

**Data Collection and Preprocessing Phase**

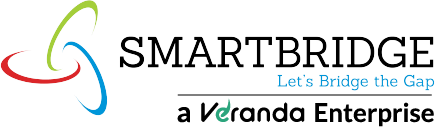


|  |  |
| --- | --- |
| Date | 05 july 2025 |
| Name | Sanika Tanaji Patil |
| Project Title | Restaurant Recommendation System |
| Maximum Marks | 2 Marks |

**Data Quality Report Template**

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Data Quality Issue** | **Severity** | **Resolution Plan** |
| Dataset (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 (User  reviews)      Dataset (Timesta mp field) | Duplicate user review entries            Inconsistent or incorrect timestamp format | Low        Moderate | Remove duplicate records using drop\_duplicates() in pandas or SQL DISTINCT queries. Use datetime parsing libraries  (e.g.,  pandas.to\_datet ime) to standardize all date/time fields. |





|  |  |  |  |
| --- | --- | --- | --- |
| Dataset (Restaura nt  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 (User preferenc es) | Sparse data or insufficient user history | High | Implement fallback strategies such as popularity-based or content-based recommendations when user data is lacking. |