zomato

EXPLORATORY DATA ANALYSIS

ANSHUM BANGA
WWW.GITHUB.COM/ANSHUMBANGA

PROJECT OVERVIEW

ZOMATO EXPLORATORY DATA ANALYSIS

About Zomato,
Zomato is an Indian startup that has grown into one of the leading food delivery and restaurant discovery platforms globally. Founded in 2008 by Deepinder Goyal and Pankaj Chaddah, the company started as an online restaurant discovery platform in Delhi, India. Over the years, it has evolved and expanded its services to become a comprehensive food-tech platform.

Goal of the Project:

The primary goal of my Zomato Exploratory Data Analysis (EDA) project is to gain comprehensive insights into the dataset provided by Zomato. Through meticulous exploration and analysis of the data, I aim to uncover patterns, trends, and meaningful information that can provide valuable perspectives on the restaurant industry, customer preferences, and the overall dynamics of the food delivery market.

DATA OVERVIEW

Dataset Description:

The dataset comprises information related to restaurants listed on Zomato, collected through web scraping in the year 2022. The dataset includes the following columns:

- name: The name of the restaurant.
- establishment: The type or establishment category of the restaurant (e.g., cafe, fine dining, etc.).
- url: The URL or link to the restaurant on the Zomato platform.
- address: The physical address of the restaurant.
- city: The city where the restaurant is located.
- city_id: An identifier for the city.
- locality: The specific locality or neighborhood where the restaurant is situated.
- latitude: The geographical latitude coordinate of the restaurant's location.
- longitude: The geographical longitude coordinate of the restaurant's location.

DATA OVERVIEW

Dataset Description:

- zipcode: The postal code or ZIP code of the restaurant's location.
- locality_verbose: A detailed description of the locality, providing additional context.
- cuisines: The types of cuisines offered by the restaurant.
- timings: Operating hours or timings of the restaurant.
- average_cost_for_two: The average cost for dining for two people at the restaurant.
- price_range: A numeric indicator of the price range (e.g., 1 for low, 4 for high).
- highlights: Noteworthy features or services offered by the restaurant.
- aggregate_rating: The overall rating of the restaurant based on user reviews.
- rating_text: A text representation of the restaurant's rating (e.g., "Excellent").
- votes: The number of user votes or reviews received by the restaurant.
- photo_count: The count of photos associated with the restaurant.
- delivery: A binary indicator of whether the restaurant offers delivery services.
- takeaway: A binary indicator of whether the restaurant offers takeaway services.

Dataset Source:

The dataset has been obtained through web scraping methods, capturing data from Zomato's platform in the year 2022. It provides a snapshot of information about a diverse range of restaurants, allowing for a comprehensive analysis of the restaurant landscape, customer preferences, and operational aspects within the food industry. The variety of columns in the dataset enables a multifaceted exploration, contributing to a rich and detailed understanding of the Zomato ecosystem.

Dataset Size:

The dataset used for your Zomato Exploratory Data Analysis (EDA) project is substantial, containing a total of 211,944 rows and 26 columns. This indicates a rich and extensive collection of information regarding various aspects of restaurants listed on Zomato in the year 2022.

Missing Value Treatment:

1. Address Column:
- Identified 134 inconsistencies in the address column.
- No discernible patterns for meaningful replacement.
- Pragmatic approach: Used "Unknown" as a placeholder for inconsistencies in the address column.

2. Zip Code Columns:
- Detected 163,187 inconsistencies in zip code-related columns.
- In some cases, localities had both null and assigned values in the zip code column.
- Approach: Determined the mode of non-null zip codes associated with localities and replaced null values with this mode.

3. Cuisines Column:
- Found 1,391 inconsistencies in the cuisines column.
- Absence of discernible patterns for replacement.
- Solution: Replaced null values with "No Description" as a placeholder.

4. Timings Column:
- Identified 3,874 inconsistencies in the timings column.
- 'Timings' column includes strings with both timings and days of the week.
- Approach: Addressed inconsistencies individually as needed.

5. OpenTable Support Column:
- 'OpenTable Support' column had all zero values, indicating no support.
- Approach: Filled null values with zero for consistency.

These strategies for handling missing values aim to maintain data integrity while acknowledging the specific characteristics of each column. Pragmatic solutions, such as using "Unknown" or "No Description" as placeholders, were employed where discernible patterns or meaningful replacements were challenging to identify.

Outlier Detection:

1. Unusual Average Cost for Two:
   - Identified an outlier in the average cost for two, with a recorded value of 30,000.
   - Research suggests a data entry error; the correct value is likely 3,000.
   - Corrected the outlier to ensure data accuracy.

2. High Votes Entry:
   - Noticed an entry with an exceptionally high number of votes, exceeding 40,000.
   - Research indicates that votes under 20,000 are within a reasonable range.
   - Considered the entry with over 40,000 votes as a likely false entry and corrected it to 4,000 for data consistency.

3. Reputable Hotel Names:
   - Observed names associated with reputable hotels.
   - Research suggests that the observed average cost for two may be justified due to the nature and prestige of these establishments.
   - No corrective action taken as the data aligns with the nature of high-end dining experiences.

The outlier detection process revealed minimal anomalies in the dataset. Corrections were made where data entry errors were evident, ensuring the accuracy and reliability of the information. Research played a crucial role in justifying certain outliers, particularly when associated with reputable establishments or when values fell within acceptable ranges based on the nature of the data. Overall, the dataset is considered acceptable with minimal outliers, contributing to the robustness of subsequent analyses.

Feature Selection:

1. Dropping "country_id" Column:
   - Decision: The "country_id" column is unnecessary since all the data is from the same country.
   - Action: The "country_id" column will be dropped for simplification.

2. Disregarding "res_id" Column:
   - Observation: Significant duplicacy in the "res_id" column, resembling a region-specific bubble.
   - Decision: "res_id" will be disregarded, and index numbers will be used for clarity.

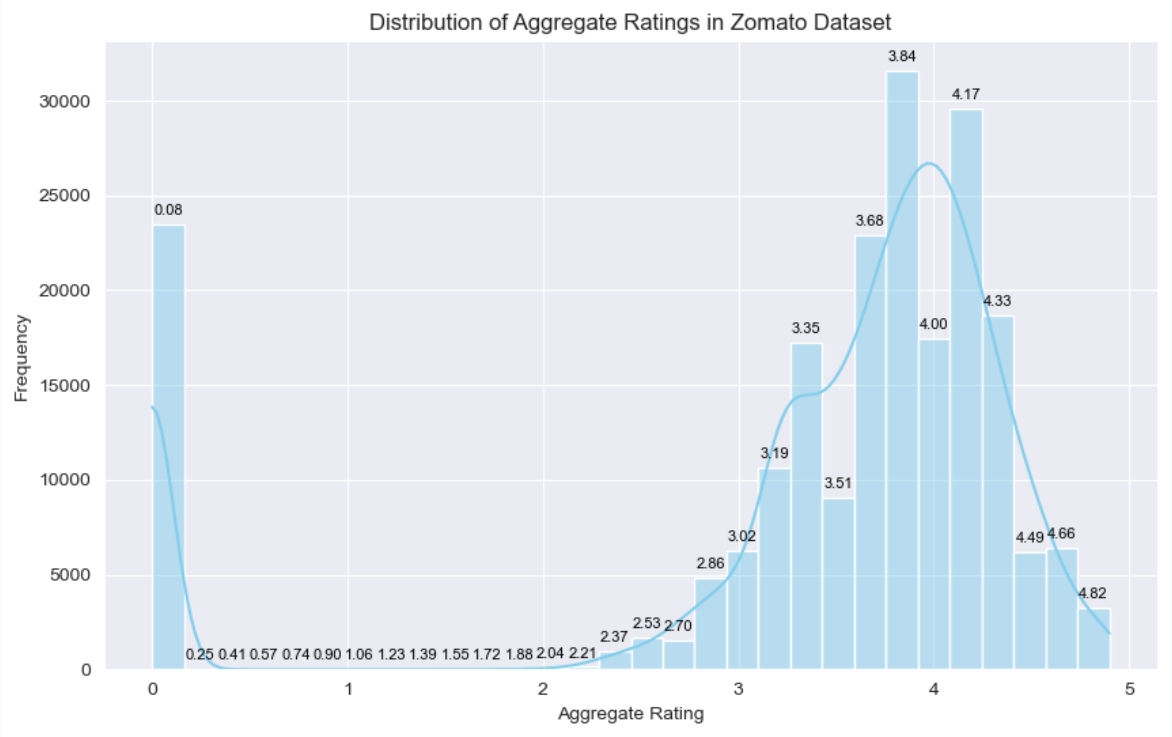
3. Removing "currency" Column:
   - Rationale: Country consistency and the exclusive use of Rs./INR make the "currency" column redundant.
   - Action: The "currency" column will be dropped for simplicity.

4. Dropping "opentable_support" Column:
   - Observation: The "opentable_support" column consistently holds zero values across all rows.
   - Decision: This column does not provide valuable insights and will be dropped for efficiency.

The feature selection process aims to enhance the simplicity, clarity, and efficiency of the dataset by removing redundant or non-informative columns. By dropping these columns, the remaining features become more focused and relevant to the analysis, contributing to a streamlined and effective exploratory data analysis (EDA).

Distribution of Aggregate Ratings in Zomato Dataset:

The distribution of aggregate ratings in the Zomato dataset displays a diverse range, with a concentration around the average. This reflects varied user perceptions of restaurant experiences within the platform.



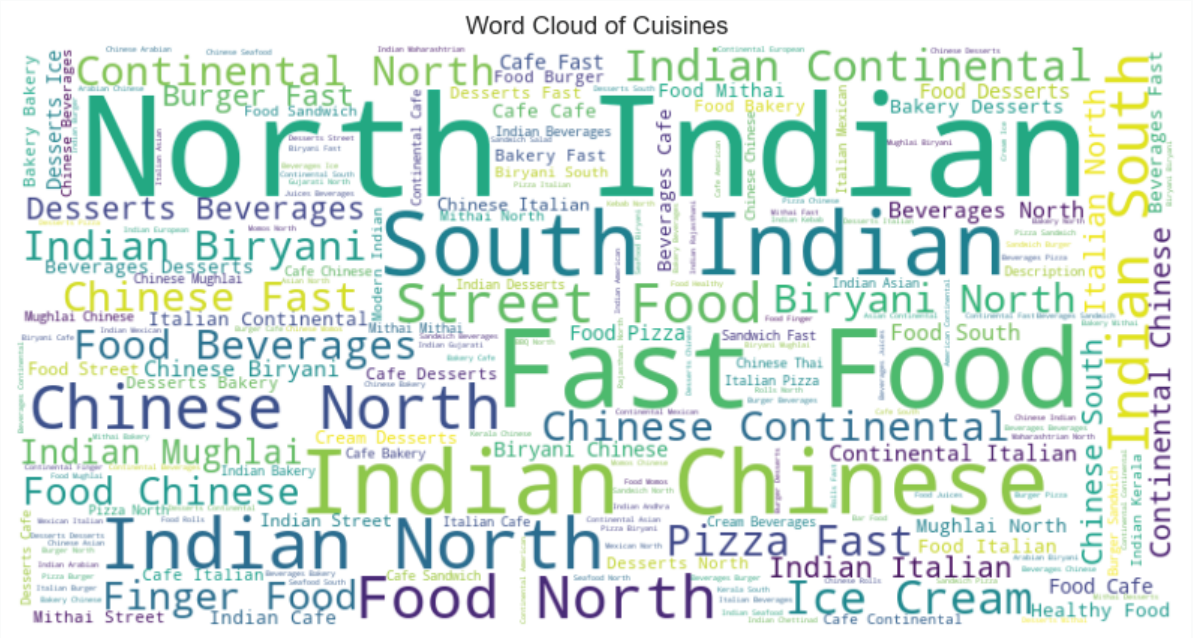
Observations and Insights:

- A noteworthy observation reveals that a substantial majority of restaurants in the dataset attained a **rating peak at 3.84.**
- The rating trend unfolds with a gradual ascent from 2.37 to 3.19, exhibiting a slight spike at 2.86. Subsequently, a prominent surge is observed at 3.35, followed by a pronounced decline around the 3.5 mark.
- The dataset indicates that the **second-highest rating attained by restaurants is 4.1.** Following this peak, a decline in ratings is notable. **It's observed that achieving a rating of 4.81 is challenging for a substantial number of restaurants.**

The challenge of achieving a **rating of 4.81** suggests that only a select **few restaurants excel in meeting or surpassing these higher expectations, possibly requiring qualitied food, exceptional offerings or service to attain such a rating.**

WordCloud of cuisines from the 'cuisines' column

The WordCloud generated from the 'cuisines' column in the Zomato dataset visually represents the diversity of cuisines offered by restaurants. It provides a quick and intuitive overview of popular culinary options.



Observations and Insights:

1. **North Indian Cuisine Dominance:**
   - The word cloud prominently features "North Indian," indicating a significant demand for this cuisine, showcasing **North India as a major market for Zomato.**

2. Indian Chinese and Tibetan Influence:
   - The inclusion of "Indian Chinese" suggests a considerable demand, possibly influenced by Tibetan cuisine, particularly popular in North India.

3. South Indian Appeal Across Regions:
   - "South Indian" stands out, indicating substantial demand not only in the South but also in the North, **reflecting Zomato's nationwide presence.**

Observations and Insights:

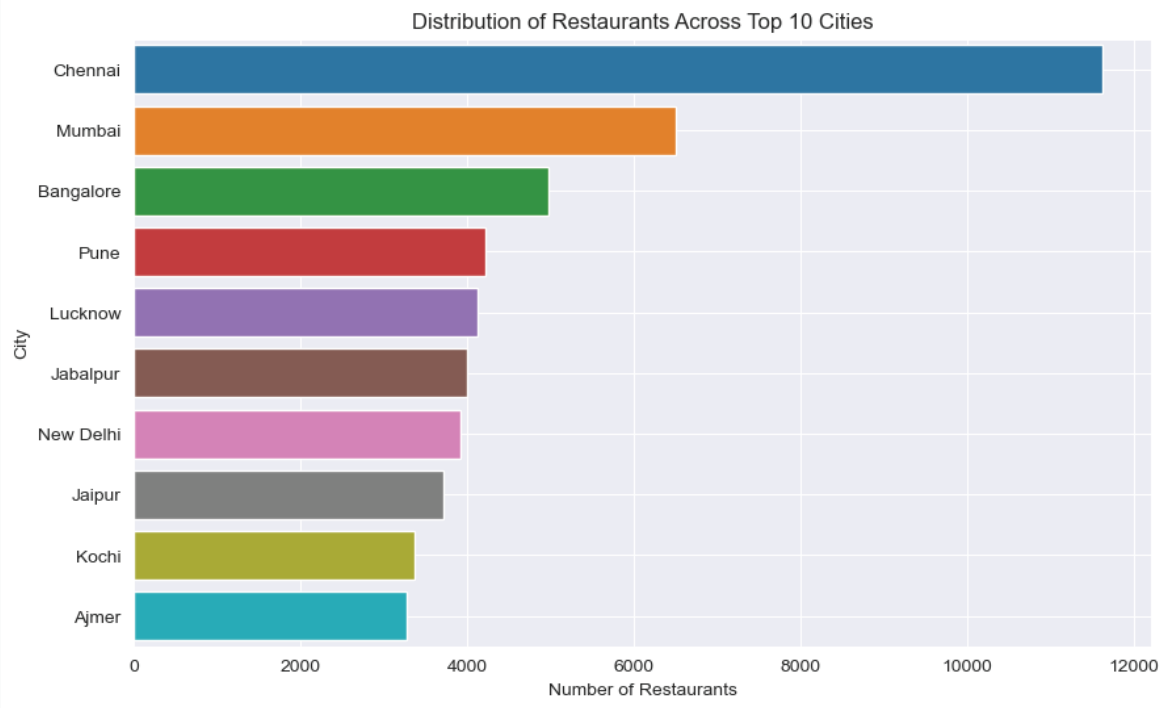
4. Diverse National Presence:
- The word cloud showcases nationwide preferences with mentions of "Indian Biryani," "Pizza," "Burger," "Fast Food," "Mughlai," and "Ice Cream," indicating the diverse and widespread appeal of these cuisines across the country.

5. Opportunities for Feature Promotion:
- The larger representations of certain foods like **Biryani, Pizza, Burger, and Ice Cream suggest their popularity.** Zomato could strategically **promote these items through carousels or featured sections on their platform,** capitalizing on their mass appeal.

6. Strategic Business Insights:
- Zomato can leverage these insights to enhance user experience, tailor marketing strategies, and strategically position popular cuisines to maximize engagement and satisfaction across the diverse culinary landscape of India.

Distribution of Restaurants Across Top 10 Cities:

The distribution of restaurants across the top 10 cities in the Zomato dataset showcases the varied culinary landscape in different urban centers. The analysis reveals the concentration and diversity of dining establishments, offering insights into regional dining preferences and market dynamics..



Observations and Insights:

1. Chennai Dominance:
- Chennai leads with the highest number of listed restaurants, suggesting a robust market presence.

2. South Indian Cities in Top Positions:
- Mumbai, Bangalore, and Pune, all from the southern region, follow Chennai. This contradicts the initial assumption of a major appeal in the North, emphasizing the widespread popularity of North Indian food in both regions.

3. Mass Appeal of North Indian Cuisine:
- The combined observations from the word cloud and countplot suggest a high demand for North Indian cuisine, potentially attributed to its ease of preparation.

Observations and Insights:

4. Fast Food Appeal in South India:
- The significant presence of fast food establishments in South Indian cities, possibly catering to a younger demographic and migrants, **presents an opportunity for targeted promotions and pricing strategies.**

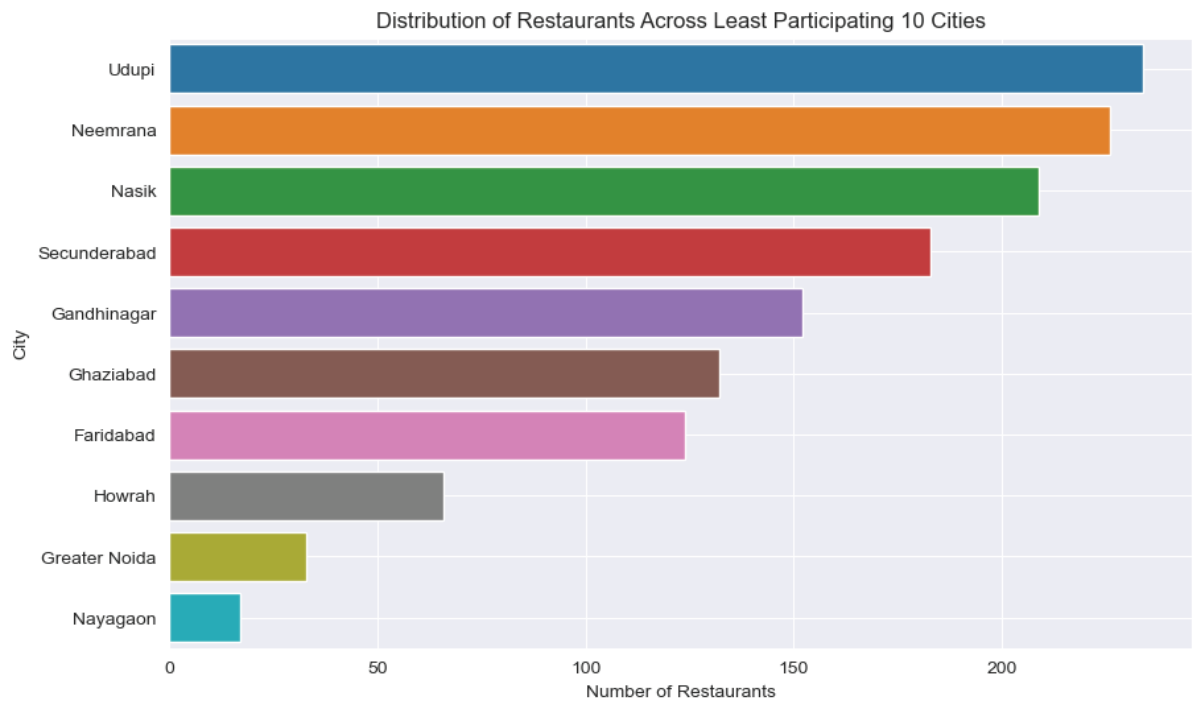
5. Lucknow's Biryani Dominance:
- Lucknow emerges as a major player in North India, indicating a high demand for Biryani, **potentially a focal point for prominent restaurants in the city.**

6. New Delhi's Remarkable Presence:
- Despite its smaller geographical size, New Delhi holds a noteworthy position, with over 3,800 listed restaurants, underscoring its importance as a key market.

These insights can guide strategic decisions for Zomato, such as tailoring offerings to regional preferences, adjusting pricing strategies, and targeting specific demographics to further enhance business in different cities.

Distribution of Restaurants Across Least Participating 10 Cities:

The distribution of restaurants across the least participating 10 cities in the Zomato dataset illustrates the sparse culinary landscape in these areas. It highlights lower restaurant density and suggests potential opportunities for market expansion.



Observations and Insights:

1. Nayagaon's Limited Presence:
- Nayagaon, representing the least number of listed restaurants, **can be attributed to its status as a small village near Pune.** Given Pune's already higher restaurant count, the limited presence in Nayagaon may be expected and could be reasonably neglected.

2. Greater Noida's Opportunity for Growth:
- Despite its name suggesting size, Greater Noida surprisingly holds the second-to-last position in terms of listed restaurants. **This presents an opportunity for Zomato to stimulate growth by incentivizing residents with offers and targeting residential areas.**

Observations and Insights:

3. Challenges in Ghaziabad, Faridabad, and Gandhinagar:
- Ghaziabad, Faridabad, and Gandhinagar, despite being notably larger in size, have a relatively low number of listed restaurants. Zomato could investigate and address potential challenges, encouraging restaurant partnerships, and focusing on marketing strategies targeting residential areas to boost its presence in these locations.

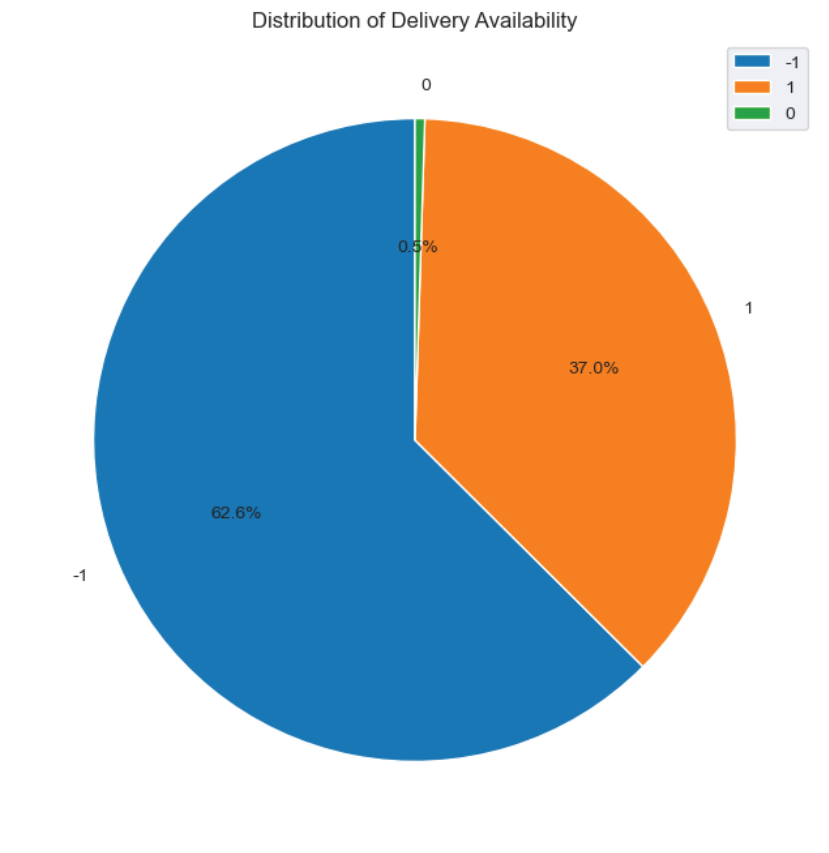
4. Udupi, Neemrana, Nasik and Secunderabad:
- Cities like Udupi, Neemrana, Nasik, and Secunderabad, with relatively fewer listed restaurants, offer opportunities for Zomato to enhance its presence through tailored strategies, including local collaborations, targeted marketing, and community outreach.

The identified areas for improvement underscore the significance of understanding local dynamics and tailoring strategies to specific geographical nuances. Implementing targeted marketing, promotional offers, and strategic collaborations can be instrumental in enhancing Zomato's presence and engagement in these regions.

Distribution of Aggregate Ratings in Zomato Dataset:

The pie chart depicting the distribution of delivery availability in the Zomato dataset reveals three distinct segments.

-1: Indicates a condition or status where delivery is not avail able or not applicable.
1: Represents a condition or status where delivery is available.
0: May represent a neutral or undefined condition, potentially used for cases where delivery information is not specified or not applicable.



Observations and Insights:

The pie chart analysis of the 'delivery' column unveils valuable insights

1. Majority without Delivery (62.6%):
- The substantial portion of restaurants (62.6%) not providing delivery services suggests the presence of reputed establishments or cafes. Zomato could strategically enhance its dine-in services by collaborating with these venues. Initiatives like special offers, featured promotions on the platform, and blog features can help promote and spotlight these restaurants.

Observations and Insights:

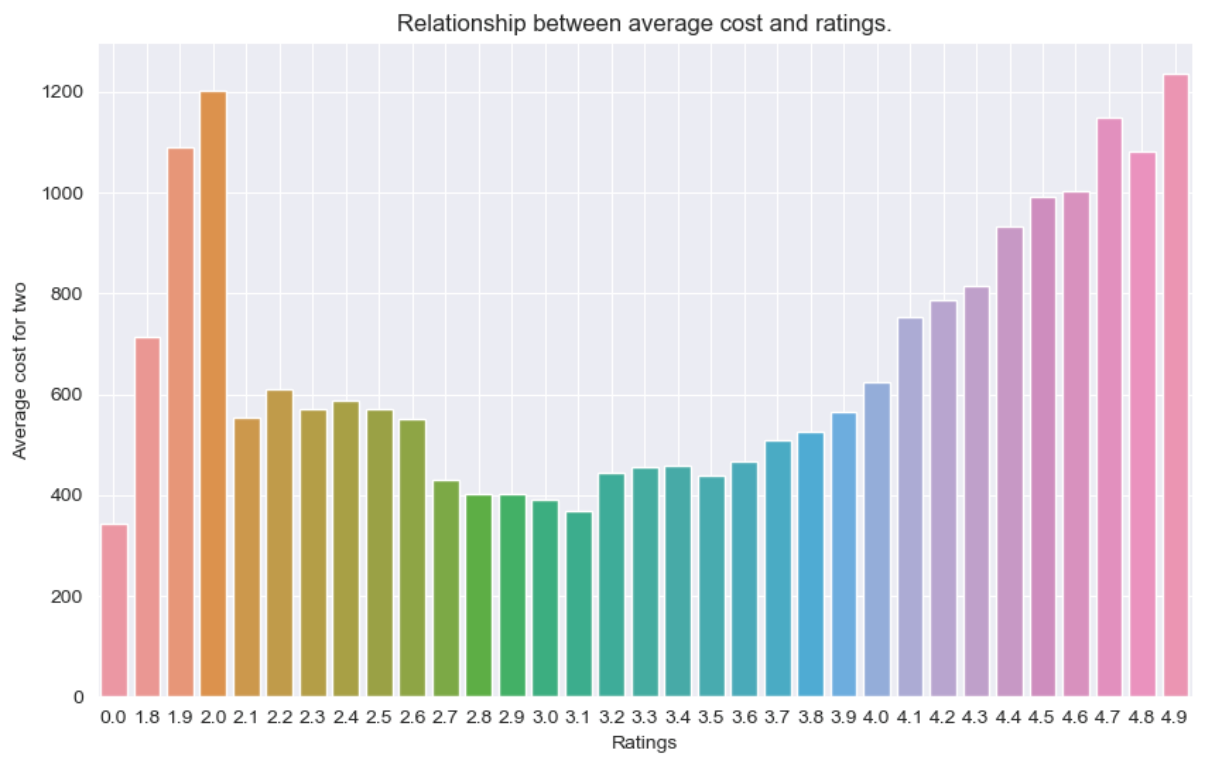
The pie chart analysis of the 'delivery' column unveils valuable insights

2. Restaurants Offering Delivery (37.0%):
- Restaurants providing delivery services represent a significant growth opportunity for Zomato. Promoting these establishments on the app, facilitating seamless deliveries, and offering exclusive deals could amplify customer engagement and contribute to the overall expansion of Zomato's delivery services.

Strategically tailoring approaches for both segments—focused marketing for dine-in experiences with non-delivery restaurants and robust promotion of delivery services for restaurants offering deliveries—can contribute to a balanced and successful growth strategy for Zomato.

Distribution of Aggregate Ratings in Zomato Dataset:

The relationship between average cost and ratings in the Zomato dataset suggests a nuanced correlation. While there may be patterns, individual preferences and restaurant characteristics contribute to diverse associations between cost and ratings.



Observations and Insights:

1. Competitive Focus at 1200 Average Cost:
- Regardless of the rating (2 stars or 4.9), significant competition is observed around an average cost per two of 1200.

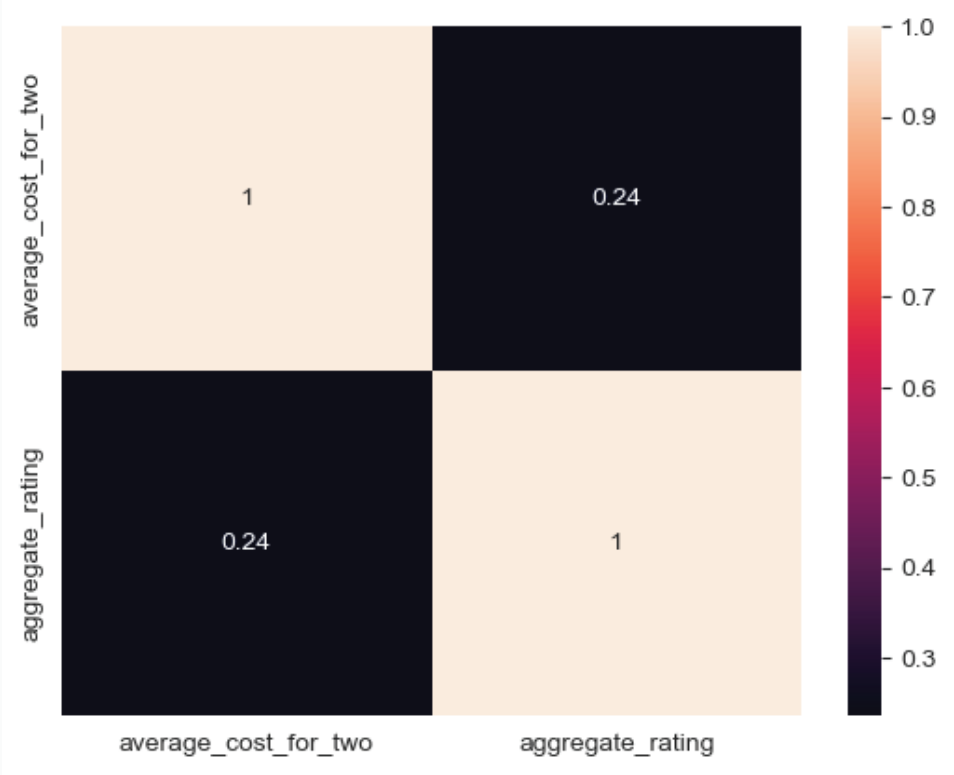
2. Customer Behavior Insight:
- The trend suggests that customers with higher expectations tend to choose restaurants with a higher average cost.

3. Impact on Ratings:
- Restaurants meeting or exceeding these expectations often receive close to 5-star ratings, while falling short may result in lower ratings (1, 2, or 3 stars).

4. Correlation with Perceived Value:
- The observed correlation highlights the connection between customer satisfaction, perceived value, and the pricing tier of higher-end establishments.

Heatmap of Aggregrate Ratings and Average cost of two:

The heatmap of aggregate ratings and average cost for two in the Zomato dataset visually displays the correlation between these variables.



Observations and Insights:

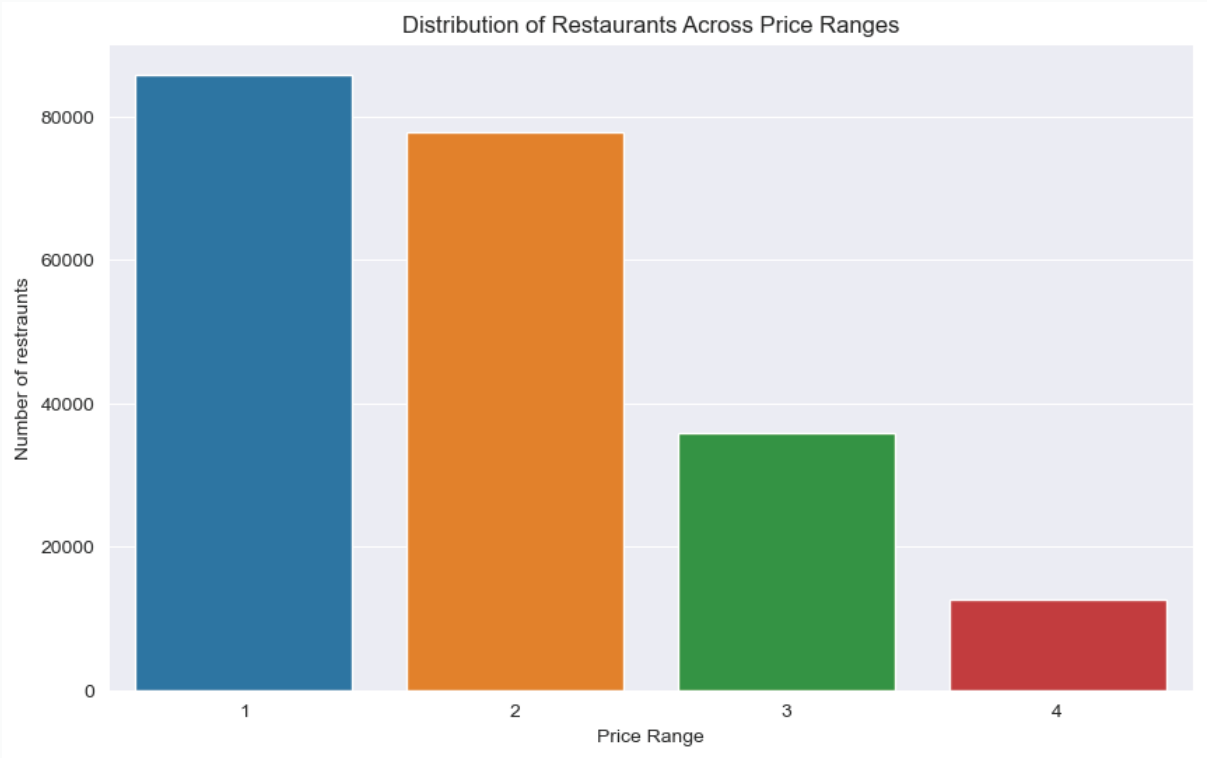
Heatmap suggests there is no significant correlation between 'average_cost_for_two' and 'aggregate_rating', which indicates that the two variables are not strongly linearly related. In other words, **the cost of a meal for two people does not have a consistent impact on the aggregate rating of a restaurant.**

This lack of correlation could be due to various factors. It's possible that customers prioritize other aspects such as food quality, service, ambiance, or specific cuisine types over the cost when assigning ratings.

Distribution of Restaurants Across Price Ranges:

The distribution of restaurants across price ranges (ranging from 1 to 4) in the Zomato dataset depicts the diversity of dining options. It offers a quick overview of pricing categories.

- **1:** Indicates a relatively lower price range.
- **2:** Suggests a moderate price range.
- **3:** Represents a higher price range.
- **4:** Indicates the highest or premium price range.

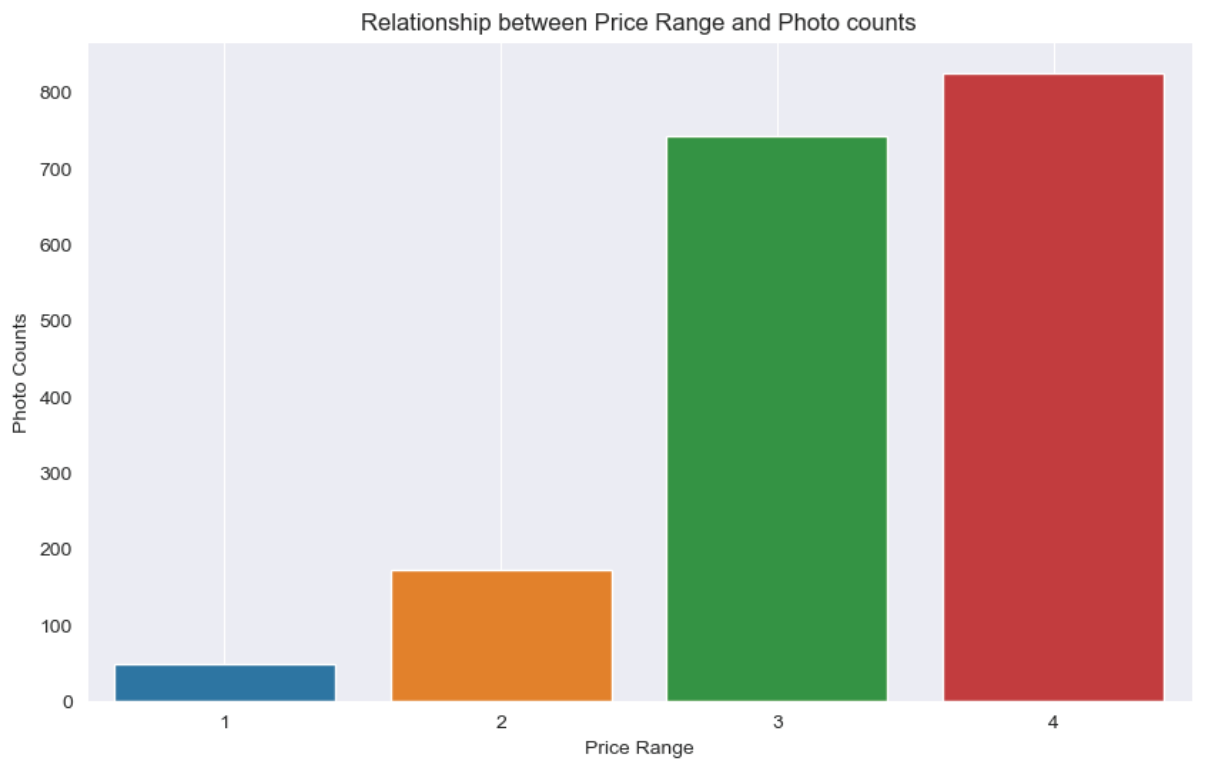


Interpretation of Price Range Distribution:

The distribution of restaurants across different price ranges on Zomato's platform **reveals a notable concentration in lower and moderate pricing categories.** This observation implies that **Zomato has a significant number of partnerships with small businesses and establishments operating in these price segments. To foster growth in this sector, **Zomato could strategically focus on this region by offering tailored incentives, reduced marginal charges, and targeted promotional campaigns.** Encouraging and supporting smaller businesses could contribute to the platform's overall success and market expansion.

Relationship between Price Range and Photo counts:

The relationship between price range and photo counts in the Zomato dataset exhibits potential insights into how pricing influences user engagement with restaurant visuals.



Observation on Relationship Between Price Range and Photo Counts:

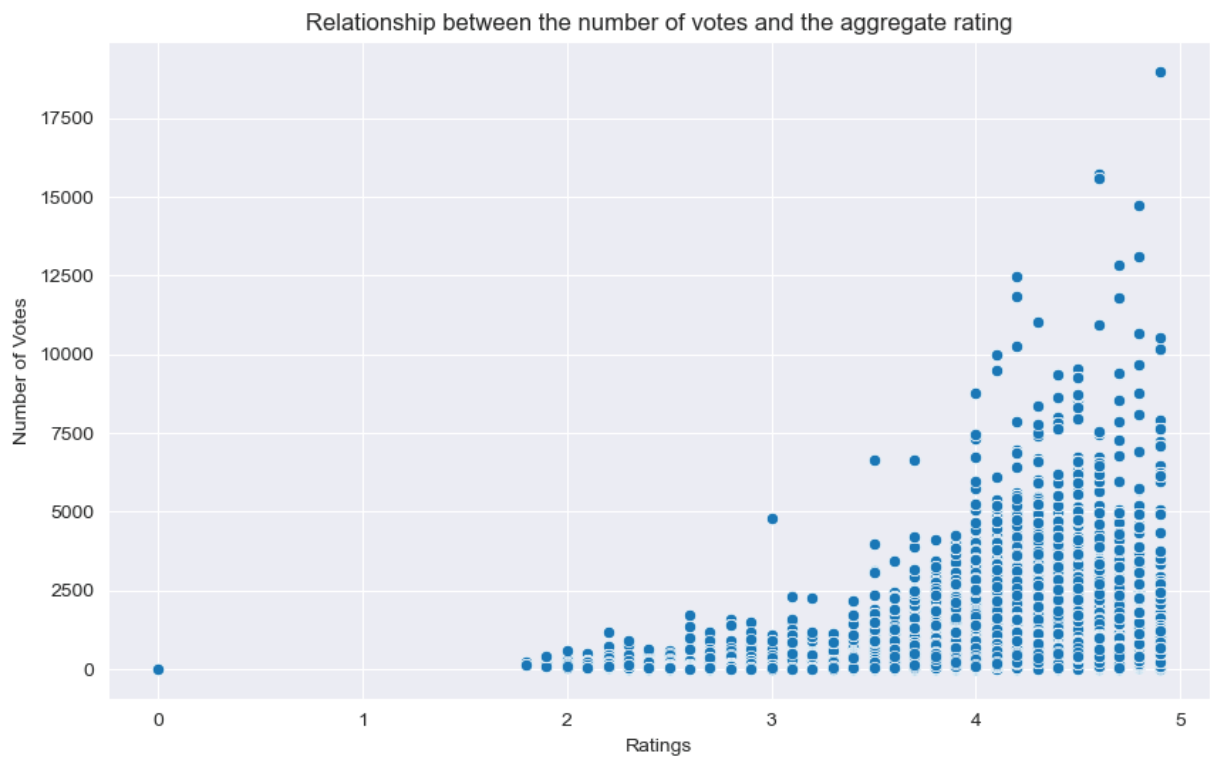
- An analysis of the data reveals a distinctive trend – **restaurants and hotels with higher price ranges tend to provide a greater number of photos compared to their moderately or less priced counterparts.**

- This pattern suggests a potential correlation between pricing tiers and the emphasis on visual representation. Establishments in the higher price range **likely invest more in showcasing their offerings through a richer visual experience.**

- This insight can guide Zomato's understanding of the marketing strategies employed by businesses in different price categories, enabling more targeted and effective promotional approaches.

Relationship between the number of votes and the aggregate rating:

The relationship between the number of votes and aggregate rating in the Zomato dataset signifies the correlation between user engagement and restaurant satisfaction. Higher votes may influence and reflect overall ratings.



Observations and Insights:

The graphical analysis of the relationship between the number of votes and the aggregate rating reveals a notable trend – **as the number of votes increases, so does the aggregate rating.** This pattern signifies a positive correlation, indicating that **customers are more likely to provide ratings for restaurants they find satisfactory.** The behavior suggests a tendency to share positive feedback.
Understanding this customer feedback behavior, **Zomato can devise incentive programs for restaurants that consistently maintain high ratings. Encouraging positive reviews can not only enhance the brand value of Zomato** but also contribute to a positive and trustworthy user experience for its customers.

Key Findings and Recommendations:

1. Regional Dynamics:
- The distribution of restaurants across cities reflects strong competition in major urban centers, with Chennai leading, followed by Mumbai, Bangalore, and Pune. Surprisingly, strong Zomato presence in South Indian cities contradicts initial assumptions of a Northern market dominance.

2. Dine-In vs. Delivery:
- A significant percentage of restaurants (62.6%) do not provide delivery, potentially indicating reputed or dine-in-focused establishments. Zomato can capitalize on this by enhancing dine-in services through special offers and strategic collaborations.

3. City-Specific Strategies:
- Cities like Greater Noida, Ghaziabad, Faridabad, and Gandhinagar, despite their size, have a lower number of listed restaurants. Targeted marketing and promotion strategies can help Zomato establish a stronger presence in these locations.

4. Cuisine Preferences:
- The word cloud analysis reveals North Indian, Indian Chinese, and South Indian cuisines as popular choices. Zomato can leverage these insights for targeted promotions and partnerships with popular cuisines.

5. Pricing and Customer Behavior:
- Most restaurants fall into lower and moderate price ranges, suggesting a dependence on small businesses. Focusing on incentives, reduced charges, and promotional campaigns for smaller establishments can be beneficial.

Key Findings and Recommendations:

6. Visual Presentation and Pricing:
- Establishments in higher price ranges tend to provide more photos, emphasizing the importance of visual representation. Zomato can consider this in marketing strategies to enhance the user experience.

7. Customer Feedback Behavior:
- The positive correlation between the number of votes and aggregate rating indicates customers are more likely to rate positively. Incentive programs for consistently high-rated restaurants can enhance Zomato's brand value.

Incorporating these insights into Zomato's strategic planning can optimize marketing efforts, strengthen regional presence, and improve overall customer satisfaction, contributing to the platform's success and growth.

SUMMARY

Summary:

The analysis of the Zomato dataset provides a comprehensive overview of various aspects within the platform. Notably, a significant majority of restaurants peak at a rating of 3.84, with a distinctive trend unfolding from 2.37 to 3.19, a surge at 3.35, and a decline around the 3.5 mark. The dataset indicates that the second-highest rating achieved is 4.1, and attaining a rating of 4.81 poses a notable challenge for many restaurants.

The project identified the dominance of North Indian cuisine, influenced by Indian Chinese and Tibetan flavors, with South Indian cuisine appealing across regions. The word cloud showcased diverse national preferences, offering strategic insights for promoting popular cuisines. In terms of restaurant distribution, Chennai leads, while Mumbai, Bangalore, and Pune follow, contradicting assumptions of North Indian cuisine dominance. The analysis suggests strategic opportunities for targeted promotions in areas with limited restaurant presence.

Furthermore, the examination of delivery availability reveals that a significant portion (62.6%) of restaurants does not offer delivery services. This presents an opportunity for Zomato to enhance dine-in experiences through strategic collaborations. Restaurants offering delivery services represent growth potential, and Zomato can amplify engagement through promotions and exclusive deals.

The relationship between average cost and ratings indicates that competition is prominent around an average cost of 1200, with customers with higher expectations choosing higher-cost restaurants. The heatmap analysis suggests no significant correlation between average cost and aggregate rating, indicating other factors influencing customer satisfaction.

SUMMARY

Summary:

The distribution of restaurants across price ranges underscores a concentration in lower and moderate pricing categories. This suggests a substantial partnership with small businesses, and Zomato could strategically focus on this sector for growth. The correlation between price range and photo counts reveals that higher-priced establishments tend to provide more photos, suggesting a visual emphasis in marketing strategies.

In examining the relationship between the number of votes and aggregate ratings, a positive correlation is evident – as votes increase, so does the aggregate rating. This behavior highlights the tendency of customers to share positive feedback. Zomato can leverage this insight by implementing incentive programs for consistently highly-rated restaurants, enhancing brand value and user experience.

Overall, the project offers valuable insights into user preferences, regional dynamics, and strategic opportunities for Zomato's growth and improvement. The platform can leverage these findings to enhance customer experiences, tailor marketing strategies, and optimize its services in the highly competitive food delivery and restaurant discovery market.

Thanks for Reading!

For Code,

https://github.com/anshumbanga/Zomato-Data-Analysis