‘Location and customer satisfaction analysis on restaurant Industry’

Student Name: Pratik Prakash Brahmapurkar

Student Number: 40331504

Word Count: 10,021

Submitted in part fulfilment of the degree of

Master of Science in Business Analytics

2022

Queen’s Management School

CANDIDATE DECLARATION

Declaration

This is to certify that:

i. The dissertation comprises only my original work;

ii. Due acknowledgement has been made in the text to all other materials used;

iii. No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

---------- PRATIK BRAHMAPURKAR

[Candidate's Signature] **Printed Name**

Date: 9/9/2022

ABSTRACT

Many people in the city of Bangalore are interested in opening a restaurant, but they lack the necessary data to do so. This study assists in investigating the importance of location in the establishment of a restaurant and the factors that contribute to a high level of customer satisfaction. One of the most appealing aspects of the restaurant business is its location, and as such, selecting the right site for a restaurant may be rewarding for both the restaurant owner and its customers. A restaurant's owner or a would-be restaurateur could use this information to better market their establishment and boost customer satisfaction. The goal of this report's findings is to give restaurant owners the knowledge they need to shape their decision-making and make sure key aspects are met.

As a result, this report focuses on a location analysis of the ratings for Zomato restaurants in the city of Bangalore, as well as figuring out what the most important factors are that affect the ratings. In Bangalore, a set of characteristics are used to evaluate a set of machine learning models as well as Centrality Network Analysis. The data for this report was downloaded from kaggle.com. A restaurant's customer satisfaction rate was being calculated from its overall rating. The customer satisfaction rate was found to be higher for restaurants that were closer to the city centre than those that were further away.

The research shows that decreasing distance from the center of the city increases customer satisfaction. In light of this, it is suggested that distance be taken into account as a primary factor. Besides satisfaction rate and rating, other important criteria are the restaurant's types and the price for two people at that restaurant, votes, restaurant type, online order and booking of table. There will be a comprehensive conclusion, including limitation, and areas for future study.

Table of Contents

1. INTRODUCTION

1.1. Overview

The restaurant industry is worth multiple billions of dollars and is expanding at a rapid rate every day. In 2014, the food service industry's market value in India was 38 billion USD, while projections for 2025 indicate that it will be increased to 110 billion USD (Statista, 2022) In 10 years, the restaurant is about to grow thrice. The introduction of a brand-new restaurant into a neighbourhood can result in a multitude of beneficial outcomes. A new restaurant will benefit a town economically by creating several jobs (McDonough, 2022), this would support the local food producers, and it will aid local food producers by purchasing their products (Harris, 2017) Over the past seven years, restaurant employment has expanded at a rate that is consistently higher than that of the entire economy of the USA (The Atlantic, 2017).

The ability to eat is a basic requirement for human in day to day life, but the ability to eat wisely is a skill. In order to be competitive and successful, restaurants of all types need to develop strategies for retaining and acquiring customers. It should come as no surprise that today's savvy diners are looking for restaurants that provide not just a delicious meal at a fair price, but also a pleasant, memorable experience due to their appealing ambiance and friendly staff (Canny, 2014). The modern customer thinks that going to a restaurant is about a lot more than just eating. They know what's going on and make careful decisions based on what they know. Consumers often try to find new product-related information from a variety of sources to make sure their purchases go well. Before the Internet became so popular, people used traditional media sources to find out about any product on the market. Since the Internet is becoming more popular, restaurant owners are looking for new ways to draw customers and make them more likely to buy, no matter

what online platform they use like Zomato, UberEats or Swiggy (Vajjhala & Ghosh, 2021).

Zomato is an online platform that connects customers, restaurant partners, and delivery partners to meet the various requirements of each group (Zomato, 2022) The Indian restaurant industry would not be complete without Zomato. The restaurants can use the site's ratings and reviews to improve their visibility and quality of service. Zomato's robust local advertising platform is being used by restaurants to reach more potential customers. In addition, the development of tools like Online Ordering and Online Booking has helped restaurants increase their delivery revenue and streamline their table reservation operations (Raman, 2018).

India has always been known as a food-loving country, and each state has its specialities. Indians, on the other hand, aren't known for going out to eat. Instead, they spend more time cooking food at home (Biswas & Verma, 2022; Mondurailingam & Subramani, 2015) The practice and tradition of only eating home-cooked food are changing quickly, and the number of restaurants has gone up steadily and sharply by about 20% over the last 30 years (Biswas & Verma, 2022).

The way of life of Indian consumers has been a significant contributor to the maturation and expansion of the fast food sector over the past few years. Other factors, such as an increase in the number of single-parent households, greater familiarity with western cuisine and international media, and a rise in the proportion of employed women, have also played a significant role in the development of eating trends and the expansion of the restaurant industry. The professionals in this sector believe that the young population of the middle class, which has a significant amount of disposable income, would spend more dining out at restaurants serving fast food. Additionally, it is anticipated that the demand for takeaway would record a significant rise in the country (Mondurailingam & Subramani, 2015).

1.2. Purpose and Scope

For this report, the city of Bangalore has been chosen as the place to look into the Zomato Restaurant's Dataset. Bangalore has around 12.7 million people and is the fourth most populous city in India (Statistic Times, 2021). It is also known as the Silicon Valley of India. India's Information Technology (IT) is primarily focused in Bangalore. According to (Financial Times, 2022) twenty of the list's top 500 companies are based in Bangalore, which is 4% of the total. In most single-parent families, both parents work. The husband's and wife's hectic schedules contribute to the rising popularity of ordering food online (Shetty & S, 2020). Because of this, there is a great need for restaurants. Seeing the situation, opening a restaurant is a good business idea, but choosing the right spot is one of the most important factors in its success. Other than location, customer satisfaction also plays an important role in the long term success of any restaurant.

Since some restaurant owners don't realise enough about location and can't think of it in the best way, this study uses Machine Learning (ML) models and Centrality Network Analysis to look at the location and try to find out which factors are best for a higher satisfaction rate in Bangalore. First, there will be a review of the literature on Customer Satisfaction and Loyalty, Restaurant Location Analysis, and Ambience. Then, the satisfaction rate will be analysed, and parameters for things like distance from the city centre and high activity places will be made. The models that are made are then looked at in terms of the restaurant business. Then, attempts will be made to bridge the gap between the literature. Subsequently, a critical discussion of the findings will be done, and a conclusion will be written that talks about the report's theoretical and practical implications.

2. LITERAURE REVIEW

2.1. Customer Satisfaction and Loyalty

Every business needs satisfied customers to stay in business. Customer satisfaction is seen as a factor in how a customer feels after making a purchase. This attitude can be either positive or negative depending on the customer's personal experiences (Canny, 2014). Customer satisfaction is a measure of how happy customers are with the products and services they buy and use. It is driven by two things: prospects and actual service performance. Customer satisfaction can be defined as the feeling of liking or disliking the outcomes with expectations (Othman & Harun, 2021). Customer satisfaction is seen as a way to get people to use the service again, but it's not a guarantee that a happy customer will come back to buy again. For restaurant owners to be able to influence their customers' restaurant choices, they need to know how customers make decisions about which restaurants to choose. (Haghighi, et al., 2012)

The research conducted by (Andaleeb & Conway, 2006) it suggests that owners and managers of full-service restaurants should focus on three major factors: service quality (responsiveness), price, and food quality (or reliability) if they want to have better customer satisfaction. And according to (Chun & Nyam-Ochir, 2020), the quality of the food, the quality of the service, the price, and the atmosphere of a restaurant were the four most important things that made customers happy. The results show that customer satisfaction will have a positive effect on both customers' plans to come back and the likelihood of recommending the restaurant to others.

There are three main sources of customer satisfaction with restaurant services: positive emotions, negative emotions, and perceived service quality. The effect of perceived service quality on satisfaction is mediated by both positive and negative feelings (Ladhari, et al., 2008). But according to (Liljander & Strandvik, 1997) the negative

emotions affect satisfaction more than positive ones. As long as the customer doesn't have too many good or bad feelings, these feelings don't seem to affect how happy they are with the service. Satisfaction is not explained by a strong feeling of happiness. But it's important to pay attention to situations where the customer has strong negative feelings. If service providers don't deal with this kind of behavior in the right way, it can have bad effects. When customers are very unhappy, they may try to spread negative word of mouth as a means of getting back (Andaleeb & Conway, 2006). According to (Biswas & Verma, 2022) Customer satisfaction is not something that comes with the product response that is set up by society. The more impactful service providers are, the happier ones customers would be. When potential customers see positive reviews, they analyze the votes and they will be more inclined to make a choice restaurants favour. Alternately, when customers come across bad reviews that have a growing number of helpful votes, they look for other available restaurant options (Lee, et al., 2020).

According to (Oracle, 2022) customer loyalty is a long-term emotional relationship between your company and your customer. It shows up in how willing a customer is to interact with you and buy from you again and again instead of your competitors. Loyalty comes about when a customer has a good experience with you and helps build trust. And as per (Othman & Harun, 2021), Customer loyalty is when a customer comes back to a business or buys from it more than once. It also includes an emotional commitment or a positive attitude toward the business.

Many researchers have found positive relationship between customer loyalty and customer satisfaction (Andaleeb & Conway, 2006; Fornell, et al., 1996; Biswas & Verma, 2022; Othman & Harun, 2021). This link between customer satisfaction and customer loyalty means that in the hotel and restaurant industries, customers have to be happy in order to be loyal. The satisfaction of the service should be reflected in the

loyalty of the service, and that all elements of satisfaction have the most impact on loyalty responses (Othman & Harun, 2021).

Offers provided by restaurants were not discussed in any of the literature, offers like buy 1 get 1 free or some discount provided for a short term.

A more in-depth investigation of the question "does a restaurant present any offer?" was ignored. And based on this attribute a proper satisfaction rate could have been determined.

2.2. Restaurant Location Analysis

Almost all the researchers have always claimed that a restaurant's location is one of the most important factor for restaurant's long-term success (Yang, et al., 2017; Kim, et al., 2022; Bhatia & Sneha, 2021; Hanaysha, 2016; Chen & Tsai, 2016). According to (Love, 1972) Location is the most important thing for a fast food business to do well. Without a good location, good management and good products don't matter. As markets get more crowded, location becomes even more important. But according to (Haghighi, et al., 2012) location of a restaurant is not a significant influence in customer happiness.

In 1933, Walter Christaller established Central Place Theory (CPT) as a mechanism to explain the location, number, and size of communities, in which these locations served as service hubs for surrounding areas (Altawee, 2021). It says that retail location is determined by the range and threshold of a product (Chen & Tsai, 2016). It says that a store's location is based on its "range," which is the farthest distance a customer is willing to go to get a product. This distance determines the store's market area's outer limit. And "minimum amount of demand," which is the minimum amount of demand that must exist in an area for a store to be financially viable. So, from a spatial point of view, a store will

be in an area with enough people to support it and within a reasonable distance for customers to travel (Prayag, et al., 2010; Chen & Tsai, 2016). Spatial interaction theory, on the other hand, says that customers might make trade-offs between store-specific differences in products and services and the location's attractiveness. This theory says that customers go to casual restaurants not only because of their location, but also because of price, cleanliness, service, type of cuisine, and food quality (Prayag, et al., 2010). Central place theory and spatial interaction theory both explain why restaurants are placed where they are in a city. As per the research done on fast food restaurants in the center of the Jakarta city by (Widaningrum, et al., 2018) suggests that fast food restaurants are located close to their competitors.

According to (Donga, et al., 2019) different kinds of restaurant businesses have very differing views about where they want to be. High-budget restaurants are more likely to benefit from agglomeration and be clustered near the city centre, whereas Low-budget restaurants, on the other hand, are more sensitive to the price of land, so they are spread out. (Chen & Tsai, 2016) Developed a Rough Set Theory (RST) based data mining system for extracting potentially relevant rules from geographical data. Using a case study of a restaurant chain, the authors displayed and proved the efficacy of their proposed strategy. Store performance in relation to locational factors was predicted using Rough Set Theory. When scouting locations for a new restaurant, it was concluded that management should pay closer attention to issues such as store size, parking availability, store exposure, and population growth rate in the neighborhood.

According to (Bhatia & Sneha, 2021) More than 50% of people would rather have their food delivered to their home, office, or business, while about 35% would rather eat out. After looking at the data, it was found that people in Bangalore prefer a delivery and dining out over other options at restaurants. In 28 of the 29 places that were looked at for the Zomato dataset, this pattern was found to be true. So, it could be said that a mix of

delivery and eating is followed and location might not be completely an important factor. One of the most important aspects of marketing management services is the location of service provision that makes the service possible. It helps to expedite and improve the key services in addition to simplifying them (Othman & Harun, 2021).

Researchers often neglected to consider how far away from the city centre the restaurants were located. However, (Sedov, 2022) analyzed the distance variable but didn't take into account other variables like restaurant types or prices.

2.3. Ambience

According to (Kotler, 1974) Ambience is the purposeful design of a space to make buyers feel certain emotions that make them more likely to buy. The atmosphere can affect how people buy things in at least three different ways. First, it can be used to get people's attention. Second, it might use a medium that sends a message by elaborating actual potential in a restaurant. And third, the atmosphere may be a way to make an effect, it can create a sensation in buyers. The quality of the surrounding environment as experienced by customers is referred to as an ambience (Omar, et al., 2015).

It's possible that individuals won't visit a restaurant for the first time because of the cuisine alone, but if the atmosphere, the reviews, or the recommendations of their friends and family are very positive, they will be more likely to go there. The customer's opinion of their eating experience can be influenced in some way by the restaurant's design, atmosphere, and level of service (Omar, et al., 2015; Wade, 2006). Consumers' emotional responses may be encouraged to be stimulated by a nice restaurant atmosphere provided by the mix of lighting, music, and color, which can serve to encourage customers to make impulsive purchases. On the other hand, the utilitarian aspects of a restaurant's

atmosphere can assist consumers in forming their cognitive appreciation, which in turn assisted them in forming logical purchasing decisions (J.DONOVAN, et al., 1994). Lighting has the potential to be one of the most influential physical stimuli in restaurants, especially more upmarket establishments. In fast-food restaurants, the use of bright lighting may signify quick service and relatively low costs, whereas the use of quiet and warm lighting may symbolically represent full service and premium prices (Omar, et al., 2015). Warm light appears to make individuals more comfortable, which leads to their remaining longer in the place, which leads to increased food intake. On the other hand, the bright light causes people to spend less time in the location where they are eating, which leads to decreased food intake (Stroebele & Castro, 2004).

According to individual interests and tastes, different kinds of travellers gravitate toward different kinds of cuisine. For instance, a luxury traveller would like a proper restaurant with a fantastic sensory eating experience (such as exquisite waiting staff, fine silverware, the perfume of excellent wine, and extraordinary food), which might offer passengers a sense of significance. Owners of businesses and those in charge of marketing should take the seemingly contradictory opinions of each traveller into consideration (Hlee, et al., 2019). As per (Kothari & Shah, 2018; Gupta, et al., 2021) restaurant reviewers are always on the lookout for establishments that offer superior quality in terms of cuisine, ambience, and service.

According to the findings of (Cheng, et al., 2016), the only way for a customer's surroundings to influence their intention to make a purchase is for it to generate best to promote values and a positive picture of the restaurant. It was observed by (Priya, 2020) that as the average price of restaurant increases the rating also increases. In premium restaurants, (Ryu & Jang, 2007) investigated the interplay between the effects of a number of environmental factors on diners' intents and behaviors. Their findings confirmed the idea that atmosphere (including music, fragrance, and temperature), as well

as personnel appearance, had the most significant impact on the emotional responses of customers, which in turn altered customers' post-dining behavioural intentions. But as per (Voon, 2017) if the level of service that is offered by the personnel of the restaurant (human service) is poor, then a renovation in the restaurant's physical environment will not necessarily boost the level of happiness and loyalty among young people.

3. METHODOLOGY

The CRISP-DM, which stands for the Cross-Industry Standard Process for Data Mining, is a strategy that is commonly utilised to enhance the level of success achieved by data mining activities. The method outlines a non-linear sequence of six phases that, when followed, make it possible to create and implement a DM model in a real-world environment, hence assisting with the implementation of business decisions (Chapman, et al., 2000; IBM, 2021).

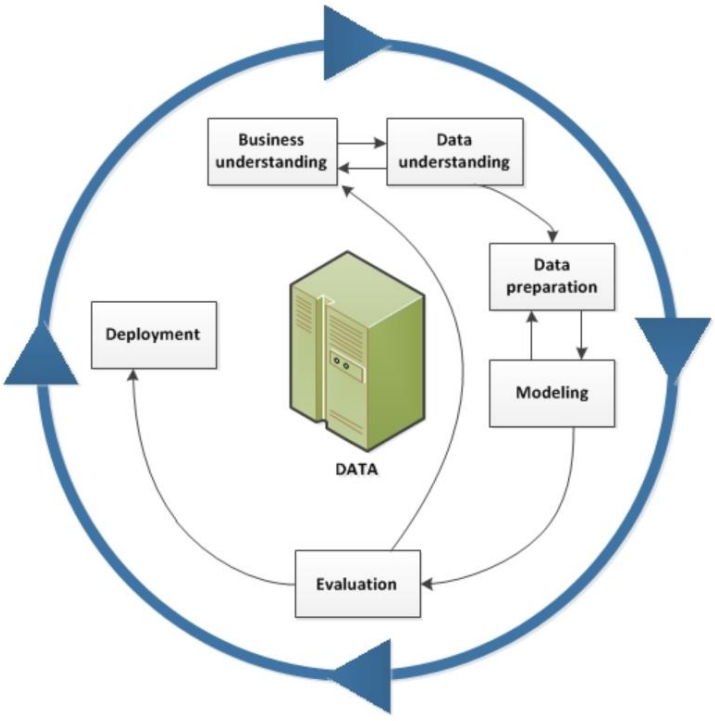

Figure 1 shows data mining life cycle (IBM, 2021)

3.1. Data Understanding

The data understanding phase of CRISP-DM begins with the initial collection of data and continues with activities to get to know the data, find problems with the quality of the data, get first insights from the data, or find interesting subsets to form hypotheses about hidden information (Wirth & Hipp, 2000).

This report aims to predict customer satisfaction and location analysis for Zomato Restaurants in Bangalore. The dataset used for analysis is downloaded from Kaggle and contains 51717 restaurants. The dataset contains 17 variables but only 11 variables are taken into consideration

Variable	Description
name (chr)	Contains the name of the restaurant.
online_order (chr)	Whether online ordering is available in the restaurant or not.
book_table (chr)	Table book option available or not.
rate (chr)	Contains the overall rating of the restaurant out of 5.
votes (int)	Contains total number of rating for the restaurant as of the above mentioned date.
location (chr)	Contains the neighborhood in which the restaurant is located.
rest_type (chr)	Restaurant type.
dish_liked (chr)	Dishes people liked in the restaurant.
cuisines (chr)	Food styles, separated by comma.
approx_cost(for two people) (chr)	Contains the approximate cost for meal for two people (₹).
listed_in(city) (chr)	Contains the neighborhood in which the restaurant is listed.

Table 1 shows Independent Variables which are taken into consideration

3.2. Data Preparation

The data preparation phase in CRISP-DM includes everything that needs to be done to turn the raw data into the final dataset, which is the data that will be fed into the modelling tool(s). Tasks for preparing data are likely to be done more than once, and not in any particular order. Some of the tasks are choosing tables, records, and attributes, cleaning the data, making new attributes, and transforming the data for modelling tools (Wirth & Hipp, 2000).

New Variables which are taken into consideration are:

- Satisfaction Rate: The dataset does not contain a dependent variable; a dependent variable is being created called as "Satisfaction rate". The ratings that were submitted by the customers are used in the calculation of the "Satisfaction Rate." If the ratings that were given by the customers were 4 or higher, then the specific restaurant in particular was considered to have satisfied customers. On the other hand, if the rating was less than 3, then the restaurant was not considered to have satisfied customers. The overall score, out of 5, is provided in the dataset. Initially customers with ratings above 3 was considered as Positive, but due to overfitting and better results 4 is been used.

- High Activity Place: Regarding the location, there is also attention given to a new variable that goes by the name of "happening_place". It is referred to be a high activity place when the location in consideration is surrounded by an information technology park, university or colleges, a tourist destination, or a commercial area. If it is neither of those things, then it would refer to the location as a Non-High Activity Place. All the locations in neighborhoods of Bangalore City are been studied and analyzed carefully. All of these changes have been made in a separate Excel worksheet.

- Distance from center of the city: The list of cities and high-activity place is been uploaded to Google Sheets. An Add-on called "Geocode by Awesome Table" is been added. With the help of the add-on, the coordinates of longitude and latitude were retrieved from google maps. There were a few data inconsistencies, such as incorrect coordinates being collected from locations other than the city of Bangalore; however, this problem was addressed by updating the coordinates manually. The distance was calculated based on longitude and latitude using the below formula in google sheets.

  Mathematical Formula:

$$\text{Distance} = \text{ACOS}(\ \text{SIN}(\text{lat1}) * \text{SIN}(\text{lat2}) + \text{COS}(\text{lat1}) * \text{COS}(\text{lat2}) * \text{COS}(\text{lon2} - \text{lon1})\ ) * 6371000$$

  Where lat1 and lon1 are reference latitude and longitude. They are the coordinates of the centre of the city in Bangalore. In this scenario MG Road location is considered a reference point (Interview Area, 2022; Movable-type, 2022).

  And lat2 and lon2 are the locations whose distance needs to be calculated from the centre of the city. The distance is calculated in miles.

Location	Comment	High Activity Place?	Latitude	Longitude	Distance from Center of City(Miles)
Banashankari	Famous Temple a bit	Yes	12.9255	77.54676	5.422224417
Bannerghatta Road	Nearby to schools, colleges and IT parks	Yes	12.7943	77.62084	12.49099692
Basavanagudi	residential and commercial locality	Yes	12.9406	77.57376	3.365333712
Bellandur	It is the largest lake in Bangalore, and separates Bellandur from the HAL Airport.	No	12.9304	77.6784	5.560495249
Brigade Road	It is a large commercial centre and one of the busiest shopping areas in the heart of Bangalore	Yes	12.971	77.6069	0.308332227
Brookefield	IT Companies like IBM, SAP have their offices here.	Yes	12.9655	77.71846	7.368134875
BTM	Residential	No	12.9083	77.60508	4.595909475
Church Street	Busiest streets in the Central Business District of Bangalore	Yes	12.9751	77.6047	0.32090771
Electronic City	IT hub in Bangalore	Yes	12.8452	77.66017	9.577011462
Frazer Town	Suburb of Bangalore	No	12.997	77.61441	1.580021823
HSR	Several educational institutions have also spring up in the area,	Yes	12.9121	77.64455	4.928229336
Indiranagar	Residential Area	No	12.9784	77.64084	2.128587593
Jayanagar	Residential and commercial neighbourhood in Bangalore	Yes	12.9308	77.58383	3.491095099
JP Nagar	Residential area	No	12.9063	77.58568	4.986950799
Kalyan Nagar	Residential area	No	13.024	77.64329	4.098067898
Kammanahalli	Famous for Food joints, street foods, super markets, Afro Arabian sweet spot,	Yes	13.0159	77.63786	3.430238991
Koramangala 4th Block	Residential area	No	12.9315	77.62999	3.292342647
Koramangala 5th Block	Residential area	No	12.9352	77.61996	2.818510735
Koramangala 6th Block	Residential area	No	12.9382	77.62283	2.67945002
Koramangala 7th Block	Residential area	No	12.9363	77.6128	2.663844293
Lavelle Road	commercial street in the city of Bangalore	Yes	12.9712	77.59787	0.816802653
Malleshwaram	One of the oldest areas in the city and it boasts of a rich cultural and social life.	Yes	13.0055	77.56924	3.444331473
Marathahalli	Residential area	No	12.9569	77.70113	6.294270618
MG Road	It is a hub of recreational and commercial activity in the city.	Yes	12.9747	77.60945	0.1
New BEL Road	Commercial and residential street	Yes	13.0302	77.57057	4.644939365
Old Airport Road	It just a highway	No	12.9592	103.8858	1768.536615
Rajajinagar	Residential neighborhood and business hub	Yes	12.9982	77.55304	4.129561516
Residency Road	It's near city Center	Yes	12.9715	77.60621	0.310293461
Sarjapur Road	Outskirt of the city	No	12.9035	77.70386	8.039695236
Whitefield	Tech Park, shopping and entertainment hub.	Yes	12.9714	77.75013	9.475430473

Table 2. Shows new attributes that are been added

Later the file consisting with location, happening place and distance was saved with the name zom_loc.csv

3.2.1. Database development in Microsoft Access

Microsoft Access was used to import two different text files, zomato.csv and zom_loc.csv. When uploading the text file into MS Access, it is checked to ensure that the data types of the foreign keys are consistent with one another. After the importation of all of the necessary data, the tables are prepared to join using SQL query. The Zomato Table is the master table, which includes all of the essential information. A foreign key is a column that helps connect two different tables, and each table has its unique foreign key. In this context, location is a foreign key that is associated with the listed in.city column. A new analytical base table (ABT) with the name "zomato_new" is produced by using a left join to connect both of the existing tables. Because all of the records

from the left table, which is the main zomato table, need to be retrieved for ABT, a left join is the type of join that is utilized. After joining the table the text file is then exported.

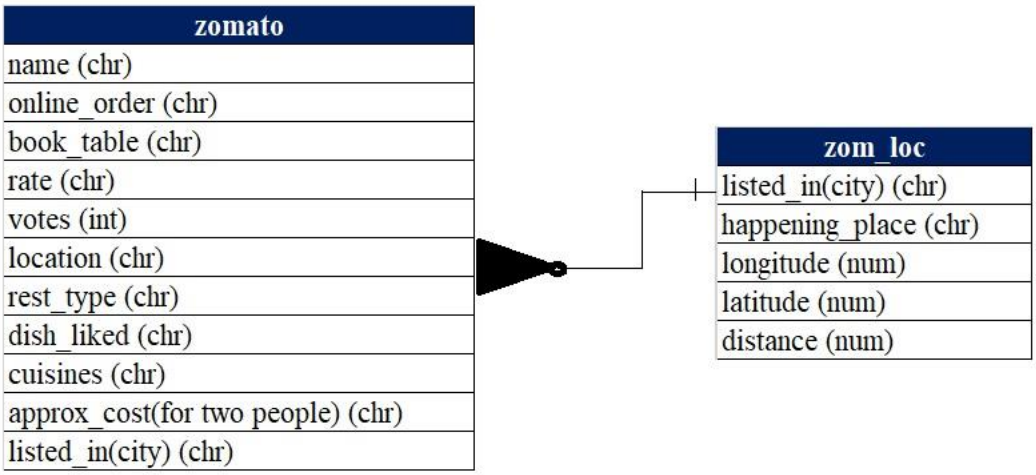

Fig 2 shows zomato table connecting zom_loc table

3.2.2. Cleaning of Data:

After loading the data set into R-studio the data is validated if there are any NA values are not. The row with NA and duplicate values are removed. After removing the unwanted rows the dataset contains 17187 observations. Columns 'name', 'online_order', 'book_table', 'location', 'rest_type' and 'listed_in.city.' are converted into factors.

- rate: The ratings given are in strings. The rating needs to be either in numeric or integer. The ratings consist of "X/5" for example "4.1/5". '/5' extra string from the rating column is been removed. And then the rate column is converted into numeric.
- rest_type : There are 73 levels in the column of Restaurant Type. A few of the levels are repeated with a different synonymous name. All 73 labels are been renamed as shown below. The unused labels are also been dropped as shown in the below table.

New Variable	Existing Variables
Bakery	Bakery, Cafe', 'Bakery, Dessert Parlor' and 'Bakery, Quick Bites'
Bar	Bar, Casual Dining', 'Bar, Pub' and 'Pub'
Beverage Shop	Beverage Shop, Cafe', 'Beverage Shop, Dessert Parlor' and 'Beverage Shop, Quick Bites'
Café	Cafe, Bakery', 'Cafe, Dessert Parlor' and 'Cafe, Quick Bites'
Casual Dining	Cafe, Casual Dining', 'Casual Dining, Bar', 'Casual Dining, Pub', 'Casual Dining, Cafe', 'Casual Dining, Lounge' and 'Food Court, Casual Dining'
Dessert Parlor	Dessert Parlor, Bakery', 'Dessert Parlor, Beverage Shop', 'Dessert Parlor, Cafe', 'Dessert Parlor, Kiosk' and 'Dessert Parlor, Quick Bites'
Food Court	Food Court, Quick Bites'
Quick Bites	Quick Bites, Cafe', 'Quick Bites, Dessert Parlor', 'Quick Bites, Food Court', 'Quick Bites, Bakery' and 'Quick Bites, Beverage Shop'
Sweet Shop	Sweet Shop, Quick Bites' and 'Quick Bites, Sweet Shop'
Takeaway	Takeaway, Delivery' and 'Takeaway'

Table 3 shows renaming the levels for restaurant type.

3.2.3. Post-Cleaning Descriptive Statistics

rest_type	
Bakery	383
Bar	57
Beverage Shop	334
Cafe	2424
Casual Dining	5710
Delivery	704
Dessert Parlor	1360
Food Court	244
Quick Bites	5393
Sweet Shop	234
Takeaway	344

online_order	
No	Yes
3756	13431

book_table	
No	Yes
15482	1705

happening_place	
No	Yes
9210	7977

Satisfaction_rate	
Negative	Positive
9993	7194

Table 4 Shows Descriptive Statistics of categorical variables

Variable	Mean	Std. Dev	Min	25%	Median	75%	Max
rate	3.825	0.421	2	3.7	3.9	4.1	4.9
votes	397.9	804.23	0	89	179	397	14726
approx_cost.for.two.people.	511.4	197.76	40	400	500	650	950
Distance_from_Heart	3.827	2.618	0.1	2.664	3.365	4.928	12

Table 5 Shows Descriptive Statistics of Numerical Variables

3.3. Data Modelling

After cleaning of the data, the data would be analyzed by using different statistical techniques.

3.3.1. Centrality Measures:

Centrality measures are one of the most important ways to understand networks, which are often also called graphs. Graph theory is used by these algorithms to figure out how important each node in a network is. They sort through noisy data to find parts of the network that need attention, but they all operate differently (Disney, 2020). A measure of

centrality is an index that gives each node in a network a numerical value. The more central a node is, the higher its value (SCHOCH, 2018).

There are different measures of centrality:

A. Degree: The simplest measure of centrality to figure out is degree centrality. Recall that a node's degree is just a count of how many edges it has, or social connections. The degree centrality of a node is simply its degree. The degree centrality of a node with 10 connections to other nodes is 10. The degree centrality of a node with one edge is 1 (Golbeck, 2015). More central nodes have higher degrees. This can be a beneficial method for determining centrality since many nodes with high degrees also have high centrality by other measures (Golbeck, 2013).

B. Betweenness: Betweenness centrality is a way to figure out how much a node affects the way information flows through a graph. It is usually used to find nodes that connect two different parts of a graph. The algorithm figures out the shortest paths between every pair of nodes in a graph that don't have any weights. Each node gets a score that is based on how many shortest paths go through it. The betweenness centrality score of a node will be higher if it is more often on the shortest path between other nodes (NEO4J, 2022).

C. Closeness: Closeness centrality is a way to find the nodes in a graph that can spread information quickly and well. The closeness centrality of a node measures how far it is from all other nodes on average. Nodes that have a high "closeness" score are closest to all other nodes. The Closeness Centrality algorithm figures out how far each node is from every other node by finding the shortest path between every pair of nodes. The sum is then turned around to find the score for that node's closeness centrality (NEO4J, 2022).

3.3.2. Logistic Regression

Logistic regression is a way to figure out how likely it is that something will occur, like voting or not voting, based on a set of independent variables. The dependent variable is between 0 and 1 because the outcome is a probability. In logistic regression, the odds are given a logit transformation. This is done by dividing the probability of success by the probability of failure. This is also known as the log odds or the natural logarithm of odds. The following formulas show how to use this logistic function (IBM, 2022) (Molnar, 2022) :

$$\text{logistic}(\eta) = \frac{1}{1 + \exp(-\eta)}$$

In logistic regression, as shown in Figure 3, we want to find the classification boundary line, which is shown by the regression formula. In the regression formula, the training classifier uses the optimization algorithm to find the best regression coefficient. Logistic regression-based classification takes any set of data as input and uses a function to determine how the data should be categorised (Zou, et al., 2019).

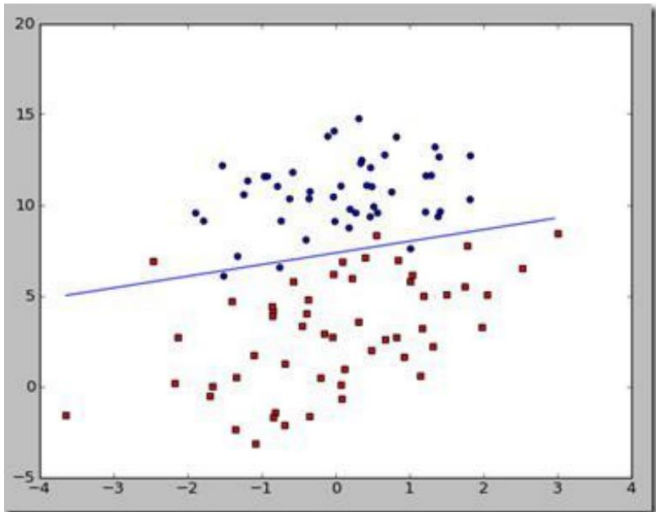

Figure 3 Shows Logistic Regression

Measures of fit and predictive power are used to figure out how good a logistic regression model is. R-squared is a number that ranges from 0 to 1 that shows how well the independent variable in a logistic function can be predicted based on the dependent variables. There are many ways to figure out R-square, such as the Cox-Snell R2 and the McFadden R2. On the other hand, tests like the Pearson chi-square, Hosmer-Lemeshow, and Stukel tests can be used to measure how well something fits. The right type of test to use will depend on things like how p-values are spread out, how interactions and quadratic effects work, and how the data are grouped (NVIDIA, 2022).

3.3.3. Naive Bayes

Naive Bayes Classifier is one of the simplest and most effective classification algorithms. It helps build fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means that it makes predictions based on how likely the probability is (javaTpoint, 2022). Nave Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem. It is used for a wide variety of classification tasks (Chauhan, 2020).

Bayes Theorem: Bayes' Theorem is a simple formula for determining conditional probabilities. Conditional probability is a way to figure out how likely it is that something will happen based on the fact that something else has already happened.

The mathematical formula is given as (CFI, 2022):

$$P(A|B) = P(A)\ P(B|A)\mathbf{P(B)}$$

P(A|B) – the probability of event A occurring, given event B has occurred

P(B|A) – the probability of event B occurring, given event A has occurred

P(A) – the probability of event A

P(B) – the probability of event B

The Naive Bayes algorithm (NB) is a Bayesian graphical model with nodes that match each column or feature. It is called "naive" because it doesn't take into account how parameters were distributed before and assumes that all features and rows are independent. The benefit is that it can plug in any kind of distribution over individual features and use the data to find the most likely features. It need not restrict the class of prior distributions to exponential family in order to simplify algebra of product of likelihood and prior (Yeturu, 2020).

Different kinds of Naive Bayes models (Rekhit Pachanekar, 2022):

1. Multinomial: This model is used to classify data that is made up of separate pieces. As a simple example, we can use the weather (cloudy, sunny, or raining) as our input and then check to see when a tennis match takes place.

2. Gaussian: As the name suggests, it used for continuous data that has a Gaussian distribution in this model. The temperature of the stadium where the game is played is one example.

3. Binomial: What if the input data is just yes or no? This is called a "boolean value." It is used for the binomial model in this case.

3.3.4. Random Forest:

Random forest is a popular method of machine learning that can be used to build models that can predict. Breiman first brought up the idea in 2001. According to (Breiman, 2001) **"**Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest." Random forests are groups of classification and regression trees, which are simple models that use binary splits on predictor variables to make predictions outcomes. Random forest is an excellent method for making predictions because it can handle datasets with a lot of predictor variables. However, in practice, the number of predictors required for obtaining outcome predictions should be minimized to improve efficiency (Speise, et al., 2019).

By sampling with replacement, about a third of the cases are left out of the training set for the current tree. As more trees are added to the forest, this "out-of-bag" (oob) data is used to get an unbiased estimate of the classification error. It is also used to figure out how important a variable is. After each tree is made, all of the data are run down the tree, and for each pair of cases, the distance between them is calculated. If two cases share the same terminal node, they are one step closer to each other. At the end of the run, the distances are normalised by dividing by the number of trees. Proximities are used to fill in missing data, find outliers, and make low-dimensional views of the data that are clear and helpful (Berkeley.edu, 2022).

Random Forests allow us to look at feature importance, which is how much the Gini Index for a feature goes down at each split. The mean decrease in Gini coefficient is a way to figure out how much each variable affects how similar the random forest's

nodes and leaves are to each other. The more important a variable is in the model, the higher it is mean decrease accuracy or mean decrease Gini score (Tat, 2017).

3.3.5. Extreme Gradient Boost Classification

Gradient boosting is a type of machine learning algorithm that uses a group of algorithms to solve classification or regression modelling problems. Decision tree models are used to build ensembles. One tree at a time is added to the group of trees, which is then fit to correct the prediction errors made by prior models. This type of ensemble machine learning model is called "boosting" (Brownlee, 2020). Boosting Unlike regular Gradient Boosting, Extreme Gradient Boosting has its own way of building trees, where the Similarity Score and Gain are used to figure out where the best node splits should be. (Dobilas, 2021).

Extreme Gradient Boosting, also known as XGBoost, is an open-source version of the gradient boosting algorithm. It is designed to be both computationally efficient (fast to run) and very effective, potentially even more efficient than other open-source implementations. Most of the time, XGBoost is faster than other ways of using gradient boosting (Brownlee, 2020).

3.3.6. Decision Tree

A decision tree is a diagram that shows how a set of choices might affect each other. It lets a person or group compare different possible actions based on their attributes, chances of success, and benefits. They can be used to start a general discussion or to make a plan for a mathematical formula that predicts the best choice

(LucidChart, 2022). The decision to make strategic splits has a big effect on how accurate a tree is. Multiple algorithms are used to decide whether or not to split a node into two or more sub-nodes. When sub-nodes are made, they become more similar to each other. In other words, we can say that the node's purity goes up as the target variable goes up. The decision tree divides the nodes into sub-nodes based on all of the available variables and then chooses the division that makes the most similar sub-nodes. The type of target variables is one of the main factor for choosing the algorithm (Chauhan, 2022).

When using decision trees to make a choice, there are two steps. The first step is to build the decision tree. This is done by drawing the tree and putting all of the probabilities and outcome values on it. Throughout, the principle of relevant attributes is used, which means that only relevant costs and revenues are taken into account. The second step is to look at what has been done and make suggestions. Here, the decision is "rolled back" by figuring out all the expected values at each of the outcome points and using those to make decisions as you move back across the decision tree. Then, management is given advice on what to do next (ACCA, 2022).

4. FINDINGS

4.1. Data Understanding

Following the evaluation of each variable on its own, it became abundantly evident that corrective action was required before any measurements of association could even be considered. Once the data has been prepared a proper relation between customer satisfaction rate and location can be examined. To achieve a more in-depth comprehension of these connections, several statistical analyses and representations of the data have been utilised. Using the Chi-Square test, which yields a p-value of less than 0.05, it is evident that the distance from the city center variable has a significant impact on the satisfaction rate.

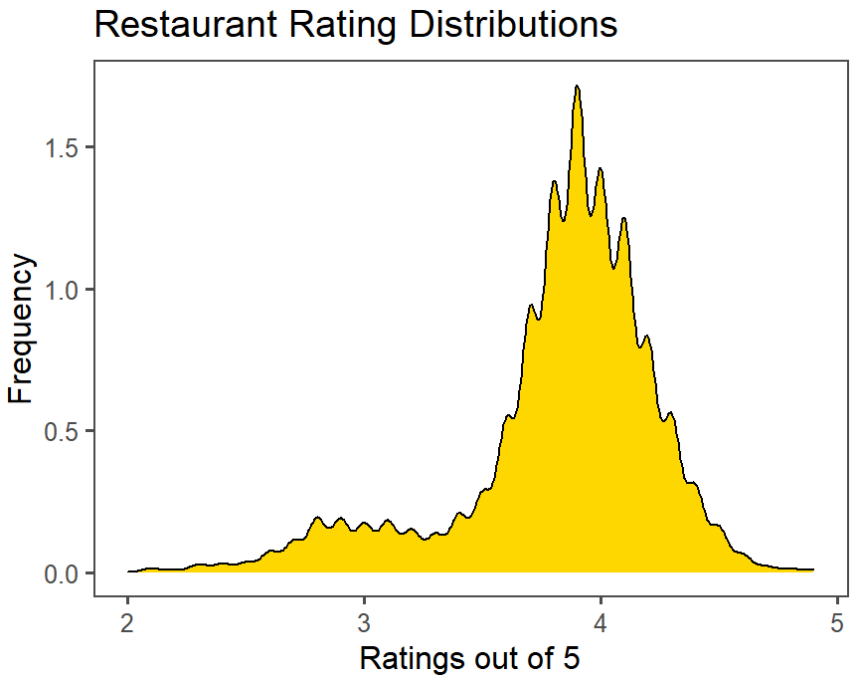

Figure 4. Shows Rating Distribution

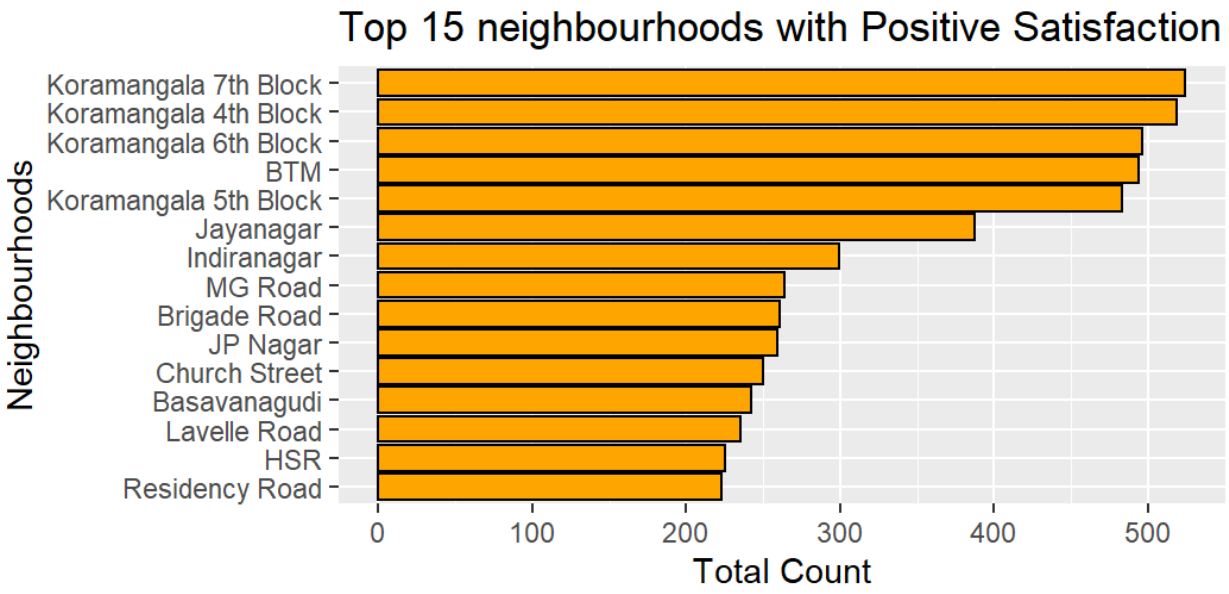

Figure 5. Top Locations in Bangalore with maximum Positive Satisfaction Rate

Major Neighbourhoods	Positive	Negative	% Positive
BTM	494	638	0.436396
Jayanagar	387	521	0.426211
Indiranagar	299	376	0.442963
Koramangala 4th Block	519	542	0.489161
Koramangala 5th Block	483	522	0.480597
Koramangala 6th Block	496	488	**0.504065**
Koramangala 7th Block	524	541	0.492019

Table 6. Top Neighbourhoods in Banglore with maximum % in Positive ratings

It is possible to deduce from figure 1 that the maximum customers have given restaurant ratings falls between 3.5 and 4.5. One further thing that stands out is the fact that the vast majority of restaurants have positive feedback from customers. It can also be observed that BTM, Koramangala, Jayanagar and Indiranagar are the top spots having maximum satisfaction rates restaurants.

4.2. Data Findings

4.2.1. Logistic Regression:

After taking into account all of the variables—including 'listed in.city','rest type', 'online order', 'book table', 'votes', and 'approx cost.for.two.people '— an initial logistic regression model has been developed. The 'Satisfaction rate' variable is the one that is being looked at as a potential target variable. Since the variables 'happening place' and 'Distance_from_heart' are derived from location variable and the variable 'Satisfaction rate' is derived from 'rate,' which has ratings of all the restaurants, Location and rate variables were not considered for the next model. The second logistic regression has been created consisting of all the variables except for location and rate variables, because of the multi-collinearity factor, the variables location and rate are disregarded as redundant. In addition, the VIF test is applied to investigate the nature of the relationship between the variables in question and locate any possible instances of multicollinearity. It was discovered that the VIF score for each of the characteristics included in the model was lower than 2 for all except the Restaurant Types. The multicollinearity of the model's attributes increases proportionally with the value of the VIF parameter. The total count of standard residuals that are more than 1.96 is 129. The computed Cook's distance is also equal to zero. In most cases, an outlier is defined as having a Cook's Distance that is more than 4/n, where n refers to the total number of data points.

The accuracy rate for model 1 is 74.06 percent, while the accuracy level that was anticipated for model 2 was 74.21 percent. It is possible to remark that location is a significant factor that ought to be taken into consideration to achieve a higher rating from customers. From figure 6 it can be observed that MG Road, Malleshwaram, Lavelle Road and Church Street are highly significant and important location point where a better satisfaction rate can be observed. From figure 7 it is probable to determine that the attribute of "distance from heart" is highly statistically significant

because the p-value is close to 0. On the other hand, the attribute of "happening place" is not statistically significant because the value of the p-value tends to be greater than 0.05. Other variables, including varied levels in restaurant kinds such as beverage shop, café, dessert parlour, and take away, have p values that are less than 0.5 and are statistically significant. Examples of these sorts of restaurants include: In addition to this, it was found that the cost of booking a table and votes were statistically significant. By looking at the model's coefficients, it's evident that not all variables are statistically significant.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.11129982	0.20593970	-5.396	0.00000006805228	***
listed_in.city.Bannerghatta Road	-0.08517114	0.19212430	-0.443	0.657540	
listed_in.city.Basavanagudi	0.28893108	0.18563420	1.556	0.119600	
listed_in.city.Bellandur	0.05043734	0.19889026	0.254	0.799809	
listed_in.city.Brigade Road	0.57321190	0.18500175	3.098	0.001946	**
listed_in.city.Brookefield	-0.23240805	0.19595182	-1.186	0.235604	
listed_in.city.BTM	0.27979618	0.16565742	1.689	0.091219	.
listed_in.city.Church Street	0.71183528	0.18542902	3.839	0.000124	***
listed_in.city.Electronic City	-0.71539139	0.23963433	-2.985	0.002833	**
listed_in.city.Frazer Town	0.57950372	0.19476490	2.975	0.002926	**
listed_in.city.HSR	-0.03092534	0.18049299	-0.171	0.863958	
listed_in.city.Indiranagar	0.26537423	0.17772049	1.493	0.135382	
listed_in.city.Jayanagar	0.19997253	0.17019922	1.175	0.240022	
listed_in.city.JP Nagar	0.03989427	0.17851188	0.223	0.823160	
listed_in.city.Kalyan Nagar	0.18575604	0.19364968	0.959	0.337439	
listed_in.city.Kammanahalli	0.14667519	0.19377822	0.757	0.449096	
listed_in.city.Koramangala 4th Block	0.43986807	0.16680378	2.637	0.008363	**
listed_in.city.Koramangala 5th Block	0.42340807	0.16802377	2.520	0.011738	*
listed_in.city.Koramangala 6th Block	0.42174272	0.16901761	2.495	0.012587	*
listed_in.city.Koramangala 7th Block	0.45018766	0.16649161	2.704	0.006852	**
listed_in.city.Lavelle Road	0.65484517	0.18851001	3.474	0.000513	***
listed_in.city.Malleshwaram	0.64743587	0.19501037	3.320	0.000900	***
listed_in.city.Marathahalli	-0.18255963	0.19324974	-0.945	0.344821	
listed_in.city.MG Road	0.61408367	0.18356360	3.345	0.000822	***
listed_in.city.New BEL Road	0.17163911	0.22445287	0.765	0.444450	
listed_in.city.Old Airport Road	0.05265112	0.18706385	0.281	0.778357	
listed_in.city.Rajajinagar	0.53070714	0.20873997	2.542	0.011008	*
listed_in.city.Residency Road	0.57958524	0.19228468	3.014	0.002577	**
listed_in.city.Sarjapur Road	-0.03270252	0.20472757	-0.160	0.873088	
listed_in.city.Whitefield	-0.19072068	0.20144389	-0.947	0.343757	

Figure 6. Shows Summary of Logistic Regression model 1 consisting of only locations

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.508217250	0.148088147	-3.432	0.000599	***
happening_placeYes	0.004096057	0.040869054	0.100	0.920167	
Distance_from_Heart	-0.090136043	0.008154077	-11.054	< 0.0000000000000002	***
rest_typeBar	-2.178108559	0.574011688	-3.795	0.000148	***
rest_typeBeverage Shop	0.834586279	0.179130144	4.659	0.00000317586074	***
rest_typeCafe	0.345394652	0.139744816	2.472	0.013451	*
rest_typeCasual Dining	-1.266343939	0.138358304	-9.153	< 0.0000000000000002	***
rest_typeDelivery	-0.238866780	0.158133658	-1.511	0.130906	
rest_typeDessert Parlor	1.012060326	0.144433310	7.007	0.00000000000243	***
rest_typeFood Court	-0.846773731	0.220116121	-3.847	0.000120	***
rest_typeQuick Bites	-0.671183522	0.132238006	-5.076	0.00000038633197	***
rest_typeSweet Shop	-0.335408651	0.204646666	-1.639	0.101221	
rest_typeTakeaway	-0.511866251	0.183651301	-2.787	0.005317	**
online_orderYes	-0.036729285	0.050951555	-0.721	0.470991	
book_tableYes	1.158116032	0.080520082	14.383	< 0.0000000000000002	***
votes	0.003065745	0.000089288	34.336	< 0.0000000000000002	***
approx_cost.for.two.people.	-0.000005535	0.000152629	-0.036	0.971074	

Figure 7. Shows Summary of Logistic Regression model 1 consisting of only locations

It can also be observed from the Logistic Regression model summary that with each 1 unit increase in distance from city center (in miles) the log odds of the customer getting satisfied and giving a positive rating decreases by 0.09. Similarly, for each unit of happening place, the log odds of a positive rating is been increased by 0.004. 1 unit increase in booking of table and votes with the log odds of the customer getting satisfied was an increase in 1.15 and 0.003 respectively. It was seen as a 1 unit increase in approximate cost for 2 people in the restaurant had a decrease in customer satisfaction by 0.000005 in the attribute.

4.2.2. Naives Bayes

It was found that the accuracy of a single Naive Bayes model, which included all of the variables apart from Distance from Heart and happening place, was 0.7290. This model was first generated using all of the data. The accuracy of the model improved to 0.7296 after the variables happening place and Distance from Heart were added to it.

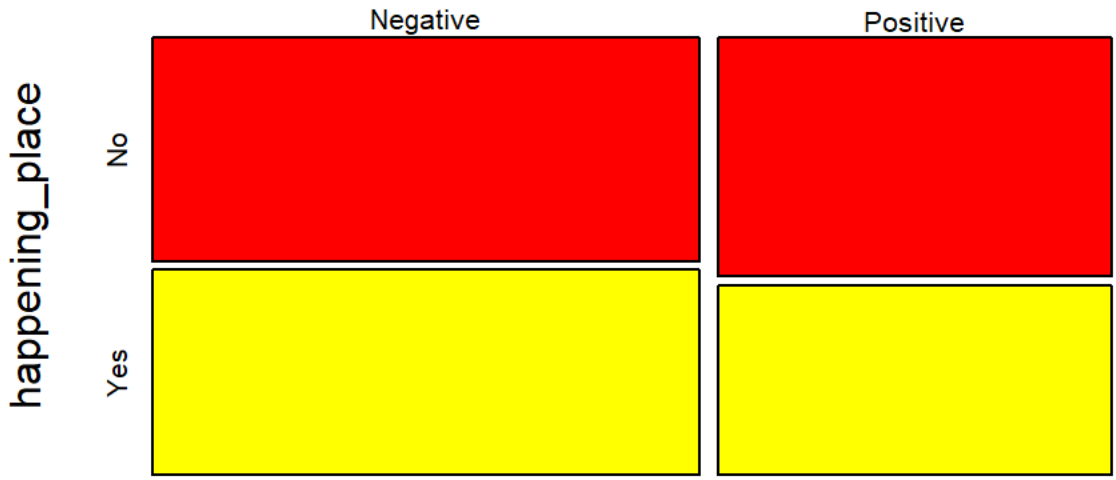

Figure 8 Naïve Bayes Model Plot – happening_place

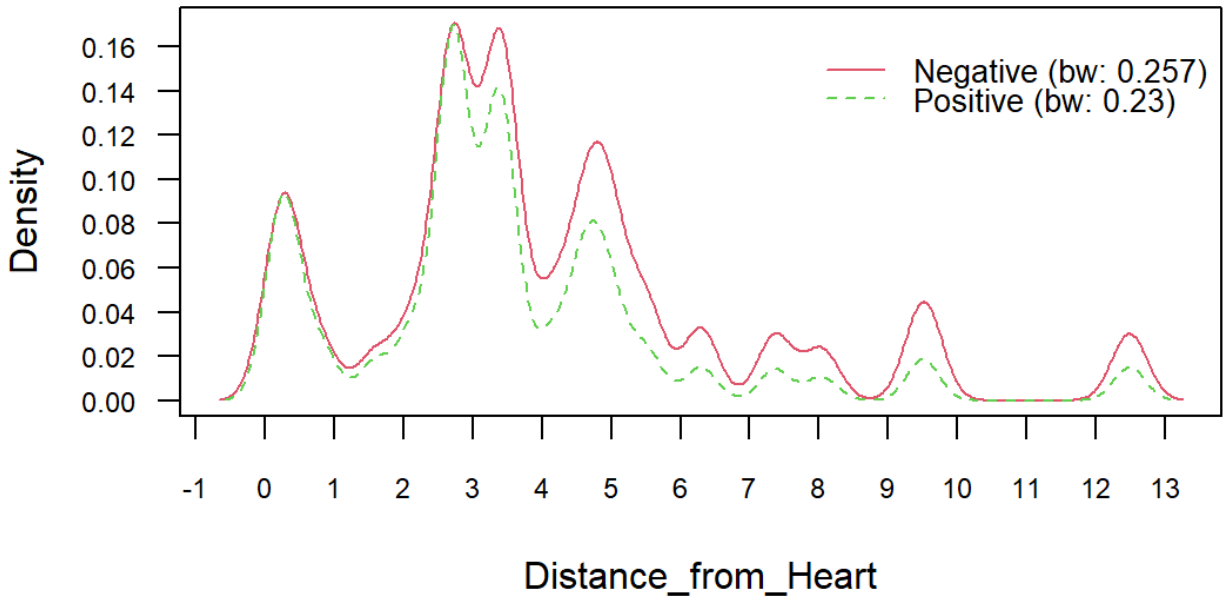

Figure 9 Naïve Bayes Model Plot – Distance_from_Heart

From figure 8. it is possible to conduct an analysis which demonstrates that the level of customer satisfaction is essentially identically distributed regardless of whether the venue in question is a happening place or not. Concerning the level of pleasure of the customers in restaurants, the

happening place is not an ideal parameter to take into consideration as a factor. In figure 9, we can see a density plot that shows the satisfaction rate is almost similar from 0 to 1 mile. The trend of satisfaction rate between different distances is almost similar, as there are maximum negative values in the dataset. This is the optimal range that one might take into consideration.

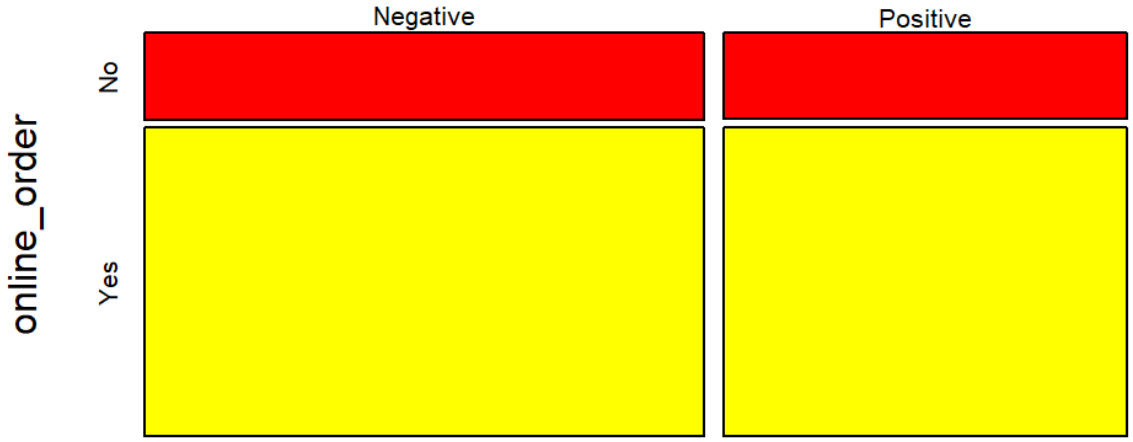

Figure 10 Naïve Bayes Model Plot – online_order

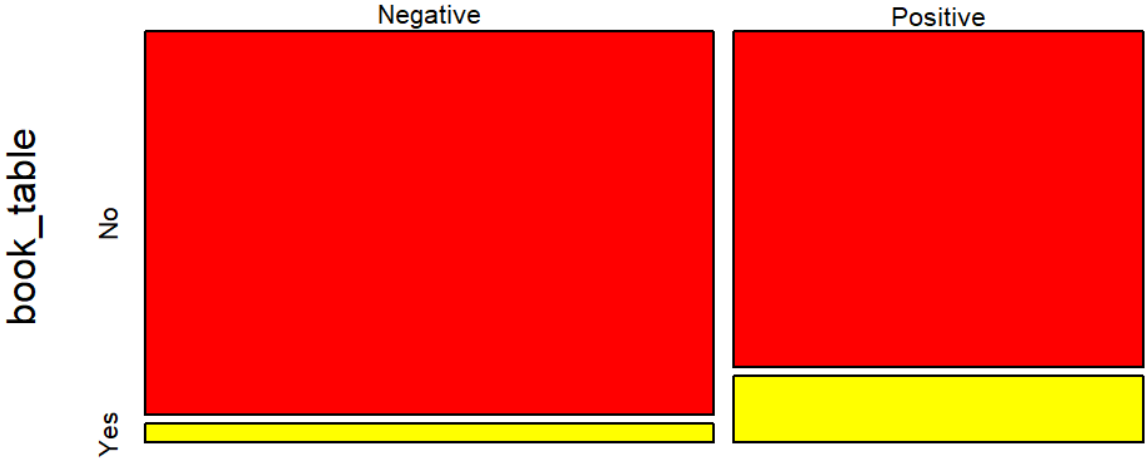

Figure 11 Naïve Bayes Model Plot – book_table

When compared to restaurants that do not offer an online delivery option, it is clear from Figure 10 that ordering online service doesn't matter to have a better customer rating. It is possible to conclude from figure 11 that the option of booking a table will not have much of an effect on the rate of customer satisfaction. As can be observed, there is a maximum possible amount of positivity in the customer satisfaction percentage even though there is no option for booking a table.

4.2.3. Random forest

The accuracy of the first model in Random Forest that was created was 93.55, and it consisted of the variable "location" but did not include any of the newly created variables. In comparison, the accuracy of the second model that was created consisted of all of the variables except location, and it had a score of 89.06. Within the context of this particular algorithm, model 1 achieves a higher level of accuracy than model 2. And in this particular case, the listed in.city characteristic has been deemed more significant than the additional variables that have been included.

It was also possible to observe that the Error rate was stabilised as the number of trees increased for both models. This was something that could be done.

It is clear from looking at figures 12 and 13 that the number of votes is the most essential factor that a restaurant takes into account because it has the largest mean decrease Gini out of all the variables in both models. The listed in.city variable had a mean decrease Gini of 886.8, whereas the distance from city center and high activity place variable each had a mean decrease Gini of 443 and 61 respectively. The more significant a variable is, the greater it is mean decrease Gini. It is also possible to conclude that the high activity place variable is the one that has the lowest impact on the customer satisfaction rating. The type of restaurant had a moderate impact on the customer satisfaction rating, while the ability to book a table and online order is seen as the least important of all the variables.

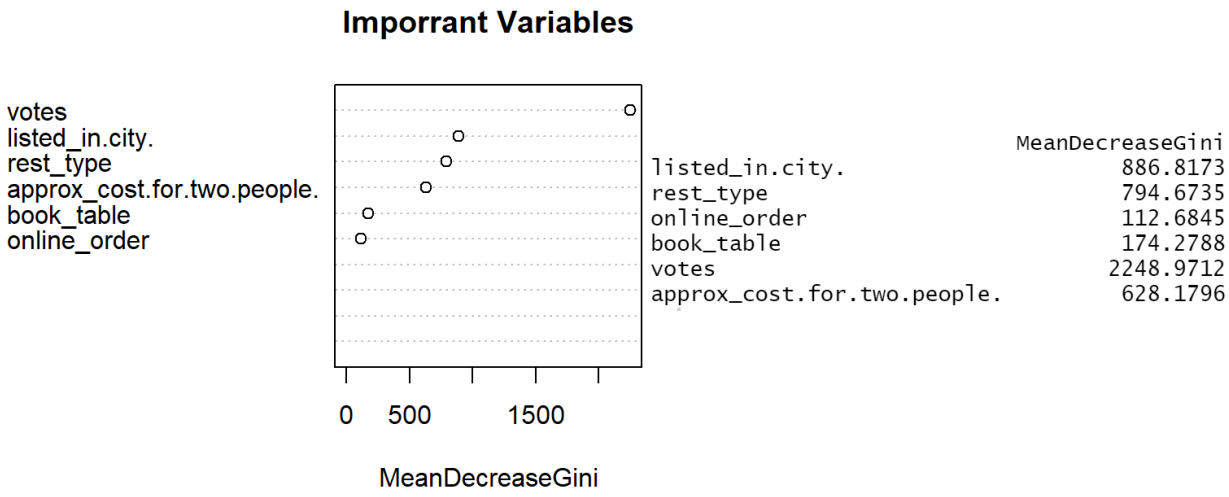

Figure 12. Shows Random forest Model 1 – Mean Decrease Gini

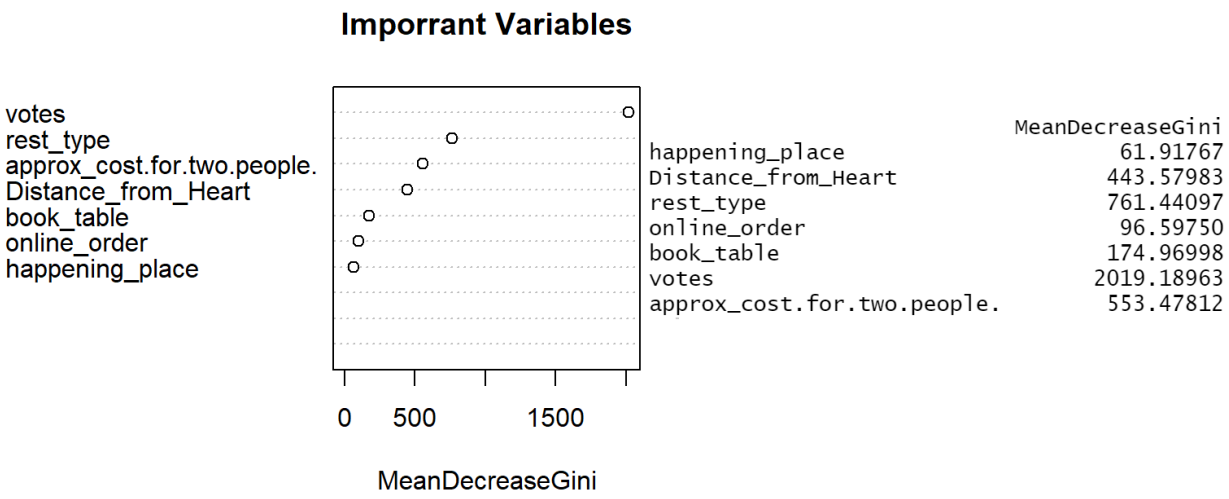

Figure 13. Shows Random forest Model 2 – Mean Decrease Gini

4.2.4. Extreme gradient boost Tree

The accuracy level of model 1, which consists of location, tends to be 0.8205, but the accuracy level of model 2, which consists of distance from city center and high activity place, is 0.8286.

The aggregate result, which is depicted in figures 14 and 15, makes it abundantly evident that almost all of the variable models had an influence that was greater than zero, except restaurant type Delivery. The data presented below, presented in the form of figures, tells us how significant each characteristic was in the process of constructing the boosted decision trees that were incorporated into the model. Nevertheless, it is reasonable that certain factors have a much greater impact on the total level of customer satisfaction. It is possible to deduce, using model 2, that the distance from the city center is another significant factor. The restaurant types Café, Casual Dining, Quick Bites, Beverage Shop and Dessert parlour were important levels while constructing a tree. And it was noticed that cities such as Electronic City and Marathahalli were important levels.

	Overall
votes	100.0000
approx_cost.for.two.people.	20.7132
rest_typeDessert Parlor	14.1825
rest_typeCafe	13.4627
book_tableYes	12.7355
rest_typeCasual Dining	8.4496
online_orderYes	3.8684
rest_typeBeverage Shop	3.1950
rest_typeQuick Bites	2.2257
listed_in.city.Electronic City	0.8190
rest_typeFood Court	0.7242
listed_in.city.Marathahalli	0.6210
rest_typeDelivery	0.5780
listed_in.city.Old Airport Road	0.5652
listed_in.city.HSR	0.4538
listed_in.city.Lavelle Road	0.4244
listed_in.city.Brookefield	0.4198
listed_in.city.Bannerghatta Road	0.4128
listed_in.city.Indiranagar	0.4021
listed_in.city.Church Street	0.3945

Figure 14. XG Boost – Importance of Variable Model 1

xgbTree variable importance

	Overall
votes	100.00000
approx_cost.for.two.people.	18.11006
Distance_from_Heart	11.64250
rest_typeDessert Parlor	10.42992
rest_typeCasual Dining	9.22809
rest_typeCafe	8.75885
book_tableYes	7.46889
online_orderYes	3.06536
rest_typeBeverage Shop	2.57954
rest_typeQuick Bites	2.14379
happening_placeYes	0.86911
rest_typeFood Court	0.24214
rest_typeSweet Shop	0.23246
rest_typeTakeaway	0.13065
rest_typeBar	0.04342
rest_typeDelivery	0.00000

Figure 15. XG Boost – Importance of Variable Model 2

4.2.5. Decision Tree

The data presented in figure 16 reveals that a positive customer satisfaction rate is found in 42 percent of the restaurants. 71% of the restaurants have customer satisfaction rate negative with votes lesser than 336. In the second node, if the votes are more than 336 and the type of restaurant specializes in Casual Dining, Delivery, Food Court, or Quick Bites, then the overall positive customer satisfaction rate is seen to be 21%. At node 4, it was also possible to observe that the customer satisfaction rating tends to be negative if the approximate cost for two persons is greater than or equal to 325. In this particular modelling effort, the identical variables

that are depicted in figure 17 were applied. Therefore, the graphical representations of the decision trees were also similar.

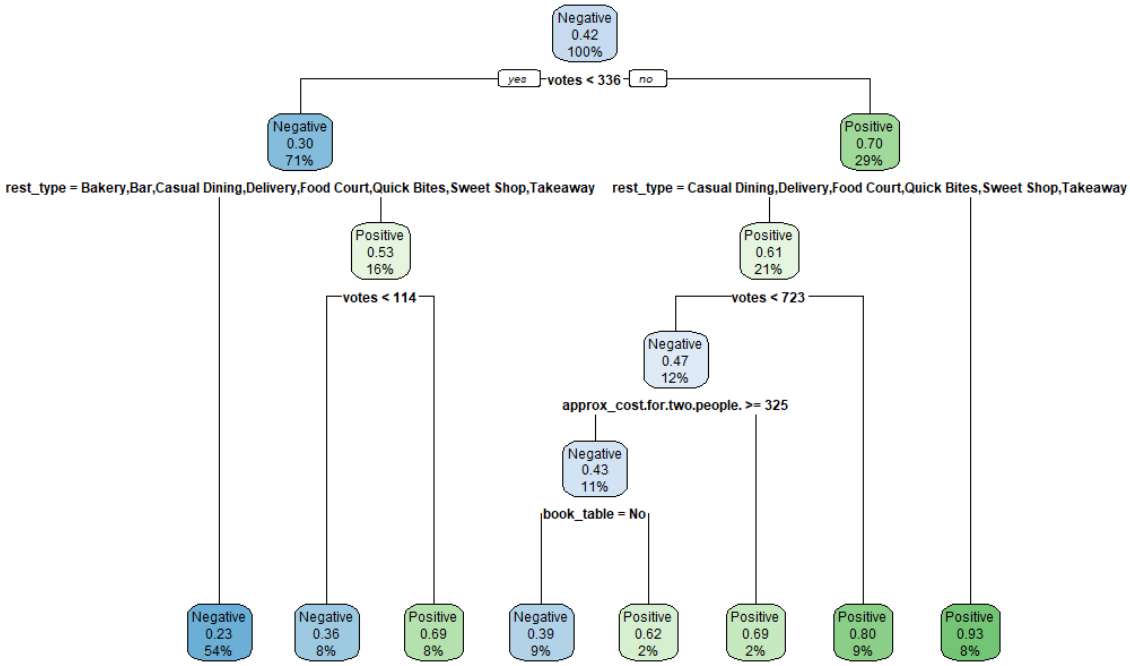

Figure 16. Decision Tree – Model 1 and Model 2

```
Variables actually used in tree construction:
[1] approx_cost.for.two.people.
[2] book_table
[3] rest_type
[4] votes
```

Figure 17. Variables used for both the models were similar

4.2.6. Models Comparison

Models	Model 1	Model 2
Logistic Regression	0.7406868	**0.7421**
Naïve Bayes	0.7290379	**0.7296924**
Random Forest	**0.9355**	0.8906
Extreme Gradient Boost Tree	0.8205949	**0.828667**
Decision Tree	0.75	0.75

table 7 Shows Accuracy Level of models

As the accuracy is reported in the same units as the dependent variable, the ratio of the number of accurate predictions to the total number of input samples is used to calculate accuracy scores for all of the different models. It can be observed from the above table that Model 2 had a better accuracy level as compared to model 1. But the overall best model with a better accuracy was Random Forest Model 1 with an accuracy level of 0.9355.

4.3. Centrality Network Analysis

By measuring how central a node or edge is, we may get a sense of how significant it is to the overall connection and flow of information in the network. It is an efficient way for targeting various locations. From below figure 18 it can be observed that all the node locations are been distributed. Table 8 shows different types of Centrality for different locations with average ratings.

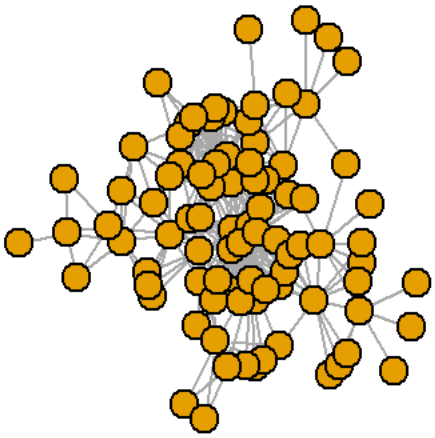

Figure 18.Shows Centrality Plot of location variable

City	degree	closeness	betweenness	eigenvector	Average_Rating
Banashankari	8	0.003921569	278.8772212	0.096355890	3.81
Banaswadi	4	0.004329004	4.8111111	0.112104472	3.66
Bannerghatta Road	9	0.004424779	11.9716077	0.331959299	3.63
Basavanagudi	14	0.005181347	333.3345642	0.457503169	3.85
Basaveshwara Nagar	2	0.003484321	0.0000000	0.020429990	3.82
Bellandur	8	0.004405286	65.5839707	0.150740051	3.57
Bommanahalli	7	0.004255319	1.8778092	0.284403861	3.05

Table 8. Shows different Centralities with average rating

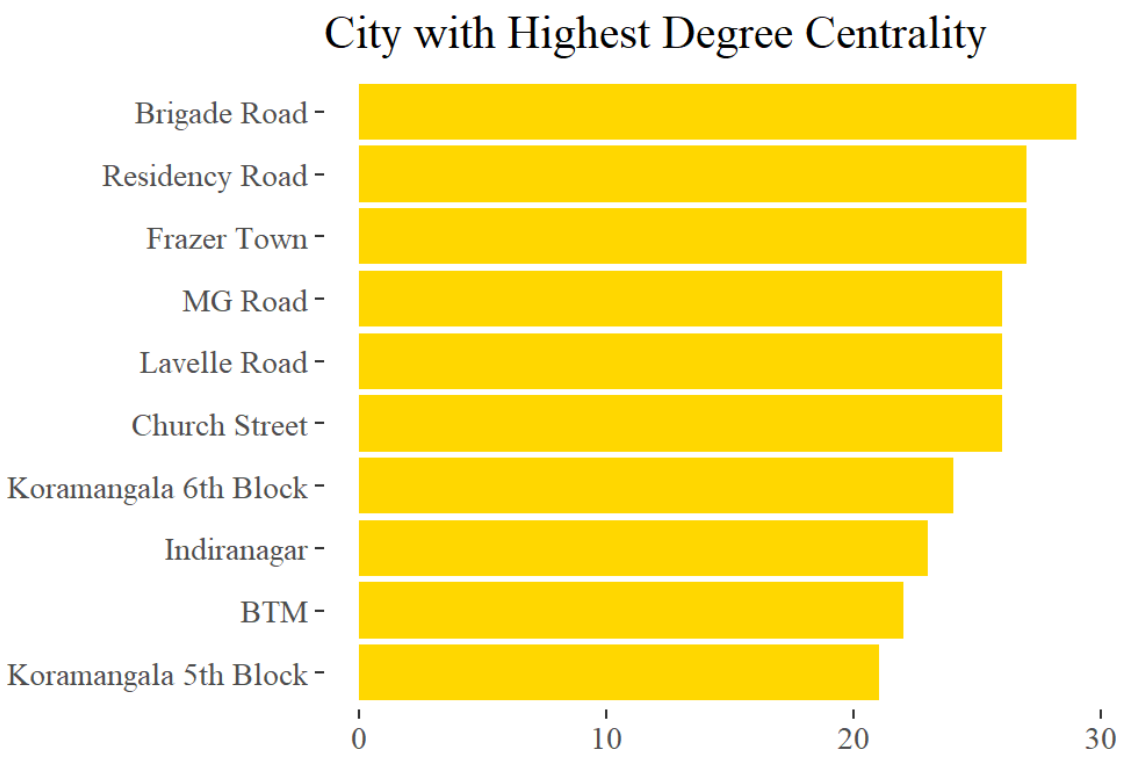

Figure 19. Shows Degree centrality in Descending order

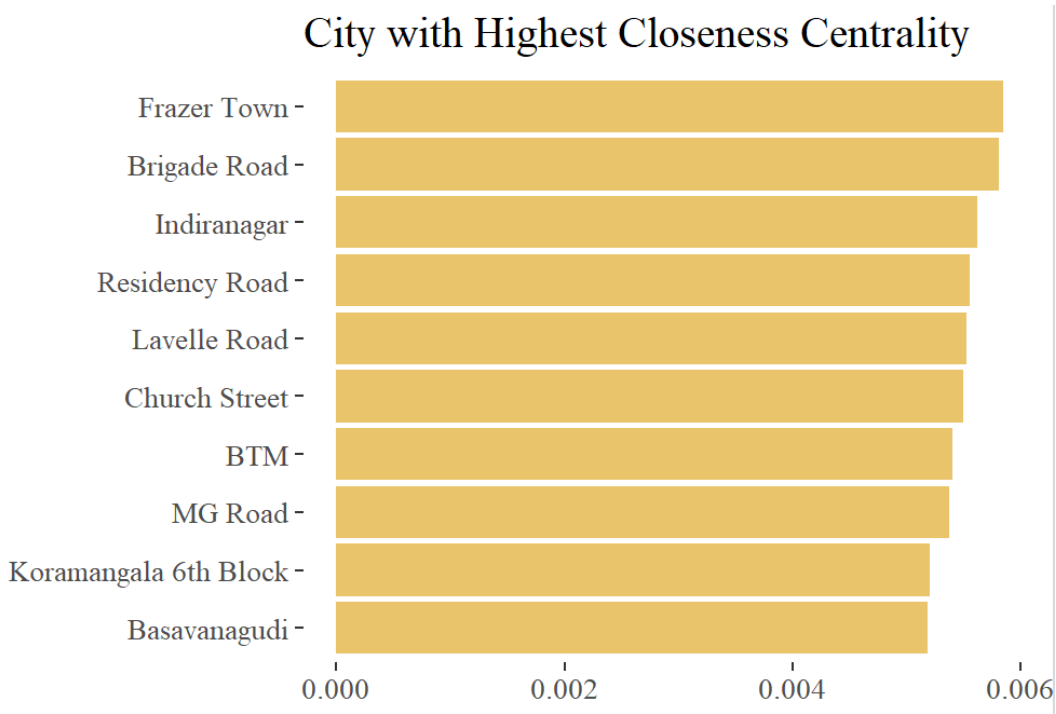

Figure 20 Show Closeness Centrality in descending order

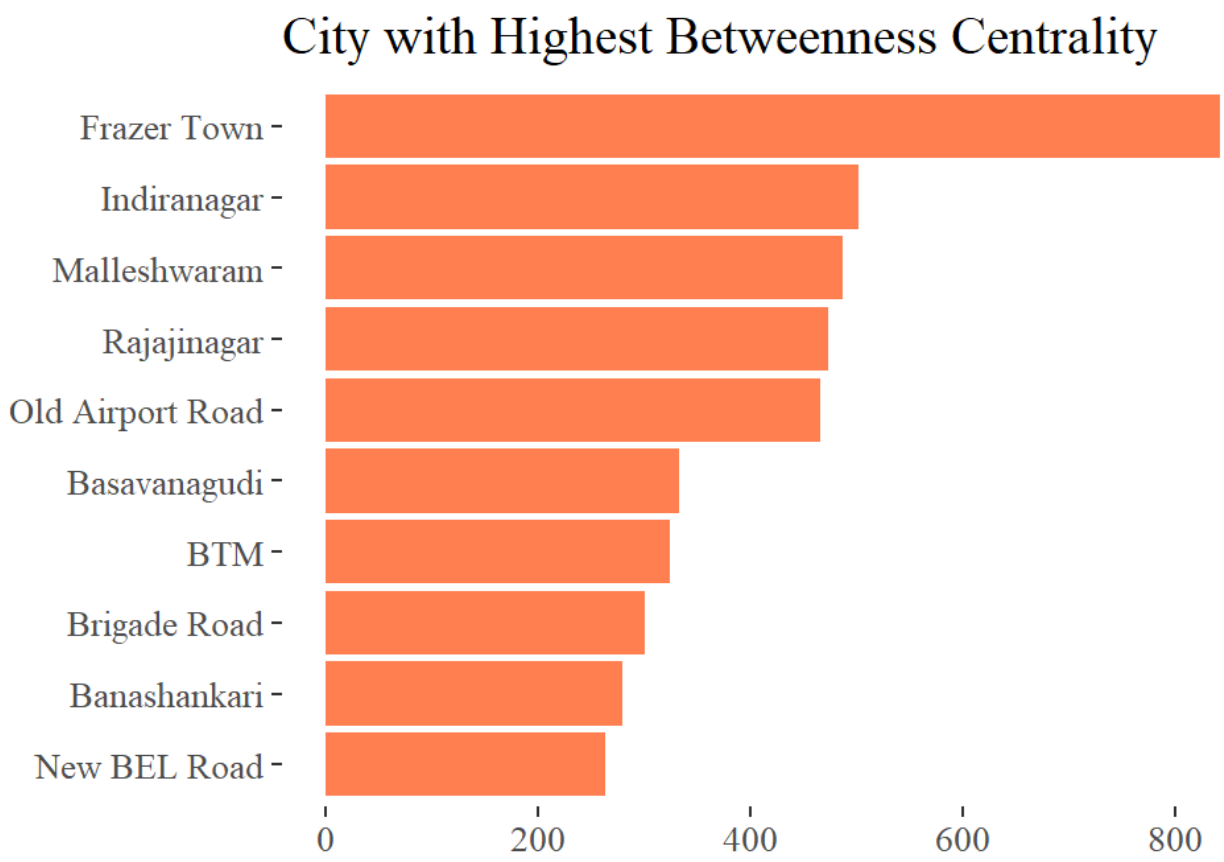

Figure 21 Shows Betweenness Centrality in Descending Order

Various representations of centrality are shown in figures 19, 20, and 21. It is clear from examining this particular figure 2 that Brigade Road, Residency Road, and FrazerTown are the three locations that rank highest in terms of their degree of centrality. It indicates the highest possible number of edges that it possesses, and the node for the locations is in the centre. If the degree is higher, then the node must be closer to the centre.

It is clear from looking figure 20 that the locations of Frazer Town, Brigade Road, and Indira Nagar have the highest closeness centrality. High-scoring nodes are the ones that have the shortest pathways to each other node in the network. As per Figure 21, Frazer Town, Indira Nagar, and Malleshwaram were the top three locations with high betweenness centrality. High

betweenness means that the nodes are likely to know about happenings in a wide variety of location groups. It was generally agreed that Frazer Town was considered to be a top in every centrality aspect.

Correlation Between Locations Average Rating and Centrality		
Degree	Closeness	Betweenness
0.312684	0.308622	0.1444511

Table 8 shows correlation between centrality and ratings

From the above table it can be noted that Degree Centrality has the highest correlation between location and restaurant rating. And a positive correlation between location centrality measures and ratings could be observed.

5. DISCUSSION

The primary objective of this research is to determine whether or not the geographical location of a restaurant in Bangalore plays a role in the level of satisfaction experienced by its customers, as well as to investigate other elements that might contribute to higher customer ratings for various types of restaurants. It is essential to have a complete understanding of the perceived value of the distance from the city's core as well as the geographic qualities, as well as how these factors eventually influence the level of customer satisfaction.

According to the findings of machine learning models, location is one of the most important factors in determining how customers rate a restaurant and their overall level of satisfaction. It is possible to use Walter Christaller's (1933) Central Place Theory to analyse the fact that if a restaurant is located in an area that is commercially active and can be reached with relative ease, then there is a greater likelihood

that the maximum number of customers will favour the restaurant, and the level of customer satisfaction may also improve. According to the findings of the model, the level of customer satisfaction steadily drops in direct proportion to the higher distance from the city centre or heart of the city. It is also possible to deduce that the majority of restaurants with high ratings and a high proportion of satisfied customers are located within a short distance of the city Centre. The findings offered by (Yang, et al., 2017; Prayag, et al., 2010) also denotes that location plays a key conscious choice for restaurants.

Mostly the nearby neighbourhoods to the city centre area are well connected with public transport in Bangalore, every customer might not have a car and customers can commute easily with public transport. Bangalore city is also known for the time-consuming traffic all over India, so the general public does not prefer to travel by car (Bardia, 2018). The marketing power of location can also play a significance to attract more customers to a restaurant. If the location is outskirt which is above 8 miles and having less frequency of public transport may impact new customers trying to explore new restaurants. The only advantage of a restaurant being far away from the city centre is rent might be a bit cheaper.

Frazer Town was considered to be in the best place when it came to the centrality analysis of locations because it had practically all of the centrality attributes. It was the most important quality in terms of both the closeness and the betweenness centralities. It is possible that having a restaurant open in this location will affect both the number of customers and their ratings. A link of around 30 percent was found to exist between the average ratings based on a specific location and centrality measures. It demonstrated that there is a favourable connection between the average rating and the metrics of centrality.

Location Koramangala was the one with the highest number of restaurants even though it was a residential area. It had the highest percentage of positive. Koramangala 4th, 5th, 6th and 7th block would be considered the best place to invest as around 50% of the restaurants present there have ratings above 4 out of 5. This location was the the one to attract many customers.

Intuitively, one would anticipate paying a greater price at a restaurant with a better rating. It was shown that the rating had a significant positive link with the approximate cost of two persons. Although earlier researchers have observed that customers give higher ratings to establishments that charge higher prices (Priya, 2020). One possible explanation for the high cost of dining at a restaurant is that more money is spent on providing a pleasant ambience and highly trained staff. If the employees are competent and have had adequate training, then the customers will unquestionably receive quality service. In addition, a higher level of service quality may result in the loyalty of some customers (Andaleeb & Conway, 2006; Fornell, et al., 1996; Biswas & Verma, 2022; Othman & Harun, 2021). Additionally, if a client is a repeat customer, they are likely satisfied with the service they have received, which makes it possible to anticipate a high level of positive customer satisfaction.

According to the findings of another piece of research, the factor known as "Restaurant Type: Dessert Parlour, Casual Dining and Quick Bites" are the one that has the most impact on the ratings. Bangalore is the world's fastest-growing IT hub (The Times of India, 2021), so many people who work in corporations may not have time to cook their meals. They spend a lot of time at work, they don't have time to make fancy meals at home anymore. Quick Service Restaurants are a more convenient option. They now spend their extra money on going out to eat (Krishna, 2014). However, (Bhatia & Sneha, 2021) reports that the majority of individuals in Bangalore like ordering food to be delivered to their homes or dining out rather than selecting any of the other dining restaurant type options available at restaurants.

When measuring the level of customer happiness, the number of votes cast is another crucial consideration to take into account. Customers might favour a restaurant more if it has received the most votes.

Therefore, if the restaurant has received a greater number of votes, there is a good possibility that it will attract new customers.

Therefore, rather than location as an important factor, other attributes like ratings, votes, and prices all play an important role in the development of a restaurant in a way that is more beneficial. The vast population are cash-strapped, so they have no choice except to search for inexpensive dining options. To be considered a budget restaurant, prices should be kept to a minimum or less than 325 INR, and the establishment should have ratings that are higher than four stars. Because of this, it is possible to classify such a place as a reasonably priced restaurant.

6. CONCLUSION

The study's overarching purpose was to determine if there is a correlation between a restaurant's location and the quality of service it offers its customers. Several machine learning classification methods agree that location and distance are crucial factors for customers' satisfaction. As per the centrality Analysis, Frazer Town which is a suburb of Bangalore is considered the best location to start a restaurant business it is not a high-activity area but is 1.6 miles away from the city centre. The restaurant would benefit from being situated closer to the centre of the city to attract a larger customer base. And another important location which was identified was Koramangala which is around 3 miles from the city centre. The location of a new restaurant should be carefully considered for a better customer satisfaction rate, although it seems likely that a lively area close to the city centre would be the best option. It must also be noted that famous places where the crowd gathers like shopping malls, and movie theatres, which increase the area's appeal to customers can also be considered (Prayag, et al., 2010). The fast food chains can also be targeted in a location with their competitors. As suggested by (Widaningrum, et al., 2018) that fast food chains located near the city centre with their competitors have a high chance of getting success. The customers who might visit a location regularly might try some new restaurant.

Although other factors like menu prices must also be kept in consideration, customers might not visit a place if the prices are too high and the customer satisfaction is not up to the mark. The average price for 2

people is recommended not more than 400 INR, if the restaurant is new and if it wants to attract more customers. But as per research by (Priya, 2020), suggests that if the rating of a restaurant is high then it's alright to have premium rates.

The number of people voting and the review count of the restaurant was also important factor to be considered. Even if the reviews are positive, the customer also checks for the total vote count of how many people have voted (Lee, et al., 2020). If a customer is visiting a newly opened restaurant then the restaurant manager can request the customer to give ratings or vote after the customer has their food. As the vote gets increased then there is a high probability of new customers willing to try the food in the restaurant.

Before opening any restaurant it must also be decided what type of restaurant needs to be opened. And what type of ambience is required. If the area has a high-class crowd, then it would be advised to have a restaurant with better ambience and much better service because if only the ambience is better but no positive customer service then the money spent on ambience would be of no use (Voon, 2017).

6.1. Limitation and future Work

However, only full-service restaurants in Bangalore, India, were included in the analysis. As a result, it's possible that the findings can only be applied to Bangalore restaurants and not to any others. The co-ordinates were identified on the basis of city location, co-ordinates could also be found out with respect to restaurant address for a better results and better understanding for the factor distance. In this research the dataset was been analyzed using 2 classes – Positive or Negative customer satisfaction rate, but it would have been more insightful if it consisted of 3 classes – Positive, Neutral and Negative. If the rating of a restaurant is 3.9 out of 5 then it would be considered as a negative customer satisfaction in this research paper.

This particular data set had ratings based on the customers feedback, but it might be a case that customer service at start might not be that good but have a majority of feedbacks from past. But recently if the restaurant service and staff has change and the quality has increased tremendously, then the majority of the negative feedback won't make much sense. This particular situation might not arise always but in few cases. To make the results more reliable, future studies should look at recent customer feedback from restaurants to analyze it in an appropriate way.

Restaurant reviews were not been analyzed, analyzing the reviews in the future can help to understand better results for individual restaurants whether the customer gave a rating of 3 out 5 but the customer is satisfied or not.

7. References

ACCA, 2022. *Decision trees.* [Online]

Available at: https://www.accaglobal.com/pk/en/student/exam-support-resources/fundamentals-exams-study-resources/f5/technical-articles/decision-trees.html

[Accessed 6 September 2022].

Altawee, M., 2021. *Central Place Theory.* [Online]

Available at: https://www.geographyrealm.com/central-place-theory

[Accessed 20 August 2022].

Andaleeb, S. S. & Conway, C., 2006. Customer satisfaction in the restaurant industry: an examination of the transaction-specific model. *Journal of Services Marketing,* pp. 3-11.

Bardia, J., 2018. *Car sales dip in city as more opt for public transport.* [Online]

Available at: https://www.newindianexpress.com/cities/bengaluru/2018/apr/30/car-sales-dip-in-city-as-more-opt-for-public-transport-1808144.html

[Accessed 2 September 2022].

Berkeley.edu, 2022. *Random Forests.* [Online]

Available at: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[Accessed 7 September 2022].

Bhatia, A. & Sneha, S., 2021. Location Based Restaurant Preferences in Bangalore. *Proceedings of the International Conference on Innovative Computing & Communication.*

Bhatia, D. A. & Sneha, M., 2021. LOCATION BASED RESTAURANT PREFERENCES IN BANGALORE. *SSRN.*

Biswas, A. & Verma, R. K., 2022. Augmenting service quality dimensions: mediation of image in the Indian restaurant industry. *JOURNAL OF FOODSERVICE BUSINESS RESEARCH .*

Breiman, L., 2001. Random Forests. *Machine Learning,* pp. 5-32.

Brownlee, J., 2020. *Extreme Gradient Boosting (XGBoost) Ensemble in Python.* [Online]
Available at: https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/#:~:text=Number%20of%20Features-,Extreme%20Gradient%20Boosting%20Algorithm,constructed%20from%20decision%20tree%20models.
[Accessed 30 August 2022].

Canny, I. U., 2014. Measuring the Mediating Role of Dining Experience Measuring the Mediating Role of Dining Experience Behavioral Intentions of Casual Dining Restaurant in Jakarta. *International Journal of Innovation, Management and Technology,* V(1), pp. 25-29.

CFI, 2022. *Bayes' Theorem.* [Online]
Available at: https://corporatefinanceinstitute.com/resources/knowledge/other/bayes-theorem/
[Accessed 29 August 2022].

Chapman, P. et al., 2000. *CRISP-DM 1.0: Step-by-step data mining guide.* [Online]
Available at: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

Chauhan, N. S., 2020. *Naïve Bayes Algorithm: Everything You Need to Know.* [Online]
Available at: https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html
[Accessed 29 August 2022].

Chauhan, N. S., 2022. *Decision Tree Algorithm, Explained.* [Online]
Available at: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
[Accessed 6 September 2022].

Cheng, J.-S., H.-Y. S. & Wu, M.-H., 2016. Ambience and Customer Loyalty of the Sport-themed Restaurant. *Universal Journal of Management,* pp. 444-450.

Chen, L.-F. & Tsai, C.-T., 2016. Data mining framework based on rough set theory to improve location selection decisions: A case study of a restaurant chain. *Tourism Management,* pp. 197-206.

Chun, S.-H. & Nyam-Ochir, A., 2020. The Effects of Fast Food Restaurant Attributes on Customer Satisfaction, Revisit Intention, and Recommendation Using DINESERV Scale. *MDPI,* 12(18).

Disney, A., 2020. *Social network analysis 101: centrality measures explained.* [Online]
Available at: https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/
[Accessed 29 August 2022].

Dobilas, S., 2021. *XGBoost: Extreme Gradient Boosting — How to Improve on Regular Gradient Boosting?.* [Online]
Available at: https://towardsdatascience.com/xgboost-extreme-gradient-boosting-how-to-improve-on-regular-gradient-boosting-5c6acf66c70a
[Accessed 30 August 2022].

Donga, L., Ratti, C. & Zheng, S., 2019. Predicting neighborhoods' socioeconomic attributes using restaurant data. *PNAS,* 116(31), p. 15447–15452.

Financial Times, 2022. *Bangalore keeps its crown as India's high-growth tech hub.* [Online]
Available at: https://www.ft.com/content/022aa805-3699-4bac-a845-81c95d015bc2
[Accessed 4 September 2022].

Fornell, C. et al., 1996. The American Customer Satisfaction Index: Nature, Purpose, and Findings. *Journal of Marketing,* 60(4), pp. 7-18.

Golbeck, J., 2013. Chapter 3 - Network Structure and Measures. In: *Analyzing the Social Web.* s.l.:s.n., pp. 25-44.

Golbeck, J., 2015. Chapter 21 - Analyzing networks. In: *Introduction to Social Media Investigation.* s.l.:s.n., pp. 221-235.

Gupta, R. et al., 2021. Sentiment Analysis on Zomato Reviews. *IEEE,* pp. 34-38.

Haghighi, M., Dorosti, A., Rahnama, A. & Hoseinpour, A., 2012. Evaluation of factors affecting customer loyalty in the restaurant industry. *African Journal of Business Management,* VI(14), pp. 5039-5046.

Hanaysha, J., 2016. Restaurant Location and Price Fairness as Key Determinants of Brand Equity: A Study on Fast Food Restaurant Industry. *Business and Economic Research,* 6(1).

Harris, L., 2017. *Farm to Restaurant: Exploring the Availability of Locally Grown Food and Obstacles to Its Use in Seacoast New Hampshire Restaurants.* [Online]
Available at: https://www.unh.edu/inquiryjournal/spring-2017/farm-restaurant-exploring-availability-locally-grown-food-and-obstacles-its-use-seacoast
[Accessed 3 September` 2022].

Hlee, S., Lee, J., Yang, S.-B. & Koo, C., 2019. The moderating effect of restaurant type on hedonic versus utilitarian review evaluations. *International Journal of Hospitality Management,* pp. 195-206.

IBM, 2021. *CRISP-DM Help Overview.* [Online]
Available at: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview
[Accessed 26 August 2022].

IBM, 2022. *What is logistic regression?.* [Online]
Available at: https://www.ibm.com/uk-en/topics/logistic-regression
[Accessed 29 August 2022].

Interview Area, 2022. *Which is the heart of Bangalore city?.* [Online]

Available at: https://www.interviewarea.com/faq/which-is-the-heart-of-bangalore-city

[Accessed 27 August 2022].

J.DONOVAN, R., ROSSITER, J. R., ROSSITER, J. R. & ROSSITER, J. R., 1994. Store Atmosphere and Purchasing Behavior. *Journal of Retailing,* 70(3), pp. 283-294.

javaTpoint, 2022. *Naïve Bayes Classifier Algorithm.* [Online]

Available at: https://www.javatpoint.com/machine-learning-naive-bayes-classifier#:~:text=Na%C3%AFve%20Bayes%20Classifier%20is%20one,the%20probability%20of%20an%20object.

[Accessed 29 August 2022].

Kim, J. et al., 2022. Why am I satisfied? See my reviews – Price and location matter in the restaurant industry. *International Journal of Hospitality Management,* Volume 101.

Kothari, K. & Shah, A., 2018. ZOMATO REVIEW ANALYSIS USING TEXT MINING. *IJCIRAS,* 1(1).

Kotler, P., 1974. Atmospherics as a Marketing Tool. *Journal of Retailing,* 49(4).

Krishna, K., 2014. Analysing Competition in the Quick Service Restaurant Industry. *SSRN.*

Ladhari, R., Brun, I. & Morales, M., 2008. Determinants of dining satisfaction and post-dining behavioral intentions. *International Journal of Hospitality Management,* pp. 563-573.

Lee, M., Kwon, W. & Back, K.-J., 2020. Artificial intelligence for hospitality big data analytics: developing a prediction model of restaurant review helpfulness for customer decision-making. *International Journal of Contemporary Hospitality Management,* 33(6).

Liljander, V. & Strandvik, T., 1997. Emotions in service satisfaction. *International Journal of Service Industry Management,* pp. 148-169.

Love, D. H. G., 1972. Fast Food Store Location Factors: A Comparison with Grocery Store Location Factors. *Journal of Food Distribution Research,* pp. 40-43.

LucidChart, 2022. *What are your decision tree needs?.* [Online]
Available at: https://www.lucidchart.com/pages/decision-tree
[Accessed 30 August 2022].

McDonough, T., 2022. *£3m Liverpool bar/restaurant to create 200 jobs.* [Online]
Available at: https://lbndaily.co.uk/3m-liverpool-bar-restaurant-to-create-200-jobs/
[Accessed 3 September 2022].

Molnar, C., 2022. *Logistic Regression.* [Online]
Available at: https://christophm.github.io/interpretable-ml-book/logistic.html
[Accessed 29 August 2022].

Mondurailingam, M. & Subramani, V. J. A. K., 2015. COMPARATIVE STUDY ON CUSTOMER SATISFACTION TOWARDS KFC AND MCDONALDS, CHENNAI. *ZENITH International Journal of Multidisciplinary Research,* 5(6).

Movable-type, 2022. *Calculate distance, bearing and more between Latitude/Longitude points.* [Online]
Available at: https://www.movable-type.co.uk/scripts/latlong.html
[Accessed 27 August 2022].

NEO4J, 2022. *Betweenness Centrality.* [Online]
Available at: https://neo4j.com/docs/graph-data-science/current/algorithms/betweenness-

centrality/#:~:text=Betweenness%20centrality%20is%20a%20way,of%20nodes%20in%20a%20graph.

[Accessed 30 August 2022].

NEO4J, 2022. *Closeness Centrality.* [Online]

Available at: https://neo4j.com/docs/graph-data-science/current/algorithms/closeness-centrality/#:~:text=Closeness%20centrality%20is%20a%20way,distances%20to%20all%20other%20nodes.

[Accessed 30 August 2022].

NVIDIA, 2022. *What is Logistic Regression?.* [Online]

Available at: https://www.nvidia.com/en-us/glossary/data-science/linear-regression-logistic-regression/

[Accessed 6 September 2022].

Omar, M. S., Ariffin, H. F. & Ahmad, R., 2015. The Relationship between Restaurant Ambience and Customers' Satisfaction in Shah Alam Arabic Restaurants, Selangor. *International Journal of Administration and Governance,* pp. 1-8.

Oracle, 2022. *What is customer loyalty?.* [Online]

Available at: https://www.oracle.com/uk/cx/marketing/customer-loyalty/what-is-customer-loyalty/

[Accessed 20 August 2022].

Othman, B. & Harun, A. B., 2021. The Influence of Service Marketing Mix and Umrah Service Quality on Customer Satisfaction and Customer Loyalty towards Umrah Travel Agents in Malaysia. *Technium Social Sciences Journal,* Volume 22, pp. 553-618.

Prayag, G., Landre, M. & Ryan, C., 2010. Restaurant location in Hamilton, New Zealand: clustering patterns from 1996 to 2008. *International Journal of Contemporary Hospitality Management,* 24(3), pp. 430-450.

Priya, J., 2020. Predicting Restaurant Rating using Machine Learning and comparison of Regression Models. *IEEE.*

Raman, P., 2018. Zomato: a shining armour in the foodtech sector. *Journal of Information Technology Case and Application Research,* pp. 130-150.

Rekhit Pachanekar, 2022. *Naive Bayes Model: Introduction, Calculation, Strategy, Python Code.* [Online] Available at: https://blog.quantinsti.com/naive-bayes/
[Accessed 7 September 2022].

Ryu, K. & Jang, S. (., 2007. THE EFFECT OF ENVIRONMENTAL PERCEPTIONS ON BEHAVIORAL INTENTIONS THROUGH EMOTIONS: THE CASE OF UPSCALE RESTAURANTS. *JOURNAL OF HOSPITALITY & TOURISM RESEARCH,* pp. 56-72.

SCHOCH, D., 2018. *http://blog.schochastics.net/post/network-centrality-in-r-introduction/.* [Online] Available at: http://blog.schochastics.net/post/network-centrality-in-r-introduction/
[Accessed 29 August 2022].

Sedov, D., 2022. Restaurant closures during the COVID-19 pandemic: A descriptive analysis. *Economics Letters,* Volume 213.

Shetty, D. & S, P. J., 2020. A Study on Impact of Covid-19 on Buying Behaviour of Consumer on Online Food Delivery with Reference to Zomato. *International E Conference on Adapting to the New Business Normal .*

Speise, J. L., M. E. M., Tooze, J. & Ip, E., 2019. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems With Applications,* pp. 93-101.

Statista, 2022. *Market value of the food service industry in India in 2014 and 2019, with an estimate for 2025.* [Online]

Available at: https://www.statista.com/statistics/1299232/india-food-service-market-size/

[Accessed 4 September 2022].

Statistic Times, 2021. *Population of Cities in India.* [Online]

Available at: https://statisticstimes.com/demographics/country/india-cities-population.php

[Accessed 4 September 2022].

Stroebele, N. & Castro, J. M. D., 2004. Effect of Ambience on Food Intake and Food Choice. *Nutrition,* 20(9).

Tat, M. J., 2017. *Seeing the random forest from the decision trees: An explanation of Random Forest.* [Online]

Available at: https://towardsdatascience.com/seeing-the-random-forest-from-the-decision-trees-an-intuitive-explanation-of-random-forest-beaa2d6a0d80#:~:text=Random%20Forests%20allow%20us%20to,100%20being%20the%20most%20important.

[Accessed 30 August 2022].

The Atlantic, 2017. *Restaurants Are the New Factories.* [Online]

Available at: https://www.theatlantic.com/business/archive/2017/08/restaurant-jobs-boom/536244/

[Accessed 4 September 2022].

The Times of India, 2021. *Bengaluru world's fastest growing tech hub, London second: R ...* [Online]

Available at: https://timesofindia.indiatimes.com/business/india-business/bengaluru-worlds-fastest-growing-tech-hub-london-second-report/articleshow/80262770.cms

[Accessed 3 September 2022].

Vajjhala, V. & Ghosh, M., 2021. DECODING THE EFFECT OF RESTAURANT REVIEWS DECODING THE EFFECT OF RESTAURANT REVIEWS. *Journal of Foodservice Business Research,* 25(5), pp. 535-560.

Voon, B. H., 2017. Service Environment of Restaurants: Findings from the youth customers. *Journal of ASIAN Behavioural Studies,* pp. 67-77.

Wade, D., 2006. Successful restaurant management: From vision to execution. *Journal of Vacation Marketing,* 12(4).

Widaningrum, D. L., Surjandari, I. & Arymurthy, A. M., 2018. Visualization of Fast Food Restaurant Location using Geographical Information System. *IOP Conference Series: Earth and Environmental Science,* 145(012102).

Wirth, R. & Hipp, J., 2000. CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining,* Volume 1, pp. 29-39.

Yang, Y., Roehl, W. S. & Huang, J.-H., 2017. Understanding and projecting the restaurantscape: The influence of neighborhood sociodemographic characteristics on restaurant location. *International Journal of Hospitality Management,* pp. 33-45.

Yeturu, K., 2020. Chapter 3 - Machine learning algorithms, applications, and practices in data science. In: *Handbook of Statistics.* s.l.:s.n., pp. 81-206.

Zomato, 2022. *Who are we?.* [Online]
Available at: https://www.zomato.com/who-we-are
[Accessed 4 September 2022].

Zou, X., Zou, X., Tian, Z. & Shen, K., 2019. Logistic Regression Model Optimization and Case Analysis. *IEEE,* pp. 135-139.

8. Appendix 1: R-Code/SQL

```
#SQL CODE USED TO CONNECT 2 TABLES
SELECT Zomato.ID, Zomato.address, Zomato.name, Zomato.online_order, Zomato.book_table,
Zomato.Rate, Zomato.votes, Zomato.rest_type, Zomato.dish_liked, Zomato.cuisines,
Zomato.approx_costfortwopeople, Zomato.location, Zomato.listed_incity, City.[Happening Place?],
City.Longitude, City.Latitude, City.[Distance from Heart of Bengaluru (Miles)],
Zomato.Satisfaction_Rate, * FROM Zomato INNER JOIN City ON Zomato.listed_incity =
City.Location;

#SQL CODE USED TO GET AVERAGE RATING OF CITY
SELECT Zomato.location, Round(Avg([Zomato].[Rate]),2) AS Average_Rating FROM Zomato GROUP
BY Zomato.location;

#SQL Code used to Connect Centrality and City Ratings
SELECT Centrality_imp.City, Centrality_imp.degree, Centrality_imp.closeness,
Centrality_imp.betweenness, Centrality_imp.eigenvector, Avg_Rating_City.Average_Rating
FROM Centrality_imp INNER JOIN Avg_Rating_City ON Centrality_imp.City =
Avg_Rating_City.location;
```

R- Code Used:

```
#::::: LOAD THE DIRECTORY ::::::

setwd('D:/Dissertation')


#::::: LOAD ESSENTIAL LIBRARIES ::::::

library(writexl)

library(tidyverse)

library(readr)

library(ggplot2)

library(dplyr)

library(caret)

library(logistf)

library(glmnet)

library(plotly)


#:::: READ THE ZOMATO FILE CONTAINING ALL THE ATTRIBUTES ::::::

#Read the CSV file and replace empty cells by NA

zomato <- read.csv('zomato.csv',na.strings = c("", "NA"))

#Check NA values

sum(is.na(zomato))

colSums(is.na(zomato))


#check the blank values

sapply(zomato,function(x) sum(x==""))
```

```
#Drop Unnecesary Columns
zomato <- subset(zomato, select = -c(url,phone,reviews_list,menu_item) )

#Remove Duplicate Rows
zomato <- distinct(zomato)

#Remove '/5' extra string from the rating column. so, that
zomato$rate <- gsub("/5","",as.character(zomato$rate))

#Remove NAs from Zomato
zomato <- drop_na(zomato)

#Check if Duplicates are Present
sum(duplicated(zomato))

#Check if NA values from the dataset
sum(is.na(zomato))

#Remove NAs from Zomato
zomato <- drop_na(zomato)

#Check if Duplicates are Present
sum(duplicated(zomato))

#Check if NA values from the dataset
sum(is.na(zomato))
```

```
#Write the updated Excel File for combining 2 files in MS-ACCESS USING SQL QUERIES
write_xlsx(zomato,"UPDATED_ZOMATO.xlsx")

#:::::::AFTER COMBINING THE 2 TABLES in MS-ACCESS WITH NEW VARIABLES
Distance from City Centre and High Activity PLACE :::::::
#Read the FINAL CSV file and replace empty cells by NA
zomato <- read.csv('zom.csv',na.strings = c("", "NA"))

#Drop Unnecesary Columns
zomato <- subset(zomato, select = -c(dish_liked, cuisines, listed_in.type.  ) )


#::::::: DATA CLEANING::::::::::::

#Remove '/5' extra string from the rating column. so, that
zomato$rate <- gsub("/5","",as.character(zomato$rate))

#Create a New Variable Satisfaction_Level - Positive or Negative
#Previously it was tested with rate >=3 but rate>=4 are giving better results
zomato <- zomato %>%
  mutate(Satisfaction_Rate = case_when(
    rate >= 4 ~ "Positive",
    rate < 4 ~ "Negative"
  ))
```

```
#Convert the Required Columns into Factors
zomato$name <- as.factor(zomato$name)
zomato$online_order <- as.factor(zomato$online_order)
zomato$book_table <- as.factor(zomato$book_table)
zomato$location <- as.factor(zomato$location)
zomato$rest_type <- as.factor(zomato$rest_type)
zomato$listed_in.city. <- as.factor(zomato$listed_in.city.)
zomato$location2 <- as.factor(zomato$location2)
zomato$Satisfaction_Rate  <- as.factor(zomato$Satisfaction_Rate)
zomato$happening_place <- as.factor(zomato$happening_place)

#Summary of New created Variable
summary(zomato$Satisfaction_Rate)

#Convert into number
zomato$rate <- as.numeric(zomato$rate)
zomato$approx_cost.for.two.people. <- as.integer(zomato$approx_cost.for.two.people.)

#Check and Reorder Levels
levels(zomato$Satisfaction_Rate)

#Rename the levels for Restaurant Type
zomato$rest_type[zomato$rest_type == 'Bakery, Cafe'] <- 'Bakery'
zomato$rest_type[zomato$rest_type == 'Bakery, Dessert Parlor'] <- 'Bakery'
zomato$rest_type[zomato$rest_type == 'Bakery, Quick Bites'] <- 'Bakery'
zomato$rest_type[zomato$rest_type == 'Bar, Casual Dining'] <- 'Bar'
```

```
zomato$rest_type[zomato$rest_type == 'Bar, Pub'] <- 'Bar'
zomato$rest_type[zomato$rest_type == 'Pub'] <- 'Bar'
zomato$rest_type[zomato$rest_type == 'Beverage Shop, Cafe'] <- 'Beverage Shop'
zomato$rest_type[zomato$rest_type == 'Beverage Shop, Dessert Parlor'] <- 'Beverage Shop'
zomato$rest_type[zomato$rest_type == 'Beverage Shop, Quick Bites'] <- 'Beverage Shop'
zomato$rest_type[zomato$rest_type == 'Cafe, Bakery'] <- 'Cafe'
zomato$rest_type[zomato$rest_type == 'Cafe, Dessert Parlor'] <- 'Cafe'
zomato$rest_type[zomato$rest_type == 'Cafe, Quick Bites'] <- 'Cafe'
zomato$rest_type[zomato$rest_type == 'Cafe, Casual Dining'] <- 'Casual Dining'
zomato$rest_type[zomato$rest_type == 'Casual Dining, Bar'] <- 'Casual Dining'
zomato$rest_type[zomato$rest_type == 'Casual Dining, Pub'] <- 'Casual Dining'
zomato$rest_type[zomato$rest_type == 'Casual Dining, Cafe'] <- 'Casual Dining'
zomato$rest_type[zomato$rest_type == 'Casual Dining, Lounge'] <- 'Casual Dining'
zomato$rest_type[zomato$rest_type == 'Food Court, Casual Dining'] <- 'Casual Dining'
zomato$rest_type[zomato$rest_type == 'Dessert Parlor, Bakery'] <- 'Dessert Parlor'
zomato$rest_type[zomato$rest_type == 'Dessert Parlor, Beverage Shop'] <- 'Dessert Parlor'
zomato$rest_type[zomato$rest_type == 'Dessert Parlor, Cafe'] <- 'Dessert Parlor'
zomato$rest_type[zomato$rest_type == 'Dessert Parlor, Kiosk'] <- 'Dessert Parlor'
zomato$rest_type[zomato$rest_type == 'Dessert Parlor, Quick Bites'] <- 'Dessert Parlor'
zomato$rest_type[zomato$rest_type == 'Food Court, Quick Bites'] <- 'Food Court'
zomato$rest_type[zomato$rest_type == 'Quick Bites, Cafe'] <- 'Quick Bites'
zomato$rest_type[zomato$rest_type == 'Quick Bites, Dessert Parlor'] <- 'Quick Bites'
zomato$rest_type[zomato$rest_type == 'Quick Bites, Food Court'] <- 'Quick Bites'
zomato$rest_type[zomato$rest_type == 'Quick Bites, Bakery'] <- 'Quick Bites'
zomato$rest_type[zomato$rest_type == 'Quick Bites, Beverage Shop'] <- 'Quick Bites'
zomato$rest_type[zomato$rest_type == 'Sweet Shop, Quick Bites'] <- 'Sweet Shop'
```

```
zomato$rest_type[zomato$rest_type == 'Quick Bites, Sweet Shop'] <- 'Sweet Shop'
zomato$rest_type[zomato$rest_type == 'Takeaway'] <- 'Takeaway'
zomato$rest_type[zomato$rest_type == 'Takeaway, Delivery'] <- 'Takeaway'

#Remove Restaurant Type with Dhaba, Food Truck, Mess and Kiosk as there are very few restaurants for the same
zomato <- zomato[!(zomato$rest_type == "Food Truck" | zomato$rest_type == "Kiosk" | zomato$rest_type == "Mess" | zomato$rest_type == "Dhaba"),]

#drop unused factor levels for Rest Type
zomato$rest_type <- droplevels(zomato$rest_type)


#:::::::MEASURES OF ASSOSCIATION::::::::

#Chi Square Tests in R for Diiferent Variables with Respect to satisfaction Rate or RATE
#chisq.test() function to perform the test

# With respect to City
table(zomato$Satisfaction_Rate,zomato$listed_in.city.)
chisq.test(zomato$Satisfaction_Rate,zomato$listed_in.city., correct = FALSE)

# With respect to Online Order
table(zomato$Satisfaction_Rate,zomato$online_order)
chisq.test(zomato$Satisfaction_Rate, zomato$online_order, correct = FALSE)
```

```
#WIth respect to book_table
table(zomato$Satisfaction_Rate,zomato$book_table)
chisq.test(zomato$Satisfaction_Rate,zomato$book_table, correct = FALSE)



#::::::::::VISUALISATION IN GGPLOTS::::::::::::

#1. Top Restaurants Food Chain in Bangalore
top_restaurant <- zomato %>% select(name) %>% group_by(name) %>% count() %>%
arrange(desc(n))
top_restaurant <- top_restaurant[1:10,]

top_restaurant %>%
  ggplot(aes(x=reorder(name,n),y=n))+
  geom_bar(stat = "identity", color="black",fill="orange") +
  coord_flip() +
  labs(title = "Top 10 Food Chains in Bangalore", y="Total Count", x= "Food Chains")
scale_fill_brewer(palette = "Dark2")



#2. Top Locations in Bangalore with Positive Satisfaction Rate
top_location <- zomato %>% filter(Satisfaction_Rate == 'Positive') %>%  select(listed_in.city.)
%>% group_by(listed_in.city.) %>% count() %>% arrange(desc(n))
top_location <- top_location[1:15,]

top_location %>%
```

```
  ggplot(aes(x=reorder(listed_in.city.,n),y=n))+
  geom_bar(stat = "identity", color="black",fill="orange") +
  coord_flip() +
  labs(title = "Top 15 neighbourhoods with Positive Satisfaction", y="Total Count", x=
"Neighbourhoods")
scale_fill_brewer(palette = "Dark2")

#3. Restaurant Rating Distributions
ggplot(zomato) +
  geom_density(aes(x = rate), fill = '#FFD700') + #geom_histogram
  labs(
    title = 'Restaurant Rating Distributions',
    x = 'Ratings out of 5',
    y = 'Frequency'
  ) +
  ggthemes::theme_few()

#4. Online Order Count
ggplotly(zomato %>%
           group_by(online_order) %>%
           summarise(total = n()) %>%
           ggplot(aes(x=online_order,y=total,fill = online_order))+
           labs(
             title = 'Online Order Count',
             x = 'Online Order?',
             y = 'Total Count'
```

```
    )+
    geom_bar(stat = 'identity'))



#::::::: NETWORK CENTRALITY FOR LOCATION :::::::

#LOAD IMPORTANT LIBRARIES FOR CENTRALITY
library(ggraph)
library(igraph)
library(echarts4r)
library(ggthemes)

#load the location columns
NetworkEL_loc <- select(zomato,c(10,11))

#Class of NetworkEL_loc
class(NetworkEL_loc)

#Convert it into a Matrix format
Network_Matrix <- as.matrix(NetworkEL_loc)
class(Network_Matrix)

#g is the variable to which we are assigning the network
g <- graph_from_edgelist(Network_Matrix, directed=FALSE)

##if gD should be a directed network
```

```
gD <- graph_from_edgelist(Network_Matrix, directed=TRUE)

#simplify the network to remove duplicates.
g <- simplify(g)
g

gD

#Try this out with the two networks (g, and gD)
Degree <- degree(g)
Indegree.Undirected <- degree(g, mode="in")
Outdegree.Undirected <- degree(g, mode="out")

Degree.Directed <- degree(gD)
Indegree <- degree(gD, mode="in")
Outdegree <- degree(gD, mode="out")

#use the cbind command to combine the measures for comparison
CompareDegree <- cbind(Degree, Indegree.Undirected, Outdegree.Undirected, Degree.Directed,
Indegree, Outdegree)

#Eigenvector Centrality
Eig <- evcent(g)$vector

Hub <- hub.score(g)$vector
Authority <- authority.score(g)$vector
```

```
#Closeness Centrality
Closeness <- closeness(g)

# Reach at k=2
Reach_2 <- (ego_size(g, 2)-1)/(vcount(g)-1)

## Reach at k=3
Reach_3 <- (ego_size(g, 3)-1)/(vcount(g)-1)

#Betweenness Centrality
Betweenness <- betweenness(g)

centralities <- cbind(Degree, Eig, Hub, Authority, Closeness, Reach_2, Reach_3, Betweenness)

round(cor(centralities), 2)

V(g)$degree <- degree(g)                    # Degree centrality
V(g)$eig <- evcent(g)$vector                # Eigenvector centrality
V(g)$hubs <- hub.score(g)$vector            # "Hub" centrality
V(g)$authorities <- authority.score(g)$vector   # "Authority" centrality
V(g)$closeness <- closeness(g)              # Closeness centrality
V(g)$betweenness <- betweenness(g)          # Vertex betweenness centrality

centrality <- data.frame(row.names   = V(g)$name,
                         degree      = V(g)$degree,
```

```
                    closeness   = V(g)$closeness,
                    betweenness = V(g)$betweenness,
                    eigenvector = V(g)$eig)

centrality <- centrality[order(row.names(centrality)),]

head(centrality)

# Top ten
head(centrality[order(centrality$betweenness),], n=10)




lay <- layout_with_kk(g)

#Plot Layout
plot(g, layout = lay,
    vertex.label = NA)

#View the Attributes
View(centrality)

plot.igraph(gD, layout=lay,
          vertex.size=degree(gD, mode="in"),
          main="Indegree")
```

```
#Write the Centrality Measures in a CSV format
write.csv(centrality, file = "Centrality.csv")

#Modify the Centrality CSV file with average ratings of the location
centrality <- read.csv('centrality.csv')

# Highest Degree Centrality
centrality %>%
  arrange(desc(degree)) %>%
  slice(1:10)%>%
  ggplot(aes(x = fct_reorder(City,degree), y = degree))+
  geom_col(fill = '#e63946')+
  coord_flip() +
  labs(x = " ,y = ",title = 'City with Highest Degree Centrality') +
  theme_tufte() +
  theme(axis.text = element_text(size = 10),
        plot.title = element_text(size = 15))

#Cities with High Closeness Centrality
centrality %>%
  arrange(desc(closeness)) %>%
  slice(1:10)%>%
  ggplot(aes(x = fct_reorder(City,closeness), y = closeness))+
  geom_col(fill = '#e9c46a')+
  coord_flip() +
  labs(x = " ,y = ",title = 'Cities with Highest Closeness Centrality') +
```

```
  theme_tufte() +
  theme(axis.text = element_text(size = 10),
        plot.title = element_text(size = 15))



#Betweeness Centrality
centrality %>%
  arrange(desc(betweenness)) %>%
  slice(1:10)%>%
  ggplot(aes(x = fct_reorder(City,betweenness), y = betweenness))+
  geom_col(fill = '#2a9d8f')+
  coord_flip() +
  labs(x = " ,y = ",title = 'Location with high Betweeness Centrality') +
  theme_tufte() +
  theme(axis.text = element_text(size = 10),
        plot.title = element_text(size = 15))



centrality_rating <- read_excel('centrality_rating.xlsx')

# Correlation Between Centrality Measures
cor(centrality_rating$degree,centrality_rating$Average_Rating, method = c("pearson", "kendall",
"spearman"))
cor.test(centrality_rating$degree,centrality_rating$Average_Rating, method=c("pearson",
"kendall", "spearman"))
```

```
cor(centrality_rating$closeness,centrality_rating$Average_Rating, method = c("pearson",
"kendall", "spearman"))
cor.test(centrality_rating$closeness,centrality_rating$Average_Rating, method=c("pearson",
"kendall", "spearman"))

cor(centrality_rating$betweenness,centrality_rating$Average_Rating, method = c("pearson",
"kendall", "spearman"))
cor.test(centrality_rating$betweenness,centrality_rating$Average_Rating, method=c("pearson",
"kendall", "spearman"))

cor(centrality_rating$eigenvector,centrality_rating$Average_Rating, method = c("pearson",
"kendall", "spearman"))
cor.test(centrality_rating$eigenvector,centrality_rating$Average_Rating, method=c("pearson",
"kendall", "spearman"))



#::::: SPLIT THE ZOMATO DATA INTO TRAINING AND TEST::::
#to create a partition with 80%
set.seed(235) #generate a sequence of random numbers
index <- createDataPartition(zomato$Satisfaction_Rate, p = 0.8, list = FALSE,)
train <- zomato[index, ] #first 80% for training
test <- zomato[-index, ] #bottom 20% for testing

#Formula 1 consisting of listed_in.city
formula1 <- Satisfaction_Rate ~ listed_in.city. + rest_type + online_order + book_table + votes +
approx_cost.for.two.people.
```

```
#Formula consisting of happening_place + Distance_from_Heart
formula <- Satisfaction_Rate ~ happening_place + Distance_from_Heart  + rest_type +
online_order + book_table + votes + approx_cost.for.two.people.

#1. :::::::::::: Logistic Regression MODEL::::::::::

#::LR Model1 Consisting of Location Factors::

# Training the Logistic Regression model1
LRM1 <- glm(formula1, data = train, family = "binomial")
LRM1

#Summary of Logistic Regression MODEL
summary(LRM1)

# TRAIN DATA ACCURACY
# Predict test data based on model
LRM_predictions1 <- predict(LRM1,test,type ="response")

#Convert probabilities to Positive or Negative
LRM_class_pred1<-as.factor(ifelse(LRM_predictions1 > 0.5,"Positive","Negative"))
#evaluate the accuracy of the predictions
postResample(LRM_class_pred1,test$Satisfaction_Rate)
```

```
#Confusion Matrix
confmat_log <- table(actual_value = train$Satisfaction_Rate, Predicted_Value =
LRM_class_pred1 )
confmat_log



#Logistic Regression Model 2
#::LR Model Consisting of Distance and happening_place::

# Training the Logistic Regression model1
LRM <- glm(formula, data = train, family = "binomial")
LRM

#Summary of Logistic Regression MODEL
summary(LRM)

# TRAIN DATA ACCURACY
# Predict test data based on model
LRM_predictions <- predict(LRM,test,type ="response")

#Convert probabilities to Positive or Negative
LRM_class_pred<-as.factor(ifelse(LRM_predictions > 0.5,"Positive","Negative"))
#evaluate the accuracy of the predictions
postResample(LRM_class_pred,test$Satisfaction_Rate)

#Confusion Matrix
```

```
confmat_log <- table(actual_value = train$Satisfaction_Rate, Predicted_Value =
LRM_class_pred )
confmat_log

#Assessing Model R-Square
logisticPseudoR2s <- function(LogModel) {
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <- length(LogModel$fitted.values)
  R.l <- 1 - dev / nullDev
  R.cs <- 1- exp ( -(nullDev - dev) / modelN)
  R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemeshow R^2 ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2 ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2 ", round(R.n, 3), "\n")
}
logisticPseudoR2s(LRM)
#Odds Ratio (Exponential of coefficient)
exp(LRM$coefficients)

#confidence interval
exp(confint(LRM))

#::Logistic Model ASSUMPTIONS::
#Add the predicted probabilities to the data frame
```

```
train$predictedProbabilities <- fitted(LRM)

#This shows the probability of churn, and the actual outcome.
head(data.frame(train$predictedProbabilities, train$Satisfaction_Rate))

#Add the standardised and Studentised residuals can be added to the data frame
train$standardisedResiduals <- rstandard(LRM)
train$studentisedResiduals <- rstudent(LRM)

#count the residuals above 1.96
sum(train$standardisedResiduals > 1.96)

#COOKs Distance
train$cook <- cooks.distance(LRM)
sum(train$cook > 1)

train$leverage <- hatvalues(LRM)

#check if any values are above 0.0009
sum(train$leverage > 0.0009)

library(car)
#VIF to identify if there is a potential problem with multicolinearity
vif(LRM)
```

#2.::::::: Naives Bayes ::::::

```
library(naivebayes)
library(dplyr)
library(ggplot2)
library(psych)
set.seed(1234)

#model 1 without Distance Variable
model1 <- naive_bayes(formula1, data = train, usekernel = T)
plot(model1)
p <- predict(model1, train, type = 'prob')
#Confusion Matrix - train data
p1 <- predict(model1, train)
#Confusion Matrix
(tab1 <- table(p1, train$Satisfaction_Rate))
#Accuracy
sum(diag(tab1)) / sum(tab1)
#Confusion Matrix - test data
p2 <- predict(model1, test)
#Confusion Matrix
(tab2 <- table(p2, test$Satisfaction_Rate))
#Accuracy
(sum(diag(tab2))/sum(tab2))
```

```
#Model 2 With DIstance from Heart and happening place
formula <- Satisfaction_Rate ~ happening_place + Distance_from_Heart  + rest_type +
online_order + book_table + votes + approx_cost.for.two.people.

model <- naive_bayes(formula, data = train, usekernel = T)
model
plot(model)
p <- predict(model, train, type = 'prob')

#Confusion Matrix - train data
p1 <- predict(model, train)
#Confusion Matrix
(tab1 <- table(p1, train$Satisfaction_Rate))
#Accuracy
sum(diag(tab1)) / sum(tab1)
#Confusion Matrix - test data
p2 <- predict(model, test)
#Confusion Matrix
(tab2 <- table(p2, test$Satisfaction_Rate))
#Accuracy
(sum(diag(tab2))/sum(tab2))



#3.::::: Random Forest :::::
library(randomForest)
library(datasets)
```

```
library(caret)
set.seed(222)
#Model 1 with Location but without dist_from_heart and happening place
rf1 <- randomForest(formula1, data=train, proximity=TRUE)
rf1
print(rf1)
#Prediction & Confusion Matrix - train data
p1 <- predict(rf1, train)
confusionMatrix(p1, train$Satisfaction_Rate)
#Prediction & Confusion Matrix - test data
p2 <- predict(rf1, test)
confusionMatrix(p2, test$Satisfaction_Rate)
#Error Rate of RF
plot(rf1)

#No. of nodes for the trees
hist(treesize(rf1),
    main = "No. of Nodes for the Trees",
    col = "orange")
#Variable Importance
varImpPlot(rf1,
          sort = T,
          n.var = 10,
          main = "Imporrant Variables")
#MeanDecreaseGini
importance(rf1)
```

```
#Model 2 with  dist_from_heart and happening place


rf <- randomForest(formula, data=train, proximity=TRUE)
rf
print(rf)
#Prediction & Confusion Matrix - train data
p1 <- predict(rf, train)
confusionMatrix(p1, train$Satisfaction_Rate)
#Prediction & Confusion Matrix - test data
p2 <- predict(rf, test)
confusionMatrix(p2, test$Satisfaction_Rate)
#Error Rate of RF
plot(rf)


#No. of nodes for the trees
hist(treesize(rf),
    main = "No. of Nodes for the Trees",
    col = "orange")
#Variable Importance
varImpPlot(rf,
          sort = T,
          n.var = 10,
          main = "Imporrant Variables")
#MeanDecreaseGini
```

```
importance(rf)



# 4. Extreme gradient boost Tree
# Fit the model on the training set
library(dplyr)
library(caret)
library(xgboost)
library(e1071)

#Model 1 without happening_place and Heart_of City Paramert instead Listed City is Considered
set.seed(456)
model <- train(
  formula1, data = train, method = "xgbTree",
  trControl = trainControl("cv", number = 2)
)
# Best tuning parameter
model$bestTune



# Make predictions on the test data
predicted.classes <- model %>% predict(test)
head(predicted.classes)

# Compute model prediction accuracy rate
mean(predicted.classes == test$Satisfaction_Rate)
```

```
#The function varImp() [in caret] displays the importance of variables in percentage:
varImp(model)

p <- predict(model, train, type = 'prob')
#Confusion Matrix - train data
p1 <- predict(model, train)
#Confusion Matrix
(tab1 <- table(p1, train$Satisfaction_Rate))
#Accuracy
sum(diag(tab1)) / sum(tab1)
#Confusion Matrix - test data
p2 <- predict(model, test)
#Confusion Matrix
(tab2 <- table(p2, test$Satisfaction_Rate))
#Accuracy
(sum(diag(tab2))/sum(tab2))



#Model 2 without location
set.seed(456)
model <- train(
  formula, data = train, method = "xgbTree",
  trControl = trainControl("cv", number = 2)
)
# Best tuning parameter
```

```
model$bestTune



# Make predictions on the test data
predicted.classes <- model %>% predict(test)
head(predicted.classes)

# Compute model prediction accuracy rate
mean(predicted.classes == test$Satisfaction_Rate)

#The function varImp() [in caret] displays the importance of variables in percentage:
varImp(model)

p <- predict(model, train, type = 'prob')
#Confusion Matrix - train data
p1 <- predict(model, train)
#Confusion Matrix
(tab1 <- table(p1, train$Satisfaction_Rate))
#Accuracy
sum(diag(tab1)) / sum(tab1)
#Confusion Matrix - test data
p2 <- predict(model, test)
#Confusion Matrix
(tab2 <- table(p2, test$Satisfaction_Rate))
#Accuracy
(sum(diag(tab2))/sum(tab2))
```

#5. :::: Decision Tree::::::

```
#Model 1 with location variable
library(DAAG)
library(party)
library(rpart)
library(rpart.plot)
library(mlbench)
library(caret)
library(pROC)
library(tree)
tree <- rpart(formula1, data = train)
rpart.plot(tree)
printcp(tree)



p <- predict(tree, train, type = 'class')
confusionMatrix(p, train$Satisfaction_Rate, positive="Positive")

#ROC Curve
p1 <- predict(tree, test, type = 'prob')
p1 <- p1[,2]
r <- multiclass.roc(test$Satisfaction_Rate, p1, percent = TRUE)
roc <- r[['rocs']]
```

```
r1 <- roc[[1]]
plot.roc(r1,
        print.auc=TRUE,
        auc.polygon=TRUE,
        grid=c(0.1, 0.2),
        grid.col=c("green", "red"),
        max.auc.polygon=TRUE,
        auc.polygon.col="lightblue",
        print.thres=TRUE,
        main= 'ROC Curve - Model 1')



#Model 2 with distance from city center and High ACtivity Place
library(DAAG)
library(party)
library(rpart)
library(rpart.plot)
library(mlbench)
library(caret)
library(pROC)
library(tree)
tree <- rpart(formula, data = train)
rpart.plot(tree)
printcp(tree)
```

```
p <- predict(tree, train, type = 'class')

confusionMatrix(p, train$Satisfaction_Rate, positive="Positive")


#ROC Curve

p1 <- predict(tree, test, type = 'prob')

p1 <- p1[,2]

r <- multiclass.roc(test$Satisfaction_Rate, p1, percent = TRUE)

roc <- r[['rocs']]

r1 <- roc[[1]]

plot.roc(r1,

        print.auc=TRUE,

        auc.polygon=TRUE,

        grid=c(0.1, 0.2),

        grid.col=c("green", "red"),

        max.auc.polygon=TRUE,

        auc.polygon.col="lightblue",

        print.thres=TRUE,

        main= 'ROC Curve - Model 1')
```

9. Appendix 2: Visualisation

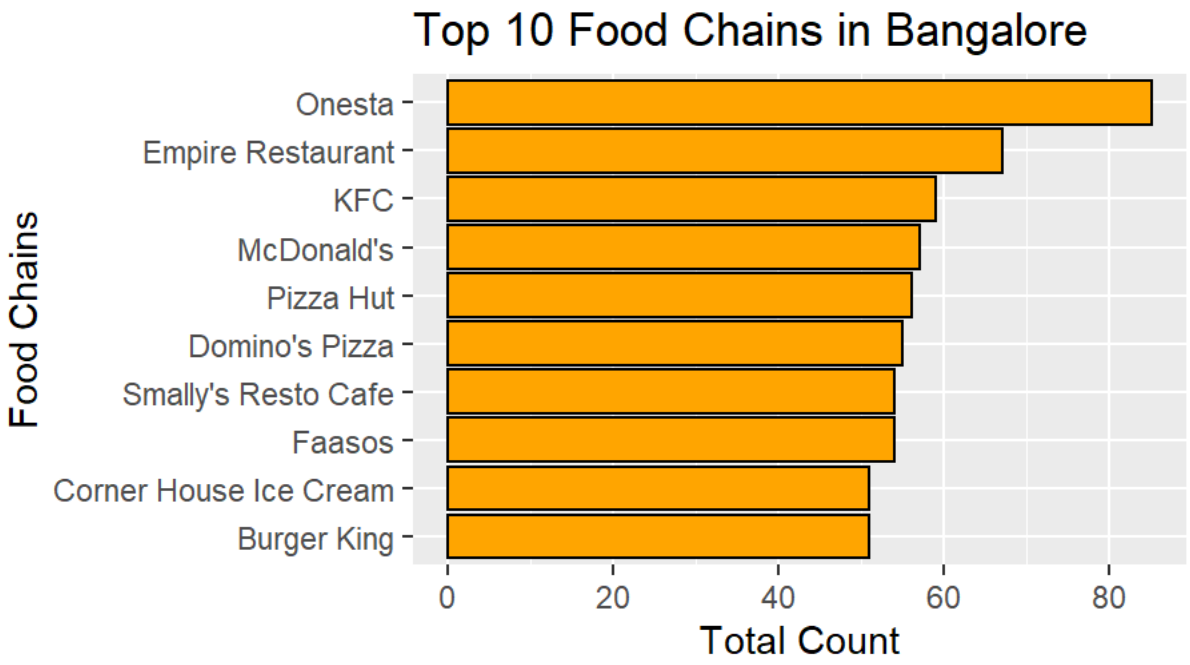

Figure shows Top Restaurants Food Chain in Bangalore

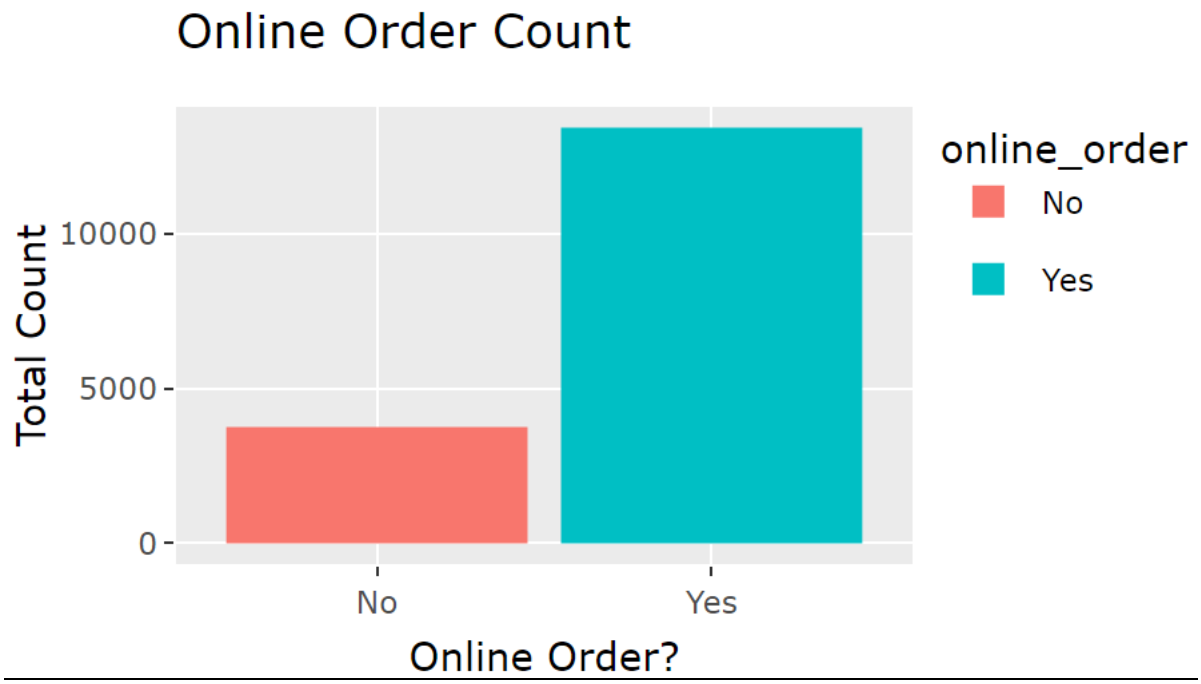

Figure Shows Online Order Count

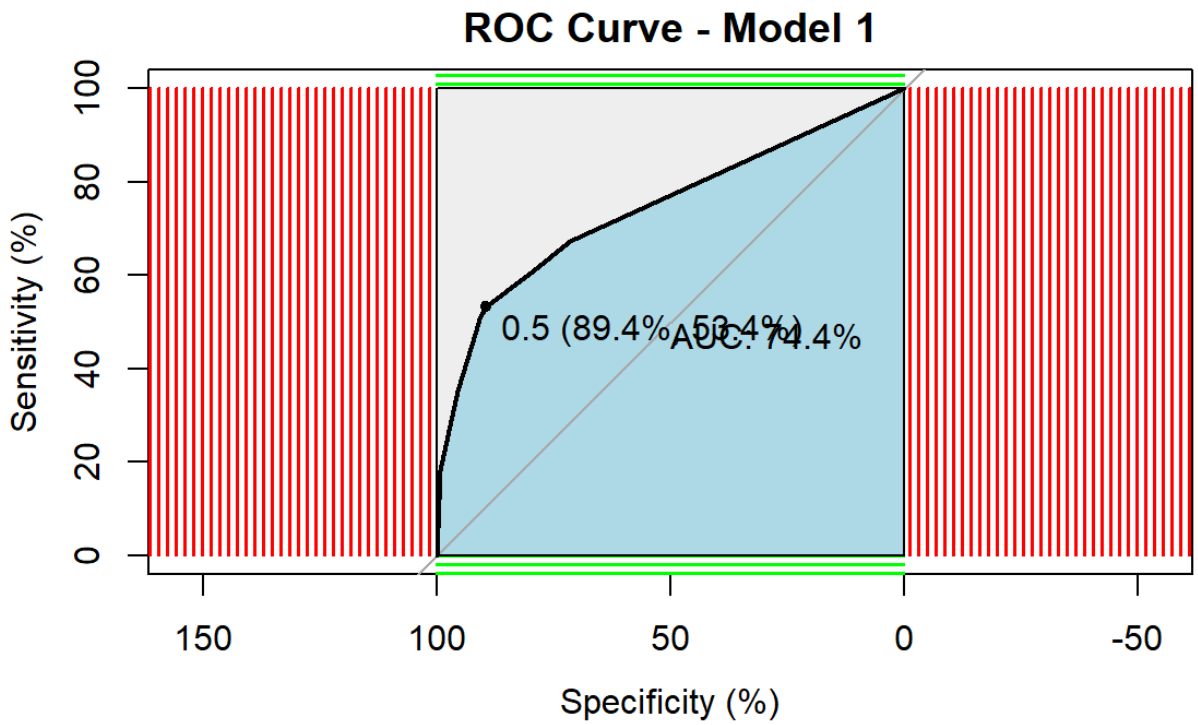

Figure Shows ROC Curve for Decision Tree Model 1

10. Appendix: Ethical Approval Form

PratikB_Student
Ethical Approval Form