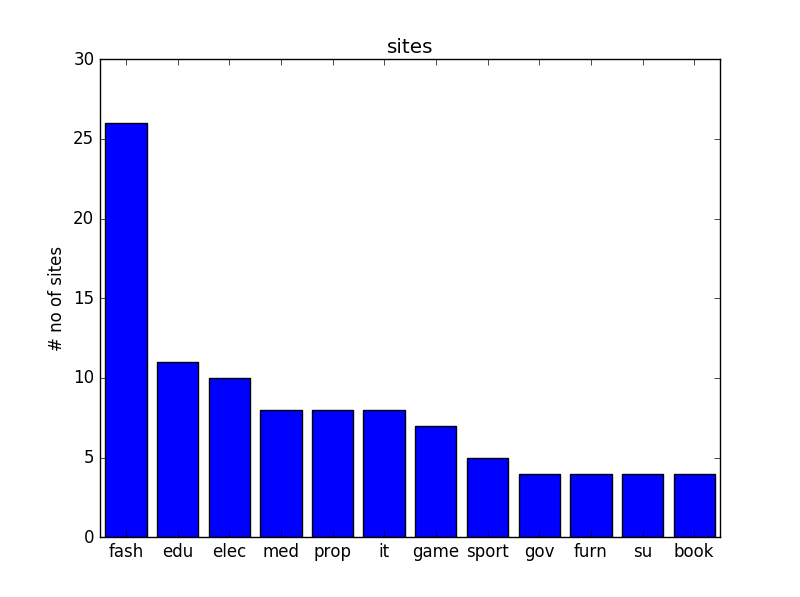
# Description of the data

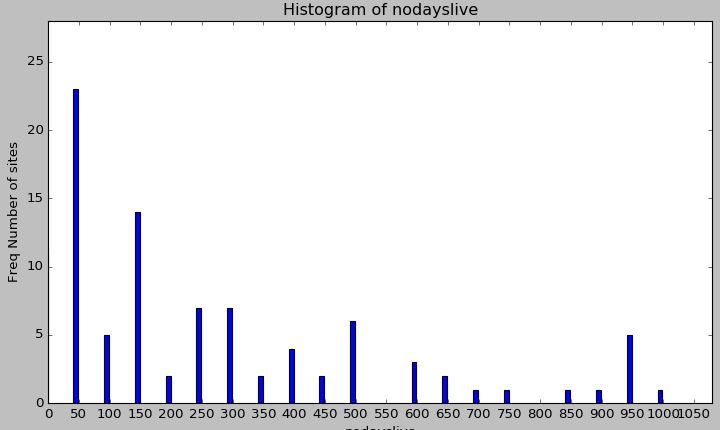
I started my CA by formatting the data from excel to lists in python. Some of the data sets had values N/A I decided to get the average of the data set it belonged to and replaced the n/a entry with the average value. For each of the data sets I generated box plots and observed scatter plots to assist in finding potential outliers as these would have effect on the mean range and standard deviation. I did calculations of the mean with only extreme outliers removed but for other calculations I kept the outliers. In one instance in avg age the value of 138 has been removed for all calculations as it must be an error

To get a description of the data I plotted histograms for each of the data sets. The histogram plots showed the underlying frequency distribution (shape) of the data. This allows the inspection of the data for its underlying distribution (e.g., normal distribution), outliers, skewness, etc.



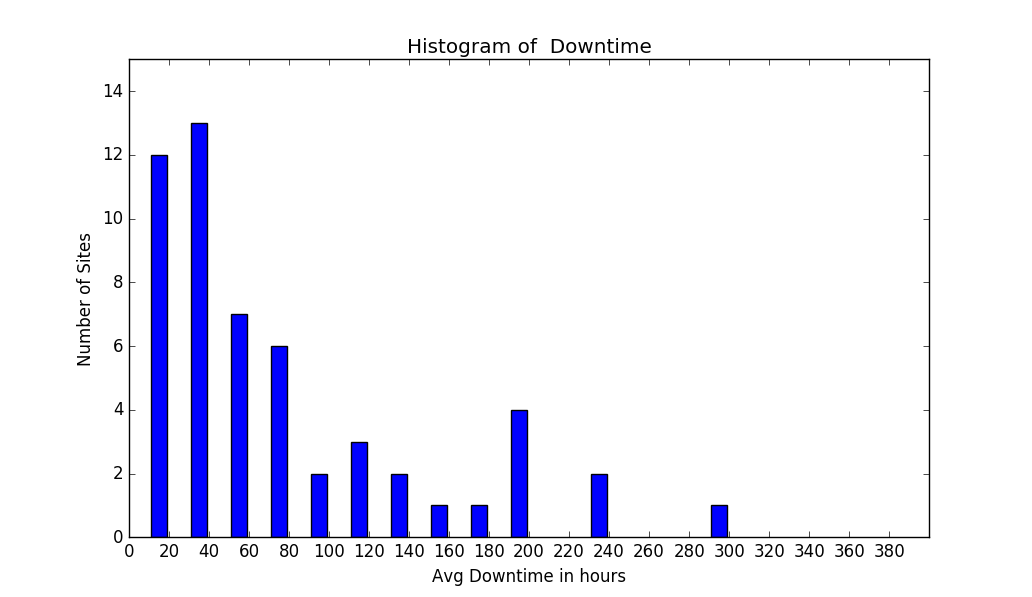
**Bar chart showing the proportion of types of websites present in the data set**

**Histogram No of Days Live**

****

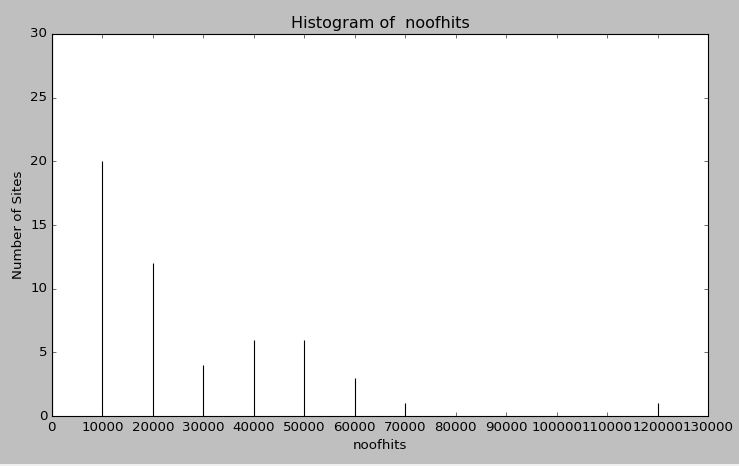
Left skewed Distribution Histogram most of the no days live value lie between 0 and 150 days

**Histogram Downtime**

****

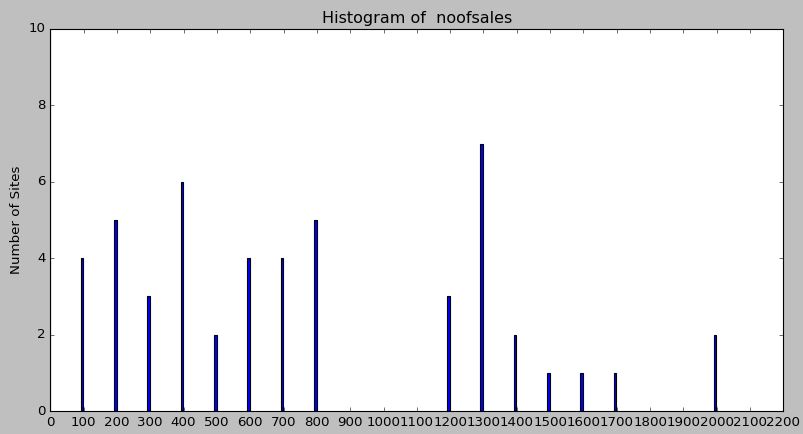
Left skewed Distribution most of the down times lie between 0 and 80 hours

**Histogram No of hits**

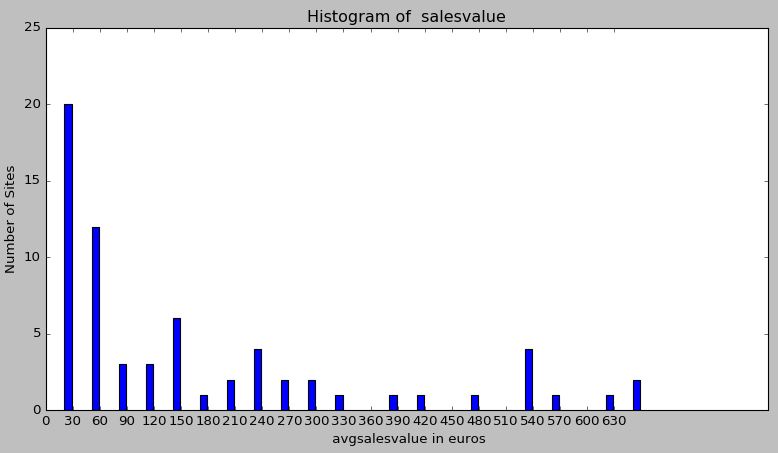
****

Left skewed(left tail) Distribution Histogram most of the no of hits value lie between 0 and 20000

**Histogram no of Sales**

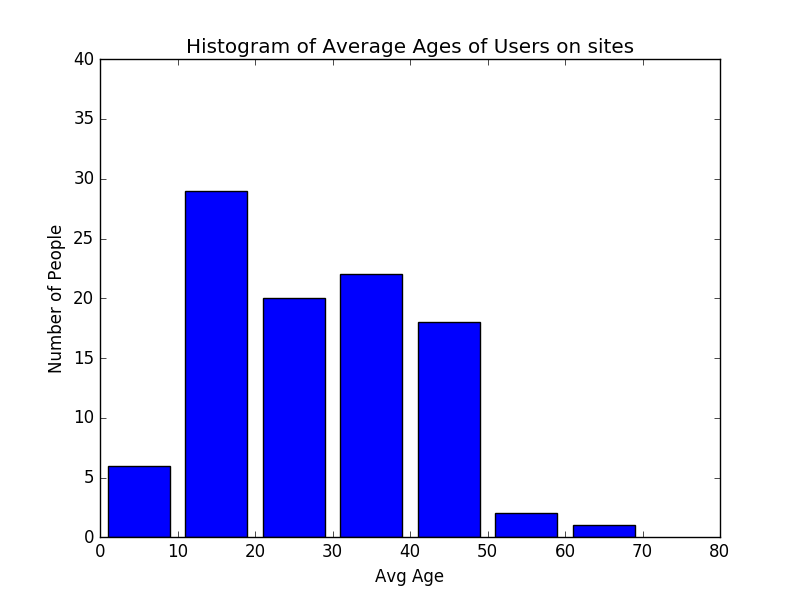
****

**Histogram Average Sales Value**

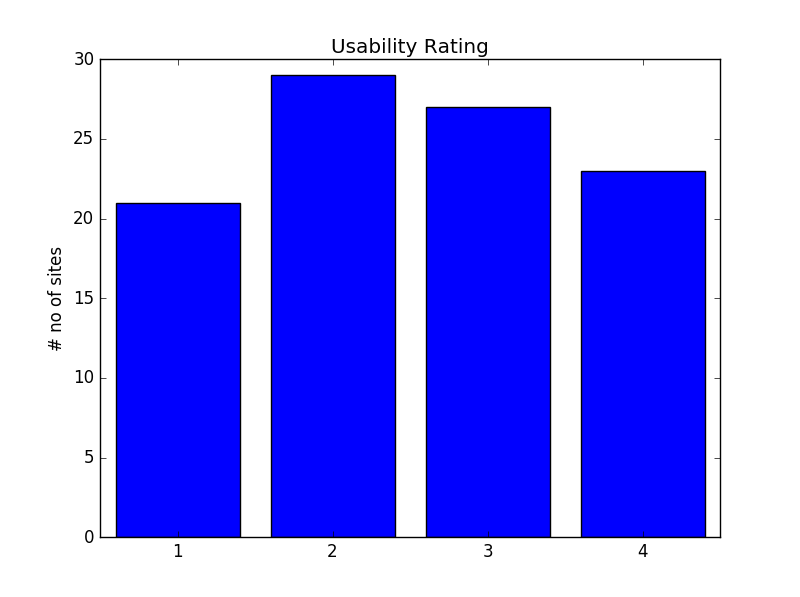
****

Left skewed Distribution Histogram most of the average sales value lie between 0 and 60

**Histogram Average Age**

****

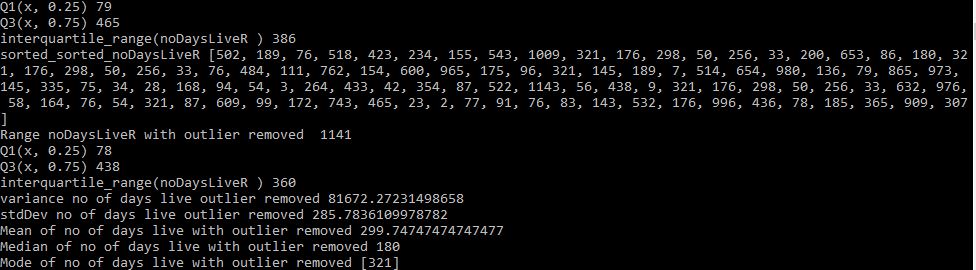
**Bar Chart Usability Rating**

****

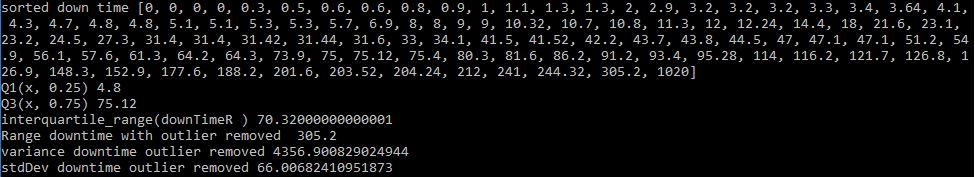
# Descriptive Statistics

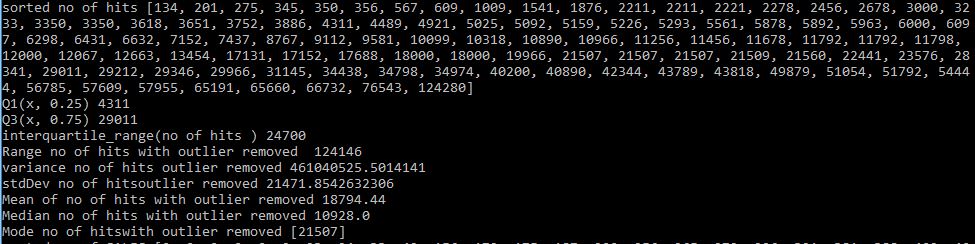
The CA required Measures of Spread and central tendencies for each of the variables in the data set.

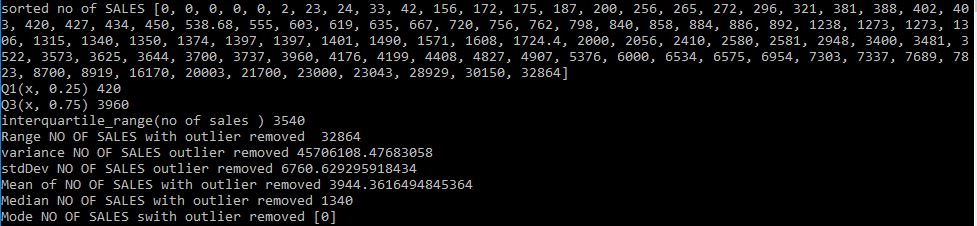
**NO OF DAYS LIVE**

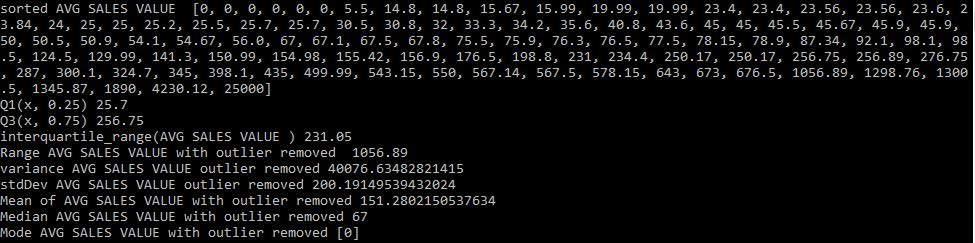


**DOWN TIME**

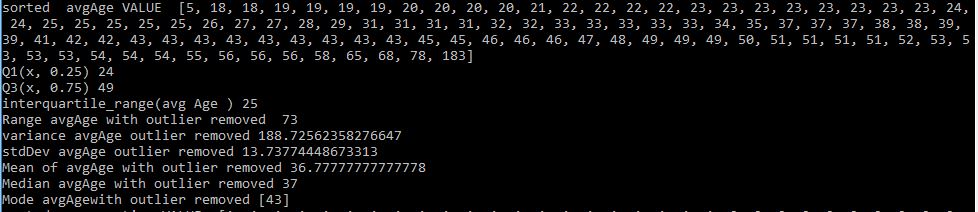


 **NO OF HITS**

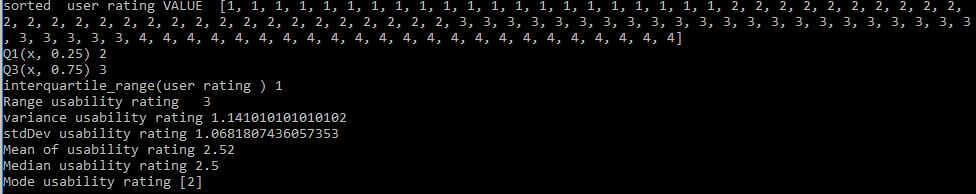
**** **NO OF SALES**

 **AVG SALES VALUE**

**AVGAGE**



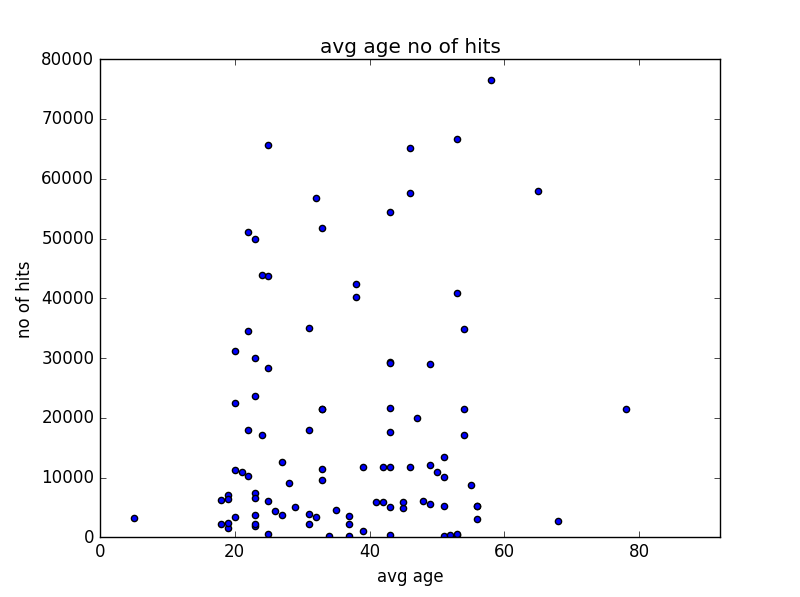
**USABILITY RATING**

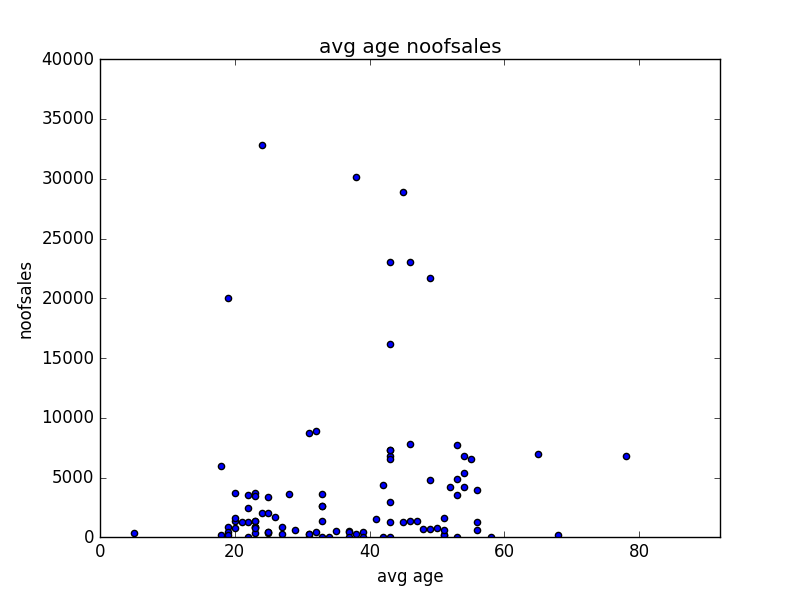


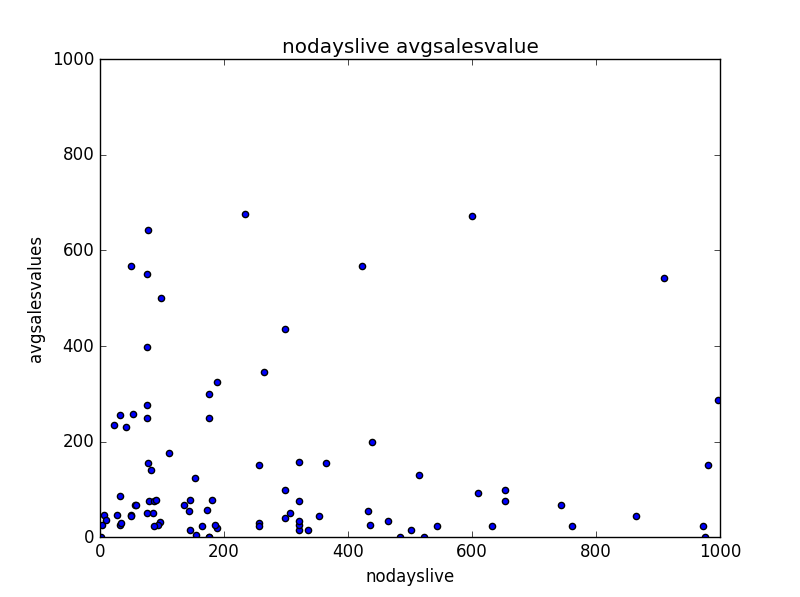
# Correlation and Linear Regression.

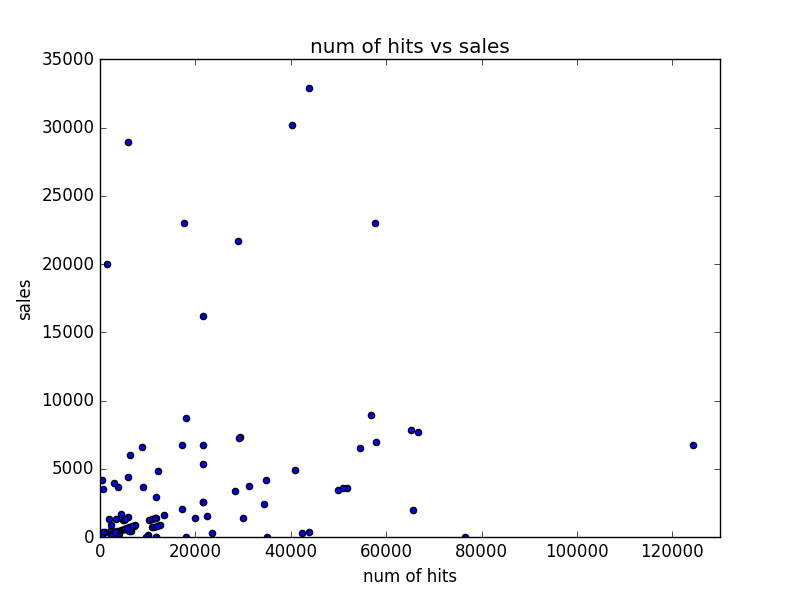
Generally, regression analysis can be used for prediction of the target variable (forecasting), modelling the relationship between x and y and testing of hypotheses.

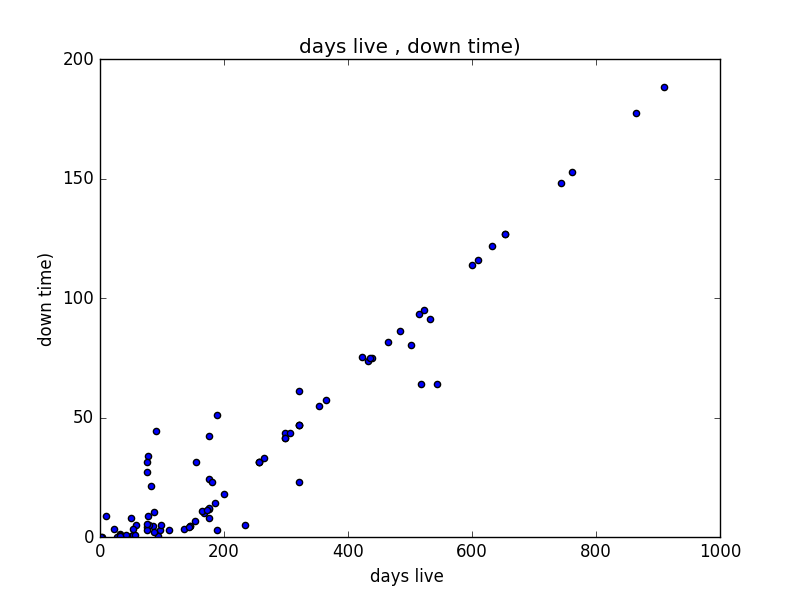
To start this part of the ca I plotted several scatter plots to help identify data sets with a strong positive correlation. A strong positive correlation would mean the least squares regression line I would fit to the data would be usable to infer something from the da tie make some projections of what may happen the future. The scatter plots where also of use earlier in the ca in the measure of central tendency and spread. They helped for observing possible outliers in the data.

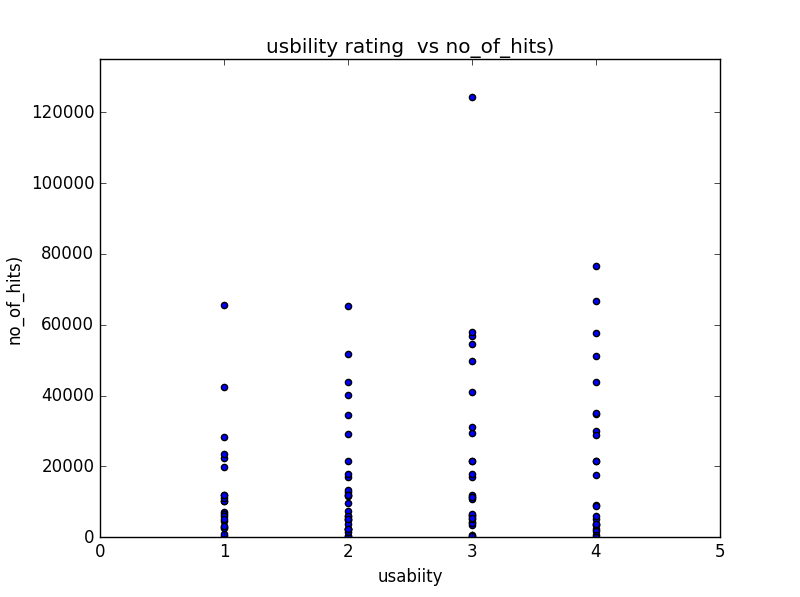
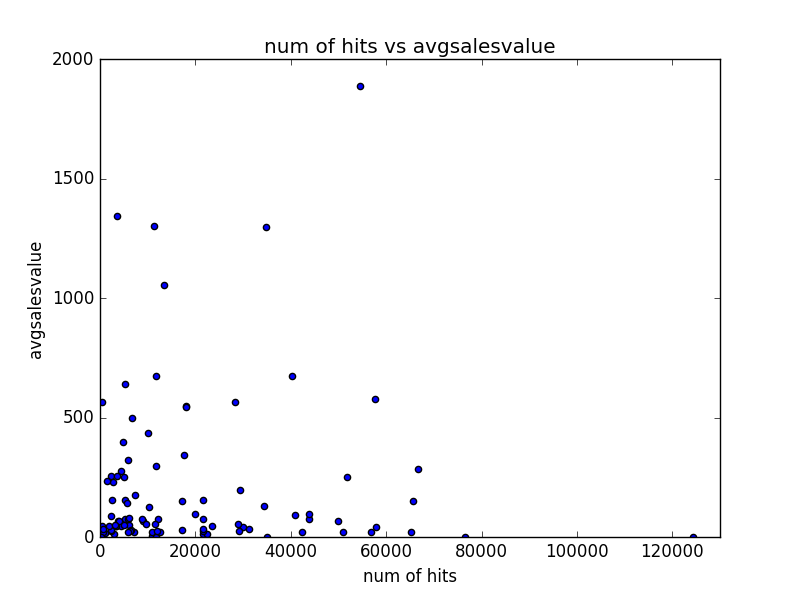


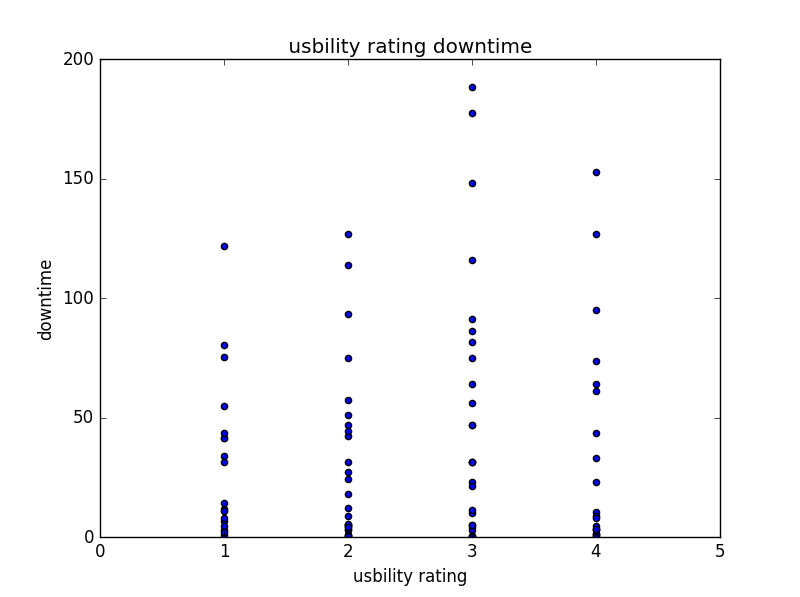


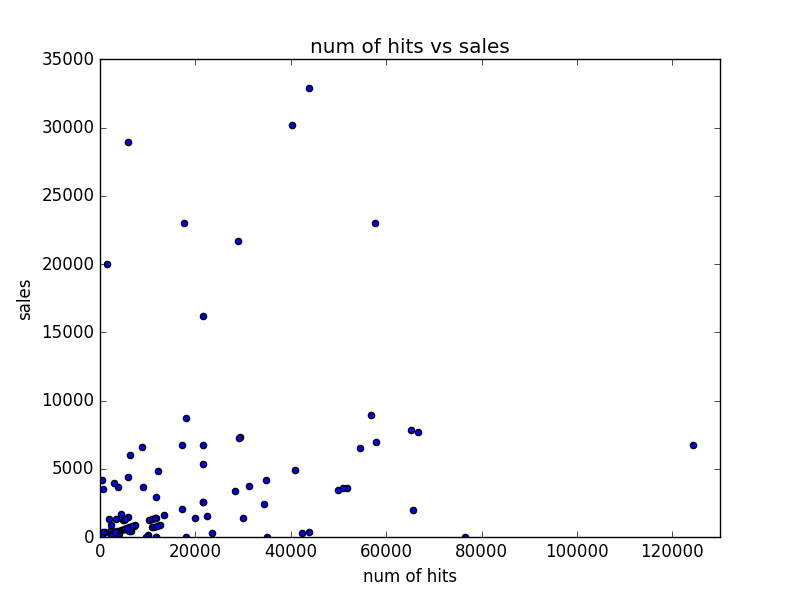












From observation of my scatter plots. It appears that many of the pairs of data sets I choose do not have a strong correlation. Apart from Days live versus downtime There did not appear to be a linear relationship with the pairs of data sets. However, the result of the correlation between Days live and downtime was .0094 although technically a positive correlation it was not strong enough for me to undertake regression analysis on.

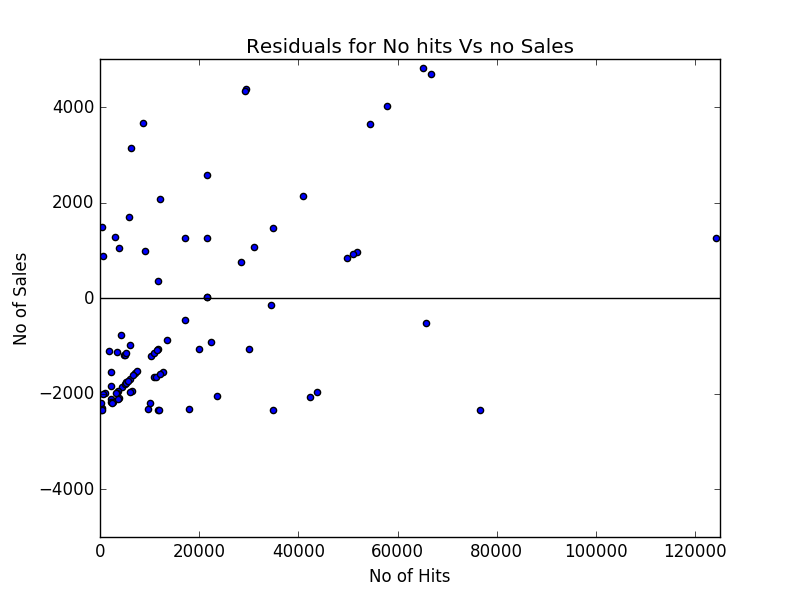
The two pairs of data that I will undertake linear regression on are **1. no of hits** with **no of sales** and **2.** **usability rating** with **no of sales**

1. **no of hits** with **no of sales**

The correlation was (num of hits, amount of sales) 0.274711882586 indicating a weak correlation and therefore there is not much indication that as no of hits increase then so do sales. No real strength of relationship between the two variables. There is also be a large scatter of data from the regression line.

Nofhits = independent variable

Noofsales= dependent variable (predictors)



The least squares regression line for the data is **y = .085 x + 2344**

**R Squared value.**

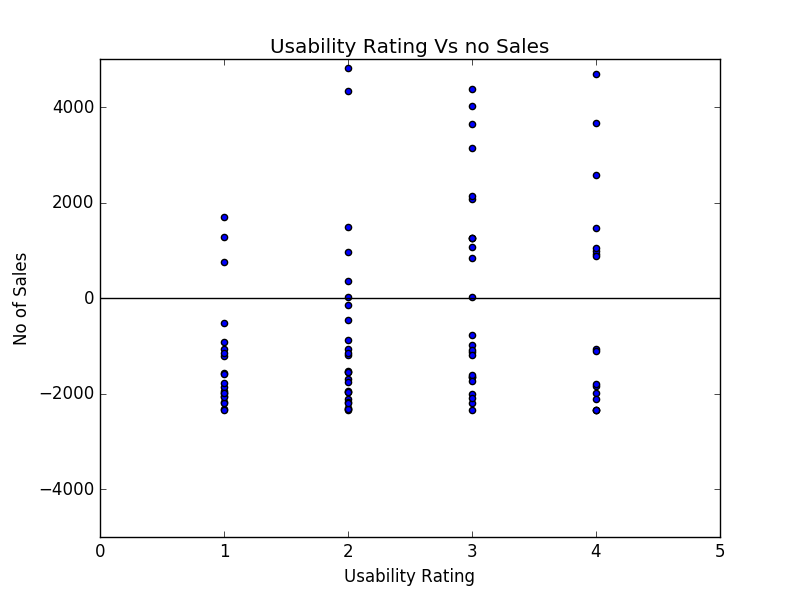
.0729 r squared value for no of hits and no of sales indicates that 7.29 percent of the variation in sales can be explained by the variation in hits.

1. **usability rating** with **no of sales**

usability rating = independent variable

no of sales = dependent variable (predictors)

the correlation was (usability rating, no of sales) 0.33 is a weak correlation, and therefore there is not much indication that as usability ratings increase then so do sales. There is no real strength of relationship between the two variables. There is also be a large scatter of data from the regression line.

 The least squares regression line for the data is **Y= 01397x +3944**

**R Squared value.**

**.01108 r** squared value for no of hits and no of sales indicates that 11.08 percent of the variation in sales can be explained by the variation in usability ratings.

# Inferential Statistics.

**Confidence Intervals** on the mean for all websites during this period.

**1** .The mean for **usability ratings** on the websites sampled for this period is 2.52. To say this is the mean for all websites i.e the population as whole I will need to examine confidence intervals for the data. N = 100 sigma = 1.068

* Constructing a 95% confidence interval for usability rating.

2.52 +- .25 \* sigma / sqrt N

2.52 +- 1.96 \* 1.068 / sqrt 100 (.25 mapped from each side i.e. outer 2.5% values)

(2.31, 2.72)

1. the mean will lie between these set of values 90 % of the time assuming the data for the population follows a normal distribution

**2** The mean for **Average Age** on the websites sampled for this period is 37.77. To say this is the mean for all websites i.e. the population as whole I will need to examine confidence intervals for the data. N = 100 sigma = 13.74. for this data, the outlier of 138 has been removed from this calculation. it appears to be an error.

* Constructing a 99% confidence interval for average age.

37.11 +- .005. \* sigma / sqrt N

37.11 +- 2.58. \* 13.74 / sqrt 100 (.005 mapped from each side i.e. outer 1% values)

(34.22 ,41.31)

1. the mean will lie between these set of values 98 % of the time assuming the data for the population follows a normal distribution All normal distributions follow a standard curve, only difference is how steep the curves are and where the mean falls

**Hypothesis testing (Spreadsheet Data)**

Now testing a claim about a population in contrast to confidence intervals where we are estimating the value of a population parameters.

**Claim 1** - The mean for **usability ratings** for all websites in the period is 2.6 using a 1 % significance level.

1. **Identify the null and alternative Hypothesis**

H0 = n 2.6 The average usability rating for websites is 2.6

H1 = n! = 2.6 The average usability rating for websites is not 2.6

1. **Calculate the test statistic**

z = observed value – hypothesised value / standard error

z = sample mean – population mean / standard error

z = 2.52 – 2.6 / (1.068/sqrt 100) = -.750

1. **Determine the Critical value.**

**This is a two-tailed test with sigma = .01 so the critical values are + Z 0.005 and -Z .005**

**-Z .005 = - 2.5758 and Z .005 = 2.5758**

**As - 2.5758 <=** -.750 <= **2.5758 we are in the fail to reject H0 or null hypothesis region**

**We fail to reject null hypothesis( We accept the claim)** that mean for **usability ratings** for all websites in the period is 2.6 using a 1 % significance level.

**Claim2** - The mean for **Average age** for all websites in the period is 39 using a 5 % significance level.

1. **Identify the null and alternative Hypothesis**

H0 = x = 39 The average age for websites is 39

H1 = x! = 39 The average age for websites is not 39

1. **Calculate the test statistic**

z = observed value – hypothesised value / standard error

z = sample mean – population mean / standard error

z = 37.77 – 39 / (13.74/sqrt 100) = -.8952

1. **Determine the Critical value.**

**This is a two-tailed test with sigma = .0 so the critical values are + Z 0.025 and -Z .025**

**-Z .025 = - 1.96 and Z .025 = 1.96**

**As – 1.96 <=** -.8952 <= **1.96 we are in the fail to reject H0 or null hypothesis region**

**We fail to reject null hypothesis (We accept claim)** The mean for **average age** for all websites in the period is 39 using a 5 % significance level.

**Hypothesis Testing the relationship of association between 2 variables using a 1 % significance level**

H0 p = 0 no linear relationship between usability rating and sales

H1 p = 1 no linear relationship between usability rating and sales

**1 Calculate the test statistic**

r = .33

**2.calculate the critical value**

**Sigma = .05 =**

**100 – 2 = 98 => .9166**

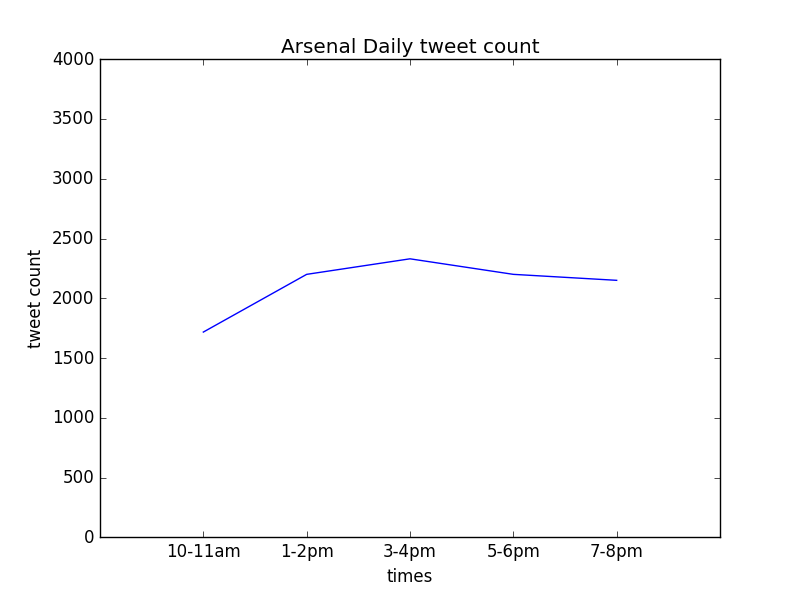
**3** If R = -.9166 r <.9166 we fail to reject the null hypothesis. Otherwise we reject h0 and support claim that a relationship exists between the variables.

As the r value of .33 is within the range of these values we accept the null hypothesis claim that no relationship exists between the variables.

# Analysis of Twitter data

**Confidence Intervals**

1. On the mean of arsenals daily hourly tweet count



**Confidence Intervals** on the mean for all websites during this period.

**1.** The mean for **hourly tweets** on daily tweets on a given number of hours sampled from the Twyton data with regards to tweets about arsenal is 2082 per hour with STD DEV of 250.54 n = 100

* Constructing a 95% confidence interval the mean of hourly tweets regarding arsenal.

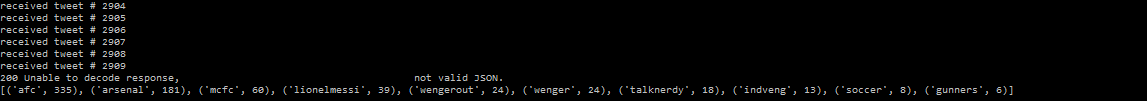
2082 +- .25 \* sigma / sqrt N

2082 +- 1.96 \* 250.54 / sqrt 100 (.25 mapped from each side i.e. outer 2.5% values)

(2032,2131)

1. the mean will lie between these set of values 95 % of the time assuming the data for the population follows a normal distribution
2. **Twitter Data (Proportion Confidence Interval)**

One of Twyton searches involved searching the keyword “wenger out” to see if the club supporters want the manager sacked, a common call from fans when the club is not performing well.



The number of hashtags with “Wenger out” was 24 from a total tweet count of 2909. As this is a sample result from a population of tweets I want to be 95 % confident that I have captured the population proportion.

Pop proportion = p p^ = sample proportion sample size = n no of occurrences = x

n = 2909 p^ = 24/2909 = .00825

95 % CI = p^ +- SQRT (p^(1-p^)/N)

= .00825+- 1.96 SQRT (.00825 (1-.00825)/2909)

= .00825 +- 1.96 ( .0017 )

= ( .005168 , .011582 )

Conclusion - we can be 95 % confident the population proportion of hashtags “Wenger out “are between (005168 , .011582)

**Hypothesis testing (Twitter Data)**

The mean for **hourly tweets** on daily tweets on a given number of hours sampled from the Twyton data with regards to tweets about arsenal is 2082 per hour with STD DEV of 250.54 n = 100

**Claim 1** - The mean number of tweets about Arsenal per hour is 2100 with a 1 % significance level.

**1)Identify the null and alternative Hypothesis**

H0 = x = 2100 The average number of tweets about arsenal is 2100

H1 = x! = 2100 The average number of tweets about arsenal is not 2100

**2)Calculate the test statistic**

z = observed value – hypothesised value / standard error

z = sample mean – population mean / standard error

z =2082– 2100 / (250.54/sqrt 100) = -718

**3)Determine the Critical value.**

**This is a two-tailed test with 1 % significance so the critical values are + Z 0.005 and -Z .005**

**-Z .005 = - 2.5758 and Z .005 = 2.5758**

**As – 2.5758 <=** -.718 <= **2.5758 we are in the fail to reject H0 or null hypothesis region**

**We fail to reject null hypothesis (We accept Claim)** - The mean number of tweets about Arsenal per hour is 2100 with a 1 % significance level.

**Claim 2** - The Proportion number of hashtags “wenger out” in a population of tweets are greater or equal to .007 assuming a significance level of 1 %

**1)Identify the null and alternative Hypothesis**

H0 = x >= .007 the population hashtags “wenger out” are greater or equal to .007

H1 = x < .007 the population hashtags “wenger out” are less than .007

Hypothesis is one-tailed test. The null hypothesis will be rejected only if the sample proportion is too small.

**2)Calculate the test statistic**

**N = 2909 P = .007 p^ = 24/2909 = .00825**

σ = sqrt[ P \* ( 1 - P ) / n ]

= sqrt[ .007 \* ( 1 - .007 ) / 2909 ]

= sqrt(0.000002389) = 0.001545

z = (p - P) / σ

= (.00825 - .007)/ 0.001545 = .809061488

where P is the hypothesized value of population proportion in the null hypothesis, p^ is the sample proportion, and n is the sample size.

**3)Determine the Critical value.**

**This is a one-tailed test with 1 % significance so the critical values is Z = 0.005**

**Z .005 = 2.5758**

**As** .809061488 <= **2.5758 we are in the fail to reject H0 or null hypothesis region**

**We fail to reject null hypothesis (We accept Claim)** - number of hashtags “wenger out” in a population of tweets is greater or equal to .007 assuming a significance level of 1 %