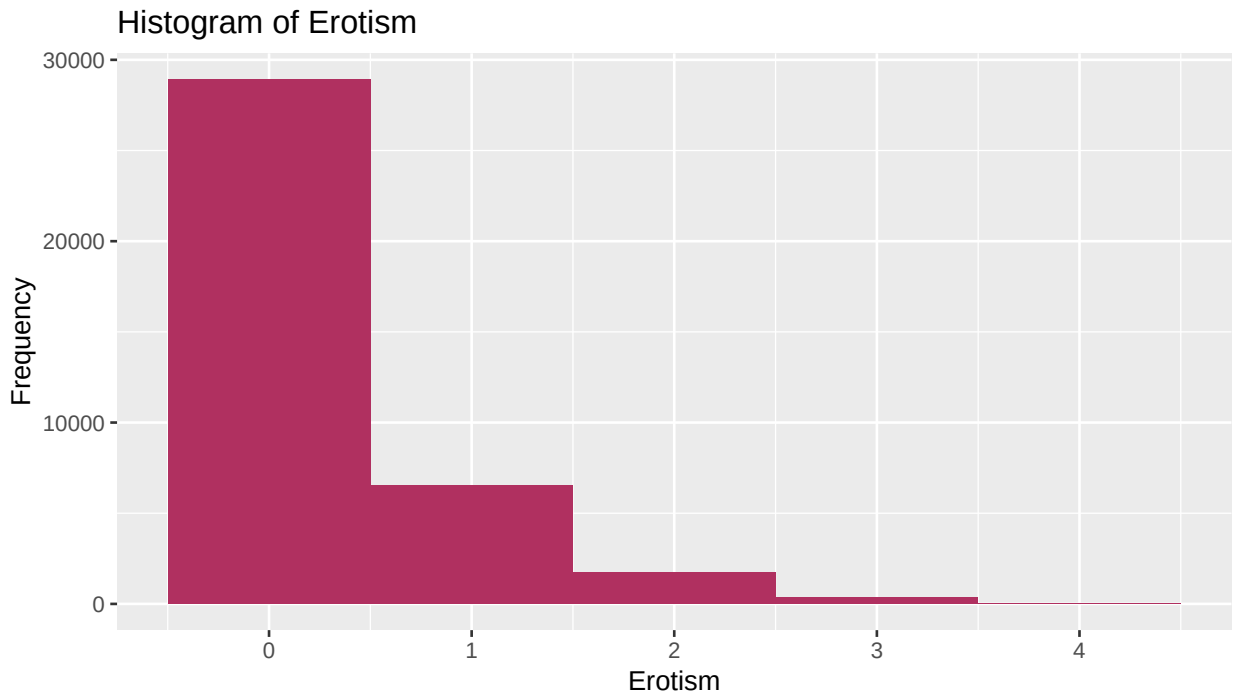

Histogram of Movie Release Year
Frequency
12000
9000
6000
3000
0
1920
1960
2000
Year


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
25000
20000
15000
10000
5000
0
Frequency
0
2
4
Effort


Histogram of Tension
15000
10000
5000
0
Frequency
0
2
4
Tension



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

PCA

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`.

Simple Regression

LASSO

Logistic Regression