Cognifyz Technologies
Level 1 – Data Science Internship Report

Intern: Mathews Henry

Duration: November 2025 – December 2025

Introduction

- Level 1 focuses on structured data exploration, cleaning, and analytical insights.
- The goal is to prepare the dataset for business analysis and ML modeling.

Task 1 – Data Exploration & Cleaning

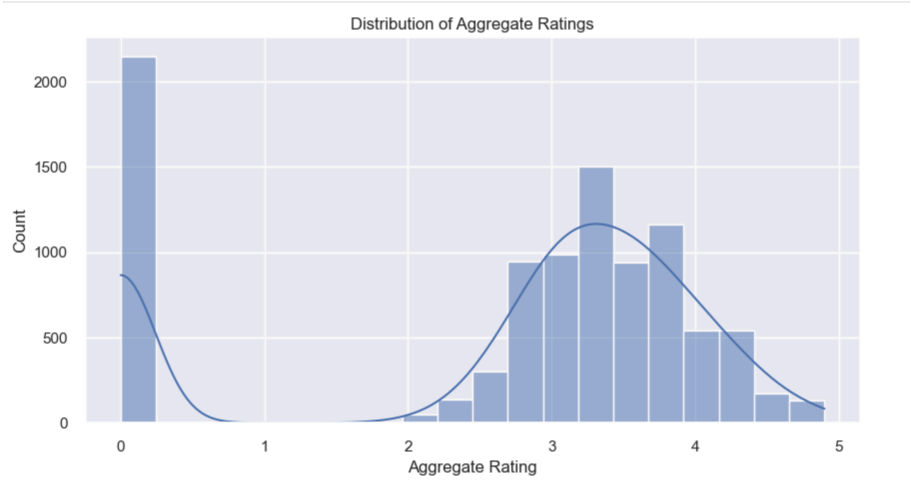
- Dataset includes 21 structured columns related to restaurant details.
- Only 'Cuisines' column had missing values (0.09%).
- Missing values were replaced with ‘Unknown’.

Data Types Analysis

- • All numeric columns (cost, rating, coordinates, votes) correctly stored.
- • All categorical variables (cuisines, city, currency) properly typed.
- • No datatype inconsistencies detected.indicating that the dataset was already well-structured.

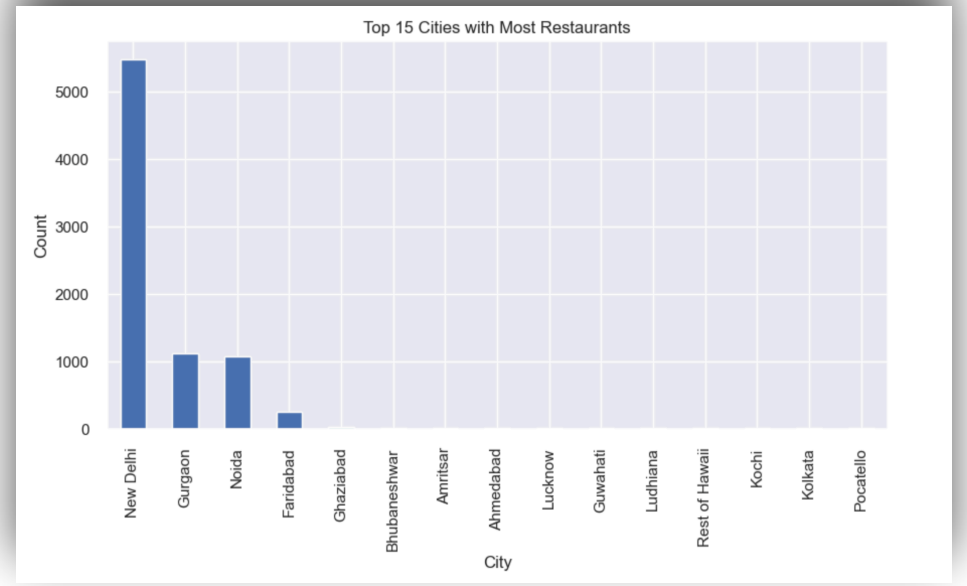
Aggregate Rating Insights

- Ratings mostly fall between 3.0 and 4.5.
- Rating 0.0 represents “Not Rated”.
- Mild class imbalance observed in the distribution.



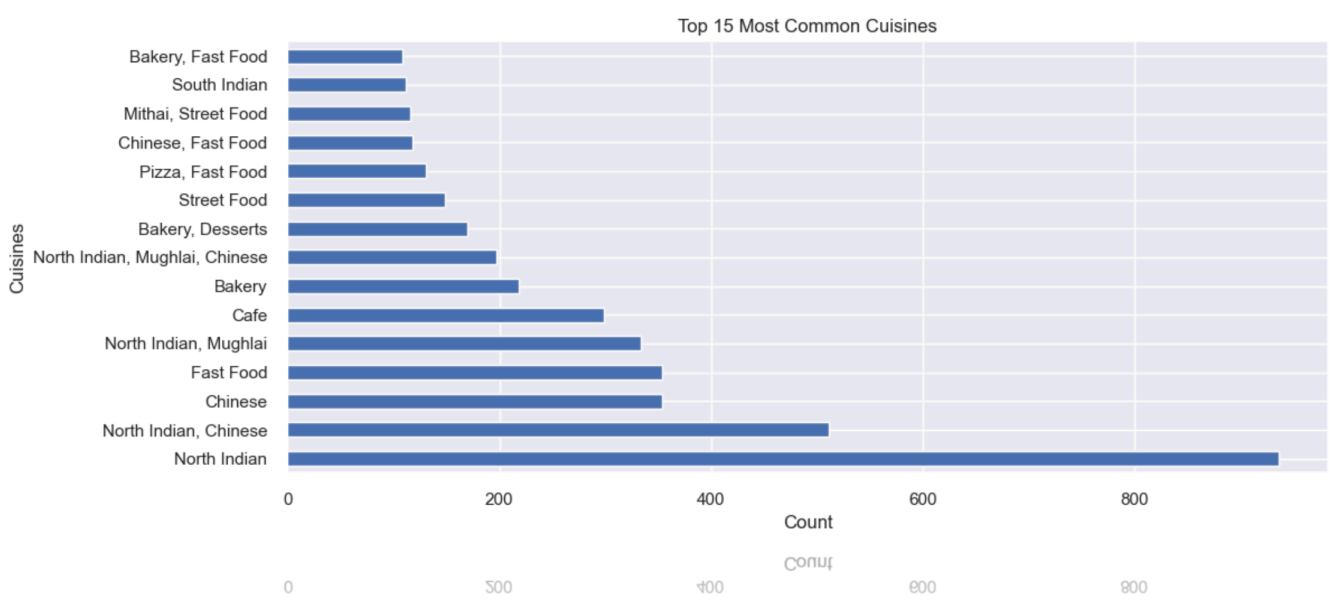
Task 2 – Descriptive Analysis

- New Delhi dominates with the highest restaurant count.
- Gurgaon and Noida follow, confirming NCR-centric dataset.
- Strong regional bias visible in city distribution.



Cuisine Analysis

- • North Indian is the most common cuisine.
- • Multi-cuisine menus are very frequent.
- • Chinese, Fast Food, and Bakery also widely represented.



Task 3 – Geospatial Analysis

- Latitude vs. Longitude plot shows dense clusters in major cities.
- No significant correlation between coordinates and ratings.
- Ratings depend more on service & food quality than raw location.

	Longitude	Latitude	Aggregate rating
Longitude	1.000000	0.043207	-0.116818
Latitude	0.043207	1.000000	0.000516
Aggregate rating	-0.116818	0.000516	1.000000

*Level*1: *Final Report*

Level 1 focused on developing a clear understanding of the dataset through exploration, cleaning, and basic analysis. The main objectives were to investigate the structure of the data, evaluate data quality, and derive initial insights about restaurant characteristics.

- The dataset contained 21 columns with information related to restaurant details
- Only the Cuisines column had minor missing values (0.09%), which were handled; all other columns had no missing data.
- Datatypes were correct and consistent
- Aggregate ratings showed a moderate imbalance, with many restaurants rated 3.0–4.5
- Descriptive analysis reveled New Delhi, Gurgaon, and Noida were the top cities with the highest number of restaurants, suggesting that the dataset is heavily centered around the NCR region.(not a good sample which represents entire restaurent population)
- Descriptive analysis also revealed popular cuisines : North Indian is the most common cuisine, followed by multi-cuisine combinations and Chinese/Fast Food.
- Geospatial visualization showed restaurant distribution and Correlation analysis showed no meaningful relationship between location coordinates and restaurant ratings.which means Ratings are influenced more by service and food quality than by raw geographic position.

Conclusion

since The dataset is largely clean, well-organized, therfore can be ready for advanced tasks such as feature engineering and predictive modeling in Level 2 and Level 3.