## Informal Reflection

**Roadblocks encountered**

It was an interesting challenge to take up the Yelp dataset to analyse. Initially we planned to run our analysis on the academic dataset which was much more manageable as compared to the challenge dataset. Our group registered on the Yelp website on July 28 to request the academic data, however our request was kept pending by them and we were not able to access the academic dataset until Aug 10. In the interim we decided to move ahead and do our analysis on the challenge dataset. The challenge dataset from Yelp was relatively large with approximately 1.6 million user review records. We initially decided to read the data from json format to csv with the intention of looking at the data using xls. However xls only supported around 1 million records. We overcame this challenge by reading the data into a dataframe object. We then ran into the issue of the selection of the separator while reading the json file. All different varieties of separators that we tried using created issue while reading the data as this data contained user reviews and these user reviews generally had all kinds of weird and unexpected combination of the special characters. Our group then decided to use a special character from other language to mark as separator and that’s when we decided to use a Chinese character. This was an interesting issue that we overcame. Once we were able to read the data into the DataFrame, we were confident that we could run any kind of analysis on the data.

**Re-framing of analysis questions**

Our first two questions remain the same as per the proposal that we submitted where a) we planned to plot the seasonal trends in the top 10 business categories and b) we planned to find out the controversial stores or restaurants based on the user ratings. However we changed our 3rd question. We initially planned to predict the user rating for the next restaurant visit based on their historical ratings. However since we were not very comfortable with the prediction model that we were planning to use, we changed the question to make it more interesting and the one which could have more business utility. We initially started out with the assumption that we will be using the academic dataset but we ended up using challenge dataset which was much more challenging as compared to academic one. We did not use any external data source for our analysis.

**Coding Issues and Efficiency**

We were not aware of how do we handle such large dataset especially as the review data after merging with the business category data was becoming really huge. We initially planned and split the data by creating 150 different text files so that we could read each one at a time and analyse. We soon figured out that it will be difficult to run aggregate functions on the different set of files and hence we returned to the approach of looking at the complete dataset together.

**Conclusion**

Fusing our python knowledge with a big dataset from Yelp gave us a good exposure of the kind of challenges that are normally encountered while working with bigger datasets. Exploring new libraries and procedures helped us build on our Python knowledge. The trend analysis helped us learn a new concept and draw business relevance from it for the restaurants. Overall, the project helped all the team members to learn effective teamwork, coordination and enhanced our experience handling large datasets.