**CURRENT STATUS:**

**Project objective and tasks:**

The objective of this project is to analyze whether the customer preference over a restaurant is high or low based on its Price range and the Parking facility. We use the number of checkin’s of each restaurant to analyze the preference. The Data Mining task, Clustering is used for this project and it is implemented using WEKA tool.

The following tasks have been completed

1. Preprocessing: Converting the Yelp datasets which is in JSON format to a CSV file.
2. Get the average count of check-in’s for each business from the Check-in dataset.
3. Join the average count of check-in’s with the Price range and Parking facility feature of attributes from Business dataset based on business id and create a CSV file

**Deliverables:**

1. The CSV files which we have converted from JSON format.
2. The code we used for conversion of JSON to CSV format.
3. The file after performing aggregation on Check-in’s.
4. The code we used to perform the aggregation.
5. The screenshot of the clustering we did in WEKA till now

**Initial Implementation:**

The Yelp datasets, Business and Check-in which are in JSON format have been converted to CSV files to be implemented in WEKA as it cannot be used in the original format. We used a conversion code from github but modified the code as per our requirements. For each business in Check-in dataset, we aggregated the number of check-in’s in a weekly basis. After that we created a new csv file joining the attributes of Business dataset, Parking and Price Range with the aggregated value of check-in based on the business ID.

Using the output of aggregation done on the checkin list we made clusters for the business ids based on the total count of checkin each business received weekly.

We used the weka tool to analyze the aggregated data. K-means algorithm was used to make 3 cluster with only total number of checkins taken as clustering attribute. Ignoring Business id is important here since each business has a unique id and does not help in deciding clusters. Also Euclidean distance is taken as the measure for the algorithm.

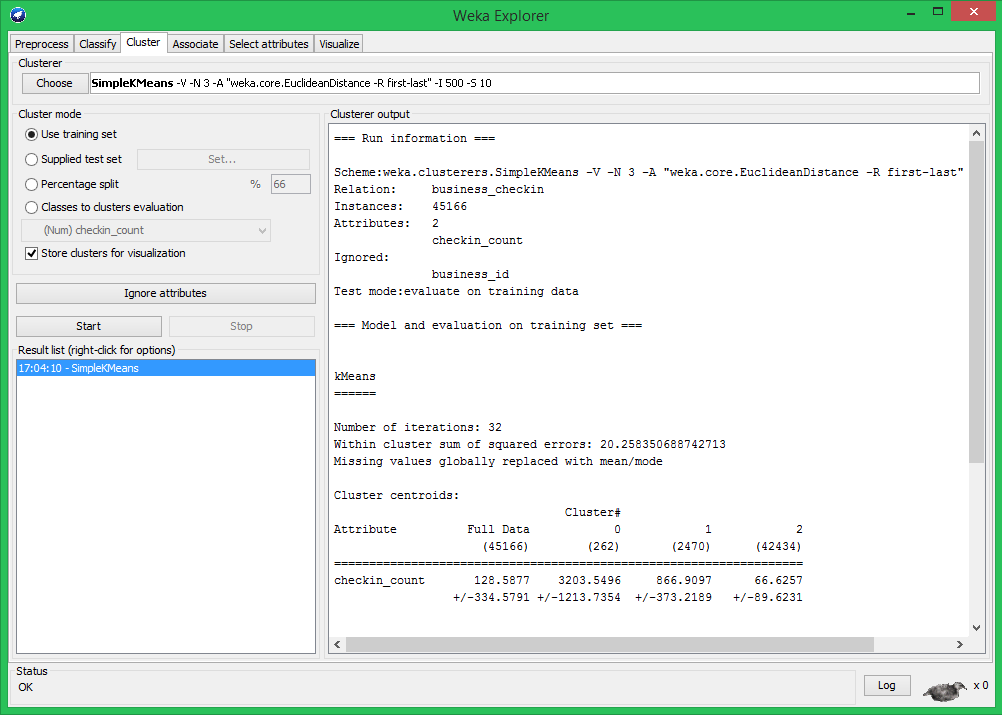
As you can see below we have deduced 3 clusters with mostly evident division between businesses. With a total of 45166 business instances it divided the set into three clusters with 1,5 and 94 percentage of data into cluster 0, 1 and 2

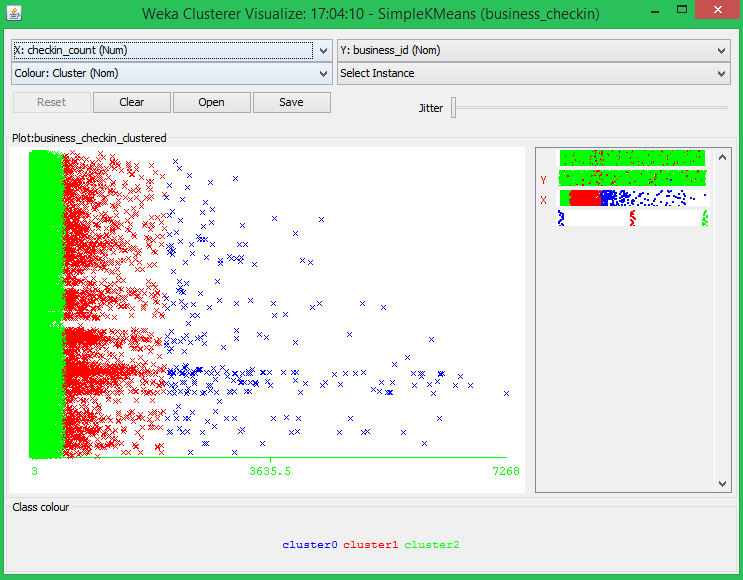
**Clustered Instances**

**0 262 ( 1%)**

**1 2470 ( 5%)**

**2 42434 ( 94%)**





The above picture describes the clustering obtained by the weka tool using Euclidean distance k means algorithm. Cluster 0 points to all the business that get around 2000 or more number of checkins weekly (which of course is much smaller set of businesses) -BLUE

Cluster 1 points to the set with total number of checkins ranging from 470 -2000 – RED

Cluster 2 points to the highest number of businesses, and these get a weekly total checkin at around 1- 470 -GREEN

There are some anomalies in the data set we have received with total number checkin being on a very large scale and we are still looking into it.

**Change of plan since the project proposal:**

Initially we planned to do the data mining task, Classification. But now we have realized that Clustering will be more relevant for this project. So we planned to use the SimpleKMeans method for clustering the aggregated Check-in value into three clusters.

**TASKS TO BE ACCOMPLISHED:**

1. Have to use the CSV file in WEKA  tool to perform clustering and analyze the results
2. A website is to be created to display the restaurants with its price range, parking availability and user’s preference over the restaurant in a bar graph
3. Visualization is to be done using Google charts API.