**CIS 5690: Advanced Systems Project Proposal – Summer 2018**

**Data Analytics with Yelp dataset**

**Problem Statement:**

Yelp is a trending application designed for the users to publish and read the reviews of local businesses such as restaurants, saloons, automobiles workshops etc. Also, it provides online reservations through Yelp reservations. With the broad range online services generating huge data day to day.

The goal is to analyze this unorganized data, design a system to generate insights from the Yelp data which could help the entrepreneurs to study the progress of the business and take wise decisions towards betterment of the business. With the Yelp’s data, the primary task is to derive insights from various reviews from customers, data being acquired from various places across USA.

Below are the few proposed questions originated after keen observation from the dataset which will be solved to develop business insights for the owners:

1. Finding the day and time in a week receiving highest number of check-ins.
2. Determining the top 50 reviewed businesses.
3. Determining the cities playing major role in making business.
4. Determining cities with highest reviews and best ratings for their businesses.
5. Determining cities with greater than 50k reviews ranked by average star ratings.
6. To determine frequently given tips(feedback) from customers.
7. Finding the closed businesses which are registered but closed
8. Taking one business with highest rating and one business with lowest rating to compare the tips recorded by the users in both the cases to compare feedback given by the users.

With the above listed factors, an existing business owner or a new entrepreneur can look at the trends solved which can help in expansion of business or establishing a new business body considering the above factors.

After deriving solutions for the above proposed questions, the data will be visualized to the end-users in an easy and captivating graphs and pie-charts for precise understanding.

To achieve these tasks, the following technologies/tools will be used.

**Technologies/Tools:**

1. **Machine Learning algorithms:** Various ML algorithms will be used to clean the data based on the determination criteria.
2. **Python**
3. **Jupyter Notebook:** IDE for Python.
4. **Dataset:** From Kaggle and Yelp official website.