Zomato Case Study

DESIGNING A MARKETING CAMPAIGN FOR A RESTAURANT CHAIN USING EXPLORATORY DATA ANALYSIS

A REPORT BY

MUHAMMAD TABISH SAMI

Table of Contents

About the Dataset

Case Study Overview

Objective

The purpose of this case study is to perform the Exploratory Data Analysis (EDA) to understand, explore, and pre-process the Zomato's dataset. In this case study, we performed EDA to understand the customer preferences, dining trends, and competitive landscape in various regions of India, and to design an effective marketing campaign for a restaurant chain.

Dataset

The dataset has been downloaded from Kaggle. It is a csv file named "**zomato_restaurants_in_India.csv**".

Context

The dataset contains information related to Zomato's restaurants, rating, votes, pricing, and other important variables. We'll use this dataset for Exploratory Data Analysis (EDA).

Data Analysis

Data Cleaning and Preparation

Basic Information

The dataset contains information about various Zomato's restaurants in India. It has 211944 observations in the dataset. Each row represents a unique restaurant with its various features such as location, votes, rating, pricing, establishments, and cuisines.
The summary of all features along it's datatype is as following:

- **res_id(int64):** Unique ID for each restaurant.
- **name(object):** Name of the restaurant.
- **establishment(object):** Type of establishment (e.g., Quick Bites, Casual Dining).
- **url(object):** URL to the Zomato restaurant page.
- **address(object):** Full address of the restaurant.
- **city(object):** City where the restaurant is located.
- **city_id(int64):** Numeric ID representing the city.
- **locality(object):** Locality within the city.
- **latitude(float64):** Latitude coordinate of the restaurant.
- **longitude(float64):** Longitude coordinate of the restaurant.
- **zipcode(object):** Zipcode of the restaurant's location.
- **country_id(int64):** Country ID (seems to be specific to India in this dataset).
- **locality_verbose(object):** More descriptive locality information.
- **cuisines(object):** Types of cuisines offered.
- **timings(object):** Operating hours of the restaurant.
- **average_cost_for_two(int64):** Average cost for two people (in local currency).
- **price_range(int64):** A coded value for price range (likely from 1 to 4).
- **currency(object):** Currency in which prices are listed (Rs. for Rupees).
- **highlights(object):** Key features or popular services of the restaurant (e.g., "Delivery", "No Alcohol Available").
- **aggregate_rating(float64):** Aggregated rating of the restaurant.
- **rating_text(object):** Text description of the rating (e.g., "Very Good", "Excellent").
- **votes:** Number of votes the restaurant has received.
- **photo_count(int64):** Number of photos posted on the restaurant's Zomato page.
- **opentable_support(float64):** Indicates if there's support for OpenTable bookings (0.0 or NaN).
- **delivery(int64):** Indicates if the restaurant offers delivery (-1, 0, 1).
- **takeaway(int64):** Indicates if takeaway is available (-1).

Handling Duplicate Values

The dataset contains **151,527** duplicate records. After removing duplicate records, it consists of **60,417** rows and **26** columns.

Missing Values

- **address:** 18 missing values.
- **zipcode**: 47,869 missing values, a significant number.
- **cuisines:** 470 missing values.
- **timings:** 1,070 missing values.
- **Opentable_support:** 19 missing values.

The column "**zipcode**" has most significant number of missing values up to **79.23**%. However, other columns have relatively low percentage of missing values.

Handling Missing Values

Removed the 'zipcode' column:

- This column was dropped from the dataset due to its high percentage of missing values.

Filled missing values in 'cuisines' and 'timings' columns:

- Missing values in the cuisines column were replaced with "Not Available".
- Missing values in the timings column were also filled with "Not Available".

Imputed Values in 'address' column:

- We have '**latitude**' and '**longitude**' column. Furthermore, we used '**geopy**' library to impute the missing values in this column.

Removed the 'opentable_support' column:

- The **'opentable_support'** column was dropped as it has not much importance due to single repeated value (i.e., 0).

After missing values imputation, the dataset contains **60417** observations and **24** features. The cleaned dataset is suitable and appropriate for the further analysis.

Handling Data Type Inconsistencies

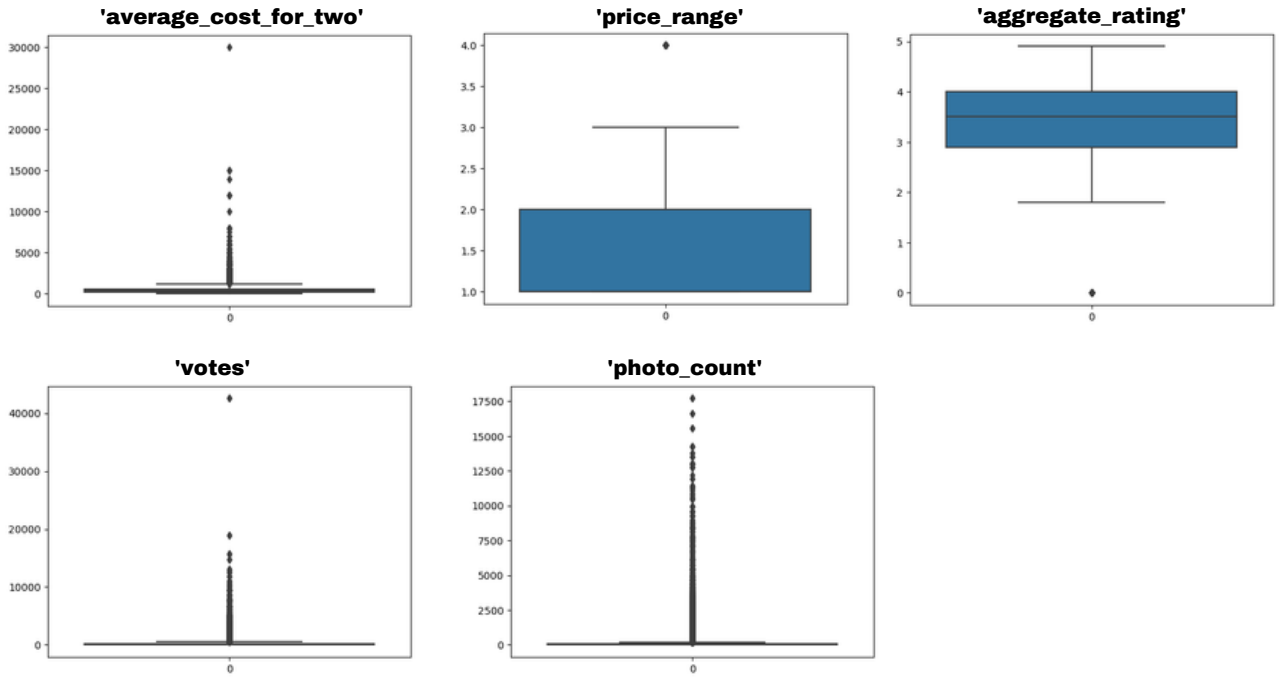
- **votes:** The **'votes'** had two negative values which were not logical and converted in to positive values.
- **timings:** Special characters **(i.e., â€“)** in timings column are replaced with **‘to’**.

Outliers Detection

Box Plots

The key numerical columns identified were 'Average Cost for Two', 'Price Range' , 'Aggregate Rating', 'Votes', and 'Photo Count'.

The box plot for these numerical columns are as following:



- **Average Cost for Two':** There are several outliers present, indicated by points far above the upper quartile.
- **'price_range':** This variable shows a few outliers. However, since price range is typically a categorical variable with limited discrete values, these 'outliers' might just represent higher-priced establishments.
- **'aggregate_rating':** There are some outliers, particularly at the lower end of the rating scale.
- **'votes':** This column shows a significant number of outliers, with some establishments having an exceptionally high number of votes.
- **'photo_count':** Similar to votes, there are many outliers with a very high photo count.

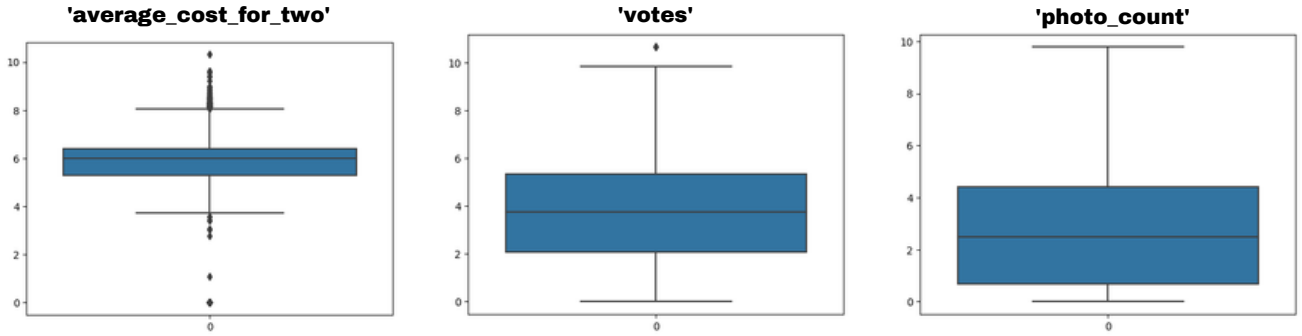
Handling Outliers

We'll apply the following steps to handle outliers:

- '**Average Cost for Two':** Applying a log to address the skewness.
- **'price_range':** Treating this column as a categorical column and leaving this column as it is.
- **'votes':** Apply a log transformation to reduce the skewness.

After handling outliers, the box plots determining the transformation of columns is as following:

Box Plots



From the transformations, we can observe the following:

- **Average Cost for Two (Original vs Log Transformed):** The log transformation reduces the skewness and brings extreme values closer to the median, making the distribution more symmetric.
- **Votes (Original vs Log Transformed):** Similar to the average cost, the log transformation of the votes significantly reduces the right-skewness and makes the distribution more uniform.
- **Photo Count (Original vs Log Transformed):** The transformation has a similar effect here, mitigating the impact of extreme high values and creating a more balanced distribution.

The transformations above helped in handling the outliers and normalizing the data distributions. This step is crucial for many statistical analysis and machine learning models that assume normally distributed inputs.

Exploratory Data Analysis (EDA)

Descriptive Statistics (Numerical Columns)

The descriptive statistics for the crucial numerical columns is as following:

	average_cost_for_two	price_range	aggregate_rating	votes	photo_count
count	60417.000000	60417.000000	60417.000000	60417.000000	60417.000000
mean	538.304517	1.730821	3.032868	261.575583	194.247414
std	593.852227	0.880462	1.440751	728.283944	705.682451
min	0.000000	1.000000	0.000000	0.000000	0.000000
25%	200.000000	1.000000	2.900000	7.000000	1.000000
50%	400.000000	1.000000	3.500000	42.000000	11.000000
75%	600.000000	2.000000	4.000000	207.000000	82.000000
max	30000.000000	4.000000	4.900000	42539.000000	17702.000000

Observation

- **'average_cost_for_two':** The average_cost_for_two varies significantly, indicated by a high standard deviation (**~593.85**) compared to its mean (**~538.30**), with values ranging from **0 to 30,000**, which suggests the presence of expensive outliers.
- **'price_range':** The price_range is a discrete variable with a small mean (**~1.73**) relative to its max value (**4**), indicating most data points are clustered in the lower price range.
- **'aggregate_rating':** The aggregate_rating shows moderate variation with a mean (**~3.03**) and a standard deviation (**~1.44**), spanning from **0 to 4.9**, which could point to a broad quality spectrum among the items rated.
- '**photo_count':** The photo_count has a mean of **~194.24** with a high standard deviation (**~705.68**), implying a skewed distribution with possibly a few entries having a very high number of photos.

After Outlier's Treatment

Descriptive Statistics (Numerical Columns)

The descriptive statistics for the crucial numerical columns after treating outliers is as following:

	average_cost_for_two	price_range	aggregate_rating	votes	photo_count
count	60417.000000	60417.000000	60417.000000	60417.000000	60417.000000
mean	5.869347	1.730821	3.032868	3.697038	2.803445
std	1.060853	0.880462	1.440751	2.134021	2.221778
min	0.000000	1.000000	0.000000	0.000000	0.000000
25%	5.303305	1.000000	2.900000	2.079442	0.693147
50%	5.993961	1.000000	3.500000	3.761200	2.484907
75%	6.398595	2.000000	4.000000	5.337538	4.418841
max	10.308986	4.000000	4.900000	10.658200	9.781489

Observation

- **'average_cost_for_two':** The average_cost_for_two has a mean of approximately **5.87** with a relatively small standard deviation of about **1.06**, which suggests costs are more standardized and variations between the costs are less extreme than before. The max value is now around **10.31**, indicating outlier costs have been removed or capped.
- **'aggregate_rating':** aggregate_rating has a mean of approximately **3.03** with a reduced max value of **4.9**, which suggests outlier treatment has moderated the extreme ratings.
- '**photo_count':** The photo_count has a mean of **~194.24** with a high standard deviation (**~705.68**), implying a skewed distribution with possibly a few entries having a very high number of photos.
- '**votes':** Votes show a mean of approximately **3.70**, which, together with a standard deviation of **2.13** and a reduced maximum value (**10.65** from the previous much larger value), indicates that outlier votes have been addressed, possibly through log transformation.

Descriptive Statistics (Categorical Columns)

The descriptive statistics for the categorical columns is as following:

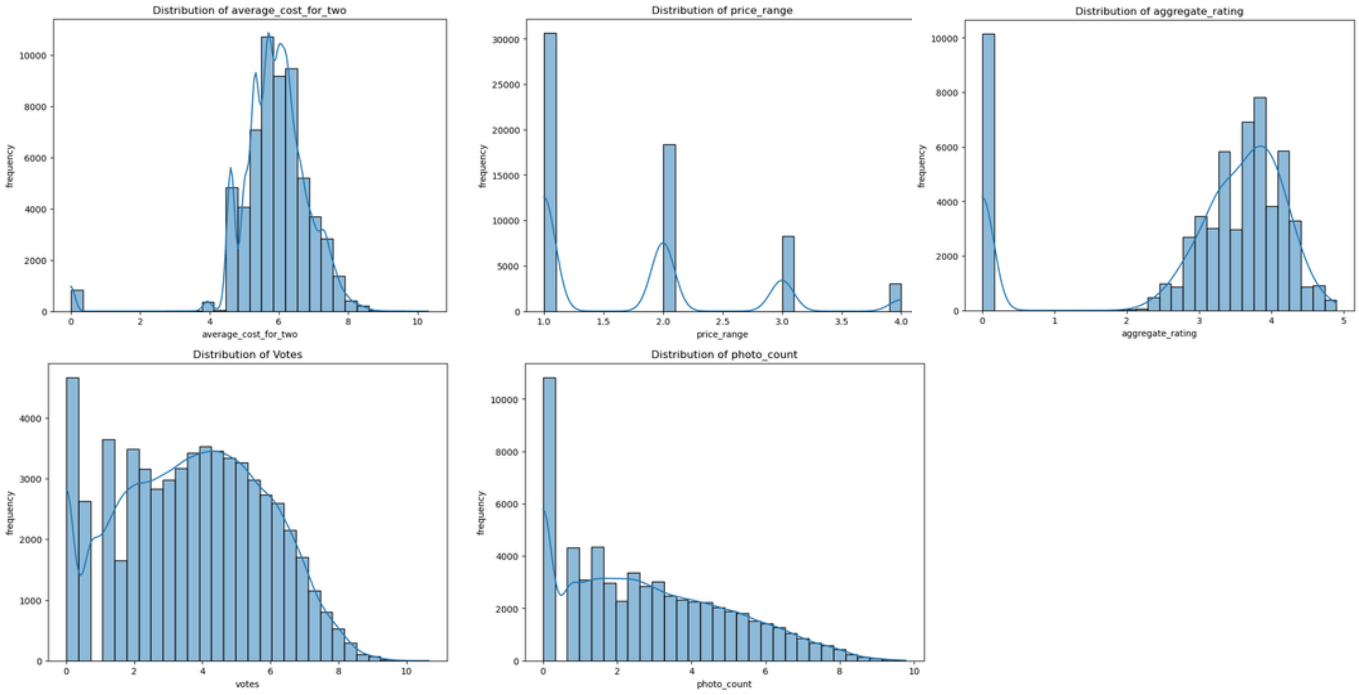
	name	establishment	url	address	city	locality	locality_verbose	cuisines	timings	currency	highlights	rating_text
count	60417	60417	60417	60417	60417	60417	60417	60417	60417	60417	60417	60417
unique	41100	27	55568	50668	99	3731	3910	9383	7741	1	31455	39
top	Domino's Pizza	['Quick Bites']	https://www.zomato.com/mumbai/candy-and-green-...	Laxman Jhula, Tapovan, Rishikesh	Chennai	Civil Lines	Gomti Nagar, Lucknow	North Indian	11 AM to 11 PM	Rs.	['Dinner', 'Takeaway Available', 'Lunch', 'Cas...	Good
freq	406	15477	9	37	2612	804	315	4587	7678	60417	925	17569

Observation

- **Name:** With **60,417** entries in the dataset and **41,100** unique names, we see a variety of establishments listed. "Domino's Pizza" appears most frequently, **406** times, which indicates it's a common chain in this dataset.
- **Establishment:** Out of the **60,417** records, there are **27** unique types of establishments. **'Quick Bites'** is the most common category, occurring **15,477** times, which suggests that fast-food or casual dining places are prevalent in this collection.
- **URL:** Each entry seems to have a unique URL pointing to its page, with **55,568** unique URLs. This suggests that most restaurants have their own page on the platform.
- **Address:** There are **50,668** unique addresses out of the total entries, and the top address is associated with **"Laxman Jhula, Tapovan, Rishikesh",** occurring **37** times**.**
- **City:** The dataset spans **99** cities, with **"Chennai"** being the most listed city, appearing **2,612** times.
- **Locality:** There are **3,731** unique localities, and **"Civil Lines"** is the top locality with 804 occurrences.
- **Locality Verbose:** This likely combines the city and locality information into a more descriptive location detail. **"Gomti Nagar, Lucknow"** is the most common, occurring **315** times.
- **Cuisines:** Out of **9,383** unique types of cuisines, **"North Indian"** is the most common, appearing **4,587** times. This indicates a preference or a larger number of North Indian cuisine offerings in the dataset.
- **Timings:** There's a wide variety of timings, with **7,741** unique entries. **"11 AM to 11 PM"** is the most common, appearing **7,678 times**, suggesting standard restaurant operational hours.
- **Highlights:** With **31,455** unique sets of highlights, this column represents features or services offered by the restaurants like **'Dinner', 'Takeaway Available', 'Lunch'**, which are frequent highlights, appearing **925** times. This variety indicates the dataset's richness in detailing what customers can expect from an establishment.
- **Rating Text:** There are **39** unique rating descriptors, with **"Good"** being the most frequent at **17,569** occurrences, suggesting that most restaurants are positively rated on the platform.

Data Distribution

Analyze the distribution of key variables (e.g., average cost for two, ratings, price range, cuisines).



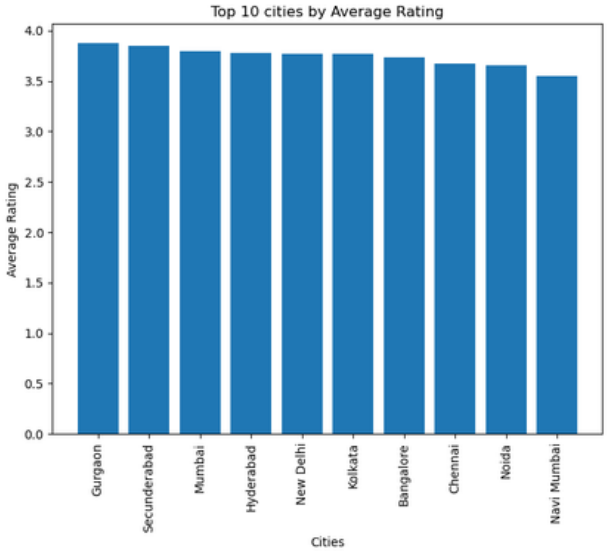
Observation

- **'average_cost_for_two':** The distribution or histogram shows that the data distribution is left-skewed.
- **'price_range':** The distribution or histogram shows that the data distribution has multimodal distribution with distinct peaks.
- **'aggregate_rating':** The distribution or histogram shows that the data distribution is left_skewed. It also shows that it has majority ratings falling at **3 to 4.5** with fewer restaurants having least ratings at **0 to 2.**
- **'photo_count': T**he distribution or histogram shows that the data distribution is slightly right_skewed.

Regional Analysis

Compare the restaurant trends and customer preferences across different cities or regions in India.

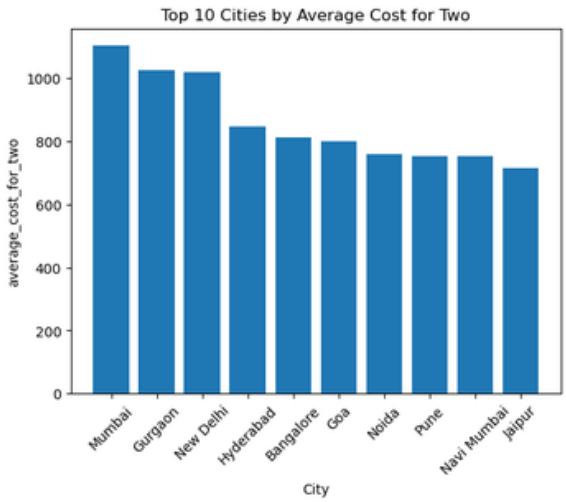
Top 10 Cities by Average Rating.



Observation:

The average rating for Gurgaon is highest at **3.87**, followed by Secundrabad at **3.85**, and Mumbai at **3.80.**

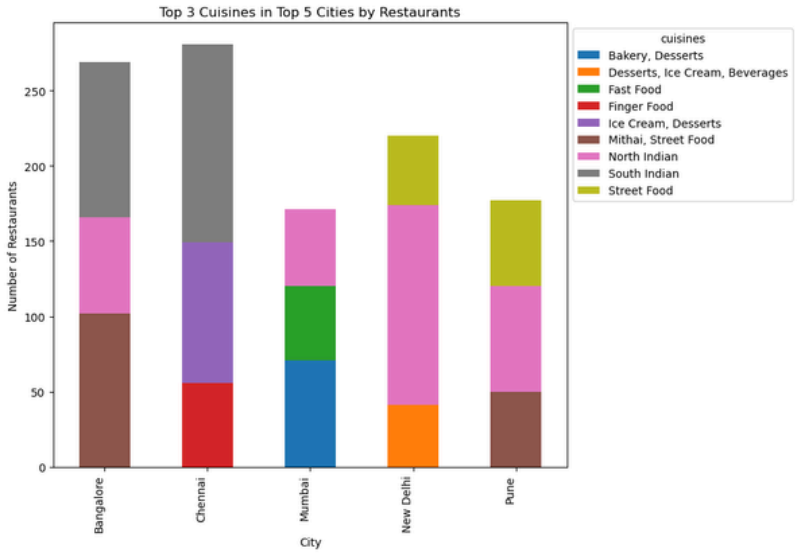
Top 10 Cities by 'average_cost_for_two'.



Observation:

The **average_cost_for_two** for **Mumbai** is highest at **1101.77**, followed by **Gurgaon** at **1025.384615**, and **New Delhi** at **1018.305360.**

Top 3 Cuisines in Top 5 Cities by Restaurants.

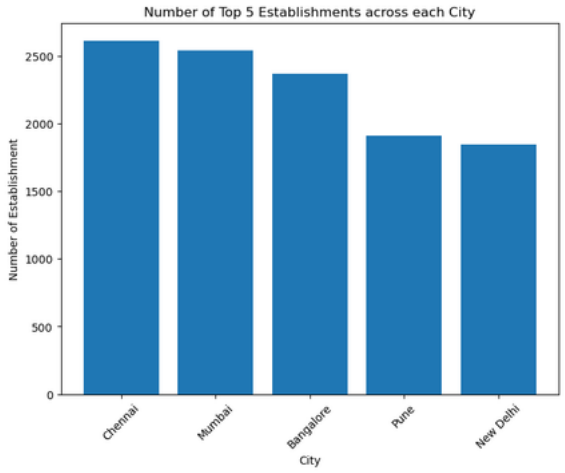


Observation:

Bangalore, Chennai, and New Delhi are Top 3 cities having top 3 cuisines as Mithai, Street Food, South Indian, and North Indian., and Finger Food, Ice Cream, Deserts, South Indian , and North Indian, Street Food, (Deserts, Ice cream, and Beverages)

Identify unique characteristics of the dining scene in each region.

Finding Number of Cities in each city to observe dining scene in each region.



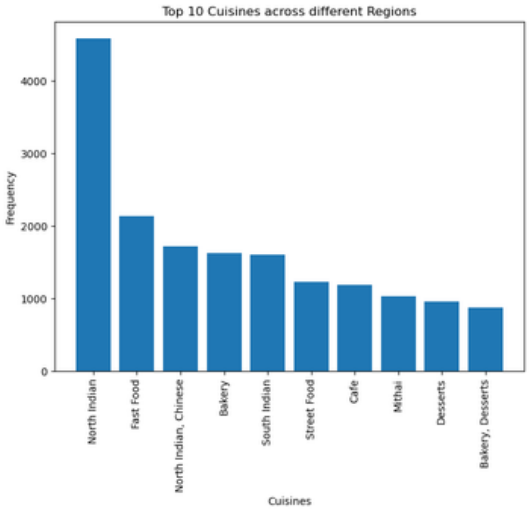
Observation:

Chennai is the leading city with highest number of establishments at **2612**, followed by **Mumbai** at **2538**, and **Bangalore** at **2365**.

Bangalore, Chennai, and New Delhi are Top 3 cities having top 3 cuisines as Mithai, Street Food, South Indian, and North Indian., and Finger Food, Ice Cream, Deserts, South Indian , and North Indian, Street Food, (Deserts, Ice cream, and Beverages)

Customer Preference Analysis

Analyze the types of cuisines that are popular in different regions.



Observation:

North Indian is the leading cuisine with highest frequency at **4587**, followed by **Fast Food** at **2137**, and **North Indian, Chinese** at **1720.**

Examine the relationship between restaurant ratings, price range, and popularity.

Plotting Correlation matrix to exmaine relationship between restaurant ratings, price range, and popularity.

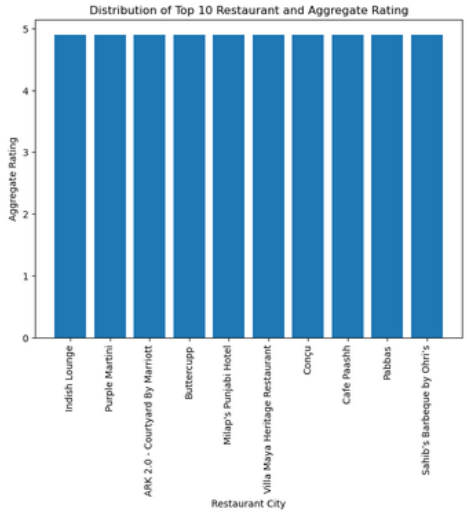
We will plot the correlation matrix to examine the relationship between different variable.

Bangalore, Chennai, and New Delhi are Top 3 cities having top 3 cuisines as Mithai, Street Food, South Indian, and North Indian., and Finger Food, Ice Cream, Deserts, South Indian , and North Indian, Street Food, (Deserts, Ice cream, and Beverages)

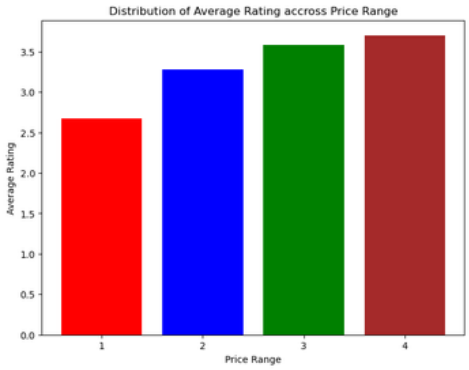
Observation:

Votes and aggregate_rating has the highest positive correlation at **0.79**, followed by **Votes and price_range** having **moderate positive correlation** at **0.36 aggregate_rating** and **price_range** has the **slightly positive correlation** at **0.26.**

Plotting Distribution of Restaurant and Aggregate Rating.



Plotting Distribution of Rating by Average Price Range.



Observation:

Price Range at **4** has the highest average rating at **3.7**, followed by Price Range at **3** at **3.57**, and Price Range at **1** has the least average rating at **2.67**

Bangalore, Chennai, and New Delhi are Top 3 cities having top 3 cuisines as Mithai, Street Food, South Indian, and North Indian., and Finger Food, Ice Cream, Deserts, South Indian , and North Indian, Street Food, (Deserts, Ice cream, and Beverages)

Competitive Analysis

Identify major competitors in each region based on cuisine, pricing, and ratings.

	city	cuisines	average_cost_for_two	aggregate_rating
1	Amritsar	Fast Food, Italian	500	4.9
2	Bangalore	Continental, North Indian, Chinese, European	2100	4.9
3	Bhubaneshwar	Tex-Mex, Fast Food	700	4.9
4	New Delhi	Asian	250	4.9
5	Mangalore	Ice Cream, Desserts, Beverages, Fast Food	1600	4.9
6	Thane	Modern Indian, North Indian, Chinese, Momos.	1000	4.9
7	Nashik	Continental, Indian, Chinese	700	4.9
8	Rajkot	North Indian, Gujarati, South Indian, Continental	600	4.8
9	Ajmer	Continental, Beverages, South Indian, Fast Food	200	4.8

Bangalore, Chennai, and New Delhi are Top 3 cities having top 3 cuisines as Mithai, Street Food, South Indian, and North Indian., and Finger Food, Ice Cream, Deserts, South Indian , and North Indian, Street Food, (Deserts, Ice cream, and Beverages)

	city	cuisines	average_cost_for_two	aggregate_rating
10	Allahabad	North Indian	200	4.8
11	Bhopal	Street Food, South Indian, Fast Food, Desserts...	400	4.8
12	Chennai	North Indian, European, Mediterranean, Contine...	1500	4.8
13	Navi Mumbai	Italian, Continental, Mexican	1600	4.8
14	Trichy	Arabian, Chinese, BBQ, Rolls	500	4.8
15	Chandigarh	European, Continental, North Indian	1600	4.7
16	Coimbatore	Biryani, South Indian	700	4.7
17	Nagpur	Cafe, Chinese, Fast Food	500	4.7
18	Patiala	Continental, Italian, Fast Food, Beverages	650	4.7
19	Surat	Beverages, North Indian	250	4.7

Bangalore, Chennai, and New Delhi are Top 3 cities having top 3 cuisines as Mithai, Street Food, South Indian, and North Indian., and Finger Food, Ice Cream, Deserts, South Indian , and North Indian, Street Food, (Deserts, Ice cream, and Beverages)

Analyze the strengths and weaknesses of these competitors.

SWOT Analysis

Strengths:

- **High Aggregate Ratings:** The aggregate rating for all competitors is high (**4.6 and above**), indicating strong customer satisfaction and positive word-of-mouth potential.
- **Diverse Cuisines:** Many restaurants serve multiple cuisines, which suggests versatility and an ability to cater to diverse tastes, potentially attracting a broader clientele.
- **Spread Across Various Cities:** The competitors are located in different cities, showing a wide market presence and the ability to appeal to different regional tastes.

Weaknesses:

- **Cost Difference:** There's a noticeable difference in pricing, which might reflect differences in quality, portion size, location, or target demographic. This could also imply inconsistency in customer experience across different price points.
- **Cuisine Concentration:** While diversity is a strength, some competitors are focused on a narrow range of cuisines **(e.g., North Indian or Fast Food)**, which could be a weakness if demand shifts towards other food trends.

Opportunities:

- **Expansion of Cuisine Options:** Competitors with a limited range could expand their menu to include more cuisines, tapping into unmet demands within their region.
- **Pricing Strategies:** Adjusting prices to fill gaps in the market could attract customers from competitors that are either too expensive or perceived as too cheap, thus capturing a different segment of the market.
- **Brand Differentiation:** By creating unique dining experiences, whether through ambiance, service quality, or specialty dishes, businesses can differentiate themselves from competitors in a crowded market.

Threats:

- **Intense Competition: T**he high ratings across the board imply that competition is fierce. New entrants and existing businesses must innovate constantly to maintain and grow market share.
- **Market Saturation:** In cities with multiple competitors offering similar cuisine, there's a risk of market saturation, making it difficult to attract new customers.
- **Economic Sensitivity:** Restaurants, especially those with higher average costs, might be vulnerable to economic downturns, which can significantly affect consumer spending habits.

Bangalore, Chennai, and New Delhi are Top 3 cities having top 3 cuisines as Mithai, Street Food, South Indian, and North Indian., and Finger Food, Ice Cream, Deserts, South Indian , and North Indian, Street Food, (Deserts, Ice cream, and Beverages)

Market Gap Analysis

Identify any gaps in the market that the restaurant chain can capitalize on (e.g., underrepresented cuisines, price ranges).

Cuisine Gaps:

- The most common cuisines across various regions are North Indian, Fast Food, and combinations of North Indian with Chinese.
- South Indian and Street Food also have a significant presence.
- Less frequent or underrepresented cuisines could represent a gap. For instance, international cuisines like Thai, Japanese, or specific regional Indian cuisines beyond North Indian and South Indian, such as Rajasthani, Goan, or North-Eastern, which could be explored.

Price Range Opportunities:

- Establishments with a wide range of **'average_cost_for_two'**, suggesting that there's competition across all price segments.
- However, the correlation analysis indicates a positive relationship between price range and aggregate rating, implying that higher-priced restaurants tend to have better ratings.
- If there is a concentration of competitors in the low to mid-range prices, there might be an opportunity in the higher price segment where the competition might be less, especially in cities where the **'average_cost_for_two'** is lower.

Regional Variations:

- The top cities by average rating are **Gurgaon, Secundrabad, and Mumbai**, suggesting these locations may already have a competitive market for high-quality dining experiences.
- Conversely, cities with lower average ratings might be ripe for high-quality establishments.
- Additionally, there is an opportunity in regions where the 'average_cost_for_two' is high but the variety of cuisines is not as broad.

Competitive Analysis in Key Cities:

- **Bangalore, Chennai, and New Delhi** are top cities with a focus on Mithai, Street Food, South Indian, and North Indian cuisines.
- The presence of these cuisines is strong, indicating that a new cuisine or a different take on these popular cuisines could be an interesting entry point.
- A restaurant chain that offers unique or less common cuisines with a focus on quality could fill a potential gap.

Bangalore, Chennai, and New Delhi are Top 3 cities having top 3 cuisines as Mithai, Street Food, South Indian, and North Indian., and Finger Food, Ice Cream, Deserts, South Indian , and North Indian, Street Food, (Deserts, Ice cream, and Beverages)

Designing the Marketing Campaign

Based on the insights from the above analyses, design a marketing campaign.

Regional Strategies:

- For regions with high 'average_cost_for_two' but less variety, such as **Mumbai and Gurgaon**, introduce fine dining with underrepresented cuisines or fusion concepts that mix local tastes with international flavors.
- In cities with lower average ratings, ensure high-quality food and service, and market these as the top reasons to try the new chain.
- Leverage regional preferences by incorporating popular local ingredients into new dishes for authenticity and familiarity.

Customer Segmentation:

- Target food enthusiasts by partnering with local food bloggers and influencers to showcase the unique offerings.
- Appeal to young professionals in tech hubs like Bangalore with modern, fast-casual dining options that offer healthy, quick meals.
- In family-centric localities, emphasize a family-friendly environment with value combo meals.

Differentiation from Competitors:

- Emphasize unique selling points (USPs) such as "First to Market" dishes, chef specials, or sustainable and organic ingredients.
- Highlight stories of the chefs, the origins of dishes, and the culinary journey to create a connection with foodies looking for authenticity.

Promotional Tactics:

- **Discounts:** Offer introductory discounts or special pricing during off-peak hours to attract initial customers and encourage them to spread the word.
- **Loyalty Programs:** Implement a loyalty program that rewards frequent visitors with points redeemable against future meals or special food experiences (e.g., cooking classes with chefs).
- **Special Events:** Host events such as themed dining nights, food festivals, or cooking demonstrations to create buzz and engagement.
- **Collaborations:** Partner with local businesses or events to offer catering or pop-up dining experiences to raise brand awareness.
- **Online Engagement:** Run a social media campaign encouraging customers to post their dining experiences. Create a hashtag campaign around the "Discover the Undiscovered" theme to track engagement.

Implementation:

- Roll out the campaign in phases, starting with cities where the potential for differentiation is highest.
- Measure the campaign's success through metrics such as customer acquisition rates, average spend per visit, social media engagement, and loyalty program enrollment.

Continuous Improvement:

- Regularly collect and analyze customer feedback to refine the menu, service, and overall dining experience.
- Stay flexible and be ready to adapt the campaign based on what is resonating with customers in different regions.