Data Collection and Preprocessing Phase

Date	4 Dec 2025
Team ID	RT
Project Title	Restaurant Recommendation System
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
Kaggle Zomato.csv	20% missing values (rate: 15%, dish_liked: 54%, cost: 10%)	High	Filled missing ratings with dataset median; dropped rows missing cuisines (essential for recommendations); reduced missing data to under 5%.
Kaggle Zomato.csv	215 duplicates (~0.4%)	Moderate	Removed all duplicate restaurant entries, retaining 51,714 unique records.
Kaggle Zomato.csv	Inconsistent formats (rate 'NEW/-', cost commas, phone NaNs)	Moderate	Standardized ratings to numeric scale (0-5); cleaned cost column by removing commas and converting to numbers; ignored irrelevant phone data.
Kaggle Zomato.csv	Outliers (rate >5/<1, cost >20k)	Low	Filtered extreme values using statistical z-score method; clipped ratings outside valid 0-5 range.
Kaggle Zomato.csv	Invalid/malformed cuisines (empty/NaN)	High	Removed restaurants without cuisine data; standardized cuisine names to lowercase and combined multiple cuisines into single searchable text field for TF-IDF processing.library