Data Collection and Preprocessing Phase

Date	30 June 2025
SkillWallet ID	SWUID20250170810
Project Title	Global Food Production Trends and Analysis: A Comprehensive Study from 1961 to 2023 Using Power BI
Maximum Marks	10 Marks

Data Exploration and Preprocessing Template

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description
Data Overview	The dataset titled "World Food Production"(sourced from Kaggle) contains over 170,000 rows of data covering global food crop production from 1961 to 2023. It includes attributes such as crop type, production quantity (in tonnes), region, country/entity, and year. This dataset is suitable for analyzing temporal and regional trends in the production of major crops and fruits.
Data Cleaning	- Handled missing values by removing rows with nulls in critical fields (e.g., Year, Item, Entity, Value). -Removed duplicate entries using Power Query. -Standardized crop and region names to ensure uniformity across filters and visuals.

Data Transformation	- Used Power Query Editor to filter data by selected crop types (e.g., wheat, maize, rice, apples, bananas, etc.). - Applied sorting to analyze trends over time. - Created pivot tables to summarize production by crop and year. - Added calculated columns for total production across selected years or regions.
Data Type Conversion	- Converted columns like Year to `Whole Number`, Value (production in tonnes) to `Decimal Number`, and Entity/Item to `Text`. - Ensured all types are consistent to avoid errors in measures and visuals.
Column Splitting and Merging	- Split “Item” column where necessary (e.g., separating fruit types from crop category if grouped). - Merged region and sub-region columns for simplified filtering (e.g., "Africa - East" into "Africa").
Data Modeling	- Defined relationships between tables if the dataset was split (e.g., crop info table and production table). - Created measures for total production, average yearly production, and growth rates using DAX. - Ensured star schema for optimal Power BI performance.
Save Processed Data	- Saved the cleaned and transformed dataset as a `.pbix` file in Power BI. - Exported intermediate cleaned data as `.csv` for backup and potential reuse. - Version-controlled processed data using Git for reproducibility.