

**AMITY UNIVERSITY**

-------------------Maharashtra-------------------

**Project Report on**

# Zomato Exploratory Data Analysis

Submitted to

**Amity Institute of Information Technology**

**In partial fulfillment of the requirements for the award of the degree of**

B.Sc. (IT)

**Submitted to Submitted By**

Dr. Preeti Gupta Shivangi Prasanna Koltharkar

Enroll No: A71004917003

Batch: 2017-2020

### DECLARATION BY STUDENT

I, **Shivangi Prasanna Koltharkar** student of B.Sc. (IT) hereby declare that the Project titled **“Zomato Exploratory Data Analysis”** which is submitted by me to **Dr. Preeti Gupta**, Amity Institute of Information Technology, Amity University Maharashtra, Mumbai, in partial fulfillment of requirement for the award of the degree of **BSC - IT**, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

The Author attests that permission has been obtained for the use of any copy righted material appearing in the Dissertation / Project report other than brief excerpts requiring only proper acknowledgement in scholarly writing and all such use is acknowledged.

Place: - Mumbai

Shivangi Prasanna Koltharkar

A71004917003

#### GUIDE CERTIFICATE

I hereby certify that the Seminar Report by **Shivangi Prasanna Koltharkar**, student of **B.Sc. (IT) Semester-6 Enrolment No- A71004917003 with title “Zomato Exploratory Data Analysis”** which is submitted to **Amity Institute of Information Technology, Amity University Maharashtra**, Mumbai in partial fulfillment of requirement for the award of the degree of **BSC-IT** is an original contribution with existing knowledge and faithful record of work carried out by him/her under my guidance and supervision and to the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Mumbai Dr. Preeti Gupta,

Date: Associate Professor

Amity Institute of Information Technology

AUM, Mumbai

(Signature)

#### ABSTRACT

In this digitalized era online food delivery is the biggest gift for every single person. The biggest company in this field is the Zomato. They are known for their services of reviews, complete restaurant menu, booking tables and most important food delivery. Some restaurants which are new get themselves registered on Zomato but surprisingly they do not start off with the online delivery service, due to their problems, some restaurants prefer dine in still. This choice of the restaurants can be producing heavy impacts as only the customers who visit the restaurants will rate it. We will find out if the facility of online delivery will affect the rates or not. Here k means clustering is used which is an unsupervised learning model which is commonly used in exploratory data analysis to find patterns and relations. Such designs were pre made for several other major cities like Bangalore and New Delhi, there was not much information found for cities in Maharashtra. Here the analysis is done for Mumbai, Nashik, Nagpur, where their restaurants with blooming business. In this project we follow the basic steps of exploratory data analysis and find the relation between the fields present. By this we will draw the conclusion if all restaurants should start online service or it will not affect their income without it. Majority of these analysis show that it is not necessary because as much as people like to eat at their own comfort, there are people who like to go out and eat at the restaurant to have a change of view as enjoy the hospitability.

### ACKNOWLEDGEMENT

It is high privilege for me to express my deep sense of gratitude to those entire faculty Members who helped me in the completion of the project, specially my internal guide **Dr Preeti Gupta** who was always there at hour of need.

I am thankful to **Dr Manoj Devare** in charge HOD (Head of Department) of AIIT, AUM.

My special thanks to all other faculty members, Batch mate & seniors of Amity Institute of Information Technology, Amity University, Maharashtra for helping me in the completion of project work and its report submission.

Shivangi Prasanna Koltharkar

A71004917003

### TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Contents** | **Page No.** |
| 1 | DECLARATION BY STUDENT | ii |
| 2 | GUIDE CERTIFICATE | iii |
| 3 | ABSTRACT | iv |
| 4 | ACKNOWLEDGEMENT | v |
| 5 | LIST OF FIGURES | vii |
| Chapter-1 | Introduction | 1-3 |
|  | 1.1 Problem Definition | 2 |
|  | 1.2 Objectives | 3 |
| Chapter 2 | Preliminary Investigation | 4 |
| Chapter 3 | Feasibility Study | 5 |
| Chapter 4 | System Design | 6 |
| Chapter 5 | Code | 7-17 |
| Chapter 6 | Implementation | 18-19 |
| Chapter 7 | Contribution | 20 |
| Chapter 8 | Maintenance | 21 |
| Chapter 9 | Conclusion and Future Scope | 22 |
|  | References | 23 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title of the Figure** | **Page No.** |
| 1 | Fig 1: Steps for Exploratory Data Analysis | 6 |
| 2 | Fig 2: Flow chart of kmeans | 6 |
| 3 | Fig 3: Dataset | 7 |
| 4 | Fig 4: Data type | 7 |
| 5 | Fig 5: Columns | 8 |
| 6 | Fig 6: Duplication | 8 |
| 7 | Fig 7: Deletion | 8 |
| 8 | Fig 8: Count plot of rating variable | 9 |
| 9 | Fig 9: Scatter plot | 10 |
| 10 | Fig 10: No of online service providers | 11 |
| 11 | Fig 11: Rating compared to online service | 11 |
| 12 | Fig 12: Number of restaurants in total | 12 |
| 13 | Fig 13: Best Restaurant | 12 |
| 14 | Fig 14: K means scatter plot | 13 |
| 15 | Fig 15: No of clusters | 14 |
| 16 | Fig 16: Forming clusters | 14 |
| 17 | Fig 17: Clusters in table form | 14 |
| 18 | Fig 18: Representing different clusters | 15 |
| 19 | Fig 19: Centroid values | 15 |
| 20 | Fig 20: Marking the centroids | 16 |
| 21 | Fig 21: Stable centroids | 17 |
| 22 | Fig 22: Elbow graph | 17 |
| 23 | Fig 23: Anaconda Website | 18 |
| 24 | Fig 24: Anaconda Navigator | 18 |

**CHAPTER 1**

INTRODUCTION

Exploratory data analysis is a technique where major relationships that are between the data fields will be analyzed through pictorial representation such as graphs, plots and tables. It becomes a very useful technique when there is a huge dataset to go through. Food analysis is one of the upcoming and necessary analysis done by machine learning in these years. People are very concerned about the price and ratings and the flexibility of receiving their food. For this Zomato is always the 1st that comes to mind for online food delivering which is convenient for any person for booking a table, find the ratings or ordering online.

There has been a rise in Zomato reviews and opinions for people due to its accuracy and convenience. Zomato is set up by Deepinder Goyal and Pankaj Chaddah in 2008 which is one of the most blooming food aggregators which provides information, menus and user-reviews of restaurants, and also has online food delivery options from all the restaurants that are registered with it across the city. Zomato was earlier known as Foodiebay and then it was relaunched and retitled as Zomato. Zomato expanded across the cities in India like Delhi NCR, Mumbai, Bangalore, Chennai, Nagpur and Kolkata at the early stages as it started getting popularity between the youth in these cities. Later on, the company decided to start operations internationally in countries like, the UAE, Sri Lanka, Qatar, the UK, the Philippines, and South Africa. Seeing the blooming business, it started developing its website, and app in different languages at it will be convenient for the regional customers along with English, like Turkish, Brazilian, Portuguese and Indonesian.

The idea of Zomato was launched when the founders found out that most of the people did not know which all restaurants were available in their surroundings, which was impacting the restaurants financially bad as the non-popular ones had got very few customers on daily basis. The founders by this got the idea if they could create an application or website where all the restaurants nearby were listed and customers could get access to the information of the restaurants and other details through one medium. by this Foodiebay was set in motion in the early days. This all was established in New Delhi within the NCR, eventually it gained popularity. This gave the founder's idea to start this around other cities too.[1]  
Zomato has many challengers like Swiggy, Dineout, JustDial, etc.

* 1. PROBLEM DEFINITION

The project deals with exploratory data analysis in the food service industry. Report of National Restaurant Association of India (NRAI) suggests that the restaurant industry in India employs about 7.3 million people and has an estimated annual turnover of Rs. 4 Lakh Crore. [2] The project tries to tap the very nature of this people specific industry through data analysis as it is an industry that have gained prominence over these years. Hence it makes it imperative to explore data on the basis of various aspects in food delivery, restaurant’s rating and affinity restaurant groups in the vicinity etc.

The case of Zomato is taken in particular, as it is the pioneer food aggregator.

* 1. . OBJECTIVE

In order to resolve the problem of exploratory data analysis in the food service industry, the case of Zomato has been taken up. The following objectives are to be met.

1. Preparation and Preprocessing of the dataset obtained from Zomato.
2. Data visualization through graphs and plots to explore data under the following criteria:

                  (i)  Online delivery services and its impact on the rating of the restaurant.

                  (ii)  Restaurant Ratings.

                  (iii) Identifying correlation/affinity between the restaurants etc.

**CHAPTER 2**

PRELIMINARY INVESTIGATION

Zomato uses all information about its rating and reviews from customers alone. There can be issues with the data due to competition and public opinion overall towards Zomato. This can cause a barrier not just between customer and the app but also Zomato and the restaurant and customer and the restaurant. This will overall affect how often will customers visit and order food using Zomato from a specific restaurant. This can cause downfall and major financial issues which will lead to low publicity and rates.

Zomato has encountered problem with the customer care calling service in the recent years. Zomato always welcomes customers to write messages and reviews about the facilities provided, services and the quality about a restaurant. This only helps them to be more reliable but claims there is no such thing as customer care number where customers can reach. Calls from the customer care are only made when there are some serious issues with the money transfer, other than that there is no number provided to the customers they can contact the company. A lady in Bangalore had called up the customer care number provided by Google and was asked to download an app and follow few steps this resulted in all the money in her savings account being deducted. This led to disbelief in the rest customers and reviews gradually become bad and this affected the restaurants that used only Zomato services too. Now the number used is not reachable but people have to be aware of such things hence forth.[3]

E-commerce has made life flexible but also dangers in such situations one needs a total reliable method and ways to find out what they search for. This is what Zomato looks forward and wants to provide customer with every little information. Customers can rely on them and visit the restaurants with an expectation and the restaurant will for sure stand up to those expectations.

It is very important for people to know about the best restaurant and their ratings and how this will improve their votes. Many regions like New Delhi and Bangalore this data analysis is done but not of places in Maharashtra. So, it was considered to do the same for few regions of Maharashtra too like Mumbai, Nagpur and Nashik. There were few data analysis done for New Delhi where they compared the cost of two of New Delhi to the cost of two to United States, number of restaurants registered overall with the result India is cheaper and has many restaurants registered to Zomato. Several factors that affected the ratings of restaurants in New Delhi was done and found out the people preferred more North Indian food hence all the ratings were based on that only.[4] In Bangalore data analysis was done to find out about how many restaurants are present and what are the ratings. [5]

**CHAPTER 3**

FEASIBILITY STUDY

1. Business feasibility

As this test was not done in the region of Maharashtra even getting the dataset was difficult, as no much record was available beforehand. People had not extracted datasets of Maharashtra much from the Zomato provided fragmented data. Even though there are much more restaurants overall Maharashtra they are not registered with Zomato but some other websites. This study will help the restaurants economically as they will do the necessary changes accordingly, as they need to make major changes of starting online delivery or not. People will visit or order only by seeing the ratings. Factors like this matter in the progress of the restaurant and affect financially too.

1. Data feasibility

Data set with enough variables and with no distortion should be procured. Either dataset which are available can be used on dataset providing websites like Kaggle or get from Zomato developers site which will be extracted in scattered form in json file. This can be worked with in the later stages of the project as of now dataset available from Kaggle is used. The Zomato developers site allows us to use the data once told for what reason it is used, for security reasons it is fragmented. We need to know the extraction of this data without any lose in the procedure or duplication. To avoid all this, we used a dataset available on Kaggle, which was used by other students for their project.

1. Execution feasibility

There should be installed python either by using an open source distributer or direct python. Preferable is anaconda as few open source software are present in it like Jupyter notebook and Spyder. There is no payable amount needed for using these software’s they are freely available on the website anaconda.com.

There should be libraries- seaborn, Sklearn, matplotlib and numpy. This can be downloaded by using anaconda prompt.

Anaconda has its own cloud-based rest packages so other libraries can just be imported while the program execution.

**CHAPTER 4**

SYSTEM DESIGN

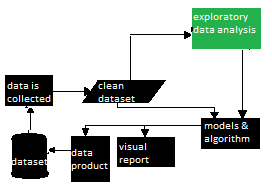
[6]

Fig 1: Steps for exploratory data analysis

# MyCodeCamp: Clustering K Means | K means clustering in Java | K ...

Fig 2: Flowchart of kmeans

**CHAPTER 5**

CODE

This is to read the data set.

import numpy as npy

import pandas as pd

df=pd. read\_csv("C://Users//user//Desktop//zomato\_data.csv", encoding='latin-1')

df.head()[7]

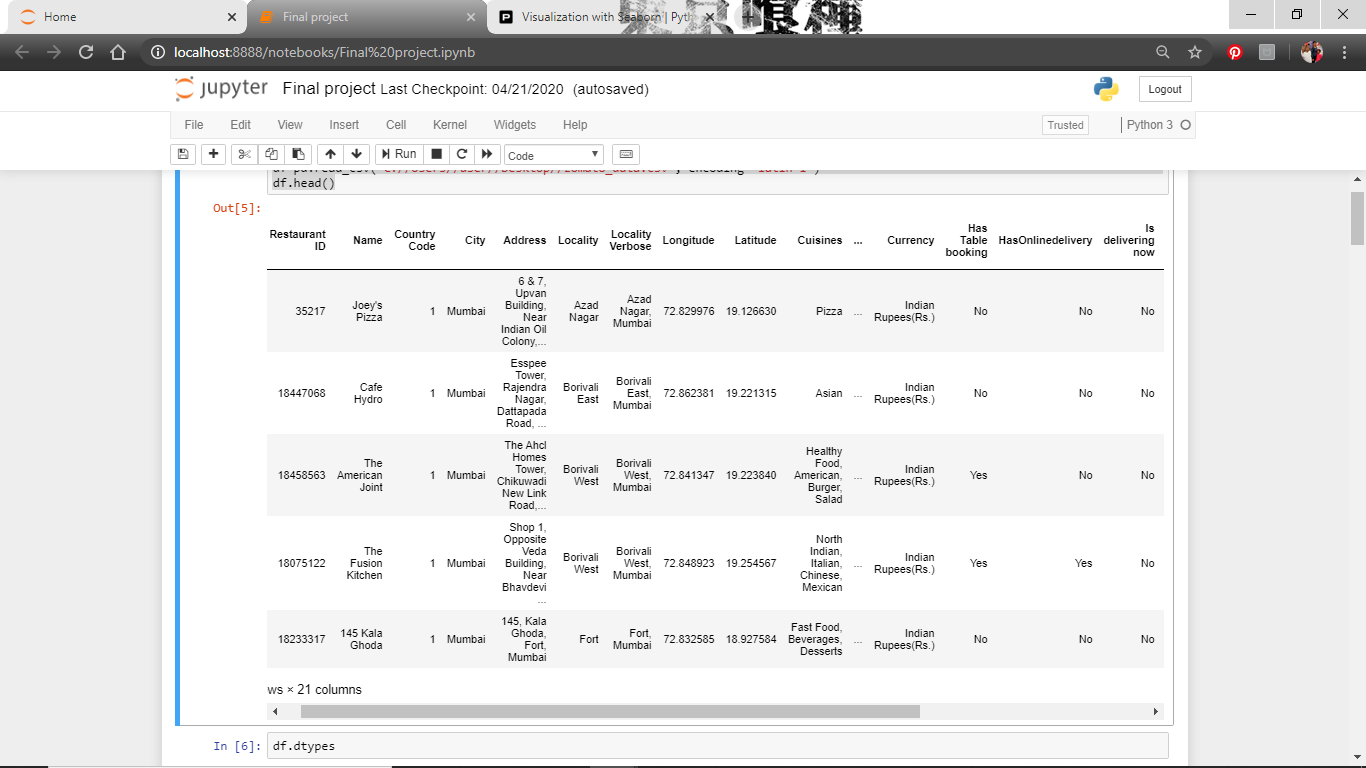


Fig 3: Dataset

The data type of each is known here

df.dtypes

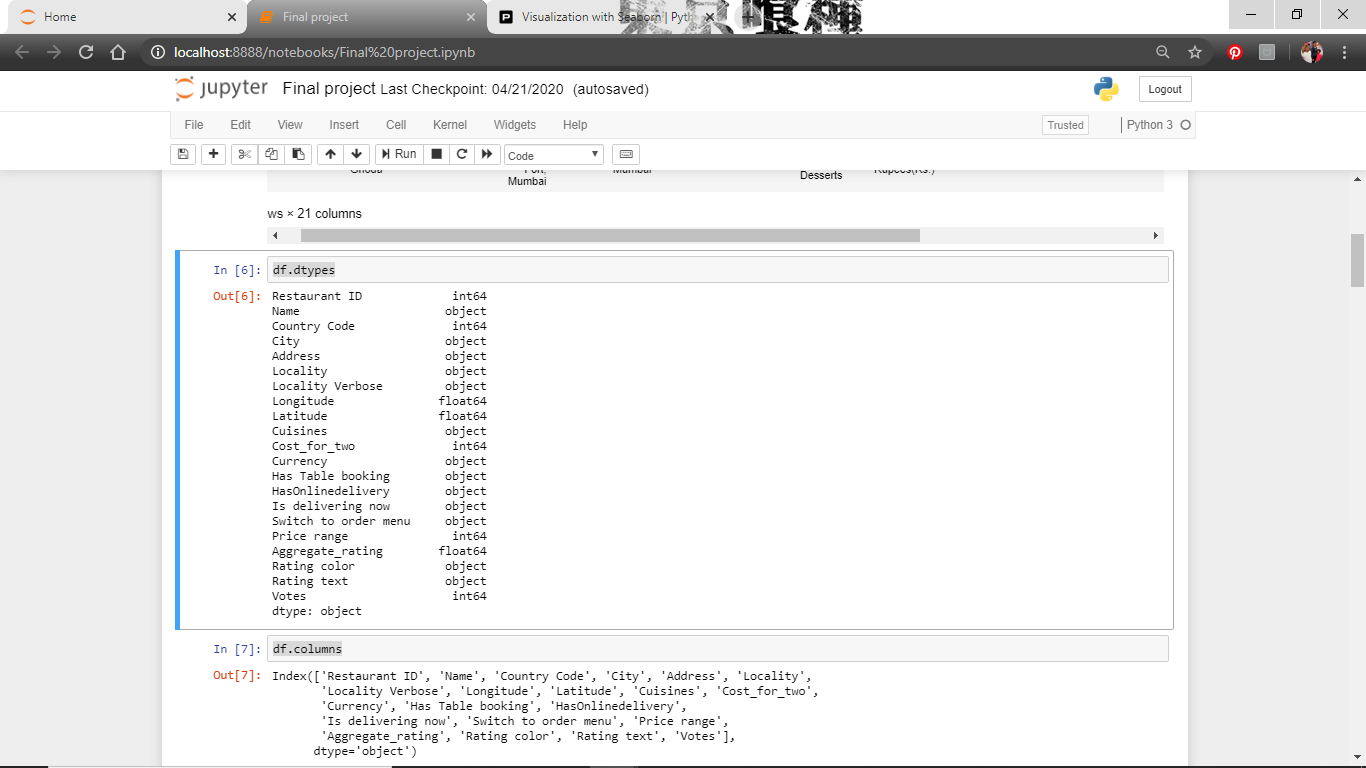


Fig 4: Data type

To know the names of the columns

df.columns

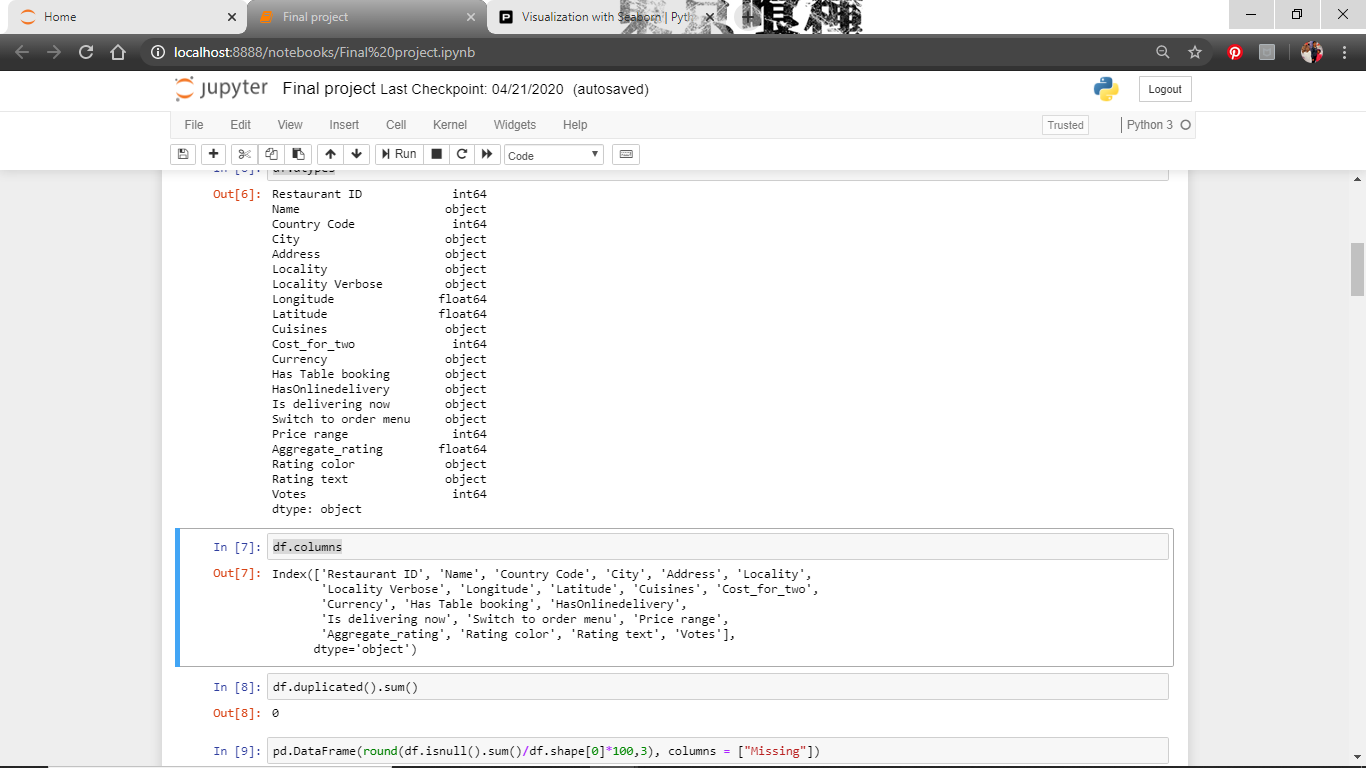


Fig 5: Column

Find if there are any duplications

df.duplicated().sum()

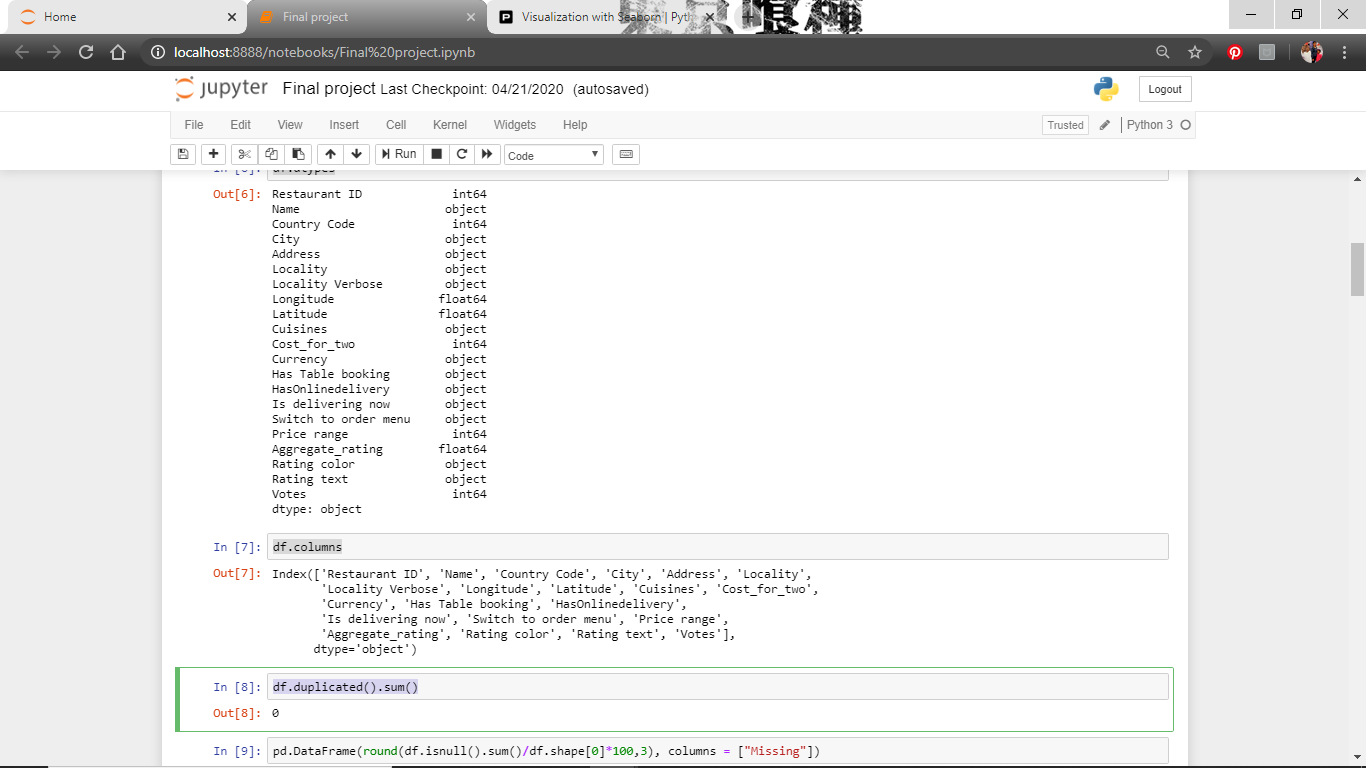


Fig 6: Duplications

Delete the unwanted columns

df.drop(columns=['Country Code', 'Restaurant ID','Latitude', 'Longitude'])

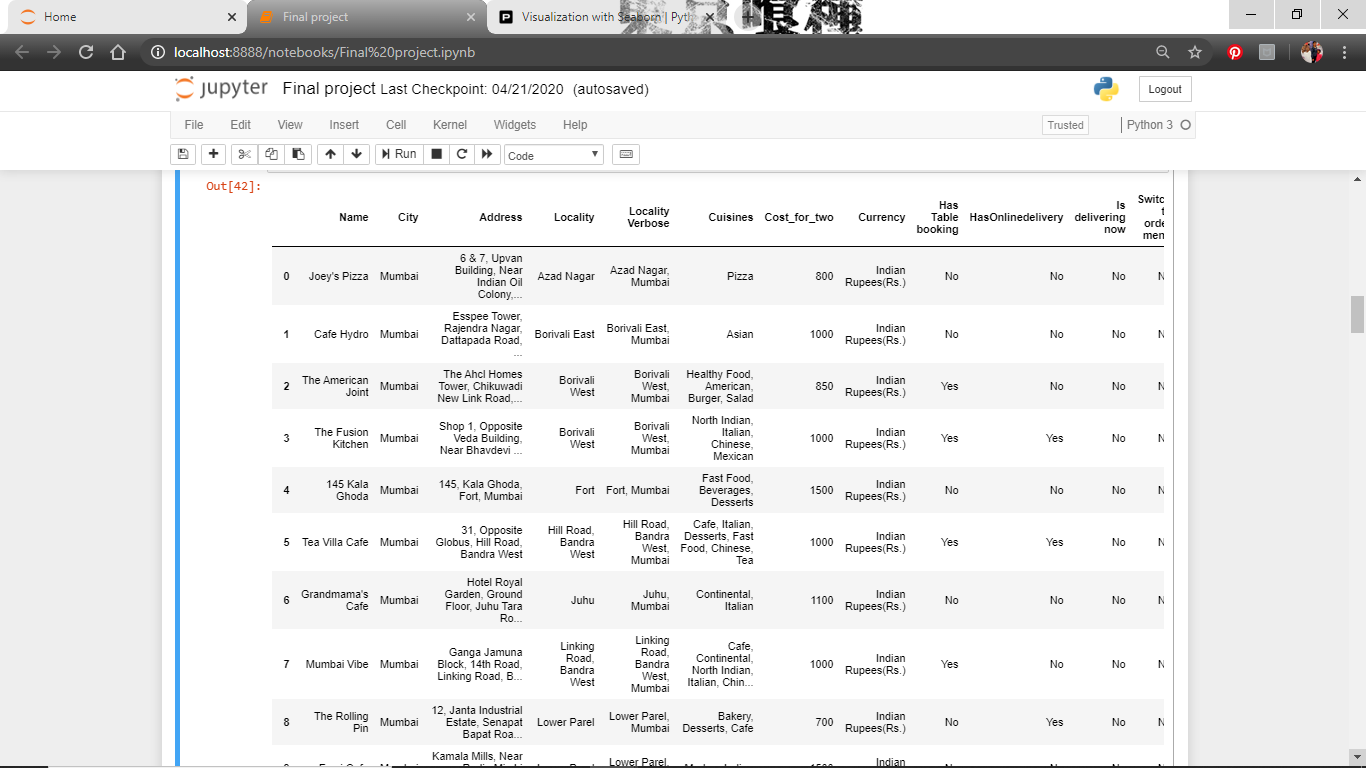


Fig 7: Deletion

We find out what is the count of rating i.e. how many 4.5 is present or how many 3.0

import matplotlib.pyplot as mplt

import seaborn as sn

df=pd.read\_csv("C:/Users/user/Desktop/zomato\_data.csv", encoding='latin-1')

mplt.rcParams['figure.figsize'] = (14,7)

sn.countplot(df["Aggregate\_rating"], palette="Set1")

mplt.title("Count plot of rating variable")

mplt.show()

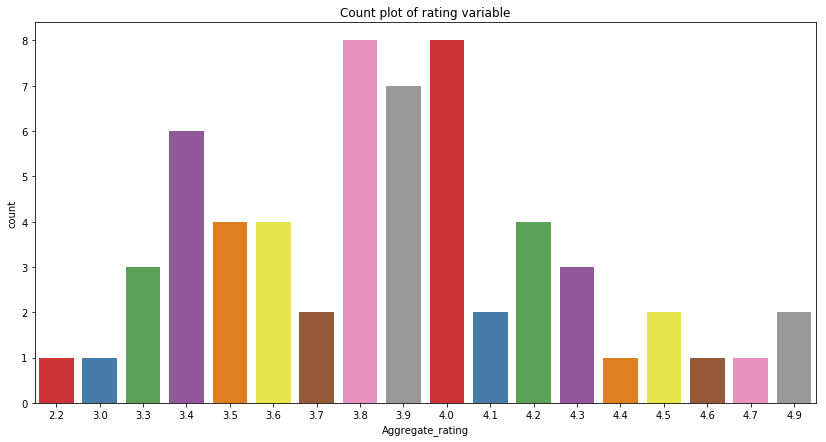


Fig 8: Count plot of rating variable

We create a joint plot to see the relation between rates and cost of two

sn.jointplot(x="Aggregate\_rating", y="Cost\_for\_two", data= df, height=8, ratio=4, color="g")

mplt.show()[8]

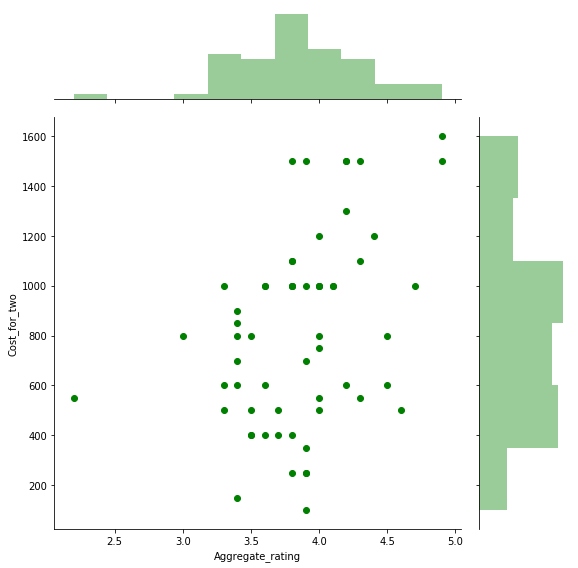


Fig 9: Scatter plot

To check how many restaurants provide online delivery services

df.HasOnlinedelivery.value\_counts().nlargest(10).plot(kind="barh")

mplt.title("Number of restaurants providing online services")

mplt.xlabel("Count")

mplt.show()



Fig 10: No of online service providers

This is to check if being an online delivery restaurant gets it more rating or it does not matter for the rating

sn.countplot(hue = df["HasOnlinedelivery"], x=df["Aggregate\_rating"])

mplt.title("Distribution of restaurants rating compare to online facility")

mplt.show()

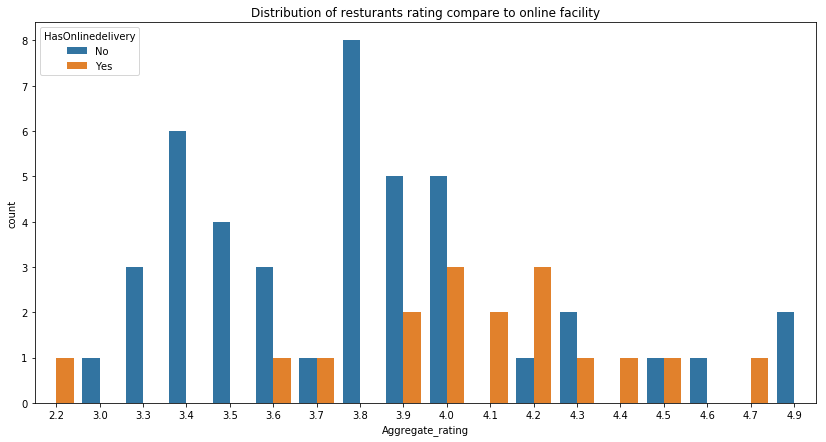


Fig 11: Rating compared to online service

Each restaurant has how many branches

df.Name.value\_counts().plot(kind = "barh", color = sn.color\_palette("hls",5))

mplt.xlabel("Number of restaurants")

mplt.title("Number of restaurants overall")

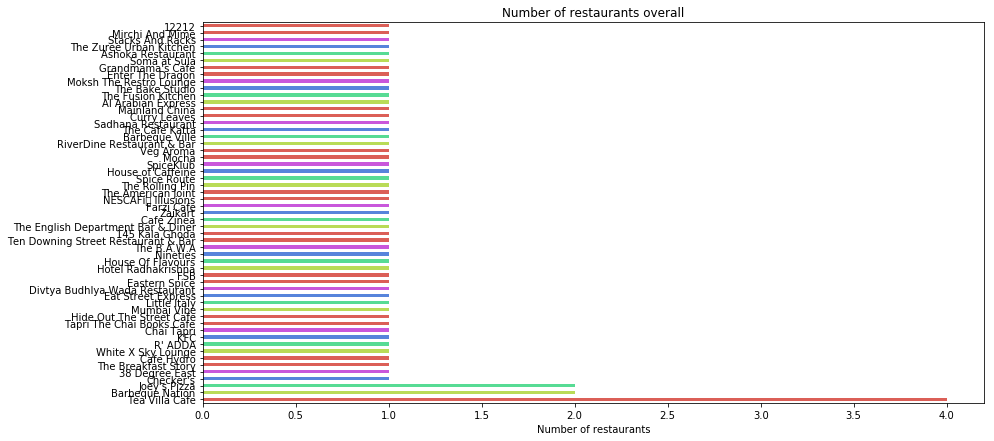


Fig 12: No of restaurants total

Best restaurant according to the rating.

df[df['Aggregate\_rating']>=4]['Name'].value\_counts().nlargest(5).plot(kind = "barh", color = sn.color\_palette("Paired"))

mplt.xlabel("Number of restaurants")

mplt.title("Best Restaurant- Rating More than 4")

mplt.plot(4,10,10) [9]

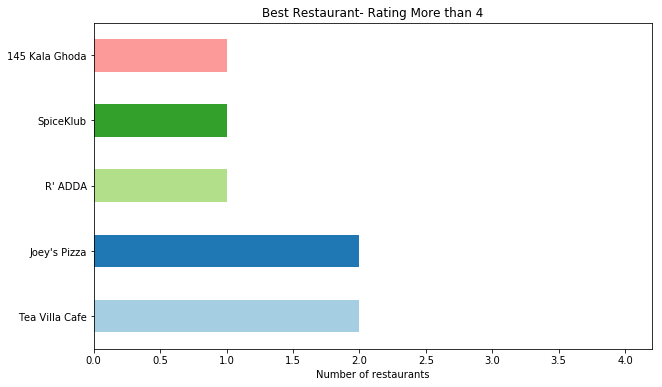


Fig 13: Best restaurant

Performing k means

import matplotlib.pyplot as mplt

import sklearn

from sklearn.cluster import KMeans

from sklearn.preprocessing import scale

import sklearn.metrics as sm

from sklearn.metrics import confusion\_matrix, classification\_report

#creating a scatterplot with the data

df=pd.read\_csv("C:/Users/user/Desktop/zomato\_data.csv", encoding='latin-1')

f1 = df['Aggregate\_rating'].values

f2 = df['Cost\_for\_two'].values

X = np.array(list(zip(f1, f2)))

mplt.scatter(f1, f2)

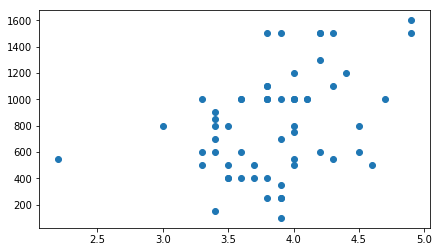


Fig 14: Kmeans scatter plot

#Giving the number of clusters we want

kmeans= KMeans(n\_clusters=3)

kmeans

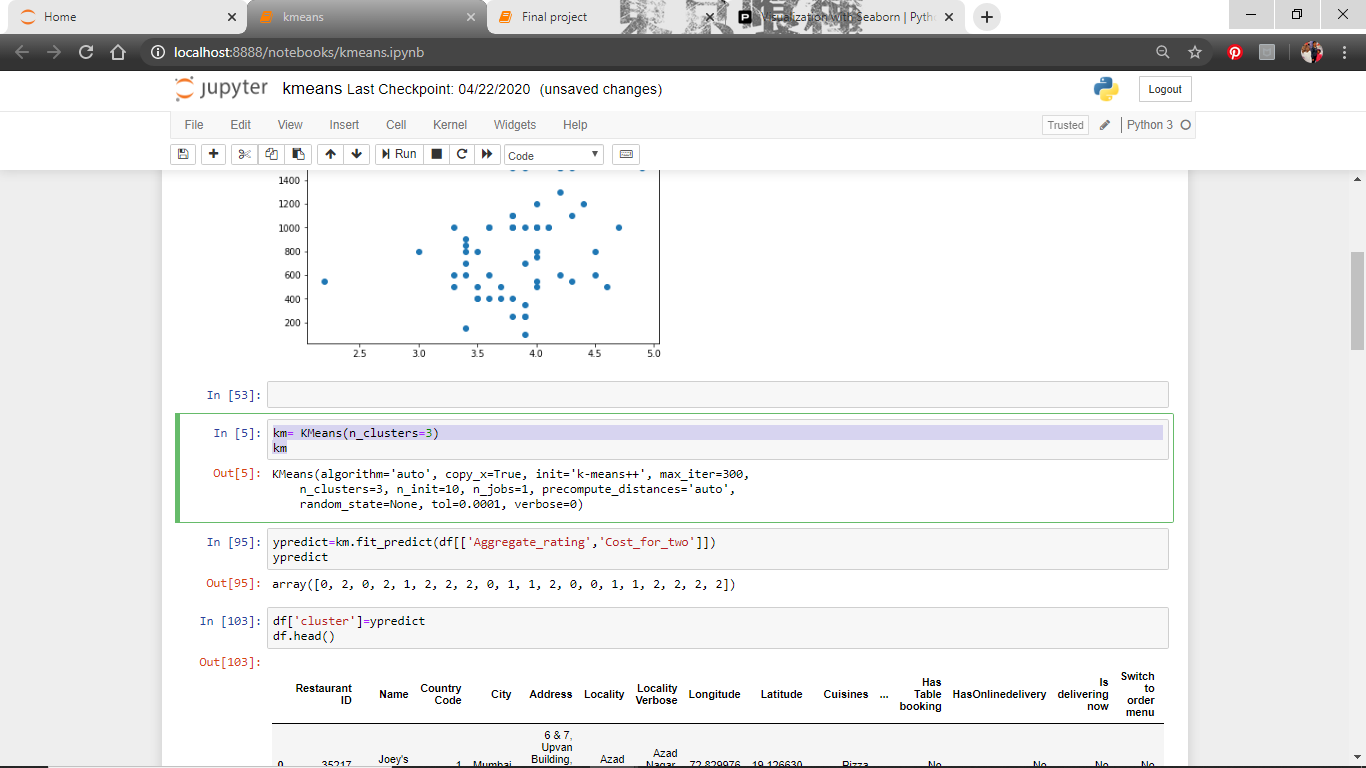


Fig 15: No of clusters

#Which cluster does a value belong to

ypredict=km.fit\_predict(df[['Aggregate\_rating','Cost\_for\_two']])

ypredict

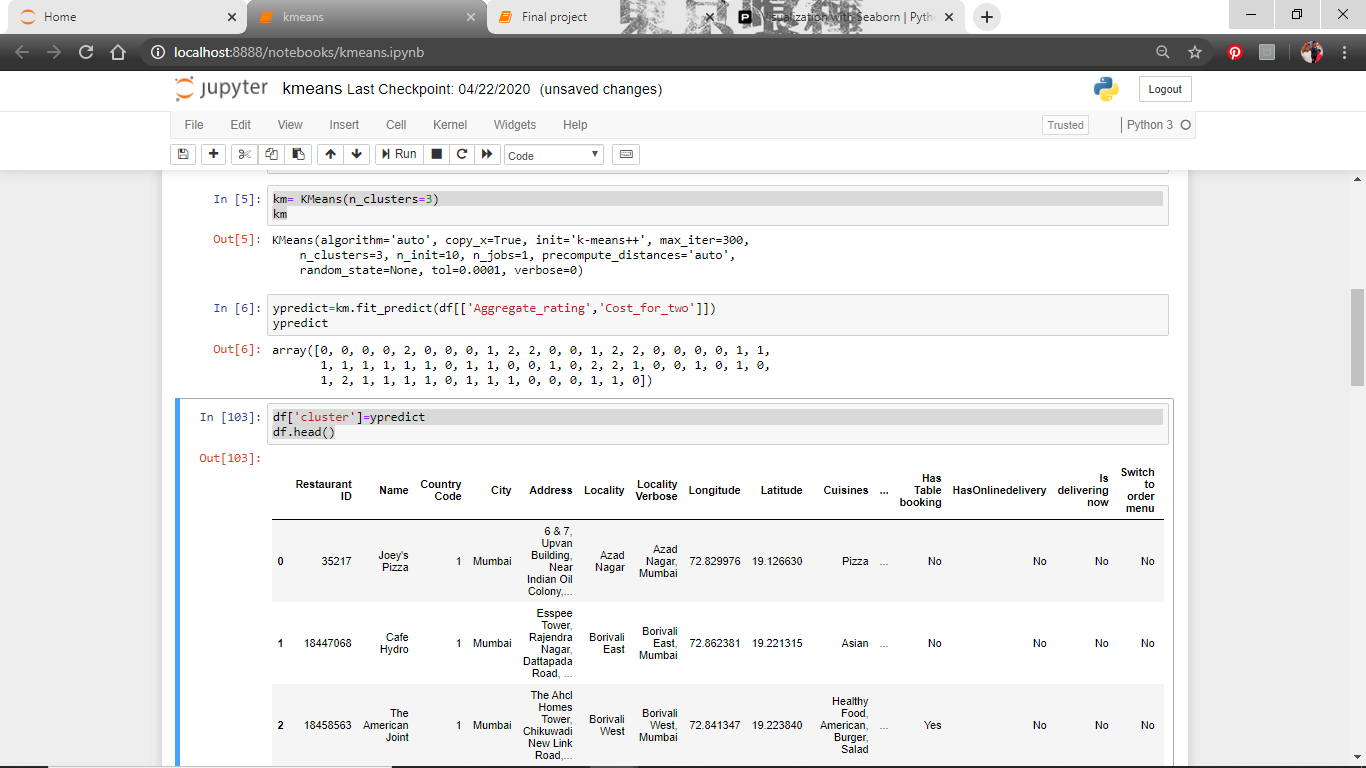


Fig 16: Forming clusters

#Clusters along with the values

df['cluster']=ypredict

df.head()

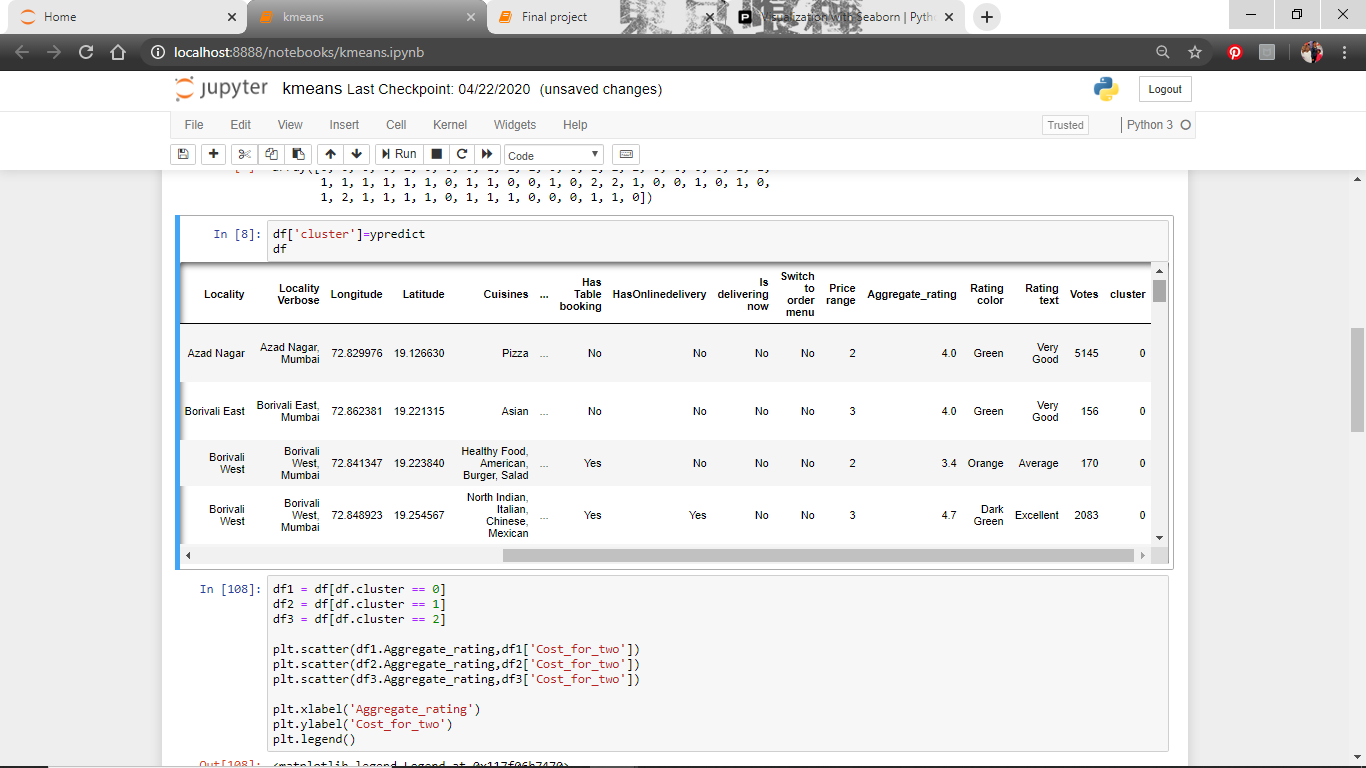


Fig 17: Clusters in table form

#To give each cluster a separate color coding to distinguish

df1 = df[df.cluster == 0]

df2 = df[df.cluster == 1]

df3 = df[df.cluster == 2]

mplt.scatter(df1.Aggregate\_rating,df1['Cost\_for\_two'])

mplt.scatter(df2.Aggregate\_rating,df2['Cost\_for\_two'])

mplt.scatter(df3.Aggregate\_rating,df3['Cost\_for\_two'])

mplt.xlabel('Aggregate\_rating')

mplt.ylabel('Cost\_for\_two')

mplt.legend()

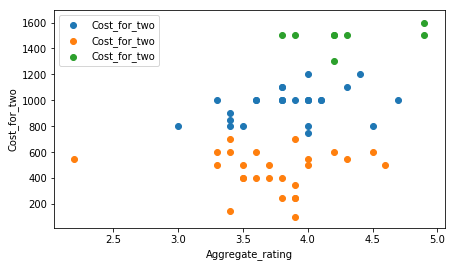


Fig 18: Representing different clusters

#finding the centroid values

kmeans.cluster\_centers\_

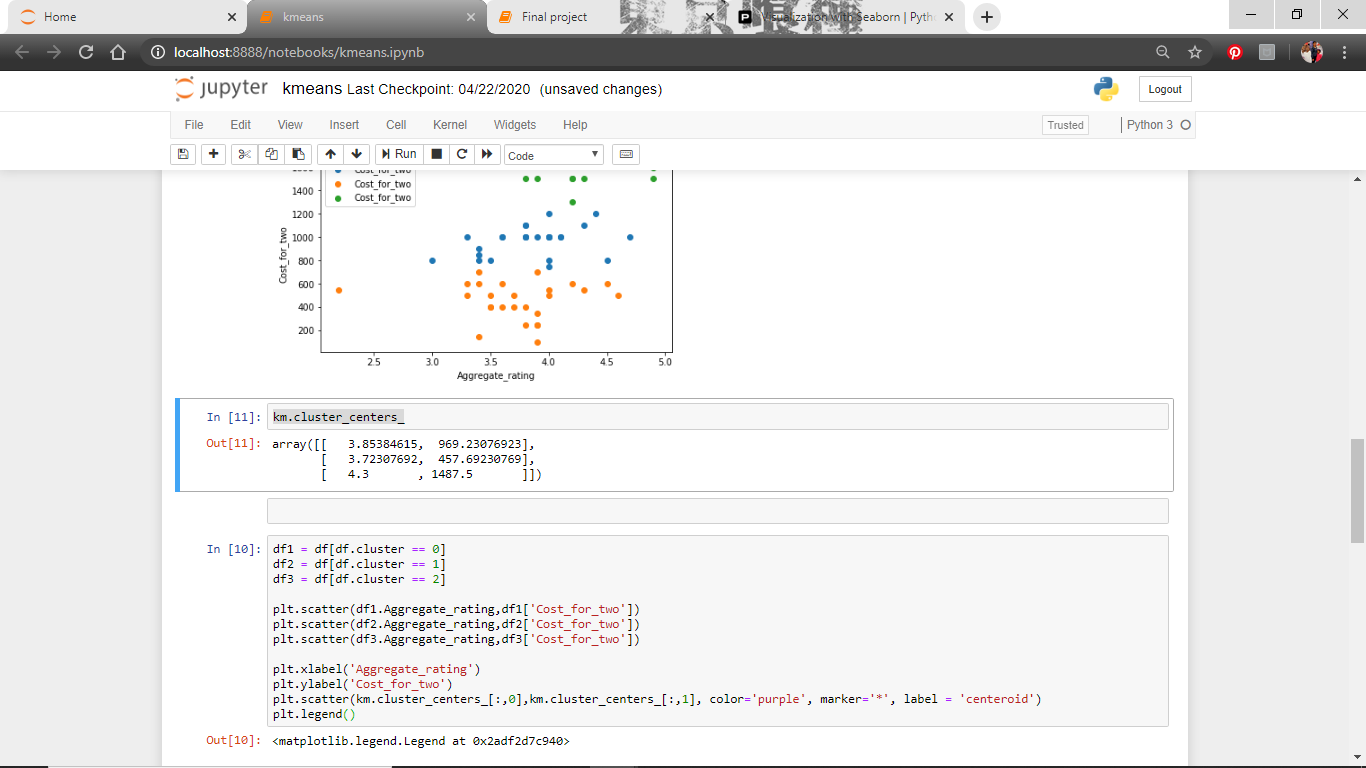


Fig 19: Centroid values

#Marking the centroids

df1 = df[df.cluster == 0]

df2 = df[df.cluster == 1]

df3 = df[df.cluster == 2]

mplt.scatter(df1.Aggregate\_rating,df1['Cost\_for\_two'])

mplt.scatter(df2.Aggregate\_rating,df2['Cost\_for\_two'])

mplt.scatter(df3.Aggregate\_rating,df3['Cost\_for\_two'])

mplt.xlabel('Aggregate\_rating')

mplt.ylabel('Cost\_for\_two')

mplt.scatter(km.cluster\_centers\_[:,0],km.cluster\_centers\_[:,1], color='purple', marker='\*', label = 'centeroid')

mplt.legend()

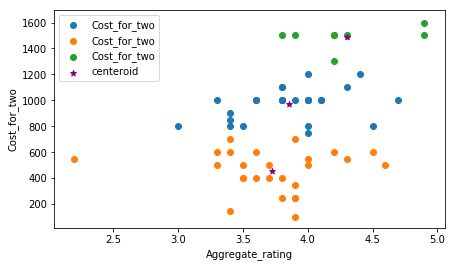


Fig 20: Marking the centroids

#Iterative method for plotting elbow plot

range=range(1,20)

s=[]

for k in range:

kmeans = KMeans(n\_clusters = k)

kmmeans.fit(df[['Aggregate\_rating', 'Cost\_for\_two']])

s.append(kmeans.inertia\_)

s

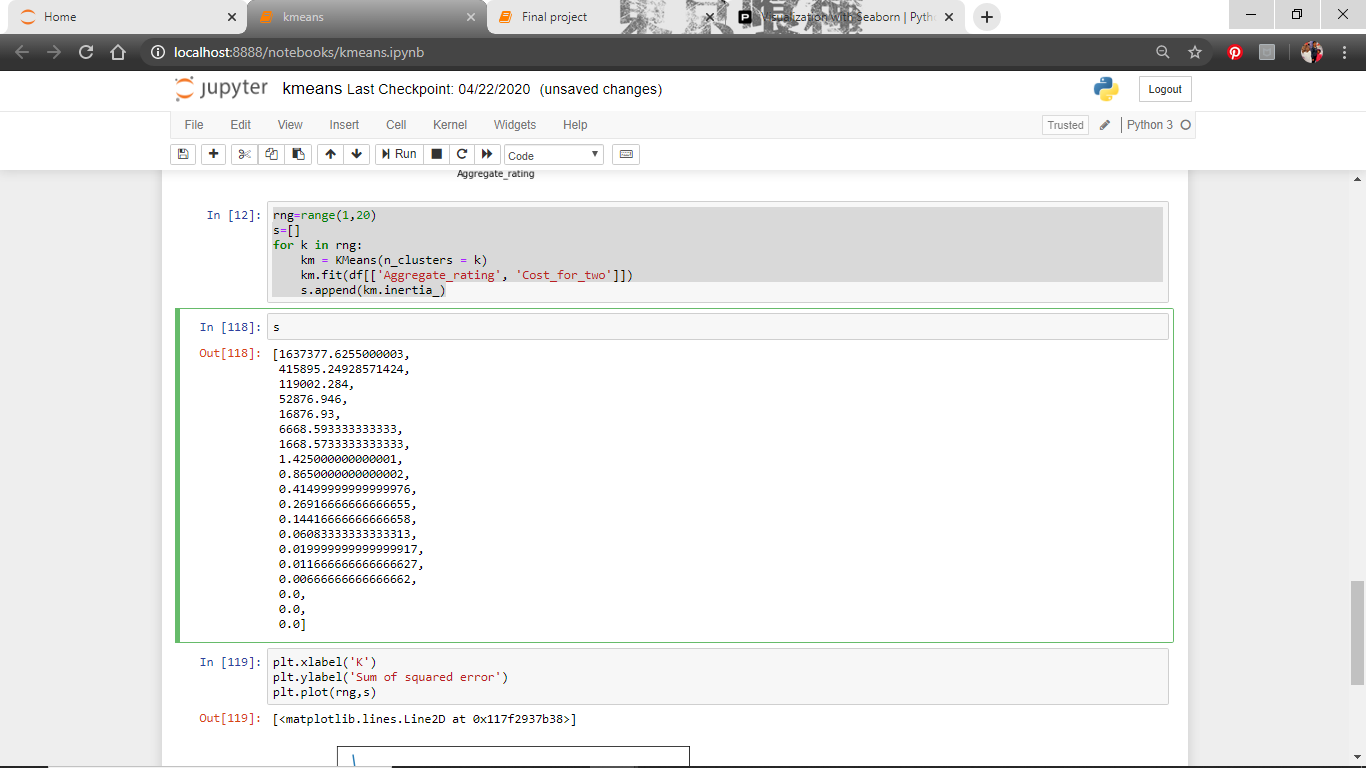


Fig 21: Stable centroids

#Ploting the elbow graph

mplt.xlabel('K')

mplt.ylabel('Errors')

mplt.plot(range,s)[10]

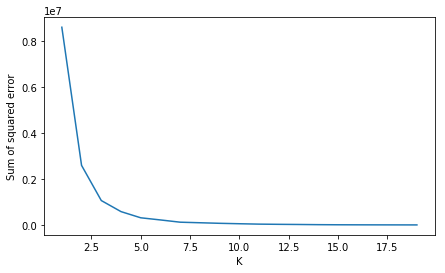


Fig 22: Elbow graph

**CHAPTER 6**

IMPLEMENTATION

This is done by using python general purpose programming language. For this installation of anaconda open source distributer is required. It can be downloaded from its official website anaconda.com/products/individual.

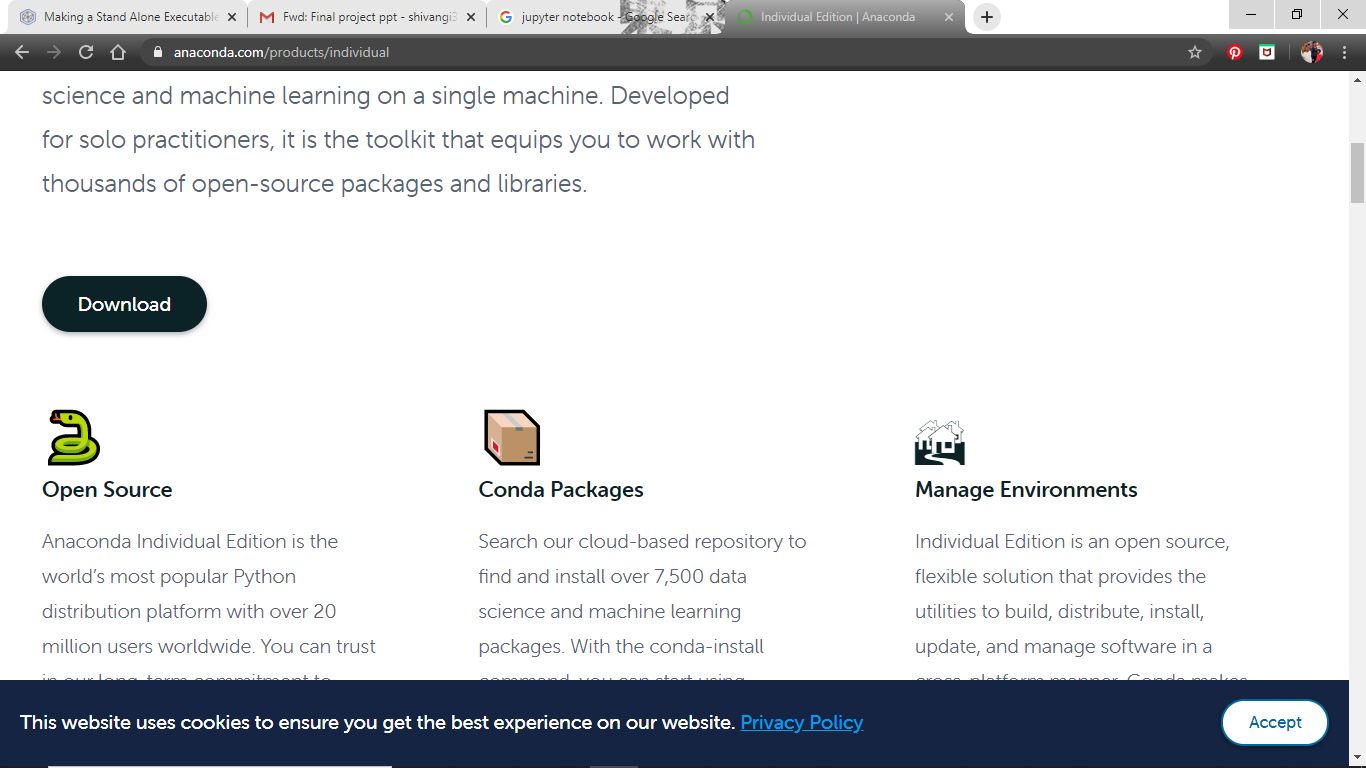


Fig 23: Anaconda website

In the anaconda navigator we will be using Jupyter Notebook. [11]

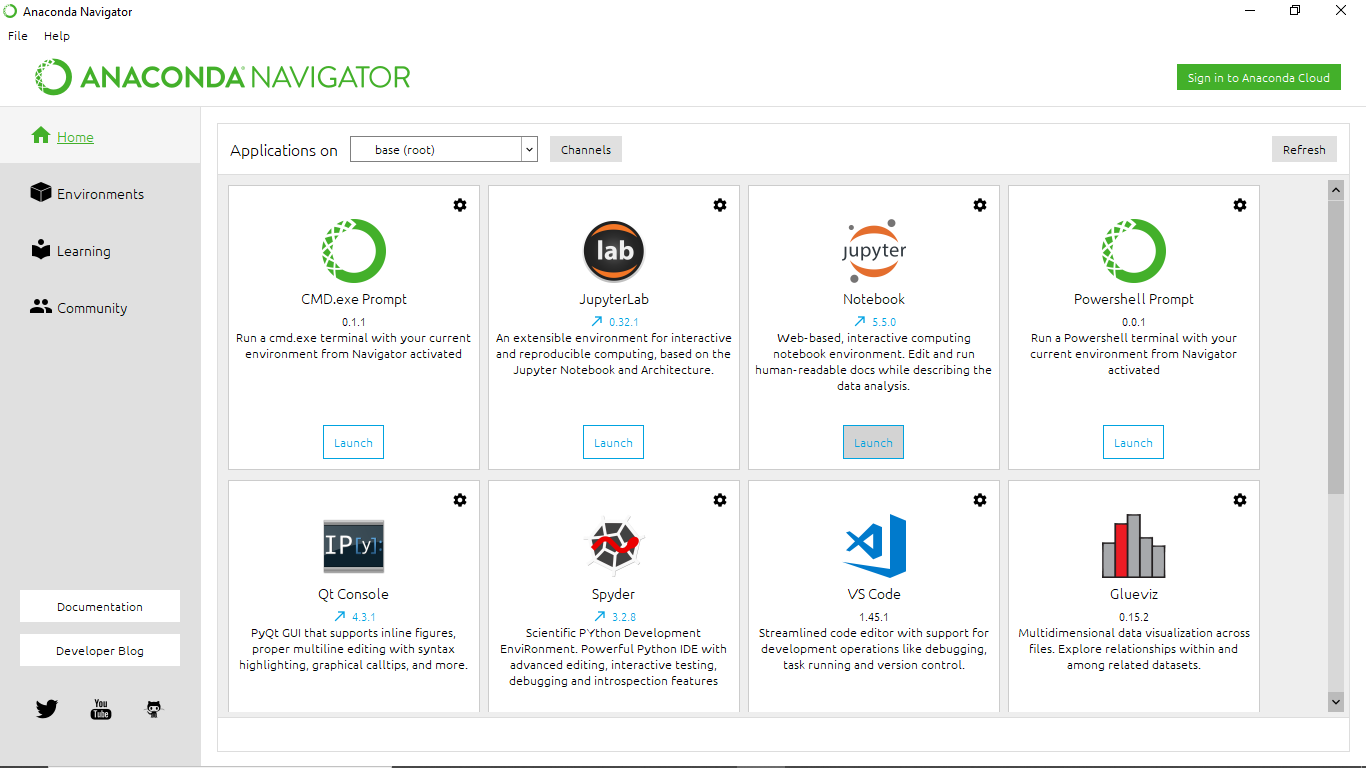


Fig 24: Anaconda Navigator

Certain libraries need to be installed as we need to call them during the execution of code.

**NumPy**

It is a library which is used for working with arrays. It is easier to access information from NumPy than in list due to its location storage method.

For installing this we need to open anaconda prompt and type **conda install -c anaconda numpy**

**Matplotlib.pyplot**

It is a collection where all command style functions are present. It is used to create plots and figures. Where it receives data and plots graphs and figures accordingly.

For installing this, on the anaconda prompt have to type **conda install matplotlib**

**Seaborn**

This library is to create plots with more interactive figures, it gives data a unique visualization based on matplotlib. By using this we can create and look into complex data by great visualization. We will have to install it by using pip or conda.

For installing this, on the anaconda prompt have to type **conda install seaborn**

**Sklearn**

It is a library where algorithms like support vector, random forest, k N and many more supervised and unsupervised algorithms are featured. This library is built upon scientific python SciPy.

For installing this, on the anaconda prompt have to type **conda install -c anaconda scikit-learn**

**CHAPTER 7**

INDIVIDUAL CONTRIBUTION

* Preprocessing the data- Data type, Column names, dropping column, searching for duplication, missing data and finding the anomalies.
* Implementation of Kmeans algorithm
* Data Visualization- Joint plot which consist scatter plot and histogram
* Report writing

**CHAPTER 8**

MAINTENANCE

It is clearly understood that for an application to be rated highly like Zomato, we need to serve to the customer in a way they are satisfied with the service. There are various factors for an app like Zomato which makes it a successful food delivering app like Legitimate Cost, Great Packing, Proper food delivery, Easy to access, Hygienic etc.

Therefore, to maintain the above program we can keep track of all the restaurants not just in the city but also throughout the state where we can analyze the data and determine which restaurant or particular venue is most popular and fulfils all the demands of the Customer. We can further link the direct dataset which will connect this model to the real world. Moreover, we can also add more datatypes like price range, best combos, exciting offers and discounts etc which further can provide more convenience to filter the ratings. As we keep feeding the model with more data it will get better in its prediction.

**CHAPTER 9**

CONCLUSION

As we can see from the analysis done that is it not necessary to have online delivery, the restaurants still got good rating without considering that factor. What is important is the quality of the food and the consistency, that is the reason customers return. The restaurants which cost 800-1000 for two tend to get good ratings as that is the optimal money anyone would like to spend at a time. If the restaurant can start online services it will eventually will be getting rating from both ways. There will be no major issue with their financial income and people seeing the rating will visit and enjoy the hospitability provided by the restaurant until then. We found out the best restaurants were Tea villa café, Joe’s pizza, R’ adda, SpiceKlub, 145 Kala Ghoda.

REFERENCES

1. https://en.wikipedia.org/wiki/Zomato
2. https://www.cnbctv18.com/retail/food-for-thought-survival-of-the-restaurant-industry-in-the-times-and-post-corona-5814141.htm, May 01 2020 08:01AM IST
3. https://www.indiatoday.in/india/story/customers-duped-zomato-call-centre-fake-customer-care-numbers-google-search-1580678-2019-08-14
4. https://rstudio-pubs-static.s3.amazonaws.com/468509\_51f0e06901204a0db7fabcf857719883.html
5. https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants
6. https://www.upgrad.com/blog/exploratory-data-analysis-and-its-importance-to-your-business/
7. https://www.kaggle.com/shrutimehta/zomato-restaurants-data
8. https://seaborn.pydata.org/generated/seaborn.jointplot.html
9. https://www.kaggle.com/bharathsangavi/zomato-exploratory-data-analysis
10. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
11. https://www.anaconda.com/