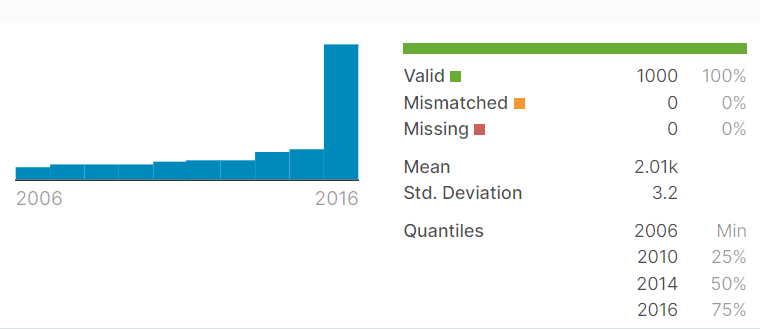
***Data set:***

We use the data from IMDb. There are about 1000 movies initially from the year 2006 - 2016. The attributes include the following:

Movie tittle, Genre, Director, Actors, Year, Running time, Rating, Votes, Revenue, Metascore.

The number of released movies by year (*Figure 1*)



(*Figure 1*)

***Data cleaning:***

To obtain valid database. We did data cleaning as following:

-- Delete invalid data.

-- Fill / Delete empty data

After data cleaning, we have the clean data set with 839 movies.

***Data Analyzing:***

The linear regression line can be presented as: *Y = a + bX* where X is the explanatory variable and Y is the dependent variable. Furthermore, the model for multiple linear regression can be presented as

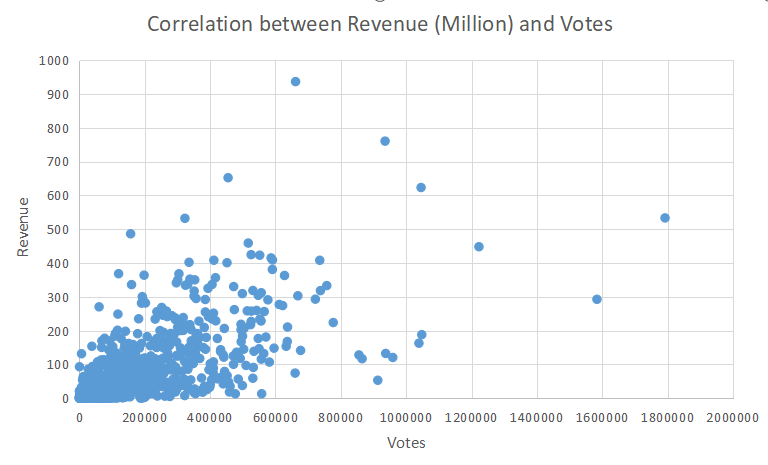
*yi*​=*β*0 + *β*1*xi*1 + *β*2*xi*2 +...+ *βpxip* + *ϵ*

where, for *i*=*n* observations: *yi* = dependent variable, *xi* = expanatory variables, *β*0 = y-intercept (constant term), *βp* =slope coefficients for each explanatory variable and

*ϵ*=the model’s error term (also known as the residuals)​

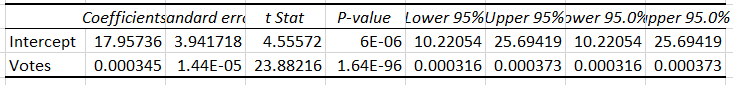
We will investigate the correlation between Revenue (movie success) and attributes (such as scoring and voting) to see which attributes may strongly influence the Revenue (movie success). We will apply Ordinary least squares (OLS) regression to verify our work.

For now, we first see the effect of Votes on revenue (*Figure 2*)



(*Figure 2)*

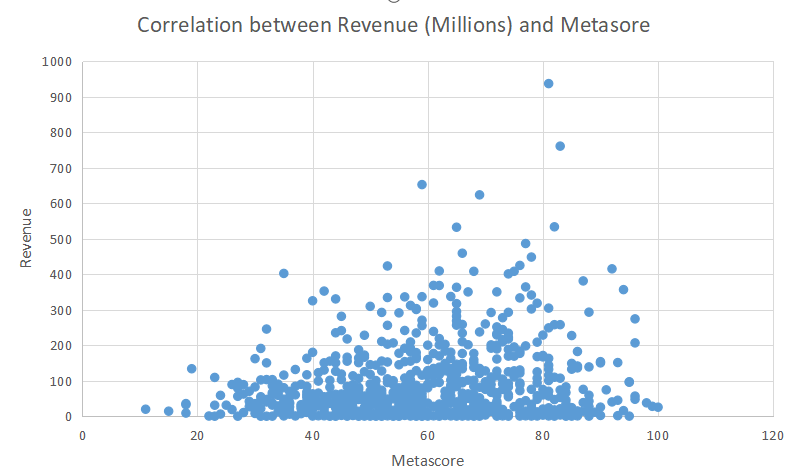
To investigate whether the number of votes has strong influence on the movie’s revenue. We use the OLS regression table (Figure 3) to check it.



*(Figure 3)*

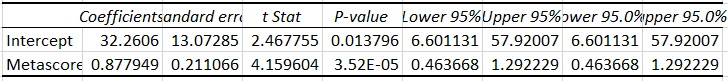
As we can see, the p-value is very small. Therefore, we know Revenue and the number of Votes are independent and the number of votes has strong influence on the movie’s revenue.

Secondly, the effect of Metascore on Revenue.



*(Figure 4)*

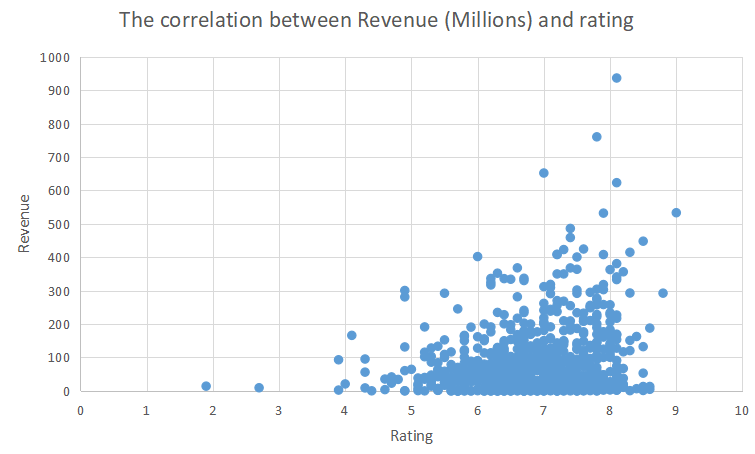
We still apply OLS regression table (Figure 5) on it.



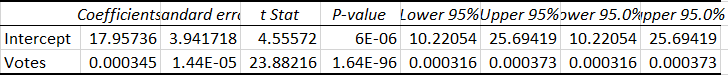
*(Figure 5)*

The p-value is very small and Revenue and Metascore are strongly correlated.

We apply the same strategy on the correlation between Revenue and Rating (Figure 6) and the OLS regression table (Figure 7).



*(Figure 6)*



*(Figure 7)*

Rating is strongly effect movie’s revenue as the p-value is still very low.

In conclusion, Rating, the number of votes and Metascore play important roles in movie success. We can easily find if a movie will be success by calculating above factors.

In the future, we will focus on data classification, data calculation and test our method to some upcoming movies.

***Reference:***

1,<https://smallbusiness.chron.com/make-regression-table-excel-30870.html>

2,<https://www.dummies.com/education/math/statistics/what-a-p-value-tells-you-about-statistical-data/>

3,<https://machinelearningmastery.com/linear-regression-for-machine-learning/>