Analysis and Visualization of Agricultural Data based on the impact of Climate Change through ETL process.

Project Description: Climate change implies significant challenges to agricultural productivity worldwide, impacting crop yields. Nevertheless, understanding the complex relation between climate and agricultural data remains a barrier, which needs an efficient data analysis and visualization framework. For this purpose, the objective of this project is to build an ETL pipeline to assess or measure the impact of climate change on agricultural data and provide necessary valuable insights. This framework will empower policymakers, and researchers on how to handle the difficulties arising from climate change in agriculture by providing them with well-informed decision-making capabilities

Project Questions

- What impact has climate change based on temperature had on crop production in countries around the world in the last years or decades? Have there been any trends or recurring patterns, in crop yields?
- Are there any considerable shifts in the geographical crops distribution as a result of changes in climatic conditions?

Data Sources

DataSource-1: Crop Production

Metadata URL: Link

Data URL:
Production_Crops_E_Africa.csv
Production_Crops_E_Americas.csv
Production_Crops_E_Asia.csv
Production_Crops_E_Europe.csv
Production_Crops_E_Oceania.csv

Data Type: CSV
Licence: CC BY-NC-SA

Source:
This dataset is taken from Food and Agriculture Organization of the United Nations (FAO).

DataSource-2: All Countries Temperature Statistics

Metadata URL: Link

Data URL: Temperature data globally

Data Type: CSV
Licence: CC0 1.0

Acknowledgment:
This data is provided by FAOSTAT and is based on publicly available GISTEMP data from the NASA GISS.

Source/Citation: International Monetary Fund. 2022. Climate Change Indicators Dashboard. Annual Surface Temperature Change. Accessed on [2024-05-23].

Note: This all work is purely non-commercial and is used for only semester project at FAU to implement the ETL pipeline and provide valuable insights.

DataSource-1 Description:

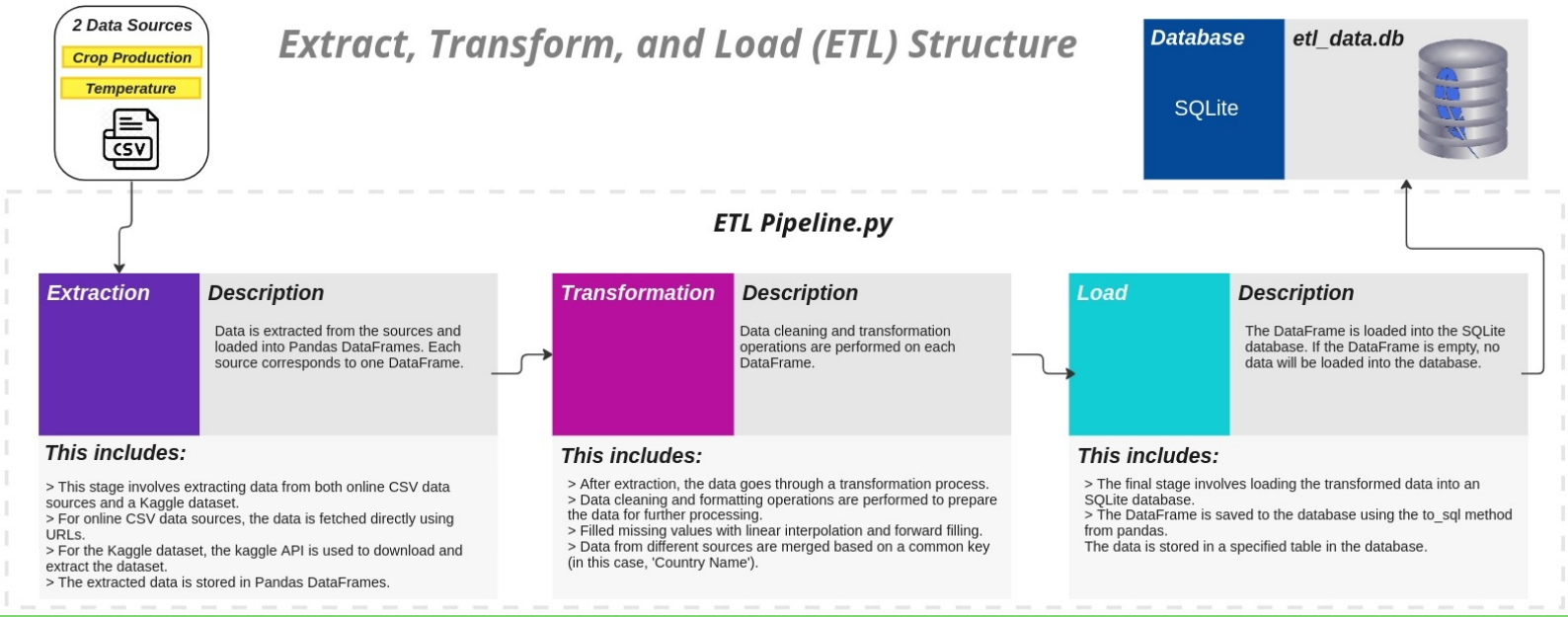
This FAO dataset provides an extensive overview of worldwide crop production statistics from 1961 to 2019. It includes 173 different products such as cereals, vegetables, fruits, tree nuts, fiber crops, oil crops, pulses, roots, and tubers. The dataset includes information on harvested areas, production quantities, and yields, offering a detailed picture of global primary crop production. This data is essential for examining agricultural productivity, food security, and related economic issues.

DataSource-2 Description:

This dataset provide details about mean surface temperature changes from 1970 to 2021 across countries around the world. It spans 51 years and compiles data from weather stations, satellites, and ocean buoys. The dataset enables analysis of temperature trends and identification of regions vulnerable to temperature changes, aiding in understanding climate change impacts and informing mitigation policies. Temperature is measured in degrees Celsius, with a +ve index indicating an increase and a -ve index indicating a decrease.

ETL Pipeline

- The ETL pipeline is built with Python3 to extract publicly available datasets including crop yield data from datasource-1 and historical temperature records from datasource-2.
- The extracted data undergoes transformations to standardize the formats, integrate relevant variables, and fill missing values using linear interpolation and forward filling.
- The "Country Name" is used as the key to merge dataframes, retaining only entries with matching country names, crops of interest, and temperature readings year-wise, focusing on the years 1970 to 2019.
- Logging is additionally configured in the ETL pipeline to capture the process flow and record errors, tracking successful steps and issues; try-except blocks handle exceptions, stopping the process, and logging critical errors as needed.
- Also ensured a clean environment by checking for and removing the Kaggle dataset file after reading.
- The loaded data will be analyzed to identify trends, correlations, and irregularities caused by climate change effects on crop yield variability.



Results and Limitations

Output Data:

The pipeline produces a merged dataset combining crop production data with mean surface temperature statistics, providing insights into the relationship between these variables across different countries.

Output Data Format:

Chose SQLite database format for flexibility, portability, integration with pandas, and scalability for medium-sized datasets.

Potential Issues:

- Data completeness issues with potential residual missing values.
- Data consistency challenges due to possible variations in source data.
- Scalability concerns with SQLite for very large datasets.
- Dependency on the reliability of source data.
- Need for adjustments if input data structures or schemas change over time.

Table 1: Showing few rows of final transformed dataframe.

	Country Name	Item	Element	Unit_x	Y1970	...	Y2019	Unit_y	Change	1970	...	2019
0	Algeria	Wheat	Area harvested	ha	6700	...	43043	Degree Celsius	Surface Temperature Change	0.114	...	1.094
1	Algeria	Maize	Yield	hg/ha	4478	...	13292	Degree Celsius	Surface Temperature Change	0.114	...	1.094
2	Algeria	Rice, paddy	Production	tonnes	3000	...	57213	Degree Celsius	Surface Temperature Change	0.114	...	1.094
3	Pakistan	Dates	Area harvested	ha	1820	...	19372	Degree Celsius	Surface Temperature Change	0.583	...	0.648
4	Pakistan	Sugar cane	Yield	hg/ha	104945	...	54069	Degree Celsius	Surface Temperature Change	0.583	...	0.648