COGNIFYZ DATA SCIENCE INTERNSHIP LEVEL 1 REPORT

About the Level

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

- Visualize the locations of restaurants on a map using latitude and longitude information.
- Analyse the distribution of restaurants across different cities or countries.
- Determine if there is any correlation between the restaurant's location and its rating.

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.

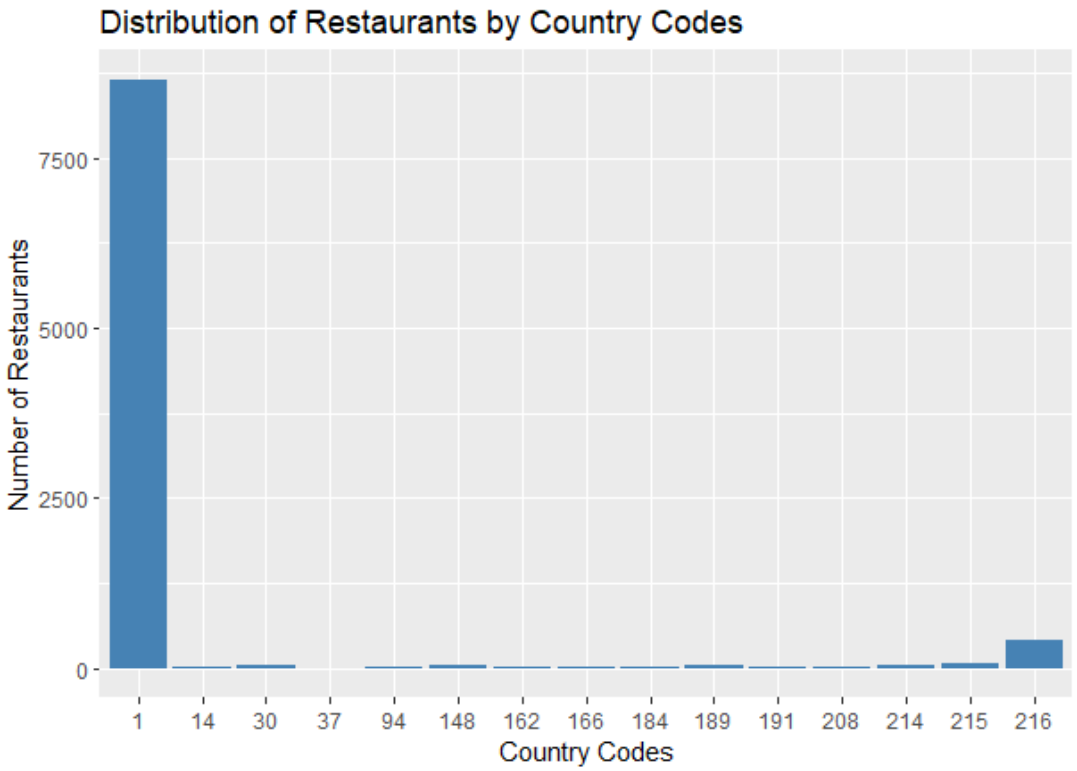
Additionally, the distribution of the target variable ("Aggregate rating") is well balanced.

Task 2: Descriptive Analysis

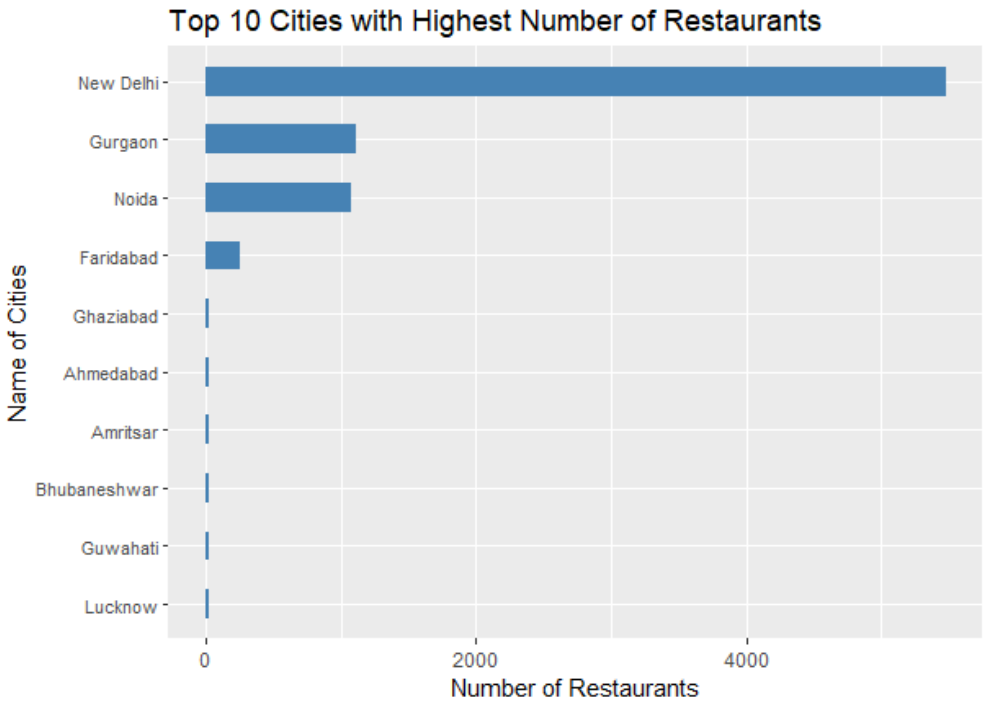
The numerical columns in the dataset are restaurant ID, country code, longitude, latitude, average cost for two, price range, aggregate rating and votes. I calculated statistical measures such as the mean, the median, the standard deviation and others of these columns.

Additionally, the following were observed:

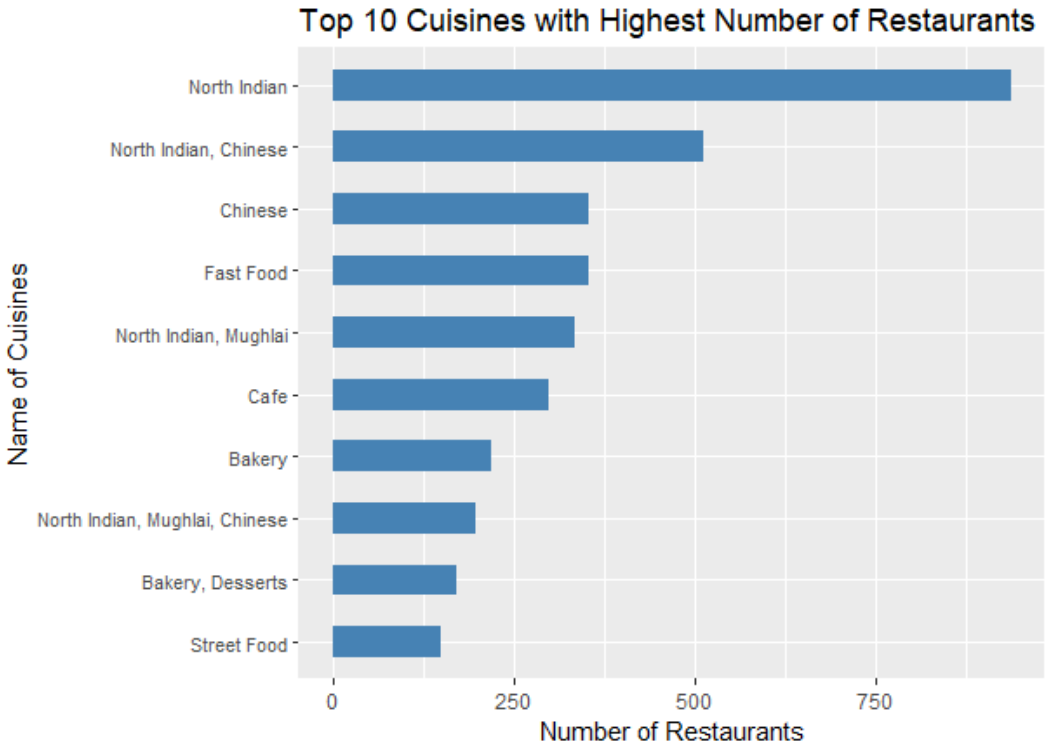
Country code 1 records the highest number of restaurants followed by 216.



New Delhi, Gurgaon and Noida are at the top with the highest number of restaurants respectively.

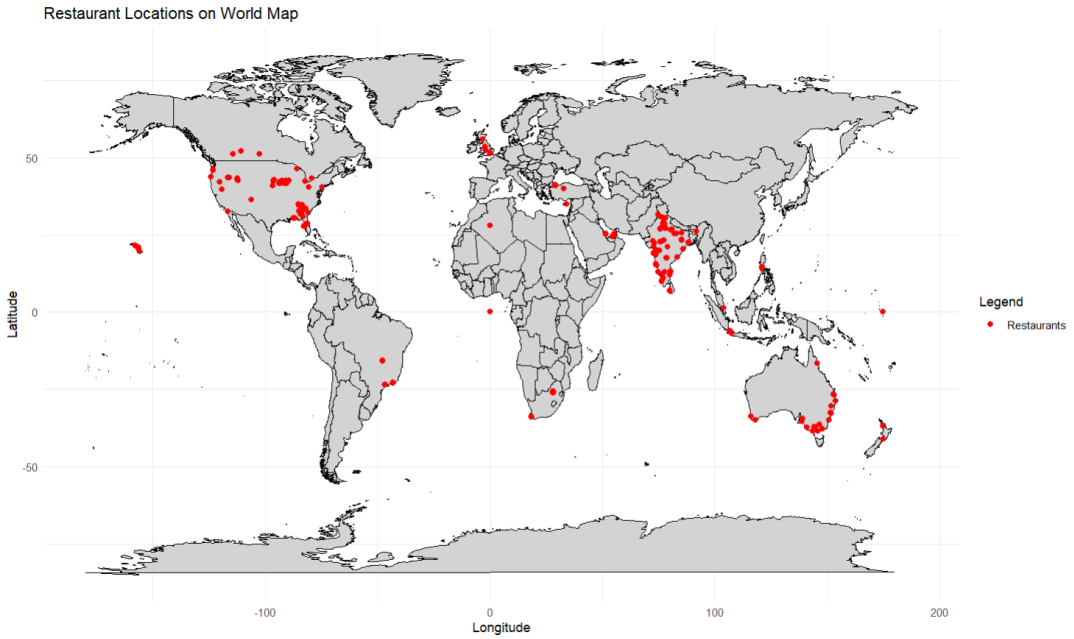


North Indian and Chinese cuisine are at the top with the highest number of restaurants.

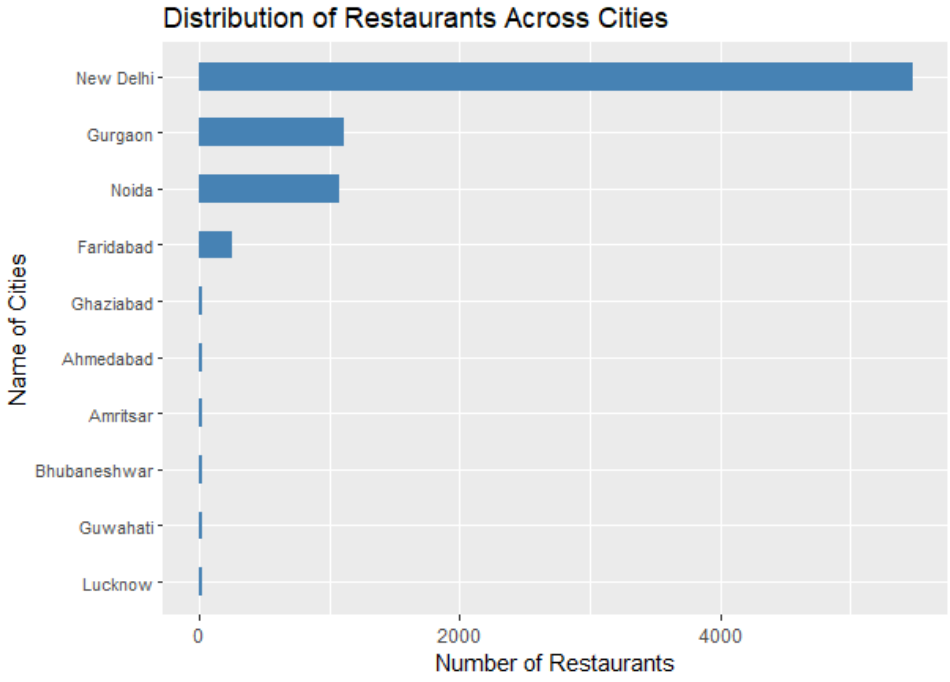


Task 3: Geospatial Analysis

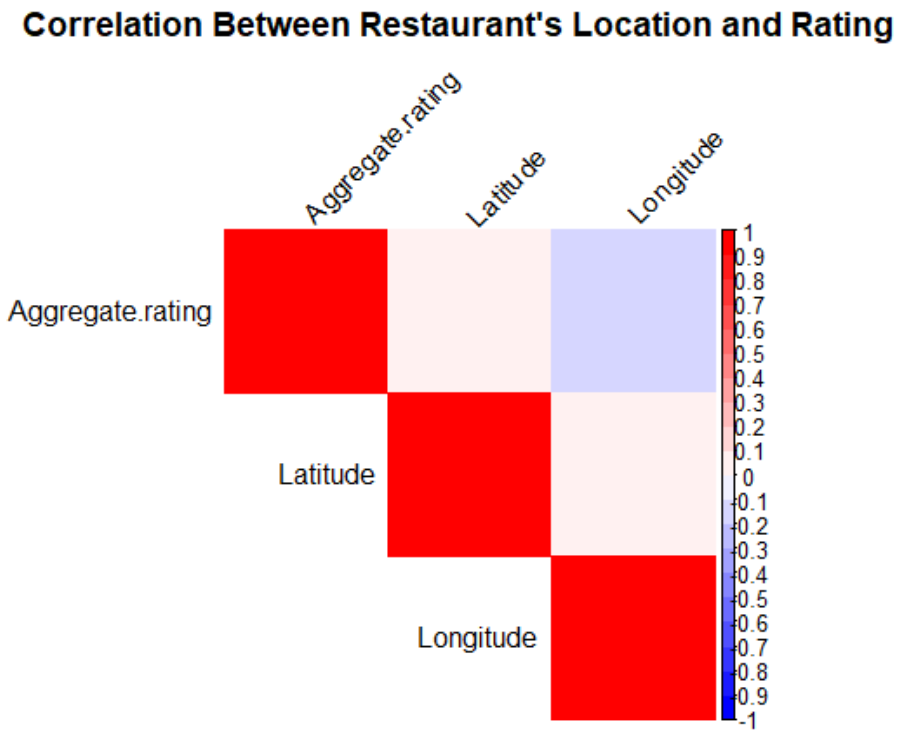
North America and Asia dominate with the highest number of restaurants followed by the Oceania continent.



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



In addition, there is no correlation between Aggregate Rating and Latitude. However, Aggregate Rating and Longitude have a weak negative correlation.



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.