# Restaurant Dataset Analysis

COGNIFYZ INTERNSHIP

Binay Kumar Naik

# Project Overview – Restaurant Dataset Analysis (Cognifyz Internship)

As part of my Data Analysis Internship at Cognifyz Technologies, I undertook a comprehensive multi-level project focused on analyzing a restaurant dataset. The dataset included structured data such as restaurant names, cuisines, cities, price ranges, delivery availability, ratings, reviews, and geolocation. The goal of the project was to extract meaningful business insights through data exploration, visualization, and analytical reasoning using Python, SQL, and data visualization tools.

The internship tasks were categorized into three levels, each increasing in analytical complexity:

- Level 1: Exploratory Data Analysis & Basic Insights
- Level 2: Intermediate Analytics (Correlations, Geo-Mapping)
- Level 3: Advanced Text and Relational Insights

# Objective

The key objectives of the project were to:

- Identify the most popular cuisines and understand their distribution among restaurants.
- Evaluate city-wise restaurant availability and rating patterns to suggest potential market opportunities.
- Analyze price range distributions and assess how pricing impacts service offerings like online delivery.
- Understand consumer preferences through ratings, votes, and reviews.
- Discover geographic and operational trends through mapping and chain identification.
- Explore sentiment and behavior insights using textual review analysis and service correlations.

# Level 1 Tasks – Basic Insights

## Task 1 – Top Cuisines:

### Identified the top 3 most common cuisines.

We will calculate the top 3 cuisnies prefered by the customers through votes (higher votes means most prefered)

```python
dataset_expanded.groupby('Cuisines')['Votes'].sum().sort_values(ascending = False).head(3)
```

```
Cuisines
North Indian    595981
Chinese         364351
Italian         329265
Name: Votes, dtype: int64
```

So the top 3 cusines are "***North Indian***", "***Chinese***", "***Italian***".

So the top 3 cusines are "North Indian", "Chinese", "Italian".

# Level 1 Tasks – Basic Insights

## Task 1 – Top Cuisines:

Calculated the percentage share of restraunts that serves each of the top cuisine each among all restaurants.

We have **7746** unique restraunts

```
# We need to fillter out 7746 restraunts which restraunts servere "North Indian", "Chinese", "Italian".
filtered_restraunts = dataset[dataset['Cuisines'].isin(["North Indian", "Chinese", "Italian"])]
```

```
filtered_restraunts['Restaurant Name'].nunique()
```

```
1221
```

```
# Percentage of restraunts serving "North Indian", "Chinese", "Italian".
round((1221/7446)*100,2)
```

```
16.4
```

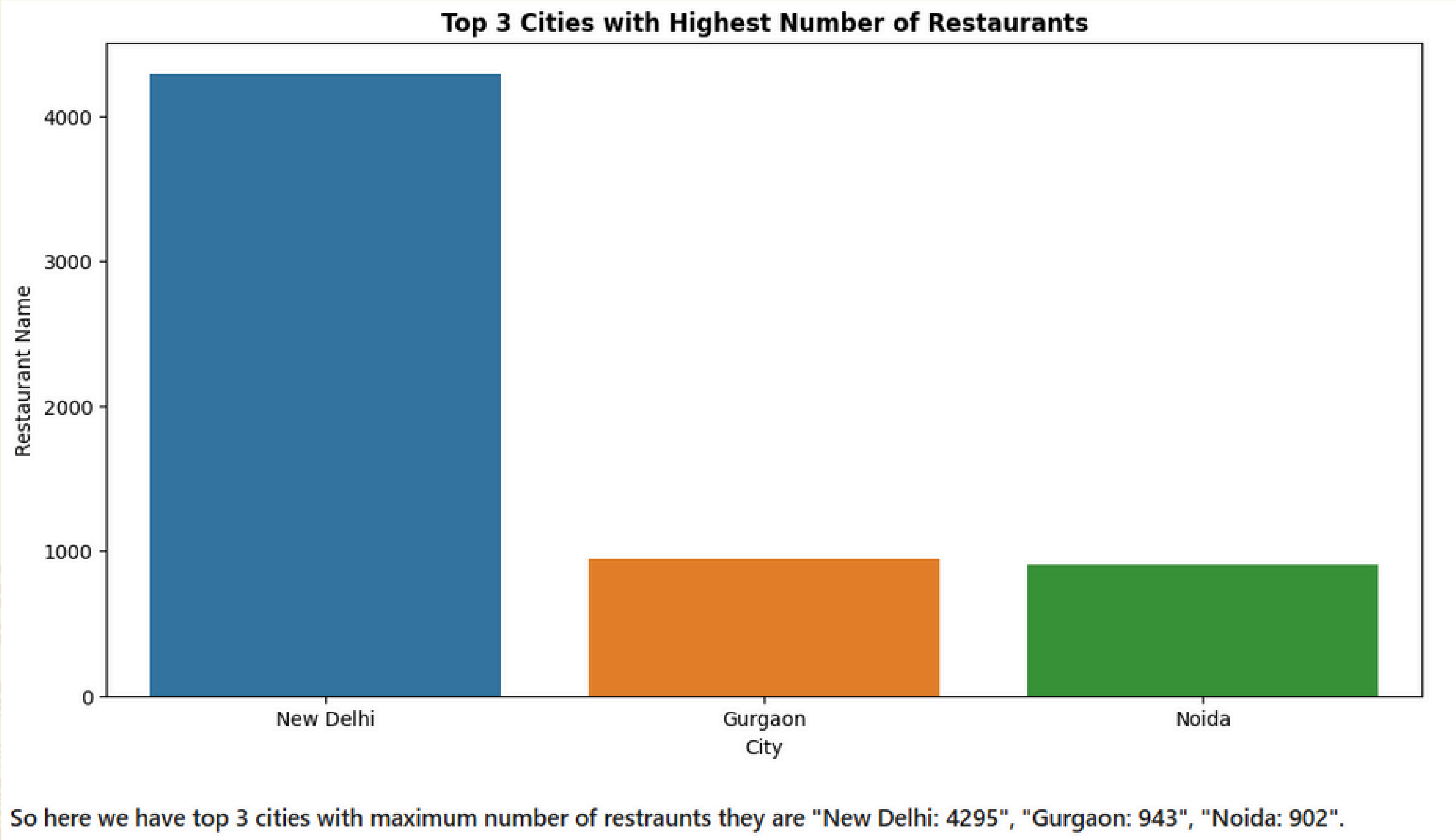
So we found 16.4 % of the total restraunts "North Indian", "Chinese", "Italian".

So we found 16.4 % of the total restraunts "North Indian", "Chinese", "Italian".

# Level 1 Tasks – Basic Insights

## Task 2 – City Analysis:

Determined the city with the highest number of restaurants.




So here we have top 3 cities with maximum number of restraunts they are "New Delhi: 4295", "Gurgaon: 943", "Noida: 902".

# Level 1 Tasks – Basic Insights

## Task 2 – City Analysis:

Calculated average restaurant ratings per city.

### 2. Calculate the average rating for restraunts in each city

```
average_rating_cities = dataset.groupby('City')['Aggregate rating'].mean().sort_values(ascending = False).round(2)
average_rating_cities
```

```
City
Inner City          4.90
Quezon City         4.80
Makati City         4.65
Pasig City          4.63
Mandaluyong City    4.62
                    ...
New Delhi           2.44
Montville           2.40
Mc Millan           2.40
Noida               2.04
Faridabad           1.87
Name: Aggregate rating, Length: 141, dtype: float64
```

# Level 1 Tasks – Basic Insights

## Task 2 – City Analysis:

Found the city with the highest average rating.

### 3. Determine the city with the highest average rating.

```
average_rating_cities = dataset.groupby('City')['Aggregate rating'].mean().sort_value
average_rating_cities
```

```
City
Inner City    4.9
Name: Aggregate rating, dtype: float64
```

"Inner City" has the highes average rating of 4.9

"Inner City" has the highes average rating of 4.9

# Level 1 Tasks – Basic Insights

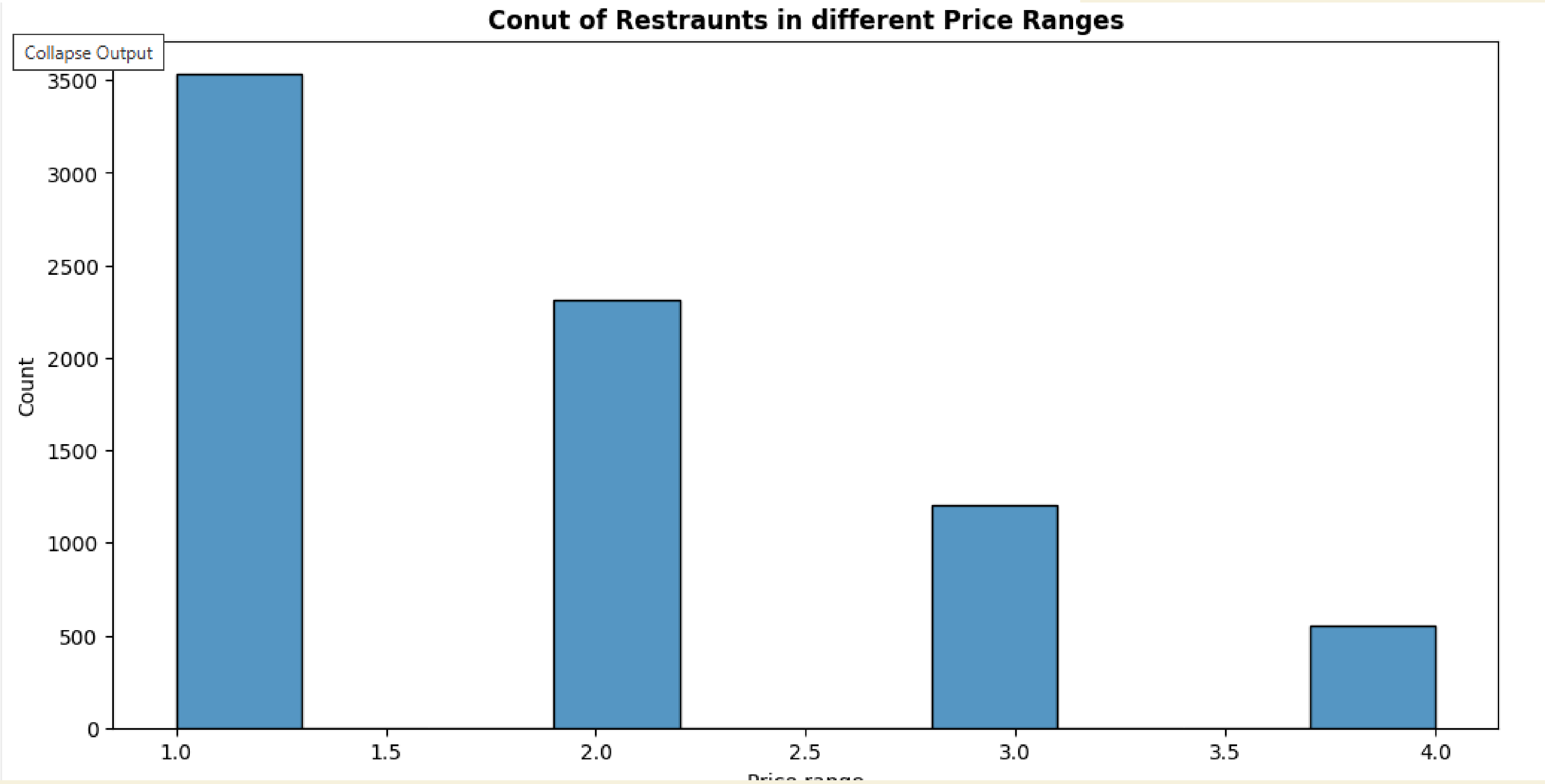
## Task 3 – Price Range Distribution:

Created a bar chart to visualize distribution across price range categories.

1. Create a histogram or bar chart to visualize the distribution of price ranges among the restaurants

```
price_range = dataset.groupby('Price range')['Restaurant Name'].nunique().sort_values(ascending = False).reset_index()
price_range.rename(columns = {'Restaurant Name':'Restaurant Count'}, inplace = True)
price_range
```

| | Price range | Restaurant Count |
|---|---|---|
| 0 | 1 | 3536 |
| 1 | 2 | 2311 |
| 2 | 3 | 1200 |
| 3 | 4 | 552 |



# Level 1 Tasks – Basic Insights

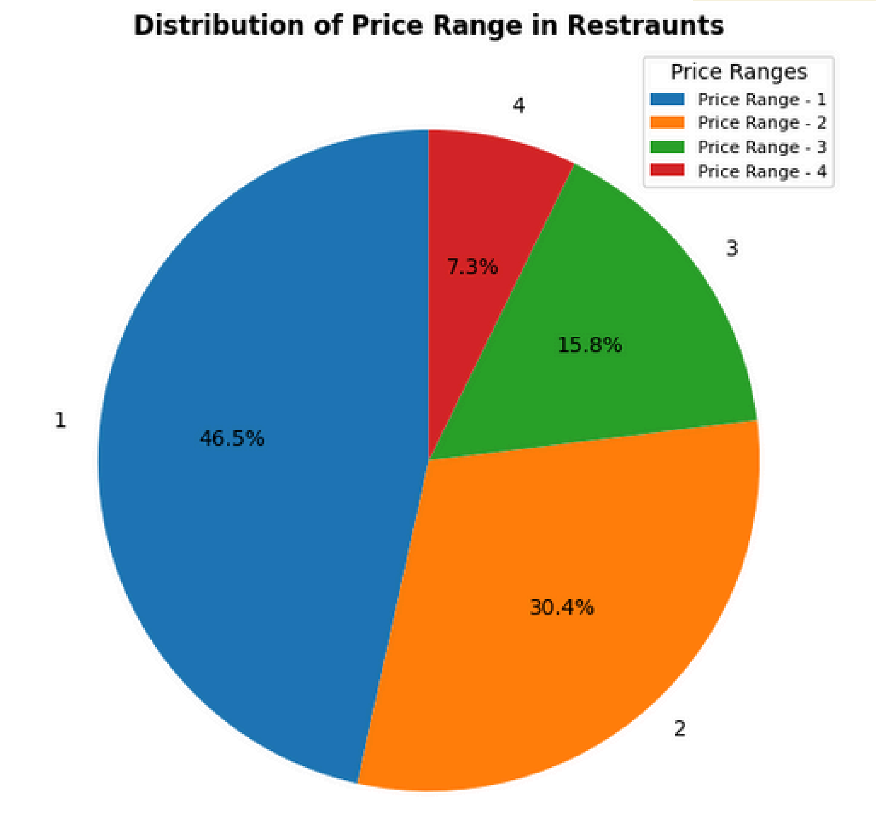
## Task 3 – Price Range Distribution:

### Computed percentage of restaurants in each price category.

2. Calculate the percentage of restaurants in each price range category.

```python
# We are trying to calculate the percentage of restaurants in each price range
price_range_distribution = (dataset.groupby('Price range')['Restaurant Name'].nunique().reset_index(name = 'Restaurant Count')
                                                                                .assign(Percentage = lambda x: (x['Restaurant Count']/x['Restaurant Count'].
                                                                                .sort_values(by = 'Percentage', ascending = False).round(2))


price_range_distribution
```

| | Price range | Restaurant Count | Percentage |
|---|---|---|---|
| **0** | 1 | 3536 | 46.53 |
| **1** | 2 | 2311 | 30.41 |
| **2** | 3 | 1200 | 15.79 |
| **3** | 4 | 552 | 7.26 |



# Level 1 Tasks – Basic Insights

## Task 4 – Online Delivery Analysis:

Calculated the proportion of restaurants offering online delivery.

### 1. Determine the percentage of restaurants that offers online delivery.

```
# dataset[['Restaurant Name', 'Has Online delivery']]
online_delivery = (dataset.groupby('Has Online delivery')['Restaurant Name'].nunique().div(len(dataset)) * 100).round(2)
online_delivery
```

```
Has Online delivery
No     62.10
Yes    18.09
Name: Restaurant Name, dtype: float64
```

18 % of restaurant provide online delivery services.

# Level 1 Tasks – Basic Insights

## Task 4 – Online Delivery Analysis:

Compared average ratings between restaurants with and without delivery options.

2. Compare the average ratings of restaurants with and without online delivery.

```
# dataset[['Restaurant Name', 'Has Online delivery', 'Aggregate rating']]
avg_rating = dataset.groupby('Has Online delivery')['Aggregate rating'].mean().reset_index(name = 'Average_rating').round(2)
avg_rating
```

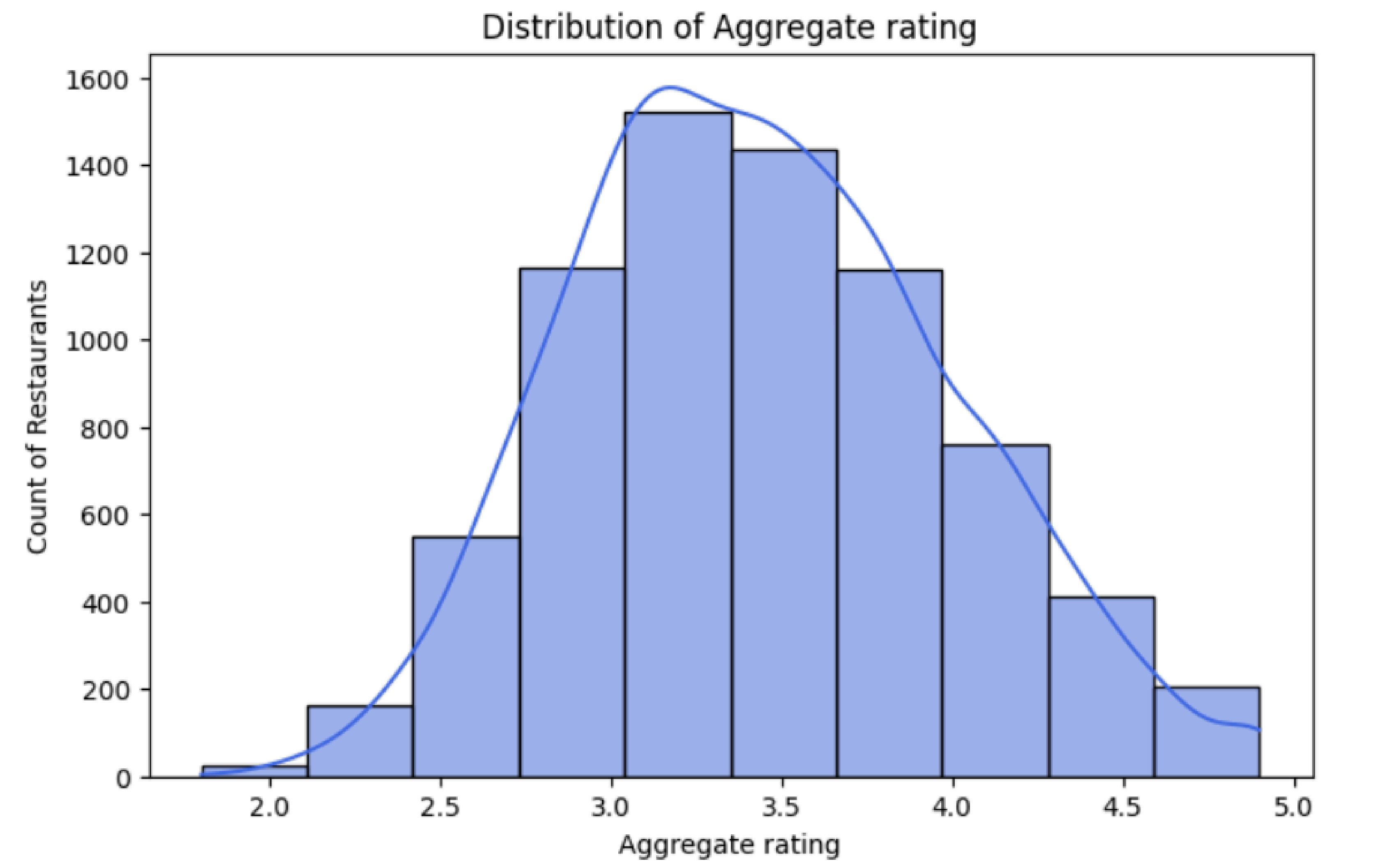
| | Has Online delivery | Average_rating |
|---|---|---|
| 0 | No | 2.47 |
| 1 | Yes | 3.25 |

Restaurants with higher average rating provide online services

# Level 2 Tasks – Intermediate Analysis

## Task 1 – Restaurant Ratings:

Analyzed rating distributions and identified the most common rating bracket.




The most common rating for the is 3.2 which is in the range (3.0 to 3.5)

# Level 2 Tasks – Intermediate Analysis

## Task 1 – Restaurant Ratings:

Explored whether certain cuisine combinations correlated with higher ratings.

| | Cuisines | Average rating |
|---|---|---|
| 53 | World Cuisine | 4.90 |
| 49 | Sunda, Indonesian | 4.90 |
| 27 | Hawaiian, Seafood | 4.90 |
| 23 | Filipino, Mexican | 4.85 |
| 13 | Chinese, Dim Sum | 4.75 |
| 31 | Indian, North Indian, Curry, Cafe | 4.70 |
| 17 | Desserts, B�_rek | 4.70 |
| 32 | Indian, Pakistani, Curry | 4.70 |
| 44 | Scottish, Cafe | 4.70 |
| 51 | Taiwanese, Street Food | 4.65 |

We can see that combination of top rated cuisines gives higher average rating.

# Level 2 Tasks – Intermediate Analysis

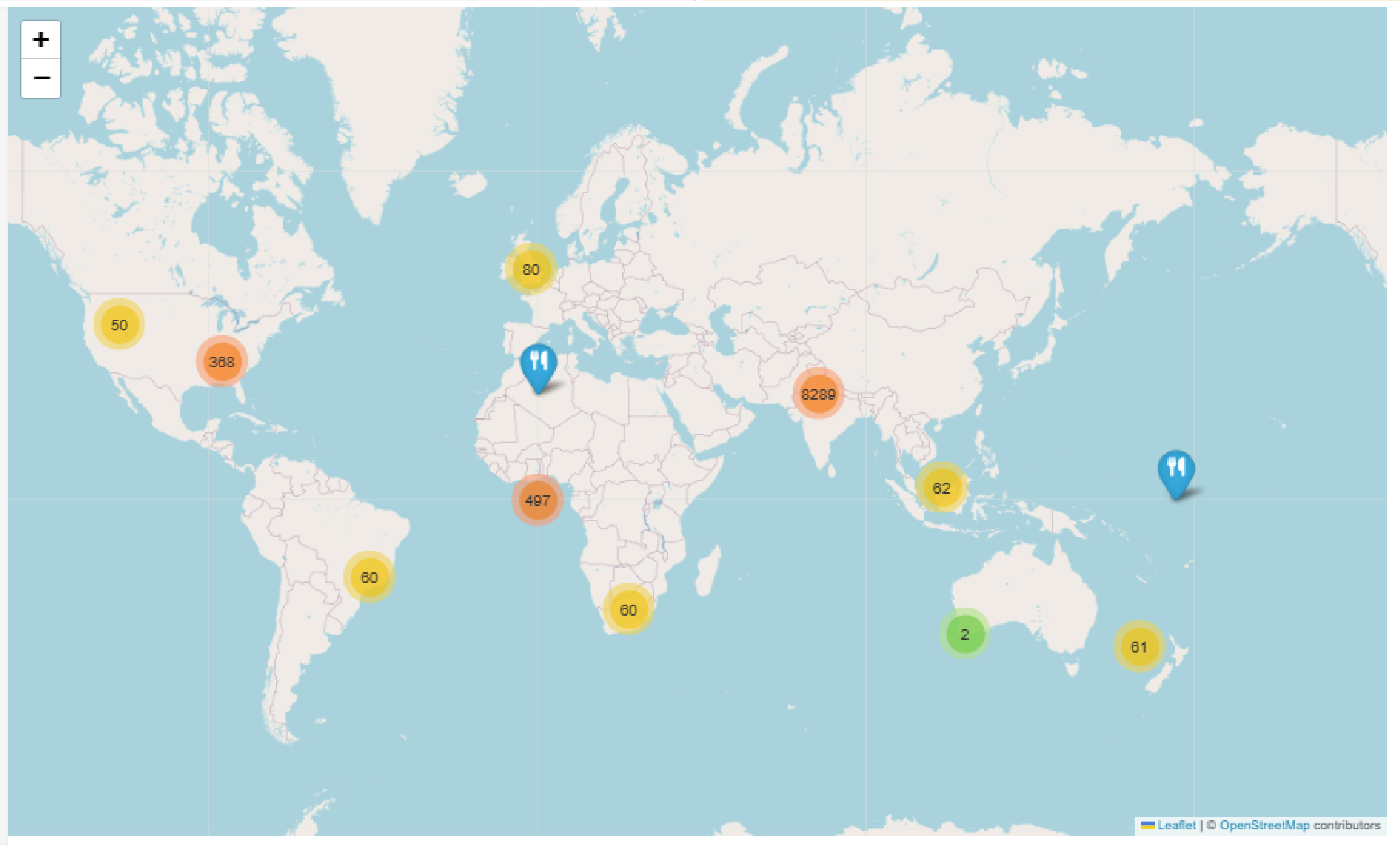
## Task 3 – Geographic Analysis:

Explored whether certain cuisine combinations correlated with higher ratings.

### TASK 3

1. Plot the locations of restaurants on a map using longitude and latitude coordinates.

```
dataset = dataset.dropna(subset=['Latitude', 'Longitude'])

map_center = [dataset['Latitude'].mean(), dataset['Longitude'].mean()]
restaurant_map = folium.Map(location = map_center, zoom_start = 2)

marker_cluster = MarkerCluster().add_to(restaurant_map)

for _, row in dataset.iterrows():
    folium.Marker(
        location = [row['Latitude'], row['Longitude']],
        popup = row['Restaurant Name'],
        icon = folium.Icon(color = 'blue', icon = 'cutlery', prefix = 'fa')
    ).add_to(marker_cluster)

restaurant_map
```



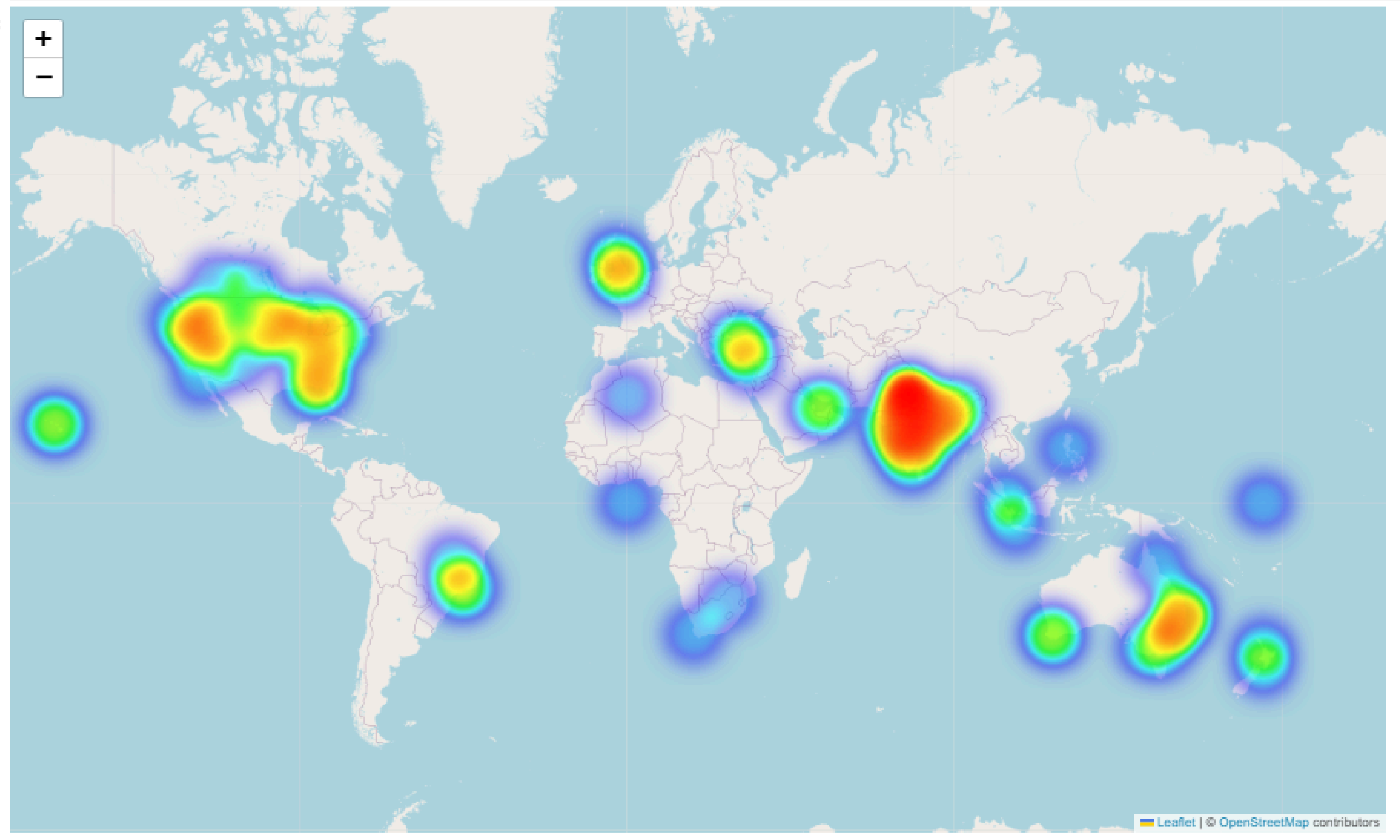
# Level 2 Tasks – Intermediate Analysis

## Task 3 – Geographic Analysis:

Identified spatial clusters and high-density zones.

2. Identify any patterns or clusters of restaurants in specific areas.

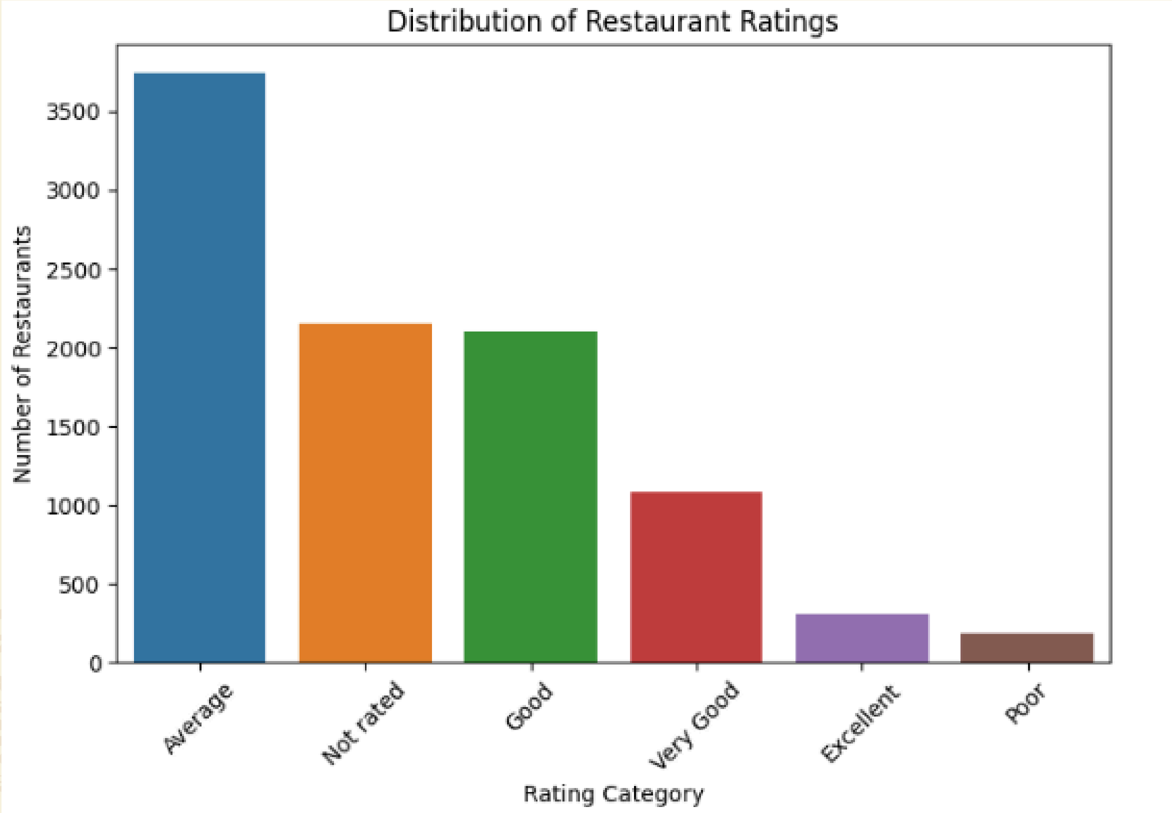
```
map_center = [dataset['Latitude'].mean(), dataset['Longitude'].mean()]
m = folium.Map(location = map_center, zoom_start = 2)

heat_data = list(zip(dataset['Latitude'], dataset['Longitude']))
HeatMap(heat_data).add_to(m)
```



# Level 3 Tasks – Advanced Insights

## Task 1 – Review Analysis:

- Conducted text analysis to extract positive/negative sentiment keywords.
- Measured average review length and its relationship to ratings.




| | review_length | Aggregate rating |
|---|---|---|
| review_length | 1.000000 | -0.599573 |
| Aggregate rating | -0.599573 | 1.000000 |

- Reveiw length does not affect aggregate rating.

- People mostly prefer giving 'Average' as the most common rating
- The most common positive would be 'Good' or 'Very Good'.
- The most common negative would be 'Poor'.

# Level 3 Tasks – Advanced Insights

## Task 2 – Votes Analysis:

- Identified highest and lowest vote-getting restaurants.
- Explored correlations between votes and ratings.

### 1. Identify the restaurants with the highest and lowest number of votes. ¶

```
max_vote_restaurants = dataset.groupby('Restaurant Name')['Votes'].mean().sort_values(ascending = False).head()
max_vote_restaurants
```

```
Restaurant Name
Toit               10934.0
Hauz Khas Social    7931.0
Peter Cat           7574.0
Big Brewsky         5705.0
The Black Pearl     5385.0
Name: Votes, dtype: float64
```

Restaurants with highest number of votes 'Toit', 'Hauz Khas Social', 'Peter Cat'

```
min_vote_restaurants = dataset[dataset['Votes'] > 0]
min_vote_restaurants.groupby('Restaurant Name')['Votes'].mean().sort_values(ascending = True).head()
```

```
Restaurant Name
4U                        1.0
Krishna Restaurant        1.0
Special Chicken Biryani   1.0
Soya Bite's               1.0
The Bakery Mart           1.0
Name: Votes, dtype: float64
```

Restaurants with lowest number of votes '4U', 'Krishna Restaurant', 'Special Chicken Biryani'

```
votes_correlation = dataset[['Votes', 'Aggregate rating']].corr()
votes_correlation
```

| | Votes | Aggregate rating |
|---|---|---|
| **Votes** | 1.000000 | 0.313691 |
| **Aggregate rating** | 0.313691 | 1.000000 |

According to the above table:

- There is a slight tendency (0.31) for restaurants with more votes to have higher ratings, but it's not a strong predictor.


- There is a slight tendency (0.31) for restaurants with more votes to have higher ratings, but it's not a strong predictor.

# Level 3 Tasks – Advanced Insights

## Task 3 – Price Range vs. Services:

- Analyzed if higher-priced restaurants were more likely to offer online delivery and table booking.



<u>From the above graphs we can conclude that</u>

- Restaurants with price range from (2 - 3) mostly have availability of online delivery
- Restaurants with price range from (3 - 4) mostly have availability of Table bookings.

# Level 3 Tasks – Advanced Insights

## Task 3 – Price Range vs. Services:

- Explored trends in service availability by price segment.

2. Determine if higher-priced restaurants are more likely to offer these services.

```
# Correlation between 'Price range', 'Online Delivery' and 'Table Booking'
price_corr = dataset[['Price range', 'Has Online delivery', 'Has Table booking']].corr()
price_corr
```

| | Price range | Has Online delivery | Has Table booking |
|---|---|---|---|
| **Price range** | 1.000000 | 0.077918 | 0.501925 |
| **Has Online delivery** | 0.077918 | 1.000000 | 0.101224 |
| **Has Table booking** | 0.501925 | 0.101224 | 1.000000 |

<u>According to the above table:</u>

- There is a strong correlation between 'Price range' and 'Table booking' (0.50) which means restaurants with higher price range tend to provide table booking services
- There is a weak correlation between 'Price range' and 'Online Delivery' (0.07) which means price range does not affect if the restaurants have online delivery as a feature.

# Thank You

You can check my github link for the detailed projrct:

https://github.com/Binay005X/Restaurant-Dataset-Analysis