# 615 final project

*Xiangliang Liu*

*December 9, 2018*

## Introduction:

*Watching movie is always a good recreation for people who live under a stressful life style. However, we are constantly facing fake revenue record regarding to the newly released movie. This project will mainly focus on exploring some fundamental information about 5000 movies in the dataset called "The Movie Data Base". Specifically, we will conduct Benford analysis on Budget, revenue and popularity variables in the dataset. Then we will find out which variable does not follow Benford distribution and the reason behind it.*

## Data visualization and EDA:

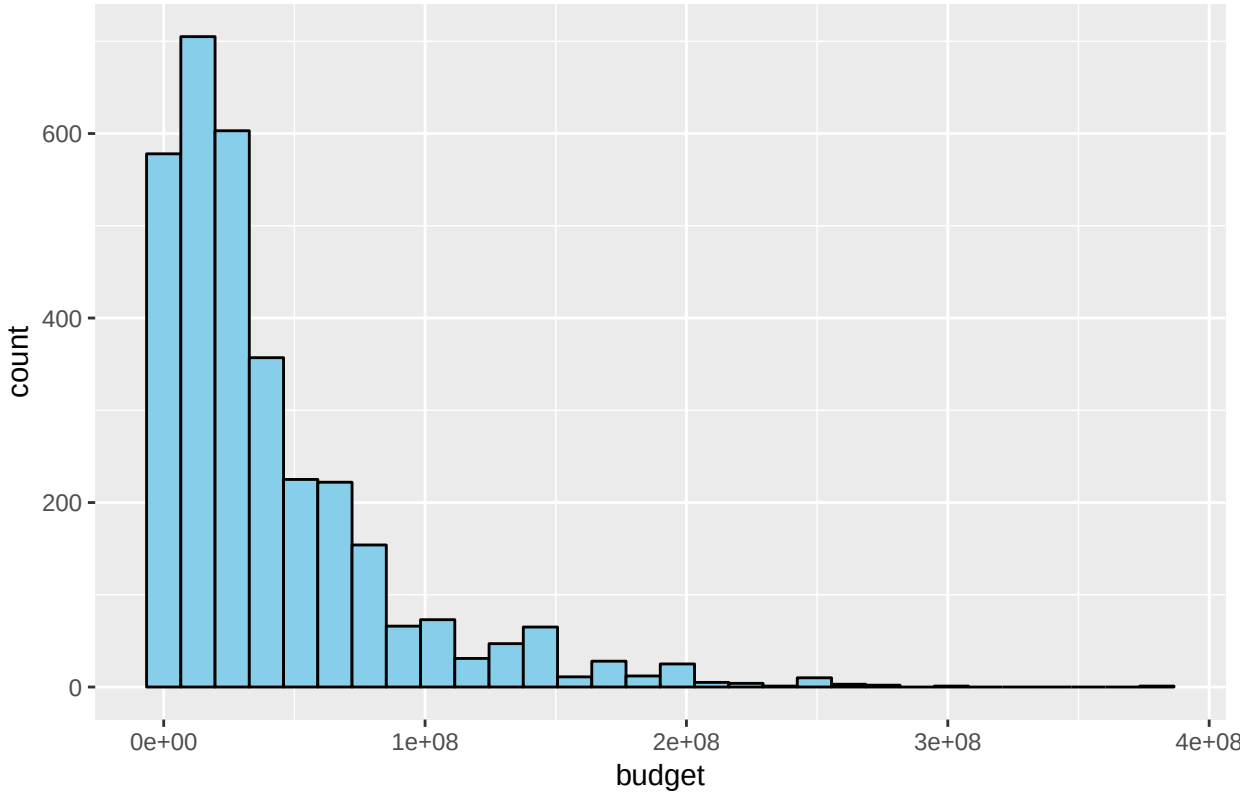
```
##      budget             genres                id           popularity
##  Min.   :        1   Length:3229        Min.   :     5   Min.   :  0.02
##  1st Qu.: 10500000   Class :character   1st Qu.:  4958   1st Qu.: 10.45
##  Median : 25000000   Mode  :character   Median : 11451   Median : 20.41
##  Mean   : 40654445                      Mean   : 44781   Mean   : 29.03
##  3rd Qu.: 55000000                      3rd Qu.: 45272   3rd Qu.: 37.34
##  Max.   :380000000                      Max.   :417859   Max.   :875.58
##     title           production_companies production_countries
##  Length:3229        Length:3229          Length:3229
##  Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character     Mode  :character
##
##
##
##     revenue            vote_average     vote_count
##  Min.   :5.000e+00   Min.   :0.000   Min.   :    0.0
##  1st Qu.:1.700e+07   1st Qu.:5.800   1st Qu.:  178.0
##  Median :5.518e+07   Median :6.300   Median :  471.0
##  Mean   :1.212e+08   Mean   :6.309   Mean   :  977.3
##  3rd Qu.:1.463e+08   3rd Qu.:6.900   3rd Qu.: 1148.0
##  Max.   :2.788e+09   Max.   :8.500   Max.   :13752.0
```

### EDA

**Check the distribution plot with variable budget, revenue, popularity and number of counts.**

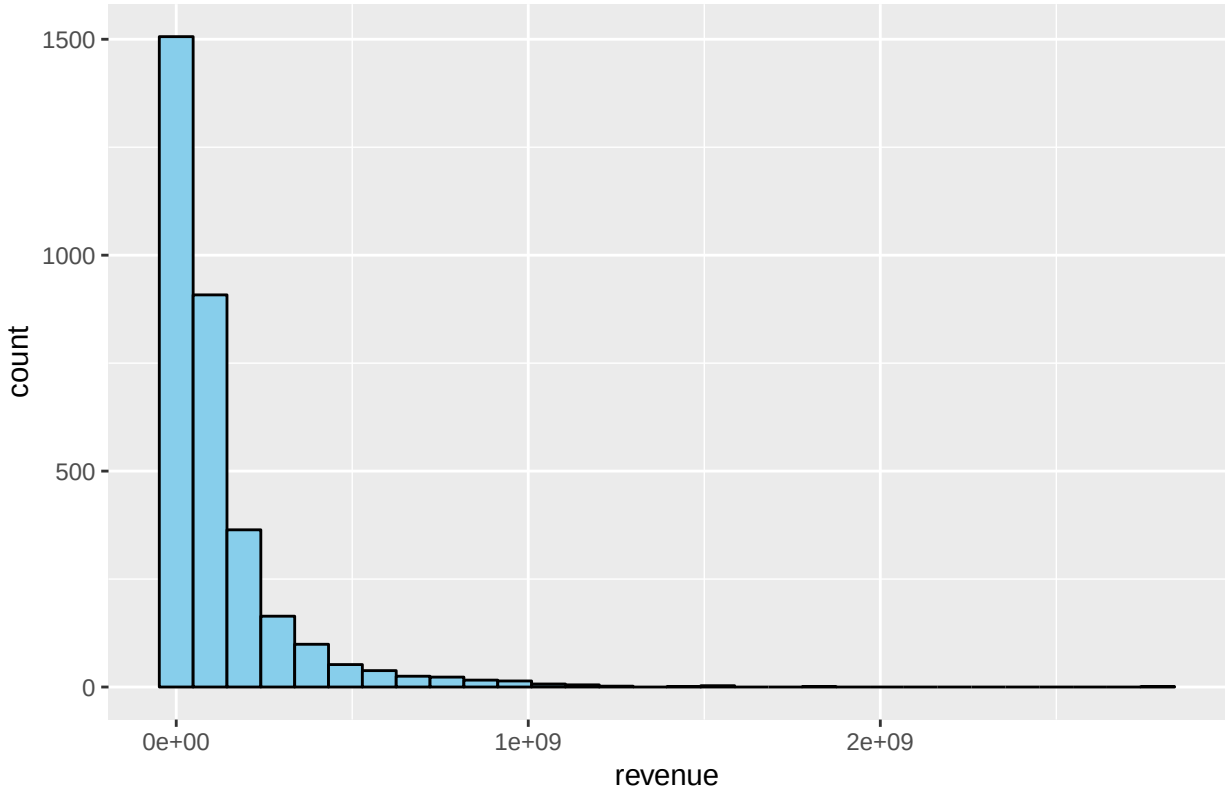
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Figure 1. The distribution of budget



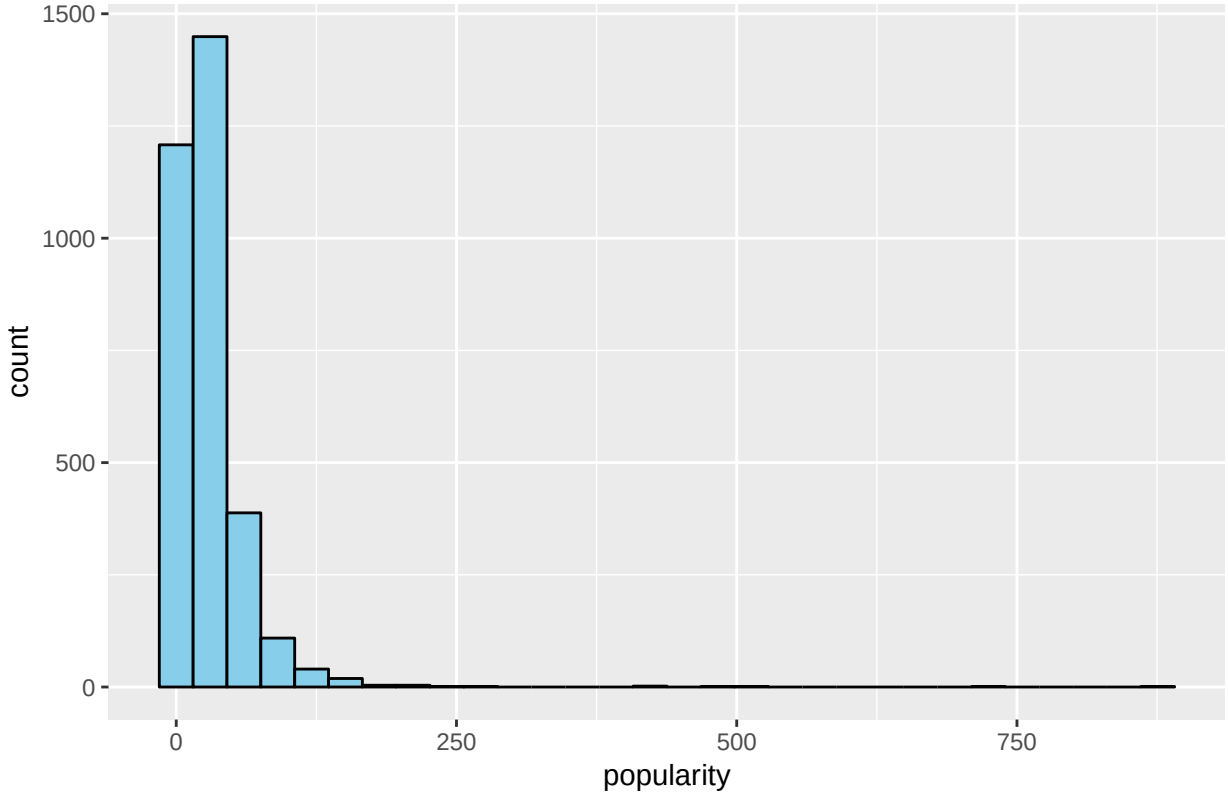

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Figure 2.The distribution of revenue



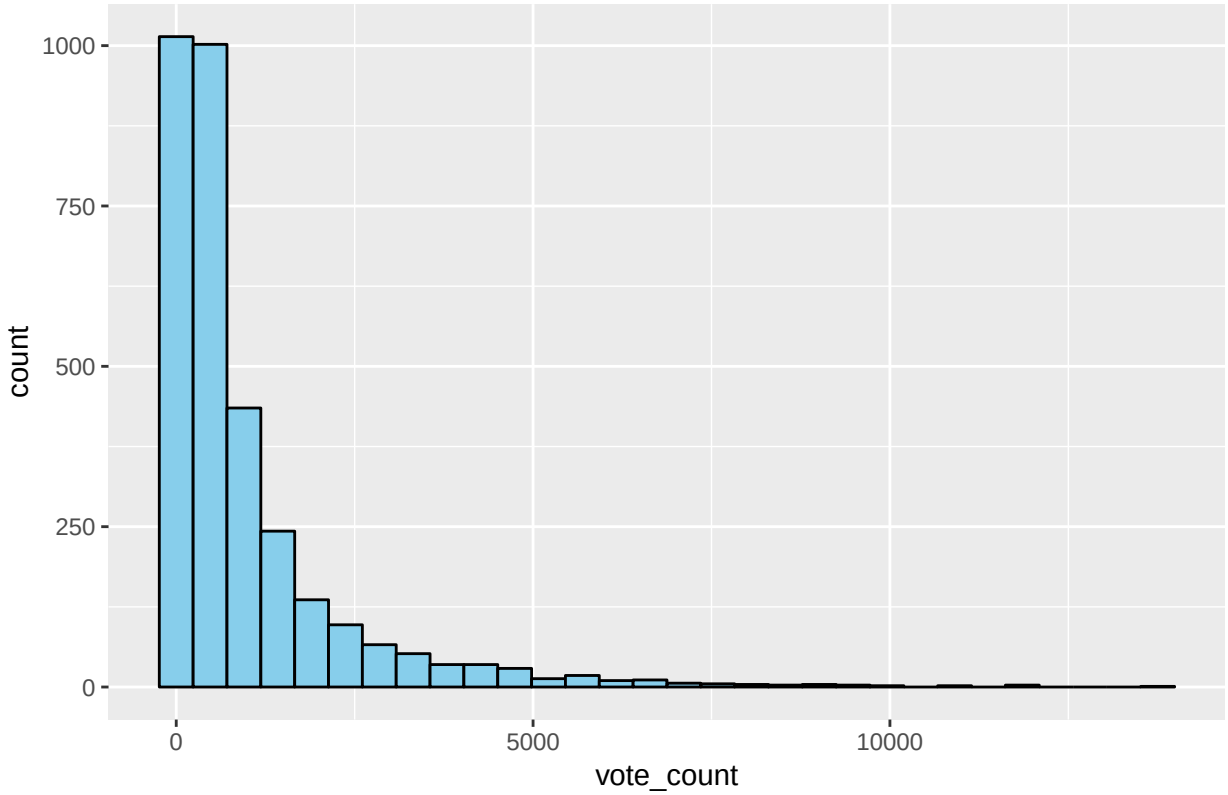

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Figure 3.The distribution of popularity




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

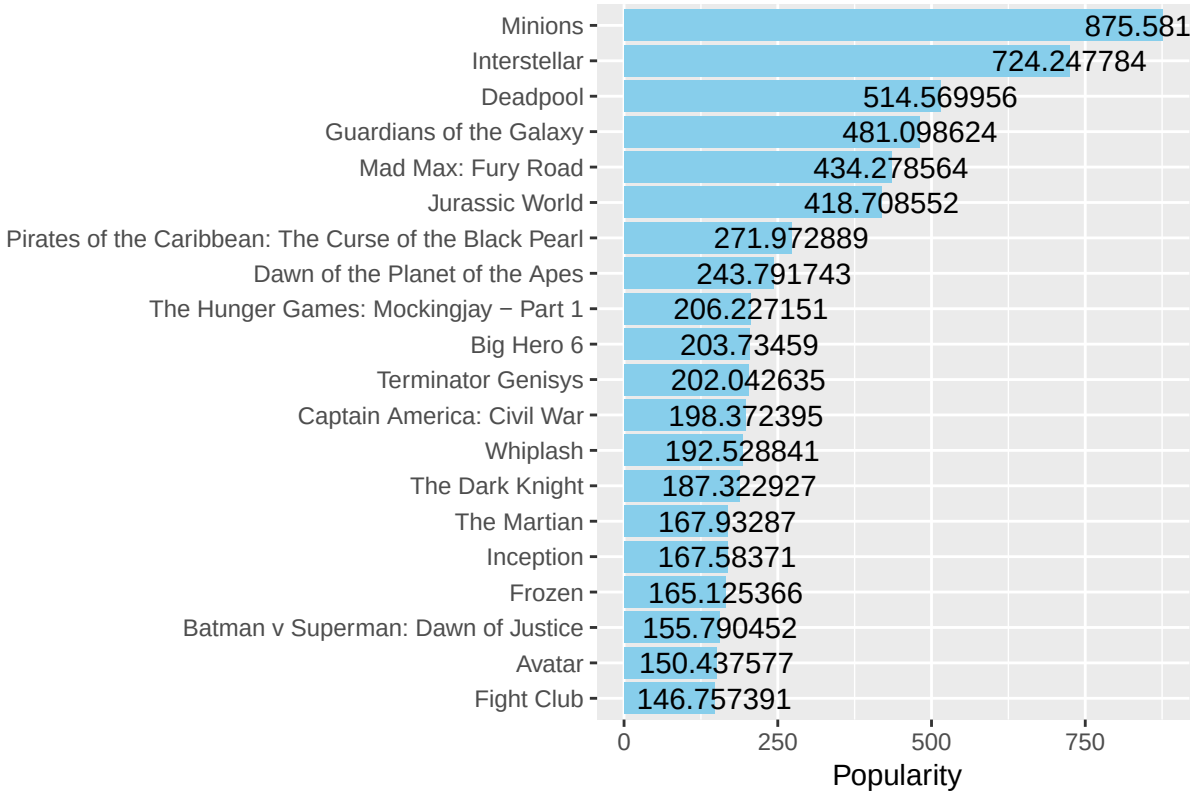
Figure 4.The distribution of number of vote




**Most popular movies in the dataset**

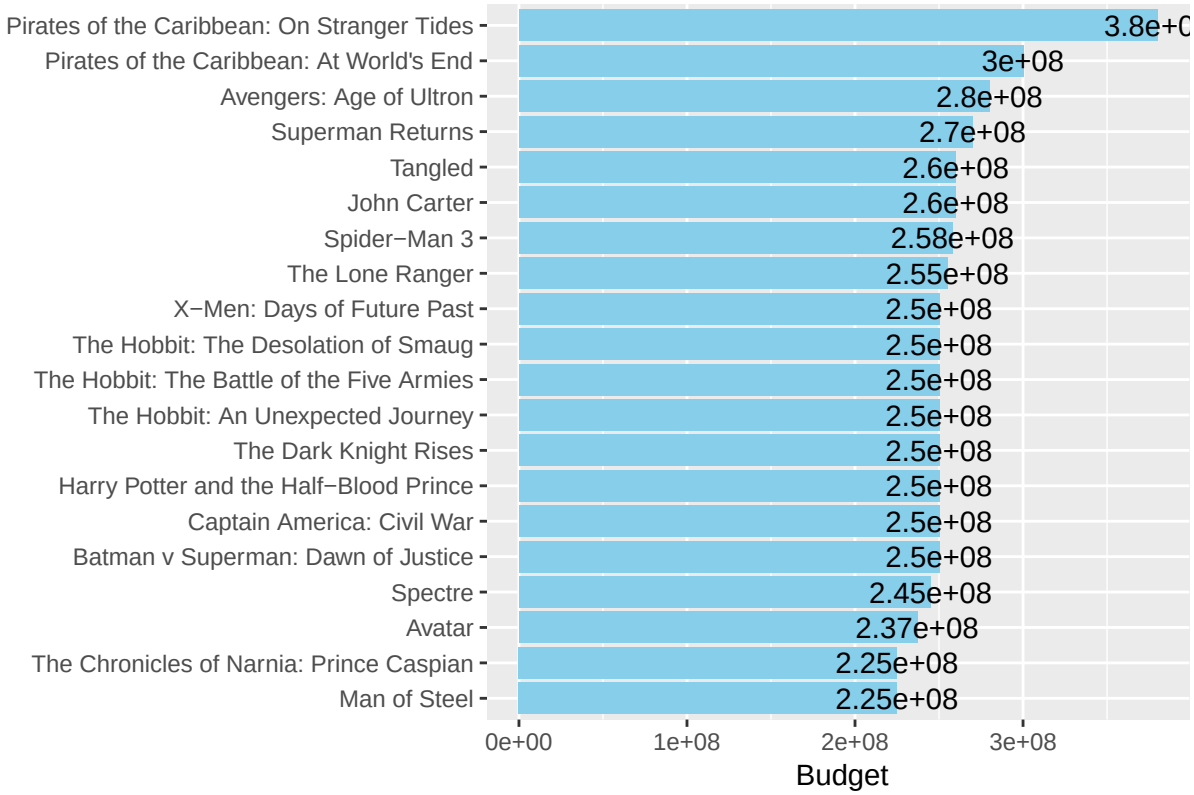
```
#Top 20 most popular movie
pop<-movie%>%
  select(title, popularity)%>%
  arrange(desc(popularity))
#chose the top 20 rows of the pop dataset
pop <- pop[c(0:20),]

ggplot(data = pop, aes(x=reorder(title, popularity ), y=popularity)) +
        geom_col(fill="sky blue") +
        coord_flip() +
        labs(x = "", y=" Popularity")+
        geom_text(aes(label = popularity))+
        ggtitle("Figure 5.Top 20 most popular movie")
```

Figure 5.Top 20 most popular movie



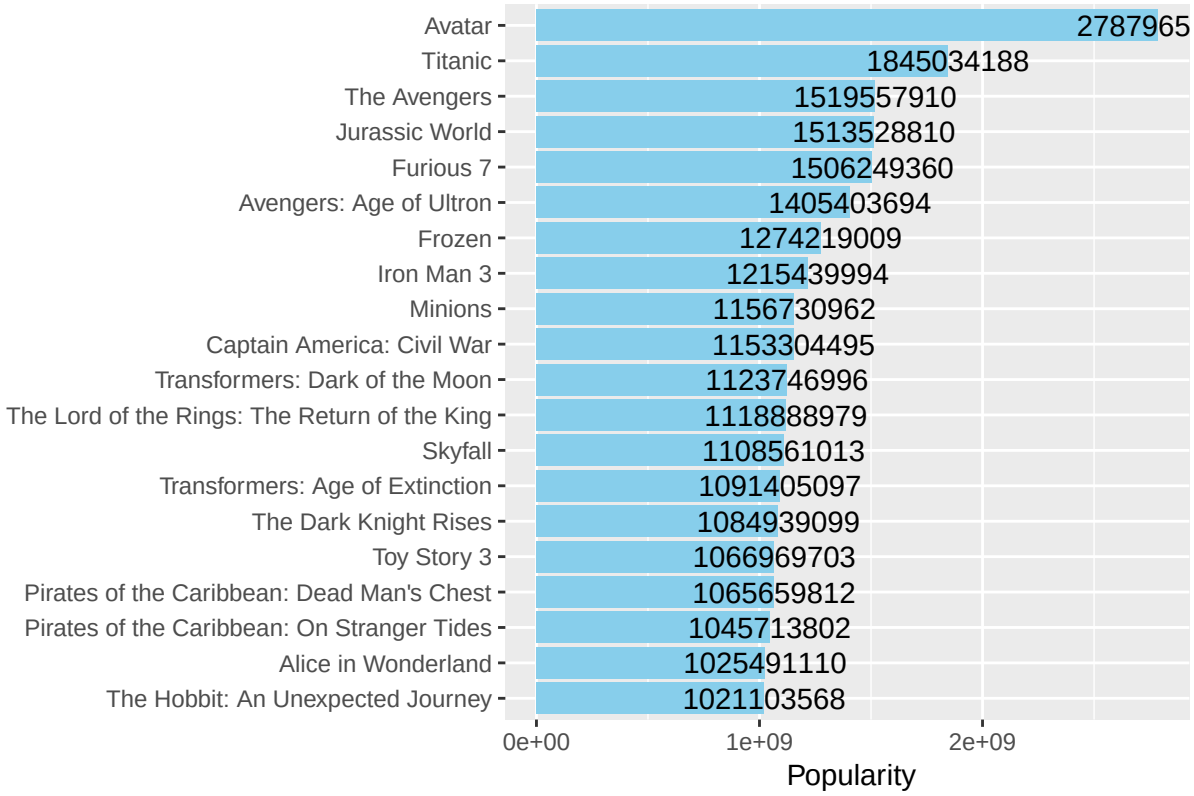

**Most expensive movies in the dataset**

Figure 6.Top 20 most expensive movie




**Most lucrative movies in the dataset**

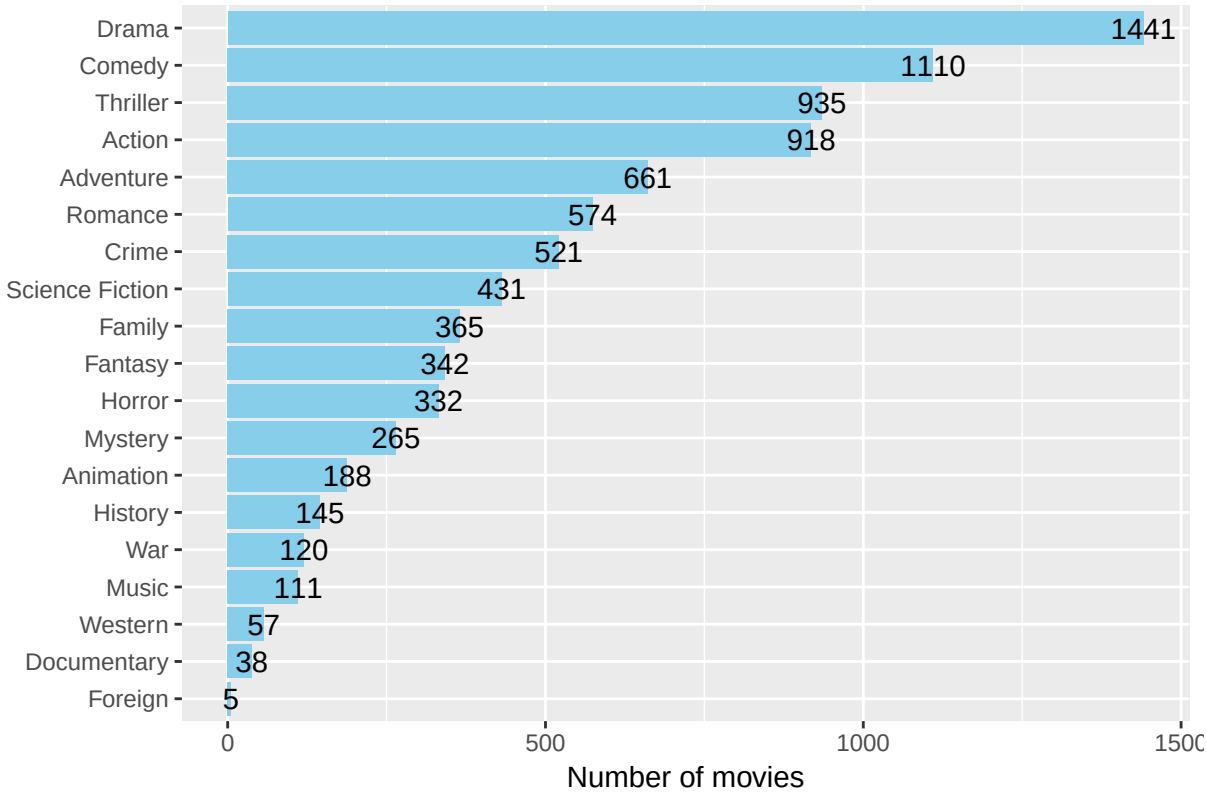
Figure 7.Top 20 most lucrative movie




*The plot showing above indicates the top 20 most popular movie in the movie data base. The movie Minions is the movie with biggest popularity*

**Create a genre wordcloud**

Figure 8. Number of movies in different genres

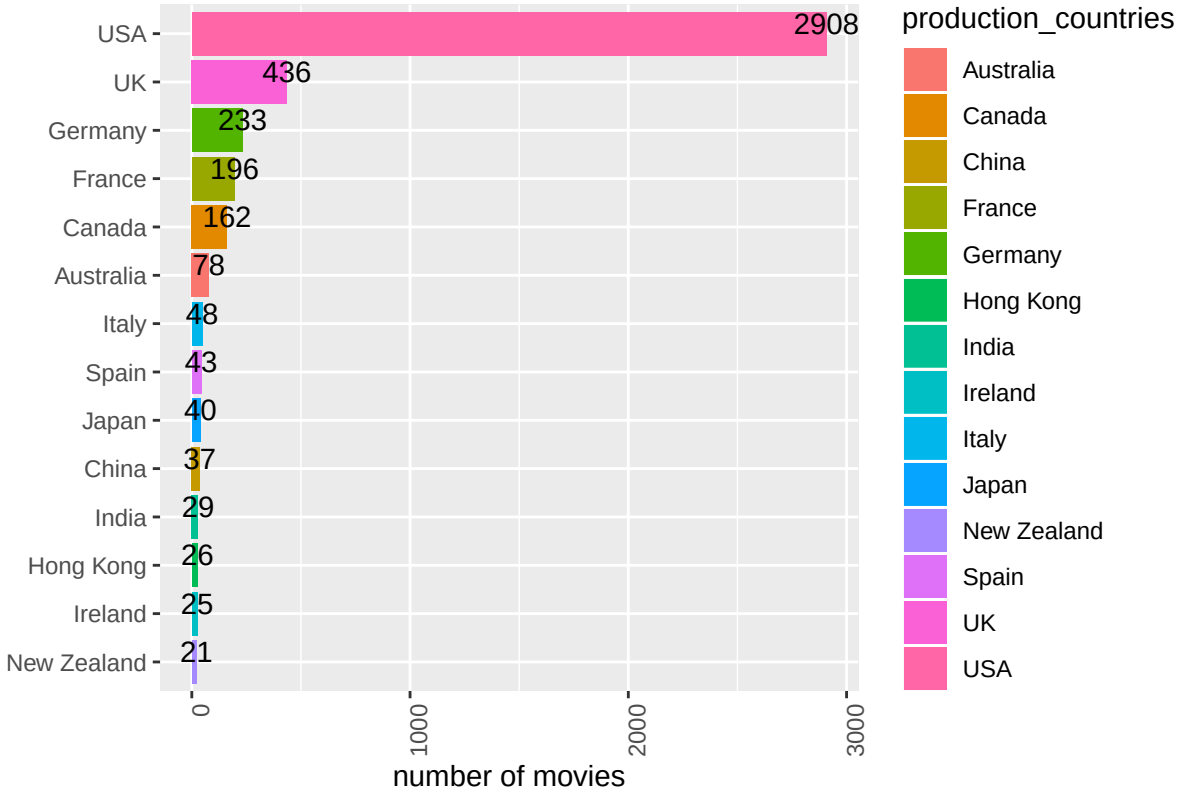


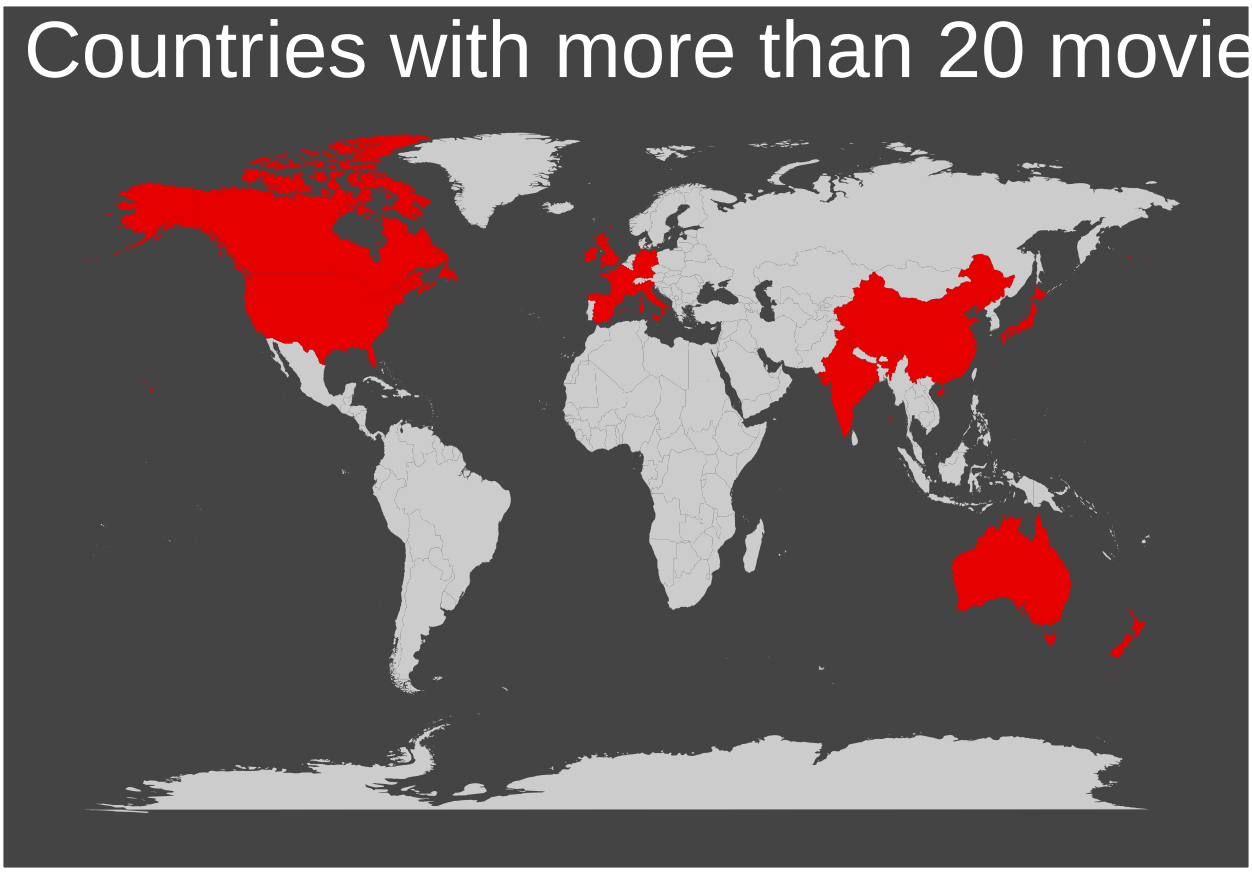

*The wordclouds gives us a general idea about which genres have the largest proportion. It turns out that Drama, Comedy and Thriller are the top 3 genres with highest number of movies*

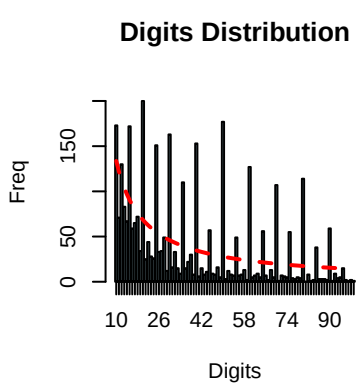
**mapping with production countries**

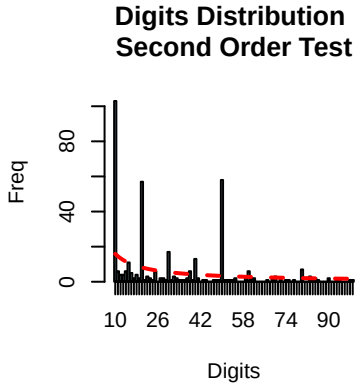
```
##
## Attaching package: 'maps'
## The following object is masked from 'package:purrr':
##
##     map
```

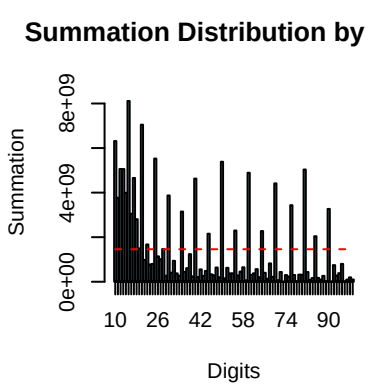
Figure 9.the distribution of movies in different countries

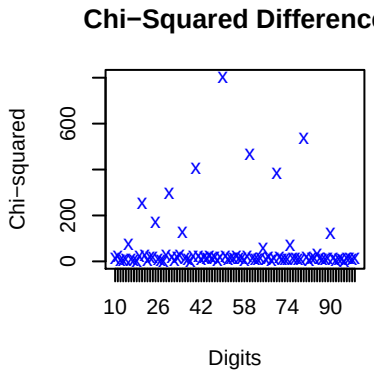


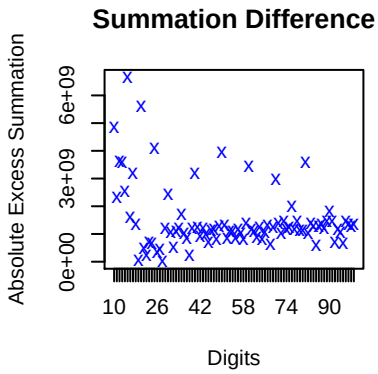

Countries with more than 20 movie

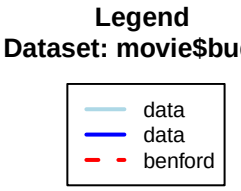


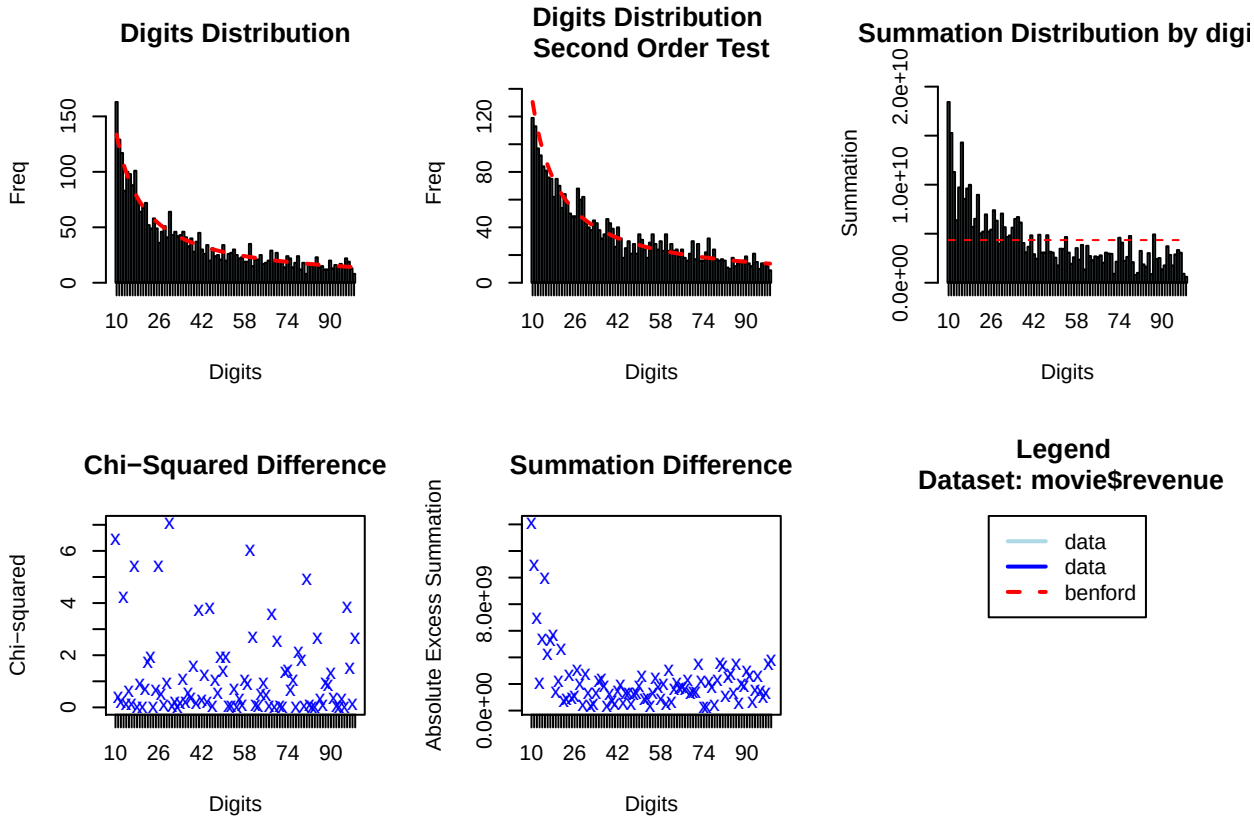




##Bnford analysis:

```
##
## Benford object:
##
## Data: movie$budget
## Number of observations used = 3229
## Number of obs. for second order = 384
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.481
##          Var  0.081
##  Ex.Kurtosis -1.174
##     Skewness -0.015
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     50        149.23
## 2     20        131.58
## 3     40        118.37
## 4     30        117.02
## 5     60        103.82
##
```

```
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  movie$budget
## X-squared = 4686, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  movie$budget
## L2 = 0.00048393, df = 2, p-value = 0.2096
##
## Mean Absolute Deviation: 0.008913238
## Distortion Factor: -4.976945
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

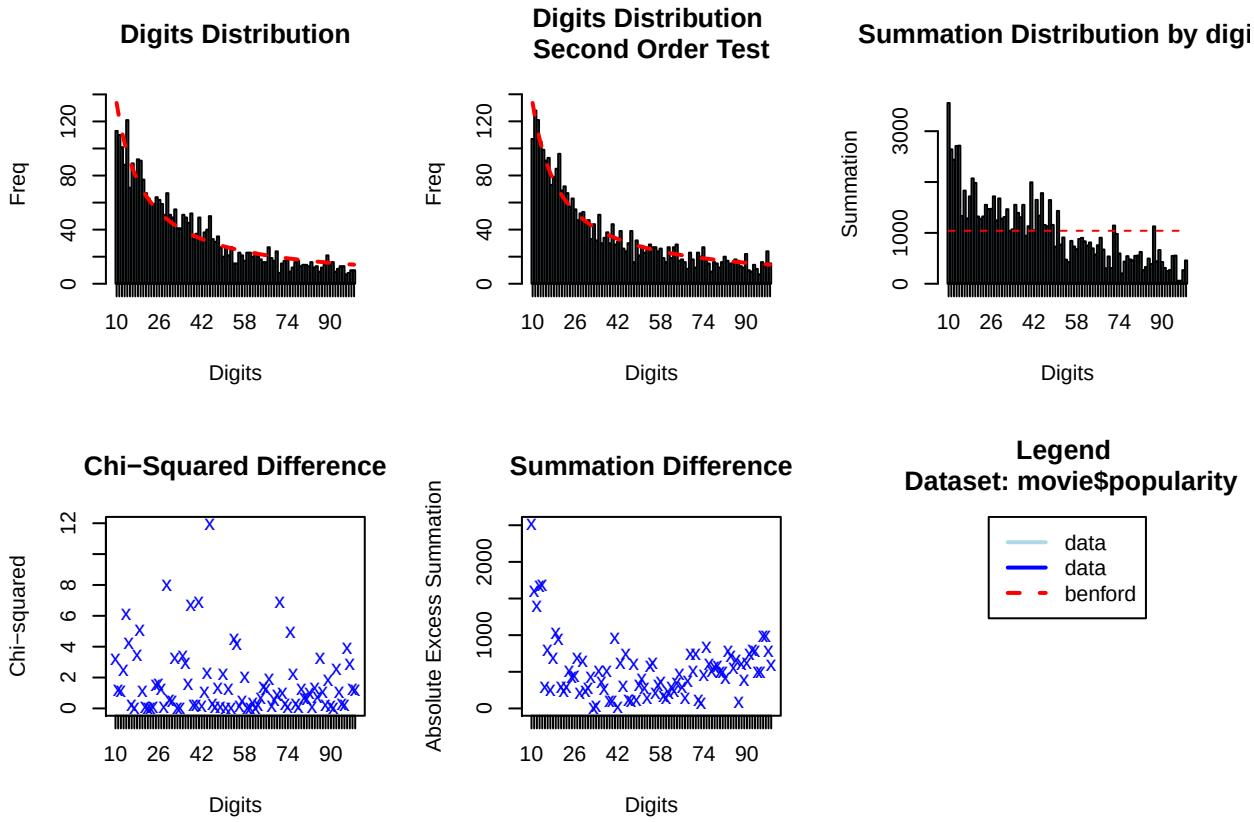
*as we can see from the distrbution and summary of the Benfore test, the budget data in movie dose not follow Benford distribution. A lot of lines exceed the red line threshold. Also, the p-value in the summary is less than 0.05, which reject the hypothesis that the distribution follow Benford distribution.*




```
##
## Benford object:
##
## Data: movie$revenue
## Number of observations used = 3229
```

```
## Number of obs. for second order = 3153
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.493
##          Var  0.086
##  Ex.Kurtosis -1.214
##     Skewness  0.015
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     10         29.34
## 2     13         20.92
## 3     17         20.84
## 4     30         18.02
## 5     26         16.92
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  movie$revenue
## X-squared = 105.71, df = 89, p-value = 0.1092
##
##
##  Mantissa Arc Test
##
## data:  movie$revenue
## L2 = 0.00038072, df = 2, p-value = 0.2925
##
## Mean Absolute Deviation: 0.00153106
## Distortion Factor: -0.9888538
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

*The Benford test on revenue indicates p-value is 0.2925 > 0.05, so do not reject the hypothesis that the distribution follow Benford distribution. This is result is what*




```
##
## Benford object:
##
## Data: movie$popularity
## Number of observations used = 3229
## Number of obs. for second order = 3227
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.488
##          Var  0.074
##  Ex.Kurtosis -1.053
##     Skewness  0.052
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     14         24.25
## 2     10         20.66
## 3     15         19.50
## 4     29         19.46
## 5     45         19.18
##
```

```
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  movie$popularity
## X-squared = 145.8, df = 89, p-value = 0.0001389
##
##
##  Mantissa Arc Test
##
## data:  movie$popularity
## L2 = 0.0086403, df = 2, p-value = 7.645e-13
##
## Mean Absolute Deviation: 0.001861221
## Distortion Factor: -20.41978
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.02   10.45   20.41   29.03   37.34  875.58
```

*As for the variable popularity, the p-values from Chi-squared test is 0.0001389, which is less than 0.05. So we do reject the hypothesis that the distribution follow Benford distribution*

## Conclusion:

### insights and findings:

*We found the budget numbers do not significantly follow Benford analysis. The budget that start with 50 and 20 have the highest deviation. This result does make sense, because when people decide to make a movie or approve a movie, they tends to give a rough number about how much money they will spend on this movie. They never specific the budget to a unit digit. For instant, the number could be 500 million or 200 million dollars. So the result of Benford analysis on variable budget is what we expected.*

*From the Benford analysis on variable revenue, we cannot reject the hypothesis that the distribution follow Benford distribution. However, though the number seems right, there's no evidence to draw the conclusion that there's no fraud in this variable. we may want to do more research on those movies.*

### limitation:

*There are a few limitations about the Benford analysis. We can only test whether the data follow Benford distribution. After that, even if we know the data does not follow the distribution, we still need to do more research on the data to explore whether there are some frauds in the data. This limitation is also intepreted in the summary: "Real data will never conform perfectly to Benford's Law. You should not focus on p-values!"*

## Acknowledge:

*Special thanks to professor Wright who give me advice to choose dataset. Thanks for his consistency in teaching data cleaning, data manipulation and all kinds of data visualization methods.*

## Reference:

https://www.kaggle.com/tmdb/tmdb-movie-metadata