**EXPLORATORY VISUALIZATION AND ANALYSIS OF RESTAURANT DATASET**

**A PROJECT REPORT SUBMITTED TO**

**SRM INSTITUTE OF SCIENCE & TECHNOLOGY**

**IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF**

**BACHELOR OF COMPUTER APPLICATIONS**

**BY**

**SALIF HAMMAD**

**Reg. No. RA1731241010051**

**R. SAHER FATHIMA**

**Reg. No. RA1731241010063**

**AMMAR MUJEEB**

**Reg. No. RA1731241010093**

**UNDER THE GUIDANCE OF**

**Mrs. A. RAJALAKSHMI M.C.A., M.Phil., M.Tech.,**

****

**DEPARTMENT OF COMPUTER APPLICATIONS**

**FACULTY OF SCIENCE AND HUMANITIES**

**SRM INSTITUTE OF SCIENCE & TECHNOLOGY**

Kattankulathur – 603 203

Chennai, Tamilnadu,

**June – 2020**

**BONAFIDE CERTIFICATE**

This is to certify that the project report titled ***“*EXPLORATORY VISUALIZATION AND ANALYSIS OF RESTAURANT DATASET*”*** is a Bonafide work carried out by **SALIF HAMMAD** (RA1731241010063), **R.SAHER FATHIMA** (RA1731241010063) and **AMMAR MUJEEB** (RA1731241010093) under my supervision for the award of the Degree of Bachelor of Computer Applications. To my knowledge the work reported herein is the original work done by these students.

**Mrs. A. Rajalakshmi**

Assistant Professor,

Department of Computer Applications

(GUIDE)

**Dr. S. Albert Antony Raj**

Associate Professor & Head,

Department of Computer Applications

**ACKNOWLEDGEMENT**

**Internal Examiner**

**External Examiner**

With the profound gratitude to the ALMIGHTY we take this chance to thank people who helped us complete this project.

We take this as a right opportunity to say thanks to our Parents who got our back always with the words “Believe in Yourself”.

We are thankful **to Dr.T.R.Paarivendhar**, Chancellor, SRM Institute of Science & Technology who gave us the platform to establish ourselves to reach greater heights.

We earnestly thank **Dr.R.Balasubramanian**, Pro-Vice Chancellor, **Dr.J.Jothi Kumar**, Dean and [**Dr. Sridhar Krishnaswami**](http://www.srmuniv.ac.in/content/dr-sridhar-krishnaswami), Deputy Dean, Faculty of Science and Humanities, SRM Institute of Science & Technology who always encourages us to do novel things.

We express our sincere thanks to **Dr. S. Albert Antony Raj** M.Sc., M.Phil., Ph.D., Associate professor and Head, Department of Computer Applications, for his valuable guidance and support to execute all incline in learning.

It is our delight to thank our project guide **Mrs. A. Rajalakshmi**, Assistant Professor, Department of Computer Applications for her help, support, encouragement, suggestions and guidance throughout development phases of the project.

We convey our gratitude to all the faculty members of the department who extended their support through valuable comments and suggestions during the reviews.

A great note of gratitude to friends and people who are known and unknown to us who helped in carrying out this project work a successful one.

**SALIF HAMMAD**

**R. SAHER FATHIMA**

**AMMAR MUJEEB**

**EXPLORATORY VISUALIZATION AND ANALYSIS OF RESTAURANT DATASET**

**ABSTRACT**

As a part of BCA course, this project is a statistical analysis of Zomato’s restaurant dataset. The objective behind doing this project is to gain insights of the food culture in Chennai and the different factors that influence it.

Exploratory Data Analysis is a method of uncovering important relationships between the variables by using bar graphs, plots, and tables. Exploratory Data Analysis (EDA) is a very useful technique especially when you are working with a large unknown dataset. It allows you to **investigate the interesting relationships between the variables,** study the different subsets of data to unlock the different patterns in the data.

Zomato is India’s largest & most successful startup company where food industry meets technology and connects thousands of restaurants in a thread.

The sole objective of this project is to collect & process the datasets and visualize them according to their ratings and other factors.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter Number** | **Chapter Name** | **Page Number** |
| 1 | Introduction | 1-4 |
| 2 | Software Requirement & System Analysis | 5-9 |
| 3 | Data Description and Pre-Processing | 10-17 |
| 4 | Exploratory Data Analysis | 18-24 |
| 5 | Regression Analysis | 25-28 |
| 6 | Source Codes | 29-53 |
| 7 | Output Screens |  |
| 8 | Conclusion |  |
| 9 | Further Enhancements |  |
| 10 | References/Bibliography |  |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Serial Number** | **Figure Number** | **Page Number** |
| 1 | Fig. 1.1 | 4 |
| 2 | Fig. 4.1 | 20 |
| 3 | Fig. 4.2 | 20 |
| 4 | Fig. 4.3 | 21 |
| 5 | Fig. 4.4 | 21 |
| 6 | Fig. 4.5 | 22 |
| 7 | Fig. 4.6 | 23 |
| 8 | Fig. 4.7 | 23 |
| 9 | Fig. 4.8 | 24 |

**LIST OF PHOTOS**

|  |  |  |
| --- | --- | --- |
| **Serial Number** | **Photo Number** | **Page Number** |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |
| 6 |  |  |
| 7 |  |  |
| 8 |  |  |
| 9 |  |  |
| 10 |  |  |

**CHAPTER 1**

**INTRODUCTION**

1. Exploratory Visualization and Analysis of Restaurant Dataset
   1. **Introduction**

As a part of BCA course, this project is a statistical analysis of Zomato’s dataset. The objective behind doing this project is to gain insights of the food culture in Chennai and the factors that influence it.

Exploratory Data Analysis is a method of uncovering important relationships between the variables by using bar graphs, plots, and tables. Exploratory Data Analysis (EDA) is a very useful technique especially when you are working with a large unknown dataset. It allows you to **investigate the interesting relationships between the variables,** study the different subsets of data to unlock the different patterns in the data.

Restaurant Rating has become the most commonly used parameter for judging a restaurant for any individual. A lot of research has been done on different restaurants and the quality of food it serves. Rating of a restaurant depends on factors like reviews, area situated, average cost for two people, votes, cuisines and the type of restaurant. The main goal of this is to get insights on restaurants which people like visit and to identify the rating of the restaurant.

Zomato is India’s largest and most successful startup organization where food meets technology and connects thousands of restaurants in a thread. The objective of this project is to collect and process the dataset and visualize them accordingly to their votes and rating based system.

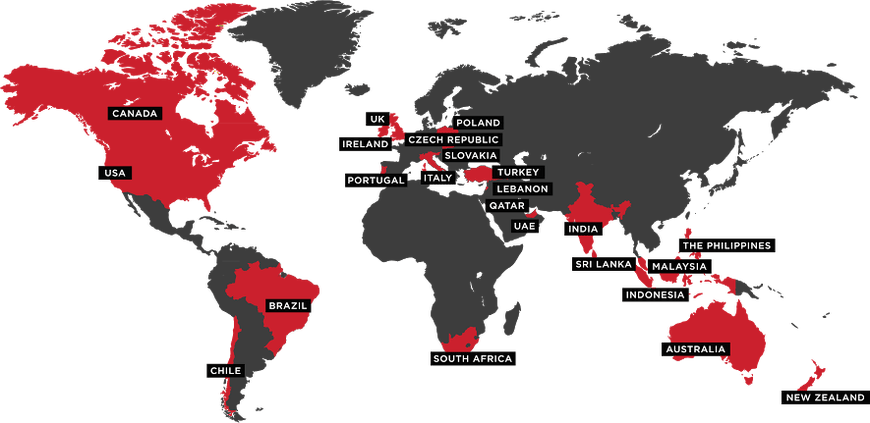
* 1. **About Zomato**

Zomato is the brainchild of Mr. Deepinder Goyal and Pankaj Chadda who were colleagues in ‘Bain And Company’.

The idea to start Zomato came from the demand for menu cards to order food among their colleagues. Mr. Deepinder Goyal and Pankaj Chadda initially started a database for food menu named “FoodieBay” in 2008 which soon had gone live with menus of1200 restaurants in Delhi NCR (India) by July 2008. Later the name was changed into Zomato (to avoid any legal implication as the name FoodieBay has EBay annexed with it) in November 2010 which is now the largest restaurant detecting platform in India listed with more than 4200 restaurant across 12 cities in the country. And internationally it has lists of about 1.4million restaurant across 10,000 cities in and at present it operates in 23 countries, including India, Australia and the United States. ZOMATO got its first funding of $1 million from InfoEdge and it’s also Zomato’s largest shareholder.

Zomato has over 2000 passionate companies across 23 countries, and this number is increasing day by day. Zomato is widely used by internet users who want information for dining and delivery options. Zomato is capturing customers by providing all the information a customer can need before going to a restaurant or ordering foods online. Zomato is following affiliate marketing model with website and mobile apps. Customers who are ordering foods online can pay through debit cards, credit cards and net banking.

Zomato has also partnered with Paytm, PayPal, and Freecharge to enable online payments. For marketing purpose, Zomato’s key strategy is it’s friendly website with up-to-date information of the restaurants. It also has a global mobile application from where it gets more than 50% of its total traffic. Zomato has partnered with Delhivery and has invested in Pickingo and Grab to quick up its delivery process.



**Fig. 1.1**

Source : Zomato

**CHAPTER 2**

**SOFTWARE REQUIREMENT & SYSTEM ANALYSIS**

**Software Requirement Analysis**

**Hardware Requirements**

RAM : 512 MB or above

I/O : Mouse, Keyboard

Processor : Any Processor above 1 Ghz

Hard Drive : Minimal Requirements

**Software Requirements**

Operating System : Windows XP and above

Coding Language : Python

Modules : matplotlib, plotLy, seaborn, etc.

**About the Software**

**Coding Language - Python**

1. Python is a widely used general-purpose, high-level programming language.
2. Python allows programming in Object-Oriented and Procedural paradigms.
3. Python programs generally are smaller than other programming languages like Java. Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.
4. Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber etc.
5. The biggest strength of the Python is large library which can be used for the following
   * Machine Learning
   * GUI Applications (like Kivy, Tkinter, PyQt etc. )
   * Web frameworks like Django (used by YouTube, Instagram, Dropbox)
   * Image processing (like OpenCV, Pillow)
   * Web scraping (like Scrapy, BeautifulSoup, Selenium)
   * Test frameworks
   * Multimedia
   * Scientific computing
   * Text processing

**Anaconda IDE**

Anaconda is a [free and open-source](https://en.wikipedia.org/wiki/Free_and_open-source) distribution of the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) and [R](https://en.wikipedia.org/wiki/R_(programming_language)) programming languages for [scientific computing](https://en.wikipedia.org/wiki/Scientific_computing) ([data science](https://en.wikipedia.org/wiki/Data_science), [machine learning](https://en.wikipedia.org/wiki/Machine_learning) applications, large-scale data processing, [predictive analytics](https://en.wikipedia.org/wiki/Predictive_analytics), etc.), that aims to simplify [package management](https://en.wikipedia.org/wiki/Package_management) and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS. It is developed and maintained by Anaconda, Inc., which was founded by Peter Wang and [Travis Oliphant](https://en.wikipedia.org/wiki/Travis_Oliphant) in 2012. As an Anaconda, Inc. product, it is also known as Anaconda Distribution or Anaconda Individual Edition, while other products from the company are Anaconda Team Edition and Anaconda Enterprise Edition, which are both not free.

**Jupyter Notebook**

Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

The Jupyter Notebook App is a server-client application that allows editing and running [notebook documents](https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html#notebook-document) via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remote server and accessed through the internet.

In addition to displaying/editing/running notebook documents, the Jupyter Notebook App has a “Dashboard” ([Notebook Dashboard](https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html#dashboard)), a “control panel” showing local files and allowing to open notebook documents or shutting down their [kernels](https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html#kernel).

**For data visualization, we will using these four libraries:**

**NumPy- NumPy**is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

**matplotlib** – Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

**seaborn** – Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**pandas- pandas data frame** is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns). A data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. pandas data frame consists of three principal components, the **data**, **rows**, and **columns**.

**import** **numpy** **as** **np**

**import** **pandas** **as** **pd**

**import** **matplotlib.pyplot** **as** **plt**

**import** **seaborn** **as** **sns**

**Data Frame** - A data frame is a two-dimensional array, with labeled axes (rows and columns). A data frame is a standard way to store data.

Data frame is well-known by statistician and other data practitioners. A data frame is a tabular data, with rows to store the information and columns to name the information. For instance, the price can be the name of a column and 2,3,4 the price values.

**Webpage Development**

**HTML –** First developed by Tim Berners-Lee in 1990, HTML is short for Hypertext Markup Language. HTML is used to create electronic documents (called pages) that are displayed on the World Wide Web. Each page contains a series of connections to other pages called hyperlinks. Every web page you see on the Internet is written using one version of HTML code or another.

HTML code ensures the proper formatting of text and images for your Internet browser. Without HTML, a browser would not know how to display text as elements or load images or other elements. HTML also provides a basic structure of the page, upon which Cascading Style Sheets are overlaid to change its appearance. One could think of HTML as the bones (structure) of a web page, and CSS as its skin (appearance).

**CSS -** Cascading Style Sheets, fondly referred to as CSS, is a simple design language intended to simplify the process of making web pages presentable.

CSS handles the look and feel part of a web page. Using CSS, you can control the color of the text, the style of fonts, the spacing between paragraphs, how columns are sized and laid out, what background images or colors are used, layout designs,variations in display for different devices and screen sizes as well as a variety of other effects.

CSS is easy to learn and understand but it provides powerful control over the presentation of an HTML document. Most commonly, CSS is combined with the markup languages HTML or XHTML.

**Visual Studio Code**

Visual Studio Code is a lightweight but powerful source code editor which runs on your desktop and is available for Windows, macOS and Linux. It comes with built-in support for JavaScript, TypeScript and Node.js and has a rich ecosystem of extensions for other languages (such as C++, C#, Java, Python, PHP, Go) and runtimes (such as .NET and Unity).

**CHAPTER 3**

**DATA DESCRIPTION AND**

**PRE- PROCESSING**

**Data Description & Pre-Processing**

Chennaiis the capital of the Indian state of Tamil Nadu. Located on the Coromandel Coast off the Bay of Bengal, it is the biggest cultural, economic and educational centre of South India. According to the 2011 Indian census, it is the sixth-most populous city and fourth-most populous urban agglomeration in India. The city together with the adjoining regions constitute the Chennai Metropolitan Area, which is the 36th-largest urban area by population in the world. Chennai is among the most-visited Indian cities by foreign tourists. It was ranked the 43rd-most visited city in the world for the year 2015.

Chennai has a unique food culture. Restaurants from all over the world can be found here in Chennai, with various kind of cuisines. Some might even say that Chennai is the best place for foodies. The food industry is always at a rise in Chennai, with 7,000 plus restaurants currently active in the city, the number is still increasing.

The growing number of restaurants and dishes in Chennai is what attracts us to inspect the data to get some insights, some interesting facts and figures.  
So, we will be analyzing the Zomato restaurant data for the city, Chennai.

The dataset is taken from Kaggle.  
Courtesy of Pranav Hari, the data is accurate to that available on the Zomato website until August 2019.

The dataset contains the following features:

1. *url* : This feature contains the url of the restaurant on the Zomato website
2. *address*: This feature contains the address of the restaurant in Chennai
3. *name* : This feature contains the name of the restaurant
4. *rate* : contains the overall rating of the restaurant out of 5
5. *votes*: contains total number of up votes for the restaurant
6. *price for two* : contains the cost for two people
7. *location* : contains the neighborhood in which the restaurant is located
8. *cuisines* : food styles, separated by comma
9. *top dishes* : contains the name of top dishes of the particular restaurant

**Link for the Dataset :**

**DATASET DESCRIPTION:**

<https://www.kaggle.com/phiitm/chennai-zomato-restaurants-data>

**Data Context**

The basic idea of analyzing this Zomato dataset is to get a fair idea about the factors affecting the establishment of different types of restaurants at different places in Chennai. This Zomato data aims at analyzing demography of the location. Most importantly it will help new restaurants in deciding their theme, menu, cuisine, cost, etc. for a particular location in Chennai. It also aims at finding similarity between neighborhoods of Chennai on the basis of food.

Data analytics allow businesses to understand their efficiency and performance, and ultimately helps the business make more informed decisions. For example, an e-commerce company might be interested in analyzing customer attributes in order to display targeted ads for improving sales. Data analysis can be applied to almost any aspect of a business if one understands the tools available to process information.

**Load the Data**

After acquiring the required dataset, we upload it into Jupyter using the appropriate command.

**import** **numpy** **as** **np**

**import** **pandas** **as** **pd**

**import** **matplotlib.pyplot** **as** **plt**

**import** **seaborn** **as** **sns**

**#load the data**

df = pd.read\_csv('cdata.csv')

dataset = df.copy()

**Basic Data Understanding**

Let’s start with basic data understanding by checking the information of the columns in which we are interested to work with.

df.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 7442 entries, 0 to 7441

**Output:**  
Data columns (total 9 columns):

Zomato URL 7442 non-null object

Name of Restaurant 7442 non-null object

Address 7442 non-null object

Location 7442 non-null object

Cuisine 7442 non-null object

Top Dishes 7442 non-null object

Price for 2 7442 non-null int64

Ratings 7442 non-null object

No of Votes 7442 non-null object

dtypes: int64(1), object(8)

memory usage: 523.4+ KB

**Data Cleaning and Manipulation (Preprocessing of Data)**

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. Preprocessing the data plays a vital role in data processing. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. It is a proven method of resolving such issues. There are many important steps in data preprocessing, such as data cleaning, data transformation, and feature selection. Data cleaning and transformation are methods used to remove outliers and standardize the data so that they take a form that can be easily used to create a model.  Data preprocessing prepares raw data for further processing. In this step we are going to eliminate all the null values and invalid data by using the appropriate commands.

*#checking for any null values*

df['Ratings'].isnull()

Output:

0 False

1 False

2 False

3 False

4 False

...

7437 False

7438 False

7439 False

7440 False

7441 False

Name: Ratings, Length: 7442, dtype: bool

The values are determined as “invalid” and not as “N/A" so we need to create a new

function for checking and determining the null values in the Data Frame.

missing\_values = ["invalid"]

df = pd.read\_csv("cdata.csv", na\_values = missing\_values)

df.isnull().sum()

Output:

Zomato URL 0

Name of Restaurant 0

Address 0

Location 0

Cuisine 0

Top Dishes 0

Price for 2 0

Ratings 2635

No of Votes 2639

dtype: int64

df.isnull().sum().sum()

Output:

5274

df.dropna()

Output:



**Feature Selection**

The next step in Data Preprocessing is to remove the additional information in the dataset.

For this we can remove the entire columns to make the dataset more compact and easy to process. To simplify the analysis, we will drop some of the columns that are not very useful like Zomato URL, Address and Top Dishes.

We did not use any feature selection algorithms but eliminated some columns due to available domain knowledge and thorough study of the system. Dropped columns mentioned below:

**del** df['Zomato URL']

**del** df['Address']

**del** df['Top Dishes']

We will rename a few columns for convenience

df = df.rename(columns = {"Name of Restaurant" : "restaurant\_name", "No of Votes" : "Votes"})

df.head()

Output:

|  | **restaurant\_name** | **Location** | **Cuisine** | **Price for 2** | **Ratings** | **Votes** |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | The Black Pearl | Sholinganallur | North Indian, European, Mediterranean, Contine... | 1500 | 4.9 | 2857.0 |
| 1 | Coal Barbecues | T. Nagar | North Indian, Mediterranean, Asian, Arabian, BBQ | 1400 | 4.9 | 7320.0 |
| 2 | Palmshore | Medavakkam | North Indian, Chinese, Arabian | 700 | 4.5 | 441.0 |
| 3 | Fromage | MRC Nagar | Cafe, Italian, French, European | 1000 | 4.5 | 1073.0 |
| 4 | Savoury Sea Shell | Anna Nagar East | Arabian, Chinese, North Indian, Lebanese, Salad | 1400 | 4.1 | 2446.0 |

With this step the Preprocessing of the Data Frame is done.

**CHAPTER 4**

**EXPLORATORY DATA ANALYSIS (EDA)**

1. **EXPLORATORY DATA ANALYSIS**

EDA plays a critical role in understanding what, why, and how of a problem statement. It’s first in the order of operations that a data analyst will perform when handed a new data source and problem statement.

Here’s a direct definition: exploratory data analysis is an approach to analyzing data sets by summarizing their main characteristics with visualizations. The EDA process is a crucial step prior to building a model in order to unravel various insights that later become important in developing a robust algorithmic model.

Let’s try to break down this definition and understand different operations where EDA comes into play. First and foremost, EDA provides a stage for breaking down problem statements into smaller experiments which can help understand the dataset. EDA provides relevant insights which help analysts make key business decisions The EDA step provides a platform to run all thought experiments and ultimately guides us towards making a critical decision

A lot of effort went into the EDA as it gives us a detailed knowledge of our data. Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to:

• maximize insight into a data set;

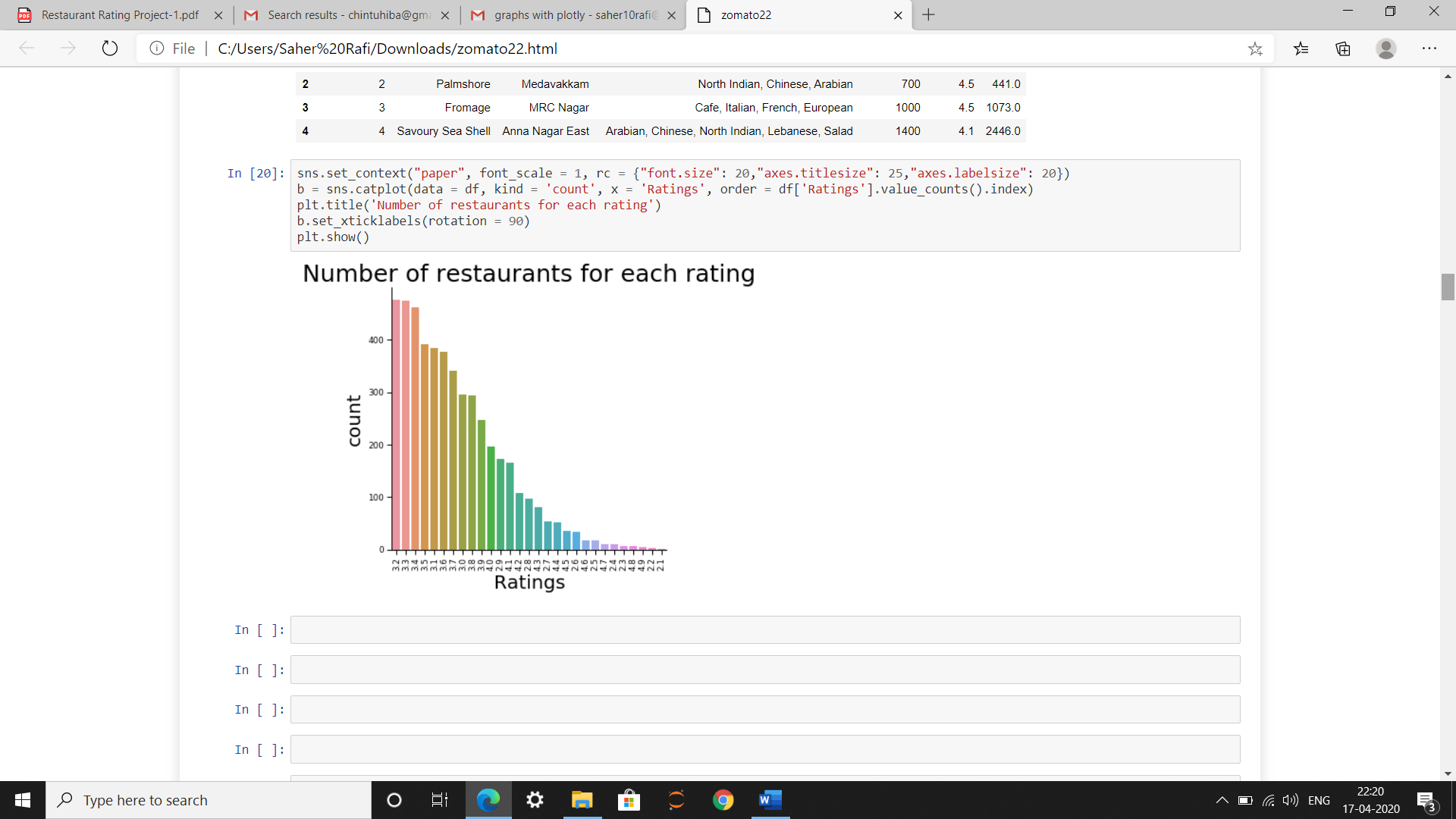
• uncover underlying structure;

• extract important variables;

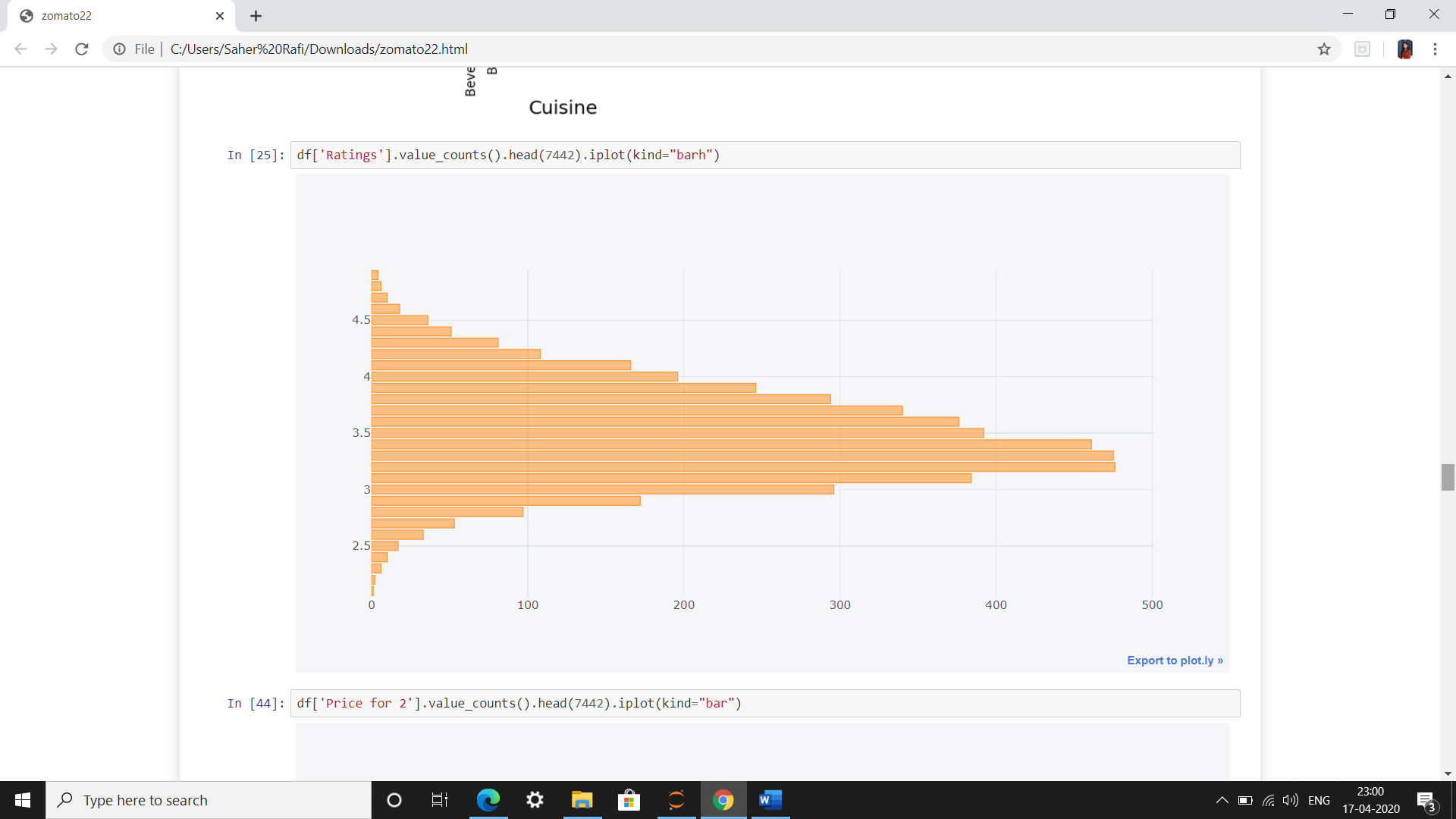
• detect outliers and anomalies;

• test underlying assumptions.

**Restaurant Rate Distribution**



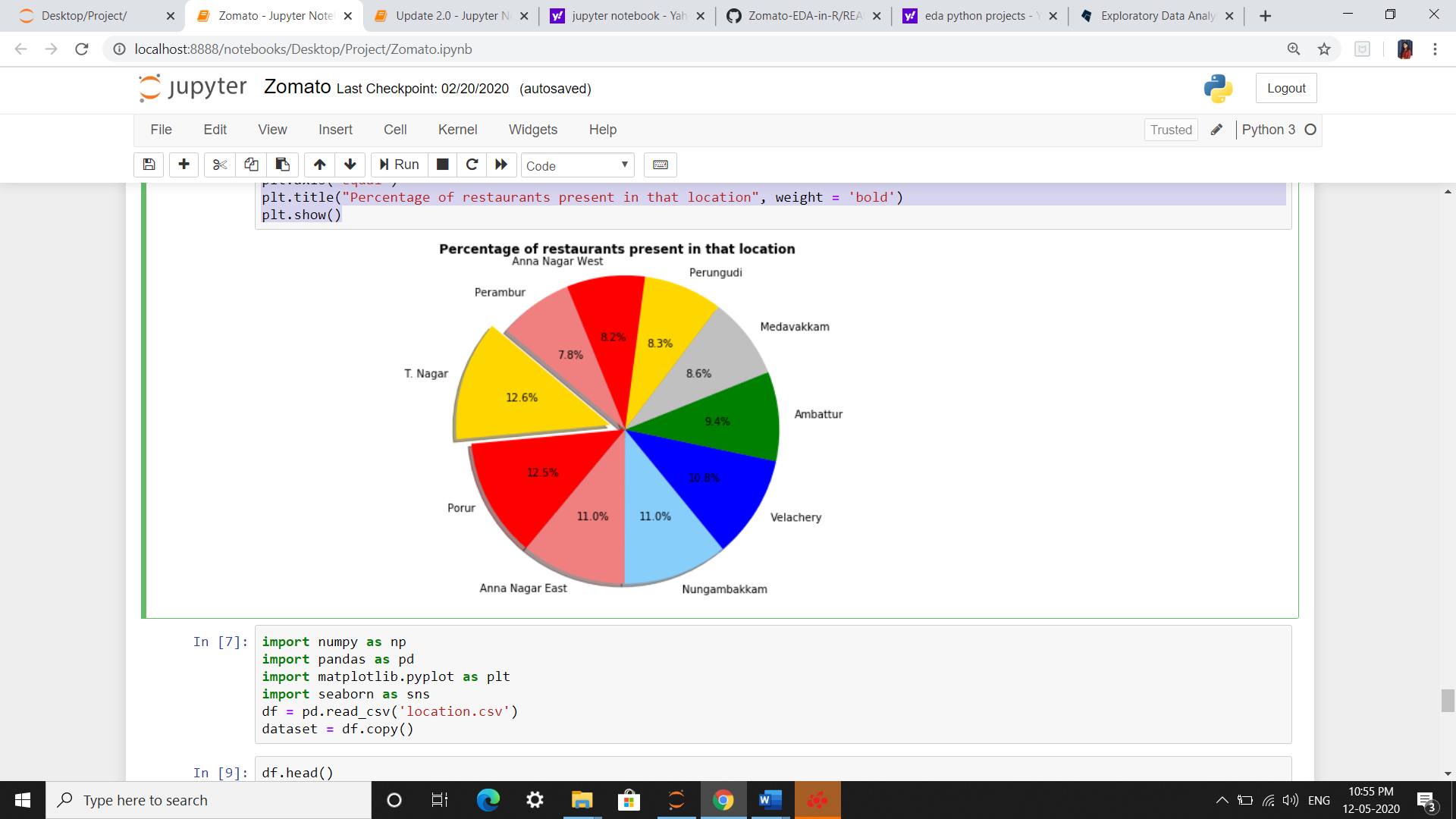
**Figure 4. 1**



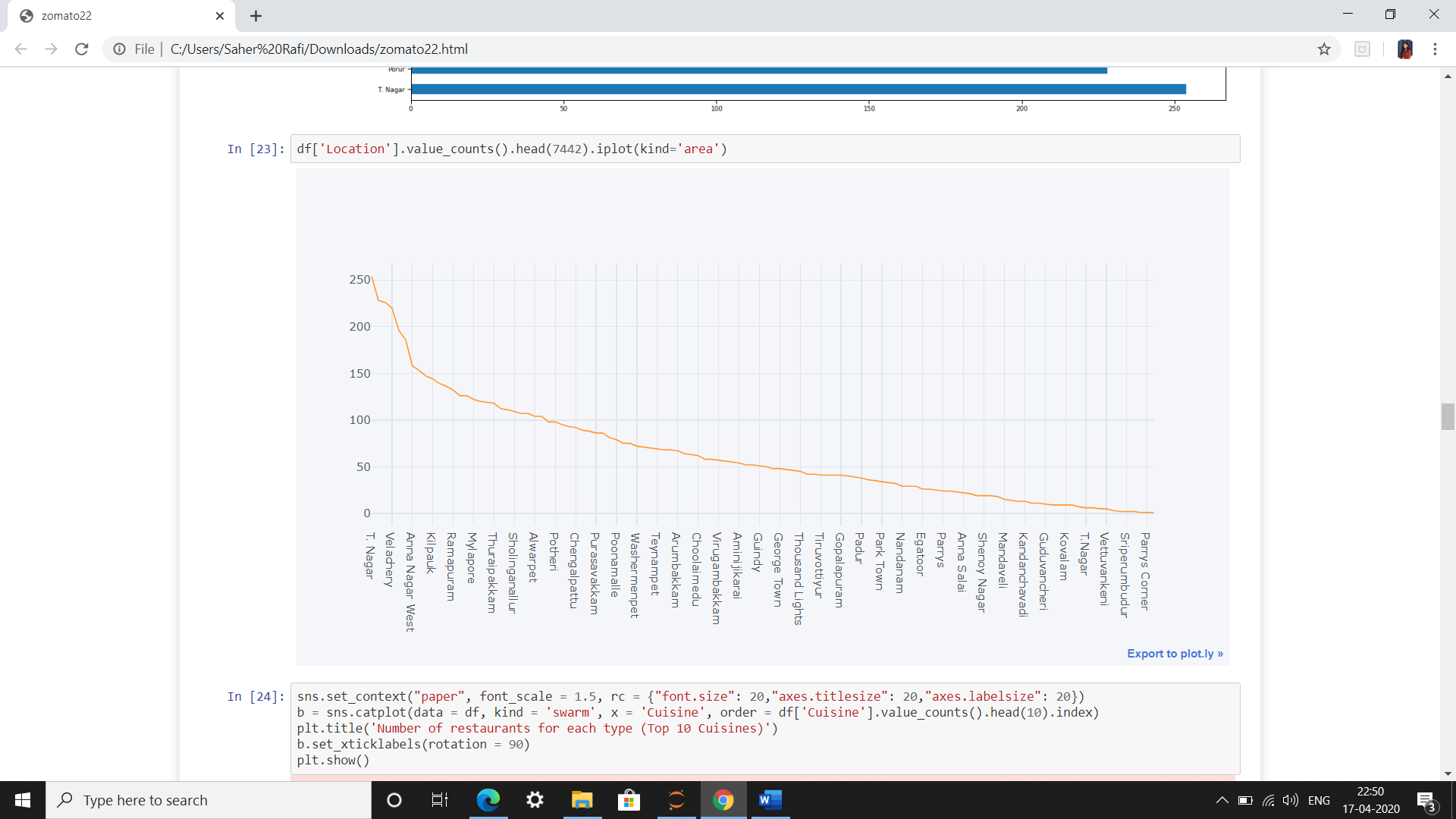
**Figure 4. 2**

The majority of restaurants are rated 3.2(out of 5), followed by 3.3 and 3.4, which are pretty decent ratings. This means that most of the restaurants at Chennai are liked by the citizens and hence rated above average. There are very few restaurants with very high rating like 4.9 or 4.8.

**Restaurant Frequency with Respect to Location**



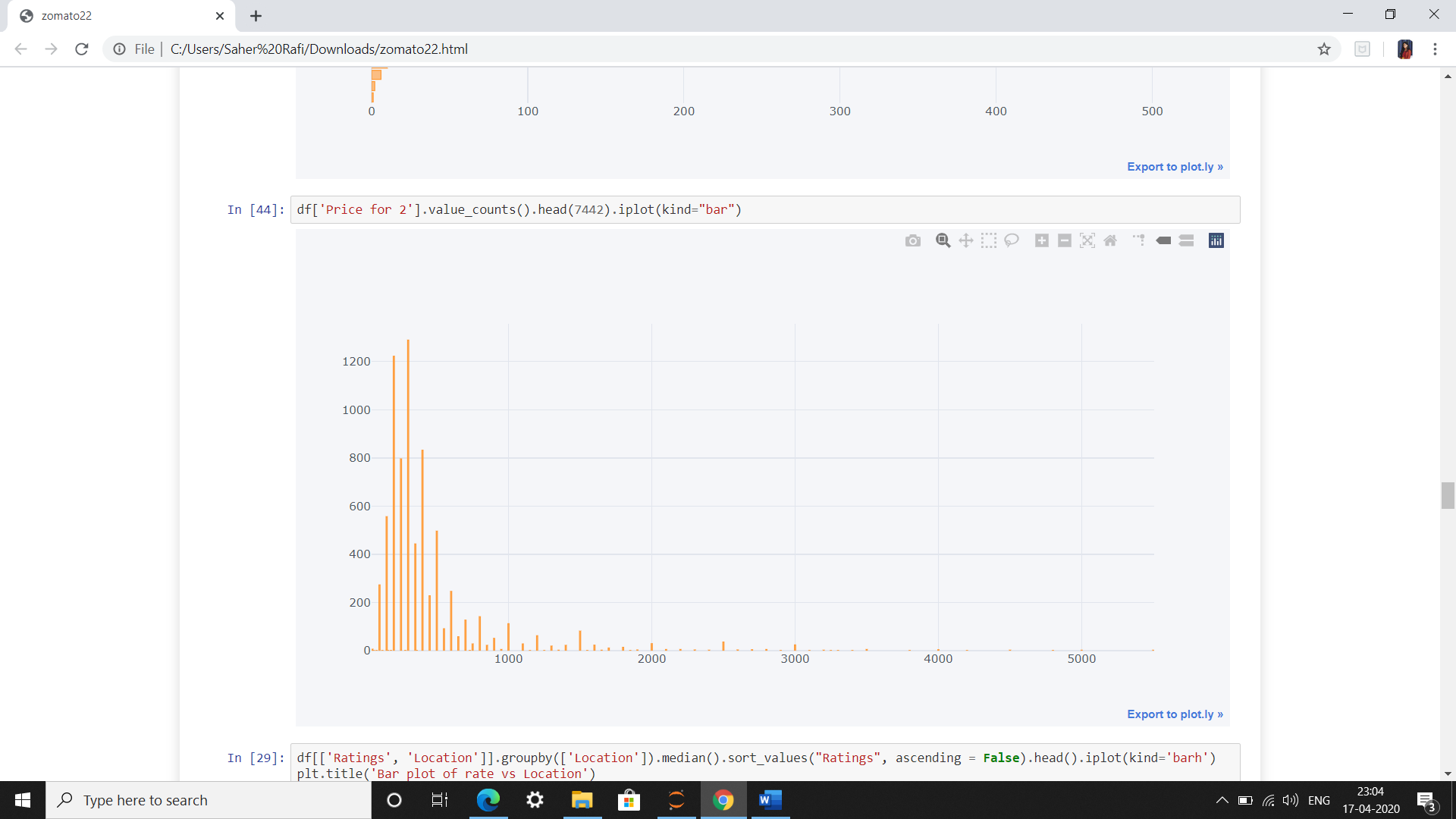
**Figure 4. 3**



**Figure 4. 4**

T.Nagar has the highest number of restaurants, followed by Velachery. Parrys Corner has the least number of restaurants, followed by Sriperumbudur. It seems that the main foodies live in T.Nagar and Velachery.

**Restaurant Frequency with Respect to Price (Price for 2)**



**Figure 4. 5**

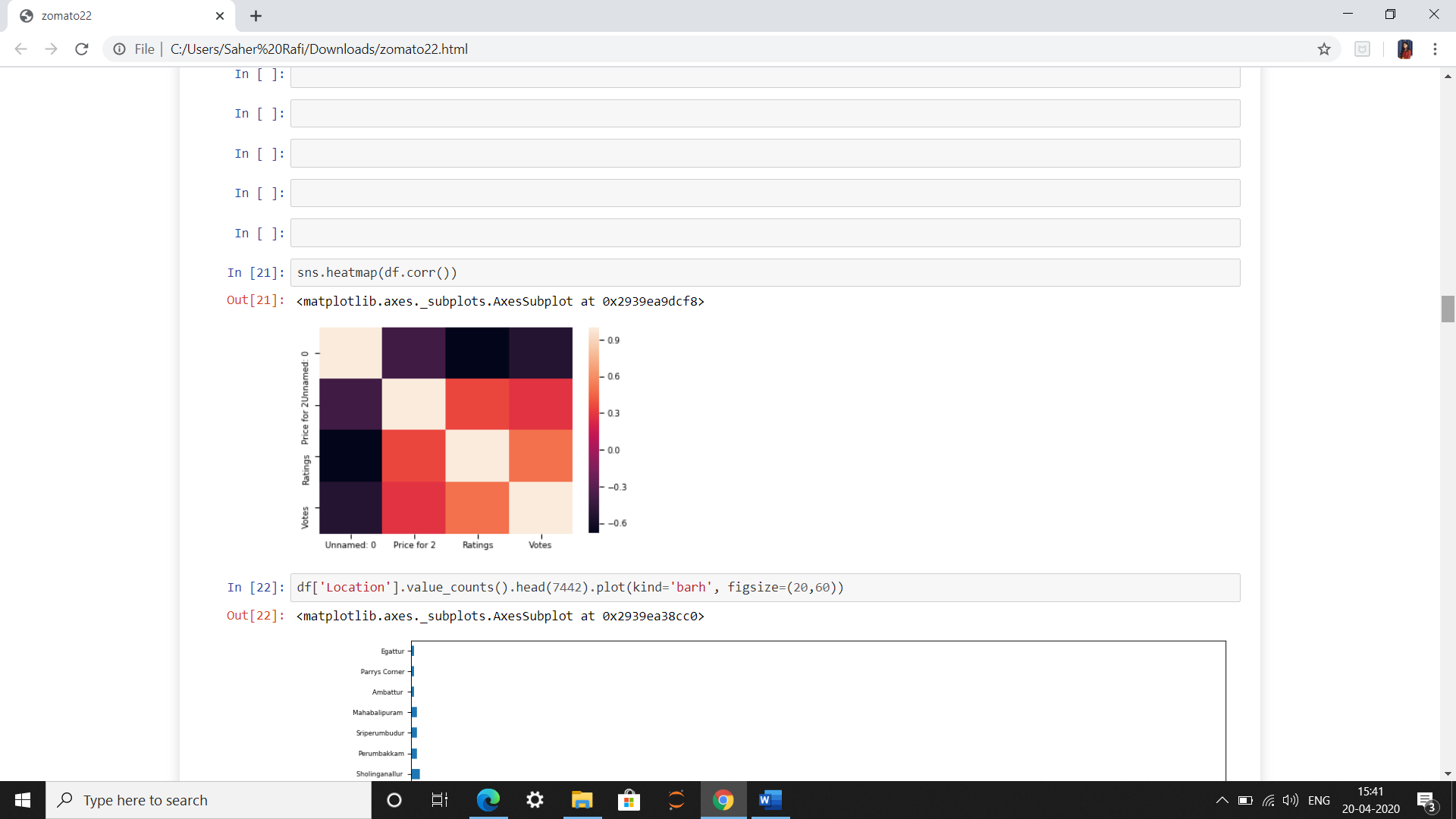
The majority of restaurants in Chennai have the average cost(for two) 300, followed by 200 and 400. This means that the competition is very high as far as the less costly restaurants are considered. The costlier ones are less in number and so the competition is slim.

**Frequency Table of Rate vs Location (Highest Price for 2)**



**Figure 4. 6**

**Correlation Heat Map**



**Figure 4. 7**

A correlation heatmap uses colored cells, typically in a monochromatic scale, to show a 2D correlation matrix (table) between two discrete dimensions or event types. The values of the first dimensions appear as rows of the table, while the values of the second dimension are represented by the columns of the table. The color value of the cells is proportional to the number of measurements that match the dimensional values. This enables you to quickly identify incidence patterns, and to recognize anomalies.

**Top Rated Restaurants:**



**Figure 4. 8**

The top rated restaurants in Chennai city is Coal Barbecues which is followed by Black Pearl. Both the restaurants are a live buffet and unlimited grill type of restaurant. We come to know that Live Barbecue and Buffet style dining is mostly preferred by the people living here.

**CHAPTER 5**

**REGRESSION ANALYSIS**

**Regression Analysis:**

Regression is another important and broadly used statistical and machine learning tool. The key objective of regression-based tasks is to predict output labels or responses which are continues numeric values, for the given input data. The output will be based on what the model has learned in training phase. Basically, regression models use the input data features (independent variables) and their corresponding continuous numeric output values (dependent or outcome variables) to learn specific association between inputs and corresponding outputs.

**Types of Regression Algorithms:**

* Linear Regression
* Logistic Regression
* Random Forest Regression
* Decision Tree Regression

**Linear Regression:**

Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

*Y=mX+bY=mX+b*

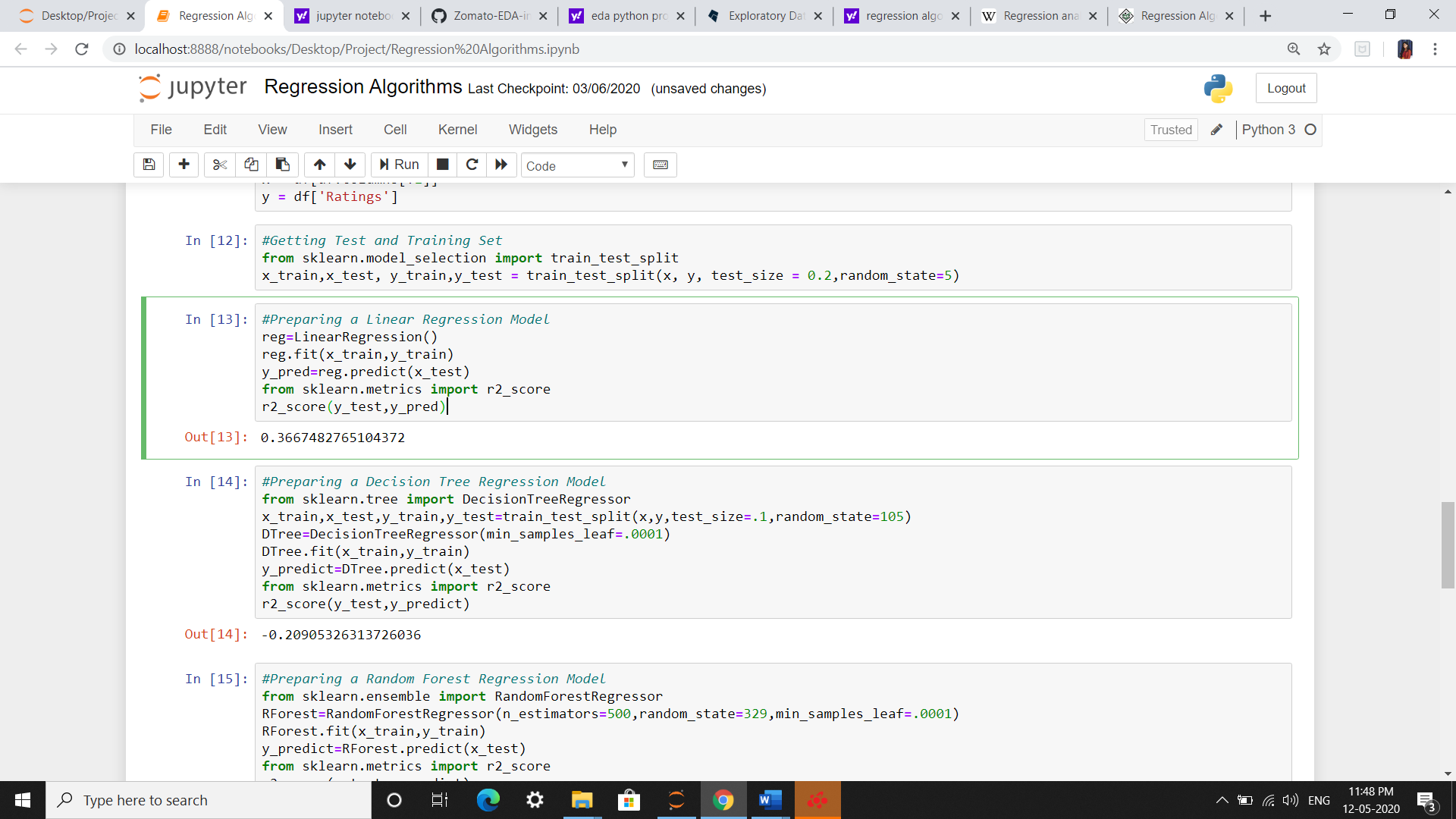
Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the 𝑌Y-intercept. If X = 0,Y would be equal to 𝑏b.

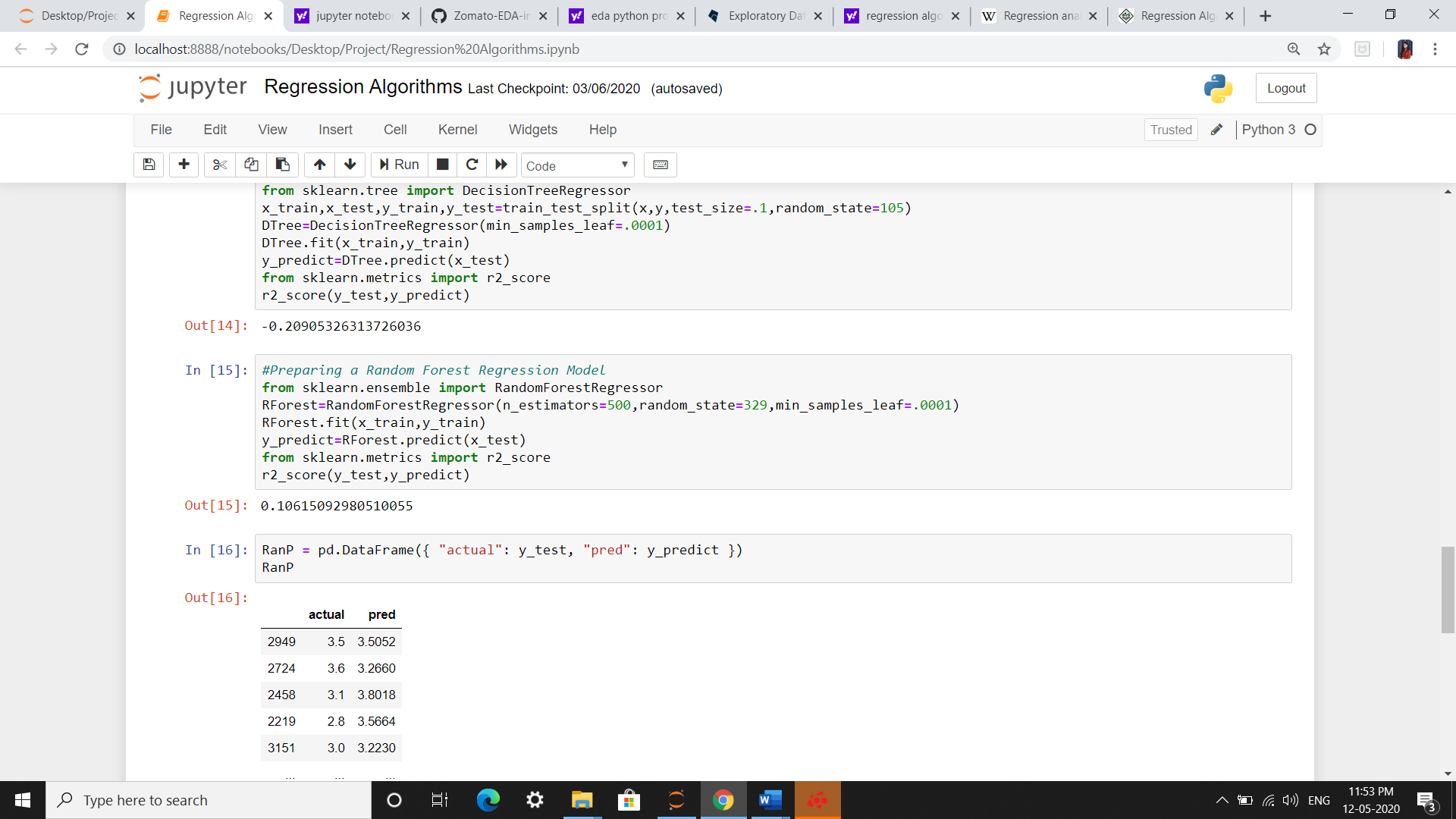
**Preparing a Linear Regression Model:**



**Random Forest Regression:**

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

**Preparing a Random Forest Regression Model:**

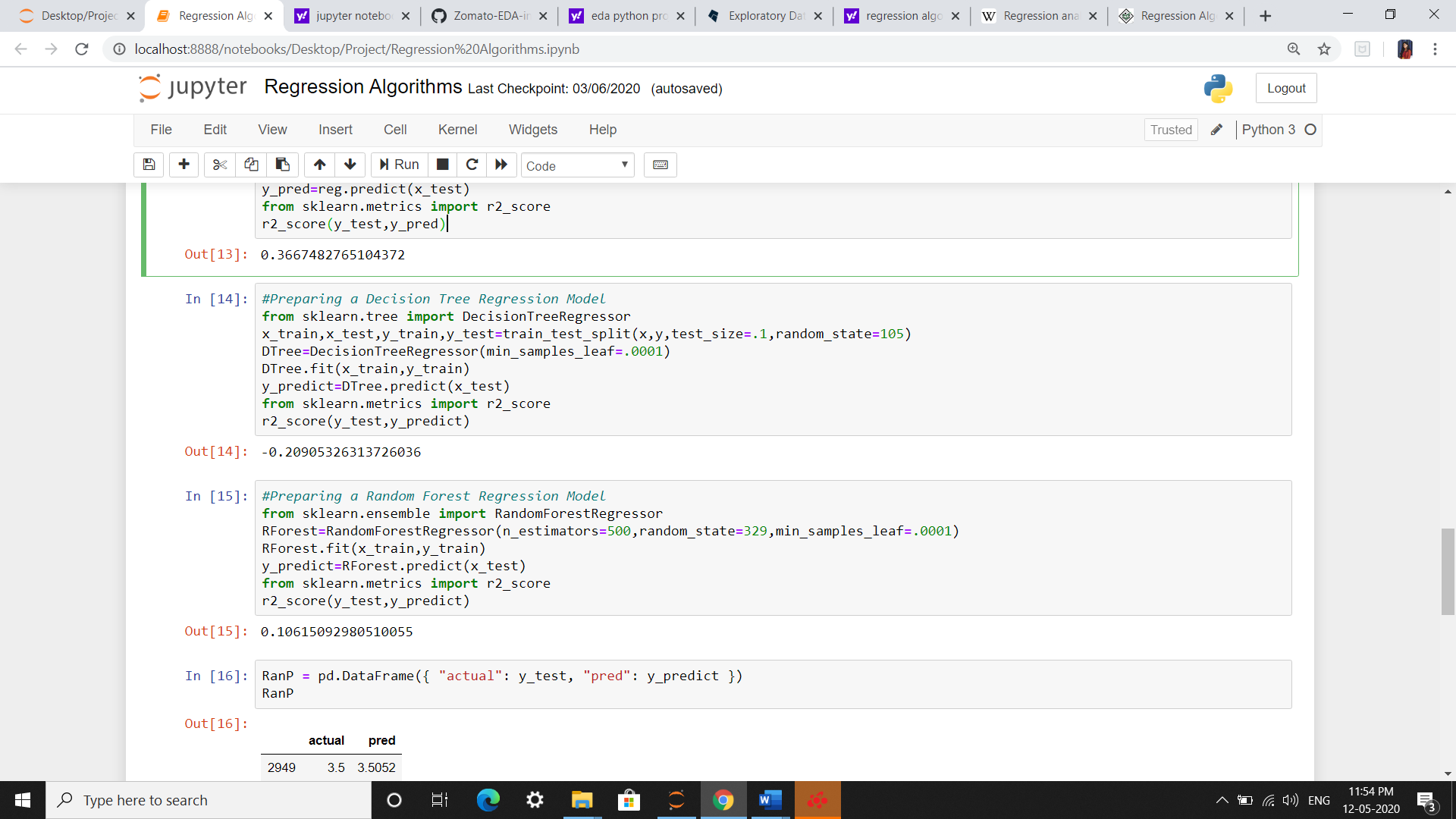


**Decision Tree Regression:**

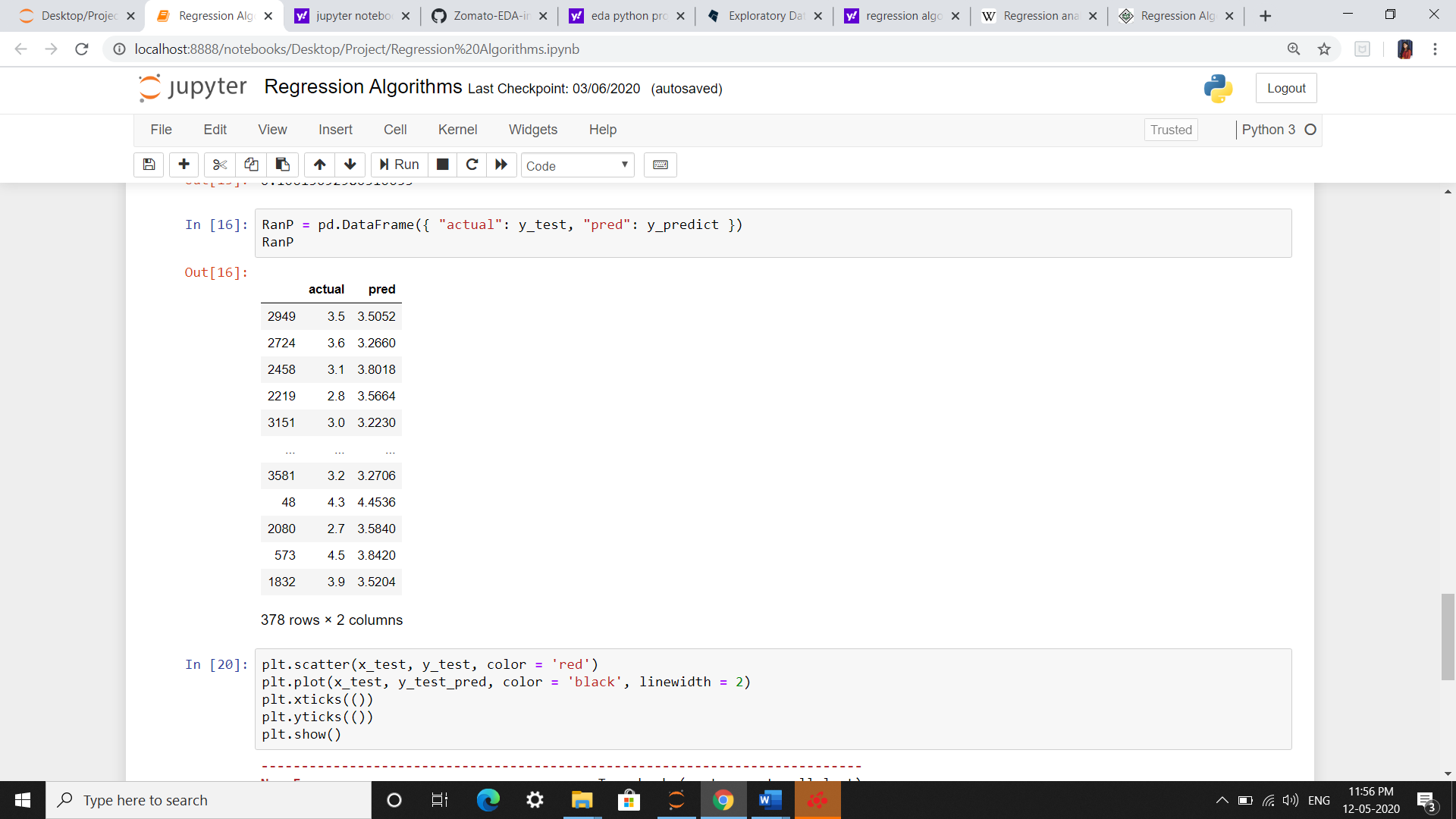
Decision tree analysis is a predictive modelling tool that can be applied across many areas. Decision trees can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions. Decisions trees are the most powerful algorithms that falls under the category of supervised algorithms.

They can be used for both classification and regression tasks. The two main entities of a tree are decision nodes, where the data is split and leaves, where we got outcome.

**Preparing a Decision Tree Regression Model:**



We take the Random Forest Regression into consideration as it gives the best predicted value and then we compare the predicted values with the actual values.



**CHAPTER 6**

**SOURCE CODES**

**Source Codes:**

**For the Data Pre-processing and Basic Charts:**

**Importing & Reading the Dataset and Modules**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_csv('cdata.csv')

dataset = df.copy()

df.info()

df.head()

**Checking and Removing the Extra and Null Values**

df['Ratings'].isnull()

missing\_values = ["invalid"]

df = pd.read\_csv("cdata.csv", na\_values = missing\_values)

df.isnull().sum()

df.isnull().sum().sum()

df.dropna()

df.to\_csv('updated.csv')

del df['Zomato URL']

del df['Address']

del df['Top Dishes']

df = df.rename(columns = {"Name of Restaurant" : "restaurant\_name", "No of Votes" : "Votes"})

len(df['Location'].value\_counts())

**For Basic Charts:**

sns.set\_context("paper", font\_scale = 1, rc = {"font.size": 20,"axes.titlesize": 25,"axes.labelsize": 20})

b = sns.catplot(data = df, kind = 'Count', x = 'Ratings', order = df['Ratings'].value\_counts().index)

plt.title('Number of restaurants for each rating')

b.set\_xticklabels(rotation = 90)

plt.show()

sns.set\_context("paper", font\_scale = 1.5, rc = {"font.size": 20,"axes.titlesize": 20,"axes.labelsize": 20})

b = sns.catplot(data = df, kind = 'Count', x = 'Cuisine', order = df['Cuisine'].value\_counts().head(10).index)

plt.title('Number of restaurants for each type (Top 10 Cuisines)')

b.set\_xticklabels(rotation = 90)

plt.show()

sns.set\_context("paper", font\_scale = 2, rc = {"font.size": 20,"axes.titlesize": 25,"axes.labelsize": 20})

b = sns.catplot(data = df, kind = 'count', x = 'Cuisine', order = df['Cuisine'].value\_counts().tail(10).index)

plt.title('Number of restaurants for each type (Bottom 10)')

b.set\_xticklabels(rotation = 90)

plt.show()

plt.figure(figsize = (12,6))

names = df['Location'].value\_counts()[:10].index

values = df['Location'].value\_counts()[:10].values

colors = ['gold', 'red', 'lightcoral', 'lightskyblue','blue','green','silver']

explode = (0.1, 0, 0, 0,0,0,0,0,0,0) # explode 1st slice

plt.pie(values, explode=explode, labels=names, colors=colors,autopct='%1.1f%%', shadow=True, startangle=140)

plt.axis('equal')

plt.title("Percentage of restaurants present in that location", weight = 'bold')

plt.show()

**For the PlotLy charts:**

**Number of restaurants for each rating**

sns.set\_context(&quot;paper&quot;, font\_scale = 1, rc = {&quot;font.size&quot;:

20,&quot;axes.titlesize&quot;: 25,&quot;axes.labelsize&quot;: 20})

b = sns.catplot(data = df, kind = &#39;count&#39;, x = &#39;Ratings&#39;, order =

df[&#39;Ratings&#39;].value\_counts().index)

plt.title(&#39;Number of restaurants for each rating&#39;)

b.set\_xticklabels(rotation = 90)

plt.show()

**Restaurant Frequency with Respect to Location**

df[&#39;Ratings&#39;].value\_counts().head(7442).iplot(kind=&quot;barh&quot;)

Restaurant Frequency with Respect to Price (Price for 2)

df[&#39;Location&#39;].value\_counts().head(7442).iplot(kind=&#39;area&#39;)

Frequency Table of Rate vs Location (Highest Price for 2)

df[[&#39;Ratings&#39;,

&#39;Location&#39;]].groupby([&#39;Location&#39;]).median().sort\_values(&quot;Ratings&quot;,

ascending = False).head().iplot(kind=&#39;barh&#39;)

plt.title(&#39;Bar plot of rate vs Location&#39;)

plt.show()

**Bar plot of average cost vs names for first 15 restaurants (ordered by Price for**

**2 feature)**

df[[&#39;Price for 2&#39;,

&#39;restaurant\_name&#39;]].groupby([&#39;restaurant\_name&#39;]).median().sort\_values(

&quot;Price for 2&quot;, ascending = False).head(15).iplot(kind=&#39;barh&#39;)

plt.title(&#39;Bar plot of average cost vs names for first 15

restaurants(ordered by Price for 2 feature)&#39;)

plt.show()

**Correlation Heat Map**

sns.heatmap(df.corr())

**Bar plot of rate vs names for top 15 restaurants**

df[[&#39;Ratings&#39;,

&#39;restaurant\_name&#39;]].groupby([&#39;restaurant\_name&#39;]).median().sort\_values(

&quot;Ratings&quot;, ascending = False).head(15).iplot(kind=&#39;scatter&#39;)

plt.title(&#39;Bar plot of rate vs names for top 15 restaurants&#39;)

plt.show()

**For Regression Algorithms:**

import numpy as np

import pandas as pd

import re

import matplotlib.pyplot as plt

import statsmodels.api as sm

from matplotlib.pyplot import figure

%matplotlib inline

import os

df = pd.read\_csv('loc2.csv')

dataset = df.copy()

**Importing Regression Models:**

from sklearn.linear\_model import LogisticRegression

from sklearn.linear\_model import LinearRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import classification\_report

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import r2\_score

df['Price for 2'] = df['Price for 2'].astype(str)

df['Price for 2'] = df['Price for 2'].apply(lambda x: x.replace(',','.'))

df['Price for 2'] = df['Price for 2'].astype(float)

df.info()

df.describe()

df.corr()

df.plot.scatter(x='Ratings', y='Votes')

**Encoding:**

def Encode(df):

for column in df.columns[~df.columns.isin(['Ratings', 'Price for 2', 'Votes'])]:

df[column] = df[column].factorize()[0]

return df

df\_en = Encode(df.copy())

x = df[df.columns[:1]]

y = df['Ratings']

**Training and Testing Data:**

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test, y\_train,y\_test = train\_test\_split(x, y, test\_size = 0.2,random\_state=5)

**Linear Regression Model:**

reg=LinearRegression()

reg.fit(x\_train,y\_train)

y\_pred=reg.predict(x\_test)

from sklearn.metrics import r2\_score

r2\_score(y\_test,y\_pred)

**Decision Tree Regression Model:**

from sklearn.tree import DecisionTreeRegressor

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=.1,random\_state=105)

DTree=DecisionTreeRegressor(min\_samples\_leaf=.0001)

DTree.fit(x\_train,y\_train)

y\_predict=DTree.predict(x\_test)

**Random Forest Regression Model:**

from sklearn.metrics import r2\_score

r2\_score(y\_test,y\_predict)

from sklearn.ensemble import RandomForestRegressor

RForest=RandomForestRegressor(n\_estimators=500,random\_state=329,min\_samples\_leaf=.0001)

RForest.fit(x\_train,y\_train)

y\_predict=RForest.predict(x\_test)

from sklearn.metrics import r2\_score

r2\_score(y\_test,y\_predict)

**Random Forest Model – Actual Values v/s Predicted Values**

RanP = pd.DataFrame({ "actual": y\_test, "pred": y\_predict })

RanP

**Website Source Code**

HTML Source Code:

<!DOCTYPE html>

<html lang="en">

<head>

<meta charset="UTF-8">

<meta name="viewport" content="width=device-width, initial-scale=1.0">

<title>zomato</title>

<link rel="stylesheet" type="text/css" href="style.css" />

<link href="css/font-awesome.min.css" rel="stylesheet" />

<link rel="shortcut icon" href="images/favicon.png" />

</head>

<body>

<section class="intro">

<nav>

<a href="#" id="menu-icon"></a>

<ul>

<li><a href="#about">About zomato</a></li>

<li><a href="#our">Exploratory Data analysis</a></li>

<li><a href="#team">My Team</a></li>

</ul>

</nav>

<div class="inner">

<div class="content">

<h1>EXPLORATORY VISUALIZATION AND

ANALYSIS OF RESTAURANT DATASET

</h1>

</div>

</div>

</section>

<a name="about">

<div class="clearfix"></div>

<section class="left-col">

<h2>About zomato</h2>

<p style="text-indent: 6%;">The idea to start Zomato came from the demand for menu cards to order food among their colleagues. Mr. Deepinder Goyal and Pankaj Chadda initially started a database for food menu named “FoodieBay” in 2008 which soon had gone live with menus of1200 restaurants in Delhi NCR (India) by July 2008. Later the name was changed into Zomato (to avoid any legal implication as the name FoodieBay has EBay annexed with it) in November 2010 which is now the largest restaurant detecting platform in India listed with more than 4200 restaurant across 12 cities in the country. And internationally it has lists of about 1.4million restaurant across 10,000 cities in and at present it operates in 23 countries, including India, Australia and the United States. ZOMATO got its first funding of $1 million from InfoEdge and it’s also Zomato’s largest shareholder. </p>

</section>

<section class="sidebar">

<img src="https://cdn.iconscout.com/icon/free/png-512/zomato-1937646-1637644.png" >

</section>

<div class="clearfix"></div>

<a name="our">

<h2>EXPLORATORY DATA ANALYSIS</h2>

<section class="one-third-our">

<img src="images/img1.png" target="\_blank">

</section>

<section class="one-third-our">

<img src="images/img2.png" target="\_blank">

</section>

<section class="one-third-our">

<img src="images/img3.png" target="\_blank">

</section>

<section class="one-third-our">

<img src="images/img4.png" target="\_blank">

</section>

<section class="one-third-our">

<img src="images/img5.png" target="\_blank">

</section>

<div class="clearfix"></div>

<a name="team">

<section class="left-col">

<h2>My Team</h2>

<p style="text-indent: 6%;">As a part of BCA course, this project is a statistical analysis of Zomato’s dataset. The objective behind doing this project is to gain insights of the food culture in Chennai and the factors that influence it.</p>

</section>

<section class="sidebar">

<section class="contact">

<p> <br class="break"> AMMAR MUJEEB <br class="break"> SAHER FATHIMA <br class="break"> SALIF HAMMAD </p>

</section>

</section>

</body>

</html>

**CSS Source Code:**

html,body{

margin: 0;

padding: 0;

height: 100%;

width: 100%;

}

h2 {

font-family: 'Playfair Display',serif;

font-size: 200%;

font-weight: 700;

color: #2c2c2c;

text-align: center;

text-decoration: underline;

}

p {

font-family: 'Raleway',serif;

color: #2c2c2c;

font-size: 140%;

line-height: 180%;

text-align: justify;

font-weight: 300;

}

img {

text-align: center;

max-width: 100%;

height: auto;

width: auto;

}

.intro {

height: 100%;

width: 100%;

background-image: url(images/food.jpg);

background-size: cover;

display: table;

top: 0;

}

.intro .inner{

display: table-cell;

vertical-align: middle;

width: 100%;

max-width: none;

}

.content{

max-width: 600%;

margin: auto;

}

h1{

font-family: 'Playfair Display',serif;

font-size: 200%;

font-weight: 700;

color: #fff;

text-align: center;

padding: 2%;

text-shadow: 0px 0px 250px #000;

}

.inner{

max-width: 600px;

margin: auto;

font-size: 170%;

line-height: 1.6;

padding: 10px;

}

/\*--- Start Nav --\*/

nav{

float: left;

padding: 5% 5% 0 0;

height: 50px;

}

nav ul{

list-style: none;

float: left;

}

nav ul li{

font-family: 'raleway',sans-serif;

font-size: 120%;

}

nav ul li a{

color: #fff;

text-decoration: none;

}

nav ul li a:hover {

text-decoration: underline;

}

nav ul, nav:active ul{

display: none;

padding: 8px 0;

left: 45px;

top: 60px;

width: 20%;

z-index: 1000;

}

nav li{

width: 100%;

padding: 8px 0 10px 7px;

margin: 0;

}

nav:hover ul{

display: block;

position: absolute;

}

#menu-icon{

width: 30px;

height: 26px;

position: fixed;

background: url(images/icon1.png) center;

display: inline-block;

margin: 50px 0 0 50px;

z-index: 2;

cursor: pointer;

}

/\*---End Nav --\*/

.left-col{

width: 60%;

float: left;

margin: 4% 0 4% 4%;

}

.sidebar{

width: 26%;

float: left;

margin: 4%;

}

.one-third-our{

width: 90.333333%;

padding: 4%;

float: left;

}

.one-third-our img:hover {

opacity: .6;

}

.contact p{

text-align: center;

letter-spacing: 4px;

text-decoration: underline;

color: #2c2c2c;

}

.clearfix{

clear: both;

height: 5%;

}

@media screen and (max-width: 780px) {

#menu-icon{

margin: 26px 0 0 26px;

}

nav ul,nav:active ul {

padding: 8px 0;

left: 23px;

top: 34px;

width: 50%;

}

.inner{

font-size: 120%;

}

.content h1{

font-size: 180%;

}

p {

font-size: 120%;

}

h2 {

font-size: 160%;

}

.left-col {

width: 95%;

margin: 0 0 3% 0;

padding: 0 2%;

}

.sidebar{

width: 96%;

margin: o auto;

}

img {

padding: 2%;

}

.one-third-our {

width: 96%;

margin: 0 auto;

padding: 0;

}

.break {

display: block;

}

}

html,body{

margin: 0;

padding: 0;

height: 100%;

width: 100%;

}

h2 {

font-family: 'Playfair Display',serif;

font-size: 200%;

font-weight: 700;

color: #2c2c2c;

text-align: center;

text-decoration: underline;

}

p {

font-family: 'Raleway',serif;

color: #2c2c2c;

font-size: 140%;

line-height: 180%;

text-align: justify;

font-weight: 300;

}

img {

text-align: center;

max-width: 100%;

height: auto;

width: auto;

}

.intro {

height: 100%;

width: 100%;

background-image: url(images/food.jpg);

background-size: cover;

display: table;

top: 0;

}

.intro .inner{

display: table-cell;

vertical-align: middle;

width: 100%;

max-width: none;

}

.content{

max-width: 600%;

margin: auto;

}

h1{

font-family: 'Playfair Display',serif;

font-size: 200%;

font-weight: 700;

color: #fff;

text-align: center;

padding: 2%;

text-shadow: 0px 0px 250px #000;

}

.inner{

max-width: 600px;

margin: auto;

font-size: 170%;

line-height: 1.6;

padding: 10px;

}

/\*--- Start Nav --\*/

nav{

float: left;

padding: 5% 5% 0 0;

height: 50px;

}

nav ul{

list-style: none;

float: left;

}

nav ul li{

font-family: 'raleway',sans-serif;

font-size: 120%;

}

nav ul li a{

color: #fff;

text-decoration: none;

}

nav ul li a:hover {

text-decoration: underline;

}

nav ul, nav:active ul{

display: none;

padding: 8px 0;

left: 45px;

top: 60px;

width: 20%;

z-index: 1000;

}

nav li{

width: 100%;

padding: 8px 0 10px 7px;

margin: 0;

}

nav:hover ul{

display: block;

position: absolute;

}

#menu-icon{

width: 30px;

height: 26px;

position: fixed;

background: url(images/icon1.png) center;

display: inline-block;

margin: 50px 0 0 50px;

z-index: 2;

cursor: pointer;

}

/\*---End Nav --\*/

.left-col{

width: 60%;

float: left;

margin: 4% 0 4% 4%;

}

.sidebar{

width: 26%;

float: left;

margin: 4%;

}

.one-third-our{

width: 90.333333%;

padding: 4%;

float: left;

}

.one-third-our img:hover {

opacity: .6;

}

.contact p{

text-align: center;

letter-spacing: 4px;

text-decoration: underline;

color: #2c2c2c;

}

.clearfix{

clear: both;

height: 5%;

}

@media screen and (max-width: 780px) {

#menu-icon{

margin: 26px 0 0 26px;

}

nav ul,nav:active ul {

padding: 8px 0;

left: 23px;

top: 34px;

width: 50%;

}

.inner{

font-size: 120%;

}

.content h1{

font-size: 180%;

}

p {

font-size: 120%;

}

h2 {

font-size: 160%;

}

.left-col {

width: 95%;

margin: 0 0 3% 0;

padding: 0 2%;

}

.sidebar{

width: 96%;

margin: o auto;

}

img {

padding: 2%;

}

.one-third-our {

width: 96%;

margin: 0 auto;

padding: 0;

}

.break {

display: block;

}

}

**CHAPTER 7**

**OUTPUT SCREENS**

**CHAPTER 8**

**CONCLUSION**

**Conclusion**

Exploratory data analysis is a powerful tool. A diligent EDA is an absolute must to put your advanced business analytics in the right direction. EDA provides a great opportunity to test your simple business hypotheses and hunches before jumping into a rigorous model building. Exploratory Data Analysis is valuable to data science projects since it allows to get closer to the certainty that the future results will be *valid*, *correctly interpreted*, and *applicable* to the desired business contexts. Such level of certainty can be achieved only after raw data is validated and checked for anomalies, ensuring that the data set was collected without errors. EDA also helps to find insights that were not evident or worth investigating to business stakeholders and data scientists but can be very informative about a particular business.

EDA is performed in order to define and refine the selection of feature variables that will be used for machine learning. Once data scientists become familiar with the data set, they often have to return to feature engineering step, since the initial features may turn out not to be serving their intended purpose. Once the EDA stage is complete, data scientists get a firm feature set they need for supervised and unsupervised machine learning.

We based our analysis keeping restaurant business in mind. We tried to figure out answers to some of the common queries when opening any new restaurant.

* We figured T-Nagar, Velachery and Nungambakkam are good places to start restaurant. T-Nagar has most number of unique restaurants and can be cheaper to get started.
* Large number of votes can ensure better rating and 1K for 2 people is good to go price.

**CHAPTER 8**

**FURTHER ENHANCEMENTS**

**Further Enchancements**

As this EDA is done using a specific dataset which was found in Kaggle, it is valid for a specific period of time. The restaurant data keeps changing and new restaurants are being updated every week and thus the whole hierarchy changes with respect to the data.

Everytime a dataset is updated, we could try to update our analysis with respect to the new data that is being put and that will help us to find new values and analysis.

This can also lead to a more effective approach and a higher prediction accuracy of the values.

**CHAPTER 9**

**REFERENCES/BIBLIOGRAPHY**

**REFERENCES/BIBLIOGRAPHY**

# 1.  [David Beazley](https://www.amazon.in/s/ref=dp_byline_sr_ebooks_1?ie=UTF8&field-author=David+Beazley&text=David+Beazley&sort=relevancerank&search-alias=digital-text), [Brian K. Jones](https://www.amazon.in/Brian-K-Jones/e/B0053YB48U/ref=dp_byline_cont_ebooks_2) - Python Cookbook: Recipes for Mastering Python 3, 2008

# 2. [Paul Barry](https://www.amazon.in/Paul-Barry/e/B001HMX52K/ref=dp_byline_cont_ebooks_1) - Head First Python: A Brain-Friendly Guide, 2010

3. <https://www.kaggle.com/search?q=zomato>

4. <https://www.analyticsvidhya.com/blog/2018/05/starters-guide-jupyter-notebook/>