EDA of Zomato Bangalore Restaurants

Hitesh Jarwani[1], Jaymin Shah[2], Dr. Smita Agrawal[3]

[1]Masters of Computer Applications, Nirma University, Gujarat, India
[2]Masters of Computer Applications, Nirma University, Gujarat, India
[3]Assistant Professor, Masters of Computer Applications, Nirma University, Gujarat, India

Abstract

The objective of this paper is to perform the exploratory data analysis on the Zomato Bangalore Restaurants. Firstly, an overview of the dataset and technical summary is summarized. The Zomato Bangalore Restaurants Datasets by Himanshu Poddar on Kaggle is a comprehensive collection of restaurants and their details scraped from Zomato website. The data is latest upto 15th march 2019. After that, data cleaning was performed to make sure that the data is fit for analysis purposes. Next, the data visualization process was undertaken and conclusions were drawn for the same.

Keywords: Data Visualization, Data Analysis, EDA, Bangalore Restaurants, Zomato Bangalore, python

1. Introduction

Bangalore, being the IT capital of the country, is home to a wide and diverse range of people. The people coming from diverse regions bring along with them their diverse eating habits, and so, Bangalore is home to one of the finest and most unique cuisines in the world. From South Indian to Mexican and Spanish, one can find here what their hearts crave and as such it is no less than a paradise for a foodie (a food loving person). The current restaurant count sits at about 12,000 restaurants, and this industry hasn't even reached a saturation point.

But this amount of market and customer base also comes with some challenges. It has become difficult for new restaurants to compete with already established restaurants. The key issues that continue to pose a challenge to them include high real estate costs, rising food costs, shortage of quality manpower, fragmented supply chain and over-licensing.

The Zomato Bangalore Restaurants Dataset by Himanshu Poddar on Kaggle is a comprehensive collection of restaurants and their details scraped from Zomato website. The data is latest up-to 15th March 2019.

Using this dataset as the guiding force, this project aims to provide insight into the restaurant industry of Bangalore city. Particularly, exploratory data analysis is performed to answer questions like popular restaurant cuisines in a locality, restaurant count in all the localities, and many more, which would ultimately help gain insights into the restaurant industry.

2. Related Work

Jagdale, S. et al. [1] used the Zomato Bangalore Restaurants dataset to compare the accuracy of various supervised machine learning algorithms in predicting the user rating for a particular restaurant. Specifically, they performed feature selection, and features such as number of votes, location, cuisine, online ordering facility, advance table booking facility, etc. were selected as input features, and the user rating was selected as the target feature. Algorithms like logistic regression, decision tree, KNN, random forest, support vector machine, gradient boosting and naive bayes were used and their accuracy was measured. They found that random forest has the best accuracy (94.90%) amongst all the algorithms.

In [2], the authors have created a food review system - Fiducia, which aims to identify information (positive or negative sentiment) on each food item mentioned in the user reviews. It takes customers' preferred food as an input, and recommends restaurants which have a reviewed acclaim about that food item as an output. Thus, it goes one step further, and performs the sentiment check on each food item mentioned in the reviews. The system also recommends a list of side foods which would go well with the input food item. The dataset for the study was retrieved using Zomato API, and a total of 3131 reviews were obtained.

3. Tools and Technologies used

The entire data analysis process, right from data wrangling to data visualization, has been undertaken using Python programming language, and its libraries. There are multiple reasons why we choose Python over its contemporaries. Firstly, the vast majority of open source libraries which are available in Python makes it very easy to undertake tasks like data manipulation, data visualization, statistical analysis, machine learning, natural language processing and data mining to name a few. Another reason is the shallow learning curve of Python, which makes it simple for experienced programmers as well as academicians to write and understand the Python code. The official documentation of Python and its libraries is well supported and maintained, with regular fixes and updates. The chief reason was the vast majority of library support is what compelled us to use Python.

2.1. Environment

The entire project was written and executed on Google Colab. We leveraged the power and reliability of Google servers to make our work easy. Additionally, there was no need to manually install packages, as the environment was already equipped with them. All we had to do was import the desired packages.

2.2. Python Packages Used

1. NumPy

   One of the fundamental packages for scientific programming in Python. It provides fast operations on arrays like shape manipulation, sorting, linear algebra, random simulations, and many more. The most pivotal object is *ndarray*, which is an n-dimensional array of homogeneous data-type.

2. Pandas

   It is a Python package which provides fast, flexible, and expressive data structures which makes working with “relational” data easy. Pandas comprises two data structures - Series (1-dimensional) and DataFrames (2-dimensional). Tasks like merging, sorting, renaming, dropping, slicing, querying, etc. can be easily performed.

3. Matplotlib

   Matplotlib is a powerful visualization library for statistical inferences. It provides a MATLAB-like interface, written in Python and is open source. Some common terminologies associated with Matplotlib are Figure, Axes, and Axis.

4. Seaborn

   Seaborn is a library which is used to create visualizations and statistical graphs in Python. It is built on top of Matplotlib. One major advantage of this library is that it integrates pretty well with Pandas library. Thus, it operates on DataFrames and Series, thus making them a very powerful tool for visualization.

5. Wordcloud

   Word cloud is a type of data visualization which represents text data (for e.g. a word) and the size of each text implies its importance or frequency. Thus, it can highlight significant text data points in a sentence.

4. Dataset Overview

Zomato Bangalore Restaurants provides a comprehensive set of data about restaurants in Bangalore city of Karnataka state of India. It was scrapped from the Zomato website by the author. The web scraping process was carried out in two phases.

In phase-1, the URL, name and address of the restaurant which were visible on the front page on Zomato were extracted. The URLs for each of the restaurants on the zomato were recorded in the csv file so that later the data can be extracted individually for each restaurant.

In phase-2, the recorded data for each restaurant and each category was read and data for each restaurant was scraped individually. 15 variables were scrapped in this phase. For each of the neighborhoods and for each category their online-order, booktable, rate, votes, phone, location, rest-type, dish-liked, cuisines, approx-cost(for two people), reviewslist, menu_item was extracted.

3.1. Technical Summary

The following command gives use the brief description of the dataset.

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51717 entries, 0 to 51716
Data columns (total 17 columns):
 #   Column                       Non-Null Count  Dtype 
---  ------                       --------------  ----- 
 0   url                          51717 non-null  object
 1   address                      51717 non-null  object
 2   name                         51717 non-null  object
 3   online_order                 51717 non-null  object
 4   book_table                   51717 non-null  object
 5   rate                         43942 non-null  object
 6   votes                        51717 non-null  int64 
 7   phone                        50509 non-null  object
 8   location                     51696 non-null  object
 9   rest_type                    51490 non-null  object
 10  dish_liked                   23639 non-null  object
 11  cuisines                     51672 non-null  object
 12  approx_cost(for two people)  51371 non-null  object
 13  reviews_list                 51717 non-null  object
 14  menu_item                    51717 non-null  object
 15  listed_in(type)              51717 non-null  object
 16  listed_in(city)              51717 non-null  object
dtypes: int64(1), object(16)
memory usage: 6.7+ MB
```

The following command shows the unique value count for each column.

```
for col in df:
    print("{} {}".format(col, len(df[col].unique())))
```

```
Column                       Unique Count

url                          51717
address                      11495
name                         8792
online_order                 2
book_table                   2
rate                         65
votes                        2328
phone                        14927
location                     94
rest_type                    94
dish_liked                   5272
cuisines                     2724
approx_cost(for two people)  71
reviews_list                 22513
menu_item                    9098
listed_in(type)              7
listed_in(city)              30
```

One might wonder how come there are more unique URLs than unique restaurants. Currently the dataset contains more than 51000 restaurants but there aren't 51k restaurants in Bengaluru. The reason for the same is because the data set was scraped individually for each category, e.g. Buffet, dine-out, pubs, bars, delivery, nightlife

etc. and so it may have happened that a restaurant was mentioned in more than one category. This can be resolved by undertaking an appropriate data cleaning process.

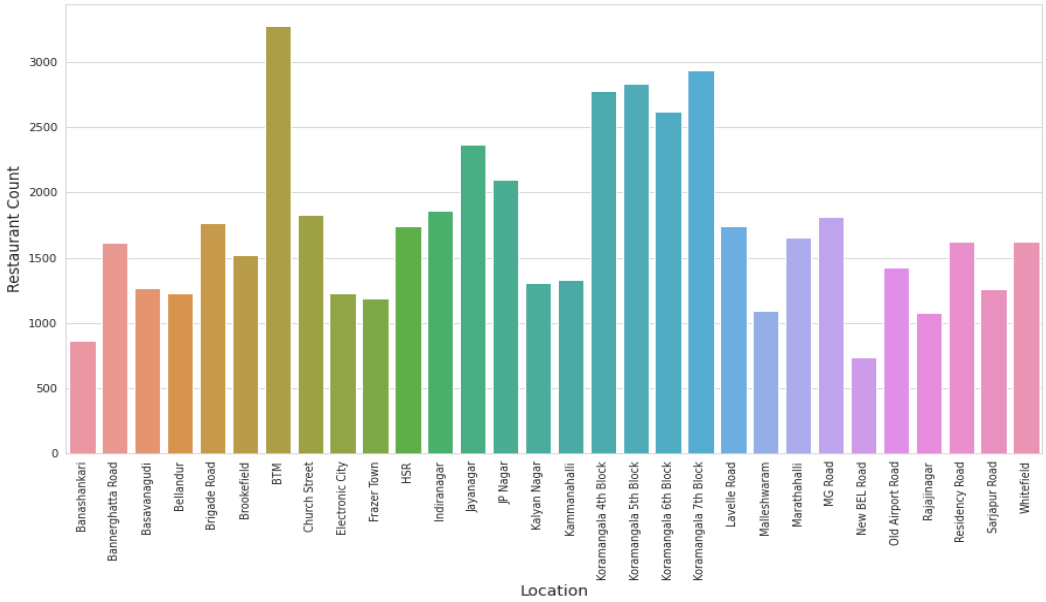
5. Objective and goal of the project

Using the process of data cleaning and data visualization, we will try to answer following statements and questions:-

1. What is the restaurant count in each locality?
2. Which are the restaurants with the most outlets?
3. Which are the most common restaurant types in the whole of Bangalore?
4. What are the most popular cuisines in Bangalore?
5. What are the most popular cuisines in each locality of Bangalore?
6. How many restaurants accept online orders, and what is the distribution of restaurant rating in both the categories?
7. How many restaurants provide advance table booking facilities? Is this number any different for high-end restaurants (average cost greater than ₹4000)?
8. What is the distribution of average cost for all restaurants?
9. Show the distribution of average cost using boxplot.
10. What is the relationship between the rating of a restaurant and its average cost?
11. Which restaurant ratings are most reliable? (most number of votes)
    a) What do the reviews of these restaurants convey?
12. An interactive recommendation system.

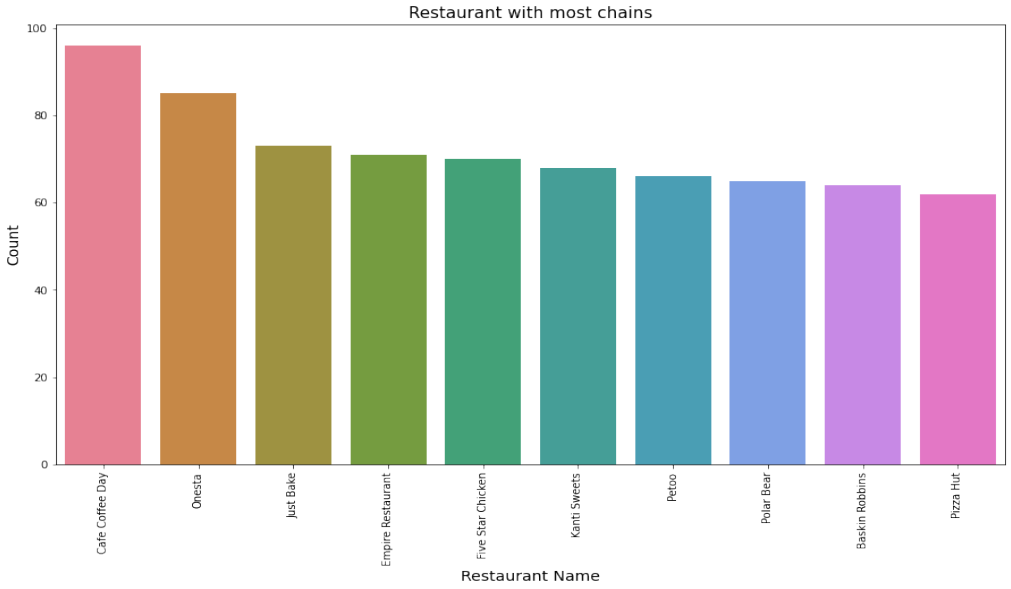
6. Data visualization and analysis

5.1 What is the restaurant count in each locality?



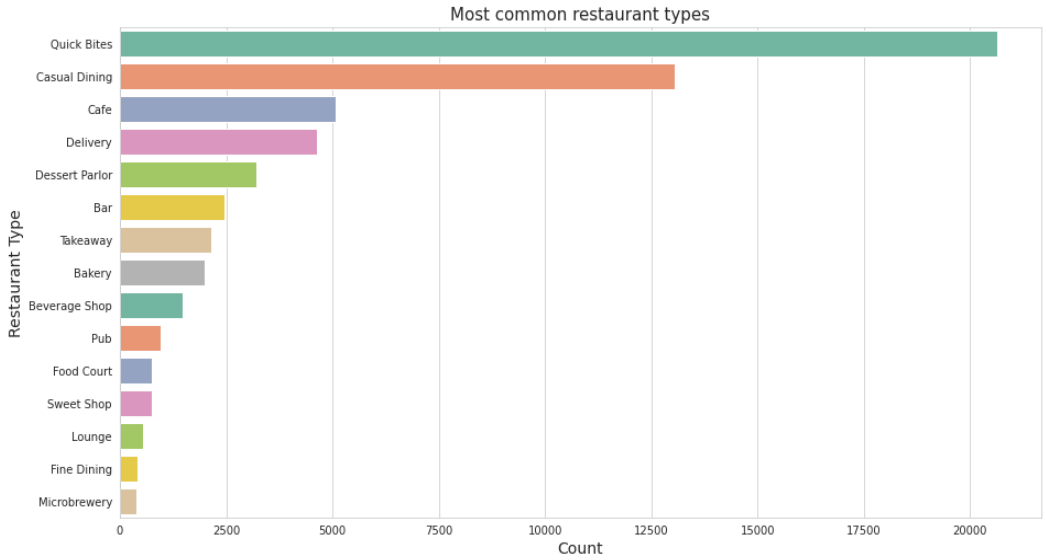
BTM is the area with the most number of restaurants. A quick Google search about this area shows that it borders many other commercial regions, making it one of the most popular residential and commercial areas in Bangalore. This could be one of the reasons for high restaurant count. Additionally, it is also one of the highest and fastest growing districts in Bangalore in terms of property prices. It is followed by Koramangala district.

5.2 Which are the restaurants with the most outlets?



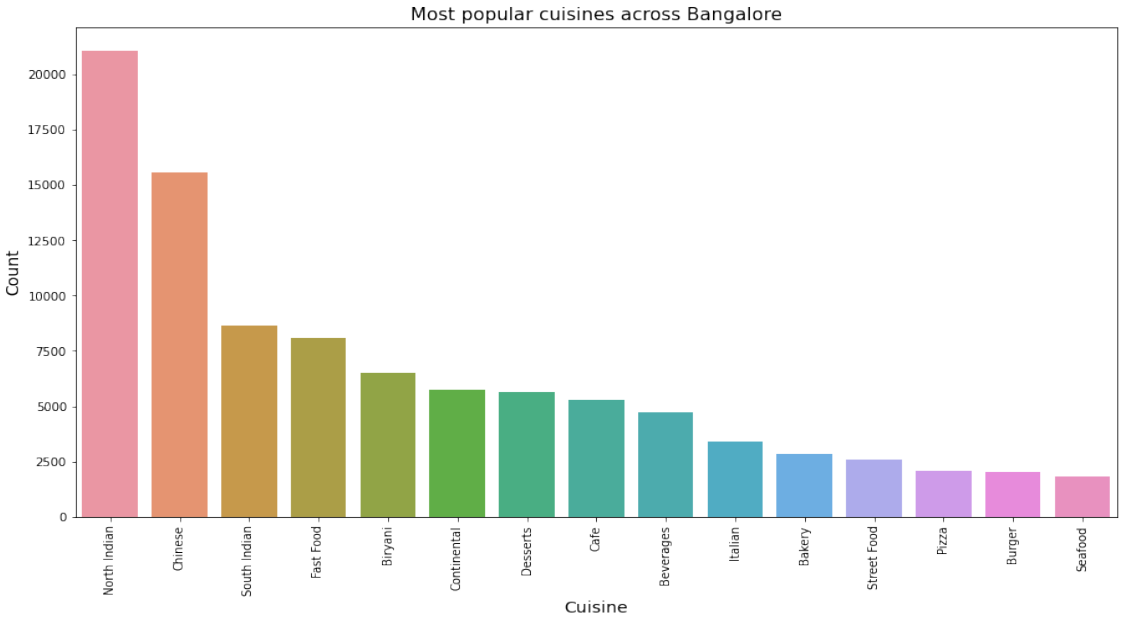
Cafe Coffee Day, Onesta and Just Bake are the top three restaurants with the most number of outlets. Almost all other restaurants fall into either casual dining, or cafes, implying the demands for such restaurants in this region. We explore this observation more in our next objective.

5.3 Which are the most common restaurant types in the whole of Bangalore?



Quick bites dominate all other restaurant types by far. Bangalore, being a huge metropolitan, can be viewed as a city with constant rush, and where everybody is in a hurry. This could explain why restaurants providing quick bites like snacks would top the list. This reasoning is further supported by the fact that Casual dining, Cafe, Delivery and Dessert parlor are other types included in the top 5.

5.4 What are the most popular cuisines in Bangalore?



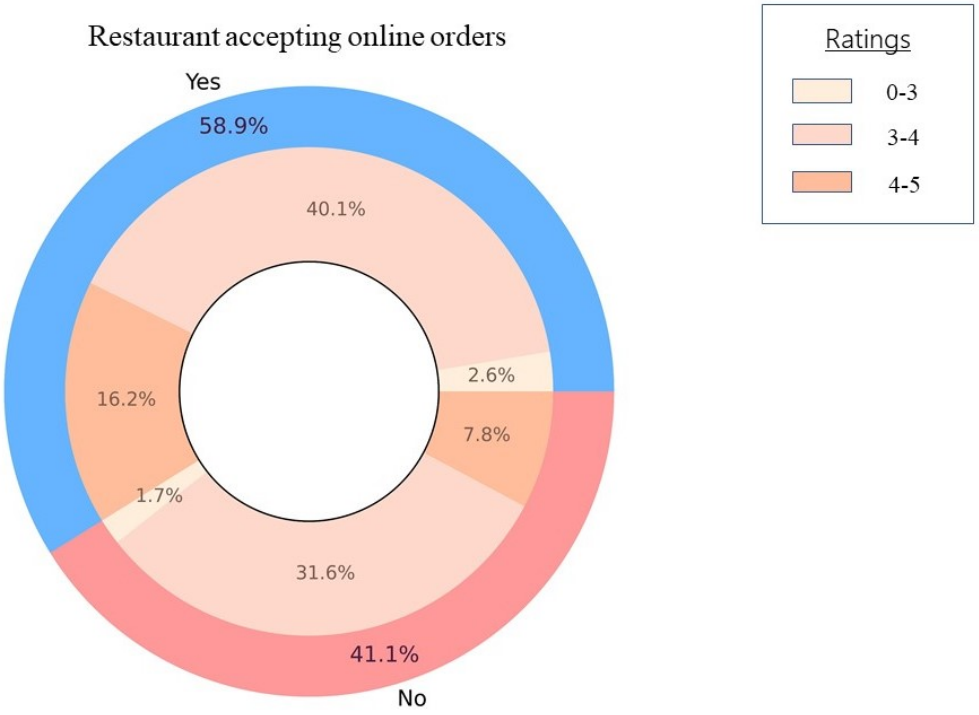
Once again, a great deal of diversity can be observed. Interestingly, North Indian cuisine tops the list. South Indian cuisine comes in third place, even beaten by Chinese cuisine. Fast food and Biryani also make it to the top 5. The “urban” and “fast” culture of Indian IT capital can be observed by the food habits of the locals. We find cuisines like seafood, continental, mexican and chinese in the most popular cuisines list, which further sheds light on the diverse demographics of Bangalore.

5.5 What are the most popular cuisines in each locality of Bangalore?

Location	0	1	2	3	4
Banashankari	North Indian	South Indian	Chinese	Fast Food	Biryani
Bannerghatta Road	North Indian	Chinese	South Indian	Fast Food	Biryani
Basavanagudi	North Indian	Chinese	South Indian	Fast Food	Beverages
Bellandur	North Indian	Chinese	Biryani	South Indian	Fast Food
Brigade Road	North Indian	Chinese	Continental	South Indian	Fast Food
Brookefield	North Indian	Chinese	Biryani	South Indian	Fast Food
BTM	North Indian	Chinese	Fast Food	South Indian	Biryani
Church Street	North Indian	Chinese	Continental	South Indian	Fast Food
Electronic City	North Indian	Chinese	South Indian	Biryani	Fast Food
Frazer Town	North Indian	Chinese	South Indian	Fast Food	Cafe
HSR	North Indian	Chinese	Fast Food	South Indian	Biryani
Indiranagar	North Indian	Chinese	Continental	Fast Food	Cafe
Jayanagar	North Indian	Chinese	Fast Food	South Indian	Beverages
JP Nagar	North Indian	Chinese	South Indian	Fast Food	Biryani
Kalyan Nagar	North Indian	Chinese	Fast Food	South Indian	Biryani
Kammanahalli	North Indian	Chinese	Fast Food	South Indian	Biryani
Koramangala 4th Block	North Indian	Chinese	Fast Food	South Indian	Biryani
Koramangala 5th Block	North Indian	Chinese	Fast Food	South Indian	Biryani
Koramangala 6th Block	North Indian	Chinese	Fast Food	South Indian	Biryani
Koramangala 7th Block	North Indian	Chinese	Fast Food	South Indian	Biryani
Lavelle Road	North Indian	Chinese	South Indian	Continental	Fast Food
Malleshwaram	North Indian	Chinese	South Indian	Fast Food	Biryani
Marathahalli	North Indian	Chinese	South Indian	Biryani	Fast Food
MG Road	North Indian	Chinese	Continental	South Indian	Fast Food
New BEL Road	North Indian	Chinese	Fast Food	South Indian	Biryani
Old Airport Road	North Indian	Chinese	Fast Food	Continental	South Indian
Rajajinagar	North Indian	Chinese	South Indian	Fast Food	Biryani
Residency Road	North Indian	Chinese	South Indian	Continental	Fast Food
Sarjapur Road	North Indian	Chinese	Biryani	Fast Food	South Indian
Whitefield	North Indian	Chinese	Biryani	South Indian	Fast Food

There are a total of 30 unique localities in the dataset and North Indian is the most popular choice in all the localities. This aligns with the fact that the North Indian cuisine is the most famous cuisine in the entire city, as observed in section *5.4*. We also observe that except for one locality (Banashankari), Chinese cuisine is the preferred cuisine after North Indian. For the 3rd, 4th and 5th most popular cuisine, different trends are observed in different localities, and cuisines like South Indian, Continental, Fast Food, Biryani etc. make it to the top of the list.

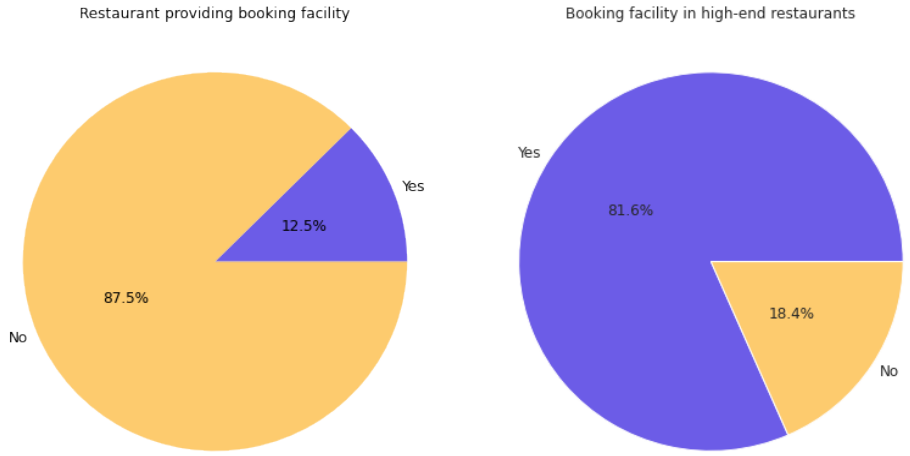
5.6 How many restaurants accept online orders, and what is the distribution of restaurant rating in both the categories?



Nearly 59% of the restaurants accept online orders, while the rest 41% only provide dine-in or takeaway services. This may be due to the fact that a number of smaller restaurants might not be able to afford the charges and expenses which incurs on associating with Zomato.

Looking at the ratings distribution in both the categories, we find that the restaurants with the ratings in the range of 3.0 to 4.0 starts form the majority, while restaurants with higher ratings (rating greater than 4.0) are more dominant in “Yes” (accepting online orders) category.

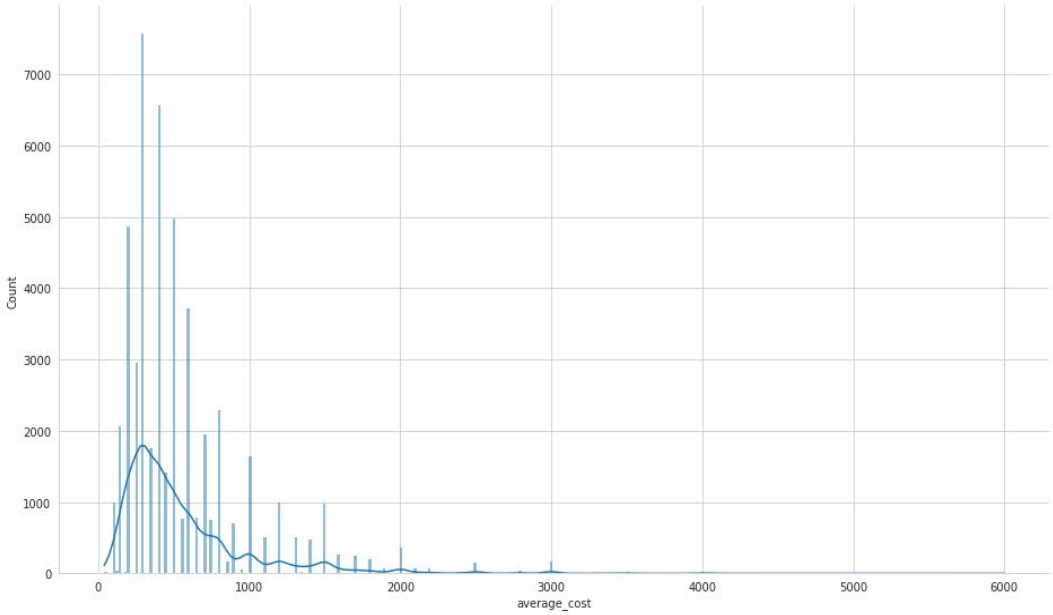
5.7 How many restaurants provide advance table booking facilities? Is this number any different for high-end restaurants (average cost greater than ₹4000)?



A huge majority of 87.5% restaurants don't provide advance table booking services. The simple reason behind this can be the fact that most restaurants are small to medium scale in size and business, and some of them might even struggle to meet ends. Complementary services like table booking prove to be a loss to their business.

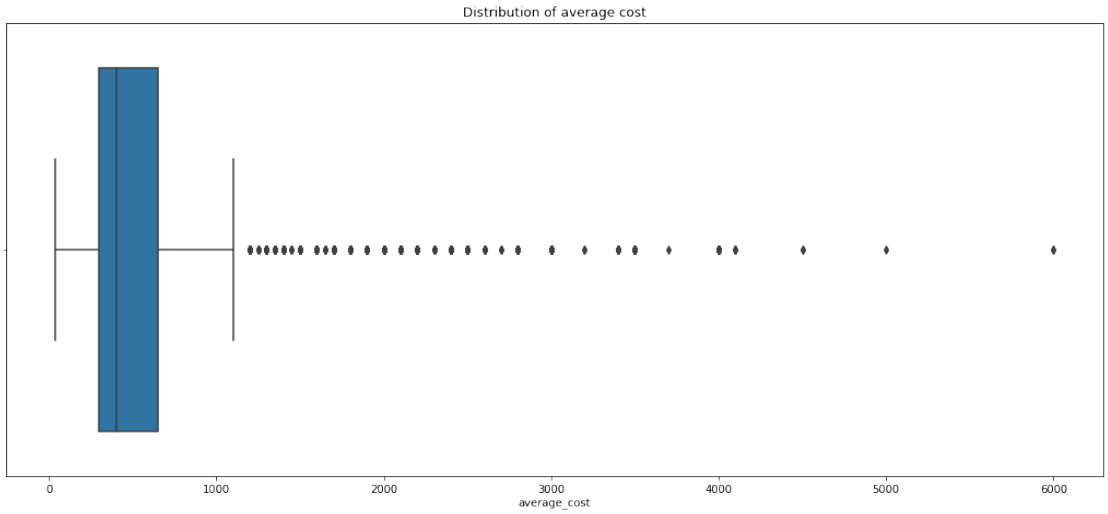
The above pie chart shows the status of table booking facilities at high-end and luxurious restaurants (*average_cost >= ₹4000).* This ratio is quite contradictory to the one we saw previously. High-end restaurants add more emphasis on qualities like ambience, vibe, food quality and table booking facility is a part of it.

5.8 What is the distribution of average cost for all restaurants?



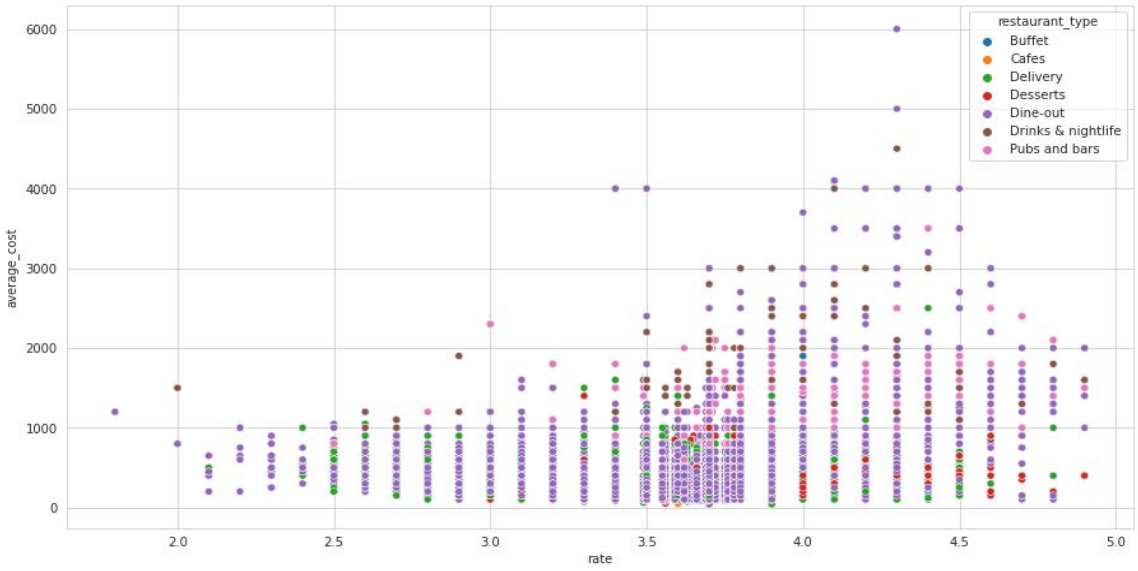
The average cost of most of the restaurants is less than ₹1000. In fact, the distribution is right skewed in nature. We can observe some spikes in the range of ₹ 1000 to ₹ 2000, but the restaurant count in this range is quite less, and the majority of the restaurants have an average cost of about ₹ 500.

5.9 Show the distribution of average cost using boxplot.



From the boxplot, we observe that the average dining cost has a median of about ₹500. Also, the Quartile-1 lies at about ₹400, and Quartile-3 lies at about ₹800, which means that 50% of all the restaurants have average cost between ₹400 and ₹800.

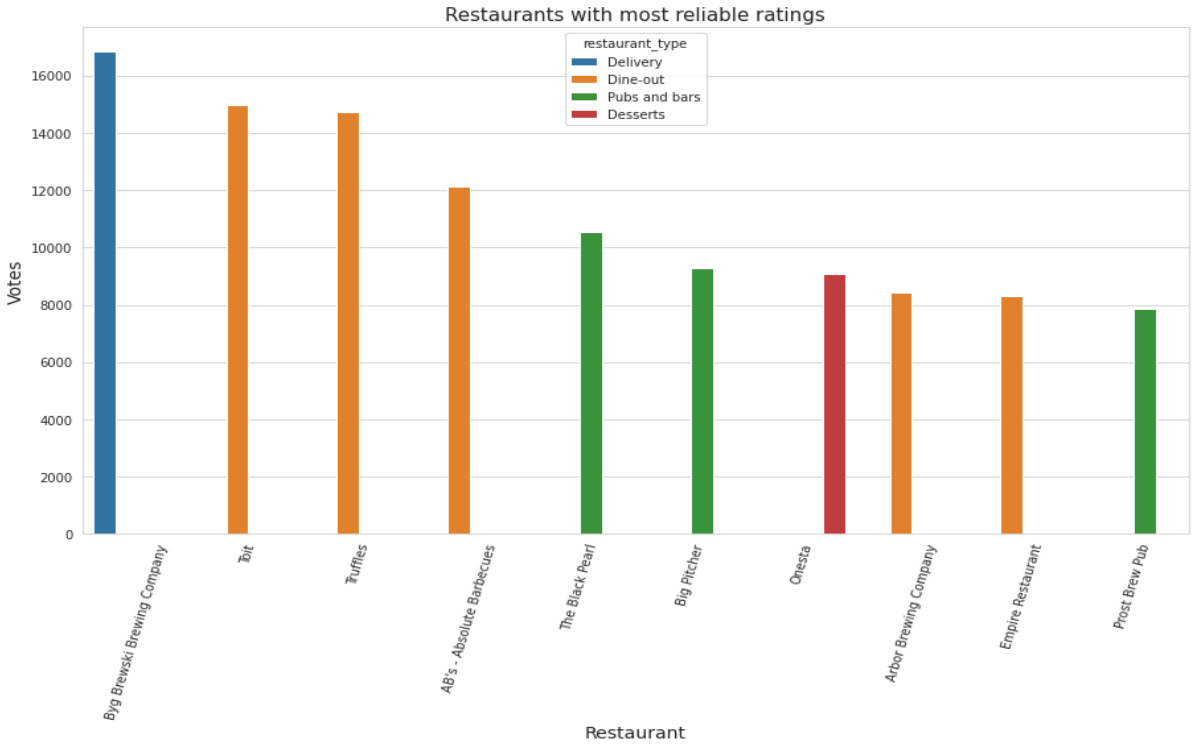
5.10 What is the relationship between the rating of a restaurant and its average cost?



We try to see if there is any relation between the rating received by a restaurant and the average cost of that restaurant. Most of the restaurants are concentrated between ratings of 3.5 and 3.8, and also their average cost is less than ₹1000, and these are mostly dine-out restaurants and cafes.

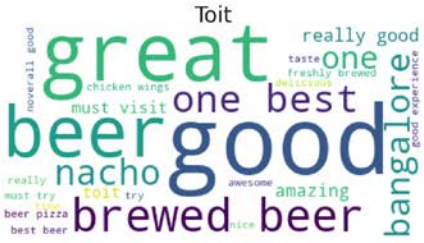
Highest rated restaurants mostly fall between the ranges of ₹1000 to ₹2000, but many restaurants with cost less that ₹1000 are also some of the highest rated. On the other side, almost all the restaurants with low ratings (< 3.0) have their average costs lower than ₹1000.

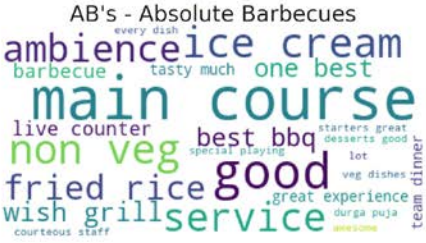
5.11 Which restaurant ratings are most reliable? (most number of votes)



We list out the top 10 most reliable restaurants in the city. Reliability in this context refers to the number of votes received by the restaurant, i.e. the top most restaurant will have the most number of votes. Additionally, the restaurant type has also been classified as a hue. Three out of most voted ten restaurants fall in the category of “pubs and bars”, while five belong to “dine-out”. All the restaurants have almost 8,000 and more number of votes.

a. What do the reviews of these restaurants convey?





The customer reviews received by the top 10 reliable restaurants have been presented in the form of a word cloud. A word cloud, also called a tag cloud, is a visual representation of a text, particularly words, wherein the size of the word indicates its importance or its frequency. In our case, the size of the word conveys its frequency or occurrence in all the reviews.

It is observed that the type of cuisine a restaurant serves does have an influence on the reviews. Bar and pubs have higher frequency of words like beer, drink, cider etc. Reviews of fast food restaurants will have words like pizza, burger, friends, unlimited pizza, dessert, etc. Almost all the reviews have positive words like good, great, must visit, must try, good service, amazing, best, etc.

5.12 An interactive restaurant recommendation system

The interactive restaurant recommendation system is such that it takes in various input parameters from the users, and based on those parameters, recommends the best restaurant the user can opt visit or order food from. Following are the parameters on the basis of which the restaurants are recommended:-

Parameter	Description	Input Type
online_order	Whether the restaurant provides online ordering facility or not	String
book_table	Whether the restaurant provides table booking facility or not	String
cuisines	The cuisines provides by the restaurant	List
average_cost	The average dining cost of the restaurant (filtered restaurant should have the cost less than the input cost)	Integer
location	The location (locality) of the restaurant	List
rate	The average rating of the restaurant (filtered restaurant should have the rating greater than or equal the input rating)	Double
dishes_liked	The popular dishes at the restaurant	List

As an example, consider a user who wants to find a restaurant with Mughlai and North Indian cuisines, in the locality of either BTM, Rajajinagar, or Brookefield, and an average cost of up-to ₹1000 with a rating of 4.0 stars and above, and also table booking facility. Then, the input to the system can be given in the following manner:-

```
d = recommend(location=['BTM', 'Rajajinagar', 'Brookefield'],
              average_cost=1000,
              rate=4.0,
              cuisines=["Mughlai", "North Indian"],
              book_table="Yes")
```

Upon execution of the above piece of code, the following output will be provided to the user:-

```
Based on your choices, following are the recommendations:-
array(['Le Arabia', 'Imperio Restaurant', "Dadi's Dum Biryani",
       'Zero Mile Punjab', 'Oldtown Dilli', 'Little Lucknow',
       'Melt - Eden Park', 'Aroma Fine Dine'], dtype=object)
```

The above mentioned restaurants conform to all the input parameters as specified by the user. Additionally, the restaurants in the output are ordered in the decreasing order as per their rating, i.e. the highest restaurant appears first in the list.

7. Conclusion and future work.

In this report, we used the Zomato Bangalore Restaurants dataset to perform data visualization and analysis. The primary objective of doing so was to understand the food sector better in the city of Bangalore. To do so, we came up with various objectives which could be answered using the dataset. We visualized those questions and derived conclusions from the same.

We started off talking about the dataset in depth. Next, the main objectives were described. Before creating visualizations to answer the questions, we performed the data cleaning process to make sure that the data is fit for analysis purposes. Next, the data visualization process was undertaken and conclusions were drawn for the same.

For future scope, we would like to incorporate maps, which could better summarize the spread and distribution of restaurants. We would also like to implement a sentiment analysis module which would perform analysis on user reviews, and find out qualities or aspects that determine the overall rating of a restaurant. Additionally, we would also like to incorporate the interactive restaurant recommendation system as a full-fledged web or a desktop application.

8. References

[1] R. S. Jagdale and S. S. Deshmukh, "Sentiment Classification on Twitter and Zomato Dataset Using Supervised Learning Algorithms," 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), 2020, pp. 330-334, doi: 10.1109/ICSIDEMPC49020.2020.9299582.

[2] Goel, M., Agarwal, A., Thukral, D., & Chakraborty, T. (2019). Fiducia: A Personalized Food Recommender System for Zomato. *arXiv preprint arXiv:1903.10117*.

[3] Poddar H. (2019) Zomato Bangalore Restaurants, Restaurants of Bengaluru, Kaggle, https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants

[4] Seaborn Authors (2021) API Reference, Seaborn, https://seaborn.pydata.org/api.html

[5] Pandas Authors (2021) API Reference, Pandas, https://pandas.pydata.org/docs/reference/index.html#api

[6] NumPy Authors (2021) NumPy Reference, NumPy, https://numpy.org/doc/stable/reference/index.html

[7] Kadam S. (2020, April 18) Generating Word Cloud in Python, Geeksforgeeks, https://www.geeksforgeeks.org/generating-word-cloud-python/