

COGNIFYZ TECHNOLOGIES

NAGPUR, MAHARASHTRA

Internship Program Data Analysis On Restaurant Dataset

Submitted By: YUMNAM PREMKUMAR SINGH

Ref.: CTI/A1/C115425

Tools & Technologies:

In this Data Analysis Internship program, I used the following tools and technologies:

- 1) Pandas: For data analysis, cleaning, exploration, and manipulation.
- 2) NumPy: For efficient numerical computations.
- 3) Matplotlib: For data visualization, including plots, charts, and graphs.
- 4) Seaborn: For statistical data visualization, building on Matplotlib.
- 5) Counter (Collections): For counting hash able objects, such as frequency analysis.
- 6) Scikit-learn:
 - DBSCAN: Density-based clustering
 - StandardScaler: Feature scaling (standardization).
- 7) Jupyter Notebook (IDEs): As the primary environment for data processing and analysis.

Why DBSCAN instead of K-Means?

1. Density-based clustering: DBSCAN groups data points based on density, not just distance.
2. Arbitrary cluster shapes: DBSCAN can identify clusters of varying shapes, whereas K-Means prefers circular shapes.
3. Noise detection: DBSCAN automatically identifies outliers (noise) that don't belong to any cluster.

LEVEL 1:

Task 1: Top Cuisines

Goal

1. Identify the top three most common cuisines in the dataset.
2. Calculate the percentage of restaurants serving each of these cuisines.

Results

1. Top three most common cuisines:
 - North Indian: 3,960 restaurants (41.50%)
 - Chinese: 2,735 restaurants (28.66%)
 - Fast Food: 1,986 restaurants (20.81%)

Task 2: City Analysis

Goal

1. Identify the city with the highest number of restaurants in the dataset.
2. Calculate the average rating for restaurants in each city.
3. Determine the city with the highest average rating.

Results

1. City with the highest number of restaurants: New Delhi (5473 restaurants)
2. Average rating for restaurants in each city:

	City	Average Rating
43	Faridabad	1.866932
88	Noida	2.036204
78	Mc Millan	2.400000
82	Montville	2.400000
87	New Delhi	2.438845
.....
75	Mandaluyong City	4.625000
94	Pasig City	4.633333
73	Makati City	4.650000
106	Quezon City	4.800000
56	Inner City	4.900000

3. City with the highest average rating: Inner City (4.90)

Task 3: Price Range Distribution

Goal

1. Create a histogram or bar chart to visualize the distribution of price range among the restaurants.
2. Calculate the percentage of restaurants in each price range category.

Results

1. Price range distribution:
 - Price Range 1: 46.51% of restaurants
 - Price Range 2: 32.62% of restaurants
 - Price Range 3: 14.72% of restaurants
 - Price Range 4: 6.14% of restaurants

Task 4: Online Delivery Analysis

Goal

1. Determine the percentage of restaurants that offer online delivery.
2. Compare the average ratings of restaurants with and without online delivery.

Results

1. Percentage of restaurants that offer online delivery: 25.69%
 2. Average ratings:
 - With online delivery: 3.25
 - Without online delivery: 2.46
-

LEVEL 2:

Task 1: Restaurant Ratings

Goal

1. Analyze the distribution of aggregate ratings.
2. Determine the most common rating range.
3. Calculate the average number of votes received by restaurants.

Results

1. Distribution of aggregate ratings:

0.0	2148	2.9	381	4.1	274
1.8	1	3.0	468	4.2	221
1.9	2	3.1	519	4.3	174
2.0	7	3.2	522	4.4	143
2.1	15	3.3	483	4.5	95
2.2	27	3.4	495	4.6	78
2.3	47	3.5	480	4.7	41
2.4	87	3.6	458	4.8	25
2.5	110	3.7	427	4.9	61
2.6	191	3.8	399		
2.7	250	3.9	332		
2.8	315	4.0	266		

Name: Aggregate
rating, dtype: int64

2. The most common rating range is 0.0 with 2148 restaurants
3. The average number of votes received by restaurants is 156.77

Task 2: Cuisine Combinations

Goal:

1. Identify the most common combinations of cuisines in the dataset.
2. Determine if certain cuisine combinations tend to have higher ratings.

Results

1. Most common cuisine combinations:

North Indian	936
North Indian, Chinese	511
Chinese	354
Fast Food	354
North Indian, Mughlai	334
...	...
Bengali, Fast Food	1
North Indian, Rajasthani, Asian	1
Chinese, Thai, Malaysian, Indonesian	1
Bakery, Desserts, North Indian, Bengali, South Indian	1
Italian, World Cuisine	1

2. Top 10 most common cuisine combinations with higher average ratings:

Cuisine Combination	No_of_Cuisine	Average Rating
North Indian	936	1.67
North Indian, Chinese	511	2.42
Chinese	354	2.04
Fast Food	354	2.12
North Indian, Mughlai	334	2.89
Cafe	299	2.89
Bakery	218	1.92
North Indian, Mughlai, Chinese	197	2.57
Bakery, Desserts	170	2.32
Street Food	149	2.16

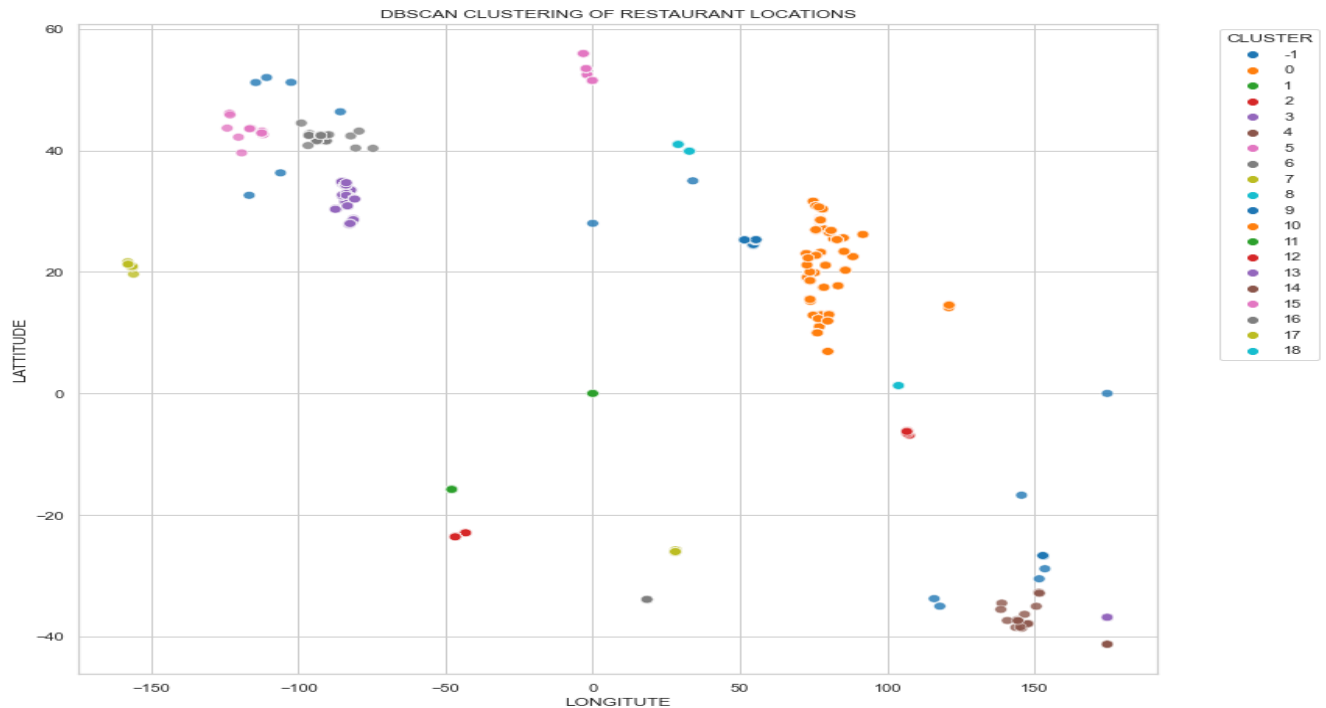
Task 3: Geographic Analysis

Goal

1. Plot the locations of restaurants on a map using longitude and latitude coordinates.
2. Identify any patterns or clusters of restaurants in specific areas.

Results

DBSCAN Clustering of Restaurant Locations is shown in the figure:



- X-axis: Longitude
- Y-axis: Latitude
- Colored dots: DBSCAN cluster labels (0-18)
- Gray dots (-1): Noise or outliers (restaurants not in any cluster)

This plot provides valuable insights into the geographic distribution and clustering of restaurants in the area.

Using DBSCAN clustering, I can identify spatial groupings of restaurants, which can be useful for:

- Understanding regional density
- Planning new outlets
- Identifying underserved locations

Task 4: Restaurant Chains Analysis

Goal

1. Identify if there are any restaurant chains present in the dataset.
2. Analyze the ratings and popularity of different restaurant chains.

Results

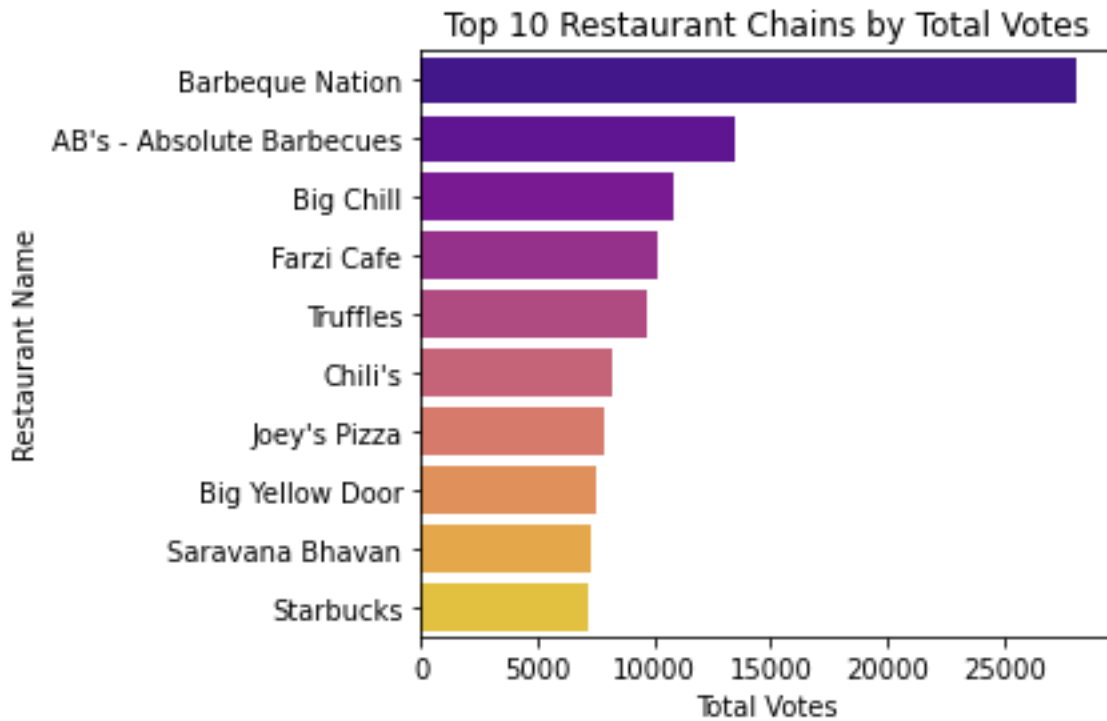
1. Restaurant chains identified:

['Cafe Coffee Day', 'Domino's Pizza', 'Subway', 'Green Chick Chop', 'McDonald's', 'Keventers', 'Pizza Hut', 'Giani', 'Baskin Robbins', 'Barbeque Nation', ... 'Nazeer Delicacies', 'Embassy', 'Delhi Biryani Hut', 'Qureshi Kabab Corner', 'KC Bakers', 'Zaika Kathi Rolls', 'The Night Owl', 'The Cheesecake Factory', 'New Kadimi', 'Ceviche Tapas Bar & Restaurant'].

2. Top 10 restaurant chains by popularity:

Restaurant Name	Aggregate Rating	Votes	Branch Count
Barbeque Nation	4.353846	28142	26
AB's - Absolute Barbecues	4.825000	13400	4
Big Chill	4.475000	10853	4
Farzi Cafe	4.366667	10098	6
Truffles	3.950000	9682	2
Chili's	4.580000	8156	5
Joey's Pizza	4.250000	7807	2
Big Yellow Door	4.266667	7511	3
Saravana Bhavan	4.133333	7238	3
Starbucks	3.805556	7139	18

For more understanding , let's see a bar plot



LEVEL 3:

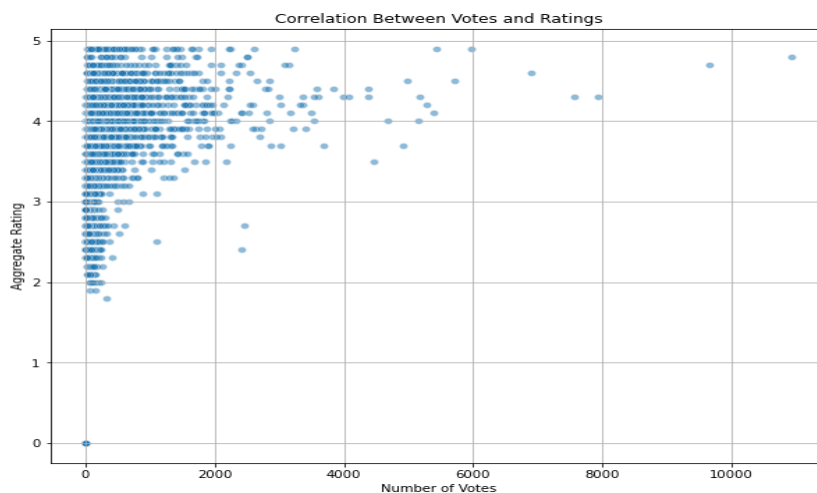
Task 2: Votes Analysis

Goal:

1. Identify the restaurants with the highest and lowest number of votes.
2. Analyze if there is a correlation between the number of votes and the rating of a restaurant.

Results

1. Restaurant with the highest number of votes: Toit (10934 votes, 4.8 rating)
2. Restaurant with the lowest number of votes: Cantinho da Gula (0 votes, 0.0 rating)
3. Correlation between votes and ratings: 0.31 (moderate positive correlation).



The above scatter plot shows that restaurants with higher ratings tend to receive more votes, but the correlation is not very strong. Some highly rated restaurants still have fewer votes.

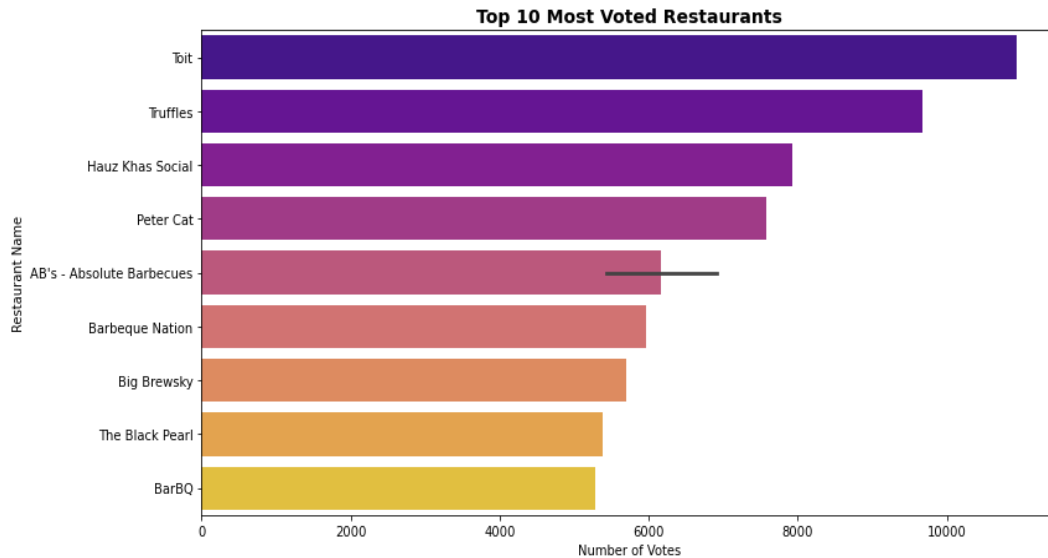


Figure: Restaurant with the highest number of votes

Task 3: Price Range vs Online Delivery and Table Booking

Goal:

1. Analyze if there is a relationship between the price range and the availability of online delivery and the table booking.
2. Determine if higher-priced restaurants are more likely to offer these services.

Results:

Online Delivery Trends:

- Mid-range restaurants (Price Range 2) are most likely to offer online delivery (41.31%).
- Higher-end restaurants (Price Range 4) rarely provide delivery service (9.04%).

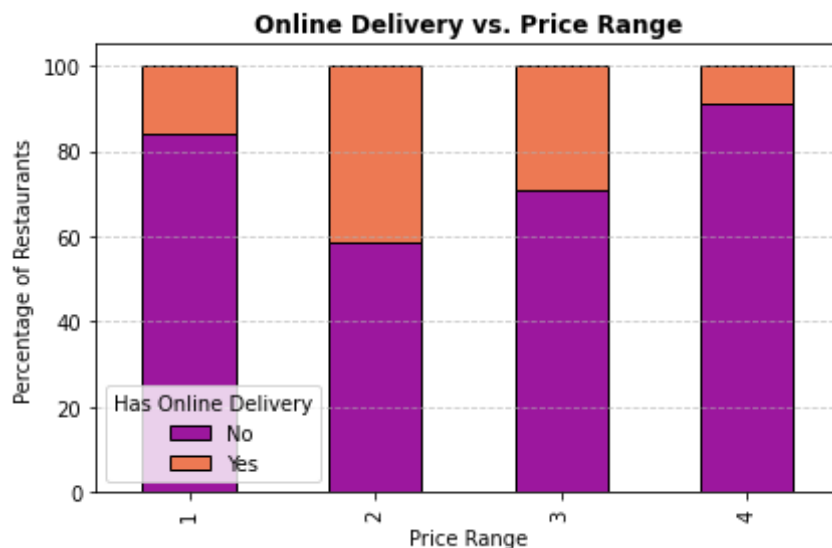


Table Booking Trends:

- Higher-priced restaurants (Price Ranges 3 & 4) are more likely to offer table booking (45.84% and 46.76%, respectively).
- Lower-priced restaurants (Price Range 1) are unlikely to offer table booking (0.02%).



The internship at Cognifyz Technologies provided a comprehensive learning experience in data analytics, blending academic concepts with real-world applications.

Through this program, I:

1. Developed hands-on expertise in Python-based data analysis.
2. Enhanced statistical analysis and data visualization skills.
3. Cultivated a structured approach to analytical problem-solving.
4. Improved ability to present complex findings in a clear, concise manner.
5. Gained insight into data-driven decision-making in the food industry.

This experience has equipped me with valuable skills and knowledge to tackle data analysis challenges. The internship's multi-level analysis approach honed my technical, analytical, and critical thinking abilities, making it a valuable step toward a data science career.

I'm grateful for the opportunity to work on this project and sharpen my data analysis skills. Thank you to Cognifyz Technologies for this amazing learning experience! I'm excited to continue growing and taking on new challenges.

Explore the full project on LinkedIn and GitHub:

- ▶ <https://www.linkedin.com/in/yumnam-premkumar-singh-6347a8145/>
- ▶ <https://github.com/premkyumnam>

I'd love to hear your thoughts and feedback!

Let's connect and explore more data-driven insights together.

Thank you!

(YUMNAM PREMKUMAR SINGH)