Zomato Restaurant Chain Analysis

Group-3

Niraj Shah	370D9284BJ
Jamima Yeasmin	PTIYQ0JV36
Kowsalya Sugumar	JY6ZD9G0B7
Nandita Malakar	L039ICT7RD
Harsh Chaudhary	AHZUJOTP1Y

Table of content:

Problem Statement:

Restaurants from all over the world can be found here in Bengaluru. From the United States to Japan, Russia to Antarctica, you get all types of cuisines here. Delivery, Dine-out, Pubs, Bars, Drinks, Buffet, Desserts you name it and Bengaluru has it. Bengaluru is the best place for foodies. The number of restaurants is increasing day by day. Currently, it stands at approximately 10,000 restaurants. With such a high number of restaurants. This industry hasn't been saturated yet. And new restaurants are opening every day. However, it has become difficult for them to compete with already established restaurants. The key issues that continue to pose a challenge to them include high real estate costs, rising food costs, shortage of quality manpower, fragmented supply chain, and over-licensing. This Zomato data aims at analyzing the demography of the location. Most importantly it will help new restaurants in deciding their theme, menus, cuisine, cost, etc for a particular location. It also aims at finding similarities between neighborhoods of Bengaluru on the basis of food. The dataset also contains reviews for each of the restaurants which will help in finding the overall rating for the place.

What problem are you solving?

This research endeavour centres around a comprehensive analysis of the provided Zomato dataset, addressing the following crucial questions:

1. **Data Loading and Library Import:**
   Commence the analysis by importing essential Python libraries and loading the Zomato dataset for Bangalore.

2. **Identifying Top Restaurant Chains in Bangalore:**
   Uncover and rank the top restaurant chains based on their prevalence and frequency in Bangalore.

3. **Analysis of Online Order Acceptance:**
   Quantify the number of restaurants that do not accept online orders, shedding light on this aspect of the business landscape.

4. **Assessing the Ratio of Table Booking Provision:**
   Determine the proportion of restaurants that offer table booking services versus those that do not, providing insights into customer preferences.

5. **Boxplot Examination for Ratings:**
   Utilize advanced techniques such as User Defined Functions (UDFs), Lambda functions, or Apply functions to create a boxplot for restaurant ratings. Extract and analyse the data before the slash ("/") to identify underlying patterns.

6. **Comparative Percentage of Online and Offline Orders:**
   Calculate the relative percentages of restaurants accepting online and offline orders to discern the prevailing consumer behaviour.

7. **Unearthing Insights from the Cost vs. Rating Scatter Plot:**
   Develop an informative scatter plot to explore the relationship between restaurant cost and ratings. Apply appropriate data preprocessing, such as removing "," in the cost values, to facilitate meaningful insight.

8. **Distribution Analysis for Votes and Approximate Cost:**
   Devise a tailored user-defined function and employ a for-loop to uncover the distribution patterns of votes and approximate cost among restaurants.

9. **Uncovering the Most Prevalent Restaurant Types in Bangalore:**
   Analyse and rank the most common restaurant types in Bangalore, aiding restaurant owners in strategic decision-making.

10. **Impact of Online Order Acceptance on Votes:**
    Investigate whether restaurants that accept online orders experience a statistically significant difference in the number of votes compared to those that do not, providing essential insights into the influence of online ordering.

11. **Identifying the Best Budget Restaurants in Specific Locations:**
    Employ data-driven techniques to identify the best budget restaurants in chosen locations, leveraging average cost and ratings as key criteria.

12. **Recognizing Top Quick Bites Restaurant Chains:**
    Ascertain the top quick bites restaurant chains in Bangalore, providing valuable information for customer-centric strategies.

13. **Analysing the Popularity of Casual Dining Chains:**
    Evaluate the popularity of casual dining restaurant chains using appropriate visualization techniques to gain a comprehensive understanding of customer preferences.

14. **Exploring Popular Cuisines in Bangalore:**
    Unearth the most popular cuisines in Bangalore through data-driven plots, aiding restaurant owners in diversifying their offerings and catering to customer demands.

How did you solve the problem?

We initiated the data analysis process by loading the dataset using the pandas library, and to enhance interpretability, we provided custom column names. Subsequently, we examined the data types of different columns, where we encountered a specific column containing numeric values represented as strings (i.e., object data type). To facilitate numeric analysis, we sought to convert this column into a numeric format. During this conversion, we encountered several error values, which we appropriately handled by utilizing the NumPy library to convert them into null values.

Further scrutiny revealed a column titled "Ratings out of 5," displaying values in the format "3.3/5" with its data type specified as "object." To perform quantitative analysis, we carefully split the values using the "/" delimiter, thereby accessing the "3.3" portion and subsequently converting it into a numeric representation.

Moreover, we detected the presence of numerous garbage values within the categorical columns (non-numeric columns), impeding meaningful analysis. To ensure data integrity, we conducted a data cleansing process, effectively eliminating the unwanted values. This cleaning process leveraged the "Regular expression (re)" library, enabling us to handle and remove the undesirable entries efficiently.

Prior to delving into in-depth analysis, we investigated the dataset's description and studied its statistical properties. This preliminary examination provided essential insights into the distribution, central tendencies, and variability of the data's features.

Additionally, we conducted a comprehensive evaluation of missing values within both categorical and numeric columns. To maintain the dataset's coherence and completeness, we devised effective strategies to replace these missing values with logical alternatives.

Following a systematic top-to-bottom approach, we thoroughly loaded and studied the dataset before undertaking essential cleansing procedures. Our aim was to establish a refined and coherent dataset, thus laying the groundwork for a thorough understanding and effective analysis of the data.

Subsequently, armed with the cleaned dataset, we proceeded to conduct detailed analysis in accordance with the problem statements provided to us.

Utilizing a combination of statistical techniques, data visualizations, and data manipulation with pandas and other libraries, we addressed various research questions and derived meaningful insights from the data.

The culmination of our efforts resulted in a robust analysis, empowering us to make data-driven decisions and provide actionable recommendations to stakeholders. This meticulous approach not only facilitated a comprehensive understanding of the dataset but also paved the way for effective problem-solving and informed decision-making.

What were the steps you took to solve

1. The dataset was loaded with custom column names to enhance clarity and understanding.

2. The data types of all columns were checked using the "info()" function.

3. The presence of missing values in different columns was assessed using the ".isnull()" method.

4. Numeric columns with error values were converted to a numeric format, treating the errors as null values using the NumPy library. The "pd.to_numeric()" function was utilized for converting non-numeric columns to numeric format.

5. The "Ratings out of 5" column contained values in the format "3.3/5" with its data type as object. To facilitate analysis, the values were split using the "/" delimiter, extracting the "3.3" part, and subsequently converted to a numeric format.

6. The "Rest_name" column exhibited numerous irrelevant values. Since restaurant names were vital for analysis, the garbage values in this column were cleaned and removed.

7. A comprehensive approach was taken to handle missing values. Logical imputation was employed to replace the majority of null values. For numeric columns, the remaining null values were replaced with the

median. In the "location" column, missing values were substituted with the values available in the "location listed in Zomato" column.

8. With the data now cleaned and prepared, attention was shifted to addressing the problem statement.

9. The top restaurant chain in Bangalore was determined by sorting the data based on "ratings" and "voters count" in descending order, and the top 10 rows were selected to identify the leading restaurant chains.

10. Analysis of the "online orders" column was performed using the ".value_counts()" method to ascertain the number of restaurants accepting and not accepting online orders.

11. The ratio of restaurants providing table booking was derived using the "value_counts()" method to gain insights into the proportion of restaurants offering table booking facilities.

12. A box plot graph for "Ratings" was created using the seaborn library to understand the distribution of rating values.

13. The percentage of restaurants accepting online and offline orders was calculated.

14. A scatter plot graph for the "Cost vs Rating" variable was generated using Seaborn and Matplotlib, taking into account the online order availability.

15. The most common restaurant types were identified from the "Rest_type" column using the "df['Rest_type'].value_counts().sort_values(ascending=False).head(10)" method.

16. The distribution of votes and approximate cost was computed using a user-defined function and a for loop.

17. The best budget restaurant was determined by employing the group by function and aggregating columns such as "Average Cost for two," "Ratings out of 5," and "Voters Count."

18. Similarly, the popular restaurant chain with a cuisine type of "Quick Bites" was also identified using the above-mentioned logic.

19. The popular casual dining restaurant was found using the group by function, and the corresponding graph was plotted using Matplotlib and Seaborn libraries.

20. To analyse the top famous cuisines, functions such as split, explode, and value counts were employed, and corresponding graphs were plotted using Matplotlib and Seaborn libraries.

Questions with code and relevant outputs

1. Import libraries that you required and Load the data set.
   We used the below libraries for the analysis of report.

```
]: import pandas as pd
   import numpy as np
   import matplotlib.pyplot as plt
   import seaborn as sns
   import re
```

Fig-1

2. Using the 'group by' and 'value_count' function we found top restaurant chains based on the distribution (frequency) in Bengaluru.

```
df_Q2 = df.groupby(['rest_name']).agg({"Voters Count" : "sum", 'Ratings out of 5' : "mean"})


## Adding new column in dfQ2 showing value counts of total hotel chain of respective brand
df_Q2["no of restaurants"] = df["rest_name"].value_counts()


df_Q2.sort_values("no of restaurants", ascending=False).head(10)
✓ 0.0s
```

```
df_Q2.sort_values(by=["Voters Count", 'Ratings out of 5'], ascending=False).head(10)
```
✓ 0.0s

rest_name	Voters Count	Ratings out of 5	no of restaurants
Onesta	64814	4.426667	15
Truffles	59814	4.580000	10
Hammered	34320	4.612500	10
Arbor Brewing Company	33583	4.500000	4
Prost Brew Pub	31435	4.500000	4
The Black Pearl	27916	4.733333	3
Empire Restaurant	25562	4.066667	9
The Biere Club	24456	4.300000	7
Barbeque Nation	22965	4.666667	6
House Of Commons	22189	4.740000	5

Fig- 2

3. How many restaurants do not accept online orders?

   We found out the number of restaurants which do not accept online orders.

Q3. How many restaurants do not accept online orders?

```
df["online_order"].value_counts()
```
✓ 0.0s

```
Yes    5423
No     3668
Name: online_order, dtype: int64
```

Fig-3

4. What is the ratio b/w restaurants that provide and do not provide table booking?

   We did the analysis for figuring out the ratio b/w restaurants that provide and do not provide table booking.

```
df['Table Booking Availability'].value_counts()
```
✓ 0.0s

```
No     7967
Yes    1124
Name: Table Booking Availability, dtype: int64
```

```
print("The ratio b/w restaurants that provide and do not provide table booking is : \n")

print(round((df['Table Booking Availability'].value_counts()[1]/df['Table Booking Availability'].value_counts()[0]), 4))
```
✓ 0.0s

```
The ratio b/w restaurants that provide and do not provide table booking is :

0.1411
```

Fig-4

5. We plotted a boxplot for the rating column and studied the 5 point summary.

```
sns.boxplot(data=df, y='Ratings out of 5')
plt.show()
```
✓ 0.1s

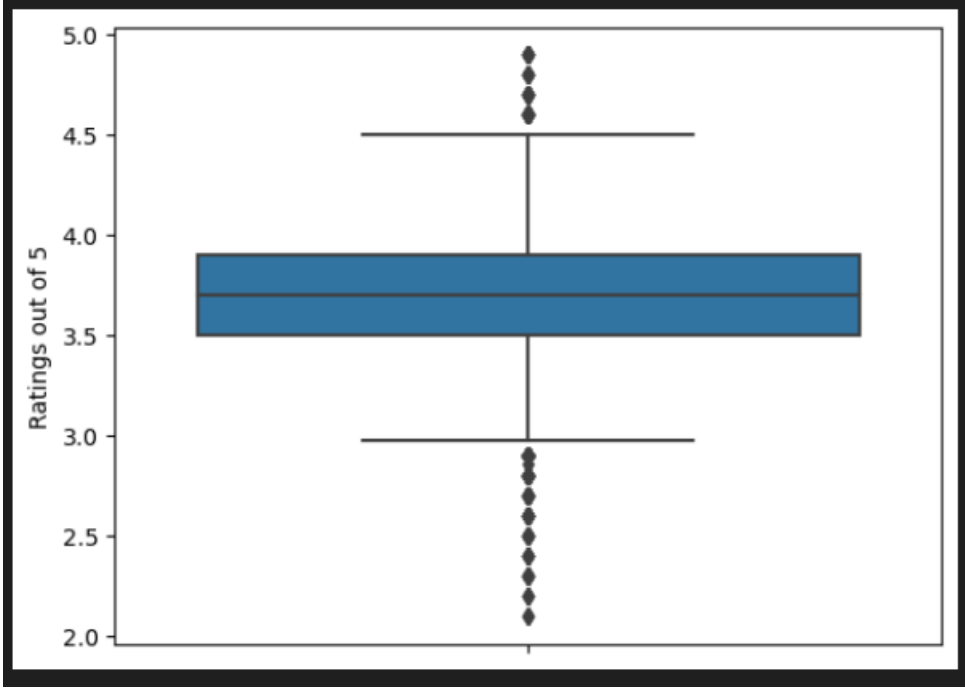

Fig-5

6. We also figured out the Online and Offline orders restaurants percentage.

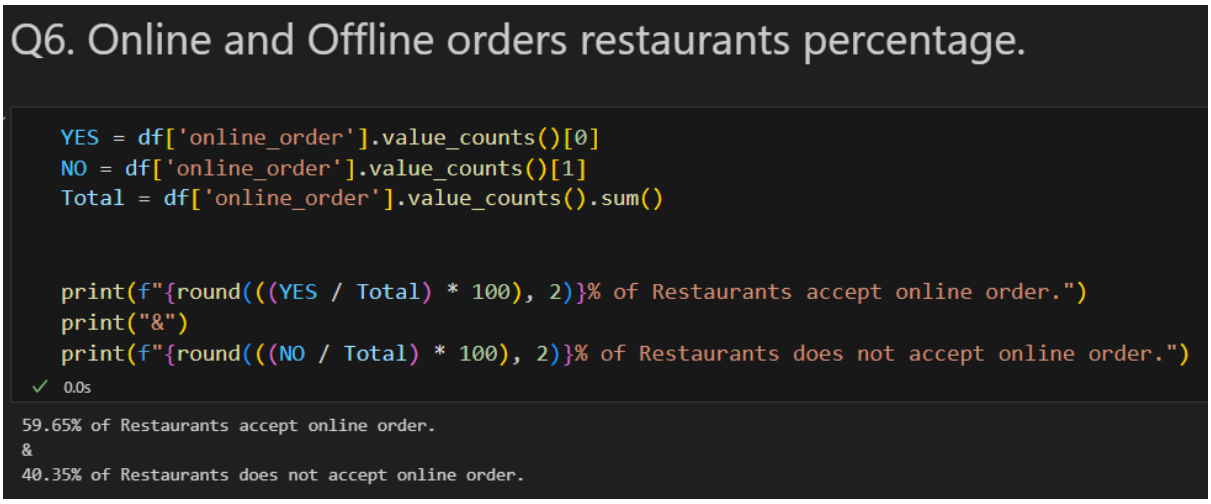

Fig-6

7. We plotted the scatter plot using the Cost vs rating variable with respect to online order.

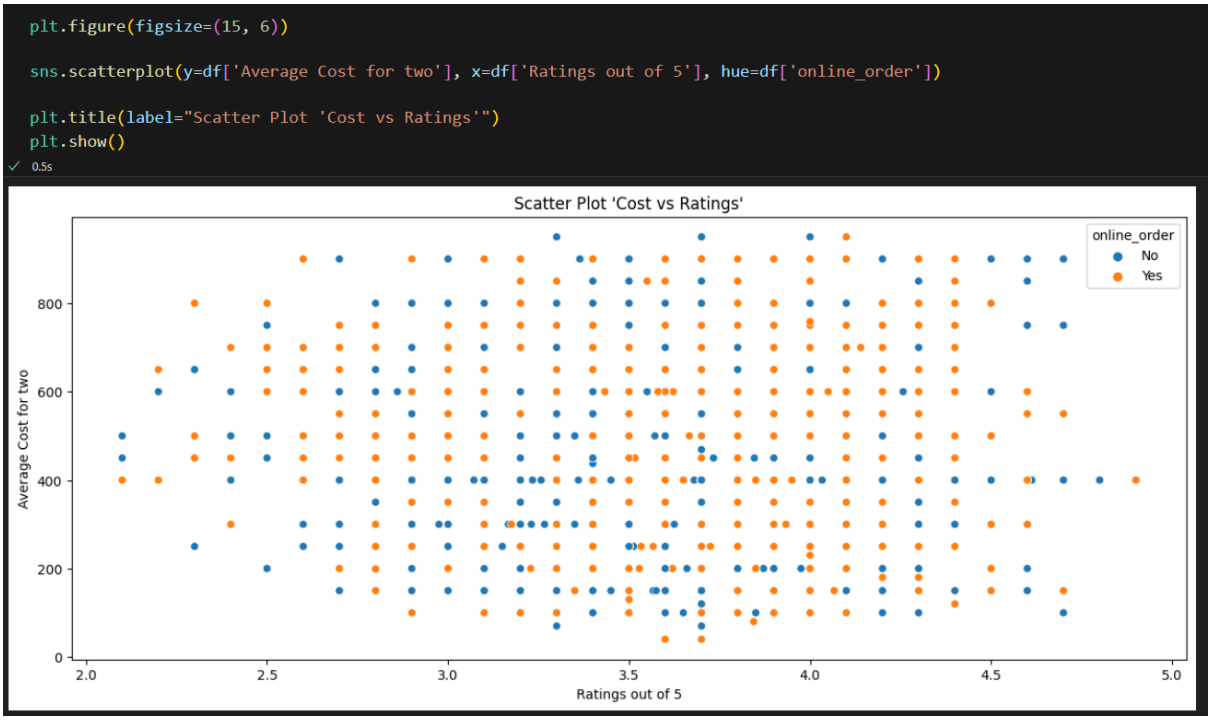

Fig-7

8. With our analysis we found that which the most common restaurant type in Bangalore city are.

9. Which are the most common restaurant type in Banglore?

```
df['Rest_type'].value_counts().sort_values(ascending=False).head(10)
```

✓ 0.0s

```
Quick Bites           3300
Casual Dining         1835
Cafe                   644
Delivery               459
Dessert Parlor         409
Takeaway, Delivery     395
Bakery                 205
Casual Dining, Bar     190
Beverage Shop          140
Bar                    120
Name: Rest_type, dtype: int64
```

Fig-8

9. We found out the difference between the votes of restaurants accepting and not accepting online orders.

Q10. Is there any difference b/w the votes of restaurants accepting and not accepting online orders?

```
ans = df[df["online_order"] == "Yes"].loc[:, "Voters Count"].value_counts().sum(
) - df[df["online_order"] == "No"].loc[:, "Voters Count"].value_counts().sum()

print("The difference between the votes of restaurants accepting and not accepting online orders is :", ans)
```

✓ 0.0s

```
The difference between the votes of restaurants accepting and not accepting online orders is : 1755
```

Fig-9

10. We find out which are the Best budget Restaurants in any location in Bangalore.

Q12. Find the Best budget Restaurants in any location.

```
budget_rest = df[((df['Average Cost for two'] >= 400) & (df['Average Cost for two'] <= 460) & (df['Ratings out of 5'] > 4) & (
    df['Voters Count'] > 5000))].loc[:, ['rest_name', 'Rest_type', 'location', 'Average Cost for two', 'Ratings out of 5', 'Voters Count']]

budget_rest.drop_duplicates(inplace=True)

budget_rest.groupby(by=['rest_name', 'location', 'Rest_type']).agg(
    {'Average Cost for two': 'mean', 'Ratings out of 5': 'mean', 'Voters Count': 'sum'})
```

✓ 0.0s

Fig-10

			Average Cost for two	Ratings out of 5	Voters Count
rest_name	location	Rest_type			
AB's - Absolute Barbecues	BTM	Casual Dining	400.0	4.9	19317
Arbor Brewing Company	Brigade Road	Pub, Microbrewery	400.0	4.5	25201
Barbeque Nation	Indiranagar	Casual Dining	400.0	4.7	7152
Biergarten	Whitefield	Microbrewery, Pub	400.0	4.7	7064
Big Pitcher	Old Airport Road	Pub, Microbrewery	400.0	4.7	9041
Church Street Social	Church Street	Lounge	400.0	4.3	7584
Fenny's Lounge And Kitchen	Koramangala 7th Block	Bar, Casual Dining	400.0	4.5	12762
Hard Rock Cafe	St. Marks Road	Casual Dining, Bar	400.0	4.5	15828
Hoot	Sarjapur Road	Microbrewery, Pub	400.0	4.2	14523
Prost Brew Pub	Koramangala 4th Block	Pub, Microbrewery	400.0	4.5	23581
Smoke House Deli	Indiranagar	Casual Dining	400.0	4.6	5446
TBC Sky Lounge	HSR	Casual Dining, Bar	400.0	4.7	6745
The Black Pearl	Koramangala 5th Block	Casual Dining, Bar	400.0	4.7	20893
	Marathahalli	Casual Dining, Bar	400.0	4.8	7023
The Boozy Griffin	Koramangala 5th Block	Casual Dining, Pub	400.0	4.6	5015
Toit	Indiranagar	Microbrewery	400.0	4.7	14956
Vapour Pub & Brewery	Indiranagar	Microbrewery, Pub	400.0	4.2	13905
Windmills Craftworks	Whitefield	Microbrewery, Pub	400.0	4.6	11782

Fig-11

11. We also found out the top quick bites restaurant chains in Bangalore.

13. Top quick bites restaurant chains in Banglore.

```
top_quick_bite_rest = df[(df['Rest_type'] == 'Quick Bites') & (df['Ratings out of 5'] > 4) & (
    df['Voters Count'] > 1000)].loc[:, ['rest_name', 'location', 'Average Cost for two', 'Ratings out of 5', 'Voters Count']]

top_quick_bite_rest.drop_duplicates(inplace=True)

top_quick_bite_rest.groupby(by=['rest_name', 'location']).agg(
    {'Average Cost for two': 'mean', 'Ratings out of 5': 'mean', 'Voters Count': 'sum'})
```

✓ 0.0s

		Average Cost for two	Ratings out of 5	Voters Count
rest_name	location			
CTR	Malleshwaram	150.0	4.7	4408
Eat Street	Koramangala 6th Block	600.0	4.2	3065
Kabab Magic	Basavanagudi	400.0	4.1	1720
Taco Bell	Koramangala 6th Block	600.0	4.1	3922
Veena Stores	Malleshwaram	150.0	4.5	2416

Fig-12

12. We also analysed the most popular casual dining restaurant chains by plotting it on the graph.

14. Which are the most popular casual dining restaurant chains, Make use of any plot related to this question?

```
popular_casual_dining_rest = pd.DataFrame(df[df['Rest_type'] == 'Casual Dining'].groupby(
    by=['rest_name']).agg({'Average Cost for two': 'mean', 'Ratings out of 5': 'mean', 'Voters Count': 'sum'}))

popular_casual_dining_rest.reset_index(inplace=True)

df2 = popular_casual_dining_rest[(popular_casual_dining_rest['Ratings out of 5'] > 4.3) & (
    popular_casual_dining_rest['Voters Count'] > 4000)]

df2['popularity'] = df2['Ratings out of 5'] * df2['Voters Count']
```

Fig-13

```
df2.plot(kind='bar', x='rest_name', y='popularity', xlabel="Restaurant Name",
         ylabel="Popularity", title="Most popular 'Casual Dining' restaurant chains", figsize=(15,5))

plt.xticks(rotation=60)

plt.show()
```

0.3s

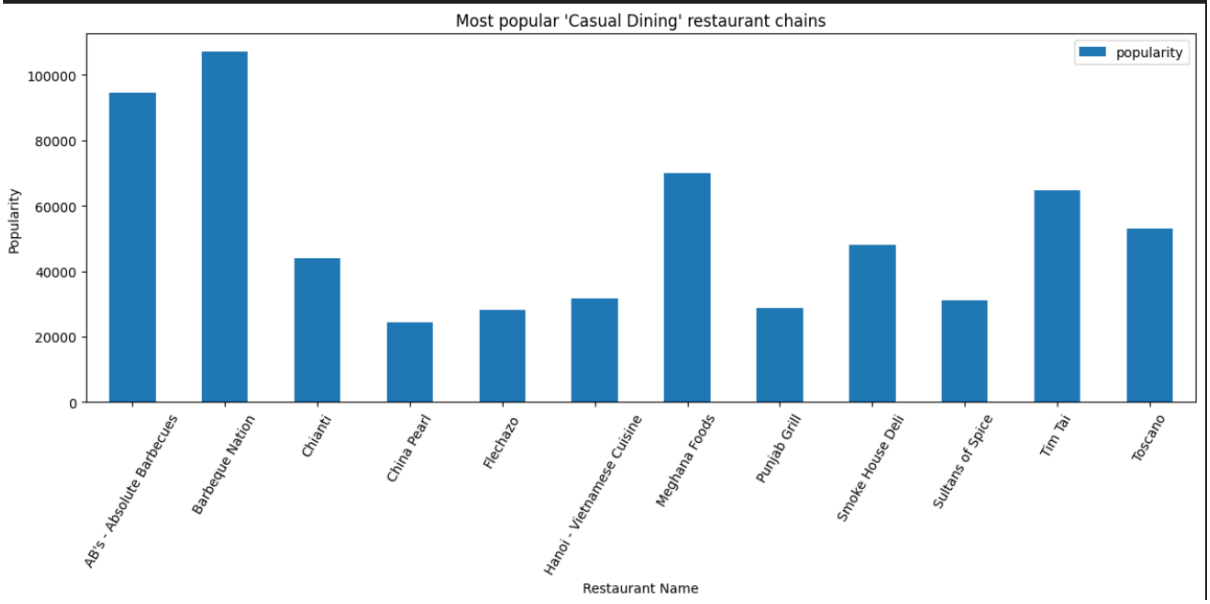

Fig-14

13. We also analysed the most popular preferred cuisines of Bangalore city and plotted it on graph.

Q15. Which are the most popular cuisines of Bangalore using a related plot?

Empty markdown cell, double-click or press enter to edit.

```
cuisine_list = df['Cuisines'].str.split(',\s*').explode()

total_count_cuisine = cuisine_list.value_counts()

total_count_cuisine

total_count_cuisine.sort_values(ascending=False).head(12).plot(
    kind='bar', xlabel= 'Cuisine Names',ylabel="Popularity", title="The most popular cuisines of the Bangalore City", figsize=(10, 8))

plt.show()
```

0.3s

Fig-15

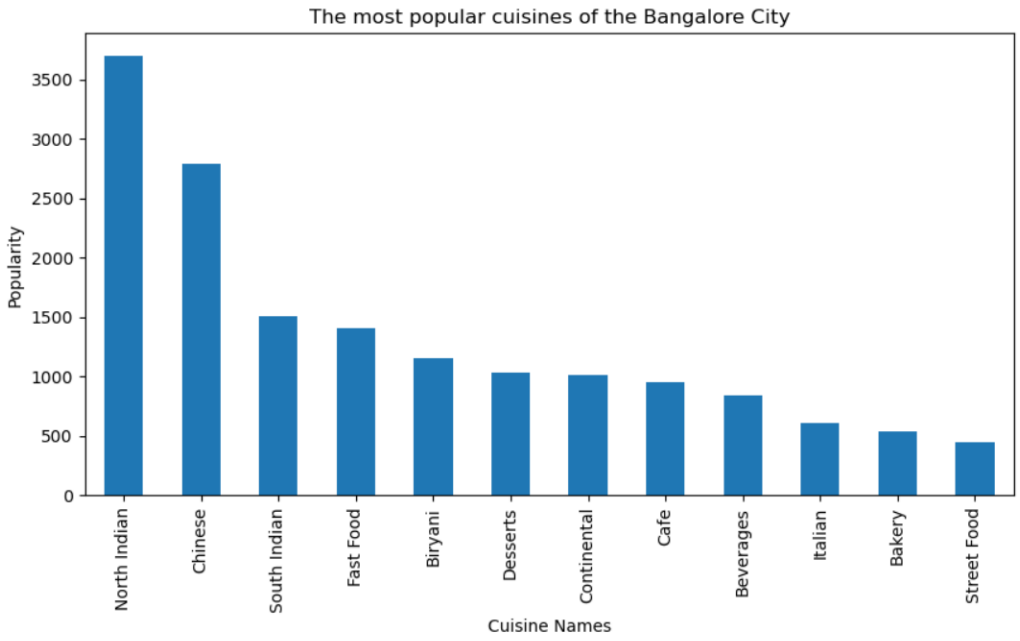

Fig-16

What could have been done better?

Looking back on our Zomato data analysis, it becomes evident that a more profound understanding of statistics and data analysis would have significantly enhanced the rigor and depth of our research. During the examination of the dataset, we came across peculiar situations where certain restaurants were assigned a rating of 2.2, despite having an accompanying voter count of 0. Naturally, such occurrences raised legitimate questions about the validity and reliability of these ratings.

In-depth knowledge of statistical principles would have allowed us to approach these anomalies more effectively. We could have utilized various statistical techniques to explore the distribution of ratings and voter counts, identifying any potential outliers or data entry errors. Moreover, we could have conducted hypothesis testing to assess whether the ratings with zero voters significantly deviated from the overall rating distribution, shedding light on the credibility of such ratings.

Additionally, early familiarity with statistical concepts would have prompted us to investigate the possibility of missing data or other underlying reasons for zero voter counts in certain instances.

By incorporating statistical analyses in our study, we could have derived more comprehensive and nuanced insights from the data. This deeper exploration could have provided valuable context and contributed to a more robust understanding of the relationships between ratings, voter counts, and other relevant variables.

Overall, having a solid foundation in statistics and data analysis would have empowered us to conduct a more insightful and rigorous Zomato data analysis, enabling us to draw more accurate conclusions and make well-informed recommendations based on the findings.

What did you learn from solving and how do you plan on using it in the future?

In the course of conducting the "Zomato Restaurant Chain Analysis," we, as Data Analysts, gained invaluable insights and experienced a robust problem-solving approach. Throughout the analysis, we realized the paramount importance of working with clean datasets, as we witnessed firsthand how unprocessed and raw data could lead to erroneous outcomes, rendering our analysis ineffective. The significance of data cleaning and preparation became evident, setting the foundation for accurate and meaningful analysis.

To better understand the dataset's characteristics, we employed the ".describe()" function, which allowed us to gain essential statistical insights and identify potential outliers within the data. This helped us make informed decisions and adjustments during subsequent analysis steps.

Furthermore, by utilizing the ".info()" function, we successfully extracted information about the data types of different columns. Armed with this knowledge, we made necessary data type changes to specific columns, thereby ensuring an efficient and coherent analysis process.

An essential aspect of our analysis involved identifying and handling missing values. By implementing strategies to filter out these missing entries, we ensured data integrity and completeness. Additionally, we effectively cleaned irrelevant or garbage values present in numerous categorical columns, thereby enhancing the quality and reliability of our analysis.

Throughout the analysis, we embraced various powerful libraries, such as "Matplotlib," "Seaborn," and "re" (Regular Expressions), which enabled us to visualize data effectively, uncover trends, and perform advanced data manipulations. For instance, we leveraged the "explode" function to obtain meaningful value counts for Cuisines, providing us with a deeper understanding of popular restaurant offerings.

In conclusion, the Zomato Restaurant Chain Analysis served as a profound learning experience, significantly upgrading our data analysis skills. As Data Analysts, we now possess a more refined problem-solving mindset, equipped with the knowledge and tools to approach complex datasets effectively. The insights gained from this analysis will undoubtedly prove invaluable in future projects, enabling us to deliver actionable recommendations and drive informed decision-making. Our commitment to continuous improvement ensures that we remain adept at tackling real-world data challenges, fostering success and impactful analysis in diverse domains.