PROJECT
R E P O R T

O n

“EDA-Analysis on Zomato Restaurants”

Submitted in partial fulfilment of the Requirements for the award of the Degree of

BACHELOR OF SCIENCE (INFORMATION TECHNOLOGY)

By

Ruturaj Bhosale - Seat No.: (1906004)

Under the esteemed guidance of

Mrs. Poonam Gajakosh

(Assistant Professor)

Designation

DEPARTMENT OF INFORMATION TECHNOLOGY

KLE SOCIETY’S SCIENCE AND COMMERCE COLLEGE

(Affiliated To University Of Mumbai)

KALAMBOLI, 410 218
MAHARASHTRA

2021-2022

PROFORMA FOR THE APPROVAL PROJECT PROPOSAL

PNR **No**.: ..... Roll no: **1906004.**

1. Name of the Student: - Ruturaj Bhosale .
2. Title of the Project: - EDA – Analysis on Zomato Restaurant
3. Name of the Guide: - Prof. Poonam Gajakosh
4. Teaching experience of the Guide: - 13 Years
5. Is this your first submission? Yes / No

Signature of the Student Signature of the Guide

Date: ..... Date: .....

Signature of the coordinator

Date: .....

KLE SOCIETY'S SCIENCE AND COMMERCE COLLEGE

(Affiliated to University of Mumbai)

KALAMBOLI, MAHARASHTRA, PIN-410 218

DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the project entitled, **"EDA – Analysis on Zomato Restaurant''**, is bonafied work of **Ruturaj Bhosale.** bearing Seat. No: (**1906004**) submitted in partial fulfillment of the requirements for the award of degree of BACHELOR OF SCIENCE in INFORMATION TECHNOLOGY from
University of Mumbai for academic year 2021-2022.

Internal Guide **Coordinator**

External Examiner

Date: **College Seal**

Name and signature of the Student

INDEX

1. INTRODUCTION

1.1 Abstract : -

Exploratory Data Analysis (EDA) is the process of visualizing and analyzing data to extract insights from it. In other words, EDA is the process of summarizing important characteristics of data in order to gain better understanding of the dataset.

So here we collect different Datasets and work on cleaning data and bring our own data models selecting the prototype and getting conclusion and insights.

Exploratory data analysis (EDA) is an iterative process where data scientists interact with data to extract information about their quality and shape as well as derive knowledge and new insights into the related domain of the dataset. However, data scientists are rarely experienced domain experts who have tangible knowledge about a domain. Integrating domain knowledge into the analytic process is a complex challenge that usually requires constant communication between data scientists and domain experts. For this reason, it is desirable to reuse the domain insights from exploratory analyses in similar use cases. With this objective in mind, we present a conceptual system design on how to extract domain expertise while performing EDA and utilize it to guide other data scientists in similar use cases. Our system design introduces two concepts, interaction storage and analysis context storage, to record user interaction and interesting data points during an exploratory analysis. For new use cases, it identifies historical interactions from similar use cases and facilitates the recorded data to construct candidate interaction sequences and predict their potential insight—i.e., the insight generated from performing the sequence.

1.2 Objective : -

The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as: a good-fitting, parsimonious model.

To Uncover a parsimonious model, one which explains the data with a minimum number of predictor variables.

To Identify the most influential variables.

To Create a list of outliers or other anomalies.

1.3 Purpose : -

The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set.

Here we will collect huge amount of data and churn it and make to it clean and understandable data which will bring insights on the basis of business driven models.

So here is an example of EDA I am working basically on Zomato restaurant dataset as an example here I have been bringing insights from a variety a data and we have build a Data Driven Model For example , Here we can analyse data and solve questions like where to open the restaurants , which restaurant has the best reviews by customers and which restaurant is influential in different types of cuisines which restaurant has more traffic on weekdays and weekends and many other questions . Also our end goal is to build a dashboard, which tells us what is lively happening and is understandable easily.

1.4 Proposed system : -

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Exploratory data analysis (EDA) is an iterative process where data scientists interact with data to extract information about their quality and shape as well as derive knowledge and new insights into the related domain of the dataset. However, data scientists are rarely experienced domain experts who have tangible knowledge about a domain. Integrating domain knowledge into the analytic process is a complex challenge that usually requires constant communication between data scientists and domain experts. For this reason, it is desirable to reuse the domain insights from exploratory analyses in similar use cases. With this objective in mind, we present a conceptual system design on how to extract domain expertise while performing EDA and utilize it to guide other data scientists in similar use cases. Our system design introduces two concepts, interaction storage and analysis context storage, to record user interaction and interesting data points during an exploratory analysis. For new use cases, it identifies historical interactions from similar use cases and facilitates the recorded data to construct candidate interaction sequences and predict their potential insight—i.e., the insight generated from performing the sequence.

Based on these predictions, the system recommends the sequences with the highest predicted insight to data scientist. We implement a prototype to test the general feasibility of our system design and enable further research in this area. Within the prototype, we present an exemplary use case that demonstrates the usefulness of recommended interactions. Finally, we give a critical reflection of our first prototype and discuss research opportunities resulting from our system design.

2. SURVEY OF TECHNOLOGIES

Technologies used in our project:

Python:
For developing the Allure we can use the Python technology.
Python is an open source, Server side web application. Python Framework Django allows developers to create web application, web services.
Given below are a few reasons why we choose the Python framework of building this system.

1) It reduces the amount of coding
    - We can simply drag and drop the necessary element.
    - We cannot write the html coding so that why it reduces the amount of coding.
2) It works fast
    - Due to the Python the code gets compiled into the "Machine Language" before any visitor view your system.
    - As Python enables data catching from database. So system works very fast.
3) It is language independent
    - Python is the platform independent language.

    Platform independent means that the application in different operating system.

Python Language: -

Python is a popular programming language.

It was created by GUIDO VAN ROSSUM and released in 1991.

It is used for:

- Mathematics and Statistics.
- Machine Learning
- Artificial Intelligence
- Web development(server-side).

- Software development.
- System scripting.

Why Python?

Python works on different platforms i.e. Windows, Mac, Linux, Raspberry Pi etc. Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.

Numpy

What Is Numpy?

Numpy is considered as one of the most popular machine learning library in Python.

TensorFlow and other libraries uses Numpy internally for performing multiple operations on Tensors. Array interface is the best and the most important feature of Numpy.

Features Of Numpy

1. Interactive: Numpy is very interactive and easy to use.
2. Mathematics: Makes complex mathematical implementations very simple.
3. Intuitive: Makes coding real easy and grasping the concepts is easy.
4. Lot of Interaction: Widely used, hence a lot of open source contribution.

Uses of Numpy?

This interface can be utilized for expressing images, sound waves, and other binary raw streams as an array of real numbers in N-dimensional.

For implementing this library for machine learning having knowledge of Numpy is important for full stack developers.

Pandas

What Is Pandas?

Pandas is a machine learning library in Python that provides data structures of high-level and a wide variety of tools for analysis. One of the great feature of this library is the ability to translate complex operations with data using one or two commands. Pandas have so many inbuilt methods for grouping, combining data, and filtering, as well as time-series functionality.

Features Of Pandas

Pandas make sure that the entire process of manipulating data will be easier. Support for operations such as Re-indexing, Iteration, Sorting, Aggregations, Concatenations and Visualizations are among the feature highlights of Pandas.

Applications of Pandas?

Currently, there are fewer releases of pandas library which includes hundred of new features, bug fixes, enhancements, and changes in API. The improvements in pandas regards its ability to group and sort data, select best suited output for the apply method, and provides support for performing custom types operations.

Data Analysis among everything else takes the highlight when it comes to usage of Pandas. But, Pandas when used with other libraries and tools ensure high functionality and good amount of flexibility.

3. REQUIREMENTS AND ANALYSIS : -

3.1 Problem Definition : -

So Here we have collected questions related to Business perspective from Zomato food platform who works as an interface between consumer and local business from both the POV.

1. Which types of restaurants do customers prefer according to rating ? Consumer do prefer example Veg and Nonveg ,contninental,Chinese?

2.In which area one should open a resto,franchise store and cloud kitchen for QSRs if there is already an restaurant?

Additional Sample question : In a area which timing do customers prefer going to resto and at what timing there is more traffic?

3.2. Software and Hardware Requirement : -

Developer Requirements: - Hardware requirement

• OS – Windows 10

• Processor – Intel(R) Pentium(R) CPU N3700 @ 1.60GHz 1.60 GHz

• Monitor - 15” color monitor

• Keyboard - 122 Keys

• System Type – 62-bit Operating System

• RAM – 4 GB or More.

Software requirement

• Operating system: – Windows 10

• Language: - Python (version 3.6)

Tools Used

• Editor : -

Colab

Jupyter

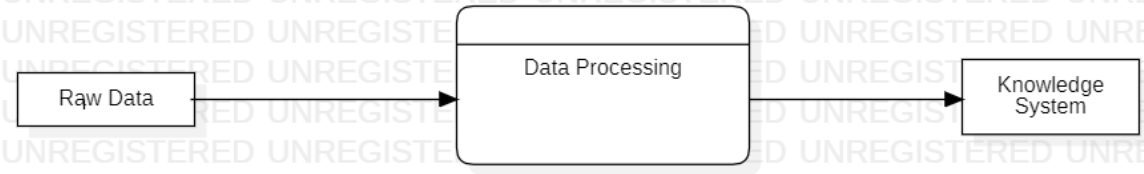
Kaggle

1. Visualization Tools Used

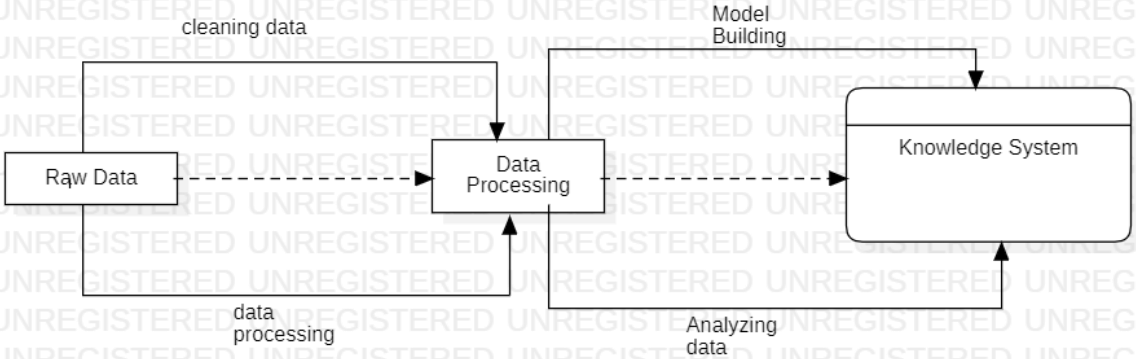
- Tabeau
- Power Bi

3.3 Conceptual Models : - UML

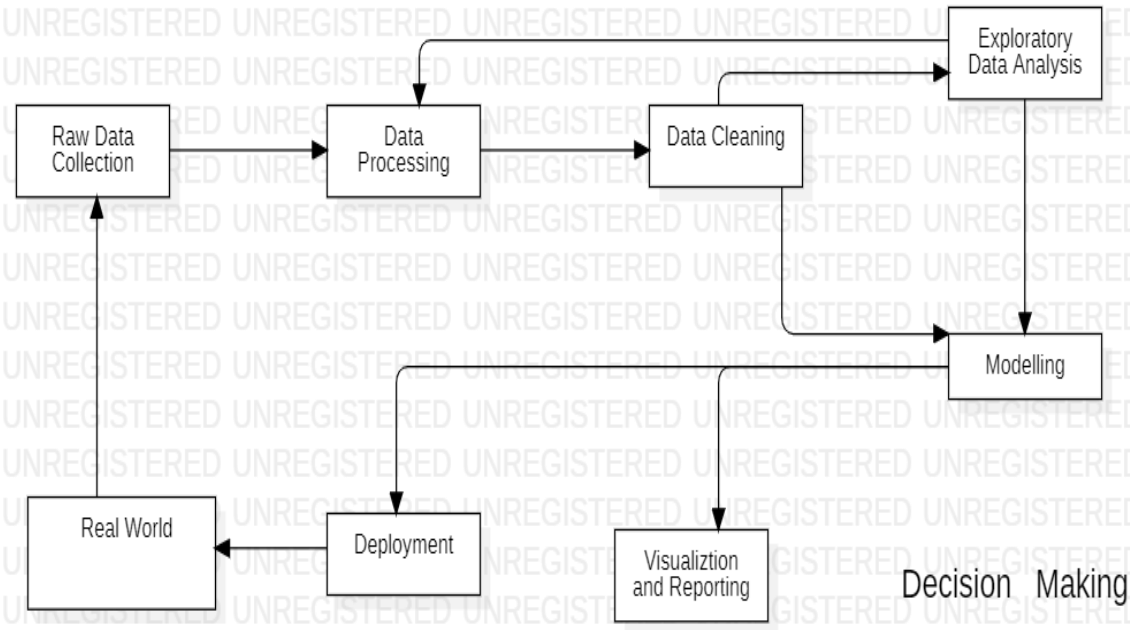
3.3.1 Data Flow Diagram 0 : -



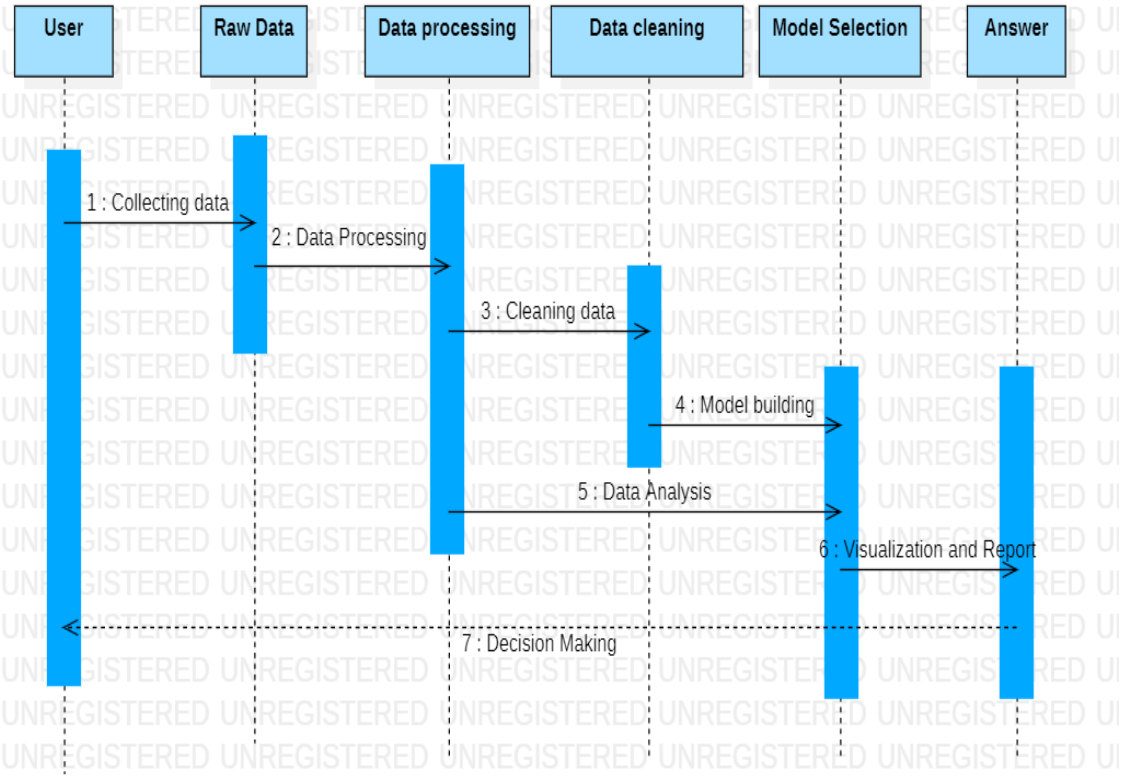
3.3.2 Data Flow Diagram 1 : -



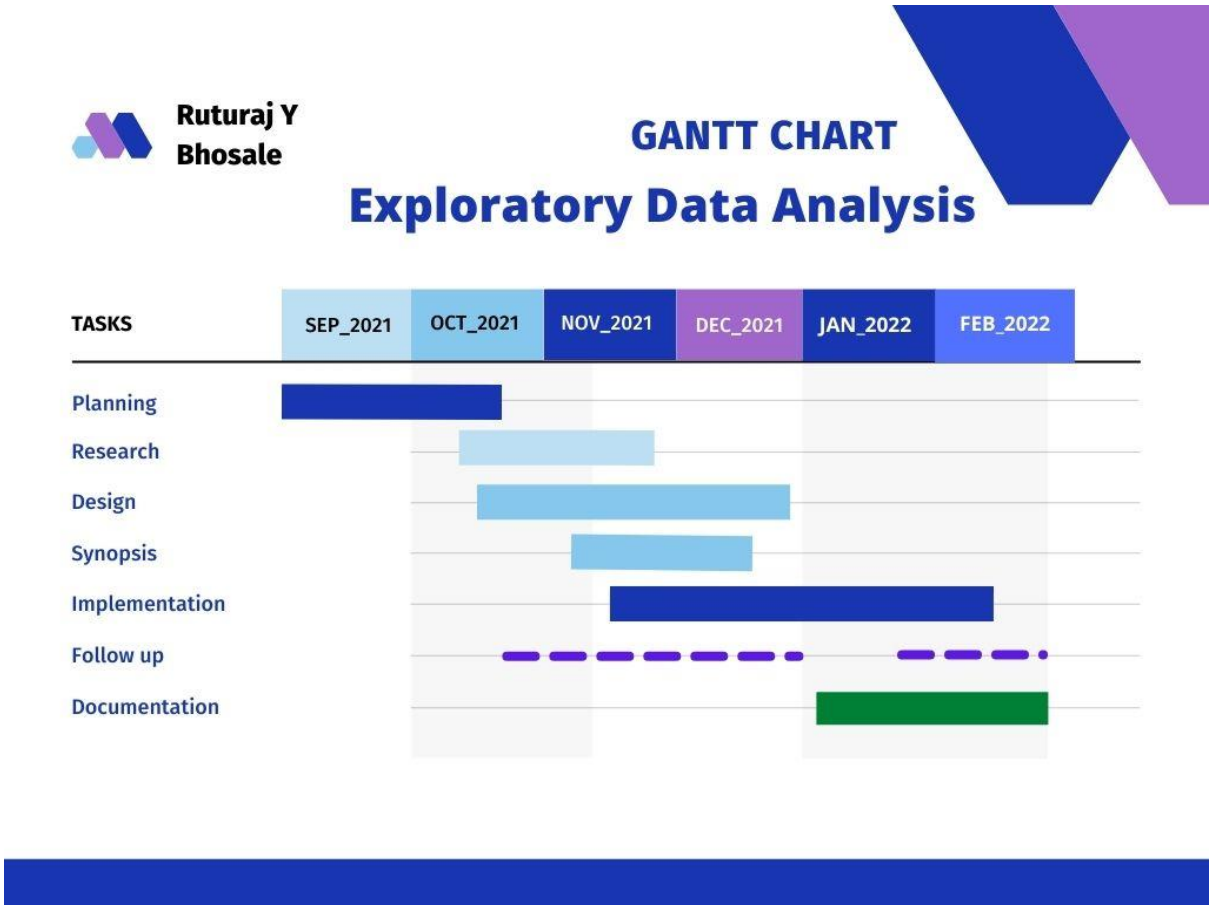
3.3.3 Data Flow Diagram 2 : -



3.3.4 Sequence Diagram : -



3.3.5 Gantt Chart : -



4. IMPLEMENTATION : -

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
```

```
df = pd.read_csv('../input/zomato-dataset-navi-mumbai/Zomato Analysis (1).csv', encoding='latin1')
df.head()
```

```
df.shape
```

```
df.columns
```

```
df = df.drop(['url','address','name','phone','reviews_list','menu_item'], axis=1)
df.head
```

```
df.info()
```

```
df.drop_duplicates(inplace = True)
df.shape
```

```
df['Dining(rate)'].unique()
df['Delivery(rate)'].unique()
```

```
def handlerate(value):
    if(value=='NEW' or value=='-' or value=='N/A'):
        return np.nan
    else:
        return float(value)

df['Dining(rate)'] = df['Dining(rate)'].apply(handlerate)
df['Dining(rate)'].head()
def handlerate(value):
    if(value=='NEW' or value=='-' or value=='N/A'):
        return np.nan
    else:
        return float(value)

df['Delivery(rate)'] = df['Delivery(rate)'].apply(handlerate)
df['Delivery(rate)'].head()
```

```
df['Delivery(rate)'].fillna(df['Delivery(rate)'].mean(), inplace = True)
df['Delivery(rate)'].isnull().sum()
```

```
df['Dining(rate)'].fillna(df['Dining(rate)'].mean(), inplace = True)
df['Dining(rate)'].isnull().sum()
```

```
df.dropna(inplace = True)
df.head()
```

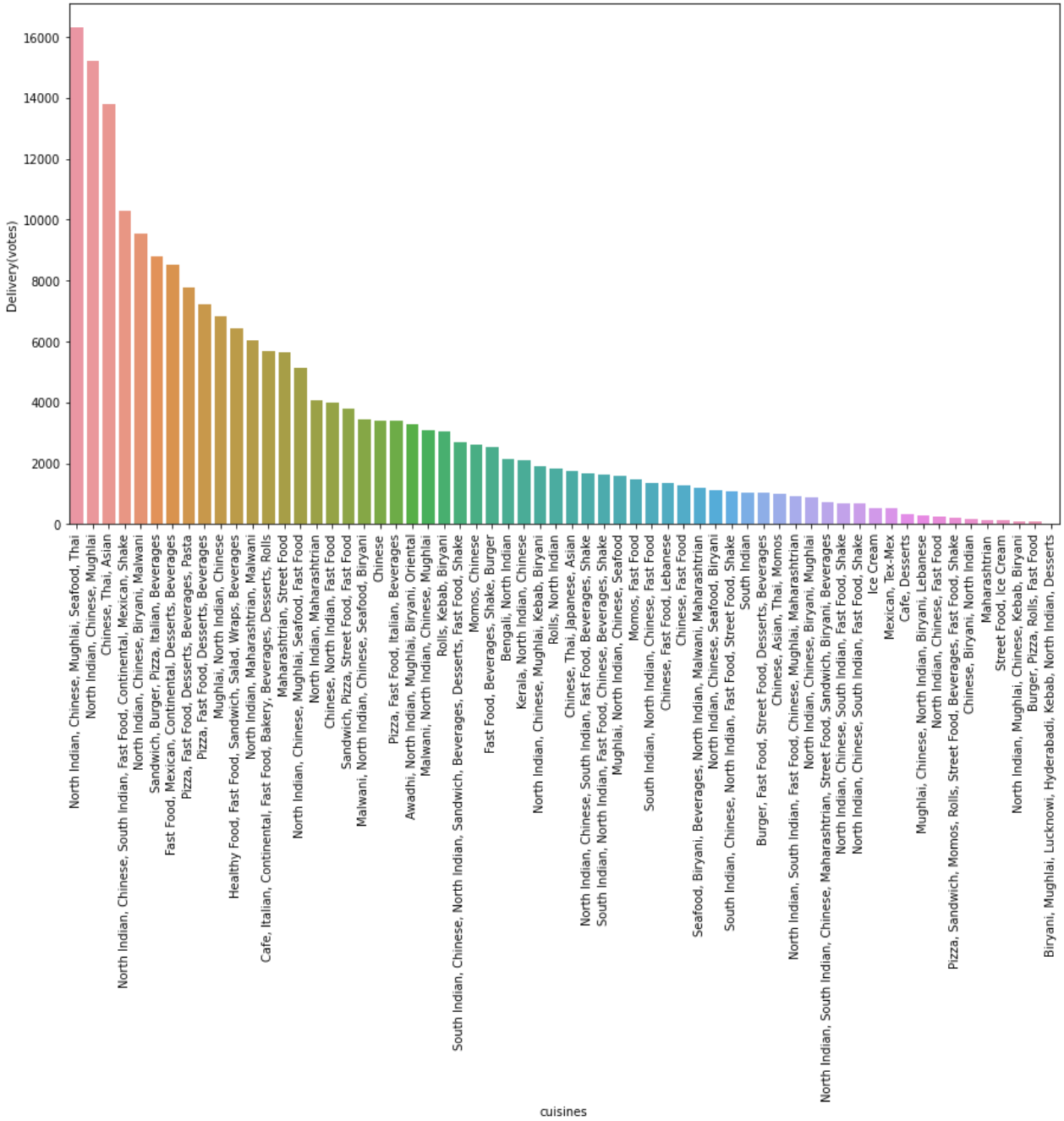
```
df.rename(columns = {'listed_in(type)':'Type','listed_in(city)':'City'}, inplace = True)
df.head()
```

```
df['location'].unique()
```

```
df.drop(['City'], axis = 1)
```

```
plt.figure(figsize = (6,6))
sns.boxplot(x = 'online_order', y = 'Delivery(rate)', data = df)

plt.figure(figsize = (6,6))
sns.boxplot(x = 'book_table', y = 'Dining(rate)', data = df)

plt.figure(figsize = (14, 8))
sns.boxplot(x = 'Type', y = 'Delivery(rate)', data = df, palette = 'inferno')

df6 = df[['cuisines', 'Delivery(votes)']]
df6.drop_duplicates()
df7 = df6.groupby(['cuisines'])['Delivery(votes)'].sum()
df7 = df7.to_frame()
df7 = df7.sort_values('Delivery(votes)', ascending=False)
df7.head()

df7 = df7.iloc[1:, :]
df7.head()

plt.figure(figsize = (15,8))
sns.barplot(df7.index , df7['Delivery(votes)'])
plt.xticks(rotation = 90)
```

FIGURE A : - Bivariate Analysis



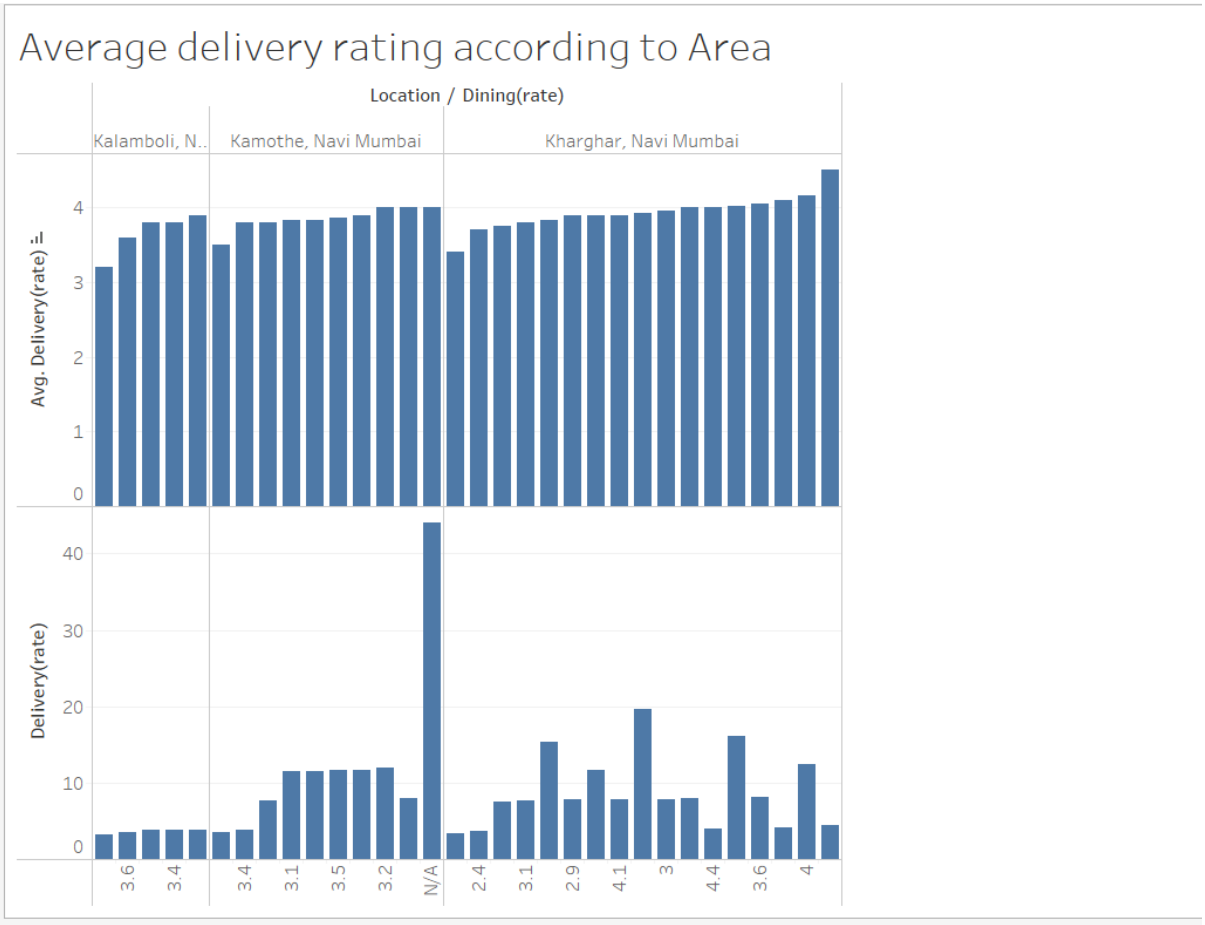
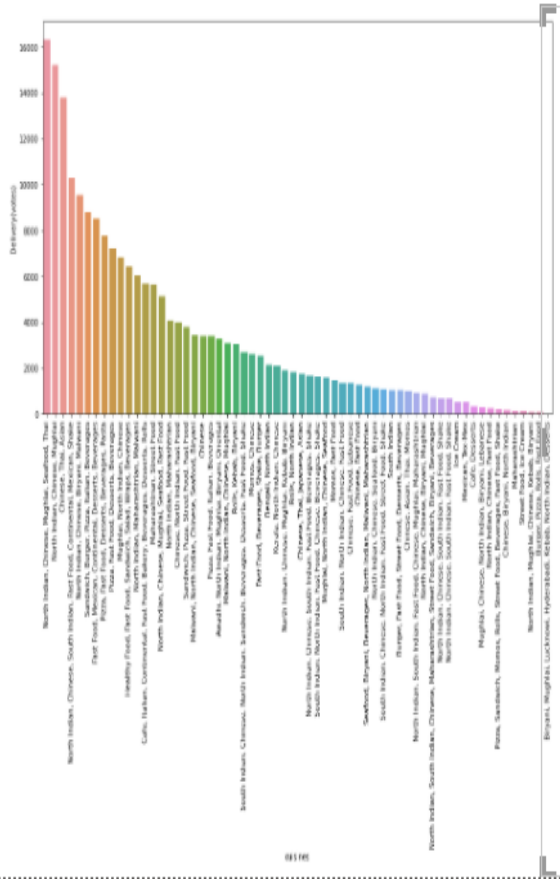
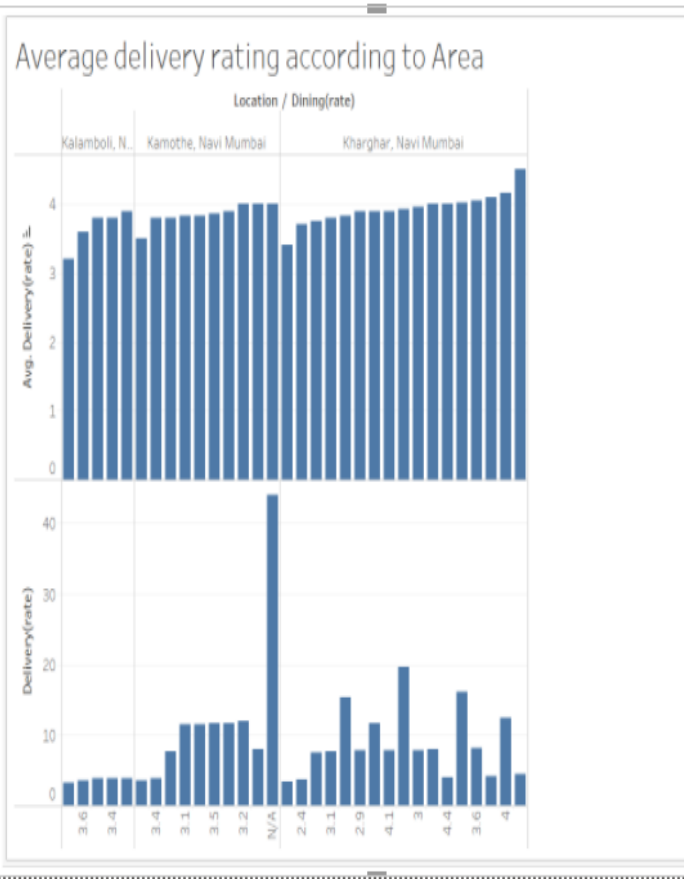


FIGURE B : - Multivariate Analysis

Dashboard : -




Average delivery rating according to Area
Location / Dining(rate)
Kalamboli, N..
Kamothe, Navi Mumbai
Kharghar, Navi Mumbai
Avg. Delivery(rate)
Delivery(rate)
3.6
3.4
3.4
3.1
3.5
3.2
N/A
2.4
3.1
2.9
4.1
3
4.4
3.6
4

5. CONCLUSION

5.1 Conclusion : -

First , we have defined the questions as problem definition given according to Business POV.

Then we have Collected the data from Zomato platform cleaned it and explained in Data analysis(plots,bar,graphs) and prepared a end goal - Dashboard which answers the Business POV.

In the end we have reach to a conclusion that the Dashboard defines Bi-variate and Multivariate Analysis which answers us as follows :-

A. Figure A defines what food people are preferring.
B. Figure B defines in which location there are a large no of quality assured restaurants according to rating.

This is also known as Descriptive Analysis. It is the initial step of Data Analytics term in Business Intelligence/Analytics which defines the question (what is currently happening).

Data Science/Analytics in practice :-

Descriptive Analysis → Diagnostic Analysis → Predictive Analysis → Prescriptive Analysis

5.2 Future Scope of the Project : -

It will be helpful for people to open restaurants according to their location .

It will be also helpful for opening franchise for QSRs according to their location .

From consumer POV it will be also helpful for knowing the type of restaurants people prefer according to Review , also type of Cuisines and types of Dishes .