1. Introduction

In a rapidly evolving world, understanding the complex interactions between climate change and economic factors is essential. My data science project aims to explore how temperature changes impact agricultural productivity and the broader economy in European countries.The project will analyze historical data on temperature variations and agricultural output across different European countries. By examining these trends, we aim to Assess how temperature change is affecting agricultural productivity and the economy in european countries.

2. Data Sources

Data-Source 1: Kaggle dataset "Global Food and Agriculture Statistics"

I've chosen the Kaggle dataset "Global Food and Agriculture Statistics" from the United Nations.

URL:https://www.kaggle.com/datasets/unitednations/global-food-agriculture-statistics?select=fao_data_production_indices_data.csv

Reason for Choice: This dataset provides comprehensive statistics on global food production indices, crucial for understanding agricultural trends in various countries, including Europe.The dataset includes metrics related to food production, such as country or area, year, and production value.

Data Structure and Quality: The data is in CSV format, well-structured with clear columns. However, it contains missing values that need to be handled.

License: This dataset is publicly available on Kaggle under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. We adhere to this license by crediting the source and sharing any derived data under the same terms.

However, one challenging aspect was that the download dataset from kaggle as it required authentication.So I used kaggle api and environment method to access it through my code.

Data-Source 2: FAOSTAT

For temperature data, FAOSTAT was chosen as the source. FAOSTAT is renowned for its vast collection of climate-related datasets, making it a suitable choice for a project focusing on temperature changes.

Data URL 2: https://fenixservices.fao.org/faostat/static/bulkdownloads/Environment_Temperature_change_E_All_Data.zip

Reason for Choice: Understanding temperature changes is vital for analyzing the impact of climate on food production.The dataset includes temperature change data by country and year.

Data Structure and Quality: The data is in a ZIP file containing CSV files. The structure includes multiple columns, some of which are redundant and need cleaning

License: This dataset is publicly accessible, and we comply with its usage terms by providing proper attribution and using the data for non-commercial purposes.

3. Data Pipeline

The data pipeline is designed to automate the download, extraction, cleaning, and storage of data from multiple sources into a unified SQLite database.We used python programming and different libraries like pandas, requests, zipfile, sqlalchemy, kaggle for our data pipline.To download and process both dataset we used two different function for both datasets.After processing and cleaning data we stored tempearture and produciton data in a sqlite database with two different tables.

For the FAOSTAT data, I made the following transformations:

- Restriction to yearly data to align with the food production data set.
- Converting year data from columns into rows facilitates compatibility with other datasets
- Transformation of year values from strings into integers to facilitate numerical analysis and comparisons.
- Restriction of the data to the following columns: Country,Year and Value
- Selected only European countries from all countries

For Kaggle dataset "Global Food and Agriculture Statistics", I undertook the following steps:

- Elimination of rows with zero values.
- Restriction of the data to the following columns: Country,Year and Value
- Selected only European countries from all countries
- Kept the countries which is present only in food datasets

Problems and Solutions:

In FAOSTAT data years was in column.So convert to to row and stored it in a country column.There were also some missing value and data integrity problem in both datasets.I solved it by removing rows with missing value and kept only countries which is present in both dataset to keep the data integrity.

Error Handling:

The pipeline includes checks for successful data download and extraction. If an error occurs (e.g., network issues, file not found), the pipeline logs the error and halts further processing.

4. Results and Limitations

- The output data is stored in an SQLite database with tables for processed food and temperature data.The resulting data is clean, with standardized column names and no missing values. This ensures high data quality for subsequent analysis.
- We used sqlite format for its simplicity and ease of integration with various data analysis tools.
- Critical Reflection

  Potential Issues: There might be inconsistencies or missing data in future datasets, which the pipeline should handle gracefully. Additionally, differences in country naming conventions across datasets could pose challenges for data integration.

Limitations: The current pipeline processes static datasets. For real-time or frequently updated data, additional mechanisms for periodic updates and checks would be necessary.

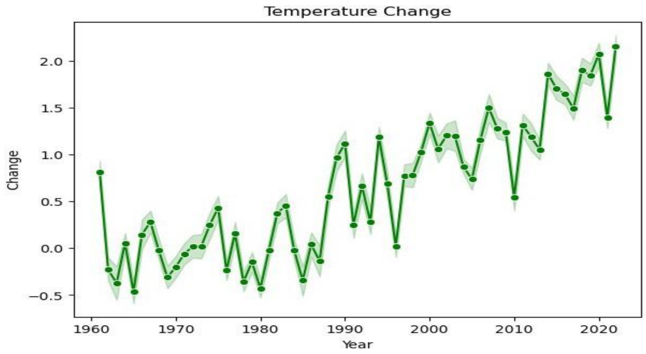
In below code snippets I showed how our data looked after processing:

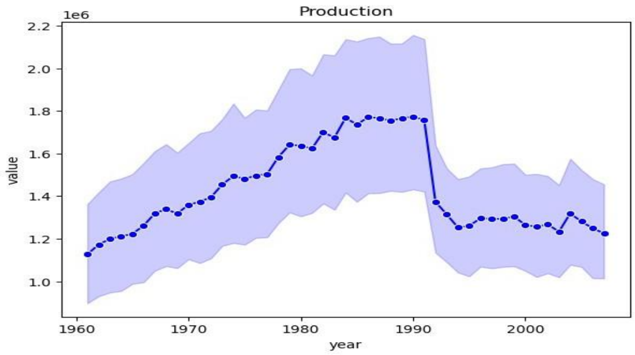
```
import pandas as pd
temp_df = pd.read_sql_table('temperature_data', 'sqlite:///C:/Users/nhbso/dataset.sqlite')
print(temp_df)
```

```
       Area         Year    Change
0      Albania      1961    0.186
1      Albania      1961    -0.611
2      Albania      1961    1.000
3      Albania      1961    2.419
4      Albania      1961    -0.515
...    ...          ...     ...
26345  Switzerland  2022    2.321
26346  Switzerland  2022    2.341
26347  Switzerland  2022    3.995
26348  Switzerland  2022    2.775
26349  Switzerland  2022    2.858
```

```
import pandas as pd
# Get data from temperature table
temp_df = pd.read_sql_table('food_data', 'sqlite:///C:/Users/nhbso/dataset.sqlite')
# Print all rows of the table
print(temp_df)
```

```
                Area      year     value
0               Albania   2007.0   824818.0
1               Albania   2006.0   858366.0
2               Albania   2005.0   813707.0
3               Albania   2004.0   819870.0
4               Albania   2003.0   789269.0
...             ...       ...      ...
42623  United   Kingdom   1965.0   65.0
42624  United   Kingdom   1964.0   64.0
42625  United   Kingdom   1963.0   65.0
42626  United   Kingdom   1962.0   67.0
42627  United   Kingdom   1961.0   68.0
```





Conclusion:

We created an automated data pipeline to process and integrate global food and temperature data. The pipeline ensures high data quality and prepares the data for subsequent analysis to assess the impact of temperature change on agricultural productivity in European countries. Future work may involve enhancing the pipeline to handle real-time data updates and addressing any inconsistencies in country naming conventions.