**Assignment #3: Latent Manifold and Regression**

**Team Name:** Codebusters

**Team Members:**

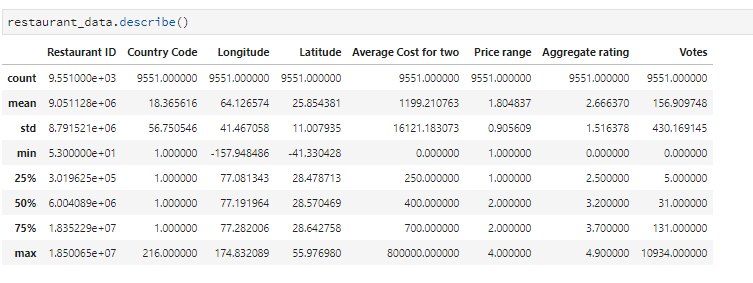
|  |  |  |
| --- | --- | --- |
| **Student Name** | **Student ID** | **Contributions** |
| Anushri Srinath Aithal | 012506897 | Data Cleaning, Data enrichment |
| Harini Balakrishnan | 010830755 | Linear Regression |
| Ravi Katta | 012127011 | KMeans Clustering and Latent Manifold |
| Sunder Thyagarajan | 011528062 | GMM and Latent Manifold |

**Dataset:** [**https://www.kaggle.com/shrutimehta/zomato-restaurants-data**](https://www.kaggle.com/shrutimehta/zomato-restaurants-data)

**Objective:**

The purpose of this assignment is to understand the ML life-cycle and apply Regression and identify the latent manifold in the dataset. The overview of the data set is as below

1. The dataset is from Zomato and has restaurant information. The dataset description is as below



1. The list of problems that can be resolved from this dataset are
   * 1. Identify the best dining place in a city
     2. Identify the popular cuisine
     3. Predict the popular times, helps restaurants with better labor management.
     4. Predict Revenue for the next month.
     5. Predict the location of the next restaurant branch.
     6. Predict the meal price for two and enable dynamic pricing.
     7. Recommendation on where to eat.

**Data Preparation:**

As a part of the data preparation phase the below steps were followed:

1. Load the data using the Pandas library.
2. The dataset contains 57% data from India and most of which are from the New Delhi region. So, we are filtering the dataset to consider restaurants from New Delhi, Gurgaon, and Noida.
3. Remove data with null values.
4. Remove data where the average cost for 2 is 0.
5. Remove data where there are no ratings available.
6. Remove data with price range less than or equal to 0
7. Convert the features Rating color, Rating text, City and Cuisine to categorical values
8. Plot the Rating Vs Average cost for two as below

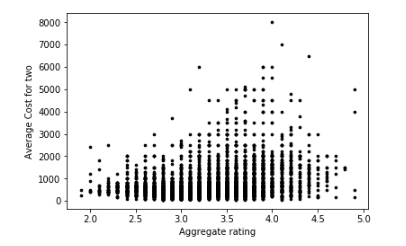


Fig.1 Dataset points

**Identify the number of clusters:**

Elbow method is used to identify the number of clusters. From the below graph it is evident that we have 3 clusters.

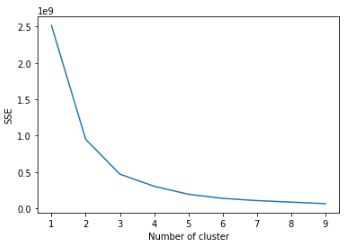


Fig.2 Graph indicating the number of clusters

**K-Means Clustering:**

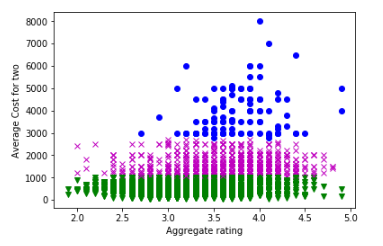


Fig.3 K-Means Clustering

**GMM:**

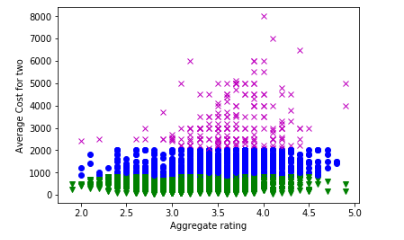


Fig.4 GMM

**Linear Regression:**

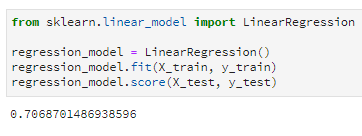


Fig.5 Model score of 70%

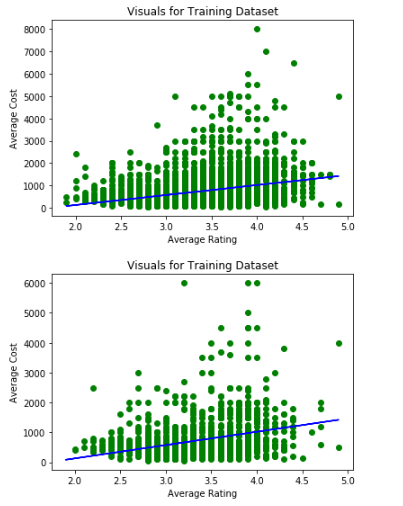
****

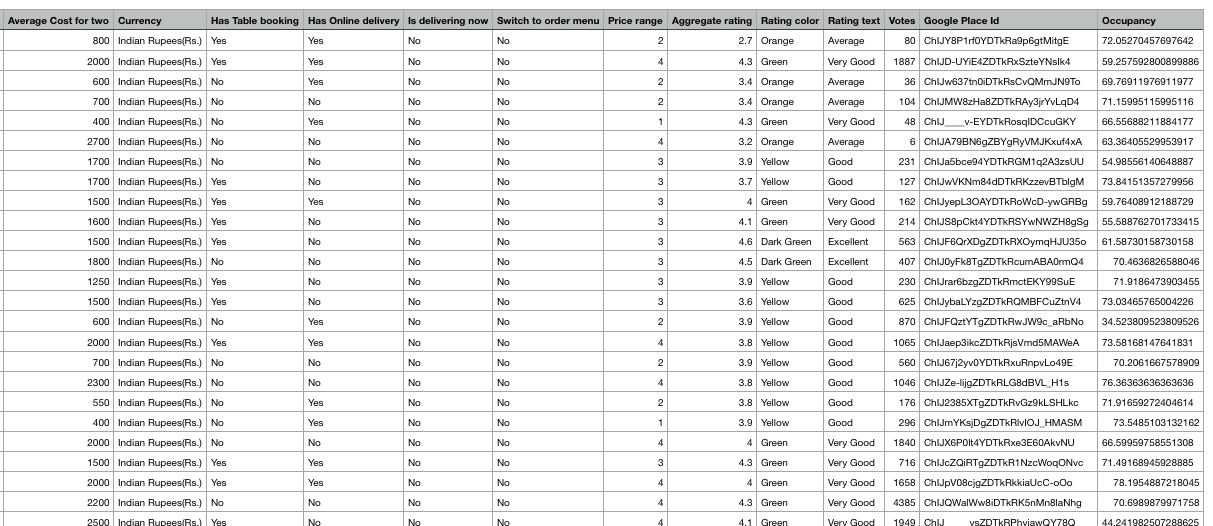
Fig.6 Linear Regression

**Data Enrichment:**

The business problem we are trying to solve using this dataset is monthly revenue prediction for restaurants. As the dataset is missing revenue details, it can be derived by enriching this dataset. The idea here is to get the weekly occupancy score from Google popular times API and then calculate the monthly revenue as a function of a number of seats in a restaurant and average meal cost for two.

As a part of data enrichment, we have populated the weekly occupancy score(percentage). To do so below steps are followed

1. Using Google’s find a place from text API, find the google place id for the restaurants in base data
2. For those restaurants where google place ID is present, using the place ID and popular times crawler (<https://github.com/m-wrzr/populartimes> ) get the popular times which is a weighted score representing how busy a restaurant is for each hour of the day for an entire week.
3. Sum the popular times for each day.
4. Find the average popular time in a week by dividing the max popular time from all values.
5. Find the percentage of popular times which represents the weekly occupancy score.
6. The newly found occupancy and place ID are added to base data and a new file output\_enrich\_data.csv is generated with the enriched data.

As revenue is not just a function of occupancy but also the number of seats in a restaurant, we have not been able to calculate the actual revenue for this dataset. Discussing with Professor on how to take this further.

**Latent Manifold:**

The latent variables that we identified for our problem set are as below

1. Yearly/Monthly Revenues of a Restaurant
2. Ingredients for their most popular dish
3. Current popular times which helps in dynamic pricing
4. Weather as weather influence the customers and dynamic pricing can vary on this factor
5. Location type – proximity to business centers, shopping complex, school district, and residential areas
6. Population density, age group, and average salaries
7. Food prices depend on fuel cost, agricultural costs these hidden cost help in predicting monthly revenue.
8. Number of seats in a restaurant