**Data Quality Report**

2.1 Introduction:

This report outlines the data quality assessment and cleaning process for the restaurant recommendation system project. High-quality data is crucial for the success of any machine learning model.

2.2 Data Sources Reviewed:

[List all data sources used: e.g., Yelp API, Google Places API, internal user database, social media data]

2.3 Data Quality Issues Identified:

Missing Values:

Identified missing values in various fields (e.g., user ratings, restaurant addresses, cuisine types).

Implemented strategies to handle missing data (e.g., imputation, removal of records).

Inconsistent Data:

Observed inconsistencies in data formats, units, and terminology across different sources.

Standardized data formats and resolved inconsistencies through data cleaning and transformation.

Data Duplication:

Detected duplicate records in user data and restaurant data.

Implemented deduplication techniques to remove duplicates and improve data quality.

Outliers and Anomalies:

Identified outliers (e.g., extremely high or low ratings, unrealistic prices) and anomalies in the data.

Investigated and addressed outliers appropriately (e.g., removal, correction, or further investigation).

Data Bias:

Assessed potential biases in the data, such as representational bias (e.g., over-representation of certain demographics) and selection bias.

Implemented strategies to mitigate potential biases (e.g., oversampling underrepresented groups, adjusting weights).

2.4 Data Cleaning and Preprocessing Techniques:

Data Cleaning:

Handling missing values (imputation, removal)

Removing duplicates

Correcting inconsistencies

Addressing outliers and anomalies

Data Transformation:

Data type conversion (e.g., string to numerical)

Feature scaling and normalization

One-hot encoding for categorical variables

Text preprocessing (e.g., cleaning, stemming, tokenization) for textual data (reviews, descriptions)

2.5 Data Quality Assessment Metrics:

Data Completeness: Measured the percentage of non-missing values for each attribute.

Data Consistency: Checked for inconsistencies and discrepancies across different data sources.

Data Accuracy: Evaluated the accuracy of data through cross-validation and comparison with external sources.

2.6 Conclusion:

Through rigorous data cleaning and preprocessing, we have addressed identified data quality issues and improved the overall quality and reliability of the data for subsequent analysis and model building.

.