**Capstone Project Submission**

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **Souvik Bhattacharyya**  Email. id- souvik.ranchi@gmail.com  Contribution role:  Data Loading (read data from the drive)  Data Exploration  Data Preprocessing  Data Analysis and Visualization  Implementing Elbow Method for finding the K value.  K- Means Clustering.  Modeling and Comparison. |
| **My GitHub link is below:** |
| GitHub Link: - <https://github.com/Souvik0651/Zomato-Restaurant-Clustering-and-sentiment-Analysis.git> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)**  **Project Summary**  We are provided with two datasets; one is the metadata dataset and the other one is the Reviews dataset.  The Project focuses on Customers and companies, we have to analyze the sentiments of the reviews given by the customer in the data and make some useful conclusions in the form of Visualizations. Also, cluster the zomato restaurants into different segments. The data is visualized as it becomes easy to analyze data in an instant. The Analysis also solve some of the business cases that can directly help the customers find the best restaurant in their locality and for the company to grow up and work in the fields they are currently lagging in.  At first, we performed Data cleaning by creating a data frame and removing unnecessary columns with null values, dropping the duplicate rows in this dataset, and started getting basic insight from the cleaned dataset.  We have performed some text preprocessing on some columns to make them efficient for us to implement our clustering models, and also performed visualization to get basic insights on the features.  We went with feature conversion in which the object datatype features are converted to numeric features in our dataset.  After feature selection, we tend to split the data into clusters and started implementing basic Machine Learning Models for clustering.  Sentiment Analysis was done on the reviews and a model was trained in order to identify negative and positive sentiments. Even though the number of false negatives is higher in the case of Logistic Regression than in Random Forest, it is performing better in terms of reducing False positives. |