**📊 Level 1 – Task 1: Data Exploration and Preprocessing**

**Objective:**

The goal of this task was to:

* Understand the **structure** of the dataset.
* Clean and **preprocess** the data for future analysis.
* Analyze the **distribution** of the **Aggregate Rating** and identify **class imbalances**.

**✅ Step 1: Dataset Shape and Structure**

**Objective:**  
Understand the **size**, **columns**, and **data types** in the dataset.

**Actions Taken:**

1. Loaded the dataset (Total rows: **9551**, columns: **21**).
2. Identified the key columns:
   * **Restaurant details** (ID, Name, Address, City).
   * **Service availability** (Has Table booking, Has Online delivery).
   * **Geolocation** (Longitude, Latitude).
   * **Aggregate rating** (Target variable for analysis).
3. Verified **data types** to ensure compatibility for analysis.

**Insights:**

* The dataset is **large and diverse**, covering multiple cities and service attributes.
* Some columns needed **type correction** (e.g., converting Yes/No to binary).

**✅ Step 2: Handling Missing Values**

**Objective:**  
Identify and fix **missing** or **incorrect** data.

**Actions Taken:**

1. Checked for **missing values** using df.isnull().sum().
2. Found **0 values** in **Longitude** and **Latitude**—considered them **missing**.
3. Fixed geolocation issues by filling **0 values** with the **mean** of each city.

**Insights:**

* No traditional missing values (NaN) were present.
* **Geolocation** errors were corrected to ensure accurate mapping.

**✅ Step 3: Data Type Conversion**

**Objective:**  
Ensure each column has the **correct** data type for analysis.

**Actions Taken:**

1. Converted **Yes/No** columns to **binary (0/1)**:
   * Has Table booking
   * Has Online delivery
   * Is delivering now
2. Changed **Average Cost for two** to **integer** for better numeric handling.
3. Verified changes using df.info() to confirm correct data types.

**Insights:**

* All columns are now in the correct format for future analysis.
* Binary columns (Yes/No) are converted to **0/1** for easy calculations.

**✅ Step 4: Analyze the Target Variable ("Aggregate Rating")**

**Objective:**  
Understand the distribution of **restaurant ratings** and identify **class imbalance**.

**Actions Taken:**

1. Counted how many restaurants fall into each **rating**.
2. Analyzed the relationship between **Aggregate rating**, **Rating color**, and **Rating text**.
3. Found **class imbalance**:
   * **2148** rows (22.5%) are **Not Rated (0.0)**.
   * Very **high ratings (4.5+)** are **rare**—only **301** restaurants.
4. Decided to **keep** all rows, including **Not Rated**, for a **comprehensive** analysis.

**Insights:**

* **Average-rated** restaurants (2.5–4.0) dominate the dataset.
* A significant portion of restaurants (22.5%) have **no ratings**.
* Keeping all data ensures a **holistic** view of both rated and unrated businesses.

**📌 Summary of Actions Completed:**

1. **Data Cleaning:** Fixed missing geolocation and confirmed no other missing data.
2. **Data Type Conversion:** Ensured accurate types for proper analysis (e.g., binary for Yes/No).
3. **Target Variable Analysis:** Explored rating distribution and identified class imbalances.
4. **Decision:** Kept **all** rows for comprehensive analysis, including **"Not Rated"** entries.