DSCC462 Final project- Selenge Tulga

Analyzing IMDB movies.

1. Loading dataset

```
movies <- read.csv("/Users/selengetulga/Downloads/Fall 2023/DSCC 462/Final/TMDB_movie_da
n <- nrow(movies)
```

```
df_movies = movies[movies$budget > 0 & movies$revenue > 0 & movies$vote_count >
    0 & movies$vote_average > 0 & movies$runtime > 0, ]
dim(df_movies)
```

```
## [1] 10331    23
```

```
summary(movies)
```

```
##        id             title            vote_average    vote_count
##  Min.   :      2   Length:951126      Min.   : 0.0   Min.   :    0.00
##  1st Qu.: 355652   Class :character   1st Qu.: 0.0   1st Qu.:    0.00
##  Median : 634594   Mode  :character   Median : 0.0   Median :    0.00
##  Mean   : 632179                      Mean   : 2.2   Mean   :   22.55
##  3rd Qu.: 927064                      3rd Qu.: 5.2   3rd Qu.:    1.00
##  Max.   :1202079                      Max.   :10.0   Max.   :34495.00
##     status          release_date          revenue              runtime
##  Length:951126      Length:951126      Min.   :       -12   Min.   :    0.00
##  Class :character   Class :character   1st Qu.:         0   1st Qu.:    2.00
##  Mode  :character   Mode  :character   Median :         0   Median :   36.00
##                                        Mean   :    770388   Mean   :   51.57
##                                        3rd Qu.:         0   3rd Qu.:   90.00
##                                        Max.   :2923706026   Max.   :14400.00
##     adult           backdrop_path          budget            homepage
##  Length:951126      Length:951126      Min.   :        0   Length:951126
##  Class :character   Class :character   1st Qu.:        0   Class :character
##  Mode  :character   Mode  :character   Median :        0   Mode  :character
##                                        Mean   :   302711
```

```
##                                       3rd Qu.:        0
##                                       Max.   :888000000
##    imdb_id          original_language  original_title       overview
##  Length:951126      Length:951126      Length:951126      Length:951126
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    popularity       poster_path          tagline             genres
##  Min.   :   0.000   Length:951126      Length:951126      Length:951126
##  1st Qu.:   0.600   Class :character   Class :character   Class :character
##  Median :   0.600   Mode  :character   Mode  :character   Mode  :character
##  Mean   :   1.414
##  3rd Qu.:   0.954
##  Max.   :2994.357
##  production_companies production_countries spoken_languages
##  Length:951126        Length:951126        Length:951126
##  Class :character     Class :character     Class :character
##  Mode  :character     Mode  :character     Mode  :character
##
##
##
```

```r
dim(df_movies)
```

```
## [1] 10331    23
```

```r
# Remove unnecessary columns
df_movies$poster_path <- NULL
df_movies$homepage <- NULL
df_movies$backdrop_path <- NULL
df_movies$imdb_id <- NULL
```

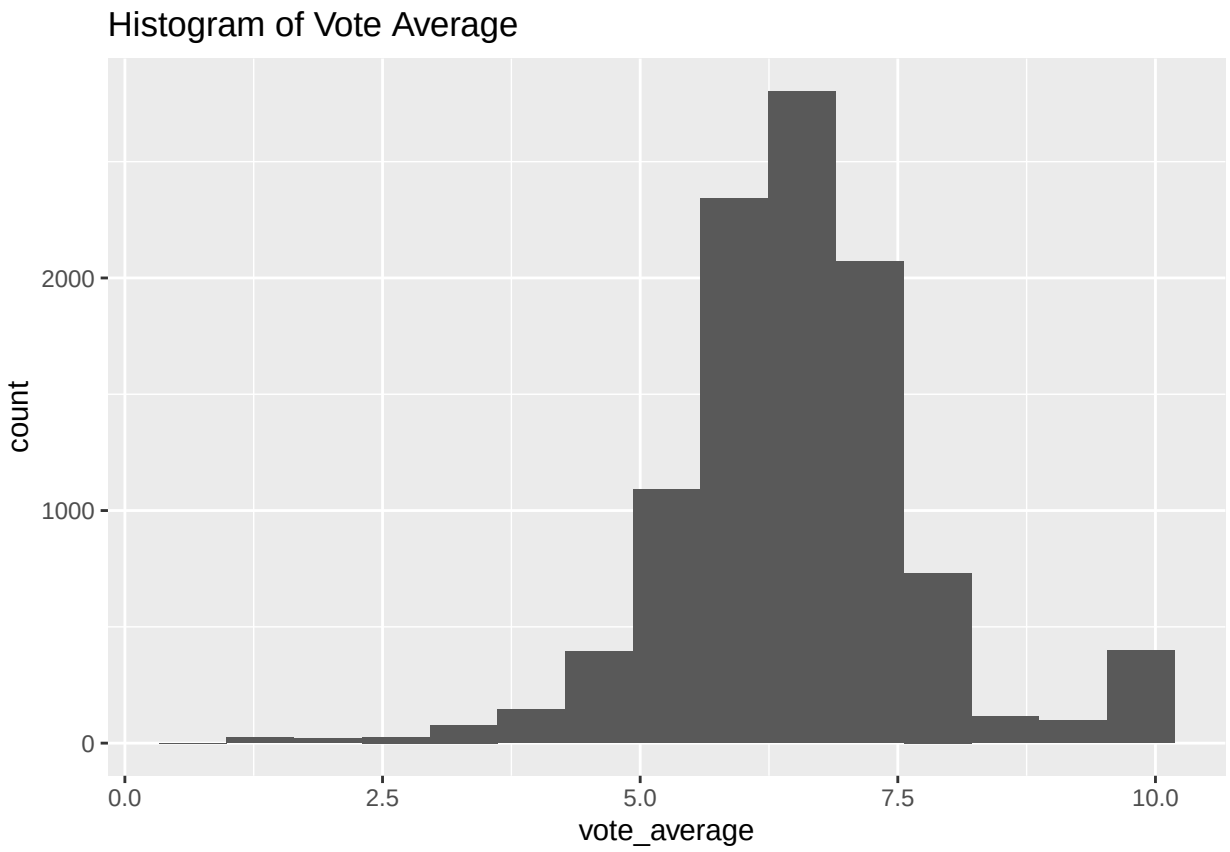
```r
dim(df_movies)
```

```
## [1] 10331    19
```

```r
vote_average <- df_movies$vote_average
ceiling(log(length(vote_average), 2)) + 1
```

```
## [1] 15
```

```r
library(ggplot2)
hist1 <- ggplot(df_movies, aes(x = vote_average)) + geom_histogram(bins = 15)
hist1 <- hist1 + ggtitle("Histogram of Vote Average")
hist1
```




```r
quantile(vote_average, c(0.2, 0.8))
```

```
##   20%   80%
## 5.700 7.281
```

```r
IQR(vote_average)
```

```
## [1] 1.255
```

```r
ss <- 1.5 * IQR(vote_average)
ss
```

```
## [1] 1.8825
```

```r
lf <- quantile(vote_average, 0.25) - ss
lf
```

```
##   25%
## 3.987
```

```r
uf <- quantile(vote_average, 0.75) + ss
uf
```

```
##   75%
## 9.007
```

```r
sum(vote_average < lf)
```

```
## [1] 188
```

```r
sum(vote_average > uf)
```

```
## [1] 411
```

```r
var(vote_average)
```

```
## [1] 1.50006
```

```r
sd(vote_average)
```

```
## [1] 1.224769
```

```r
sd(vote_average)/mean(vote_average)
```

```
## [1] 0.187842
```

```r
library(moments)
skewness(vote_average)
```

```
## [1] 0.1940057
```