**Data Collection and Preprocessing Phase**

| Date | 8 oct 2025 |
| --- | --- |
| Team ID | SWUID20250216152 |
| Project Title | Predicting plant growth with environmental and management factors |
| Maximum Marks | 10 Marks |

**Data Exploration and Preprocessing Template**

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

| **Section** | **Description** |
| --- | --- |
| Data Overview | The dataset is used to The dataset contains **193 records and 7 attributes** related to environmental and management factors affecting plant growth.  **Attributes include:**  · Soil\_Type  · Sunlight\_Hours  · Water\_Frequency  · Fertilizer\_Type  · Temperature  · Humidity  · Growth\_Milestone  analyze and predict how environmental and management parameters influence plant growth stages. |
| Data Cleaning | · Verified **no missing values** across all columns.  · Checked for **duplicates**; none were found.  · Ensured **consistent formatting** (e.g., converted categorical values like *Loam*, *Sandy*, *Clay* to uniform case).  Validated that all numeric columns (Temperature, Humidity,· Sunlight\_Hours) contain valid ranges.  · Outlier analysis using the **IQR method** confirmed no extreme outliers. |
| Data Transformation | Performed using **Power Query in Power BI**:  · Filtered irrelevant records (none found).  · Sorted data by environmental parameters for easier analysis.  · Created calculated columns such as:  o **Average Growth Rate**  o **Temperature–Humidity Index**  Applied transformations for clear visualization of growth milestones across multiple conditions. |
| Data Type Conversion | · Converted numeric fields (Temperature, Humidity, Sunlight\_Hours) to **Decimal Number** type.  · Converted categorical fields (Soil\_Type, Fertilizer\_Type, Water\_Frequency) to **Text** type.  · Ensured Growth\_Milestone column is stored as **Whole Number** for classification. |
| Column Splitting and Merging | · No column splitting was necessary as all data attributes were already well structured.  Combined certain management-related attributes (Water\_Frequency + Fertilizer\_Type) in calculated columns to create derived insights during Power BI modeling. |
| Data Modeling | · Defined relationships among fields for proper analysis hierarchy.  · Developed a **single fact table model** integrating all 7 attributes.  · Created **DAX measures** for analysis:  · Average Temperature  · Average Humidity  · Count of Growth Milestones by Soil Type  · Growth Rate Comparison across Fertilizer Types. |
| Save Processed Data | The processed and cleaned dataset was saved as **“plant\_growth\_data\_cleaned.csv”** for future visualization and modeling within Power BI.  This ensures data integrity and allows efficient refresh and reuse across multiple dashboards. |