

## Data Collection and Preprocessing Phase

|               |                                  |
|---------------|----------------------------------|
| Date          | 26 May 2025                      |
| Name          | Omkar Raju More                  |
| Project Title | Restaurant Recommendation System |
| Maximum Marks | 2 Marks                          |

### Data Quality Report Template

| Data Source                                  | Data Quality Issue  | Severity | Resolution Plan   |
|--|---|----------|---|
| Dataset<br>(Restaurant reviews and metadata) | Missing values in fields like restaurant name, location, or ratings | Moderate | Perform data imputation using techniques like mean/mode for numeric values and most frequent value for categorical data. Alternatively, remove rows with critical missing fields. |

|                              |  |          |  |
|------------------------------|--|----------|--|
| Dataset<br>(User reviews)    | Duplicate user review entries              | Low      | Remove duplicate records using drop_duplicates in pandas or SQL DISTINCT queries. Use datetime parsing libraries (e.g., pandas.to_datetime) to standardize all date/time fields. |
| Dataset<br>(Timestamp field) | Inconsistent or incorrect timestamp format | Moderate |  |

|                                  |   |          |  |
|----------------------------------|---|----------|--|
| Dataset<br>(Restaurant metadata) | Inconsistent formats (e.g., location written in different ways like "NY", "New York") | Moderate | Apply data standardization techniques, using string functions or regex patterns to unify the format.               |
| Dataset<br>(User preferences)    | Sparse data or insufficient user history  | High     | Implement fallback strategies such as popularity-based or content-based recommendations when user data is lacking. |