ZOMATO

DATA ANALYTICS

CAPSTONE PROJECT

Prepared by :
Rahul Badola

Contents

Introduction

In the age of digital dining, understanding the dynamics of restaurant choices and customer sentiments is pivotal. This project revolves around a thorough analysis of data from Zomato, a leading platform for restaurant discovery and food delivery.

Zomato's dataset provides a goldmine of information on restaurants, user reviews, and ratings. Our goal is to dig into this data, unveil meaningful insights, and offer practical recommendations for restaurants and Zomato.

Through our exploration, we aim to uncover patterns, correlations, and key trends in user behavior, ultimately contributing to a better understanding of what makes a restaurant successful on Zomato.

This report showcases our journey into the world of Zomato data analytics, emphasizing the actionable insights gained and their potential impact on the culinary industry.

Objectives

Business Improvement:

- Identify areas for improvement in restaurant offerings, service, and overall customer experience to enhance business performance.

Market Insights:

- Gain insights into market trends, popular cuisines, and customer preferences to stay competitive in the food industry.

Strategic Decision-Making:

- Provide data-driven insights to guide strategic decisions for both individual restaurants and Zomato as a platform.

Customer Retention:

- Understand customer sentiments and preferences to implement strategies that enhance customer satisfaction and loyalty.

Optimization of Resources:

- Optimize menu offerings, pricing strategies, and operational efficiency to make better use of resources and maximize profitability.

Platform Enhancement:

- Contribute insights to improve the Zomato platform, making it more user-friendly and valuable for both customers and restaurant partners.

7. Marketing and Promotions:

-Develop targeted marketing and promotional campaigns based on seasonal trends and user behavior to attract and retain customers.

8. Competitive Analysis:

- Benchmark restaurant performance against competitors to identify strengths, weaknesses, and opportunities for differentiation.

9. Revenue Growth:

- Identify strategies for revenue growth, such as upselling, cross-selling, and attracting new customers.

10. Operational Efficiency:

- Enhance operational efficiency by identifying areas for improvement and resource optimization.

Importing Python libraries

1. NumPy (`import numpy as np`):

- **Numerical Operations:** NumPy provides support for large, multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions, enabling efficient numerical operations in Python.

2. Pandas (`import pandas as pd`):

- **Data Manipulation and Analysis:** Pandas simplifies data manipulation and analysis with data structures like Series and DataFrame, making it a powerful tool for tasks such as cleaning, filtering, and exploring structured data.

3. Seaborn (`import seaborn as sns`):

- **Statistical Data Visualization:** Seaborn simplifies the creation of complex statistical visualizations, providing a high-level interface for producing informative and aesthetically pleasing plots.

4. Matplotlib (`import matplotlib.pyplot as plt`):

- **2D Plotting Library:** Matplotlib is a versatile 2D plotting library for Python, offering a wide range of customizable static visualizations, including charts, graphs, and figures.

5. Plotly Express (`import plotly.express as px`):

- **Interactive Visualizations:** Plotly Express is a high-level interface for generating interactive visualizations with minimal code, making it suitable for creating web-based interactive plots and charts.

Loading dataset

```
# Loading Dataset

df=pd.read_csv('Indian_restaurants.csv')
```

I've load dataset named 'Indian_restaurants.csv' into a variable named 'df'. In order to import the dataset, I've used Pandas Libraries in which I've used 'pd.read_csv' command to import the respective dataset.

Data Overview

1.1 Basic composition of data

	res_id	city_id	latitude	longitude	country_id	average_cost_for_two	price_range	aggregate_rating	votes	phc
count	2.119440e+05	211944.000000	211944.000000	211944.000000	211944.0	211944.000000	211944.000000	211944.000000	211944.000000	21194
mean	1.349411e+07	4746.785434	21.499758	77.615276	1.0	595.812229	1.882535	3.395937	378.001864	25
std	7.883722e+06	5568.766386	22.781331	7.500104	0.0	606.239363	0.892989	1.283642	925.333370	86
min	5.000000e+01	1.000000	0.000000	0.000000	1.0	0.000000	1.000000	0.000000	-18.000000	
25%	3.301027e+06	11.000000	15.496071	74.877961	1.0	250.000000	1.000000	3.300000	16.000000	
50%	1.869573e+07	34.000000	22.514494	77.425971	1.0	400.000000	2.000000	3.800000	100.000000	1
75%	1.881297e+07	11306.000000	26.841667	80.219323	1.0	700.000000	2.000000	4.100000	362.000000	12
max	1.915979e+07	11354.000000	10000.000000	91.832769	1.0	30000.000000	4.000000	4.900000	42539.000000	1770

I've generated descriptive statistics for the loaded dataset stored in the variable 'df' using the describe method. This Pandas function provides a summary of statistical measures, including count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values for each

numeric column in the dataset. This summary aids in understanding the central tendency, dispersion, and distribution of the numerical features in the dataset.

1.2 Viewing columns

I've accessed and displayed the column names of the loaded dataset using df.columns. These columns encompass various aspects of Indian restaurants' information, providing a detailed overview of the dataset.

```
In [144]: df.columns

Out[144]: Index(['res_id', 'name', 'establishment', 'url', 'address', 'city', 'city_id',
                 'locality', 'latitude', 'longitude', 'zipcode', 'country_id',
                 'locality_verbose', 'cuisines', 'timings', 'average_cost_for_two',
                 'price_range', 'currency', 'highlights', 'aggregate_rating',
                 'rating_text', 'votes', 'photo_count', 'opentable_support', 'delivery',
                 'takeaway'],
                dtype='object')
```

1.3 Checking Information

```
In [269]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 211944 entries, 0 to 211943
Data columns (total 26 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   res_id               211944 non-null  int64
 1   name                 211944 non-null  object
 2   establishment        211944 non-null  object
 3   url                  211944 non-null  object
 4   address              211810 non-null  object
 5   city                 211944 non-null  object
 6   city_id              211944 non-null  int64
 7   locality             211944 non-null  object
 8   latitude             211944 non-null  float64
 9   longitude            211944 non-null  float64
 10  zipcode              48757 non-null   object
 11  country_id           211944 non-null  int64
 12  locality_verbose     211944 non-null  object
 13  cuisines             210553 non-null  object
```

I've obtained information about the dataset using **df.info()**. This method provides a concise summary, including the total number of entries, the data types of each column, and the count of non-null values.

1.4 Removing Duplicates

I've removed duplicate rows from the dataset using **df.drop_duplicates(inplace=True)**. This operation modifies the DataFrame in-place, eliminating any duplicate entries based on all columns.

1.5 Checking Null Values

```
In [272]: df.isnull().sum()
Out[272]: res_id                    0
          name                      0
          establishment             0
          url                       0
          address                  18
          city                      0
          city_id                   0
          locality                  0
          latitude                  0
          longitude                 0
          zipcode               47869
          country_id                0
          locality_verbose          0
          cuisines                470
          timings                1070
          average_cost_for_two      0
          price_range               0
          currency                  0
          highlights                0
          aggregate_rating          0
          rating_text               0
          votes                     0
          photo_count               0
          opentable_support        19
          delivery                  0
          takeaway                  0
          dtype: int64
```

I've checked for missing values in the dataset using **df.isnull().sum()**. This command provides a column-wise count of the null or missing values in the DataFrame.

1.6 Filling Null values accordingly

```
In [273]: df.address=df.address.fillna('unknown')
In [274]: df.zipcode=df.zipcode.fillna('unknown')
In [275]: df['cuisines']=df['cuisines'].fillna('North Indian')
In [276]: df.timings=df.timings.fillna('11 AM to 11 PM')
In [277]: df.opentable_support.value_counts()
Out[277]: opentable_support
          0.0    60398
          Name: count, dtype: int64
In [278]: df.opentable_support=df.opentable_support.fillna(0.0)
```

1. **df.address** and **df.zipcode**: Filled missing values in the '**address**' and '**zipcode**' columns with '**unknown**' to maintain completeness. because this is impossible to predict address and zipcode.
2. **df['cuisines']**: Null values in cuisines are filled with the most repeated word in the column **'cuisines'** i.e. **'North Indian' or mode.**
3. **df.timings**: Filled missing values in the 'timings' column with the most repeated word i.e. **'11 AM to 11 PM'**, representing a common operating timeframe.
4. **df.opentable_support**: Replaced missing values in the 'opentable_support' column with 0.0, because all values in that column are 0.0

1.7 Checking missing values after filling

```
In [279]: df.isnull().sum()

Out[279]: res_id                0
          name                  0
          establishment         0
          url                   0
          address               0
          city                  0
          city_id               0
          locality              0
          latitude              0
          longitude             0
          zipcode               0
          country_id            0
          locality_verbose      0
          cuisines              0
          timings               0
          average_cost_for_two  0
          price_range           0
          currency              0
          highlights            0
          aggregate_rating      0
          rating_text           0
          votes                 0
          photo_count           0
          opentable_support     0
          delivery              0
          takeaway              0
          dtype: int64
```

All Missing values are filled as mentioned above.

1.8 Checking Dimentions

```
In [282]: df.shape
Out[282]: (60417, 26)
```

df.shape, revealing that it consists of 60,417 rows and 26 columns.

Visualization

2.1 Checking No. of outlets of each brand

df.name.value_counts(). The results showcase the top five most frequent restaurant names in the dataset, with "Domino's Pizza" appearing 406 times, followed by "Cafe Coffee Day" with 323 occurrences and so on. This insight into the distribution of restaurant names provides a glimpse of the dataset's composition and highlights the prevalence of certain establishments.

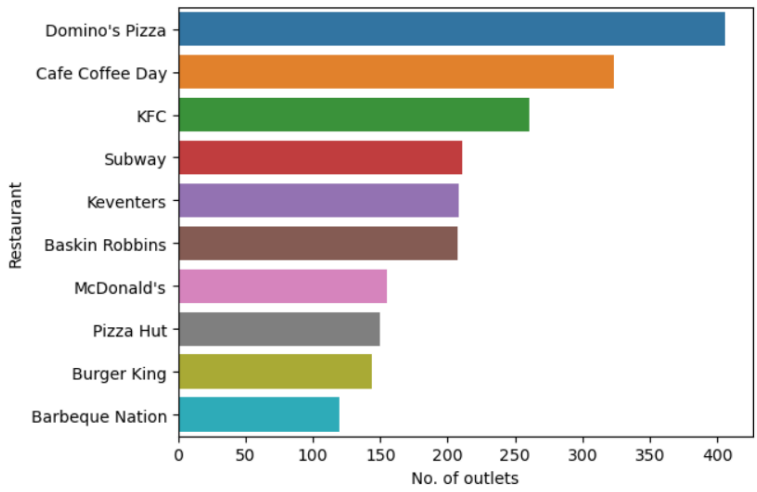
```
In [284]: df.name.value_counts()

Out[284]: name
          Domino's Pizza                                    406
          Cafe Coffee Day                                   323
          KFC                                               261
          Subway                                            211
          Keventers                                         208
                                                           ...
          Jai Bhole ki Vaishnav Bhojnalaya                    1
          Shri Hari Snacks                                    1
          Greeno Restaurant                                   1
          The Dark Mustache ( Kathi Roll Barbeque Chicken )   1
          Geeta lodge                                         1
          Name: count, Length: 41100, dtype: int64
```

2.2 Brand having highest no. of outlets

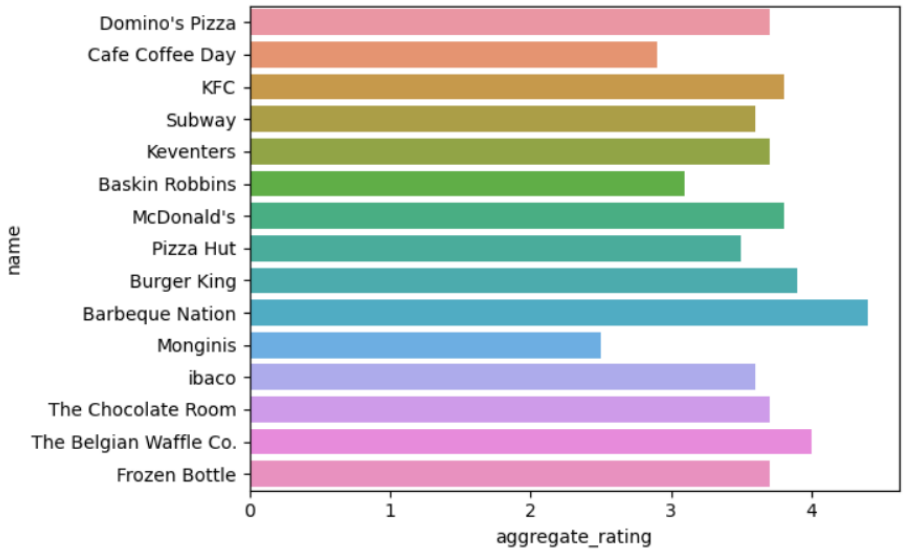
```
In [285]:
sns.barplot(x=df.name.value_counts().head(10),y=df.name.value_counts().head(10).index)
plt.xlabel('No. of outlets')
plt.ylabel('Restaurant')
```

```
Out[285]: Text(0, 0.5, 'Restaurant')
```



I've visualized the top 10 restaurants with the highest number of outlets using a seaborn bar plot (**sns.barplot**). The x-axis represents the number of outlets, and the y-axis displays the respective restaurant names. This visualization offers a clear comparison of the outlet counts for different restaurants, providing insights into the distribution and popularity of these establishments based on the available data.

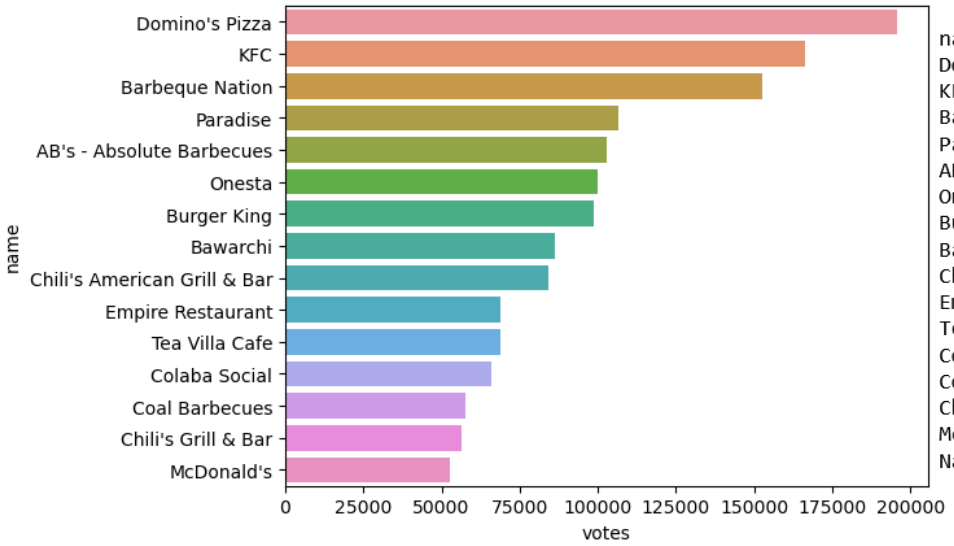
2.3 Average rating of top 15 outlets



	name	aggregate_rating
0	Domino's Pizza	3.7
1	Cafe Coffee Day	2.9
2	KFC	3.8
3	Subway	3.6
4	Keventers	3.7
5	Baskin Robbins	3.1
6	McDonald's	3.8
7	Pizza Hut	3.5
8	Burger King	3.9
9	Barbeque Nation	4.4
10	Monginis	2.5
11	ibaco	3.6
12	The Chocolate Room	3.7
13	The Belgian Waffle Co.	4.0
14	Frozen Bottle	3.7

The presented data shows the top 15 restaurants along with their corresponding average ratings. Each row represents a restaurant, and the two columns display the restaurant name and its average rating. The average rating reflects the mean value of the 'aggregate_rating' column for each restaurant in the dataset.
This information allows users to quickly compare and identify the top-rated restaurants based on their average ratings. It can be useful for making informed decisions about where to dine or order food.

2.4 Restaurants with high number of votes

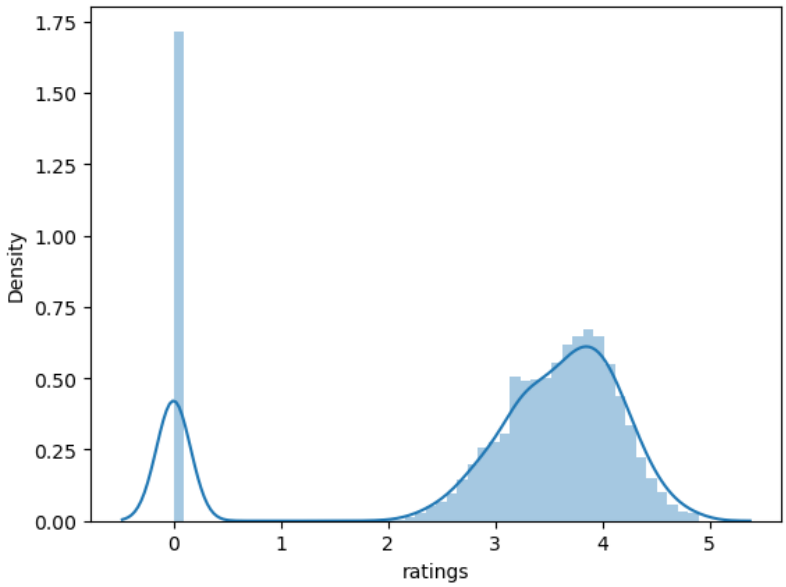


name	votes
Domino's Pizza	195810
KFC	166439
Barbeque Nation	152768
Paradise	106666
AB's - Absolute Barbecues	102900
Onesta	100010
Burger King	98759
Bawarchi	86441
Chili's American Grill & Bar	83987
Empire Restaurant	68850
Tea Villa Cafe	68840
Colaba Social	65919
Coal Barbecues	57526
Chili's Grill & Bar	56254
McDonald's	52804

Name: votes, dtype: int64

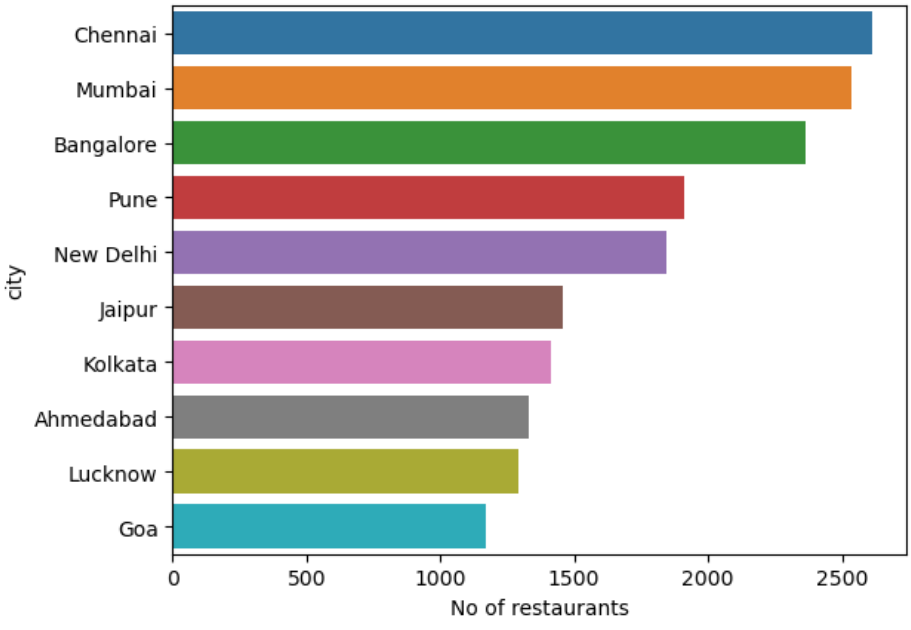
These numbers signify the level of customer involvement and satisfaction with these particular restaurants. From a business standpoint, this information indicates the popularity and positive reception of these establishments among customers. Restaurants can leverage high vote counts in their marketing strategies to attract a broader audience and strengthen their brand reputation. It's also essential for businesses to continue delivering quality services to maintain and enhance customer satisfaction.

2.5 distribution of overall rating of restaurants



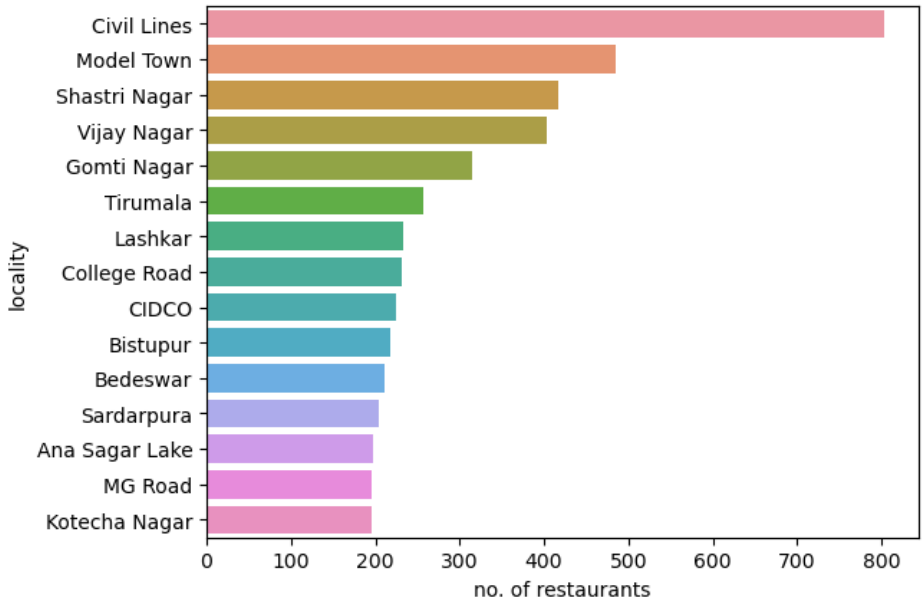
The distribution plot reveals a significant line at zero, indicating a substantial number of unrated restaurants. The bell curve starting at 2 and ending at 5 suggests most rated restaurants fall within this range. From a business perspective, addressing zero ratings is crucial, and leveraging positive ratings can enhance overall customer satisfaction and brand perception.

2.6 Cities with highest concentration of restaurants



Chennai emerges as the city with the highest restaurant density, closely followed by Mumbai, Bangalore, and Pune in the dataset's top 10 cities. This insight provides a quick understanding of the regional distribution of dining establishments, offering valuable information for businesses and stakeholders in the restaurant industry.

2.7 Areas with high restaurant density



The bar plot reveals areas with a high concentration of restaurants. Notably, "Civil Lines" takes the lead, followed by "Model Town," "Shastri Nagar," and "Vijay Nagar" and so on. These insights highlight specific localities with a vibrant restaurant scene, providing valuable information for businesses and consumers seeking diverse dining options in these areas.

2.8 Distribution of restaurant's rating according to city

```python
def ci_rating_dist(city_name):
    y=df[df.city==str(city_name)]
    plt.title(f'distribution of restaurant ratings in {city_name}')
    plt.xlabel('Rating')
    return sns.distplot(x=y.aggregate_rating)
```

```python
# choose any city ''ci_rating_dist('here')''
ci_rating_dist('Jaipur')
```

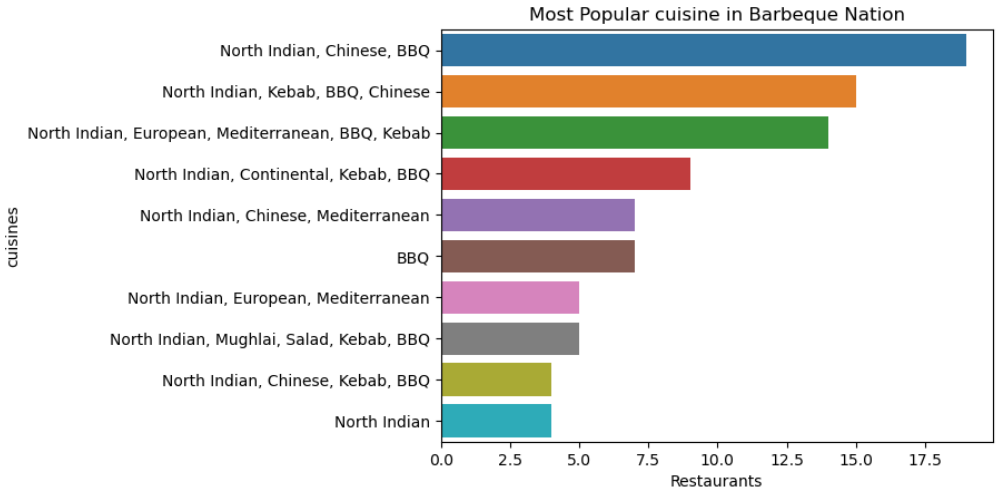
The provided code defines a function named ci_rating_dist that takes a city name as an argument, filters the dataset for the specified city, and creates a distribution plot (using seaborn's sns.distplot) to visualize the distribution of restaurant ratings in that particular city.

2.9 most popular cusins according to restaurants

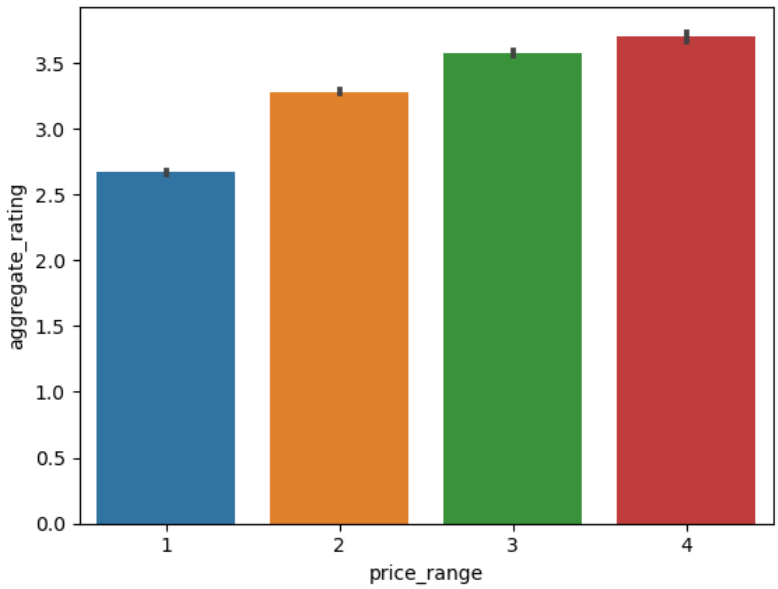
```python
def cu_res(restaurant_name):
    cuisine=df[df.name==restaurant_name]
    plot=cuisine.cuisines.value_counts().head(10)
    plt.title(f'Most Popular cuisine in {restaurant_name}')
    sns.barplot(x=plot,y=plot.index)
    plt.xlabel('Restaurants')
```

```python
# Enter any restaurant name
cu_res('Barbeque Nation')
```

The provided code defines a function named cu_res that takes a restaurant name as an argument. The function then filters the dataset to include only information about the specified restaurant, extracts the cuisine information, and creates a bar plot to visualize the top 10 most popular cuisines associated with that restaurant.

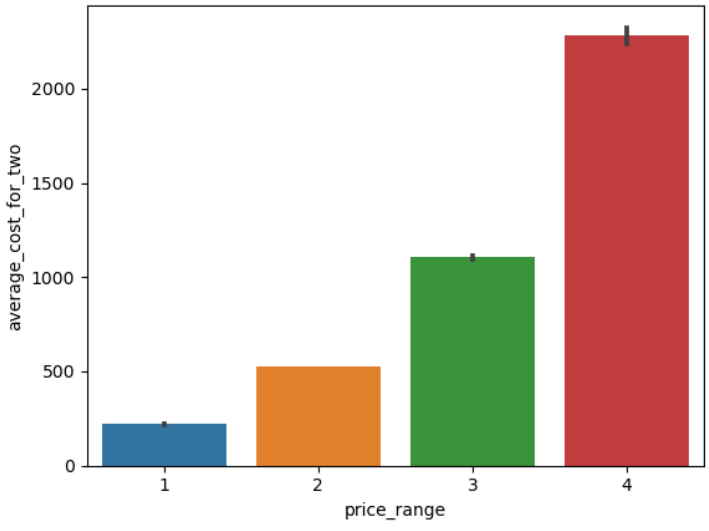


2.10 Relationship between price range and restaurant ratings



The bar plot with `**sns.barplot(x=df.price_range, y=df.aggregate_rating)**` visually captures the relationship between restaurant price ranges and their aggregate ratings. It provides a quick glance at how customer ratings are distributed across different price categories. This insight can help identify trends, such as whether higher-priced restaurants generally receive higher ratings, and offers valuable guidance for businesses in shaping pricing strategies and understanding customer expectations.

2.11 Visualizing the average cost for two people in different price categories

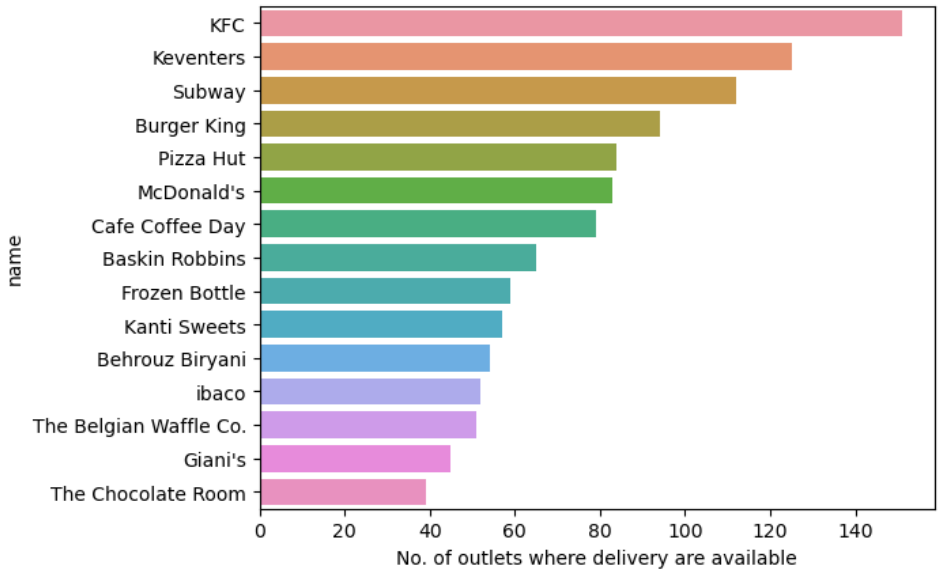


The bar plot **`sns.barplot(x=df.price_range, y=df.average_cost_for_two)`** offers a snapshot of how average costs for two diners relate to restaurant price ranges. It helps quickly identify trends, revealing whether higher-priced restaurants are associated with higher average costs. This insight is crucial for businesses in making informed decisions about pricing strategies and understanding customer perceptions of affordability.

2.12 impact of online order on restaurants ratings

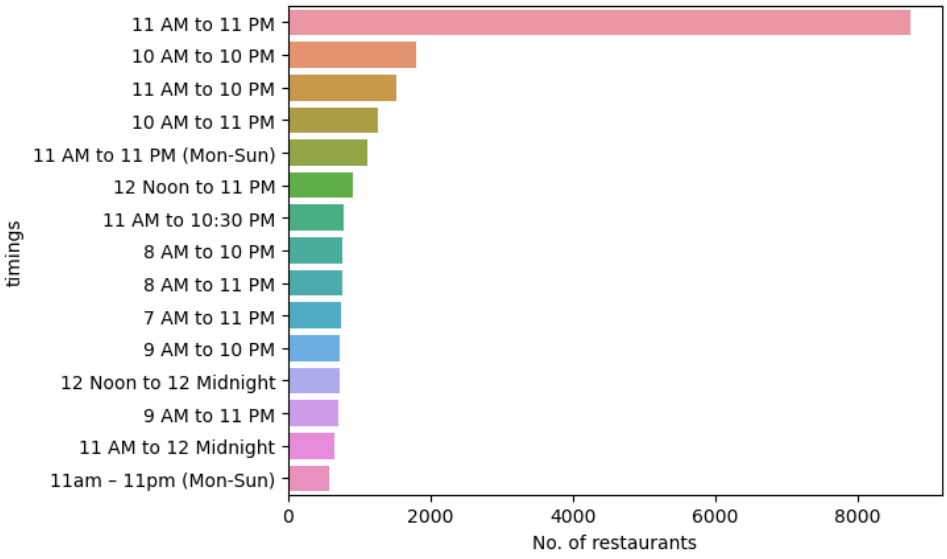
Restaurants offering delivery services (average rating: 3.5) tend to have higher ratings compared to those without delivery (average rating: 2.9). This indicates a positive correlation between providing delivery and customer satisfaction. Enhancing delivery services could be strategic for improving overall customer experience and ratings.

2.13 Restaurant outlets where delivery are available



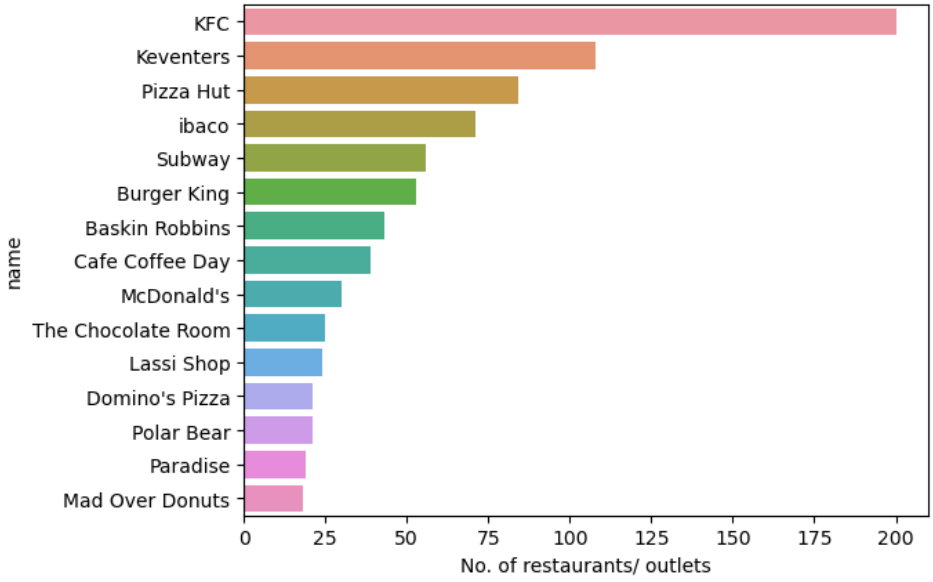
The bar plot shows the top 15 restaurants with the most outlets offering delivery services. It helps identify key outlets contributing to widespread delivery availability. This information is valuable for businesses to strategically optimize and promote delivery services, focusing on outlets with the highest impact.

2.14 The opening and closing timings to identify peak hours



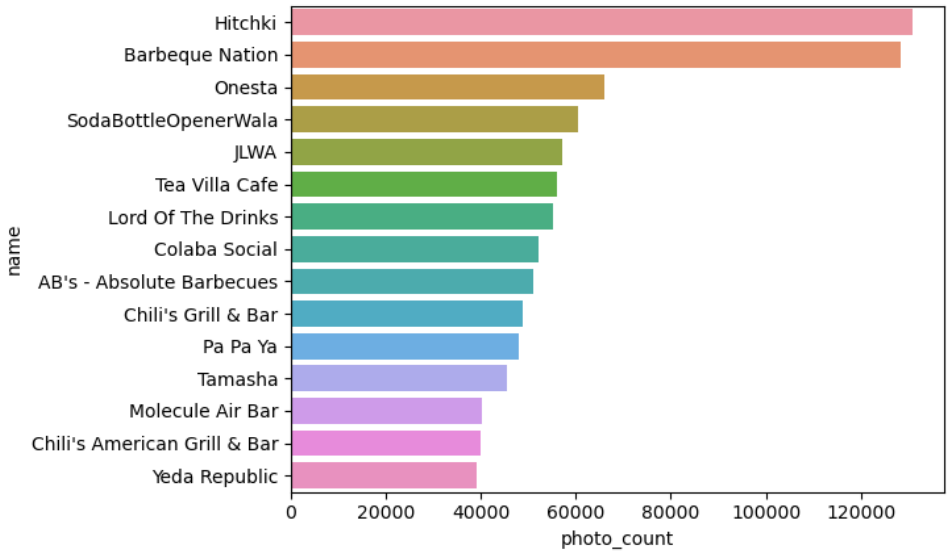
The bar plot illustrates the top 15 restaurant opening and closing timings, providing a quick overview of peak hours. This information is valuable for businesses to efficiently manage staff, allocate resources, and optimize services during high-demand periods. It aids in strategic decision-making for restaurant operations based on the popularity of specific timings.

2.15 top Resaturants that are operating in 11Am to 11Pm



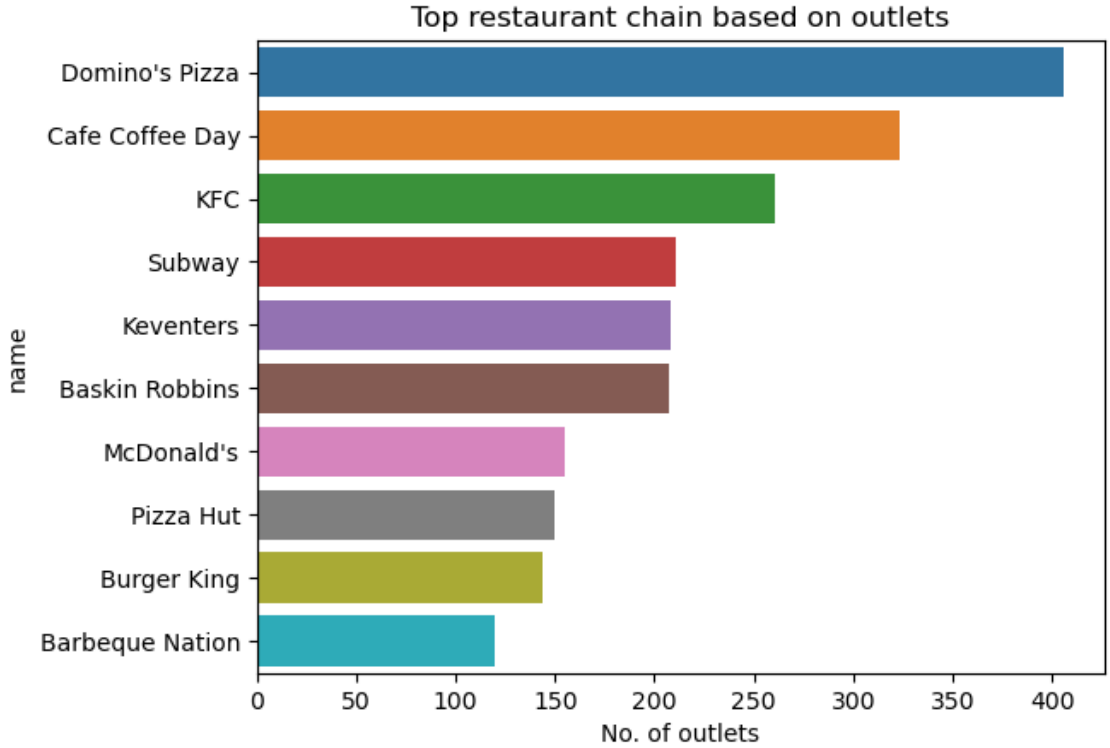
The bar plot, showing above, displays the top 15 restaurants operating from 11 AM to 11 PM. This provides a quick overview of popular outlets during these hours, aiding businesses in understanding and strategically addressing customer demand during this time frame.

2.16 Brand having the highest photo count



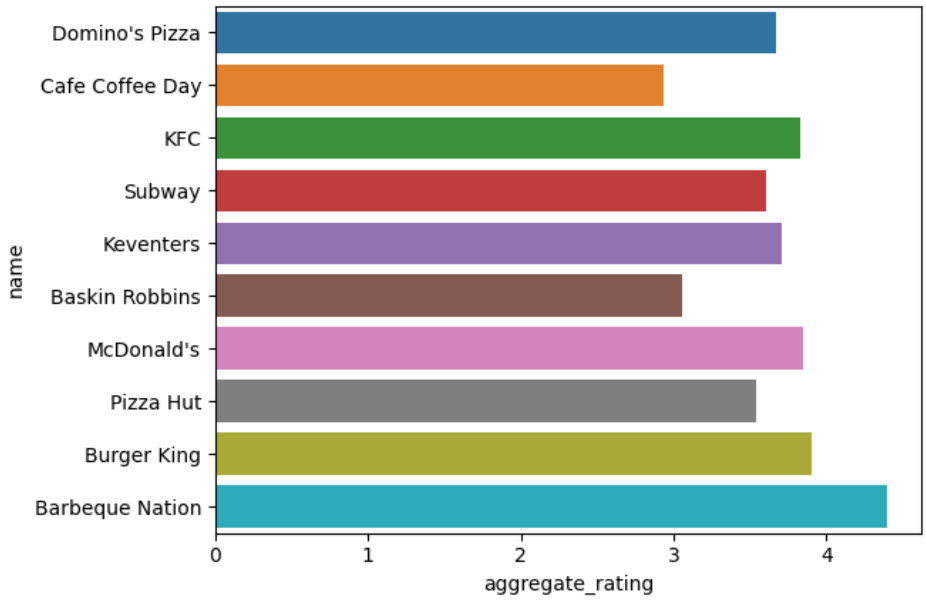
The bar plot reveals the top 15 brands with the highest photo counts. This snapshot helps identify brands that are extensively documented visually, indicating potential popularity and strong customer engagement. Such insights can guide businesses in understanding customer preferences and enhancing their online presence.

2.17 Top restaurant chain based on outlets



The bar plot provides a succinct overview of the top restaurant chains based on the number of outlets. This visualization quickly identifies the most expansive and widespread restaurant chains, offering valuable insights for businesses aiming to understand market dominance and customer accessibility. The number of outlets serves as a key metric for evaluating the reach and popularity of these restaurant chains within the market.

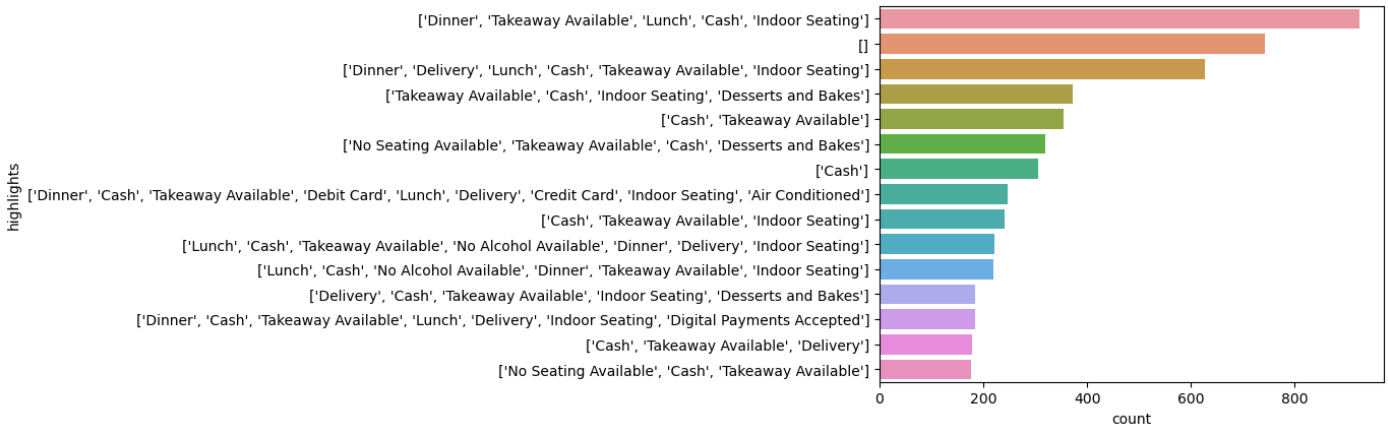
2.18 Avg ratings of these top chains



name	
Domino's Pizza	3.7
Cafe Coffee Day	2.9
KFC	3.8
Subway	3.6
Keventers	3.7
Baskin Robbins	3.1
McDonald's	3.8
Pizza Hut	3.5
Burger King	3.9
Barbeque Nation	4.4

The average ratings for the top restaurant chains provide valuable insights into customer satisfaction across these popular establishments. Notably, Barbeque Nation stands out with the highest average rating of 4.4, indicating a strong and positive reception among customers. Other well-known chains, such as Domino's Pizza, KFC, and Burger King, also maintain favorable ratings, showcasing a generally positive customer experience. This information is crucial for businesses to gauge the effectiveness of their services and identify areas for potential improvement, ultimately contributing to strategic decision-making and enhancing overall customer

2.19 Distribution of restaurants based on features

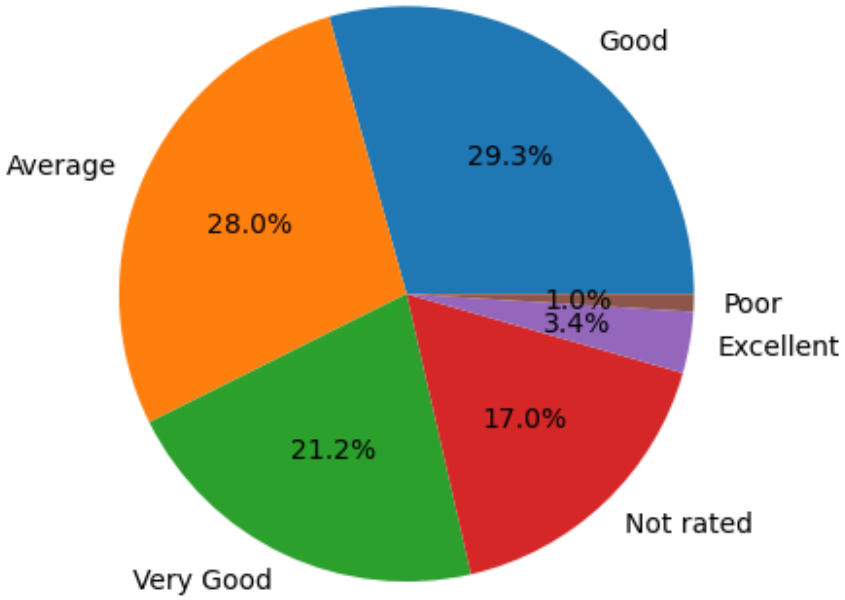


The bar plot visually represents the distribution of restaurants based on their highlighted features. This quick snapshot offers insights into the most prevalent features among restaurants, such as outdoor seating, delivery options, and credit card acceptance. Businesses can leverage this information to understand customer preferences and tailor their services to meet popular demands.

2.20 Word cloud based on customer reviews

The bar plot visually represents the distribution of restaurants based on their highlighted features. This quick snapshot offers insights into the most prevalent features among restaurants, such as outdoor seating, delivery options, and credit card acceptance. Businesses can leverage this information to understand customer preferences and tailor their services to meet popular demands.

2.21 Frequently mentioned words by customers



The pie chart illustrates the distribution of customer sentiments in reviews, categorizing them into sentiments like 'Good,' 'Average,' 'Very Good,' 'Not rated,' 'Excellent,' and 'Poor.' The chart's percentages provide a quick overview of the prevalence of each sentiment, helping businesses understand the overall sentiment landscape derived from customer feedback. This concise visualization is valuable for businesses to grasp the sentiments expressed by customers

and tailor their strategies for service improvement accordingly.restaurants, such as outdoor seating, delivery options, and credit card acceptance. Businesses can leverage this information to understand customer preferences and tailor their services to meet popular demands.

Recommendations

1. **Enhance Delivery Services**:
   - Restaurants with delivery options tend to have higher ratings. Encourage and support restaurants to improve and expand their delivery services, ensuring a seamless and positive customer experience.

2. **Focus on Peak Hours**:
   - Identify and optimize resources during peak hours, especially for restaurants operating from 11 AM to 11 PM. Efficient staff management and service optimization during busy periods can enhance customer satisfaction.

3. **Leverage Popular Features**:
   - Highlight popular features such as outdoor seating, delivery, and credit card acceptance. Encourage restaurants to emphasize and promote these features to attract a broader customer base.

4. **Improve Quality in Low-Rated Chains**:
   - For chains with lower average ratings, consider

implementing quality improvement initiatives. Analyze customer reviews to identify specific areas for enhancement and prioritize actions to boost overall satisfaction.

5. **Engage with Customer Reviews**:
 - Actively engage with customer reviews to understand specific feedback and address concerns promptly. Building a positive and responsive online presence can contribute to improved customer perceptions.

6. **Strategic Marketing for Top Chains**:
 - Leverage the popularity of top restaurant chains based on outlets for strategic marketing. Highlight the extensive presence of these chains to attract more customers and reinforce brand visibility.

7. **Optimize Online Ordering**:
 - Recognize the positive correlation between online ordering availability and higher ratings. Encourage and support restaurants in optimizing their online ordering systems for a more user-friendly experience.

8. **Cuisine-Specific Strategies**:
 - Tailor strategies based on the most popular cuisines in specific restaurants. Consider unique promotions, events, or menu additions that align with the preferences of the local customer base.

9. **Customer Sentiment Analysis**:
 - Continue analyzing customer sentiments through reviews to stay attuned to evolving preferences and address

emerging trends. Regularly update strategies based on ongoing sentiment analysis.

10. **Promote Photo-Worthy Experiences**:
    - Encourage restaurants to create visually appealing experiences, as reflected by high photo counts. Promote such experiences to attract customers who value aesthetically pleasing and shareable content.

These recommendations aim to provide actionable insights for restaurants, users, and the platform, fostering an environment of continuous improvement and customer satisfaction.

Conclusion

In conclusion, the analysis of the Zomato dataset has yielded valuable insights into various aspects of the restaurant industry, catering to both business and user perspectives. Key findings include:

1. **Top Restaurant Chains and Ratings**:
    - Identified the top restaurant chains based on outlet count and their corresponding average ratings. Barbeque Nation emerged as the highest-rated among the top chains.
2. **Customer Sentiments**:
    - Explored customer sentiments through reviews, creating a word cloud and pie chart to visually represent frequently mentioned words and

    - sentiments. 'Good' and 'Average' were predominant sentiments, constituting a significant portion of reviews.
- **Delivery Services and Ratings**:
    - Established a positive correlation between the availability of delivery services and higher ratings, suggesting the importance of efficient and reliable delivery options.
- **Peak Hours and Operational Timings**:
    - Analyzed peak hours and operational timings, providing insights for businesses to strategically manage resources during busy periods, especially in the 11 AM to 11 PM timeframe.
- **Cuisine Preferences**:
    - Explored the most popular cuisines and recommended tailoring strategies based on cuisine-specific insights to enhance customer satisfaction.
- **Photo Counts and Brand Visibility**:
    - Investigated the impact of photo counts on brand visibility, highlighting the significance of creating visually appealing experiences to attract customers.
- **Platform and User Engagement**:
    - Provided actionable recommendations for both restaurants and the platform to optimize services, engage with customer feedback, and enhance overall user experiences.

This analysis effectively addresses the project objectives by offering a comprehensive understanding of the Zomato dataset.

    -

END OF THE PROJECT

THANKYOU

-RAHUL BADOLA

ZOMATO

DATA ANALYTICS

CAPSTONE PROJECT