BenfordLaw

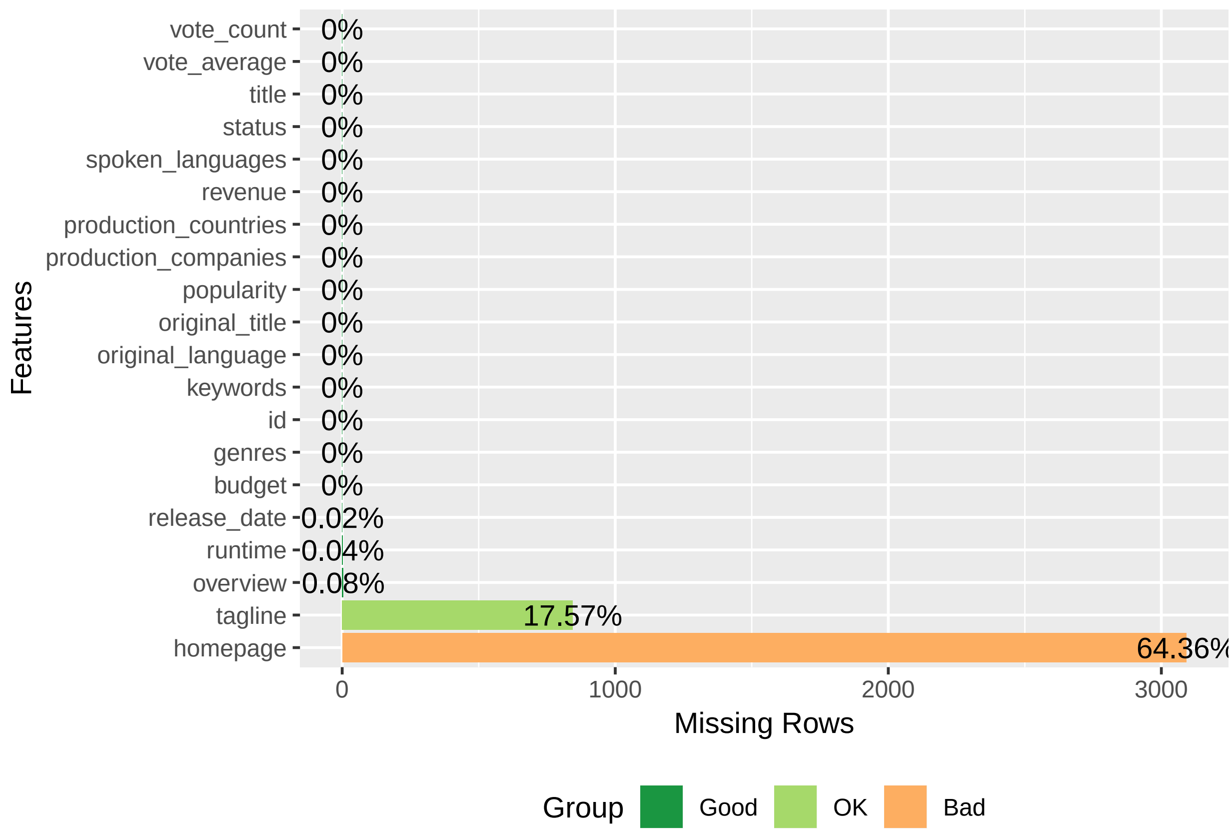
Tingrui Huang

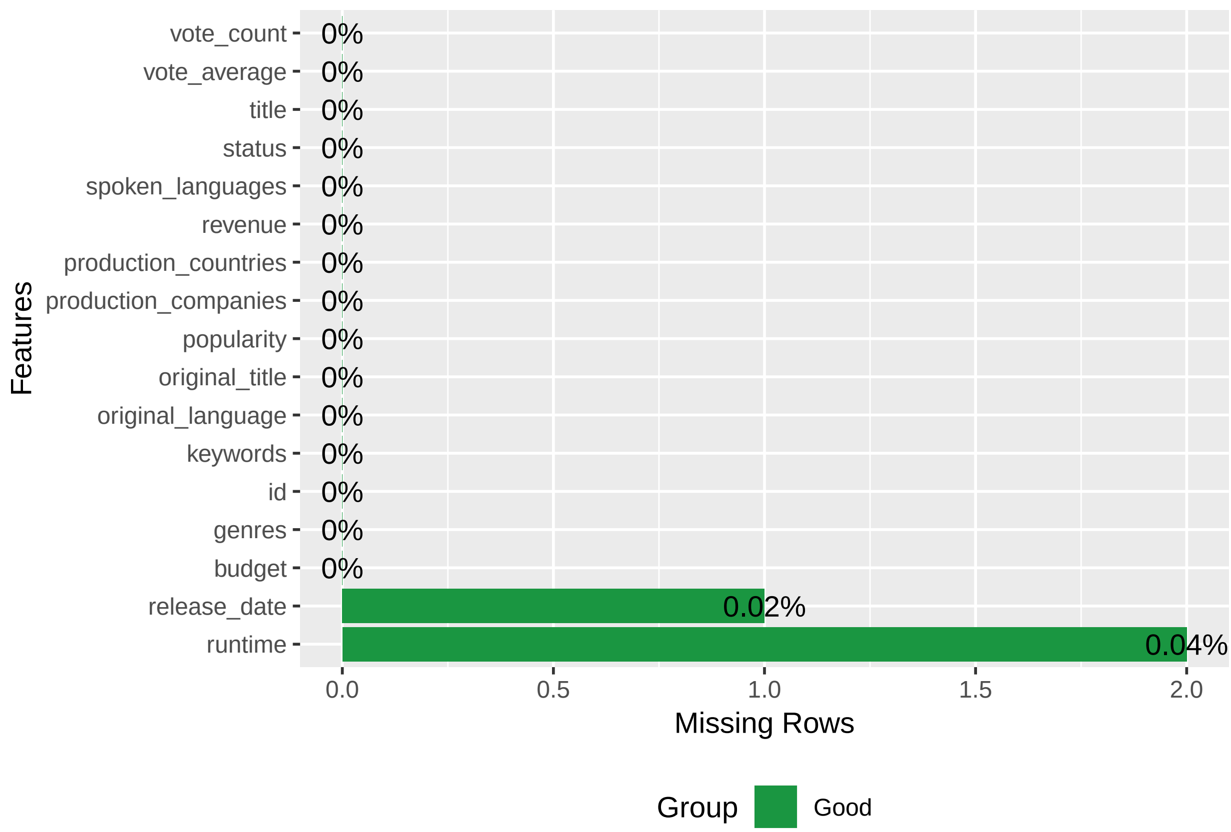
December 6, 2018

Introduction

In this project, I'm going to do the Benford analysis on the movies' budget, revenue, vote average(ratings), vote count. In the analysis, I will use 2 datasets, the movie dataset and the credit dataset. In the movie dataset, there are 4803 movies with 20 explanatory variables including title, language, released date, budget, revenue, runtime and so on. The credit dataset contains the list of crews and directors for each movie.

First of all, let's take a look at the missing values in the dataset.





After looking at the initial missing value plot, I decided to remove the "tagline", "homepage" and "overview", since for now, I'm not going to use these variables and they have too many missing values. In later analysis, I will include "overview" to do the sentiment analysis.

Data Preparation

Since the "keyword", "genres", "production company" and "country" are in JSON format, I use a package called "jsonlite" to reformat these variables and subtract these columns from the main table. Since one movie could correspond to multiple keywords and genres, therefore, if I didn't subtract those columns from the main table, they would make the table much longer and create repetitive information.

I added "released year", "released month" and "profit" into the table. And reformatted some variables for later analysis.

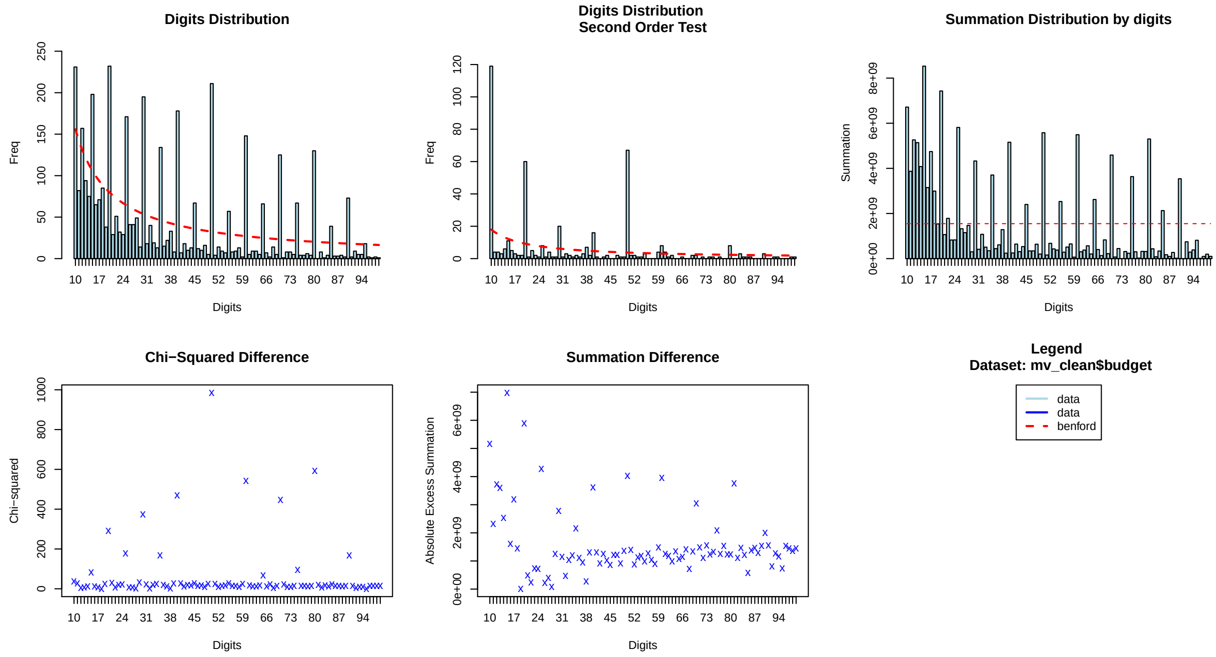
I. Benford Analysis

(i) Overview

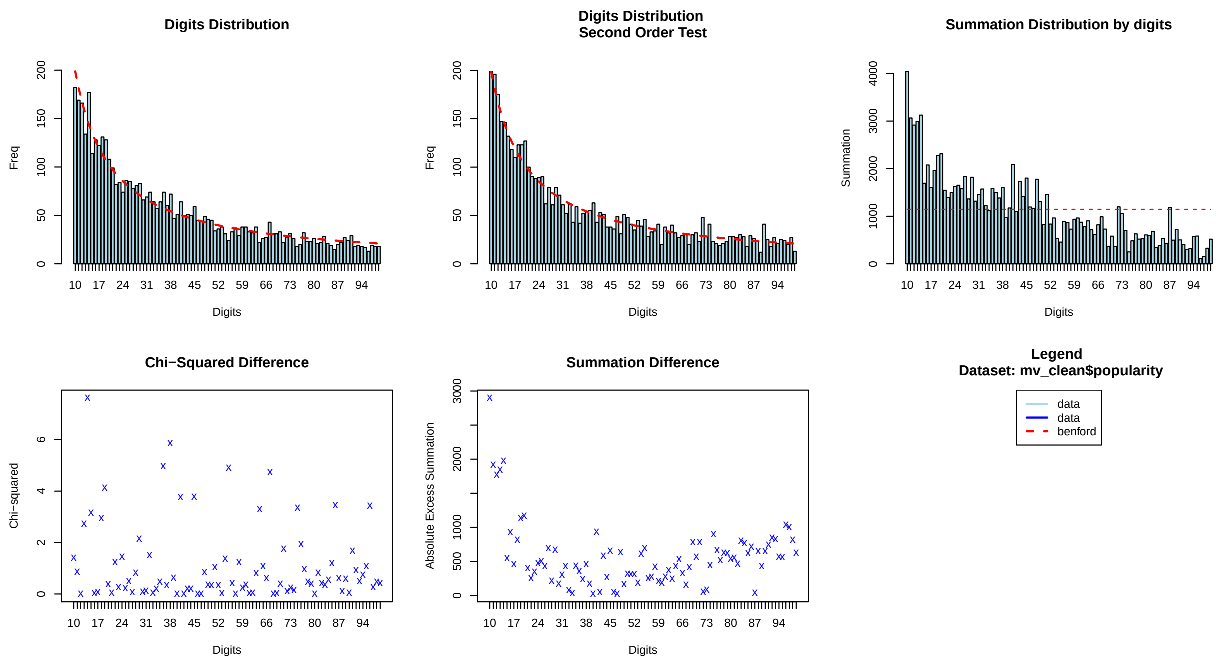
In the Benford analysis, I'm going to find out the suspecious values in "budget", "popularity", "revenue", "runtime", "vote average" and "vote count". Over the past a few years, a lot of movies are accused to be misreporting their revenue so that they could attract more attention from the public. Meanwhile,lots of production companies are accused to pay for people to write positive reviews to their movies.

Thereore, I think it will be interesting to explore if there is any movie that cheated on its statistics.

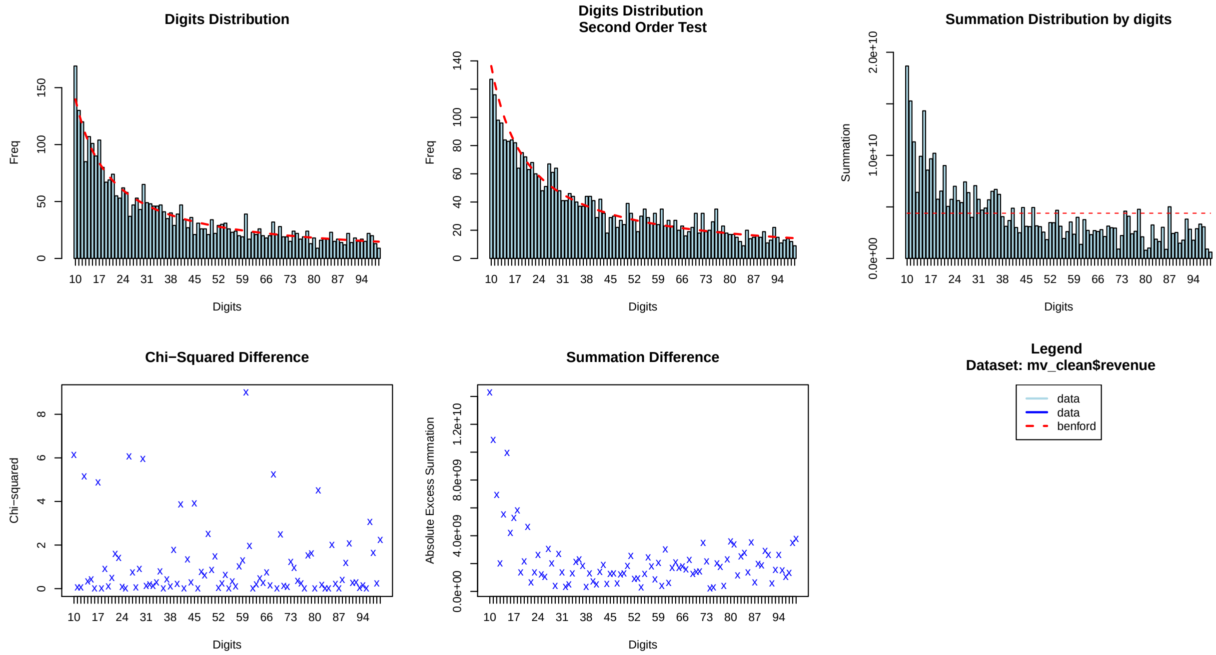
```r
# budget
bfd.budget <- benford(mv_clean$budget)
plot(bfd.budget)
```



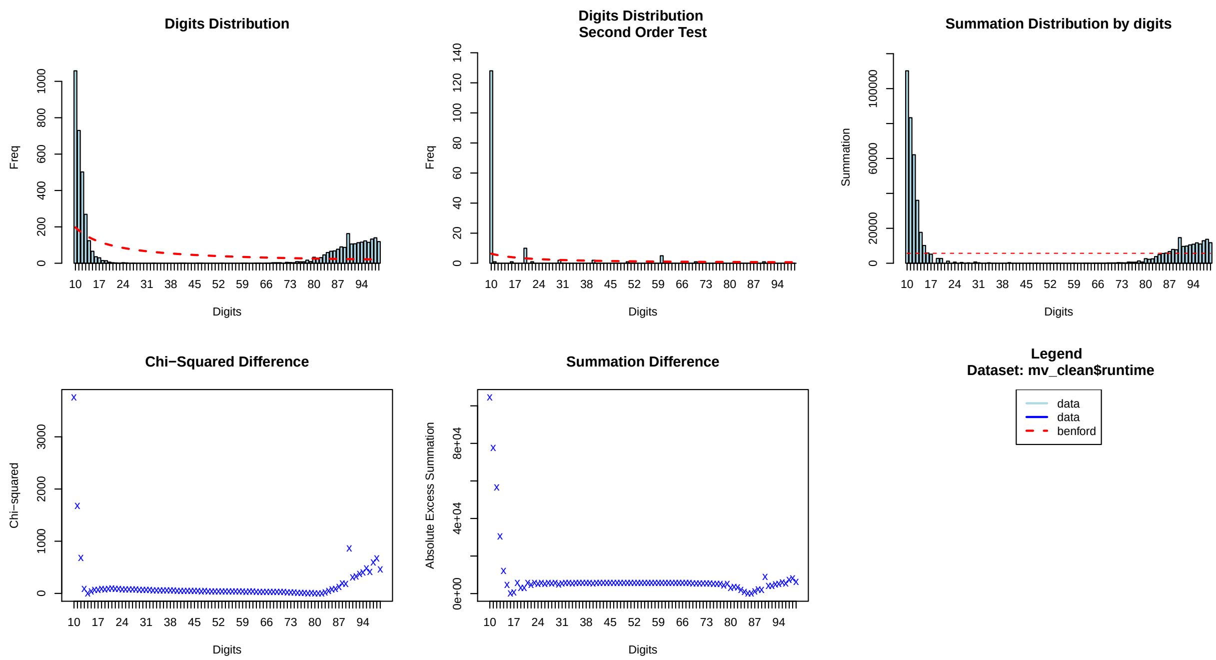

```r
# popularity
bfd.popular <- benford(mv_clean$popularity)
plot(bfd.popular)
```



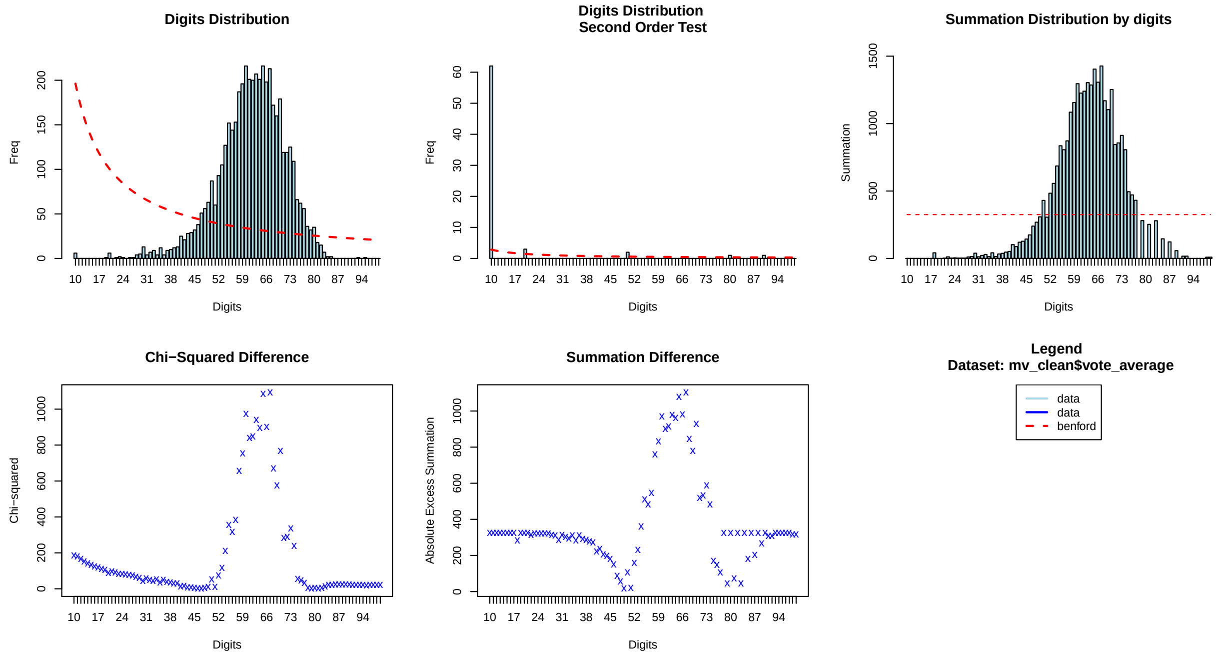
```r
# revenue
bfd.rev <- benford(mv_clean$revenue)
plot(bfd.rev)
```



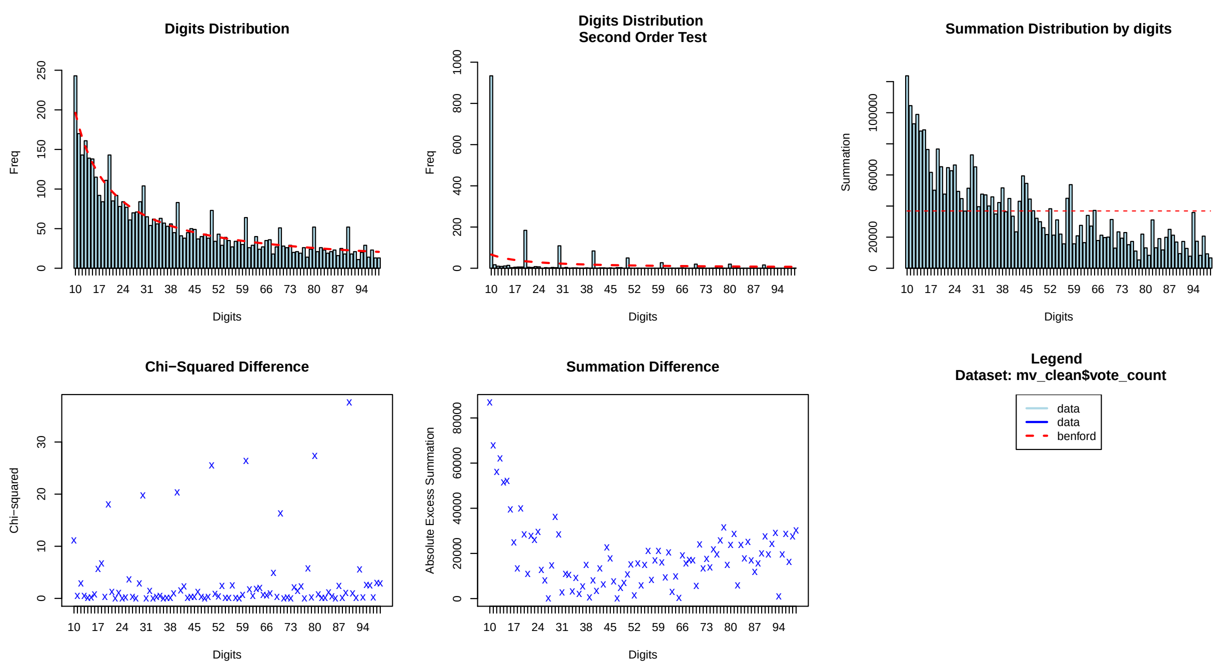

```r
# runtime
bfd.rt <- benford(mv_clean$runtime)
plot(bfd.rt)
```



```r
# vote average
bfd.votea <- benford(mv_clean$vote_average)
plot(bfd.votea)
```




```r
# vote count
bfd.votec <- benford(mv_clean$vote_count)
plot(bfd.votec)
```



Discussion on initial Benford Analysis

As we can see from the plots of Benford Analysis, "budget", "runtime" and "vote average" do not follow Benford distribution. For "runtime", I think the reason is that most movies run between 30 and 60 days, there is only a few movies run less than 10 days or more than 100 days. For "vote average", the scale of vote rating is from 1 to 10, and most people tend to give mediocre scores between 5 to 8, therefore, the average is heavily centered between 5 and 8. However, it is weird to see "budget" does not follow Benford Law, and I think if a production company cheats on the budget of their movies, they could be benefitted by doing that. Therefore, I would like to take a further look at the suspecious values in the "budget".

Based on these facts, I would not say suspecious values are the reason for "runtime" and "vote average" do not follow Benford distribution. But there could be tampering data in "budget".

On the other hand, the plots of "popularity", "revenue" and "vote count" generally follow Benford distribution and only small portion of the data do not follow it. In the following analysis, I will take closer look at the suspecious movies based on the previous Benford analysis.

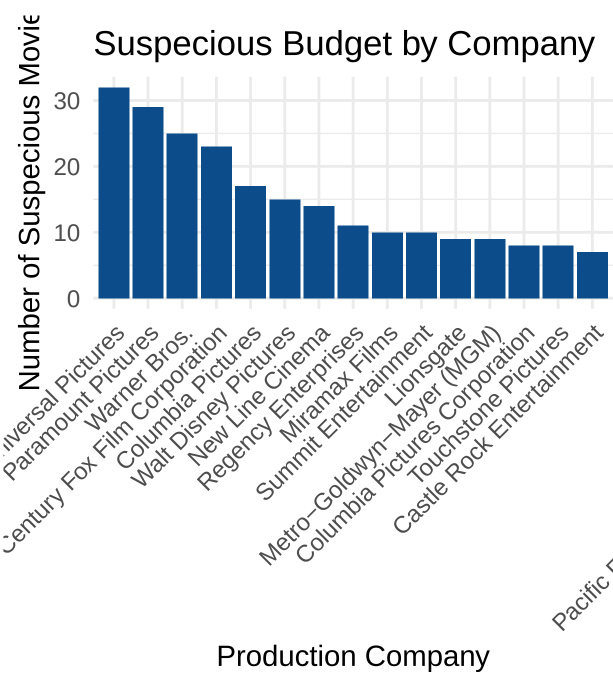
(ii) Zoom in on each topic

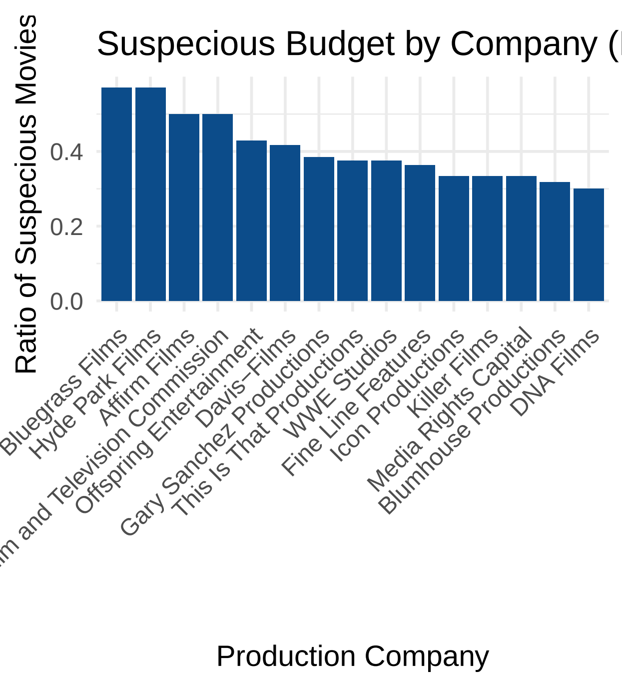
Suspecious movies in "budget"

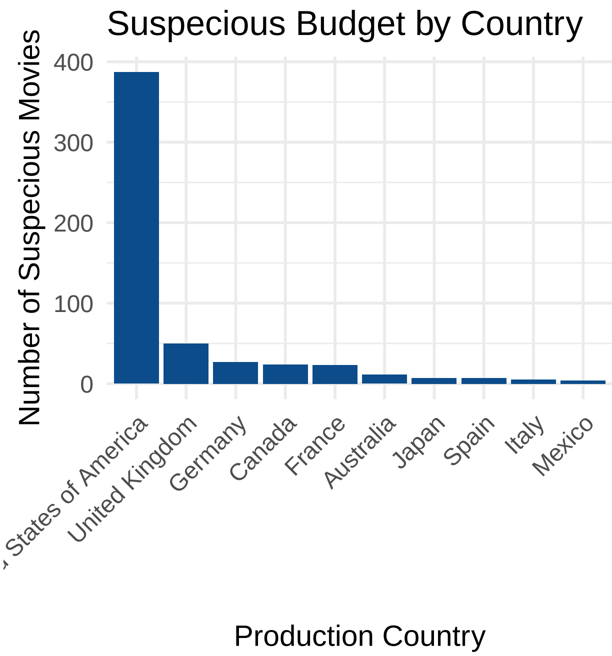
```
# Look at chisq first
chisq(bfd.budget)
```

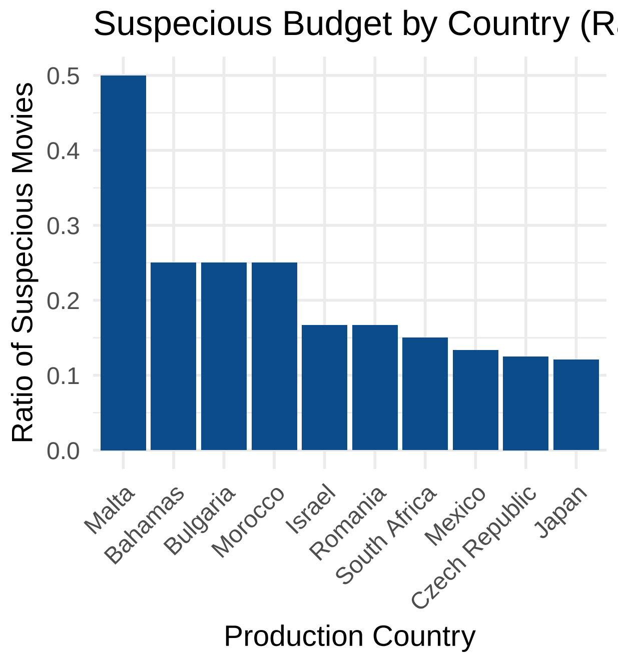
```
##
##  Pearson's Chi-squared test
##
## data:  mv_clean$budget
## X-squared = 5582, df = 89, p-value < 2.2e-16
```

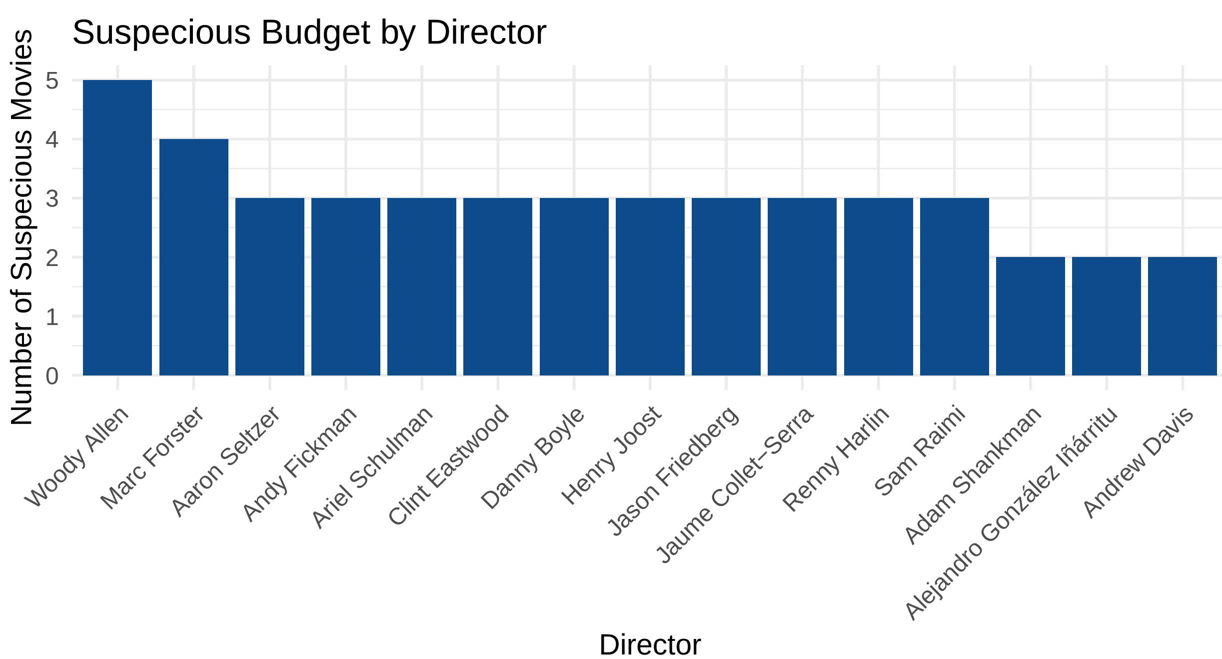
```
# As we can see, the p-value indicates the distribution of "budget"
# is not the same as Benford distribution
```











Summary:

As we can see the top 5 companies with most suspecious budgets are all bignames, including Universal Pictures, Paramount, Warner Bros. and so on. However, if we check out the ratio of suspecious movies a company made, smaller production companies popped up.

As the largest movie production contry, not quite suprisingly, the USA has the most suspecious movies in budget. If we look at the ratio, smaller countries tend to have higher ratio of suspecious movies.

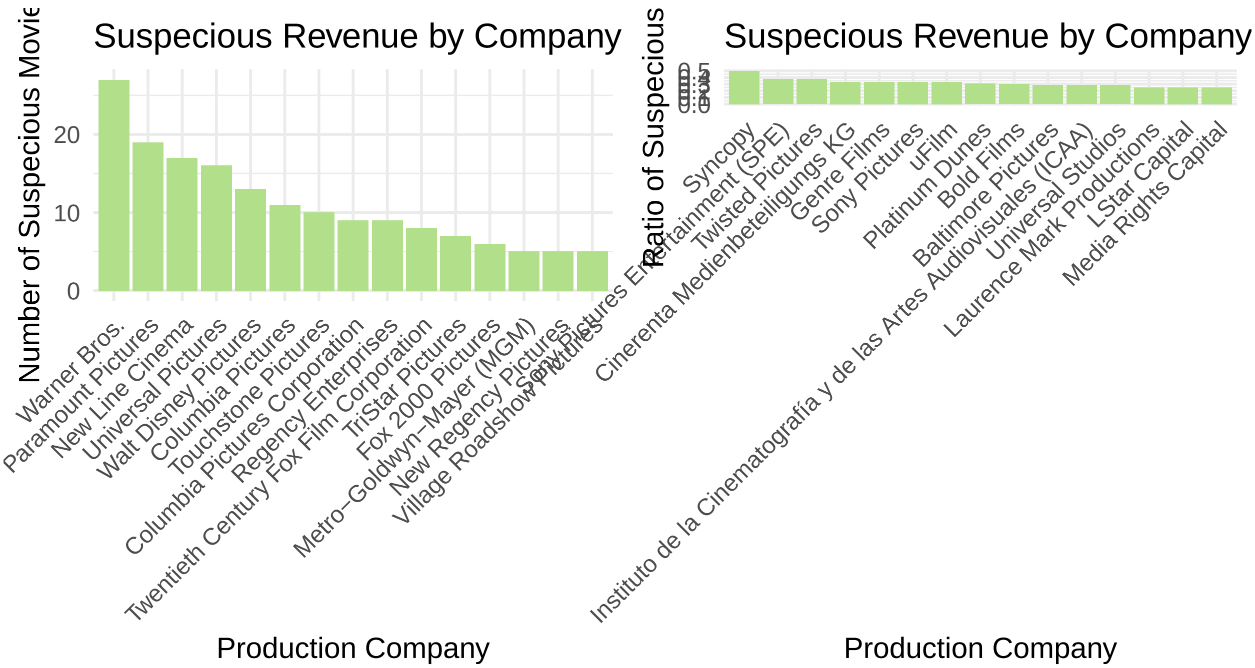
Suprisingly, bigname directors like Woody Allen showed up on the list of director with most suspecious movies in budget. However, I would not say those five movies Woody directed cheated on their budget, since this is only a simple analysis.

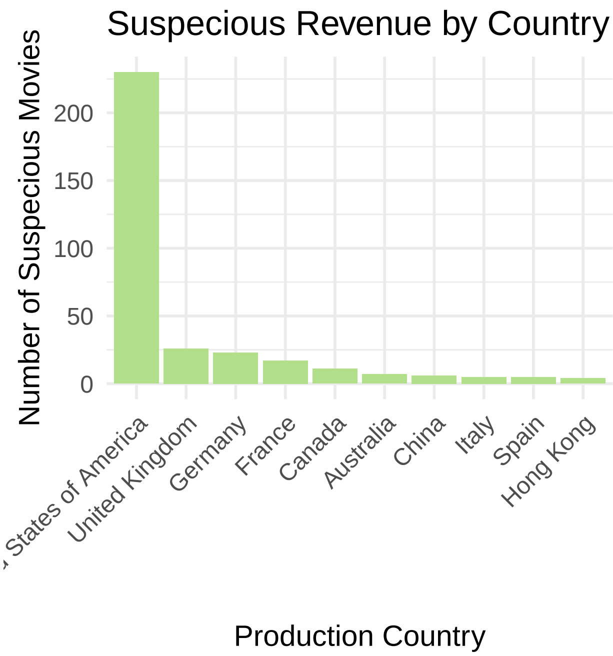
Suspecious movies in "revenue"

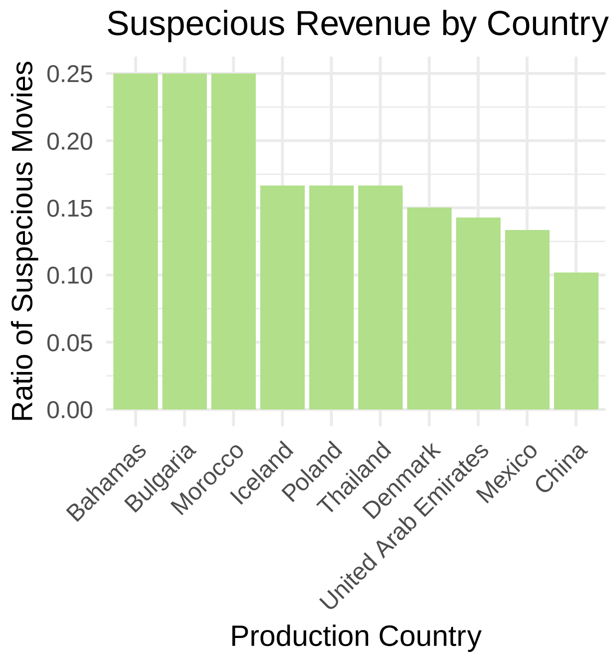
```r
# Movies with suspecious revenue
chisq(bfd.rev)
```

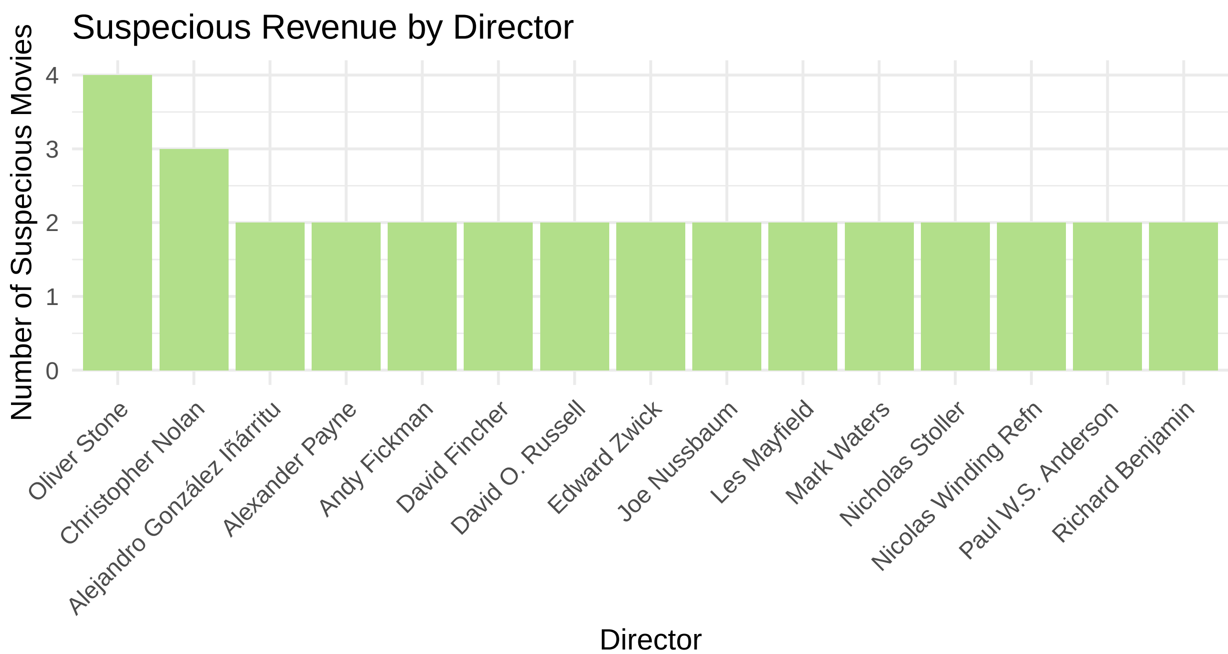
```
##
##  Pearson's Chi-squared test
##
## data:  mv_clean$revenue
## X-squared = 103.63, df = 89, p-value = 0.1376
```

```r
# As we can see, the p-value indicates the distribution of "revenue"
# is the same as Benford distribution
```









Summary:

Revenue is one of the statistics that movie production companies often misreport to attract more investors. By taking out the suspecious companies, just like what we have seen in the budget analysis, we can see lots of bignames like Warner Bros, Paramount, New Line and so on. Again, ranking companies by the ratio of suspecious movies over total movies they made, smaller production companies showed up. But the list of smaller companies is different from the list in the budget analysis.

Again, the USA surpasses other competitors by huge amount, and small countries have higher ratio of suspecious movies.

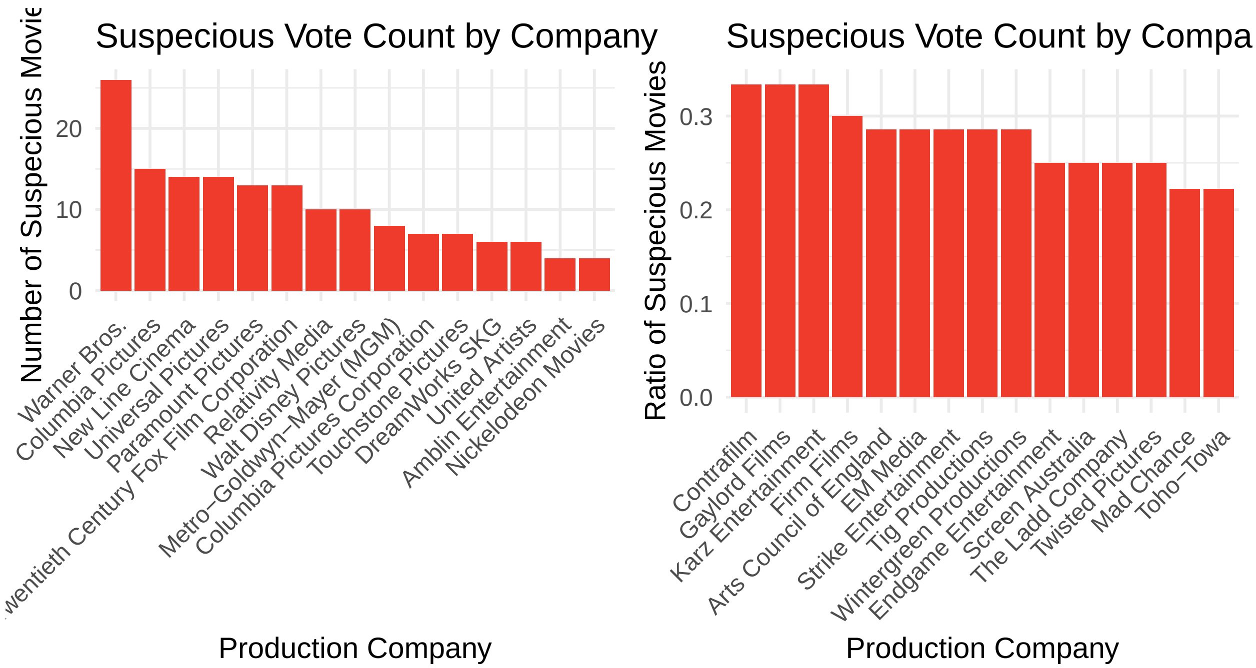
It is shocking to see Christopher Nolan on the list since he is one of my favorite directors. Again, this will not be a proof that the directors have cheated on the revenue of their movies.

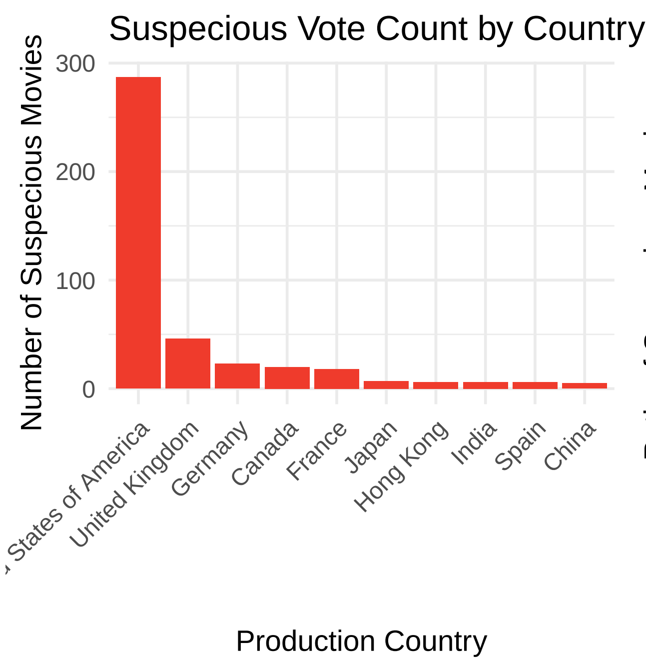
Suspecious movies in "vote count"

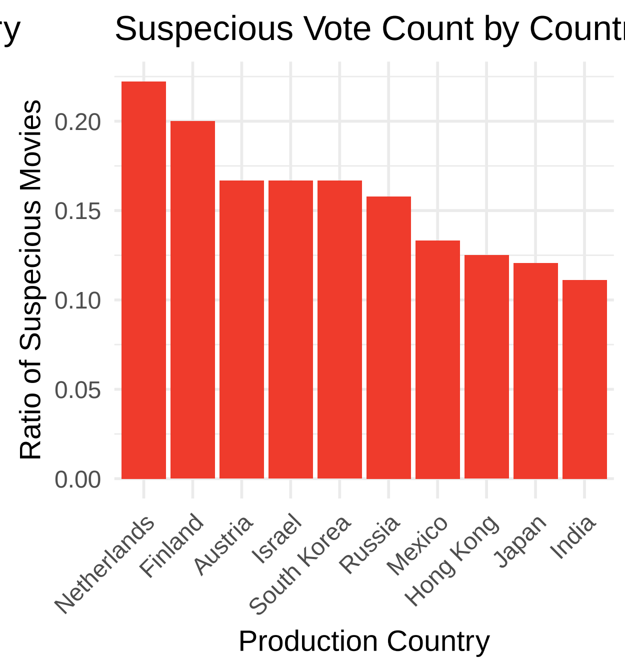
```
# Movies with suspecious vote count
chisq(bfd.votec)
```

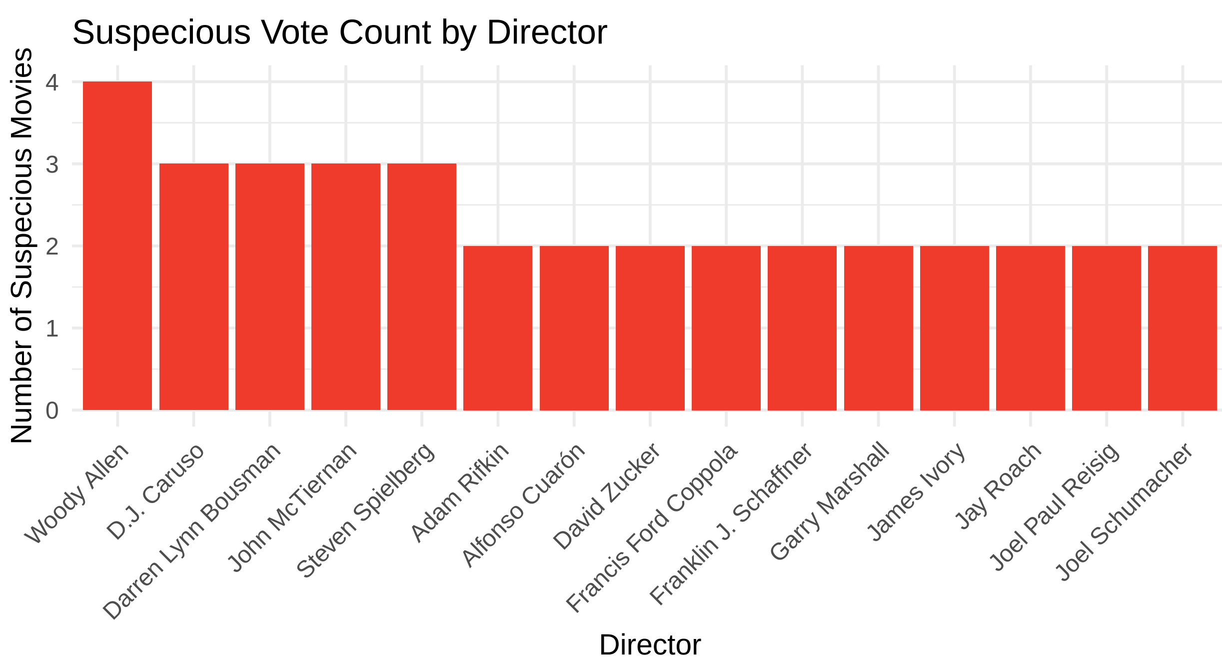
```
##
##  Pearson's Chi-squared test
##
## data:  mv_clean$vote_count
## X-squared = 294.25, df = 89, p-value < 2.2e-16
```

```
# As we can see, the p-value indicates the distribution of "vote count"
# is the not same as Benford distribution
```









Summary:

The vote count is the number of people that give ratings to a movie. Some movies have been suffering the scandal of buying voters to boost their ratings. In my analysis, the big movie companies such as Warner Bros., Columbia Picture, New Line Cinema and so on have lots of movies with suspecious vote count. If we look at the ratio, smaller companies showed up.

Surprisingly, some of the South Korea and Japan are on the list of ratio of suspecious movies. Since both of these two countries have respectful movie industry.

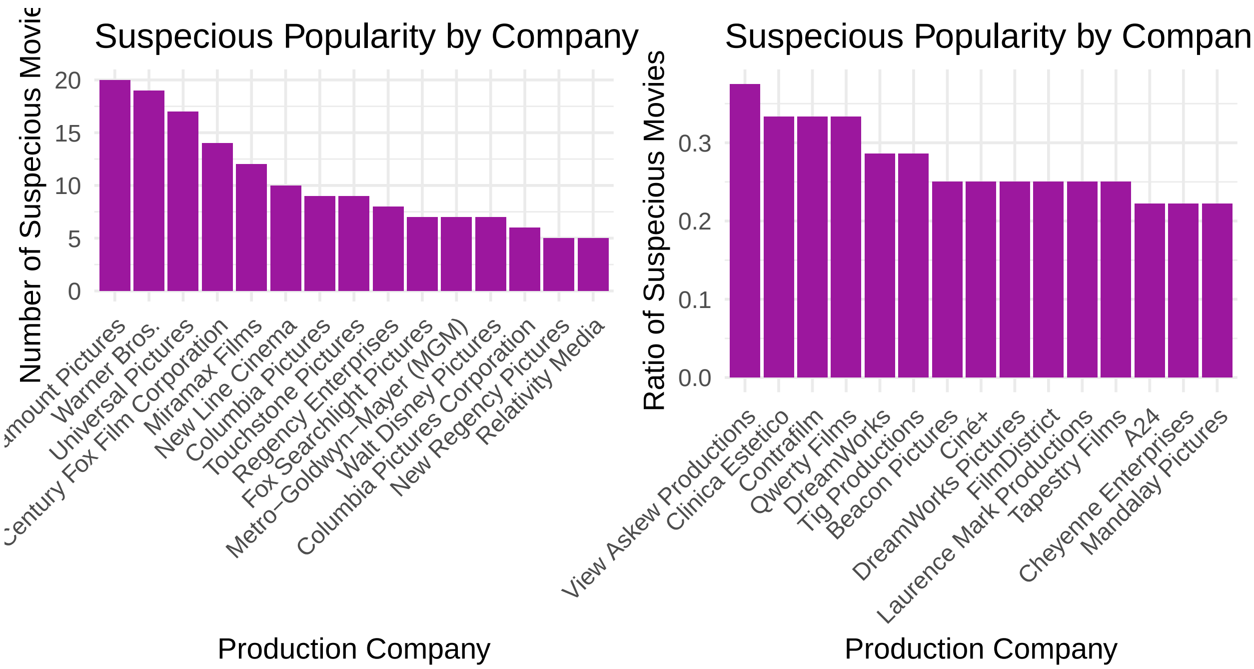
Again, we see Woody Allen's name at the top of the suspecious movie list, and we can find another famous movie director Steven Spielberg on the list too.

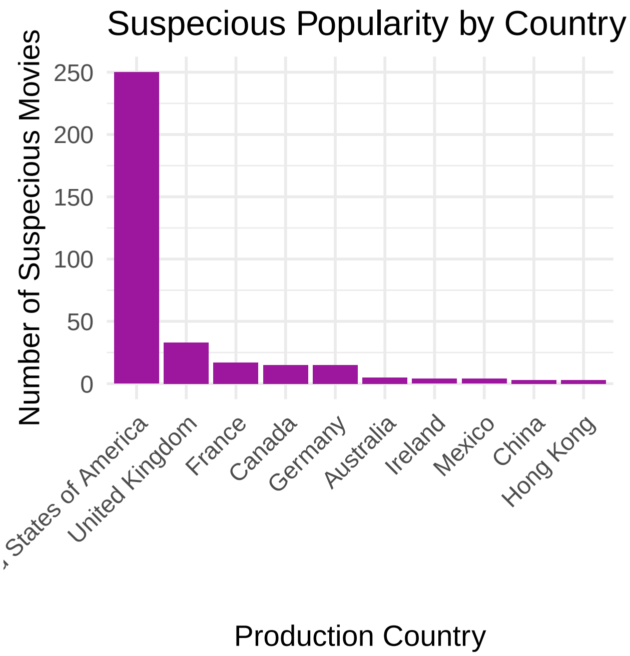
Suspecious movies in "popularity"

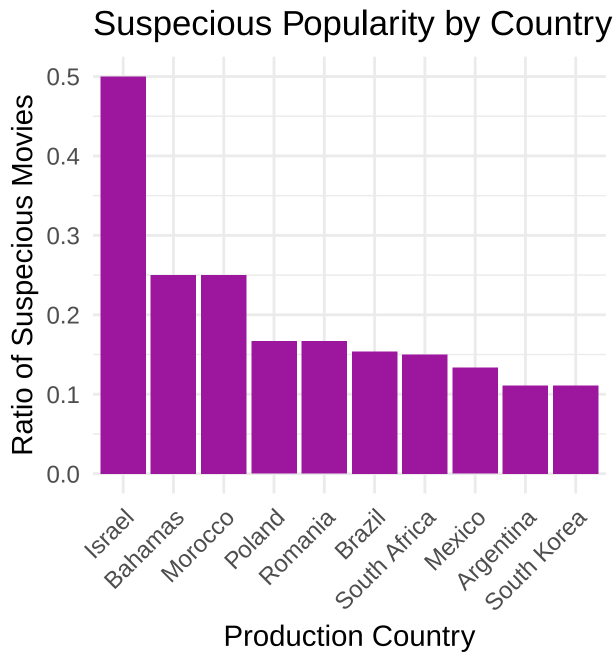
```
# Movies with suspecious popularity
chisq(bfd.popular)
```

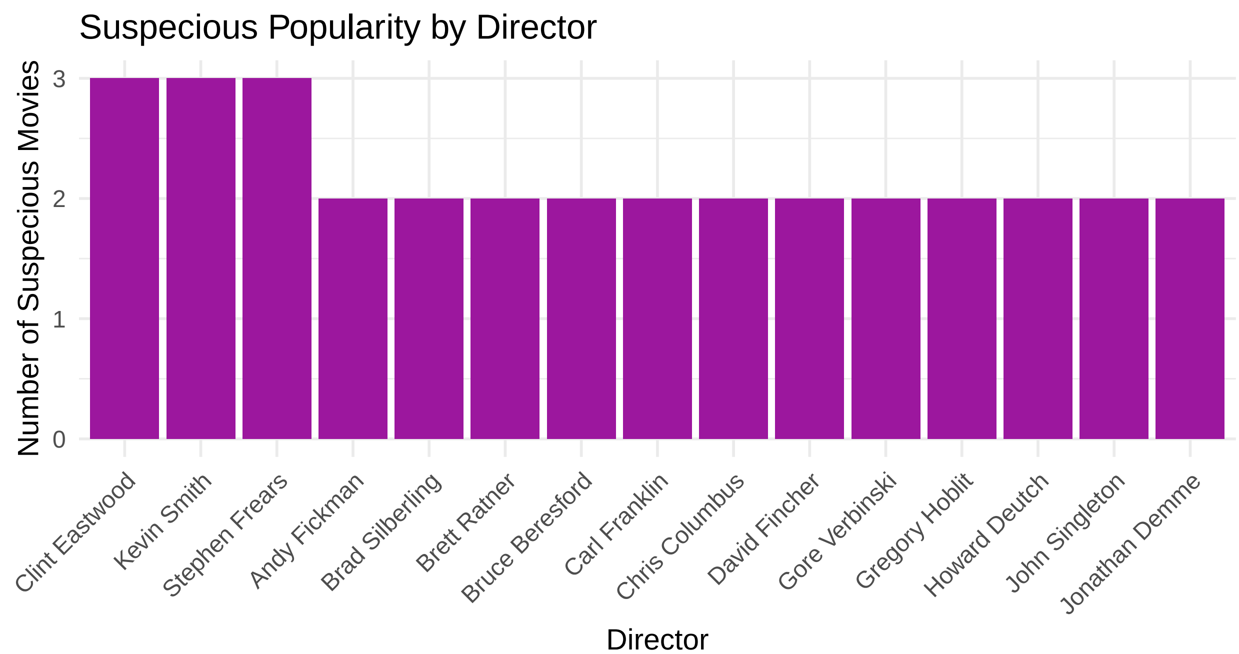
```
##
##  Pearson's Chi-squared test
##
## data:  mv_clean$popularity
## X-squared = 101.88, df = 89, p-value = 0.1655
```

```
# As we can see, the p-value indicates the distribution of "vote count"
# is the not same as Benford distribution
```









Summary:

Popularity is a score that indicates how popular a movie is. The predictors that are used to calculate popularity include # of votes for the day, # of views for the day, # of total votes, # of users who marked the movie as "favorite" and previous day score.

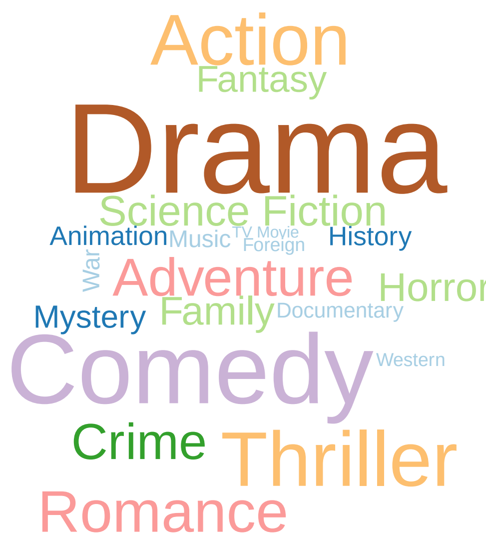
Since popularity is a score of the combination of many factors, I would say it is difficult to tell whether the value is suspecious. Even if we could tell the suspecious values, we do not have the method to decomposite the score and find out the suspecious part.

The results from the plots look similar to previous analysis.

II. Text Mining

(i) Wordcloud

Genre (all years)



As we can see from the wordcloud, the most frequent genres in these 4,800 movies are Drama, Comedy, Thriller and Action.

Genre (2015-2017)

Now, let's look at the most popular genres from 2010, and I guess the result might be different from the result of all time.

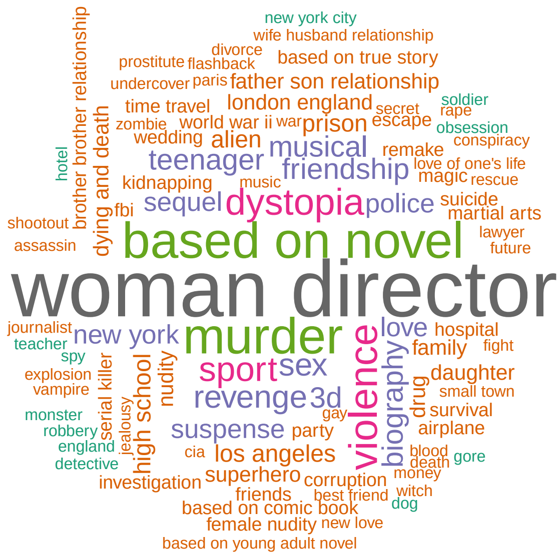
Surprisingly, the result looks quite similar to the previous result, the most popular genres are Drama and Comedy, following by Thriller and Action movies.

Keywords (all years)

```
## Warning in wordcloud(words = wc_kws$keyword, freq = wc_kws$n, max.words =
## 100, : independent film could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = wc_kws$keyword, freq = wc_kws$n, max.words
## = 100, : duringcreditsstinger could not be fit on page. It will not be
## plotted.

## Warning in wordcloud(words = wc_kws$keyword, freq = wc_kws$n, max.words
## = 100, : aftercreditsstinger could not be fit on page. It will not be
## plotted.
```

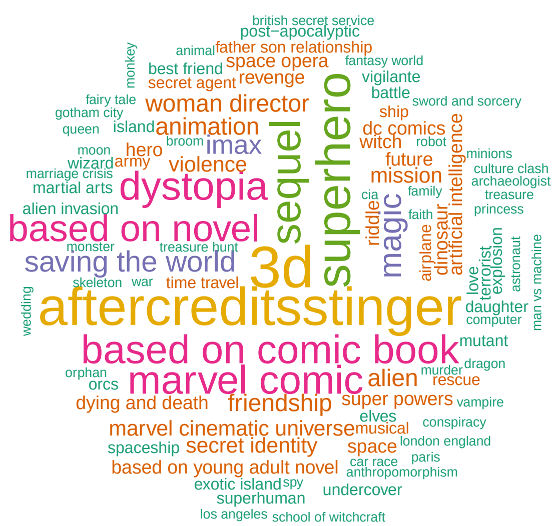


As we can see, the most popular keywords overall are “woman director”, “independent filt”, “during credits stinger”, “based on novel” and so on.

Keywords (top 300 in revenue)

Now let’s look at the popular keywords for the movies that have highest revenue. Let’s see if we can find any patterns here!

```
## Warning in wordcloud(words = wc_kws_t300$keyword, freq = wc_kws_t300$n, :
## duringcreditsstinger could not be fit on page. It will not be plotted.
```



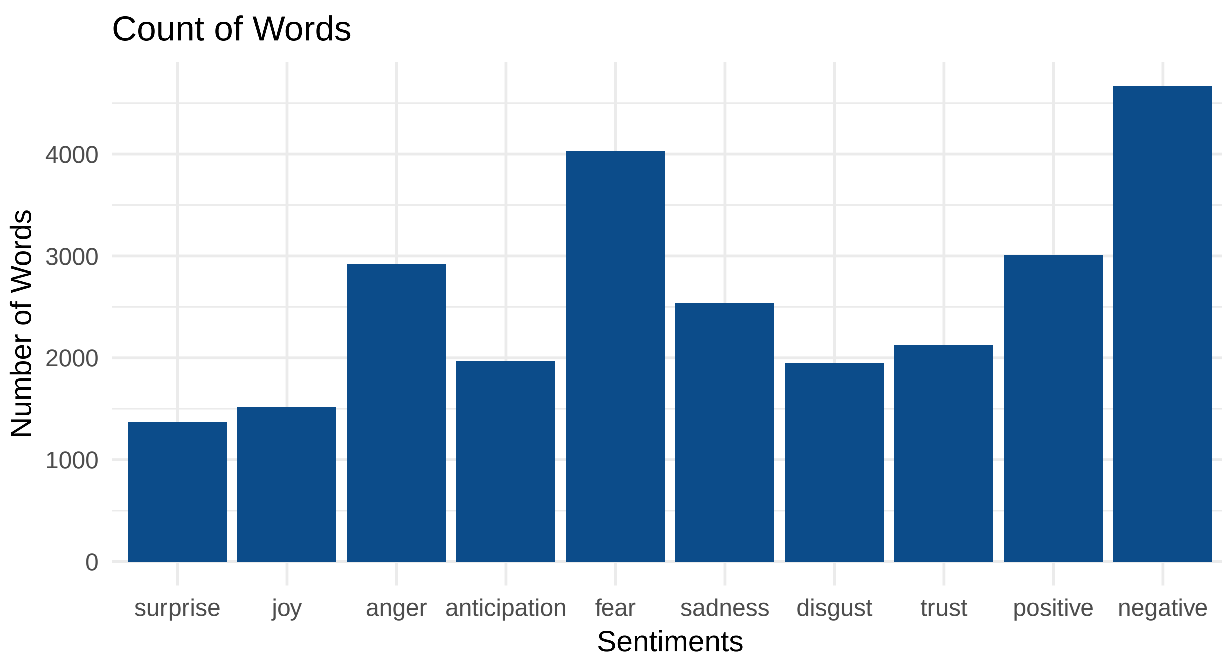
Here we can see some difference. While "during credits stinger" is still a popular keyword, we also see many new keywords showed up, such as "3d", "super hero", "alien", "magic", "saving the world" and so on. In current movie industry, I believe super hero movies are really popular and making lots of money. If you are going to make films or invest in production compnies, target on these!

(ii) Sentiment Analysis

Keywords - lexicon choice: "nrc"

The reason I chose "nrc" was because this lexicon has more sentimental levels/categories than the other two lexicons, and in the movie dataset, the keywords contains more than just negative or positive sentiments.

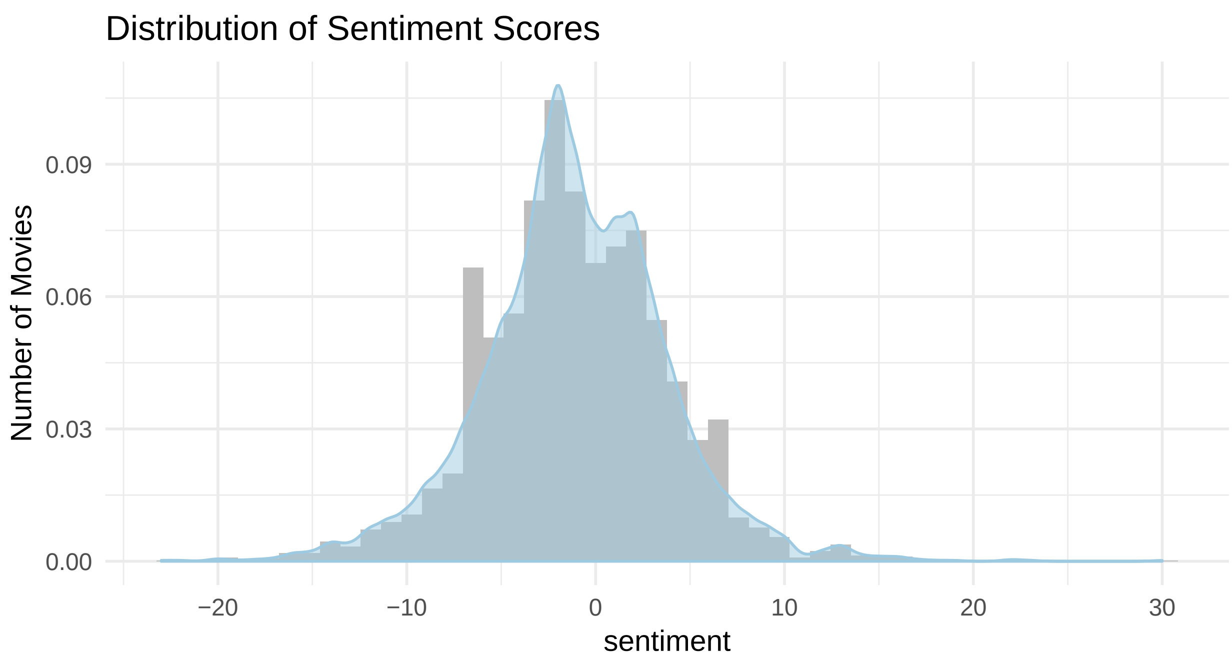
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



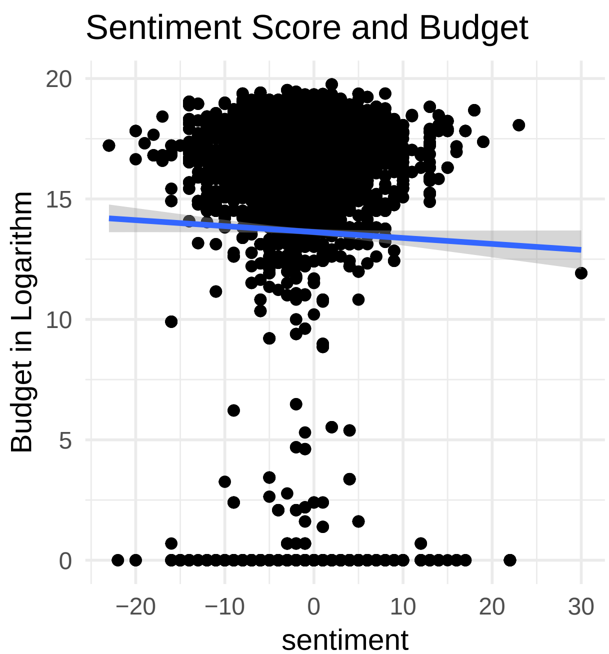
As we can see, the top four sentiments are "negative", "fear", "positive" and "anger". Let's dive into these four categories to see what words are included in them.

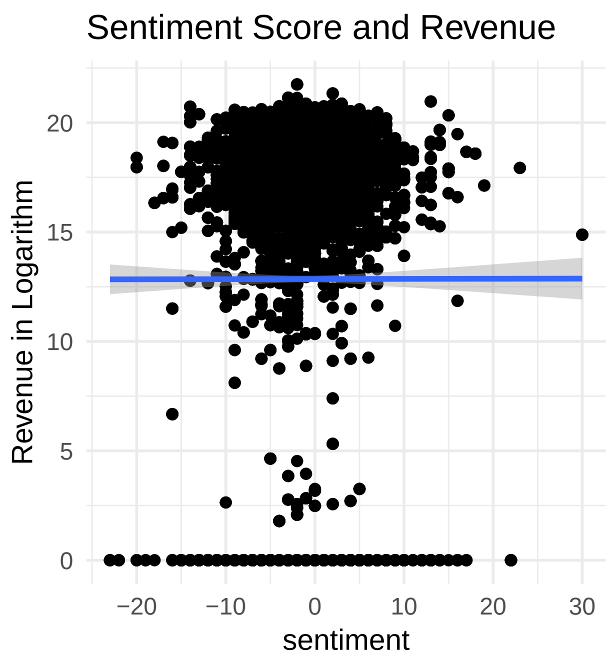
Overview (assign sentiment score to each movie) - lexicon choice: "afinn"

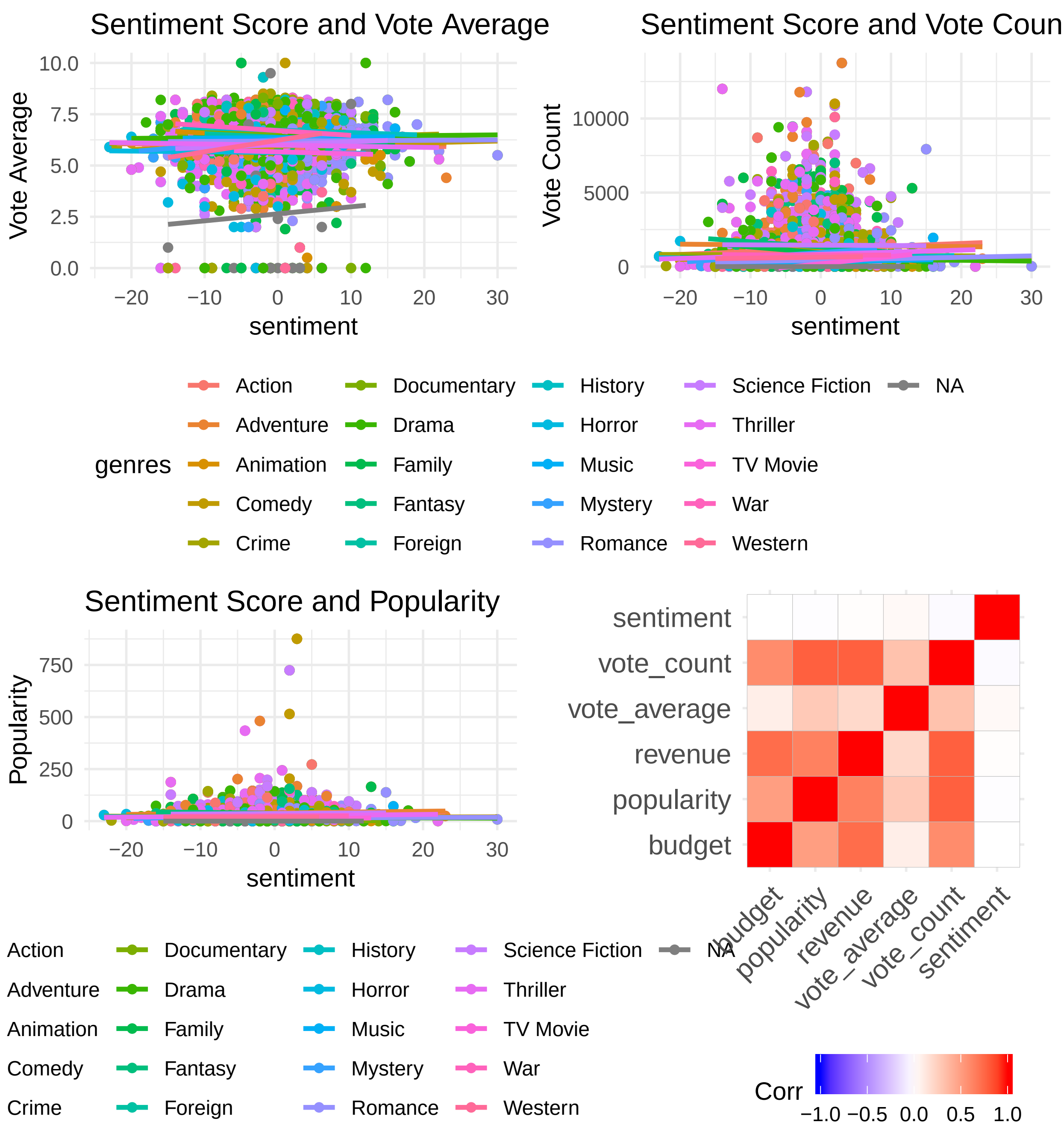
The variable "overview" contains the brief introduction of each movie, and I will assign each movie a sentiment score based on the sentiment analysis on its overview.



As we can see, most movies have a sentiment score between -7 and 7. Now, let's find out if the sentiment scores will have effects on movies' budget, popularity, revenue, vote average and vote count.







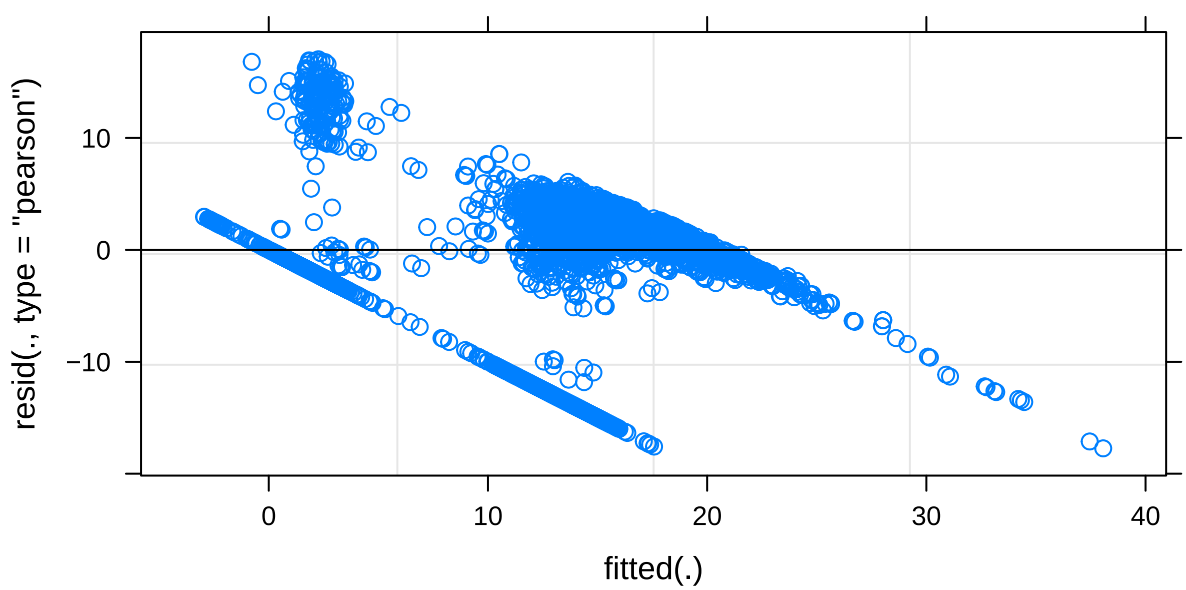
As we can see from the plots, in general, there is not much correlation between sentiment scores and other factors. However, from the correlation test, we do see some correlation between each of the factors. So, let's make a simple multilevel model to see if we could predict revenue by using these factors.

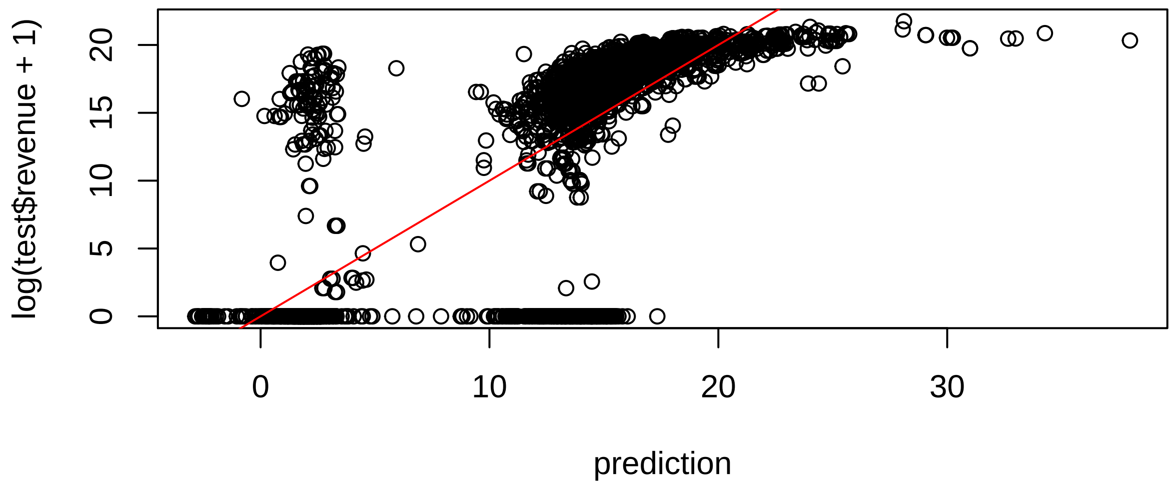
$$y_{logRevenue} \sim N(2.893 + 0.728X_{logBudget} + 0.633X_{scaledPopularity} + 0.731X_{scaledVoteAverage} + 0.957X_{scaledVoteCount} + 0.08X_{scaled}$$

$$u_{j[i]} \sim N(0, 0.22^2)$$

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
```

```
## log(revenue + 1) ~ log(budget + 1) + scale(popularity) + scale(vote_average) +
##     scale(vote_count) + scale(sentiment) + (1 | genres)
##    Data: train
##
## REML criterion at convergence: 46786.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2307 -0.2570  0.3110  0.5256  3.1017
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  genres   (Intercept)  0.05202 0.2281
##  Residual             30.14359 5.4903
## Number of obs: 7489, groups:  genres, 20
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)          2.89311    0.16774  17.247
## log(budget + 1)      0.72873    0.01020  71.439
## scale(popularity)    0.63358    0.09588   6.608
## scale(vote_average)  0.73114    0.06979  10.477
## scale(vote_count)    0.95733    0.09892   9.678
## scale(sentiment)     0.08079    0.06462   1.250
##
## Correlation of Fixed Effects:
##              (Intr) lg(+1) scl(p) scl(vt_v) scl(vt_c)
## log(bdgt+1) -0.843
## scl(pplrty)  0.082 -0.100
## scl(vt_vrg)  0.066 -0.078 -0.020
## scl(vt_cnt)  0.092 -0.115 -0.694 -0.217
## scl(sntmnt) -0.005  0.008 -0.017 -0.019     0.022
```





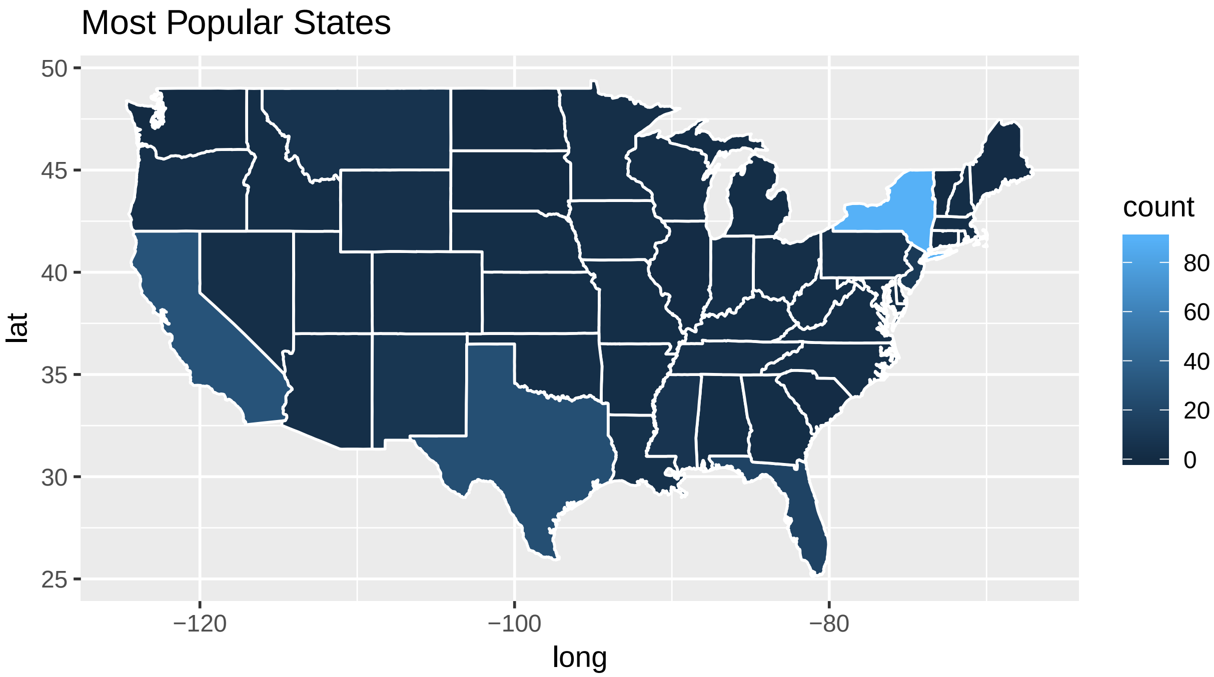

```
## integer(0)
```

From the residual plot we can tell that the model does not fit the data well. In the prediction, for none zero values, the model tends to under predict the revenue for most movies. For zero values, the model tends to over predict the revenue. Overall, I would say the model does not do a good job in both fiting and predicting.

III. Maps

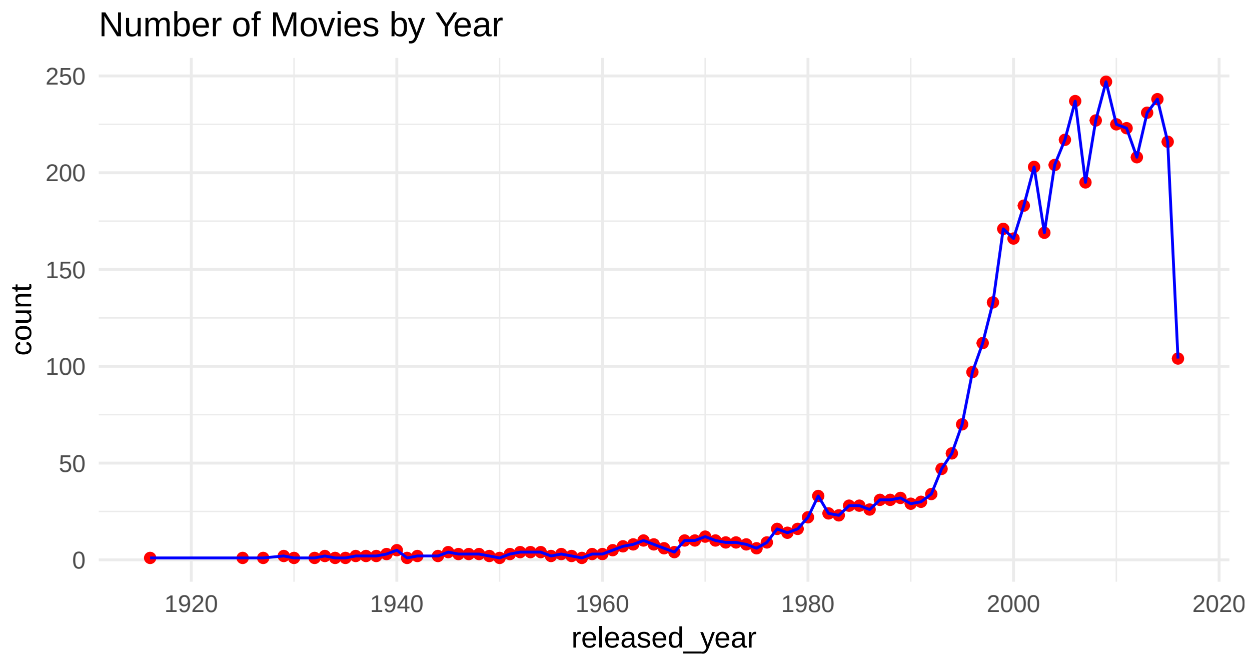
State map - extract state names from keywords of all movies

Let's see the most popular state in all movies.

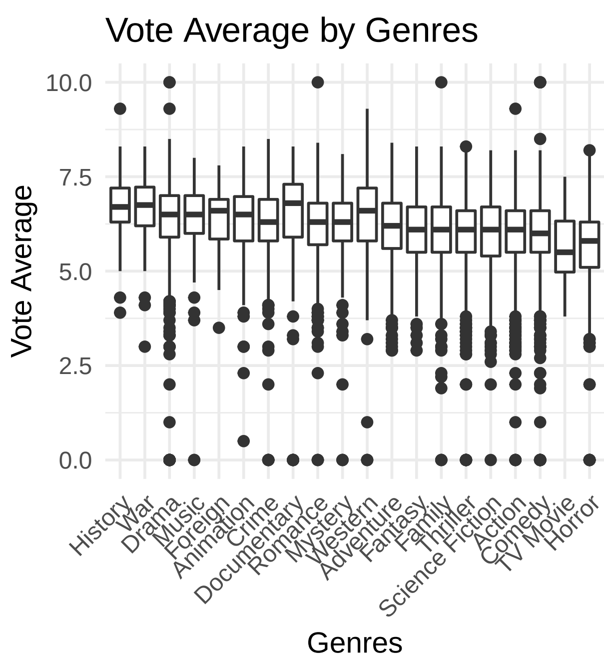

Most Popular States
50
45
40
35
30
25
lat
−120
−100
−80
long
count
80
60
40
20
0

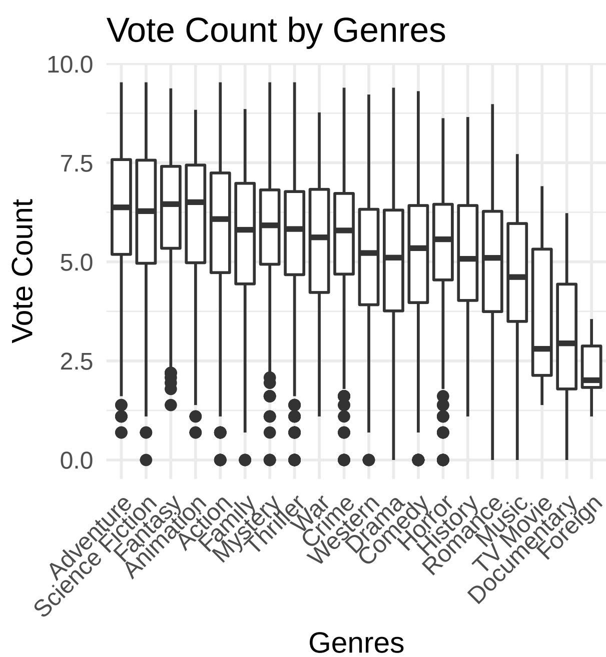
IV. General EDA

(i) Overall trend of number of movies



(ii) Vote Average and Vote Count by Genre



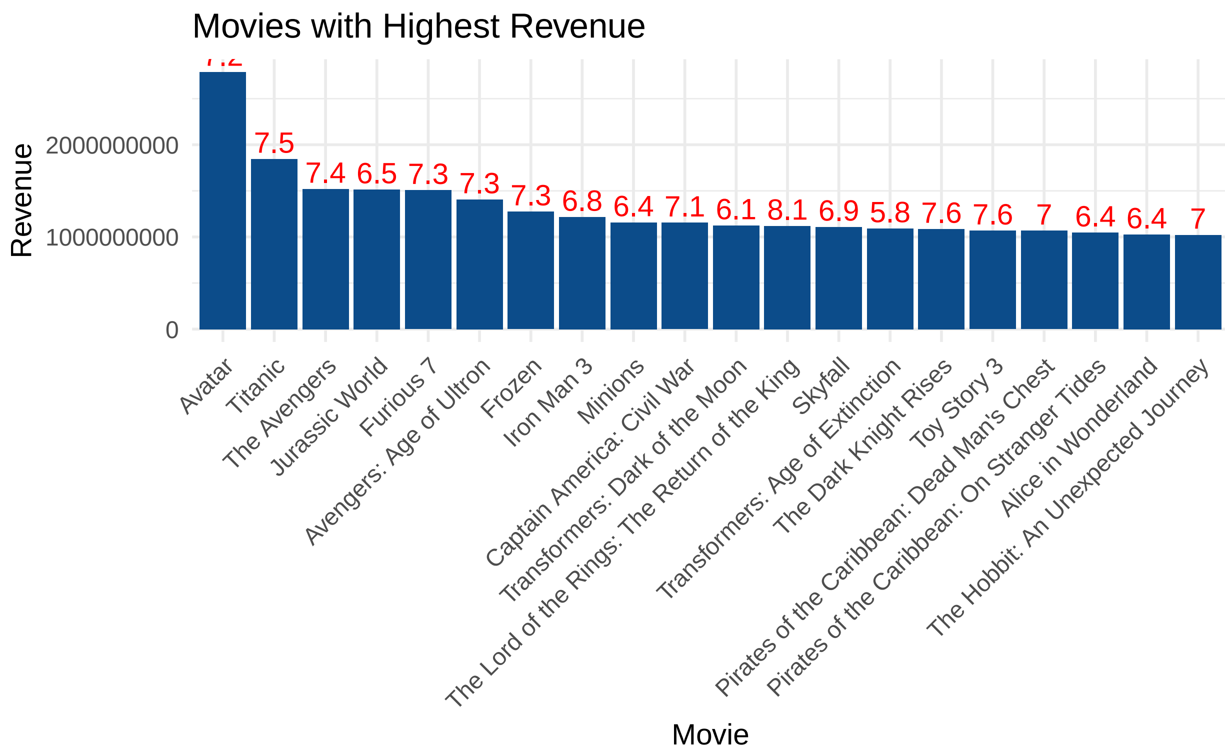


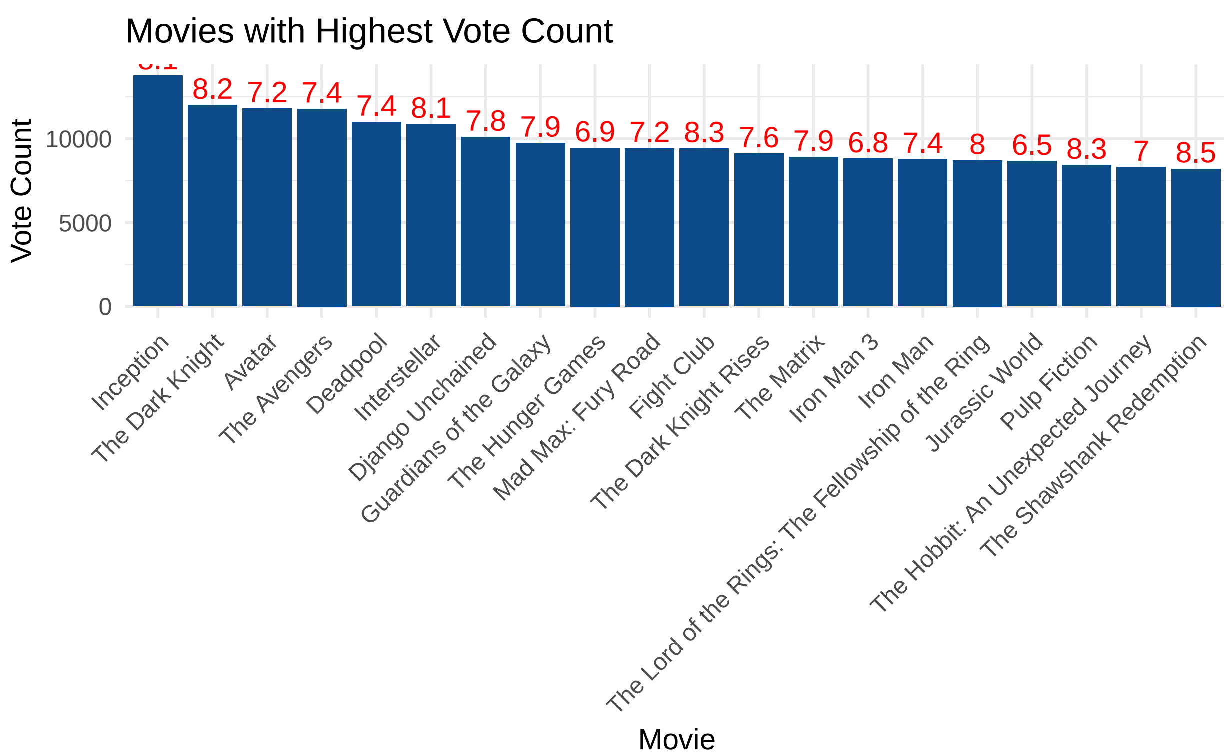
Summary:

As we can see, the medians of vote average of each genre are kind of on the same level, except for TV Movies.

In the vote count plot, foreign movies has the lowest median and quantiles comparing with other genres.

(iii) Movies with highest revenue and vote count





(iv) See more details in ShinyApp!