**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 16 March 2024 |
| Team ID | 17 |
| Project Title | VoyageVista: Illuminating Insights from Uber Expeditionary Analytics Using Tableau |
| Maximum Marks | 3 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** |
| My Uber Drives - 2016 | **Duplicate entries** in the dataset | Moderate | Identify duplicates based on all columns and remove them to ensure accurate data representation. |
| My Uber Drives - 2016 | **Missing values** in END\_DATE\*, CATEGORY\*, START\*, and STOP\* columns (1 row). | Low | Remove the row with missing values as it has minimal impact on overall data integrity. |
| My Uber Drives - 2016 | Significant missing values in the PURPOSE\* column (503 rows). | High | Replace missing values with "Unknown" to preserve the rows while maintaining clarity for analysis. |
| My Uber Drives - 2016 | **Incorrect data types** for START\_DATE\* and END\_DATE\*. | Moderate | Convert START\_DATE\* and END\_DATE\* columns to datetime format for accurate time-based analysis. |
| My Uber Drives - 2016 | **Unnecessary columns** or poorly formatted data. | Low | Combine START\* and STOP\* into a "Route" column for better clarity. Split date and time if needed. |
| My Uber Drives - 2016 | **Outliers** in the MILES\* column (e.g., unusually high or zero values). | Moderate | Use statistical techniques to detect outliers. Validate extreme values and handle them appropriately. |