Zomato Restaurants Bangalore Data Cleaning – Project Summary

Objective:

To clean and preprocess the Zomato restaurant dataset (Bangalore) by fixing missing values, handling duplicates, correcting data types, and making the dataset analysis-ready.

🔧 **Process Overview:**

1. **Data Loading & Initial Checks:**
    - Loaded the raw dataset using pandas.
    - Explored data with .info(), .nunique(), and .head() to understand structure and issues.
2. **Cleaning URLs & Identifying Duplicates:**
    - Cleaned the url column by removing extra “context” parameter.
    - Noticed multiple rows for the same restaurant (same URL) but listed under different listed_in(type) and listed_in(city) categories.
3. **Dropping Unnecessary Columns:**
    - Dropped the phone column (not useful).
    - Dropped dish_liked and menu_item due to >50% missing values.
4. **Rating Imputation :**
    - Cleaned rate values by removing /5, replacing "NEW" with 0.0, "-" with NaN, and converted to float.
    - For missing ratings, extracted rating numbers from the reviews_list (text inside review entries).
    - Left some null values to fill later for accurate imputation.
5. **Cost for Two Cleaning & Imputation:**
    - Renamed column to cost_for_two, removed commas, converted to float.
    - Left null values to fill later for accurate imputation.
6. **Restaurant Type & Cuisines Fixes:**
    - Filled missing rest_type using the mode based on the corresponding listed_in(type) group.
    - Filled missing cuisines with "Unknown".

7. **Location & listed_in(city):**
    - Filled mission locations using address column by splitting the address by commas and extracting the last second value representing the location of restaurant.
    - Dropped the address and listed_in(city) because they didn't have much significance.
8. **Grouping Duplicates:**
    - Since many restaurants had multiple rows (same URL), grouped the dataset by url.
    - Applied aggregations (like mean for ratings and cost, joined column values by comma for different listed in type) to reduce each restaurant to one row.
9. **Final Fixes:**
    - Calculated average review ratings by grouping restaurant name and used them to impute missing values.
    - For remaining ratings nulls, grouped by restaurant types and filled using grouped averages.
    - Final few nulls filled with overall mean rating.
    - Filled missing cost_for_two values by grouping restaurant type and taking their mean cost_for_two.

Final Outcome:

- Clean, complete dataset with no nulls or duplicates.
- All columns with their correct datatype.
- Ready for EDA, visualization, or modeling.