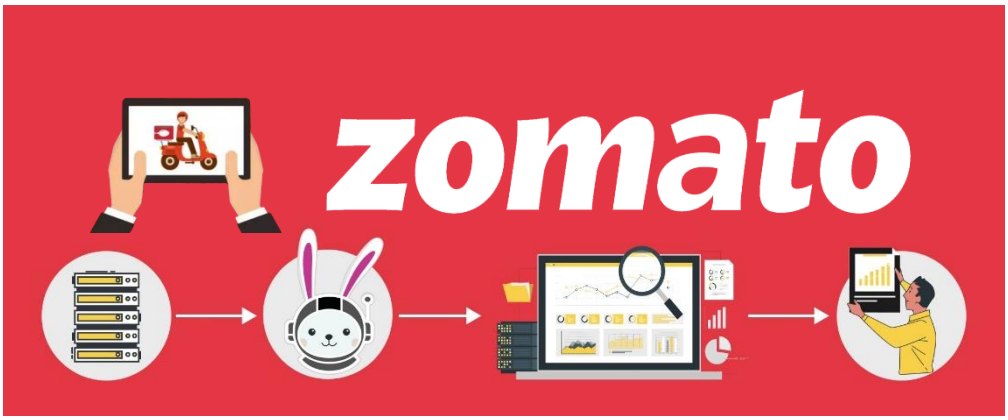
Symbiosis Skills and Professional University Kiwale, Pune

Project Report
On

Zomato Data Analysis and Recommendation on Map



SUBMITTED BY

Arshad Sheikh
Tahir Maner
Neha Rathod

REGISTERED BATCH: FST 2.0 – ML 2

UNDER THE GUIDANCE OF

Prof. Amrita Helwade

STUDENT DECLARATION AND ATTESTATION BY TRAINER

This is to declare that this report has been written by us. No part of the report is plagiarized from other sources. All information included from other sources have been duly acknowledged. I aver that if any part of the report is found to be plagiarized, I shall take full responsibility for it.

Name Of Student **Signature**

Arshad Sheikh

Tahir Maner

Neha Rathod

Signature of trainer

Prof. Amrita Helwade

Certificate

This is to Certify that, **Tahir Maner** has completed and submitted the project entitled, **"Zomato Data Analysis and Recommendation on Map"** under the guidance of Amrita ma'am, to Symbiosis Skills and Professional University, Pune, Maharashtra, India, is a record of Bonafede project work carried out by them and is worthy of consideration for the completion of certificate course in "Machine Learning".

Date: 17/09/2022

Signature of trainer **Supervisor**

Prof. Amrita Helwade

ACKNOWLEDGEMENT

We are profoundly grateful to trainer Prof. Amrita for her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

We would like to express our deepest appreciation towards SYMBIOSIS SKILLS & PROFESSIONAL UNIVERSITY Prof. Baliram whose invaluable guidance supported us in completing this project and also JP Morgan for funding and giving us this opportunity.

At last, we must express our sincere heartfelt gratitude to all staff members of Computer Science & Engineering Department who helped us directly or indirectly during this course of work.

Thanking you all…!!!

INDEX

ABSTRACT

Zomato API Analysis is one of the most useful analysis for foodies who want to taste the best cuisines of every part of the world which lies in their budget. This analysis is also for those who want to find the value for money restaurants in various parts of the country for the cuisines. Additionally, this analysis caters the needs of people who are striving to get the best cuisine of the country and which locality of that country serves those cuisines with maximum number of restaurants.

Introduction

Zomato was founded as Foodie Bay in 2008 by Deepender Goyal and Pankaj Chaddah who worked for Bain & Company. They renamed the company Zomato in 2010 as they were unsure if they would "just stick to food" and also to avoid a potential naming conflict with eBay.

In 2011, it expanded across India to Delhi NCR, Mumbai, Bangalore, Chennai, Pune, Ahmedabad and Hyderabad. In 2012, it expanded operations internationally in several countries, including the United Arab Emirates, Sri Lanka, Qatar, the United Kingdom, the Philippines, and South Africa. In 2013, expanded to in New Zealand, Turkey, Brazil and Indonesia, with website and apps available in Turkish, Portuguese, Indonesian and English languages. In April 2014, it launched in Portugal, which was followed by launches in Canada, Lebanon and Ireland in 2015.

In January 2015, Zomato acquired Seattle-based restaurant discovery portal Urban spoon, which led to the firm's entry into the United States and Australia. . In an effort to expand its business beyond restaurant listing, Zomato piloted an online payments facility in partnered restaurants in Dubai called Zomato Cashless in February 2015. This was discontinued a few months later.

Zomato started its food delivery service in India in 2015, initially partnering with companies such as Delivery and Grab to fulfill deliveries from restaurants that did not have its own delivery service. In April 2015, Zomato acquired the American online table reservation platform Next able, which was subsequently renamed as Zomato Book. In January 2016, it launched Zomato Book's table reservation facility on its application in India. In April 2015, Zomato acquired cloud-based point of sale company Maple Graph Solutions, and in April 2016, it launched its own Android point of sale system for restaurants called Zomato Base, which was built on Maple OS.

Motivation:

We really get fascinated by good quality food being served in the restaurants and would like to help community find the best cuisines around their area. Which led us to find better option for finding new options and their location. It also helps us to determine better location for new business settlement and their nature and Cuisines.

Problem Statement:

To analyze Zomato Data and built Recommendation system based on analysis.

Objective:

1)You are in the team in charge of developing the interface for the food app. you are tasked with creating something that enables #user to choose the category of the food. you have no idea what categories to use while designing the app. how can this dataset #help you solve this issue?

2)You are hearing rumors that people in India passionately dislike the restaurants offered in the app. find out whether this #rumor has any substance to it.

3)You are one of the developers working in the food app company. you decide to quit and follow your dreams of opening a restaurant. using this data how will you go about increasing your chances of succeeding in this highly competitive business?

METHODOLOGIES OF PROBLEM SOLVING

1) Algorithm:

a) Linear Regression:

Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of variables, one variable is called an independent variable, and the other is a dependent variable. The range of the coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data between two variables.

b) Decision Tree:

Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision trees are composed of three main parts— decision nodes (denoting choice), chance nodes (denoting probability), and end nodes (denoting outcomes). Decision trees can be used to deal with complex datasets, and can be pruned if necessary to avoid overfitting.

c) Random Forest:

Random forest is solid choice for nearly any prediction problem (even non-linear ones). It's a relatively new machine learning strategy (it came out of Bell Labs in the 90s) and it can be used for just about anything. It belongs to a larger class of machine learning algorithms called ensemble methods.

d) KNN:

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well –

(i) Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.

(ii) Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn’t assume anything about the underlying data

e) Naive Bayes:

Naïve Bayes algorithms is a classification technique based on applying Bayes' theorem with a strong assumption that all the predictors are independent to each other. We have the following three types of Naïve Bayes model under Scikit learn Python library –

(i) Gaussian Naïve Bayes: It is the simplest Naïve Bayes classifier having the assumption that the data from each label is drawn from a simple Gaussian distribution.

(ii) Multinomial Naïve Bayes: Another useful Naïve Bayes classifier is Multinomial Naïve Bayes in which the features are assumed to be drawn from a simple Multinomial distribution. Such kind of Naïve Bayes are most appropriate for the features that represents discrete counts.

(iii) Bernoulli Naïve Bayes: Another important model is Bernoulli Naïve Bayes in which features are assumed to be binary (0s and 1s). Text classification with 'bag of words' model can be an application of Bernoulli Naïve Bayes in this project we had implemented Gaussian Naïve Bayes.

f) Support Vector Machines:

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and it can be done in the following step − First, SVM will generate hyperplanes iteratively that segregates the classes in best way. Then, it will choose the hyperplane that separates the classes correctly.

2) Hypothesis Testing:

The Pearson's Chi-Square statistical hypothesis is a test for independence between categorical variables. In this article, we will perform the test using a mathematical approach and then using Python's SciPy module. We start by defining the null hypothesis (H0) which states that there is no relation between the variables. An alternate hypothesis would state that there is a significant relation between the two. If our calculated value of chi-square is less or equal to the tabular (also called critical) value of chi-square, then H0 holds true.

3) Roc Curve:

An ROC curve shows the relationship between sensitivity and specificity for every possible cutoff. The ROC curve is a graph with:
The x-axis showing 1 – specificity (= false positive fraction = FP/(FP+TN))
The y-axis showing sensitivity (= true positive fraction = TP/(TP+FN))
Thus, every point on the ROC curve represents a chosen cut-off even though you cannot see this cut-off. What you can see is the true positive fraction and the false positive fraction that you will get when you choose this cut-off.

4) Libraries Used:

a) **Pandas**: It provides fast, expressive, and flexible data structures to easily (and intuitively) work with structured (tabular, multidimensional, potentially heterogeneous)
b) **NumPy**: It has advanced math functions and a rudimentary scientific computing package. NumPy is a popular array – processing package of Python. It provides good support for different dimensional array objects as well as for matrices.
c) **Matplotlib**: Matplotlib helps with data analyzing, and is a numerical plotting library. Matplotlib can create such quality figures that are really good for publication. Figures you create with Matplotlib are available in hardcopy formats across different interactive platforms.
d) **Seaborn**: It provides a high-level interface for drawing attractive and informative statistical graphics.
e) **Sk-learn**: It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.
f) **Scipy**: The SciPy library, a collection of numerical algorithms and domain specific toolboxes, including signal processing, optimization, statistics, and much more. Matplotlib, a mature and popular plotting package that provides publication-quality 2-D plotting, as well as rudimentary 3-D plotting.
g) **Folium**: folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map. The library has a number of built-in tile sets from OpenStreetMap, Mapbox, and Stamen, and supports custom tile sets with Map box or Cloud made API keys. folium supports both Image, Video, GeoJSON and TopoJSON overlays.

SOFTWARE REQUIREMENT SPECIFICATIONS

1) **Assumptions And Dependences:**
   a) **Assumptions:**
      i) The end user device should be a laptop.
      ii) Additionally, the end user has an active internet connection in his/her laptop.
   b) **Dependencies:**
      i) The system browser is dependent on the end user device.
      ii) The prediction and analysis purpose are dependent on the types of algorithms used.

2) **Performance Requirements**:
   a) **Accuracy:** The system can predict with varying accuracy between 50 to 60% using one of the Algorithms c which gives maximum accuracy right now, but later on, as the number of responses will increase the accuracy will also increase.
   b) **Privacy:** Data will be totally secured and will not be leak as no personal details are asked.

3) **System Requirements:**
   a) **Database Requirement:**
      i) MS-Excel
   b) **Software Requirement:**
      i) Jupyter Notebook
      ii) Any Desktop Operating System
      iii) Programming Language- Python.
   c) **Hardware Requirement:**
      i) Any Device with Brower Support.

PROJECT IMPLEMENTATION

1) Overview Of Project Module:

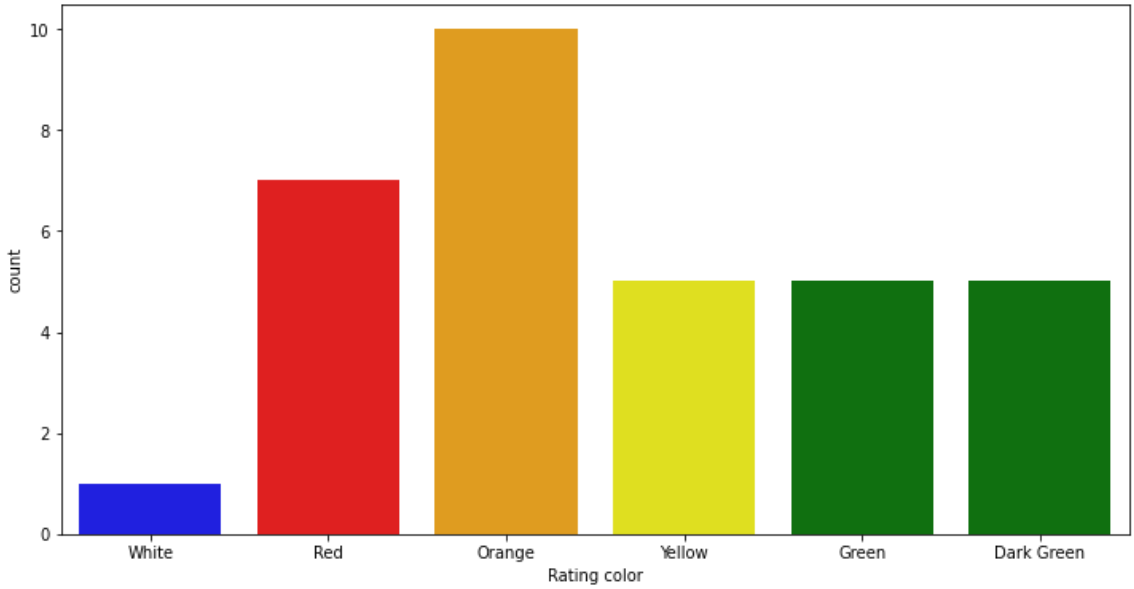
The project is comprised of following steps:

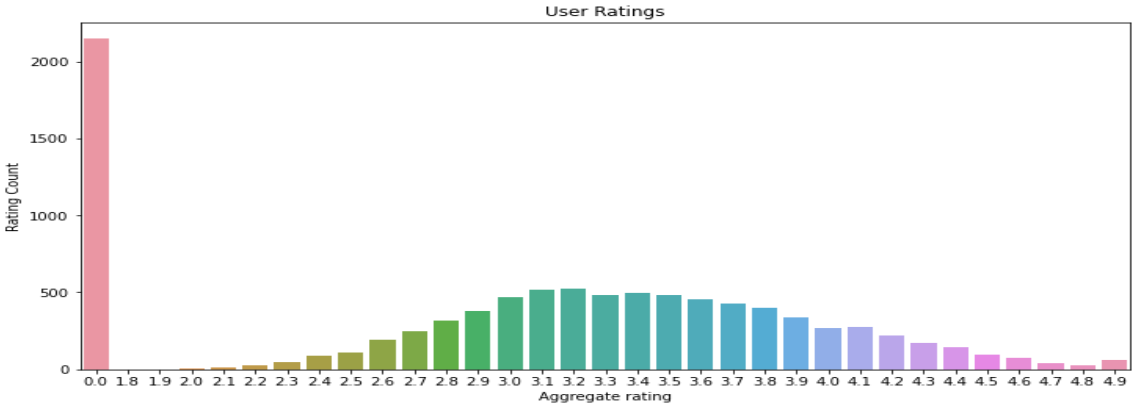
- Collection of data
- Importing data into python
- Data cleaning
- Performing exploratory data analysis
- Hypothesis testing
- Feature engineering
- Implementation of algorithms
- Create Recommendation
- Plot on Maps

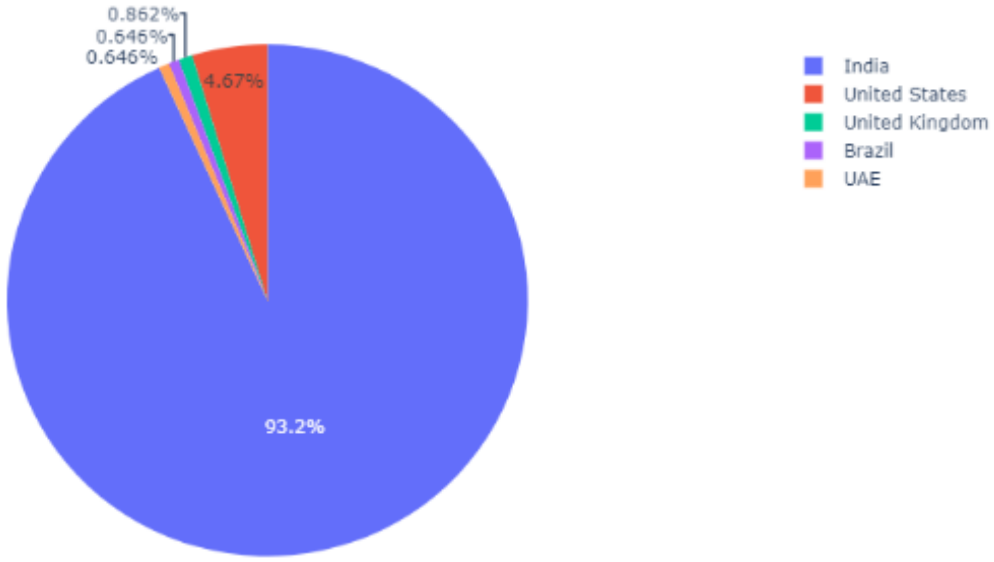
2) Data Flow Diagram:

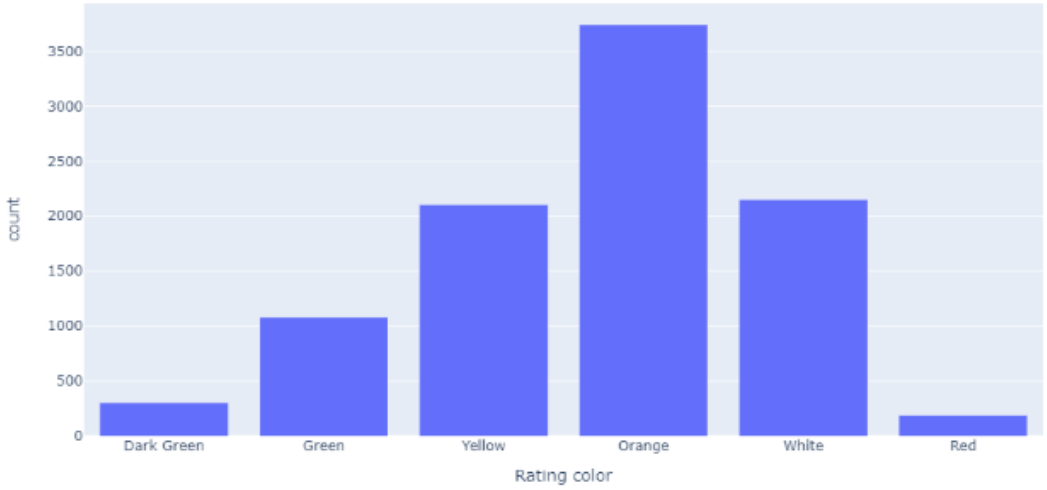
A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system. DFD is a preliminary step to create overview of system.

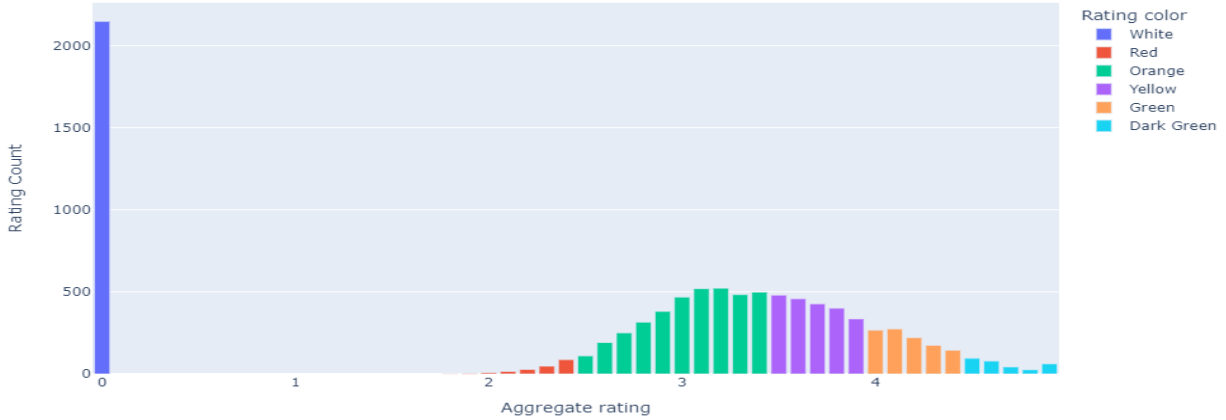
- **Collection of data:** We have taken the data collected on the cloud from sever. Data has been collected based on the order place and there.
- **Importing the data and cleaning the data:** All the implementation is done in python programming language on jupyter notebook. The dataset was imported using pandas library. Head, tail and shape of data was checked. Columns of the dataset were checked. Since we used to google forms, times stamp column was also added. Hence, we removed that column. Column names were questions from google forms, hence it was replaced by question number for better understanding. There were some null values in data set too, hence it was also replaced with not prefer to say.
- **Performing exploratory data analysis:** Before doing data exploration, we converting the categorical data set into numeric data set. After doing this, we performed data visualization techniques. Plotted each graph of every question to find the spread of the data and to check measure of tendency. We also plotted boxplots and density plots too. Group by was also done check which gives clearer picture.

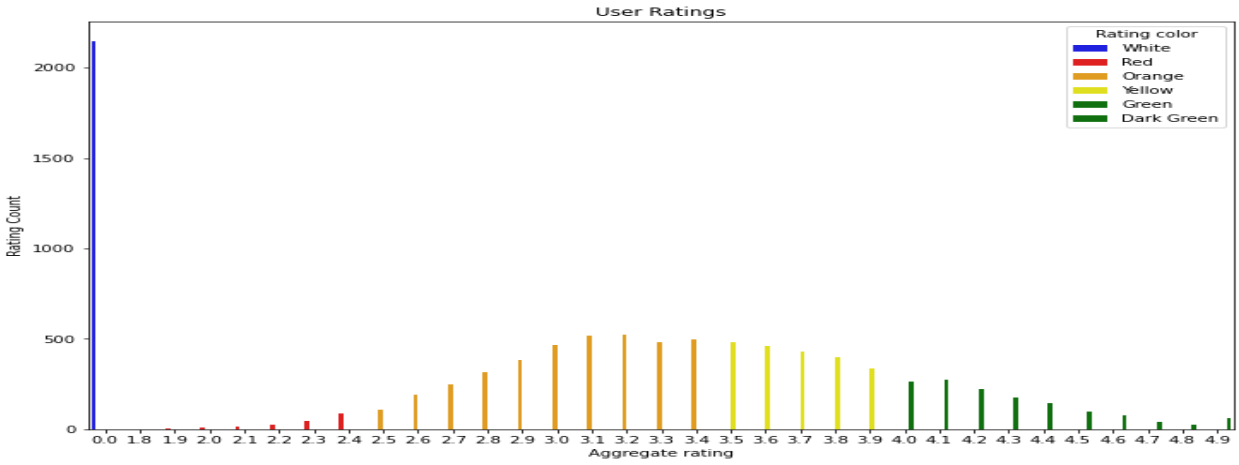

count
Rating color
White
Red
Orange
Yellow
Green
Dark Green

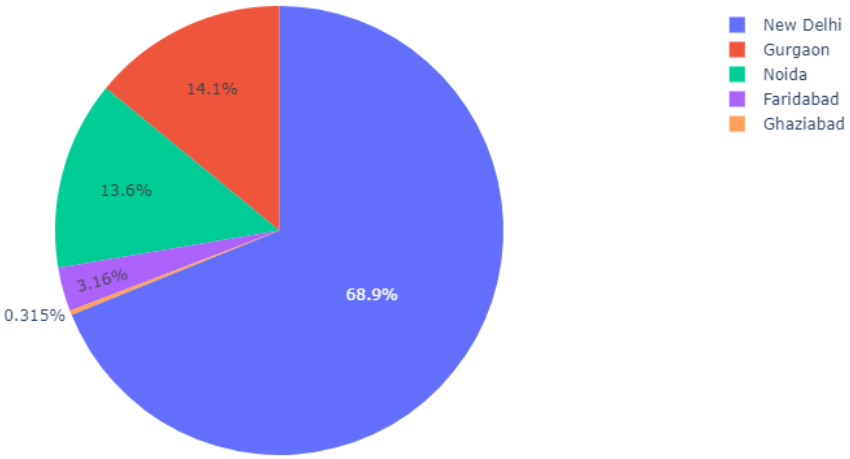

User Ratings
Rating Count
Aggregate rating

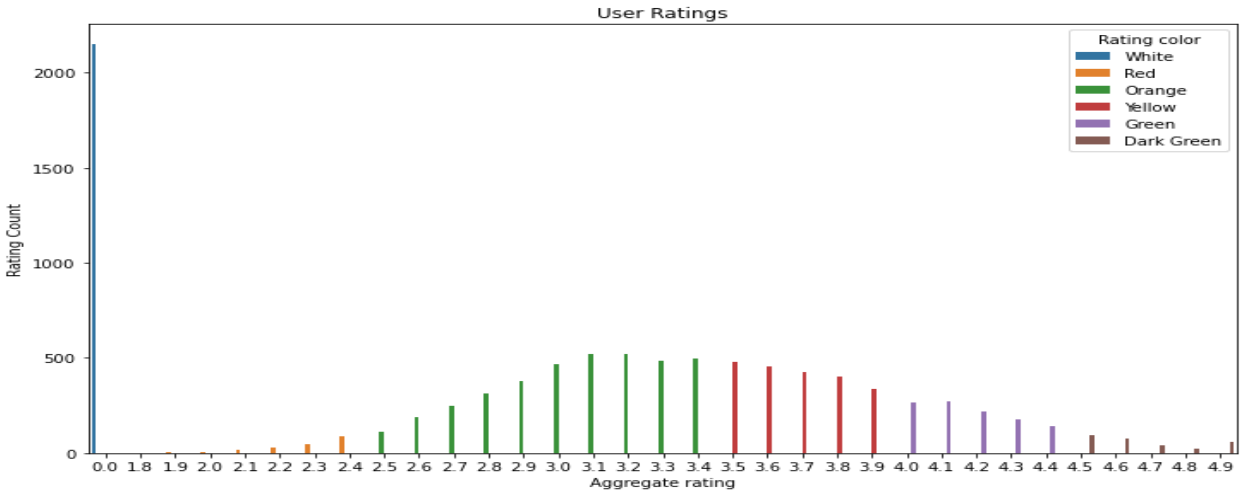

0.862%
0.646%
0.646%
4.67%
93.2%
India
United States
United Kingdom
Brazil
UAE

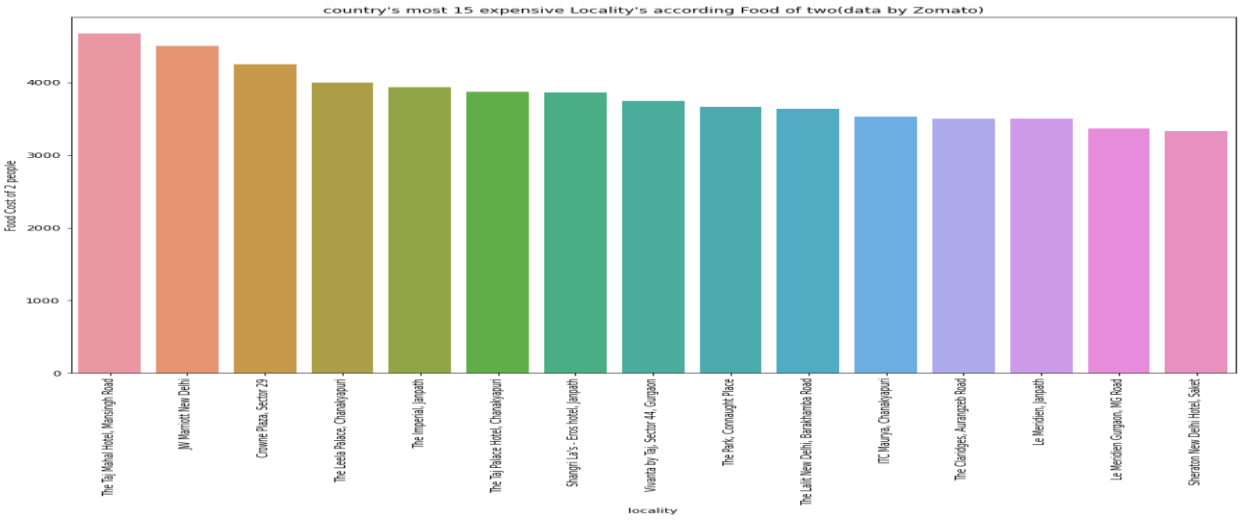

count
3500
3000
2500
2000
1500
1000
500
0
Dark Green
Green
Yellow
Orange
White
Red
Rating color

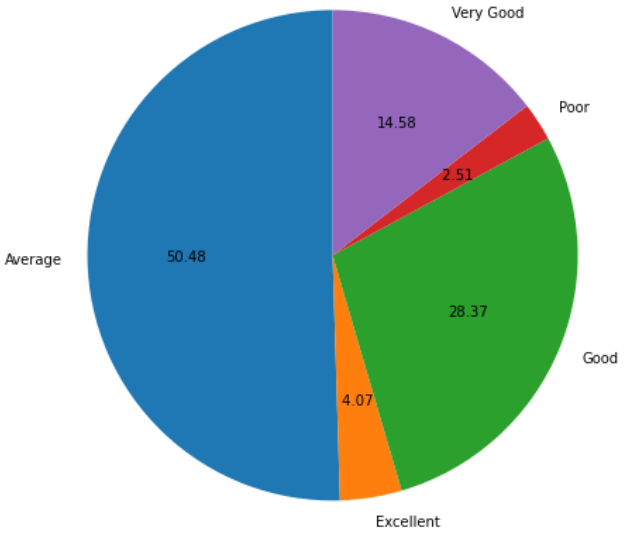

Rating Count
2000
1500
1000
500
0
0
1
2
3
4
Aggregate rating
Rating color
White
Red
Orange
Yellow
Green
Dark Green


User Ratings
Rating Count
2000
1500
1000
500
0
0.0 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 4.0 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9
Aggregate rating
Rating color
White
Red
Orange
Yellow
Green
Dark Green


New Delhi
Gurgaon
Noida
Faridabad
Ghaziabad
14.1%
13.6%
3.16%
0.315%
68.9%


User Ratings
Rating color
White
Red
Orange
Yellow
Green
Dark Green
Rating Count
Aggregate rating


country's most 15 expensive Locality's according Food of two(data by Zomato)
Food Cost of 2 people
locality


Very Good
Poor
Average
Good
Excellent
14.58
2.51
50.48
28.37
4.07

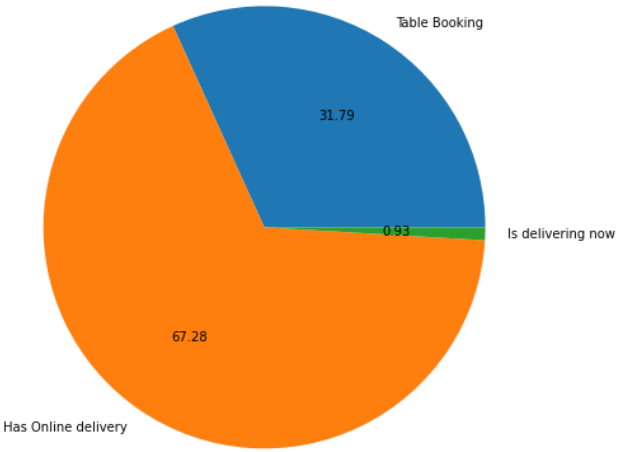
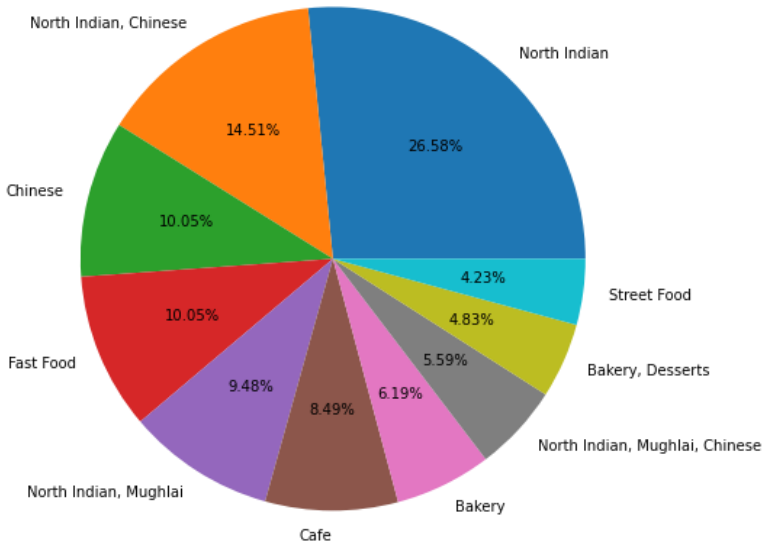
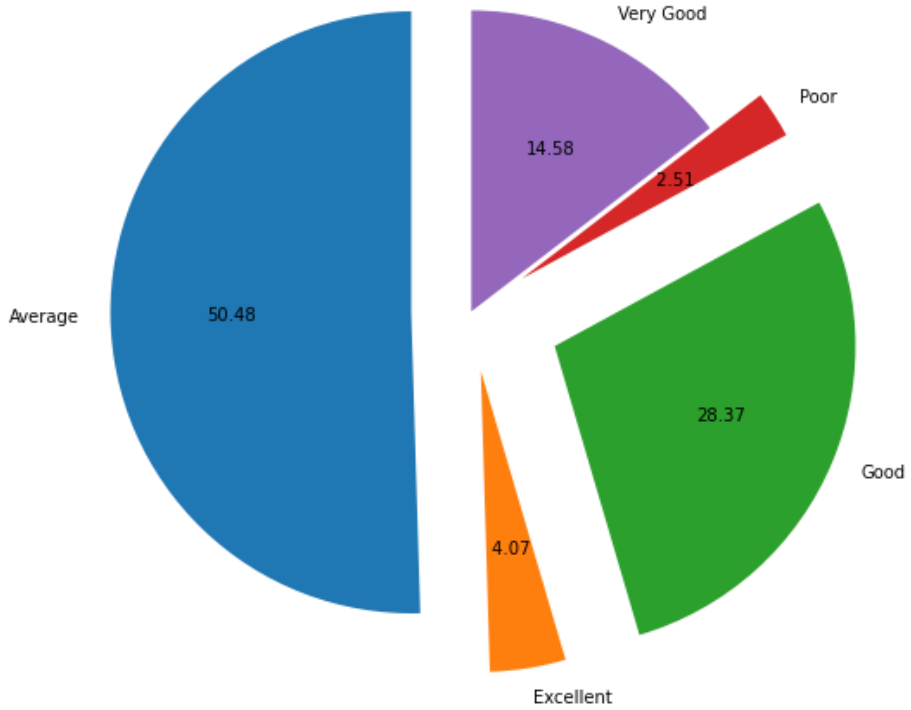
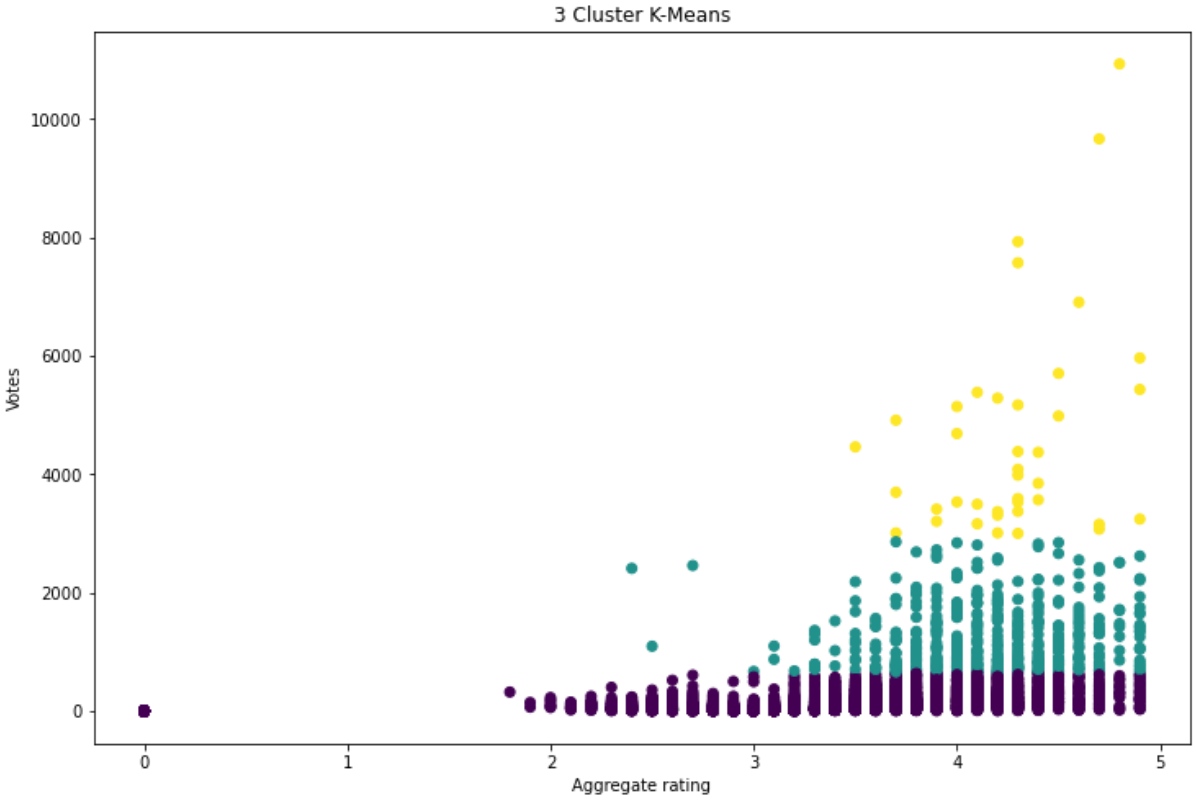


Table Booking
Is delivering now
Has Online delivery
31.79
0.93
67.28


North Indian, Chinese
North Indian
Chinese
Street Food
Bakery, Desserts
Fast Food
North Indian, Mughlai, Chinese
North Indian, Mughlai
Bakery
Cafe
14.51%
26.58%
10.05%
4.23%
4.83%
10.05%
5.59%
9.48%
6.19%
8.49%


Very Good
Poor
14.58
2.51
Average
50.48
28.37
Good
4.07
Excellent


3 Cluster K-Means
Votes
10000
8000
6000
4000
2000
0
0
1
2
3
4
5
Aggregate rating

3) Feature engineering:

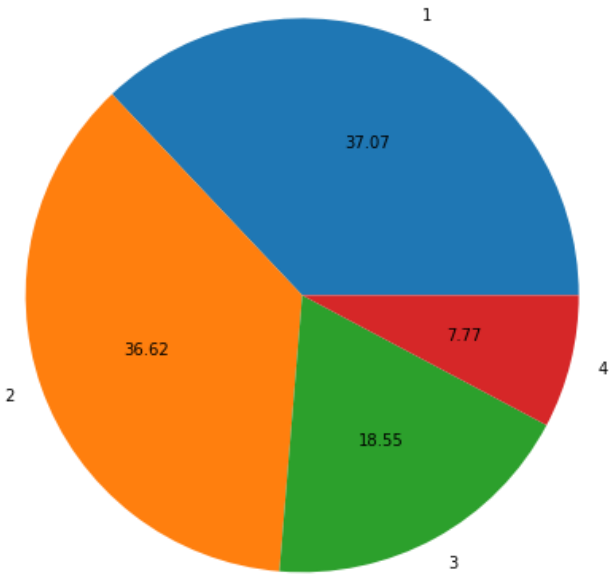
After knowing which were the independent and dependent variables, feature engineering techniques were used for X and y. it is important step as the accuracy will be dependent on X and y variable. It was trial and error method because to get the highest accuracy we needed to check the dependency.

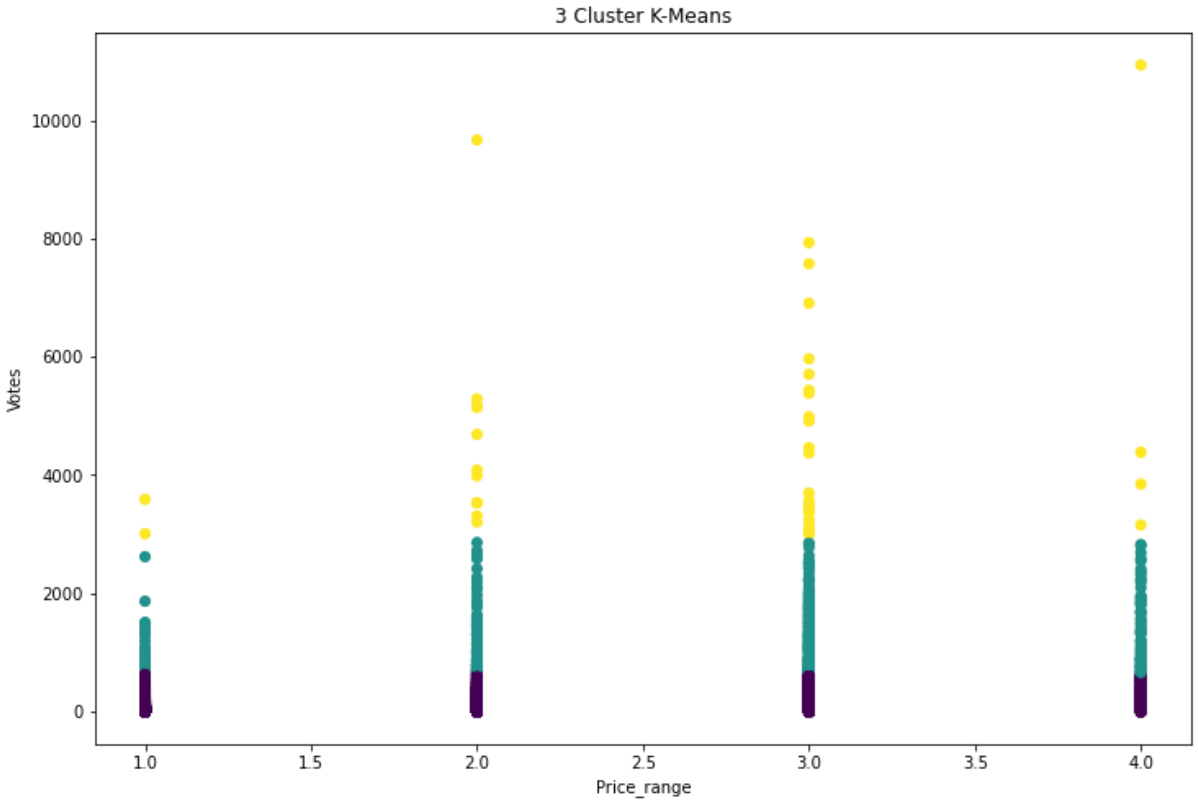
4) Implementation of algorithms:

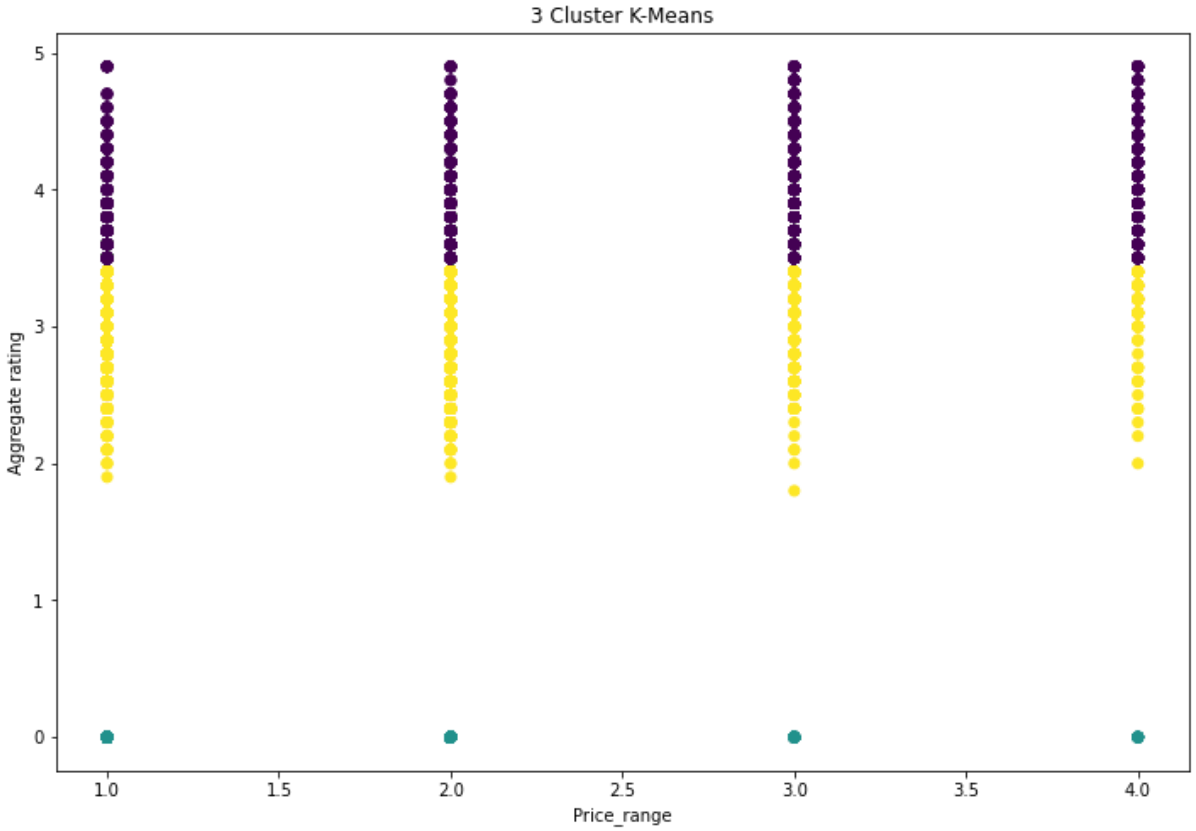
With the data defined, K-means Clustering was used to plot graph on which further analysis was going to be proceed. Clustering was performed on two different bases for better understanding and detail view of different aspect affecting outcomes or expected result. With data clustered together we get better idea how data is associated with other factors. With these, recommendation can be created.

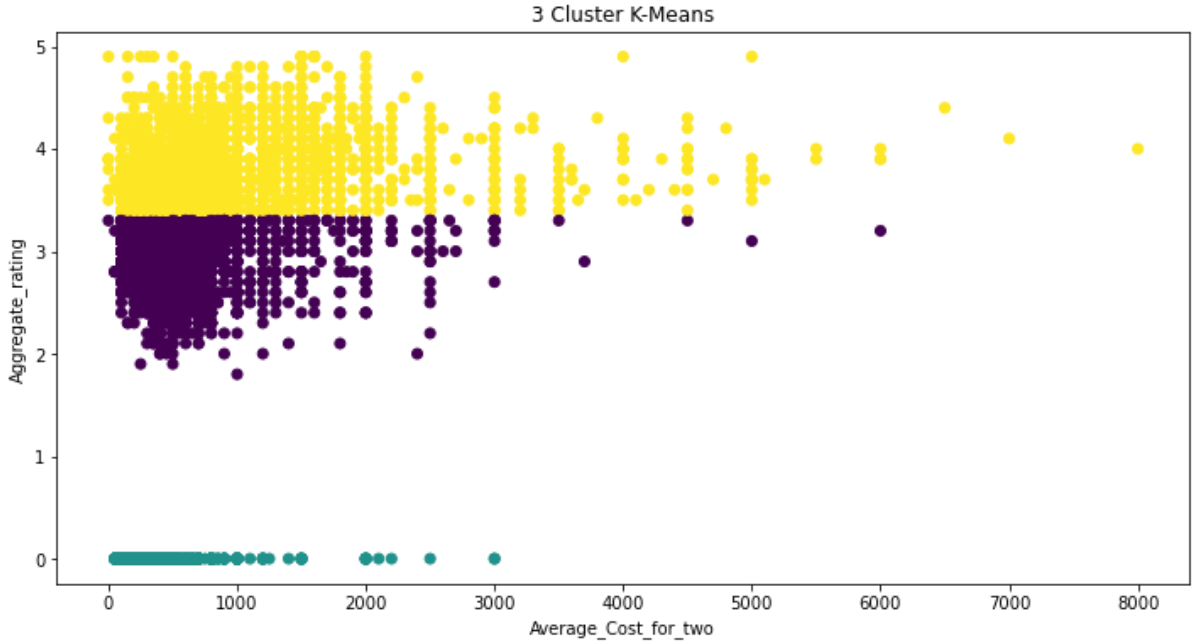
5) Implementation of Map Plotting:

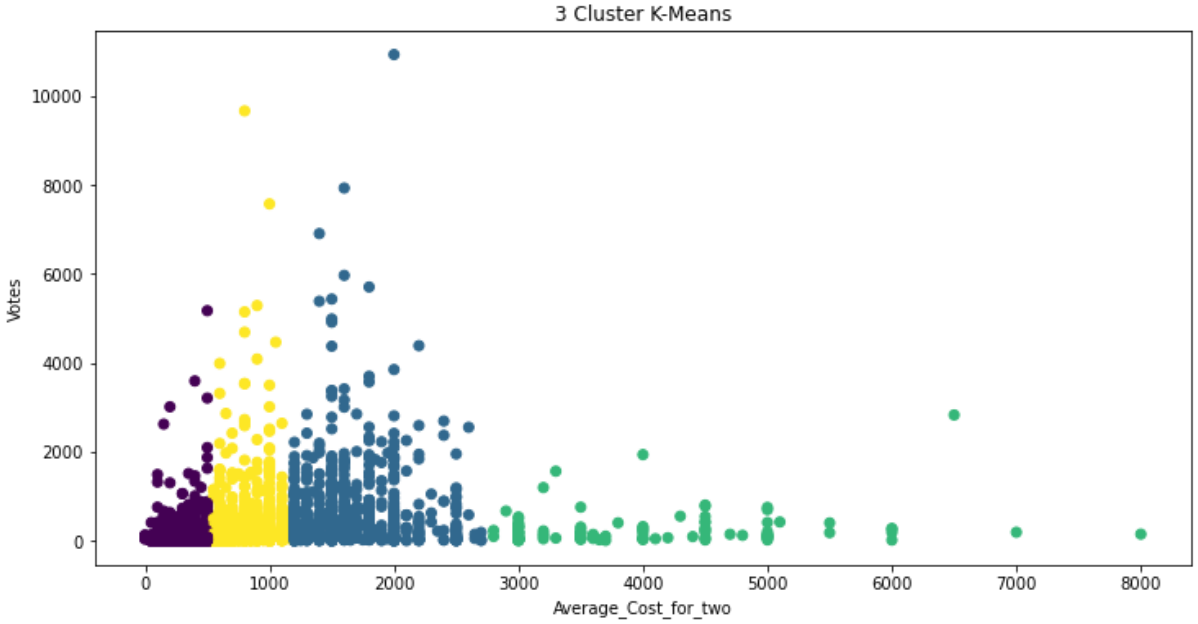
From recommendation, output come as restaurant location and their details. Form these maps can be plot for better view and understanding of the location nearby the end user.

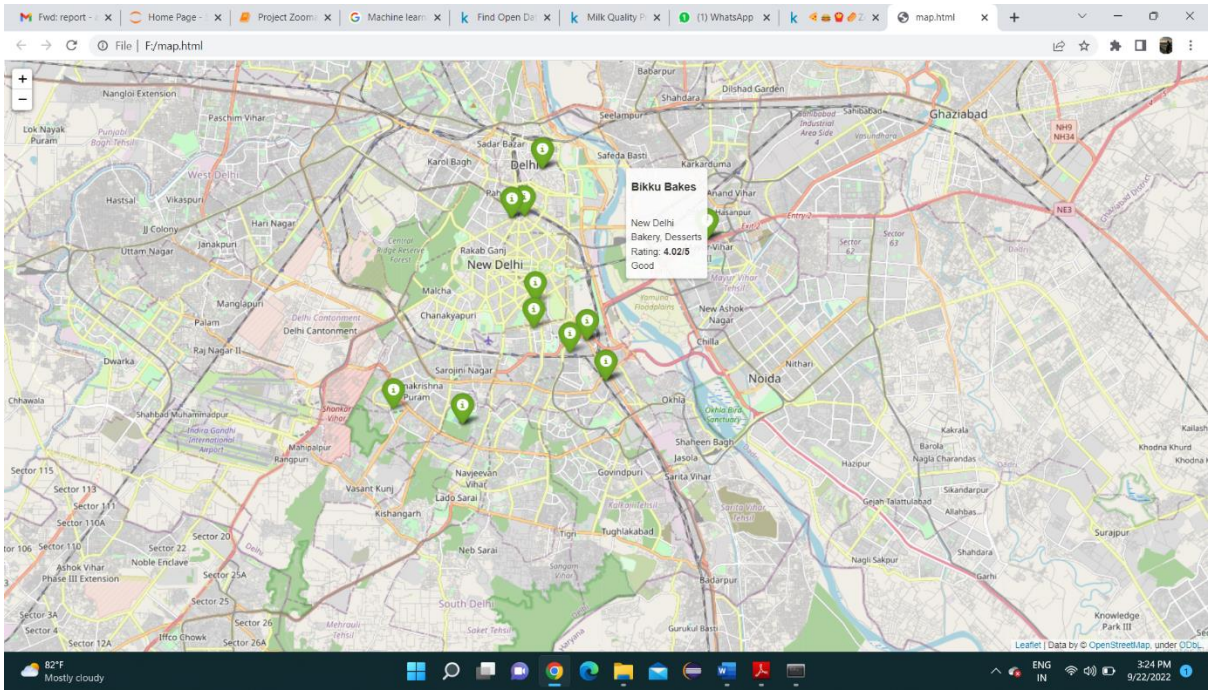



3 Cluster K-Means
Votes
Price_range
0
2000
4000
6000
8000
10000
1.0
1.5
2.0
2.5
3.0
3.5
4.0


3 Cluster K-Means
Aggregate rating
Price_range
0
1
2
3
4
5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


3 Cluster K-Means
Aggregate_rating
Average_Cost_for_two
0
1
2
3
4
5
0
1000
2000
3000
4000
5000
6000
7000
8000


3 Cluster K-Means
Votes
Average_Cost_for_two
0
2000
4000
6000
8000
10000
0
1000
2000
3000
4000
5000
6000
7000
8000



Map For Recommended Restaurant.

6) Advantages:

- This will help us to track peoples ordering and their food liking.
- Using data science, we can help to setup new business in better location and help to improve already existing ones.
- It gives better recommendation with the ease of seeing them on map for better determining and limiting their search time.
- It helps to understand customer liking and their choices based on that.

7) Limitations:

- Currently the accuracy isn't satisfactory but as the data will increase it will also increase.
- The model can't be 100% accurate, it has some limitations too.

8) Application:

For commercial use,

- It gives refresh view on peoples spending on food and their liking, which can be used for both new settlement or existing businesses.
- Used to develop new app based on these.

9) Future scope:

- By giving front-end to module, user end can contribute to data.
- Currently in front end we aren't able to select multiple options but soon we'll able to do that too.
- We can create new app based on this analysis, and can launch as commercial purpose.

10) Conclusion:

- We covered the analysis, where most people order and how the prefer ordering nowadays.
- We found pattern in this analysis.
- Checked the data distribution based on city wise and country wise.
- Based on their ordering behavior, recommendation can be employed for more business and commercial purpose.

Thank You