A Data-Driven Analysis of Restaurant Delivery Platforms

A Final report for the BDM capstone Project

Submitted by

Name: Rishabh Indoria

Roll number: 21F3001823

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Table of Contents

List of Tables

List of Figures

1 A Data-Driven Analysis of Restaurant Delivery Platforms

Restaurant industry is a field that is always in constant development. Recently, online ordering apps have been drastically growing, which leads me to identify factors that could be changed or modified to improve customer experience. Among other things, the report covers important changes and trends in the restaurant industry where data is taken from Zomato and Swiggy, the leading delivery sites. Through analysis and interpretation, I try to uncover actionable insights to address three primary problem statements:

1. **Investigating the Impact of Cuisine Type on Customer Preferences**: This problem statement focuses on understanding how different cuisine types influence customer preferences and dining experiences. By analyzing cuisine frequencies, we aim to identify major preferences.
2. **Assessing the Competitive Dynamics of the Restaurant Industry**: This problem statement explores the influence of factors influencing competition within the restaurant industry. Through the use of correlation analysis, and exploratory data visualizations, I seek to gain insights into what strategies restaurants should employ, to gain an edge in the restaurant industry.
3. **Identifying Strategic Opportunities for Restaurant Expansion**: The final problem statement examines strategic opportunities for restaurant expansion. By using geospatial analysis techniques, the aim is to identify regions for expansion.

Through our analysis, we examined valuable insights into customer preferences, and strategic opportunities within the restaurant industry. I try to propose actionable recommendations to guide decision-making and gain positive outcomes for restaurant owners.

In conclusion, this report offers a comprehensive overview of key insights and trends within the restaurant industry, as well as actionable recommendations to address different challenges. By leveraging these insights and analysis, restaurant owners could drive business growth, and deliver better experiences to their customers.

2 Proof of Originality

In this project, I have used a total of three datasets obtained from Kaggle to conduct a comprehensive analysis of the restaurant industry.

1. **Swiggy Dataset:**
   - Link: https://www.kaggle.com/datasets/abhijitdahatonde/swiggy-restuarant-dataset
   - Description: The first dataset was derived from Swiggy, which is one of the famous food delivery platforms. This dataset includes details about the restaurants listed on Swiggy, which include attributes like the restaurant name, their average rating, number of ratings, and average cost to name a few. This set of data is valuable not only because it can be used to analyze consumer behavior and restaurant attributes, but also for determining consumer preferences and characteristics of such restaurants.
2. **Zomato Dataset:**
   - Link: https://www.kaggle.com/datasets/abhijitdahatonde/zomato-restaurants-dataset
   - Description: The second dataset is obtained from Zomato, another popular food ordering app. This dataset is similar to the Swiggy dataset as the records here contain the same variables but also contains the variables if online ordering is available or not, availability of reservation service and the type of cuisines offered.
3. **Swiggy Geospatial Dataset:**
   - Link: https://www.kaggle.com/datasets/lokeshparab/swiggy-restraurant-and-item-full-datasets
   - Description: Additionally, I am using a Swiggy profiled dataset specifically for geographical analysis. It is a huge data set consisting of different cities along with their average ratings. Owing to the wide span of coverage and the notable size, this dataset reflects an analytical advantage of analyzing geographical distribution patterns and regional variation of restaurant attributes.

By using these three datasets, we aim to acquire a better understanding of the restaurant, isolate key tendencies and patterns and offer actionable insights.

3 Descriptive Statistics

Before starting to analyze the datasets, I had to perform a few preprocessing tasks. This included removing unnecessary columns such as restaurant link, restaurant address, license number and so on. There were also NaN or missing values in the dataset in key columns which included Average Rating, and Average cost for 2. This was filled with their respective means. These steps were essential for creating clean and standardized datasets.

Next, we take a look at what data we ended up using after the preprocessing step.

First, we have the Swiggy dataset which contains Average Rating, Number of Ratings. Average Cost of 2, and Cuisine Type.

Variable Name	**Description**	**Data Type**	**Range**
Average Rating	Represents the overall rating given to a restaurant.	Float	Varies from 1 to 5, with 1 being the lowest and 5 the highest. [1, 5]
Number of Ratings	Total count of ratings received by the restaurants, it also reflects the level of customer engagement.	Integer	Can be any positive number. [0, ∞)
Average Cost of 2	Represents the typical cost of a meal for two people at the restaurant.	Float	Can be any positive number. [0, ∞)
Cuisine Type	Represents the type or style of cuisine offered by the restaurant.	String	The range of cuisines varies widely, including “Mexican”, “North Indian”, etc.

Table 1: Metadata for Swiggy dataset

Next, we have the Zomato dataset which contains similar variables to Swiggy, this includes Average Rating, Number of Ratings, Average Cost of 2, and Cuisine Type. Additionally, this dataset also includes data regarding whether online ordering is available or not, and whether table booking is available or not.

Variable Name	Description	Data Type	Range
Average Rating	Represents the overall rating given to a restaurant.	Float	Varies from 1 to 5, with 1 being the lowest and 5 the highest. [1, 5]
Number of Ratings	Total count of ratings received by the restaurants, it also reflects the level of customer engagement.	Integer	Can be any positive number. $[0, \infty)$
Average Cost of 2	Represents the typical cost of a meal for two people at the restaurant.	Float	Can be any positive number. $[0, \infty)$
Online Order	Indicates whether the restaurant offers online ordering or not.	Boolean	Binary variable where "Yes" indicates online ordering is available and "No" means otherwise.
Table Booking	Indicates whether the restaurant offers table booking or reservation services.	Boolean	Binary variable where "Yes" indicates table booking is available and "No" means otherwise.
Cuisine Type	Represents the type or style of cuisine offered by the restaurant.	String	The range of cuisines varies widely, including "Mexican", "North Indian", etc.

Table 2: Metadata for Zomato

Finally, we have the geospatial dataset, which only contains the City and their Average Rating. This dataset will only be used for geospatial analysis.

Variable Name	Description	Data Type	Range
City	Represents different cities within India	String	The range of cities is any city within India
Average Rating	Represents the overall rating given to a restaurant.	Float	Varies from 1 to 5, with 1 being the lowest and 5 the highest. [1, 5]

Table 3: Metadata for Geospatial dataset

Now, we visualize each variable to gain a better understanding of the dataset and also see if we can derive some meaningful conclusions from the distributions.

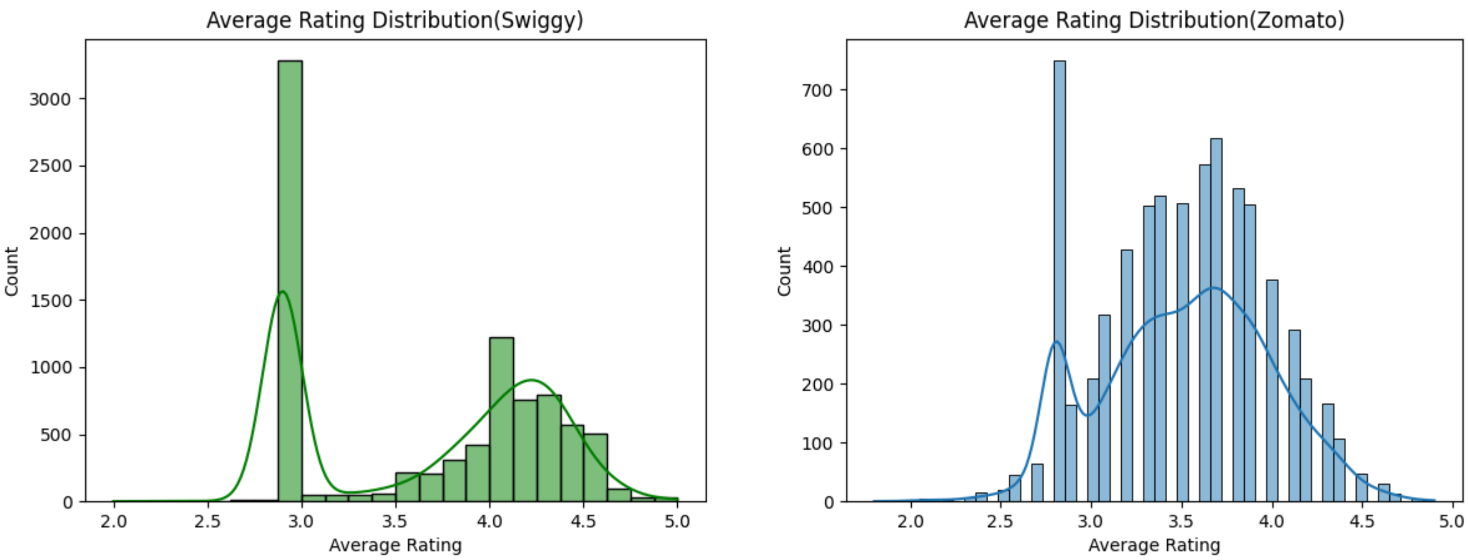


Fig 1: Distribution of “Average Rating”

We begin our investigation with the “Average Rating” variable, this represents the overall customer satisfaction rating for restaurants. This variable, as seen in metadata, is measured between 1 and 5. In general, we expect the average rating to be around the middle or 2.5 which is considered moderate. This metric also provides valuable insights into the quality of dining experiences offered by different restaurants and is a crucial factor depending on customer preferences and choices.

If we take a look at Figure 1 we can see that most restaurants have their average ratings around 2.7 to 3, which is a moderate rating and was expected. Another crucial observation is that the distribution for Swiggy seems to be more towards a higher rating than the distribution in Zomato, which could be an indication of the platforms. This fact is also evident from observing their respective means, the mean for Zomato is 3.51 whereas the average for Swiggy is 3.65. The rest of the ratings seem to follow the normal distribution.

This distribution of average ratings provides valuable insights into the satisfaction levels of customers across different restaurants listed on Zomato and Swiggy. Understanding these patterns helps to identify areas for improvement and enhance overall customer experiences.

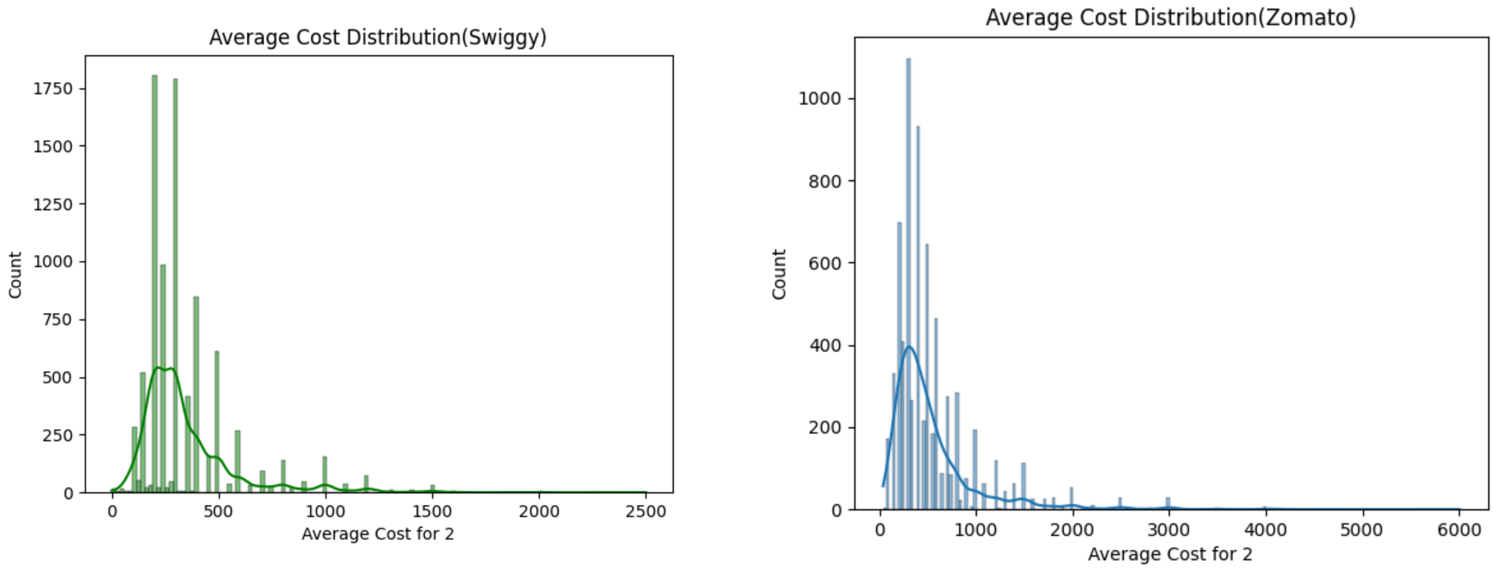


Fig 2: Distribution of “Average cost for 2”

Next, we have the “Average Cost of 2” variable, this represents the approximate amount of money required for a meal for two people at any restaurant. This variable could provide information regarding the affordability of restaurants and whether customers prefer low cost restaurants or high cost restaurants.

If we take a look at Figure 2 we observe that in both datasets the average cost is more biased towards the 0 to 1000 range. We can also observe that in the case of Swiggy the average cost is distributed between 0 and 500 with its mean being 348.4, whereas Zomato is distributed between 0 and 1000 with its mean being 540.29. We previously observed that Swiggy seemed to have slightly higher average rating, this could be related to the fact that average cost is lower in Swiggy.

This distribution seems to highlight the range of pricing options available within the restaurant industry. By understanding these patterns, we can cater to diverse customer preferences and optimize our pricing strategies to attract the target demographic.

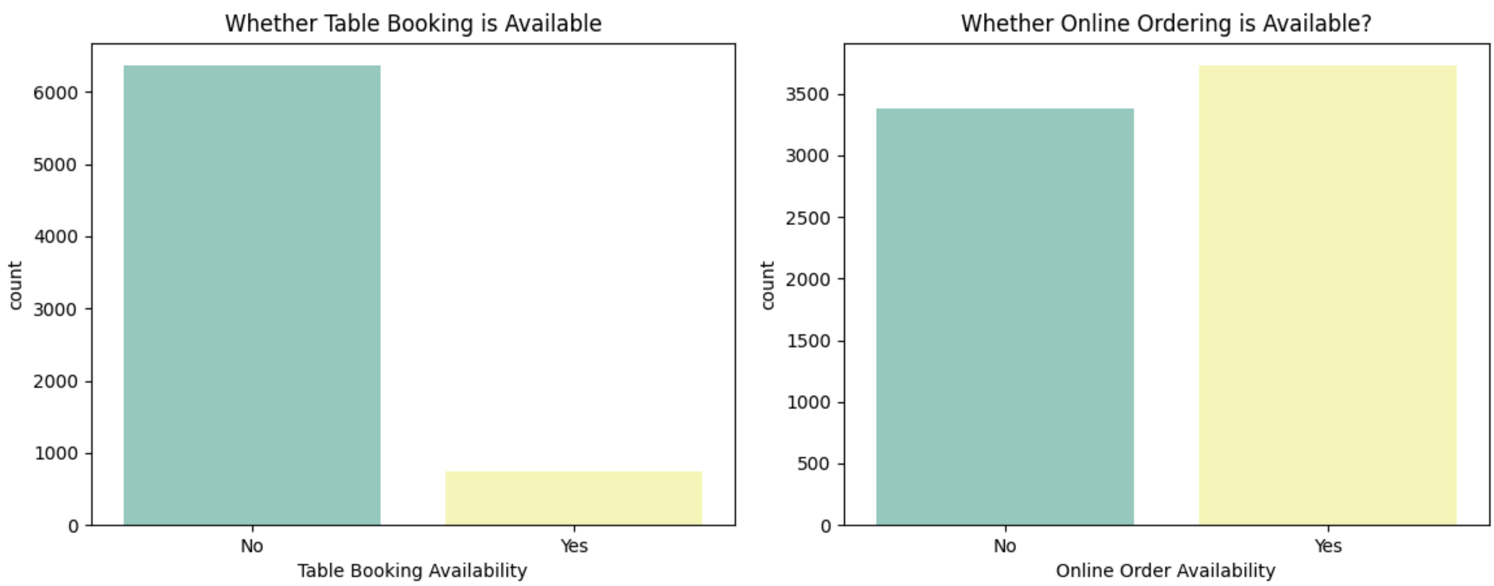


Fig 3: Frequency Distribution of "Online Ordering" and "Table Booking"

Further, we take a look at "Online Order" and "Table Booking" variables. Since only Zomato has these variables, we won't be able to get a complete overview for them, but we do get an idea for how these variables could influence the convenience and accessibility of restaurants. The "Online Order" variable indicates whether a restaurant offers the option for customers to place orders online, while the "Table Booking" variable denotes whether the restaurant provides the facility to reserve tables in advance.

Taking a look at Figure 3 we can clearly see that about equal number of restaurants seem to offer and do not offer online ordering. Whereas we see that a very small amount of restaurants seems to offer "Table Booking", this could be attributed to the fact that Zomato and Swiggy both started out as online ordering apps, and later included reservation as well. It is possible that the restaurants simply don't know about this facility.

Understanding these trends can help restaurants in optimizing their service offerings to meet customer demands.

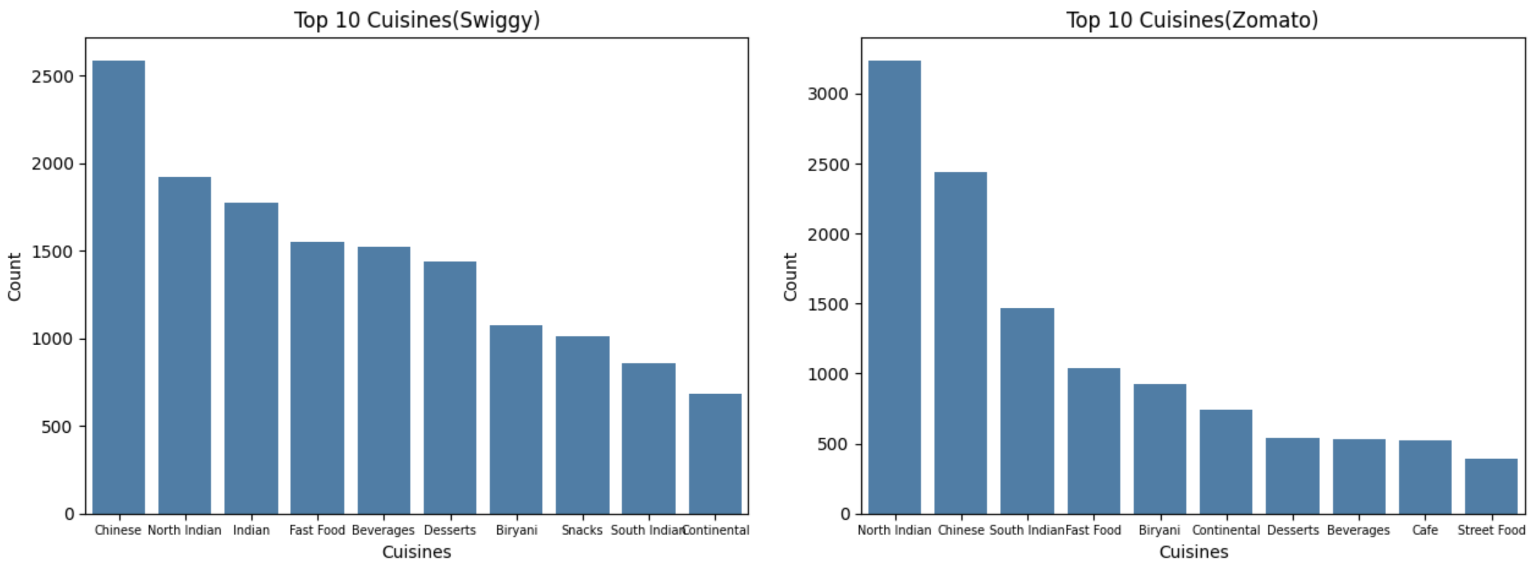


Fig 4: Frequency Distribution of Cuisine Types

Finally, we take a quick look at the cuisines offered by restaurants. This variable provides insights into the diversity of dining options available to consumers, and also reflects the customer preferences and cultural influences within the industry.

The visuals are made by counting how many restaurants offer a specific type of cuisine, and the top 10 cuisines offered by the restaurants have been displayed. If we take a look at the histogram, we see restaurants on both platforms seem to follow a similar pattern. North Indian, Chinese, South Indian, Fast Food, Beverages, Indian seem to be in top 10 for both platforms.

This distribution of cuisine types highlights the diversity and richness of dining options available within the restaurant industry. Understanding these patterns help restaurants tailor their menu offerings to cater to specific consumer tastes and preferences, thereby enhancing customer satisfaction and loyalty.

This concludes our descriptive analysis, and now we move to different analytical approaches and methodologies used.

4 Analytical Approach and Methodology

I have employed multiple analytical methods to gain important insights in the dataset.

Correlation analysis was used to identify potential relationships between the variables. It measures the degree to which two variables are linearly related to each other. The Pearson correlation coefficient for each pair of variables is calculated and a heatmap is plotted to identify relationships. A low or close to zero correlation coefficient indicates no linear relationship, whereas a high coefficient indicates strong linear relationship. This helped identify potential relationships between continuous variables.

I have also performed Exploratory Data Visualization to visually explore and interpret the datasets. This has helped in uncovering patterns, and trends that were not apparent from summaries alone. Key characteristics of the dataset were identified using different plots which included bar charts, and distribution plots.

Finally, I also used Geospatial analysis to identify potential locations where a deeper analysis could be done.

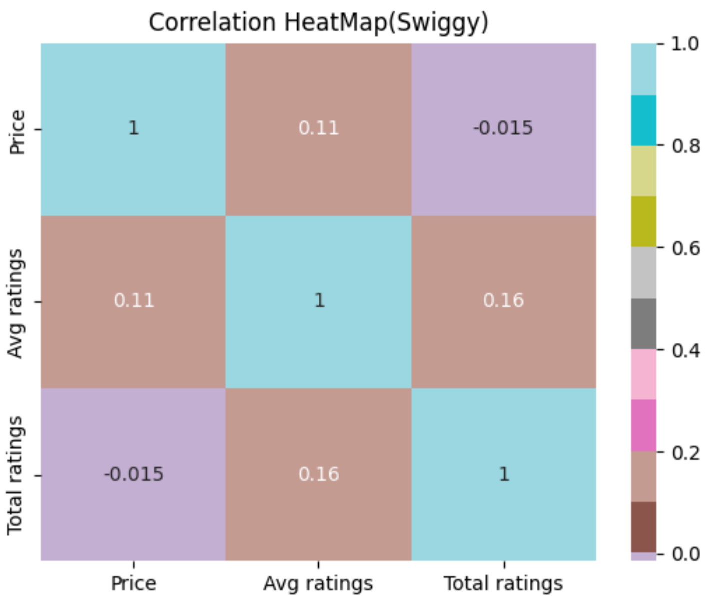
The entire analysis was performed on Kaggle, and using python. The various python libraries included NumPy and pandas for importing and manipulating the dataset, Matplotlib and seaborn to visualize the dataset. Geopy to extract latitude and longitude for various cities, and geopandas to plot and visualize the cities as a heatmap.

5 Results and Findings

Now that we have a better understanding of the data we are working with, we can start analyzing the data more thoroughly and start comparing the variables we have. This is done so we can identify any relationship between the variables. We primarily go for three statistical techniques which are Correlation Analysis, Exploratory Data Visualization, and Geospatial Analysis. Each analytical method serves to uncover different aspects of the restaurant industry, ultimately contributing to our comprehensive analysis of factors influencing customer satisfaction.

5.1 Correlation Analysis

We begin this section with correlation analysis, this statistical technique aims to measure the strength and direction of the relationship between variables. In our case, we have three numerical variables namely “Avg Rating”, “Total ratings”, and “Average Price”. These variables were selected based on their potential impact on customer satisfaction. By examining the correlation or heatmap between these variables, we aim to gain insights into the underlying patterns and dynamics within the restaurant industry.



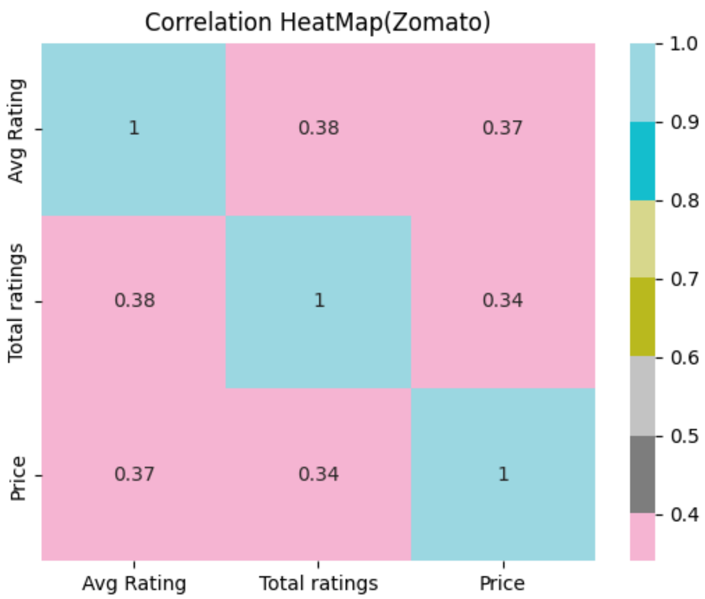


Fig 5: Correlation Heatmap

Observing the result of correlation analysis in Figure 5, we can see a notable difference between the strengths of relationships between variables across the two datasets. In the Zomato dataset, each variable exhibited a Pearson correlation coefficient of about 0.37 with each other, indicating a positive correlation. This indicates that there is a tendency for restaurants with higher average ratings to also have higher average costs and total number of ratings, and vice versa. However, in the case of Swiggy, the correlation coefficients are much weaker and closer to 0. This indicates a weaker relationship between the variables, suggesting that the variables may not be as strongly associated with each other as they are in Zomato.

These differences observed between Zomato and Swiggy highlight the importance of considering dataset specific characteristics in data analysis. The correlations in the Zomato dataset suggest a more moderate relationship between average rating, average cost, and total number of ratings, whereas the weaker coefficients in Swiggy indicate a less straightforward relationship. As a result, further investigation is required to explore additional factors affecting customer satisfaction and restaurant performance. Additionally, due to low Pearson correlation coefficients, we forego conducting Regression analysis and instead go for exploratory data visualization.

5.2 Exploratory Data Visualization

Next, we try to leverage the power of visual representations to delve deeper into our data. Through a diverse range of visualizations, we aim to gain a comprehensive understanding of the distributional characteristics, relationship between variables, and spatial patterns present in the datasets.

5.2.1 Distribution of Average Rating based on Average Price

To begin this section, we try to identify a deeper relationship between Average Rating and Average Price. We try to understand if customers give a higher rating for a higher priced restaurant or vice versa. To verify this, we plot the Average Rating distribution for restaurants having average price less than or equal to the mean of the Average Price and comparing with distribution for restaurants having Average Price greater than the mean.

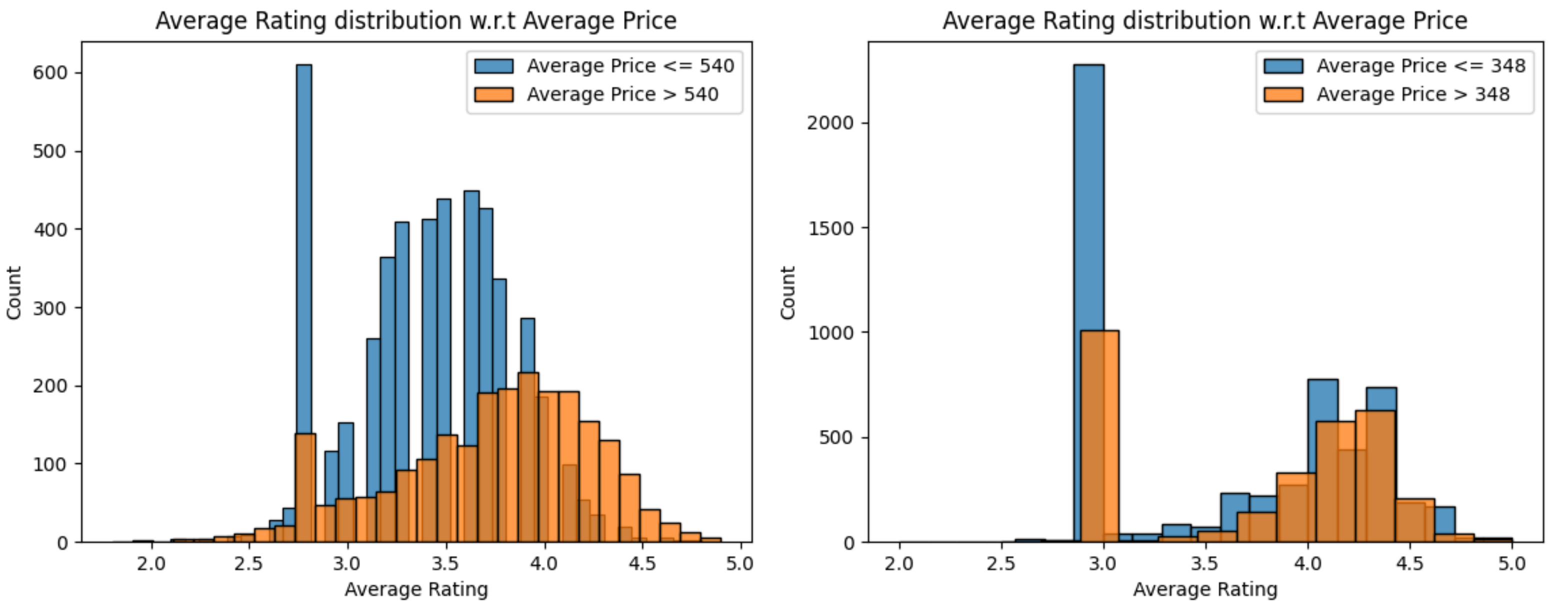


Fig 6: Average Rating distribution w.r.t Average Price(Zomato and Swiggy)

If we take a look at the visualizations in Figure 6 we observe interesting patterns in the data. For both restaurants having average price greater than and less than or equal to, there still seems to be a spike in rating at around 2.7. This simply indicates that in general, people tend to give restaurants a moderate rating. This pattern is prevalent in both Zomato and Swiggy.

However, when we compare the distribution for restaurants having average price greater than the mean, it is easy to observe that the average starts to shift a bit higher than the case for less than or equal to. This indicates a tendency for higher priced restaurants to receive comparatively higher average ratings. This also corresponds to the fact that we got a positive correlation coefficient between average rating and average price. Such findings may imply that customers perceive higher priced restaurants as having superior quality dining experience, thereby influencing their ratings.

The observed shift in the distribution of average ratings based on average price is an implication of the importance of pricing strategies in the restaurant industry and its impact on customer perceptions and satisfaction levels. This fact could be leveraged to optimize pricing strategies and enhance customer satisfaction.

5.2.2 Average Rating Distribution based on Online Ordering

Next, we examine the difference in distribution of average ratings for restaurants who offer online ordering versus restaurants who do not offer online ordering.

Fig 7: Average Rating Distribution for restaurants with and without online ordering

Taking a look at Figure 7 we observe a noticeable difference in the distributions. Restaurants offering online ordering seem to have a slightly higher distribution of average ratings compared to those without online ordering. This suggests that the availability of online ordering may positively influence customer satisfaction. This could also be attributed to the fact that once an order is placed, the app itself asks for a review. Whereas, if an online order is not placed the customer gets no such notification.

This observed difference in average rating distributions based on online ordering availability also implies the advantage of embracing technology and digital platforms in the restaurants. By offering convenient online ordering service, restaurants can enhance customer experience, attract more customers, and potentially increase their overall rating. This also highlights the strategic significance of incorporating online ordering capabilities into restaurant operations to remain competitive in today's digital marketplace.

5.2.3 Average Rating Distribution based on Table Booking

Next, we explore the distribution of average ratings for restaurants categorized based on the availability of table booking options. Specifically, we compare the distributions of average ratings for restaurants with table booking options to those without.

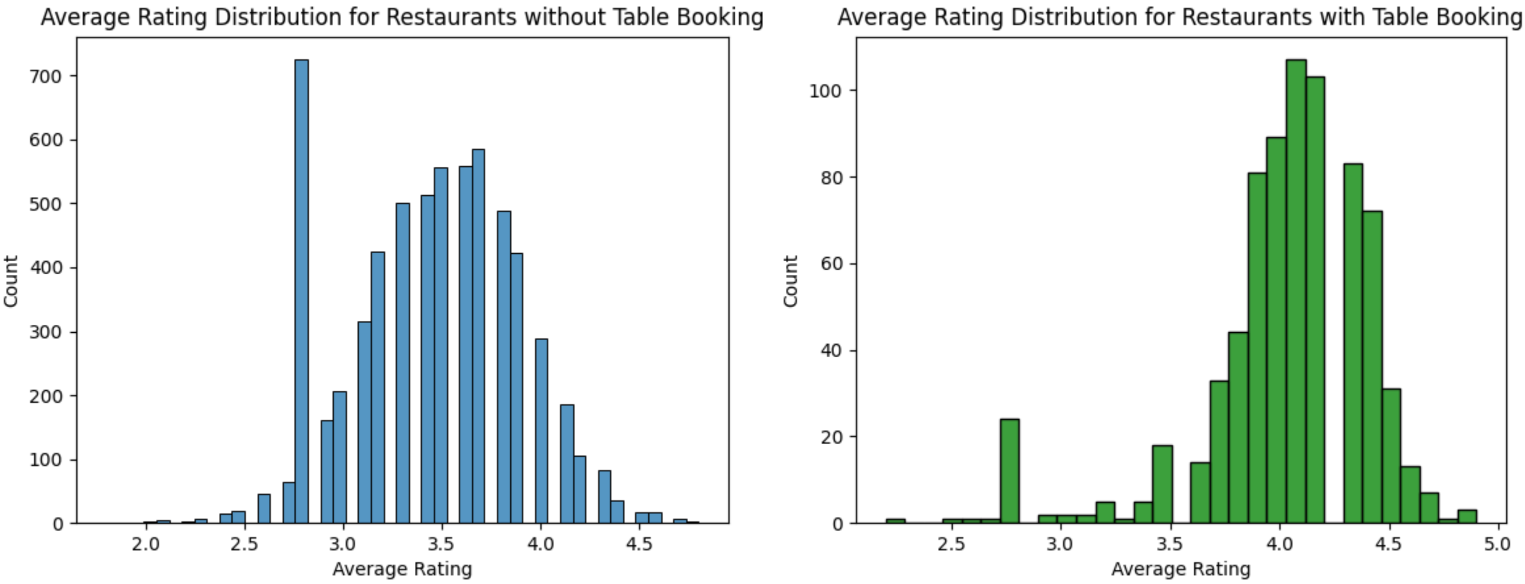


Fig 8: Average Rating Distribution for restaurants with and without Table Booking

Examining Figure 8, we observe a similar trend as seen in the case of online ordering availability. Restaurants that offer table booking exhibit a slightly higher distribution compared to those restaurants that do not offer table booking. This could imply that availability of table booking services may positively influence customer perception levels. We have already seen that a higher average rating corresponds to a higher average price, this relationship could be attributed to the availability of table booking, as upscale restaurants typically offer reservation services due to their larger space and need for table management.

This observed trend in average rating distributions based on table booking availability implies the strategic importance of providing convenient reservation options for customers. By offering booking services, restaurants can enhance their dining experience, and potentially improve their overall rating.

5.2.4 Top 10 Cuisines by Number of Ratings

Further, we try to identify and visualize the top 10 cuisines customers prefer. This ranking will primarily be based upon the number of ratings received, as it could be an indicator towards the popularity of that cuisine. We do not use Average rating for this, as the rating primarily depends upon the quality and quantity of the food and may not accurately reflect the popularity of cuisines.

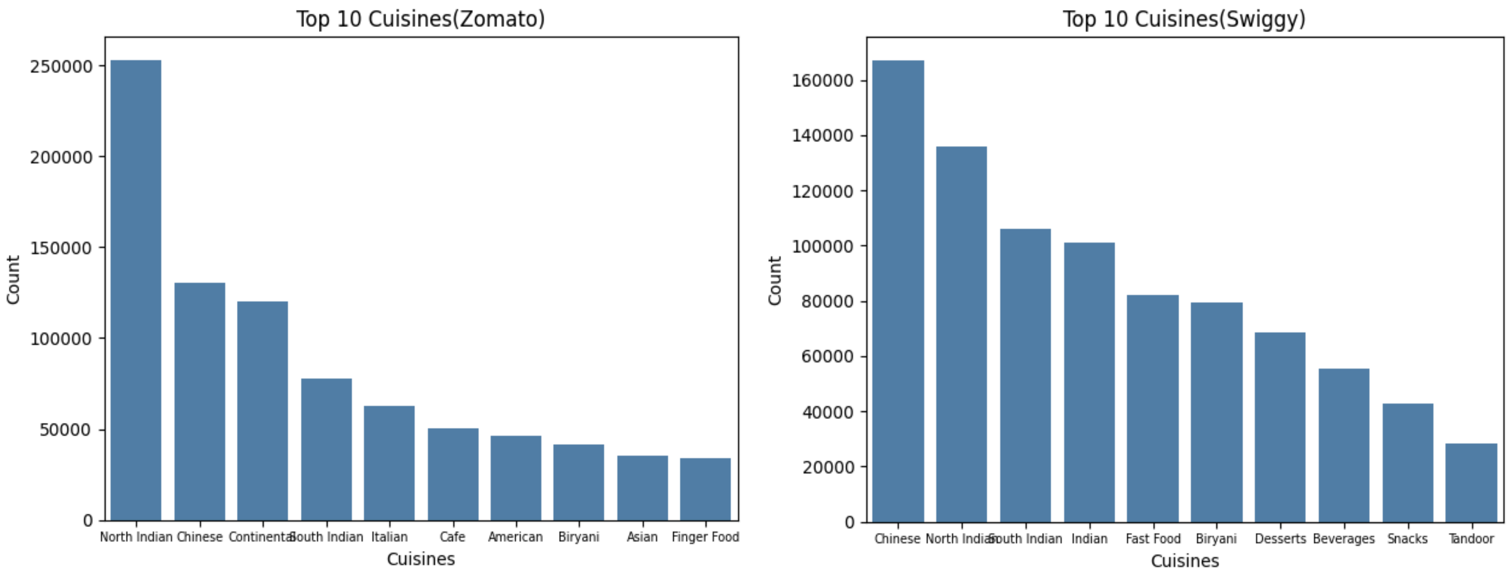


Fig 9: Top 10 Cuisines by Number of Ratings

Taking a look at Figure 9, we find that the most popular cuisines in both platforms differ slightly. We see that “North Indian” seems to be most popular, followed by “Chinese” in Zomato. Whereas “Chinese” seems to be most popular followed by “North India” in Swiggy. Another common popular cuisine is “South Indian”, which appears at rank 4 in Zomato and rank 3 in Swiggy.

By examining these findings, we can cater to customers' preferences by prioritizing the most popular cuisines in the restaurant's menu. However, it is important to note that these rankings may fluctuate periodically due to changing customer preferences. Hence, a periodic reassessing and updating the menu to ensure it meets the customer preferences may be required.

5.3 Geospatial Analysis

Finally, using our third dataset to perform geospatial analysis to identify potential locations where the average ratings is comparatively higher. The aim is to uncover regional variations in customer preferences and satisfaction.

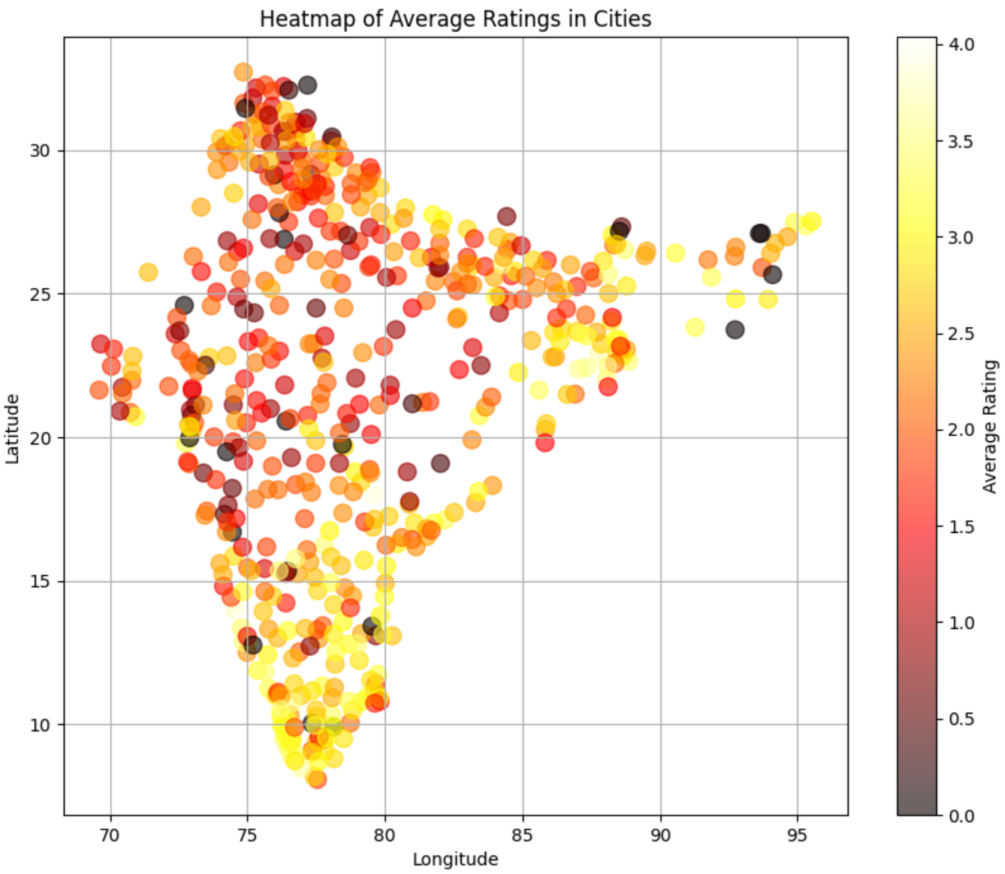


Fig 10: Heatmap of Average Ratings across cities

Now, we observe in Figure 10 a heatmap based upon the average rating across a total of 542 cities. Notably, the yellowish regions, corresponding to a higher average rating, seem to be in South India.

These findings offer insight towards potential states where a restaurant could be opened. By pinpointing areas with higher average ratings, we could conduct a deeper analysis in those locations and identify specific local addresses where the restaurants are likely to thrive.

5.4 Summary

In summary, we have performed correlation analysis, exploratory data visualization, and geospatial analysis. All these techniques gave us insight into the restaurant industry, customer satisfaction, and preferences.

First, we conducted correlation analysis and identified positive correlations between average rating, average price, and total number of ratings. We concluded that since the correlation coefficient, even though positive, was not high enough to conduct a Regression analysis.

Next, we investigated how the average rating distribution changed with different factors. We observed intriguing patterns in average rating distributions, with slight increases observed in correlation with average price and the presence of online ordering and table booking services. We also identified the most popular cuisines based upon the total number of ratings.

Finally, we performed geospatial analysis to identify potential locations where the restaurant could thrive. Using this analysis we can perform an even deeper analysis to pinpoint specific local addresses where these restaurants are likely to thrive.

6 Interpreting Results and Recommendations

Problem Statement 1: Investigating the Impact of Cuisine Type on Customer Preferences

- Result: The analysis revealed that which cuisine is most popular varies across which platform they use. Understanding this difference can help the restaurants tailor their menus towards customer preferences.
- Recommendation: Restaurant owners can leverage insights into popular cuisine preferences to optimize their menu offerings, develop targeted marketing campaigns, and enhance customer satisfaction.

Problem Statement 2: Assessing the Competitive Dynamics of the Restaurant Industry

- Result: In the competitive dynamics of the restaurant industry, a higher rating in different platforms is crucial. In general, restaurants with higher ratings are shown more frequently than restaurants with comparatively lower ratings.
- Recommendation: Restaurant owners should focus on providing online ordering services as well as table booking services, with an average or slightly higher price range to gain a higher average rating.

Problem Statement 3: Identifying Strategic Opportunities for Restaurant Expansion

- Result: Geospatial analysis uncovers strategic opportunities for restaurant expansion in regions with high growth potential. By mapping the total number of ratings across different geographic regions, restaurant owners can identify untapped opportunities for growth and market penetration.
- Recommendation: Restaurant owners can utilize these results to perform a more rigorous analysis and pinpoint specific locations for expansion.