iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XII **Month of publication:** December 2021

DOI: https://doi.org/10.22214/ijraset.2021.39303

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

Zomato Data Analysis

Arpit Saxena[1], Prof. Nidhi Sengar[2], Prof. Amita Goel[3], Prof. Vasudha Bahl[4]
[1]*Bachelor in Technology, IT Department, Maharaja Agrasen institute Of Technology*
[2]*Mentor*
[3, 4]*Panelists*

Abstract: ***Whenever we would like to visit a brand new place in delhi -NCR, we often search for the most effective restaurant or the most cost effective restaurant, but of decent quality. For looking of our greatest restaurants we frequently goes for various websites and apps to induce an overall idea of restaurants service. the foremost important criteria for all this is often rating and reviews of the those that have already got experience in these restaurants. People see for rating and compare these restaurants with one another and choose for his or her best. We restrict our data only to Delhi-NCR. This Zomato dataset provides us with enough information in order that one can decide which restaurants is suitable at which place and what kind of food they must serve so as get maximum profit. it's 9552 rows and 22 columns during this dataset. We'd wish to find the most affordable restaurant in Delhi-NCR.We can discuss various relationships between various columns of information sets like between rating and cuisine type , locality and cuisine etc. Since it's a true time data we might start first with data cleaning like cleaning spaces , garbage texts etc , then data exploratory like handling the None values, null values, dropping duplicates and other Transformations then randomization of dataset so analysis. Our target variable is that the "Aggregate Rating" column. We explore the link of the opposite features within the dataset with relevancy Rates. we'll the visualize the relation of all the opposite depend features with relevance our target variable, and hence find the foremost correlated features which effects our target variable.***

Keywords: Online food delivery, Marketing mix strategies, Competitive analysis, Pre-processing, Data Cleaning, Data Mining, Exploratory data analysis , Classification , Pandas , MatPlotLib.

I. INTRODUCTION

Digitalization has impacted the whole world and India is also remain affected by this phenomenon. Various things from classrooms to eating food gone to the internet.Customers not only use the Internet to buy product online, but also to compare costs, features and quality of the product, and any sale available[1].They would get the idea if they want to buy the product from a specific store . The Internet is becoming an pervasively common platform to facilitate searching, choosing, and buying products. Online food ordering companies offer a range of options and conveniences that enable consumers to have their favourite food on their fingertips[2]. Recently, Online Food Delivery become a new trend comforting the foodies and Zomato is the biggest name that come to mind when talking with reference to India. Zomato helps various restaurants to increase their customer base and even the concept of cloud kitchen also finds its way in India having only delivery but not dine in facilities. Many people across the country want to get into this profitable business of food delivery and wants to open restaurants and cloud kitchen in different parts of India.

The Objective of this project is to get an idea of following :

1) What type of food is like by people in various places.
2) Which restaurant they like most.
3) Which type of restraint is profitable to open in which area

II. OBJECTIVES

The study has the following objectives:

1) To identify the various strategies adopted by various restaurants.
2) To analyze how Zomato is different from its competitors.
3) To analyze the perfect location and perfect cuisines for restaurants to open at certain places.
4) To analyze which food is famous in which places
5) To analyze relationships between restaurants, food and ratings

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

III. RESEARCH METHODOLOGY

Detailed analysis of this data set composed of data cleaning , data pre-processing so that we get a proper data set to work upon. After this we are checking for various relationships between various columns so that we get to know the answers of various questions like top 10 restaurants with highest rating for particular food etc.Then after getting various results we analyze those results and get to know how one column become an affecting factor for other.

IV. COMPANIES AND FIRMS

One of the biggest advancement in the e-commerce industry worldwide is how things become online at your fingertips that are once have far reach. Food sector is also one of that advancements that makes your favourite food accessible to you from anywhere and at anytime. Various companies deals with online food delivery system some of them are given below.

Table-1: Exhibits the list of online food delivery companies

Sl. No	List of Companies	Country of origin	Year of establishment
1	Domino's Pizza	Ypsilanti, Michigan, U.S.	1961
2	Just Eat	London, England, UK	2001
3	Grubhub	Chicago, Illinois, U.S.	2004
4	Zomato	Haryana, India	2008
5	Postmates	San Francisco, California, US	2011
6	FoodPanda	Berlin, Germany	2012
7	Deliveroo	London, England, UK	2013
8	DoorDash	San Francisco, California, US	2013
9	Swiggy	Bangalore, India	2014
10	UberEats	San Francisco, California, U.S.	2014

V. DATASET DESCRIPTION

Zomato dataset is real time data set which gives information about restraunts , its cuisins , locality , ratings etc.
The data is taken from url: https://drive.google.com/file/d/1FSa_x3COvCoMODa44qXufO9CQb3ydqKw/view

The dataset contains the following features

1) *Restaurant Id:* This feature contains the Id of the restaurant on the Zomato website
2) *Address:* This feature contains the address of the restaurant in Delhi-NCR
3) *Restraunt Name:* The name of the restaurant
4) *Has online_order:* Whether online ordering is available in the restaurant or not
5) *Has Table Booking:* Table book option available or not
6) *Aggregate Rating:* Contains the overall rating of the restaurant out of
7) *Votes:* Contains total number of upvotes for the restaurant
8) *Locality:* Contains the neighbourhood in which the restaurant is located
9) *Cuisines:* Type of meal
10) *City:* Contains the neighbourhood in which the restaurant is located.
11) *Locality Verbose:* Exact place in that locality.
12) *Longitude:* Longitude of restaurant
13) *Latitude:* Latitude of restaurant
14) *Is Delivering Now:* Food delivered currently or not
15) *Switch to Order Menu:* Order menu available
16) *Price Range:* Price ranges available
17) *Average Cost for Two:* Average cost when two people eating
18) *Currency:* Currency of that country
19) *Has Online Delivery:* Online delivery option available
20) *Rating Color:* Color of rating
21) *Rating Text:* Rating numerical value
22) *Country:* Country of restaurants

VI. METHODS

A. Data Collection

Data that we got from Url above is a platform used for getting various results that are further analyse to get proper relationships among various factors.There are total of 21000 data points approx.. and calculation is done on this.

B. Data Pre-Processing

The Dataset contained following Attributes.-

1) Records having null values were dropped from ratings columns and were replaced in the other columns with a numerical value.
2) Various spaces at different places were deleted and data set get cleaned

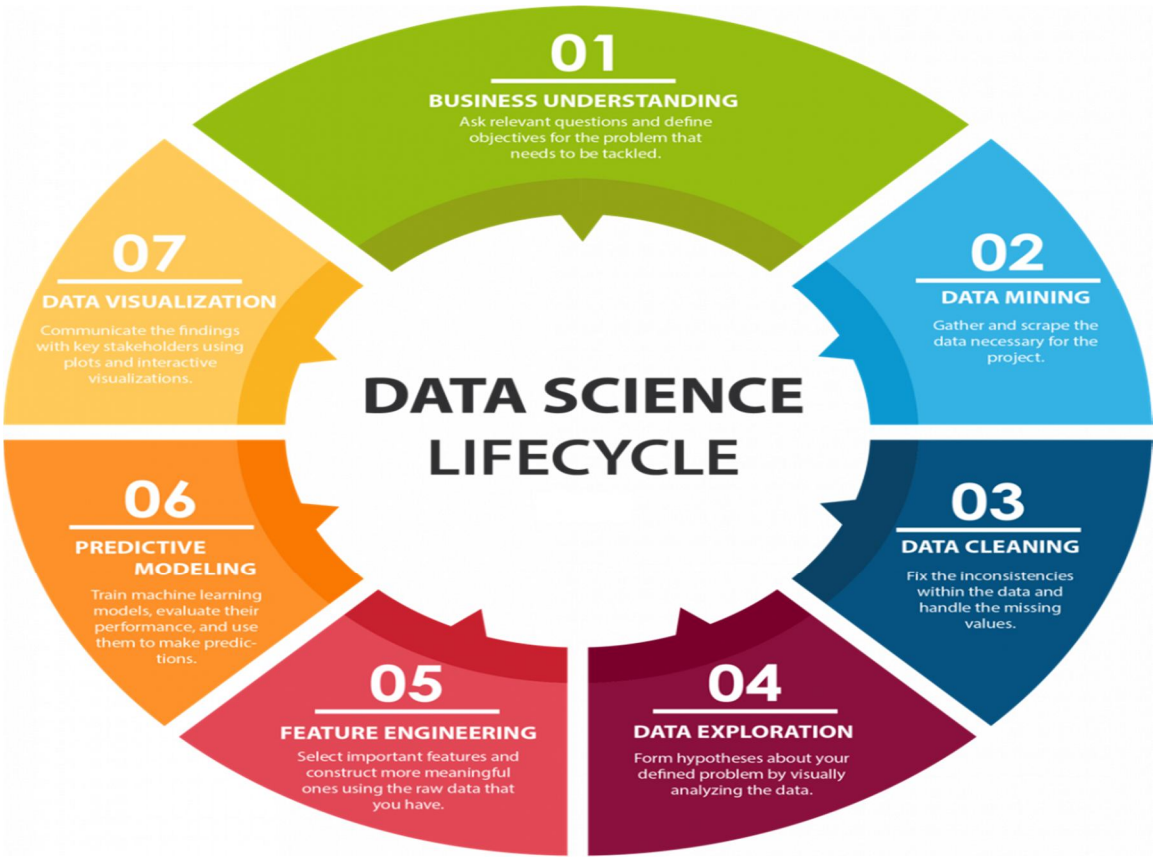
C. Exploratory Data Analysis

High amount of effort went into the EDA because it gives us an in depth knowledge of our data and its related information.
Exploratory Data Analysis (EDA) could be a technique/method for data analysis that uses a range of techniques (mostly graphical) to

1) Maximize insight into an information set;
2) Uncover underlying structure;
3) Extract important variables;
4) Detect outliers and anomalies;
5) Test underlying assumptions;
6) Develop parsimonious models;
7) Determine optimal factor settings [3]

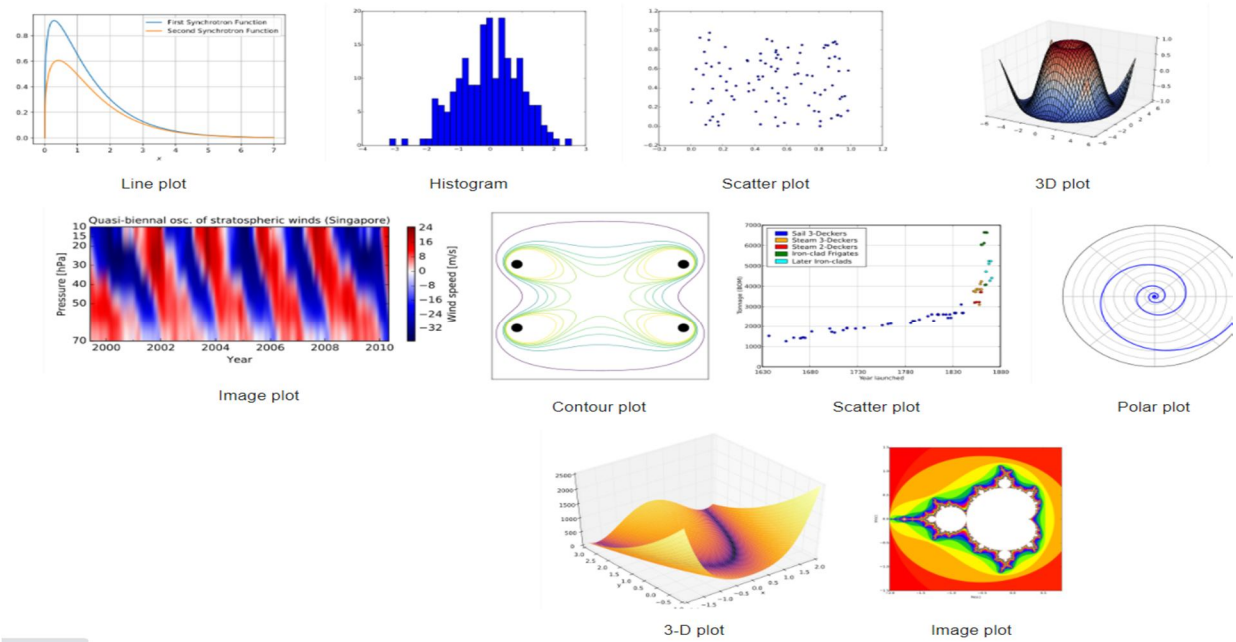
D. Randomization and Splitting Of Dataset

The features selected using above steps were used to develop classification models. Initially the dataset has to make random so to make it splitted. It was then followed by splitting of the dataset into two parts : training (70% of the dataset) and test (30%) sets.



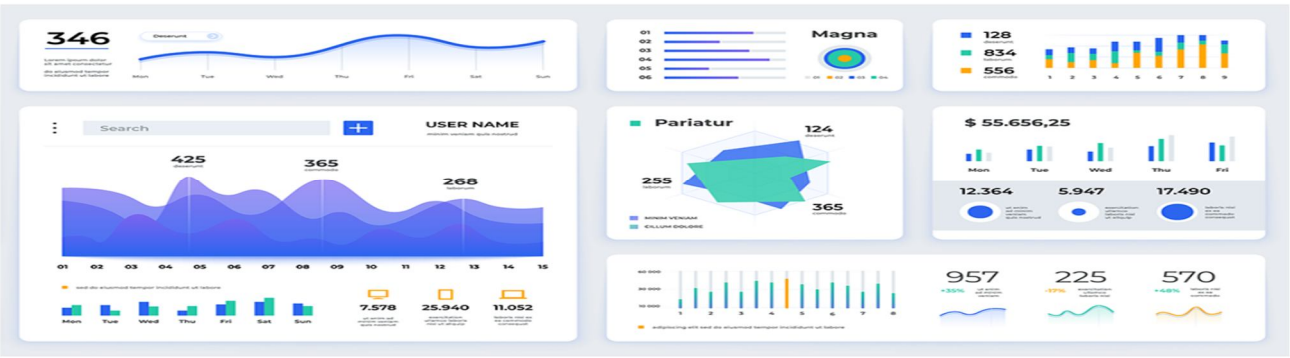
VII. TECHNOLOGIES USED

1) *Python:* Python is a high-level general-purpose programming language that works with interpreter. It emphasises more on code readability with core indentation. It is an object-oriented language which helps programmers to write clear, logical code for small and large-scale projects[4]. Python have feature of dynamically-typing and garbage-collection.It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming[5]. It is often called as a "batteries included" language due to its various comprehensive standard library. Guido van Rossum started working on Python in the late 1980s, as a successor to the ABC programming language, and it got first released in 1991 as Python 0.9.0.[6] Python 2.0 was released in 2000 and introduced some new features, such as list comprehensions and a cycle-detecting garbage collection system (in addition to reference counting system). Python 3.0 was released in 2008. Python 2 got discontinued with version 2.7.18 in 2020.Python consistently ranked among one of the most popular programming languages.
2) *Pandas:* Pandas could be a library utilized in python for data analysis. Started by Wes McKinney in 2008 thanks to the necessity for a robust and versatile measuring tool, pandas has grown itself into one in every of the foremost popular Python libraries that exists. it's a good community of contributors. Pandas is constructed on above of two core Python libraries—matplotlib for data visualization and NumPy for mathematical operations. Pandas library becomes a container for these two libraries, allowing you to access many of matplotlib's and NumPy's features and methods with considerably less code. as an example, pandas' .plot() method combines multiple matplotlib methods into one method, enabling you to plot beautiful chart in an exceedingly few lines. Before pandas, most of the analysts used Python for data munging and preparation of dataset, and so switched to a more domain specific language like R programming for the remainder of their workflow. With Pandas introduced two new sorts of objects for storing data that make analytical tasks much easier and eliminate the necessity to change between tools: Series, which contains a list-like structure, and DataFrames, which contains a tabular structure.[7]
3) *Matplotlib:* Matplotlib , a library that is arguably the top most used Python library for 2D-graphics. It provides us with both a fast way to visualize data in Python and publication-quality of figures in various formats. Matplotlib was originally written and developed by John D. Hunter. After that, it got huge development community and is distributed under a BSD-style license. Michael Droettboom was nominated as matplotlib's lead developer shortly before John Hunter's death in August 2012 and was further joined by Thomas Caswell.. Matplotlib is basically a NumFOCUS financially sponsored project. Matplotlib 2.0.x supports Python versions 2.7 through 3.10. Python 3 support started with Matplotlib 1.2. Matplotlib 1.4 is the last version to support Python 2.6.[10] Matplotlib has pledged not to support Python 2 past 2020 by signing the Python 3 Statement. Pyplot may be a Matplotlib module which provides a MATLAB-like interface. Matplotlib is meant to be as useful as MATLAB, with the power to use Python language in it, and therefore the advantage of being free and open-source software that supportd by large community. [11]



4) *Data Visualization:* Data visualization is the technique of converting large data sets and numbers into charts, graphs and other visuals that is made to represent data pictorially[12]. This visual representation of data makes it easier to identify and share real-time comparisons, trends, and new insights about the information represented in the data. It helps you to keep an eye on different events or activities in a single look by providing insights on one or more pages or screens.Various Data science techniques can be used to identify what affecting what, why it's affecting, and what will happen next. As the size of database increases, more people required data visualization tools to process their data [13]

5)



VIII.RESULTS

1) The dataset is very skewed towards Delhi ncr restaurants
2) BBQ, German, Malwani,, Cajun - these cusines are not available in ncr regions.
3) North Indian 3597, Chinese 2448, Fast Food 1866 are the most served cuisines in both nor and non nor regions
4) when we print data we see NON NCR values and NON NCR EX values are same except for 3 cases. I am just printing top 10 because after that data is very small which means chances of error are high. The first unexpected value in column 5 is "Continental" which means in NON NCR region local food is famous. Second unexpected value is in column 6 which is "Italian" which I believe can be justified by people likeness to "Pizza and Pasta". Third unexpected value is in cloumn 8 "Cafe" which shows increase in coffee culture in other parts of India.
5) When we print data we see most of cuisines are from another country. As data is very low for this segment this might not represent the actual facts. In this data we see most of the exotic food from another countries are in NCR region while NON NCR region is 0, for 7 out of 10 cases.
6) The restaurants serving 6 cuisines have highest rating. The rating first increased form 1-6 cuisines and then decreased from 6-8.
7) The graph tells us all. After the average_cost_for_two hits a target of RS 7000 the average rating is usually above 4.0 which is a good rating. We can say the more premium a restaurant gets the more rating it has.
8) The analysis is simple. The more exotic the cuisine gets the more rating it has. One more thing we can see if we print avg cost each cuisine we will see avg cost doesn't need to more for good rating. All costs are in Indian RS.
9) Restaurants chins like BBQ, AB - absolute barbeque, have highest number of votes.
10) Cities like Delhi, Gurgaon and Noida have more number of restaurants and more weighted rating.

IX. CONCLUSION

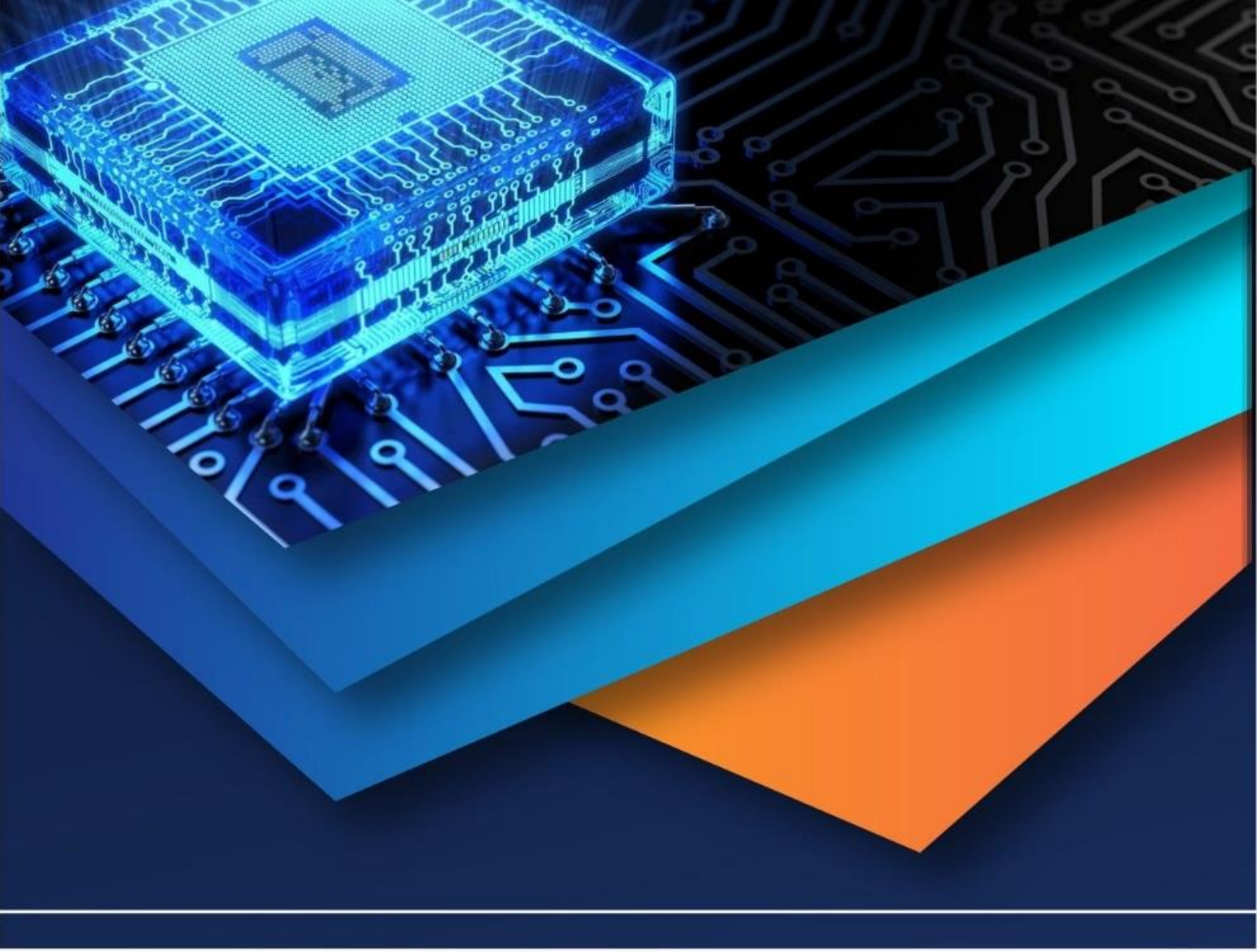
This paper have the analyses of various characteristics of current restaurants in different localities of a city in particular country and analyzes them to predict restaurant ratings related to particular food. This makes it an important thing to take into consideration before making a dining in or online ordering decision. Before creating a venture like that of a restaurant, such kind of research is an important part of planning and this paper has already done it for atleast delhi-NCR people. There has been a lot of research into variables impacting profits and the competition in the restaurant industry when someone opening its restaurants. To enhance customer satisfaction rates, various dine-scape variables have been analyzed. If data is also collected for other reviwers, such predictions could be made for accuracy.

X. ACKNOWLEDGMENT

I completed this project and Research paper under the guidance of my mentor Prof. Nidhi Sengar who constantly supports me in every sphere. I would also like to thanks my friends, family and my teachers who helped me in completing this project.

REFERENCES

[1] Das, J. (2018). Consumer Perception Towards “Online Food Ordering and Delivery Services”:
[2] Kapoor, A. P., & Vij, M. (2018). Technology at the dinner table: Ordering food online through mobile apps. Journal of Retailing and Consumer Services, 43(1), 342–351
[3] Atharva Kulkarni,Divya Bhandari,Sachin Bhoite.A study of Restaurants Rating Prediction using Machine
[4] Kuhlman, Dave. "A Python Book: Beginning Python, Advanced Python, and Python Exercises". Section 1.1. Archived from the original (PDF) on 23 June 2012.
[5] Python Software Foundation. Archived from the original on 20 April 2012. Retrieved 24 April 2012., second section "Fans of Python use the phrase "batteries included" to describe the standard library, which covers everything from asynchronous processing to zip files."
[6] Rossum, Guido Van (20 January 2009). "The History of Python: A Brief Timeline of Python". The History of Python. Archived from the original on 5 June 2020. Retrieved 5 March 2021.
[7] "License – Package overview – pandas 1.0.0 documentation". pandas. 28 January 2020. Retrieved 30 January 2020.
[8] matplotlib.org.
[9] "NumFOCUS Sponsored Projects". NumFOCUS. Retrieved 2021-10-25.
[10] "Installing – Matplotlib 2.0.2 documentation". Retrieved 2017-06-23.
[11] "Matplotlib: Python plotting — Matplotlib 3.2.0 documentation". matplotlib.org. Retrieved 2020-03-14.
[12] Nussbaumer Knaflic, Cole (2 November 2015). Storytelling with Data: A Data Visualization Guide for Business Professionals. ISBN 978-1-119-00225-3.
[13] Gershon, Nahum; Page, Ward (1 August 2001). "What storytelling can do for information visualization". Communications of the ACM. 44 (8): 31–37. doi:10.1145/381641.381653. S2CID 7666107.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY