

Assignment Report

Findings on Website Data &

Extended Research on Twitter data

Zhengxing Li | D00171597 | 27/03/2017 | Data Analysis 1

Table of Contents

[Introduction 2](#_Toc481138653)

[Missing Data & Outliers 2](#_Toc481138654)

[1. Missing Data 2](#_Toc481138655)

[2. Outliers 3](#_Toc481138656)

[Descriptive Statistics 4](#_Toc481138657)

[1. Type of Website 4](#_Toc481138658)

[2. The Number of Live Days 5](#_Toc481138659)

[3. Down Time 6](#_Toc481138660)

[4. The Number of Hits 6](#_Toc481138661)

[5. Website Still Alive 7](#_Toc481138662)

[6. The Number of Sales 8](#_Toc481138663)

[7. Average Sales Value 8](#_Toc481138664)

[8. Average User Age 9](#_Toc481138665)

[9. Usability Rating 10](#_Toc481138666)

[Correlation & Linear Regression 11](#_Toc481138667)

[1. The Number of Live Days & The Number of Hits 11](#_Toc481138668)

[2. The Number of Live Days & The Number of Sales 12](#_Toc481138669)

[3. The Number of Sales & Average Sales Value 13](#_Toc481138670)

[Confidence Interval 14](#_Toc481138671)

[Confidence Interval 1 14](#_Toc481138672)

[Confidence Interval 2 14](#_Toc481138673)

[Confidence Interval 3 14](#_Toc481138674)

[Hypothesis Testing 15](#_Toc481138675)

[Hypothesis Test 1 15](#_Toc481138676)

[Hypothesis Test 2 15](#_Toc481138677)

[Hypothesis Test 3 16](#_Toc481138678)

[Extension of Research on Twitter Data 16](#_Toc481138679)

[1. Intro 16](#_Toc481138680)

[2. Obtain & Prepare Data 16](#_Toc481138681)

[3. Analysis Results 18](#_Toc481138682)

[4. Confidence Interval & Hypothesis Test 19](#_Toc481138683)

# Introduction

In this report, I’m going to interpret my findings on analysis of given website data and report my results in the following several aspects:

1. Finding and handling missing data
2. Identification and treatment of outliers
3. Generating and interpreting descriptive statistics for each variable
4. Producing appropriate data visualization
5. Hypothesis testing

Furthermore, I’ll also extend and refine the research on Twitter data from last semester by adding hypothesis data to it and drawing some conclusions from the test results.

Before actually carrying out analysis on the data set, I’d like to provide some basic information of this data set first. This data set are activities of 100 websites captured in period from 31/12/2015 to 31/12/2016. The data set contains 100 records in total with some missing values and 9 fields which are type of website, day being live, down time, the number of hits, still alive or not, the number of sales within this period, average sales value, average user age and usability rating respectively.

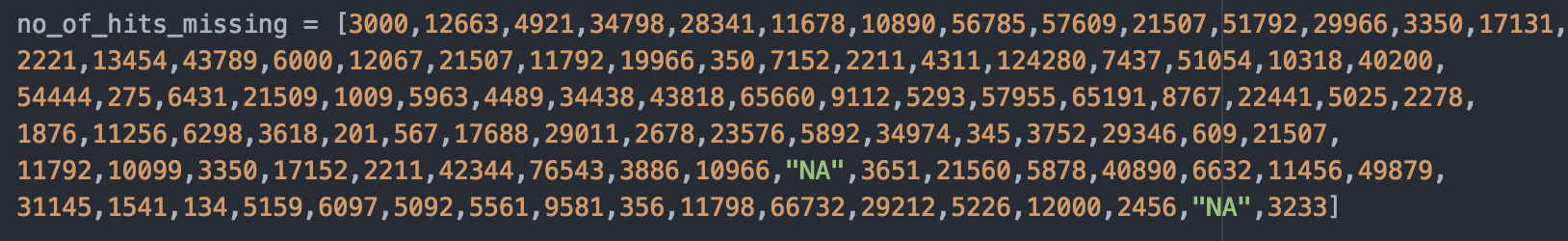
# Missing Data & Outliers

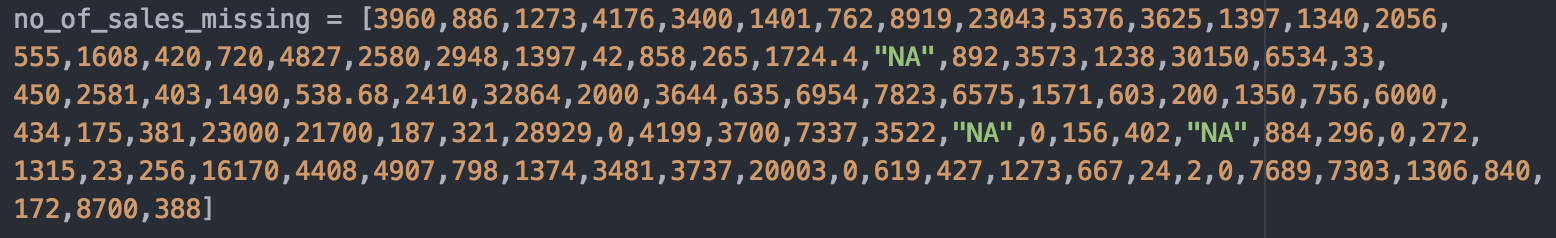
The first step of the analysis is data cleansing. In this phase, I need to identify and handle missing values and outliers to prepare data for following analysis.

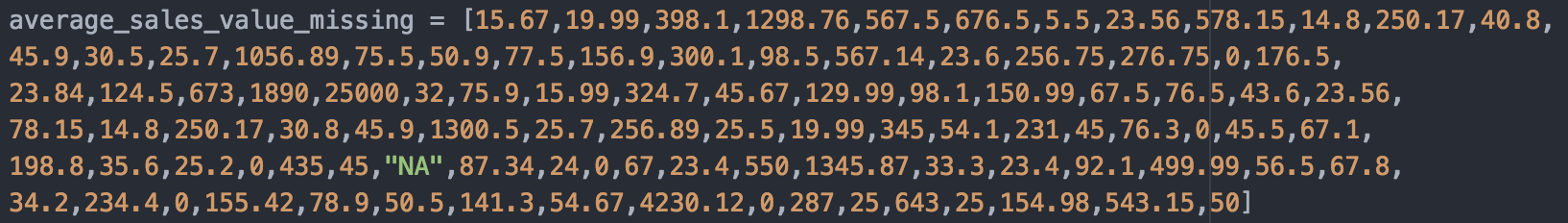
## Missing Data

Missing values or missing data are very common in statistics and it can have tremendous influence on the conclusions, so we should be really careful with them.

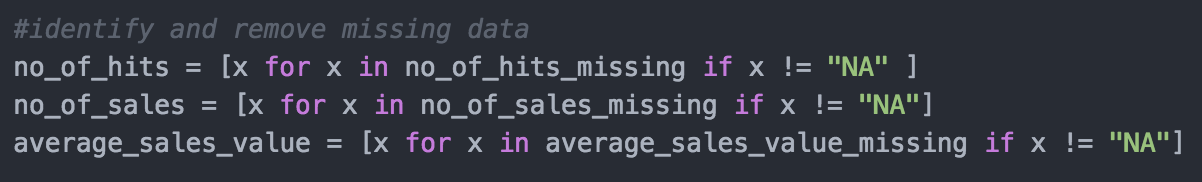
To identify the missing data, I first check whether the length of each array is 100 or not. Then I find there are several “NA” in the number of hits, the number of sales and average sales value by manually looking through the data set.







Given that there are no more than 3 missing value in the variables above, the number of missing value is relatively small compared with 100, the number of records, and I think these missing values might just have small influence on the analysis results. So, I choose to delete them by add each value if it’s not “NA” to a new list for every variable.



## Outliers

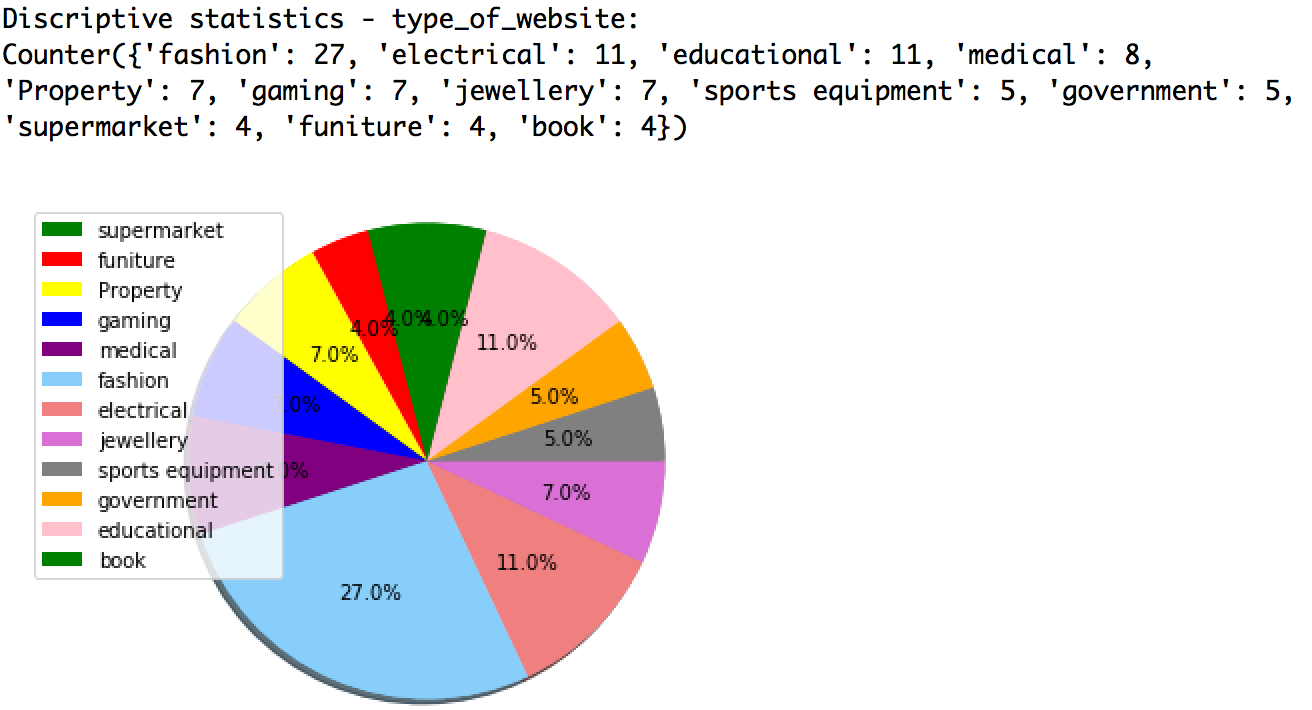
Data cleansing also includes identification and treatment of outliers, I want to know if there is any outlier in the data set. If so, I will treat the outlier properly, because it can give me different results.

To identify outliers, I chose to calculate the mild fence of each variable. Any value falls out of this range are considered as an outlier and I will provide more details on outliers of each variable in descriptive statistics part.

After identifying outliers, I need to find a proper treatment for those outliers. There are two ways in general, either simply remove it or report other statistics which are not influenced by outliers like median, mode range and etc. In this case, each variable may have various numbers of outliers. If I remove them, this will result in different lengths of lists and bring inconvenience for following analysis like producing plots, correlation and linear regression. So, I chose to keep them and report other statistics.

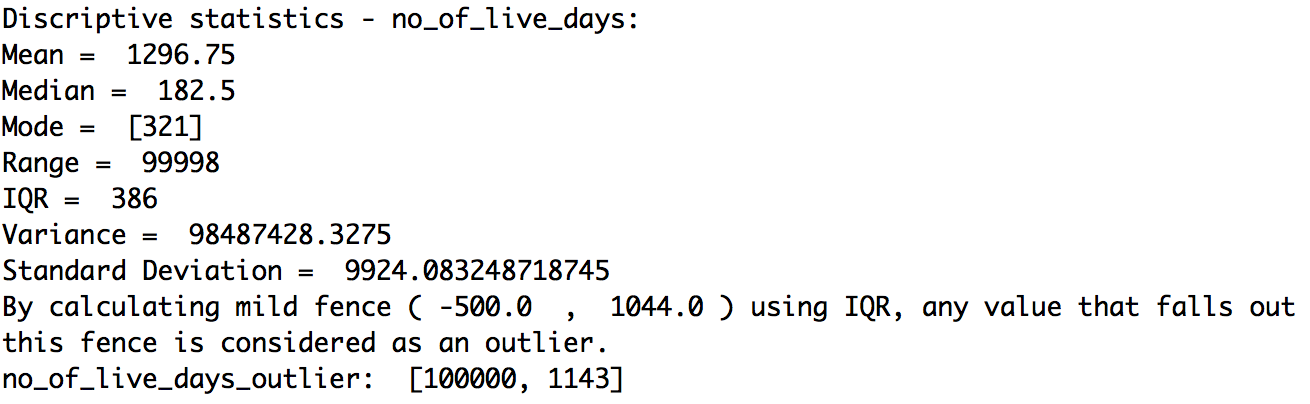
# Descriptive Statistics

## Type of Website



Because this set of data are all non-numeric data, I’m unable to generate descriptive statistics as what I did for numeric data. However, I still can tell which type of website has the largest proportion in these 100 websites. To achieve this, I made a pie chart which illustrates the proportion of each slice. The pie chart above tells me 27% of these 100 websites are fashion websites which occupies the largest proportion.

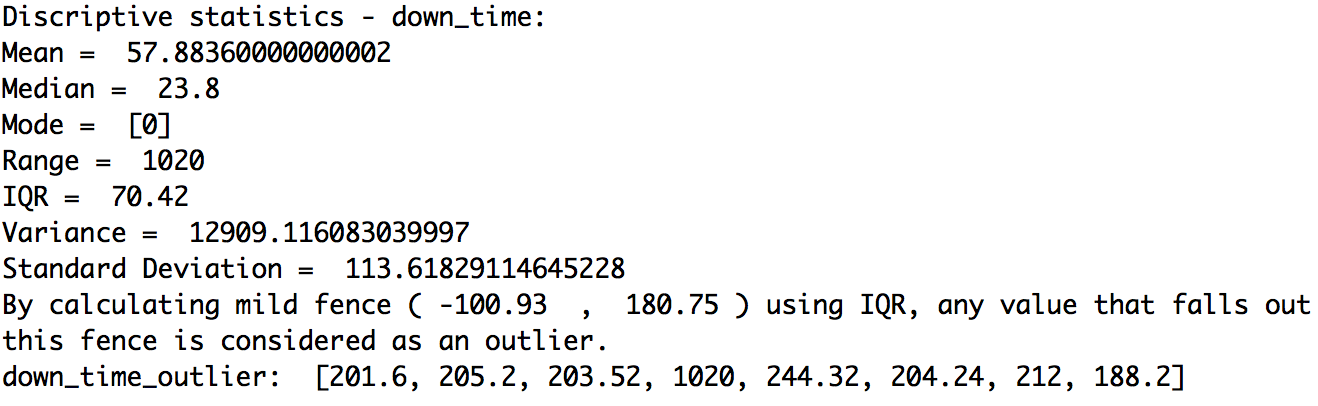
## The Number of Live Days



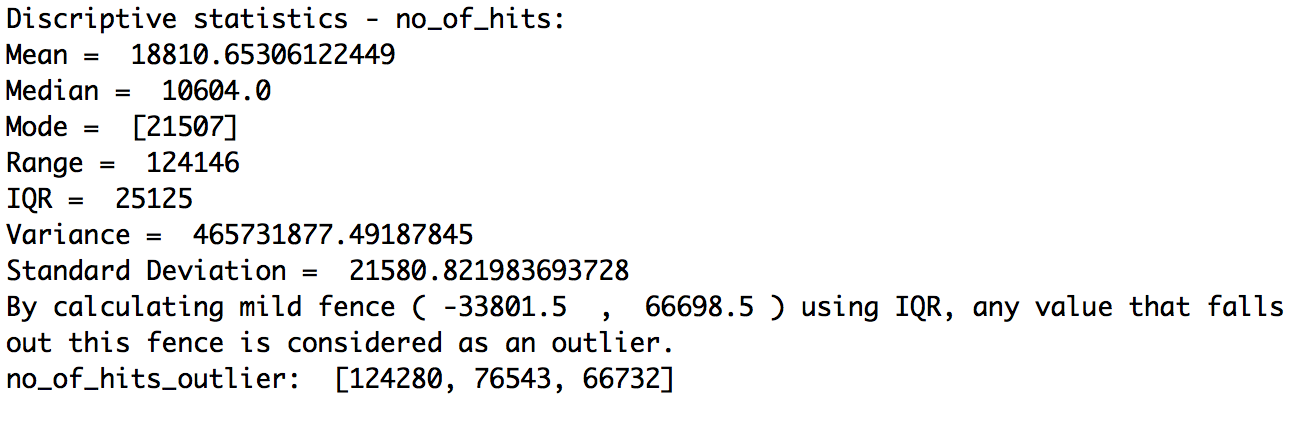
Range is the difference between the largest and the smallest values and it’s highly influenced by extreme values/outliers. The range (99998) above indicates this set of values have large fluctuation because of outliers. This may be because many websites have gone live for a long time before this period of time. By calculating mild fence ( -500.0 , 1044.0 ) using IQR, any value that falls out this fence is considered as an outlier. Mean, range, variance and standard deviation are largely affected by outliers and become imprecise for representing this set of data.

In this case, I’d like to choose the median (182.5) to represent the middle level of this set of values. Because according to the IQR (Interquartile Range) (386) which indicates the range of the middle 50% values of a data set, the median is located at the most middle place in this range and IQR are much smaller than the range, which mean the middle 50% of the number of live days of these 100 websites are really close and they are usually less than a year.

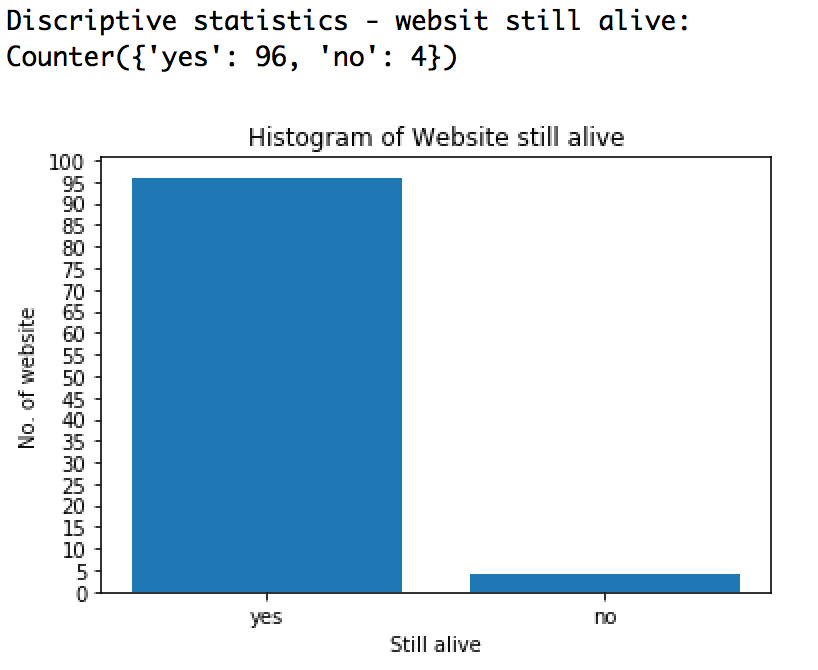
## Down Time



## The Number of Hits

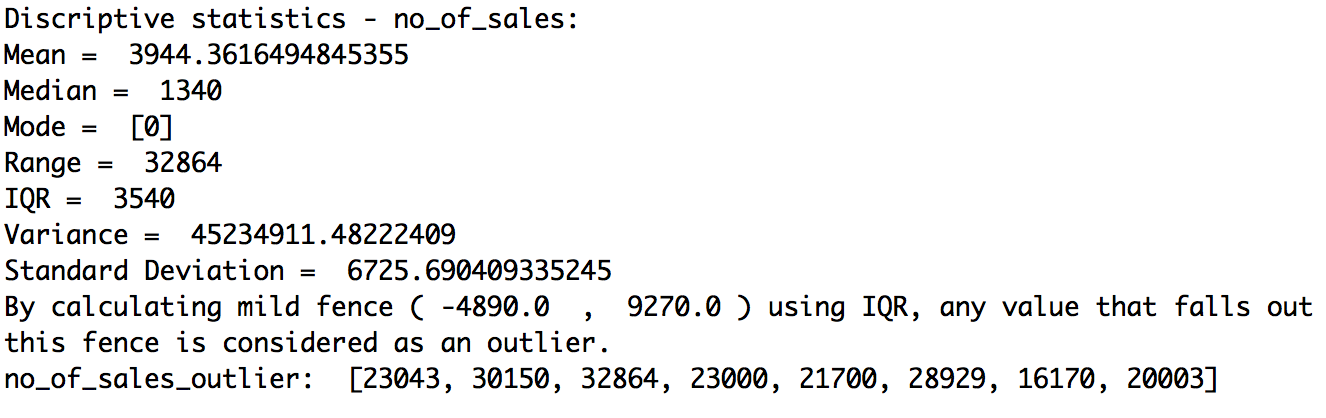


## Website Still Alive

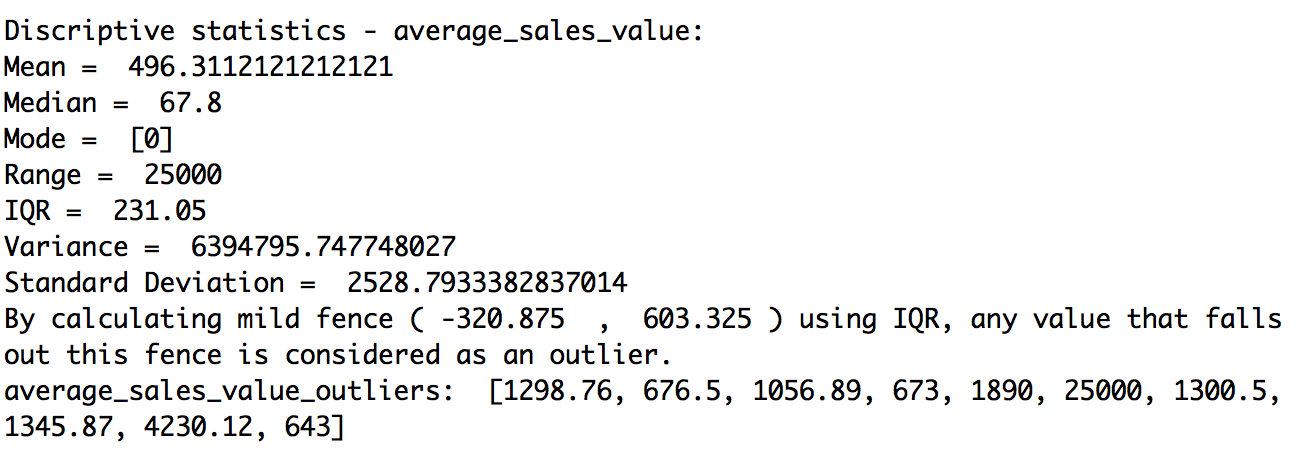


This set of data is also non-numeric. So, I produced a histogram which clearly shows me the number of websites alive and the number of those going offline. There 96 websites are still alive on 31/12/2016, while 4 websites went offline within this period.

## The Number of Sales

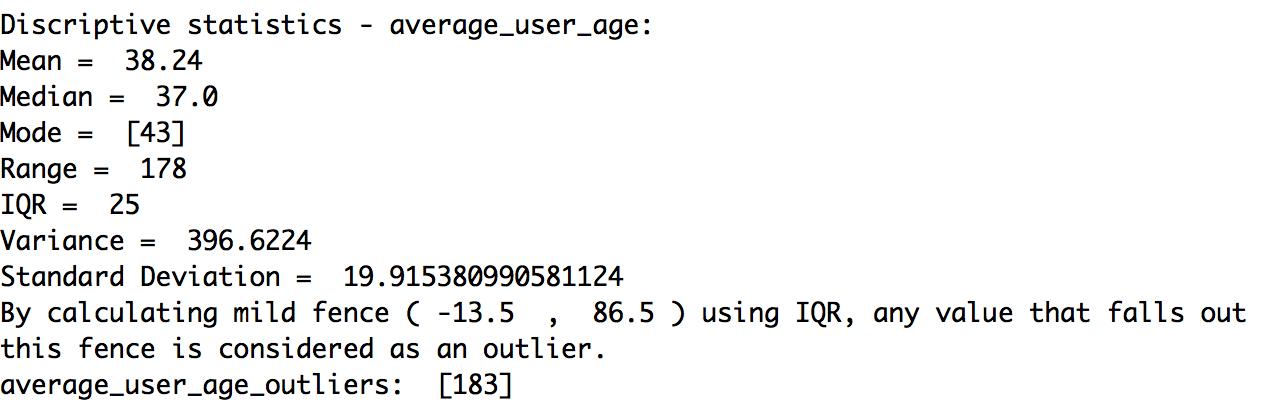


## Average Sales Value



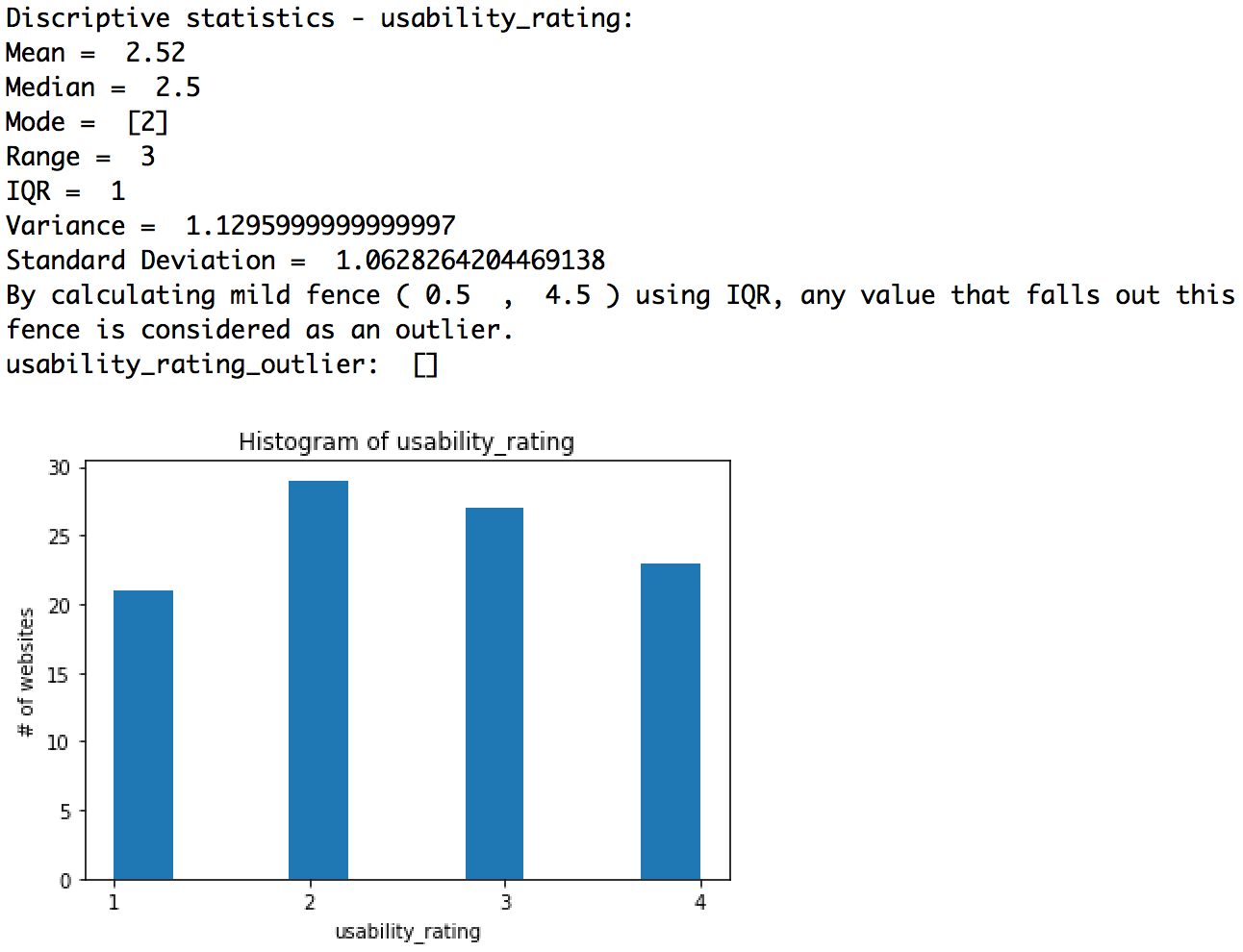
About 90% of average sales values fall into (0, 603.325), but the 10 outliers have huge influence on the mean. So, the median(67.8) in this instance better reflects the average sales values these 100 websites than the mean(496.31).

## Average User Age



The outlier (183) should be considered as an invalid value, because it’s almost impossible for a human being to live for 183 years in biology. In this case, the outlier has a subtle influence on the mean. Even if I remove the outlier, the mean without outlier (36.78) is still close to that with outlier. So, I can either choose the mean (38.24) or the median (37) to represent the common level of the average age of website users. It indicates the major user group of these 100 websites are young people.

## Usability Rating

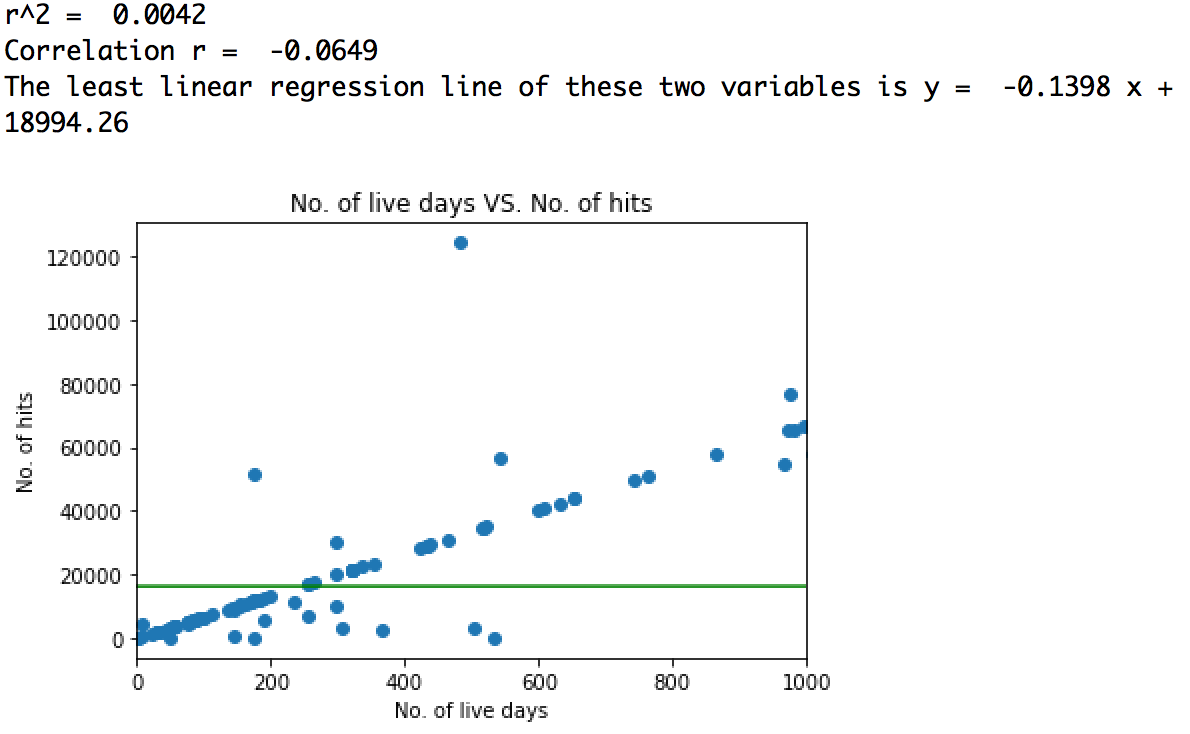


The usability rating has no outliers at all and a possible reason for this situation is that every observation is close to each other and the range of this set of values is just 3. The differences among mean, median and the mode are small. The usability of more than ¾ of these 100 websites is rated as 2 – Reasonable or even better, but there still have over 20 websites which are rated as poor usability and need improvements.

# Correlation & Linear Regression

In this part, I test three bivariate data sets of my interest to determine if there is a relation between the two variables of each data set.

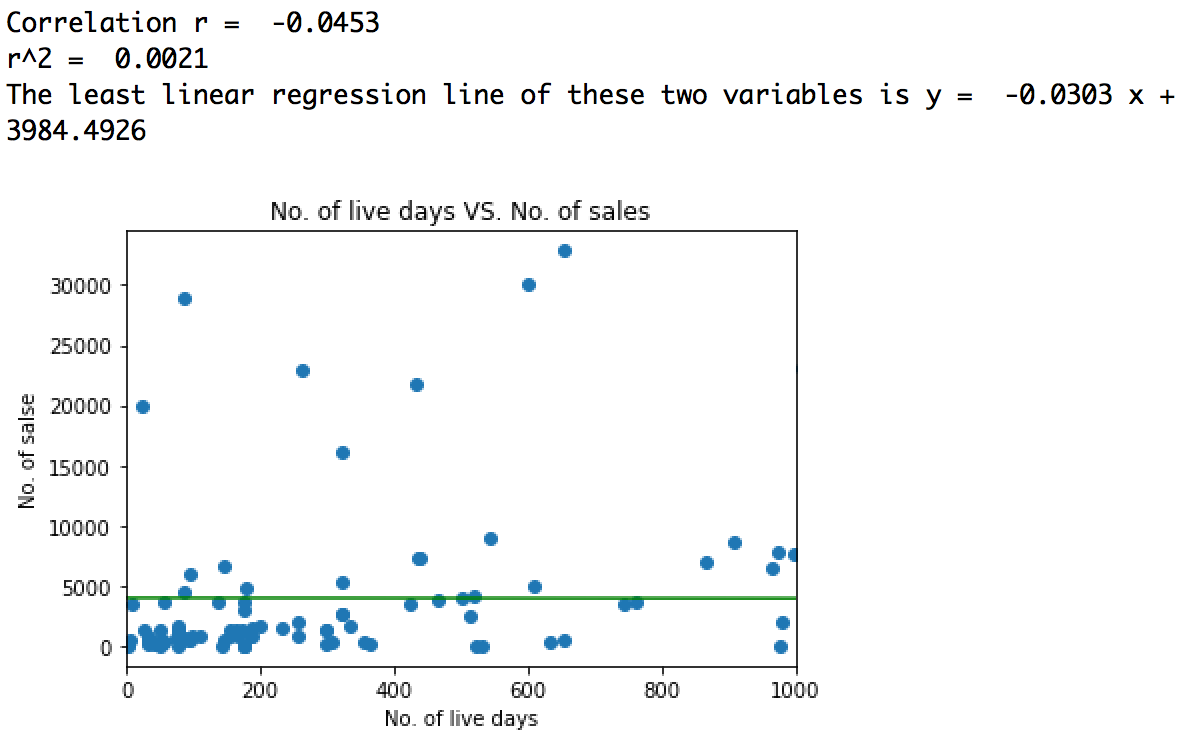
## The Number of Live Days & The Number of Hits



From the scatter plot, I can see there is a mild and positive relationship between the number of live days and the number of hits. However, the fact is different from my assumptions. After calculating the Pearson’s Correlation and the coefficient of determination, the correlation r (-0.0649) shows there is a very weak and negative or even no relation between these two variables. The r^2 (coefficient of determination) (0.0042) suggests that only 0.42% of variance in the number of hits is predictable from the number of live days. The least linear regression line also doesn’t fit the data presented in the scatter plot at all and it means this regression line can’t be used to estimate or compare the number of live days and the number of hits.

To draw a conclusion, there’s no relationship between the number of live days and the number of hits.

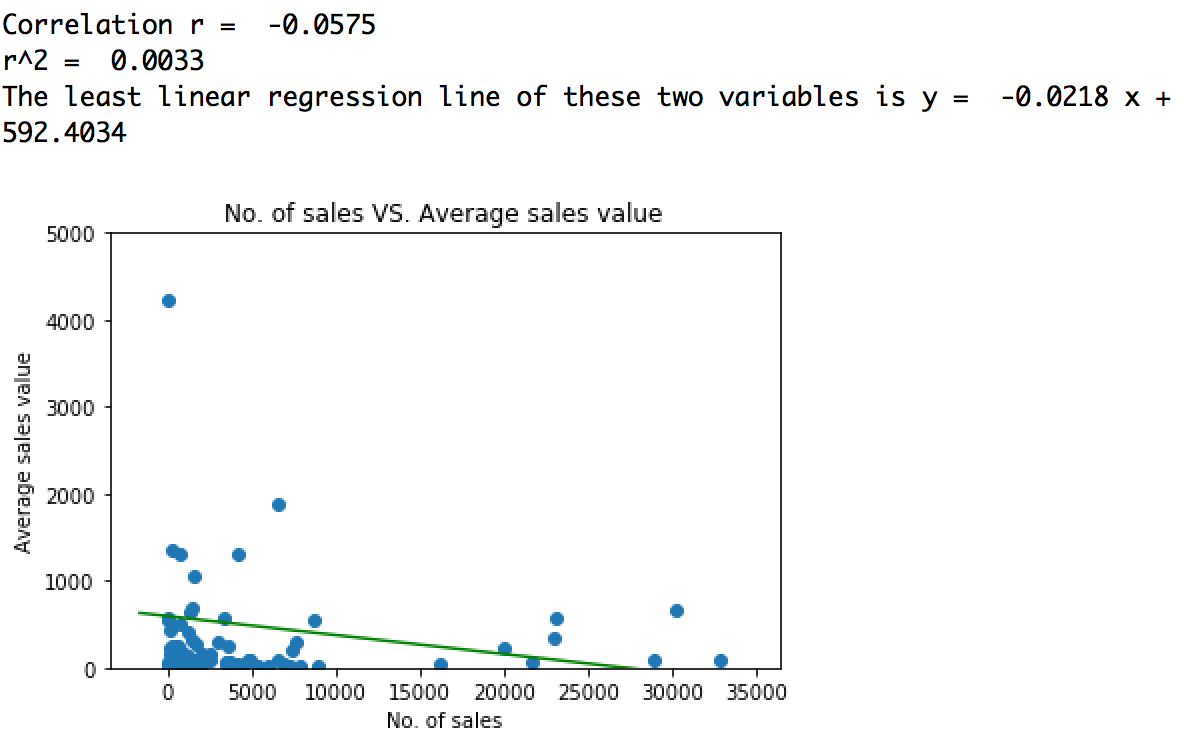
## The Number of Live Days & The Number of Sales



From the scatter plot, I can see there is no obvious relationship between the number of live days and the number of sales and the statistics calculated supports my assumption. After calculating the Pearson’s Correlation and the coefficient of determination, the correlation r (-0.0453) shows there is a very weak and negative or even no relation between these two variables. The r^2 (coefficient of determination) (0.0021) suggests that only 0.21% of variance in the number of sales is predictable from the number of live days. The least linear regression line also doesn’t fit the data presented in the scatter plot at all and it means this regression line can’t be used to estimate or compare the number of live days and the number of sales.

To draw a conclusion, there’s no relationship between the number of live days and the number of sales.

## The Number of Sales & Average Sales Value



From the scatter plot, I can see there is no obvious relationship between average sales value and the number of sales and the statistics calculated supports my assumption. After calculating the Pearson’s Correlation and the coefficient of determination, the correlation r (-0.0575) shows there is a very weak and negative or even no relation between these two variables. The r^2 (coefficient of determination) (0.0033) suggests that only 0.33% of variance in average sales value is predictable from the number of sales. The least linear regression line also doesn’t fit the data presented in the scatter plot at all and it means this regression line can’t be used to estimate or compare average sales value and the number of sales.

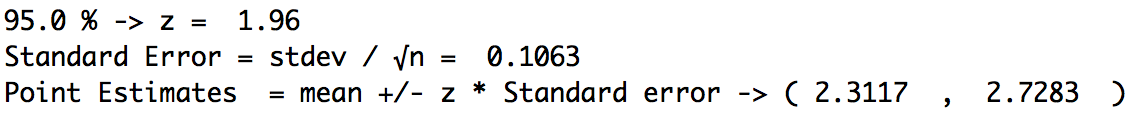
To draw a conclusion, there’s no relationship between average sales value and the number of sales.

# Confidence Interval

Basically, the confidence interval allows to know an estimated interval where the population parameter falls in based on the observed data and the confidence level.

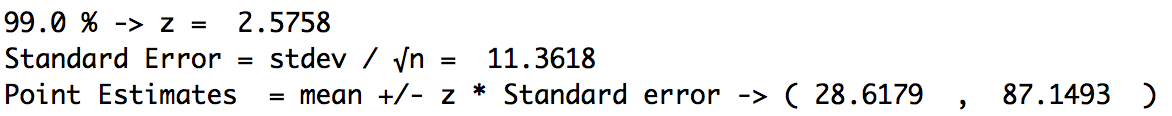
## Confidence Interval 1

Given that we have already known the mean of usability rating from the 100 samples, now I want to know the in which interval the population mean of usability ratings are in.



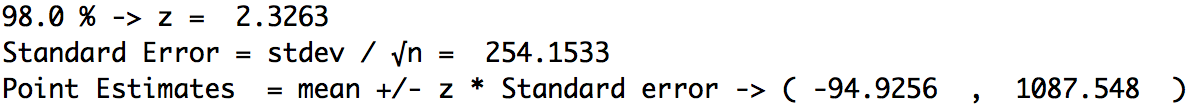
So, we’re 95.0 % sure that the population mean of Usability Rating are between 2.3117 and 2.7283.

## Confidence Interval 2



Based on the result above, I’d say that the population mean of Down Time will lie between 28.6179 and 87.1493, 99.0% of the time.

## Confidence Interval 3



So, we’re 98.0 % sure that the population mean of Average sales value are between -94.9256 and 1087.548.

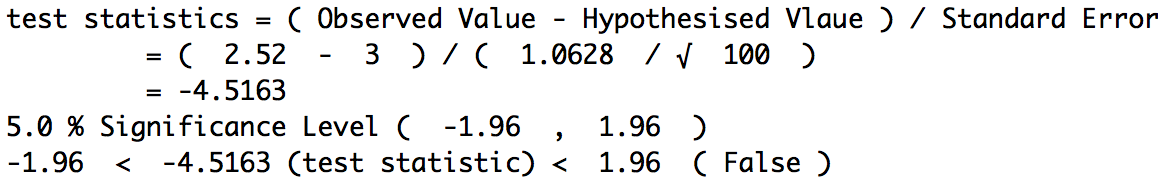
# Hypothesis Testing

In this phase, I try to make some conclusion about activities of all website during this period (31/12/2015 – 31/12/2016) based on the sample data of 100 websites.

## Hypothesis Test 1

H0: “The normal mean of usability rating of a website is 3 in 2016.” (μ = 3)

HA: “The normal mean of usability rating of a website is not 3 in 2016.”(μ ≠3)

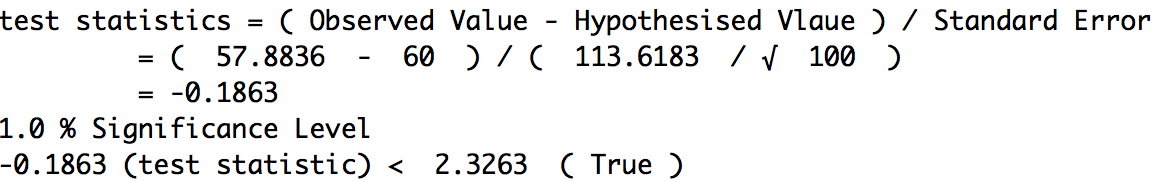


The null hypothesis is rejected and there is sufficient evidence to warrant reject of the claim that the normal mean of usability rating of a website is 3 in 2016.

## Hypothesis Test 2

H0: “The average down time of a website in 2016 is less than or equal to 60 hours.” (μ ≤ 60)

HA: “The normal mean of usability rating of a website is not 3 in 2016.”(μ 60)

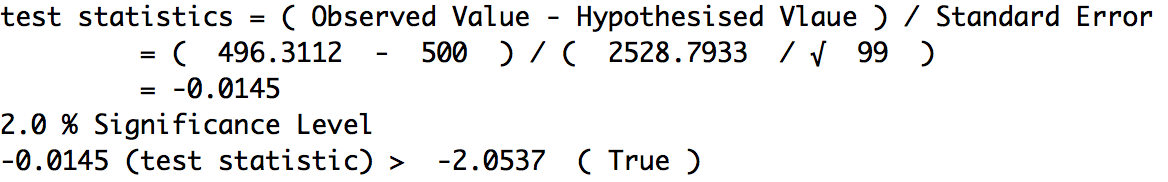


We fail to reject the null hypothesis and conclude that there is insufficient evidence to reject the claim that the average down time of a website in 2016 is less than or equal to 60 hours.

## Hypothesis Test 3

H0: “The average sales value of a website in 2016 is greater than or equal to $500.” (μ 500)

HA: “The average sales value of a website in 2016 is less than $500.”(μ 500)

****

We fail to reject the null hypothesis and conclude that there is insufficient evidence to reject the claim that the average sales value of a website in 2016 is greater than or equal to $500.

# Extension of Research on Twitter Data

## Intro

In the data science assignments from last semester, I carried out a research on how active people who also like Coldplay are on Twitter based on the data of 1000 tweets which contain Coldplay’s hash tag.

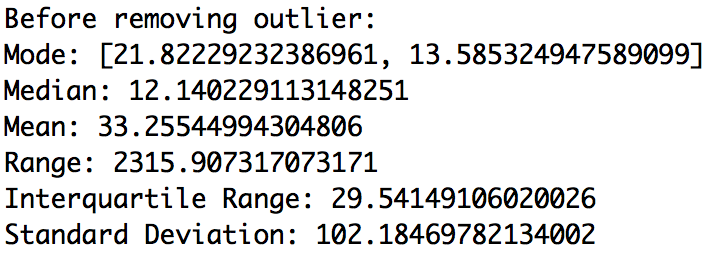
## Obtain & Prepare Data

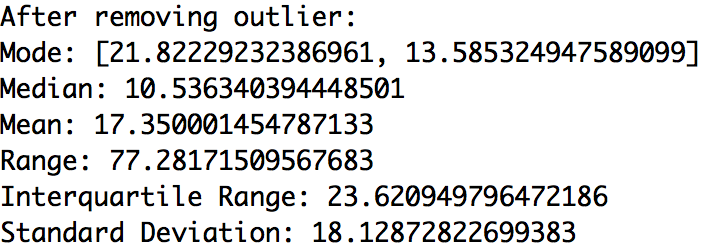
This time I want to make an assumption of the average number of tweets per day of a Twitter user who likes Coldplay since the account being created. So, I doubled the sample size which is 2000. I took advantage of Twython to use Search API to search for tweets containing the phrase "amazing" and then I used Streaming API to filter and save all tweets which contains keyword “Coldplay” into a text file.

Before carrying out analysis, I need to cope with the outliers first. To identify outliers, I calculated the mild fence and any value falling out this range will be considered as an outlier. As the screenshot below shows, the mild fence of the average number of posts per day is (-40.62, 77.54), all number outside this range are outliers.



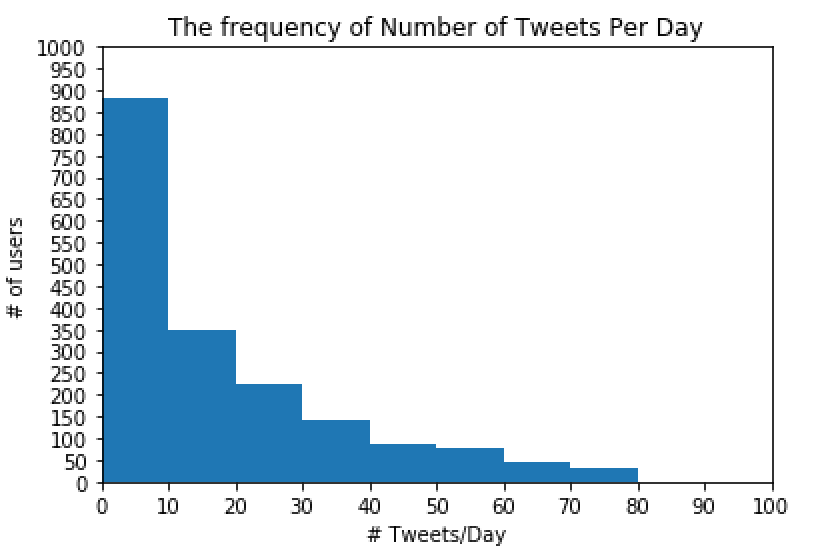
The two screenshots below indicate the result of calculating some measures with/without outlier. As shown on the screenshots, modes and medians has subtle changes which is acceptable, while means, ranges, interquartile range and standard deviation undergo significant changes after removing outliers. In this case, outliers will skew the analysis results if we choose to keep the outliers and perform the analysis. Therefore, I chose to remove the outliers.





After removing those outliers, the sample size shrinks to 1840.

## Analysis Results



I produced a histogram to present the average number of tweets per day of each user and their frequencies. As is shown to me, approximately 880 users in these 1840 samples tends to tweet less than 10 tweets per day.

Back to the results without outliers, two values were shown as modes and this means each of them has the same occurrence. In this case, I didn’t take this measure represent the majority, because it has several values and it didn’t cover all data points.

Median is 10.54 which is the middles value among all average numbers of posts per day and it represents the middle level.

Mean is 17.35 and it’s close to median. I suppose it can represent the average or middle level of how active these 1000 users who like Coldplay are on Twitter.

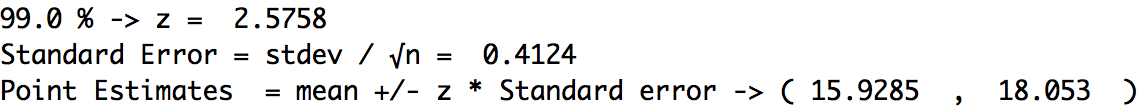
Range is 77.28 and it indicates this set of data fluctuates obviously and there is distinct difference between the maximum average number of posts per day and the minimum one.

Interquartile range is 23.62 which implies the difference between the maximum and minimum of middle 50% part of average numbers of posts per day. It is far less than range and I suppose the average numbers of posts per of the middle 50% part are very close.

Standard deviation is 18.13 and it tell me each average number of post per day is not very close to the mean and they are a little bit scattered.

## Confidence Interval & Hypothesis Test

Given that we have already known the mean of the average numbers of tweets per day of a Twitter user who likes Coldplay since the account being created from the 1840 samples, now I want to know the in which interval the population mean of the number is in.

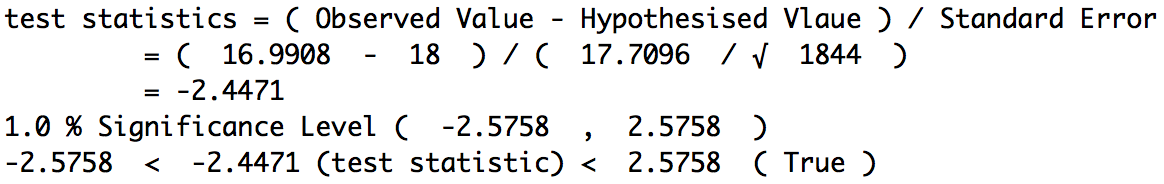


Given the result above, the population mean of the average numbers of tweets per day of a Twitter user who likes Coldplay since the account being created lies between 15.9285 and 18.053.

Based the analysis result, my null hypothesis is that the average number of tweets per day of a Twitter user who likes Coldplay since the account being created is 18, and the alternative hypothesis is that the average number of tweets per day of a Twitter user who likes Coldplay since the account being created is not 18.

H0: “The average number of tweets per day of a Twitter user who likes Coldplay since the account being created is 18.” (μ = 18)

HA: “The average number of tweets per day of a Twitter user who likes Coldplay since the account being created is not 18.”(μ ≠18)



At 1% significance level, we fail to reject the null hypothesis and conclude that there is insufficient evidence to reject the claim that the average number of tweets per day of a Twitter user who likes Coldplay since the account being created is 18.