Homework Assignment 4

Solution to Q4:

About UCI repository Balance Scale Data

Dataset: Balance Scale Dataset with no missing values
Characteristic of Data: Multivariate but not time series
Attributes: Categorical
# Attributes: 4
# Instances: 625

Using the above dataset, I have calculate the summary statistics
L: 288
R: 288
B: 49
3.0
rows: 625
cols: 1
median of left balance 8.0
median of right balance 8.0
standard deviation of left balance 6.32455532034
standard deviation of right balance 6.32455532034
minimum element in left balance 1.0
maximum element in left balance 25.0
minimum element in right balance 1.0
maximum element in right balance 25.0
median absolute deviation for left balance 4.0
median absolute deviation for right balance 4.0

L,R,B stands for left balanced, right balanced and balanced respectively.
The strange part of the data is although the values of each tuple is different but the statistics values are coming out to be same. So considering each tuple individually, you can define left balanced, right balanced and balanced respectively but as a whole it is balanced.

The same statistic has been visualized using the histogram, boxplot and scatter plot. The correlation plot of the data showed some pattern in 'maroon color' which can be thought of as strong connected components of L, R and B.

The heat map or density map of UCI data, has on its y axis the tuples and the x axis the normalized values of left balance and right balance. Looking at the heat map, there is clear visualization that considering normalized value of 1.0, we see that all the initial 100 tuples has majority of balanced data and more over, the demarcation between Left balance, Right balance and Balance is not very clear but considering the normalized value above 1.0 we see that majority of data is inclined towards left balance and right balance. Due to this, Left and Right balance are highly dark as compared to those less than 1.0 normalized value.

About the twitter data

For calculating the word cloud, I have used the word cloud library of python. The word cloud was made using the text extracted using Twitter API. Some of the frequently used words were https, Albany, NY, Job, Hiring etc. The higher frequency words have larger fonts as compared to those words which have lower frequency and lower fonts.

I have also calculated the density plots of twitter data based on sentiments. The sentiment measure is calculated either objectively or subjectively. Both measure were taken into consideration. The subjective sentiments tend to be declining in an exponential way as compared to objective sentiments. This was because, in some cases the objective sentiment measure was not coherent with subjective measure. To illustrate this incoherence, we made the parallel coordinate plot showing the number of tuples in which objective sentiment is different from subjective sentiment and by how much difference. The scatter matrix plot was showed that some of subjective measure didn't follow up objective measure. The correlation plot showed components in the sentiment matrix based on objective and subjective measure relation.
The twitter data showed similar behavior in statistics as UCI balance scale data in which considering the whole dataset, we are not able to visualize small changes which are happening at tuple level.