Cognifyz

17/02/2025

INTERNSHIP REPORT

LEVEL 1

PREPARED FOR :

Cognifyz Technologies

COGNIFYZ TECHNOLOGIES DATA SCIENCE INTERNSHIP LEVEL 1 REPORT

Level 1 Objectives

This level focuses on data exploration and data analysis of a restaurant dataset. The level comprises three key tasks:

1. Data Exploration and Preprocessing
2. Descriptive Analysis, and
3. Geospatial Analysis.

Task 1: Data Exploration and Preprocessing

- Explore the dataset and identify the number of rows and columns.
- Check for missing values in each column and handle them accordingly.
- Perform data type conversion if necessary. Analyse the distribution of the target variable (“Aggregate rating”) and identify any class imbalances.

Task 2: Descriptive Analysis

- Calculate basic statistical measures (mean, median, standard deviation, etc.) for numerical columns.
- Explore the distribution of categorical variables like "Country Code," "City," and "Cuisines."
- Identify the top cuisines and cities with the highest number of restaurants.

Task 3: Geospatial Analysis

- Calculate basic statistical measures (mean, median, standard deviation, etc.) for numerical columns.
- Explore the distribution of categorical variables like "Country Code," "City," and "Cuisines."
- Identify the top cuisines and cities with the highest number of restaurants.

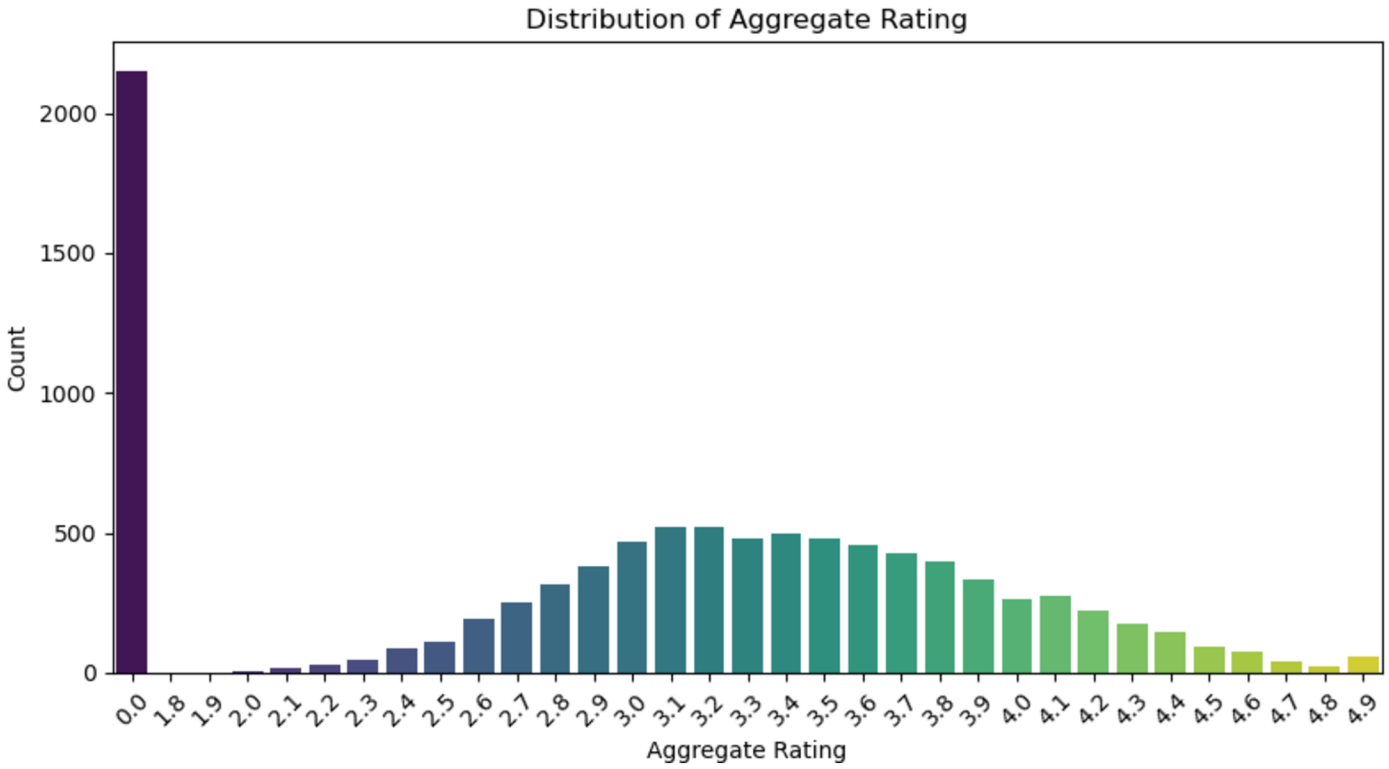
RESULTS

Task 1: Data Exploration and Preprocessing

The dataset consists of information on restaurants in different cities. It includes information such as restaurant ID, restaurant name, country code, city, address, locality, cuisines, rating, and currency among others. There are 9551 rows and 21 columns in the dataset.

The Cuisines column contains nine (9) empty values. These are very few which when removed will not affect the data hence I dropped those rows with it. There are also no duplicate values in the dataset and no data type conversion is required.

Additionally, the distribution of the target variable ("Aggregate rating") is well balanced.



Dataset contains 9551 rows and 21 columns.

Missing Values in Each Column:
Cuisines   9
dtype: int64

✅ Missing values handled. Updated dataset:
Restaurant ID   0
Restaurant Name   0
Country Code   0
City   0
Address   0
Locality   0
Locality Verbose   0
Longitude   0
Latitude   0
Cuisines   0
Average Cost for two   0
Currency   0
Has Table booking   0
Has Online delivery   0
Is delivering now   0
Switch to order menu   0
Price range   0
Aggregate rating   0
Rating color   0
Rating text   0
Votes   0
dtype: int64

Column Data Types Before Conversion:
Restaurant ID   int64
Restaurant Name   object
Country Code   int64
City   object
Address   object
Locality   object
Locality Verbose   object
Longitude   float64
Latitude   float64
Cuisines   object
Average Cost for two   int64
Currency   object
Has Table booking   object
Has Online delivery   object
Is delivering now   object
Switch to order menu   object
Price range   int64
Aggregate rating   float64
Rating color   object
Rating text   object
Votes   int64
dtype: object

📊 Basic Statistical Measures for Numerical Columns:
       Restaurant ID Country Code   Longitude   Latitude \
count   9.551000e+03   9551.000000 9551.000000 9551.000000
mean   9.051128e+06   18.365616   64.126574   25.854381
std   8.791521e+06   56.750546   41.467058   11.007935
min   5.300000e+01   1.000000 -157.948486   -41.330428
25%   3.019625e+05   1.000000   77.081343   28.478713
50%   6.004089e+06   1.000000   77.191964   28.570469
75%   1.835229e+07   1.000000   77.282006   28.642758
max   1.850065e+07   216.000000   174.832089   55.976980

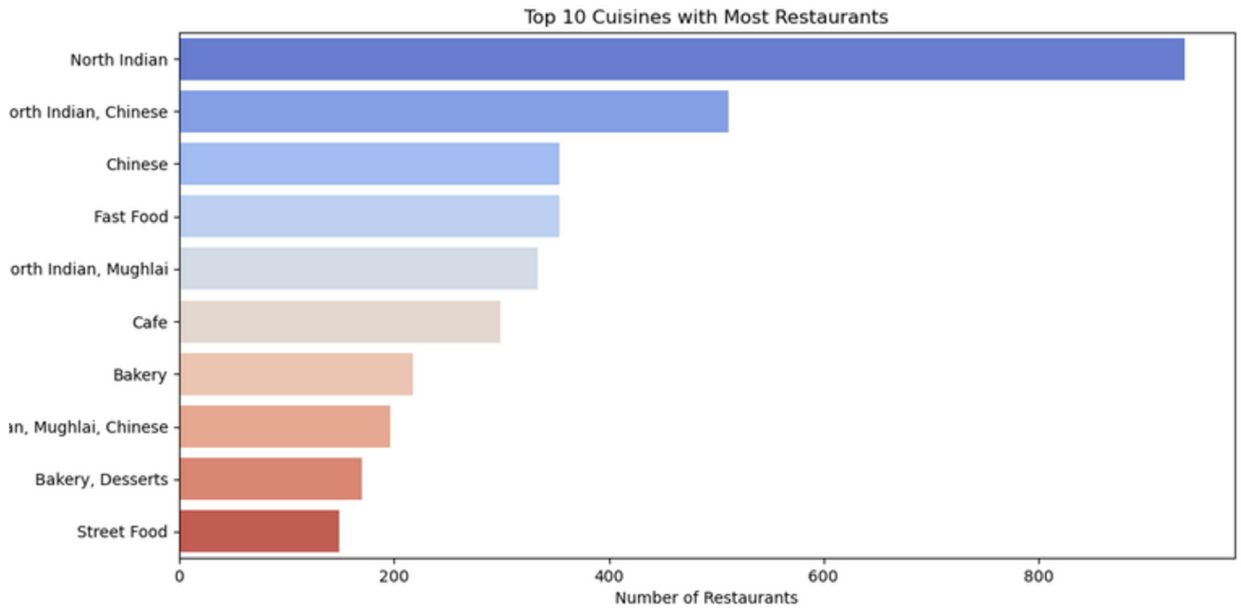
       Average Cost for two Price range Aggregate rating   Votes
count   9551.000000 9551.000000   9551.000000   9551.000000
mean   1199.210763   1.804837   2.666370   156.909748
std   16121.183073   0.905609   1.516378   430.169145
min   0.000000   1.000000   0.000000   0.000000
25%   250.000000   1.000000   2.500000   5.000000
50%   400.000000   2.000000   3.200000   31.000000
75%   700.000000   2.000000   3.700000   131.000000
max   800000.000000   4.000000   4.900000 10934.000000

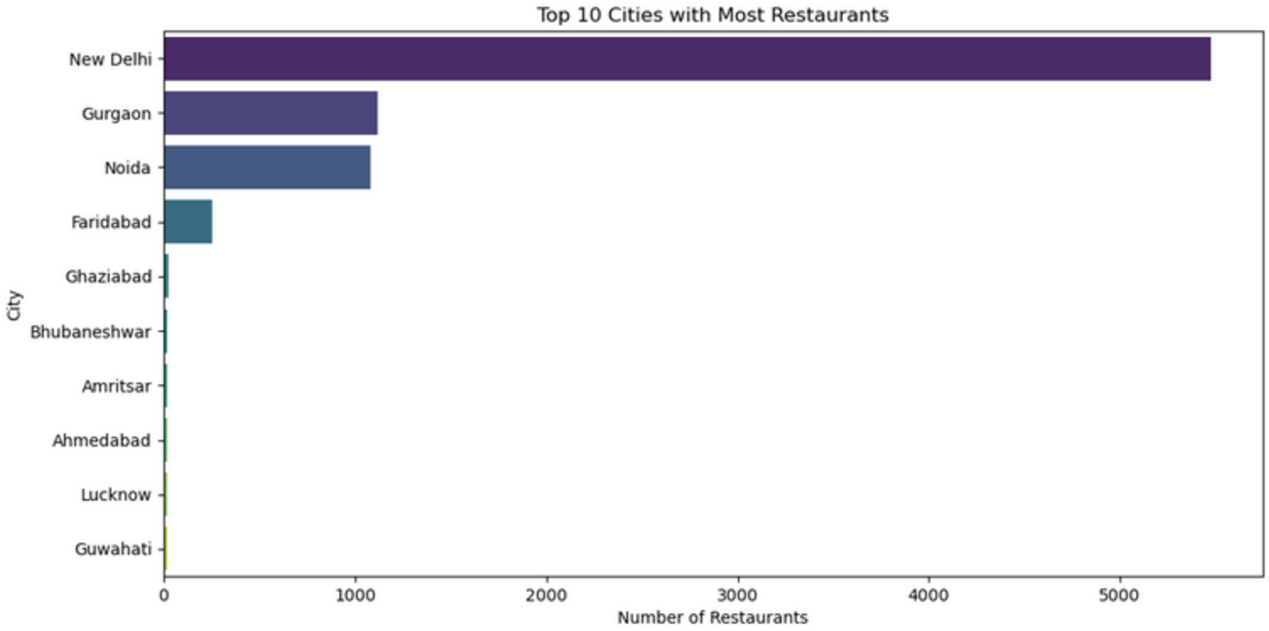
Task 2: Descriptive Analysis

The numerical columns in the dataset are restaurant ID, country code, longitude, latitude, average cost for two, price range, aggregate rating and votes. I calculated statistical measures such as the mean, the median, the standard deviation and others of these columns.

Additionally, the following were observed:

Country code 1 records the highest number of restaurants followed by 216.

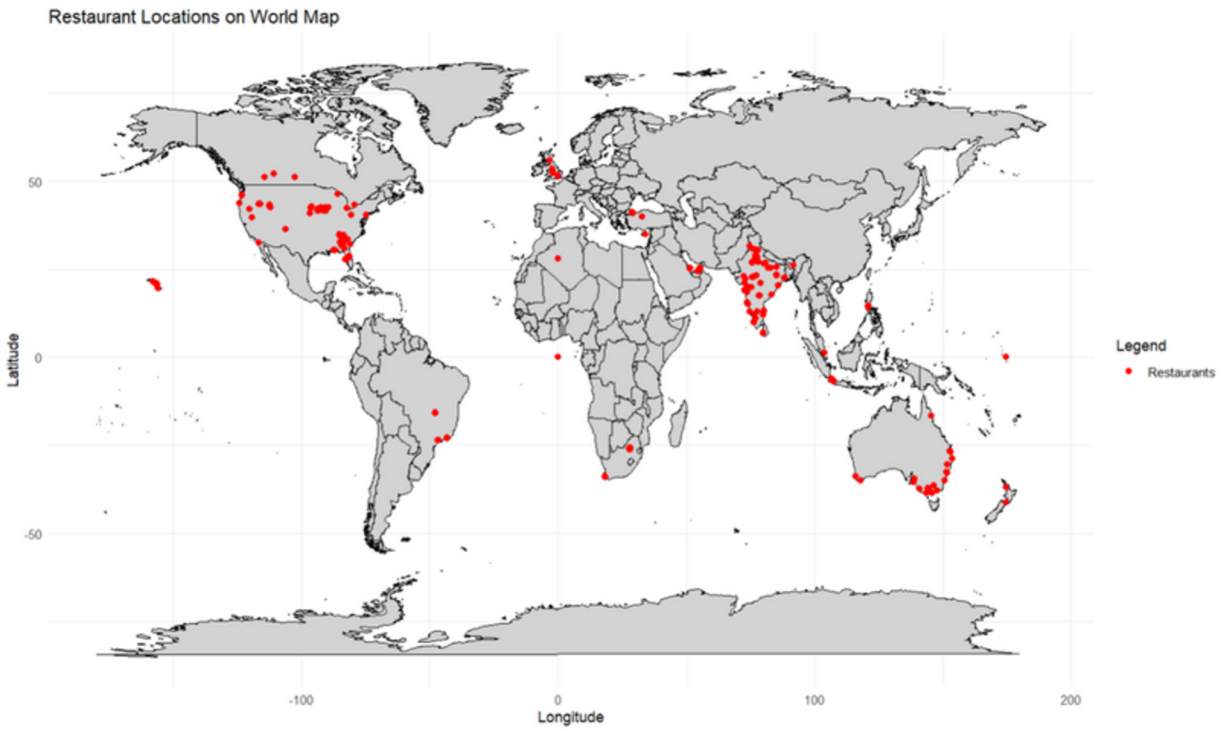




Class Distribution in 'Aggregate rating':
Aggregate rating
0.0 2148
3.2 522
3.1 519
3.4 498
3.3 483
3.5 480
3.0 468
3.6 458
3.7 427
3.8 400
2.9 381
3.9 335
2.8 315
4.1 274
4.0 266
2.7 250
4.2 221
2.6 191
4.3 174
4.4 144
2.5 110
4.5 95
2.4 87
4.6 78
4.9 61
2.3 47
4.7 42
2.2 27
4.8 25
2.1 15
2.0 7
1.9 2
1.8 1
Name: count, dtype: int64

🍽️ Top 10 Cuisines with Most Restaurants:
Cuisines
North Indian 936
North Indian, Chinese 511
Chinese 354
Fast Food 354
North Indian, Mughlai 334
Cafe 299
Bakery 218
North Indian, Mughlai, Chinese 197
Bakery, Desserts 170
Street Food 149
Name: count, dtype: int64
🌆 Top 10 Cities with Most Restaurants:
City
New Delhi 5473
Gurgaon 1118
Noida 1080
Faridabad 251
Ghaziabad 25
Bhubaneshwar 21
Amritsar 21
Ahmedabad 21
Lucknow 21
Guwahati 21
Name: count, dtype: int64
🗺️ Interactive restaurant map saved as 'restaurants_map.html'
🔥 Heatmap of restaurant locations saved as 'restaurants_heatmap.html'
🗺️ Interactive restaurant map saved as 'restaurants_map.html'
🚀 Cleaned dataset saved as 'cleaned_dataset.csv'

Task 3: Geospatial Analysis



- New Delhi has the highest number of restaurants followed by Gurgaon, Noida and Faridabad in order.

Conclusion

- This data science project has underscored the importance and effectiveness of thorough data exploration, preprocessing, descriptive analysis, and geospatial analysis in extracting valuable insights from complex datasets.
- Through exploration and preprocessing, various data quality issues such as missing values or empty values were identified and addressed, ensuring the reliability and integrity of the analysis.
- Also, the descriptive analysis part of the project provides a comprehensive understanding of the dataset's characteristics, distributions, and relationships among variables, laying the foundation for deeper insights and informed decision-making.
- Furthermore, leveraging geospatial analysis techniques uncovers the continents and cities with the highest number of restaurants helping with location-based insights.