1. Question:

1.1 Main Question:

How do climate conditions influence agricultural productivity in North and South America?

2. Data Sources:

Source 1: Climate Change: Earth Surface Temperature Data

- **Description:** Global temperature trends by city - this dataset. It has average temperatures across cities, and it's relevant, because temperature gets directly to agricultural productivity.
- **Structure and Quality**: it contains date (dt), average temperature (Average Temperature) and city, country (City, Country) columns We have checked and cleaned the data (i.e., null values removal), transforming dates into formats that can be analyzed.
- **License**: Public domain data, typically available under open-data licenses which require proper attribution.
- **Data URL:** https://query.data.world/s/x5sksfhbjl3h2xfswrbolreeaguqrg?dws=00000

- **Meta Data URL:** https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data/data

Source 2: Agricultural Productivity (FAO)

- **Description**: This data Provides info about agricultural productivity, crop yields, and types of crops grown in North and South America on yearly basis.
- **Structure and Quality**: It has a structure with columns, i.e., Year, Area, Item, Value, and Unit. Data was cleaned and only South and North America data were retained, some void values.
- **License**: data provided by FAO is Public domain data, typically available under open-data licenses which require proper attribution.
- **Data URL:** https://bulks-faostat.fao.org/production/Production_Crops_Livestock_E_Americas.zip

- **Meta Data Link:** https://www.fao.org/faostat/en/#data/QCL/metadata

3. Data Pipeline

3.1 Technologies Used:

- **Data Manipulation**: used Pandas library for data loading.
- **Storage**: data was stored using Pandas Data Frames.

- **Processing**: for processing Pandas and NumPy for data transformation and cleaning.

3.2 Process Steps:

- **Data Downloading:** Both datasets (climate change and agricultural production) are being downloaded via HTTP requests.
- **Data Loading**: CSV files are being Loaded into Pandas Data Frames.

3.3 Data Cleaning and Transformation:

1. **Climate change data** : Climate data is processed to only contain North and South America, and columns are cleaned and formatted accordingly.
2. **Agricultural Productivity data**: This data is similarly Processed for the North and South America and cleaned by removing missing values
3. **Drop Missing Values**: Fill or droped missing values based on comparison needs.
4. **Normalized data**: Standardize date formats and numerical values.

- **Data Storage**: Transformed datasets are merged and stored in an SQLite database (project_data.db).

3.4 Transformation and Cleaning Of Data:

Climate Change Data:

- Removed unneeded columns.
- Formatted the date and temperature columns, handled missing values.
- Filtered for the relevant countries in South and North America.

Agricultural Productivity Data:

- Processed for the relevant countries(South and North America) and dropped rows with missing values.

3.5 Challenges:

- **Data Source Availability**: Processing incomplete data from both datasets was an issue. The climate change data required Processing by region, and the agricultural data contains some null values that had to be cleaned.
- **Data Integration**: Merging datasets for comparison was a challenge due to mismatched timeframes and missing entries, so keeping in mind the data structures was essential**.**

4. Quality Measures:

- **Error Logging**: The pipeline includes error handling (e.g., ensuring files are downloaded before processing etc).
- **Data Validation**: Implemented validation checks at each stage of the pipeline.

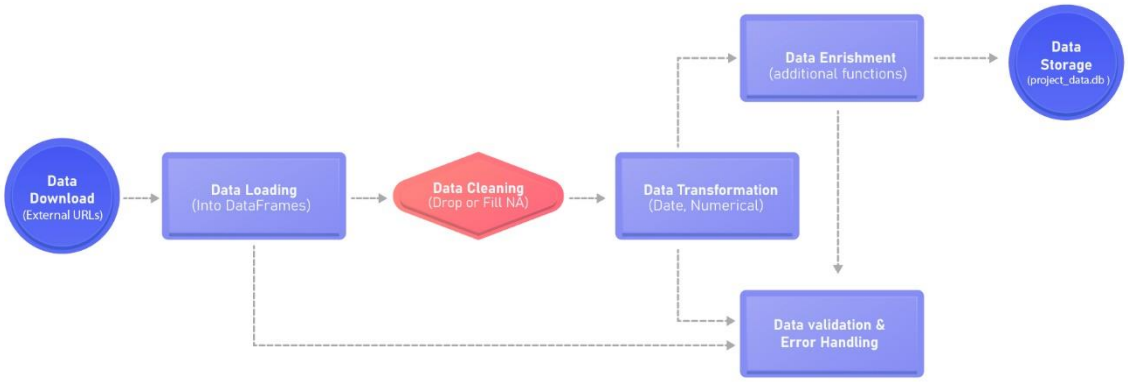
5. Result and Limitations

5.1 Output Data Structure::

- **Climate Data**: Cleaned and processed climate Change data for North and South America, with columns for City, Country, AverageTemperature, and dt.
- **Agricultural Data**: Processed agricultural Productivity data, focusing on crop yields for South and North American countries, with columns Year, Area, Item, Value, and Unit.
- **Data Quality**: Both datasets were Processed and transformed by removing missing values and sorted by needed regions. The final data output is in a structured, time-series format that can be used for further comparison.
- **Format:** The Merged data is stored in an SQLite database for easy querying and further analysis. Why SQLite? because of its simplicity and efficiency for small to medium dataset processing.

6. Figures and Tables

6.1 Figure 1: Data Pipeline Structure



6.2 Figure 2: Comparison