Climate Change and its burn on Pocket - Cost of Living Crisis

04 June, 2024

This project aims at finding trends that can help us mitigate the risks posed by climate change on availability and affordability of the basic needs for each and every person on the planet.

1. Questions addressed in the report

1.1. What are the trends in the production of staple food items (Wheat, Rice, Potatoes, Maize etc.) across the globe? How the yield (production per hectare) has changed over the years (1970-2019)?

1.2. Trends and correlation between production yields and the consumer price inflation (CPI) from year 1970-2019?

1.3. How temperatures have flared up from 1970 – 2019? Has rising global temperatures impacted the global crop production yield?

2. Data Sources

2.1. FAOSTAT – Food and Agriculture Organization of the United Nations

- Data Format: CSV
- URL: Africa, America, Asia, Europe, Oceania
- License: CC BY-NC-SA - Access, downloading, creating copies and re-disseminating datasets are subject to the following terms: "Unless specifically stated otherwise, all datasets disseminated through the databases are licensed under the Creative Commons Attribution-Non-Commercial-Share"

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, fibre 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.

2.2. Kaggle

- Data Format: CSV
- URL:https://www.kaggle.com/datasets/mdazizulkabirlovlu/all-countries-temperature-statistics-1970-2021/data?select=all+countries+global+temperature.csv
- License: **CC BY-NC-SA** – Access, downloading, creating copies and re-disseminating datasets are subject to the following terms: "Unless specifically stated otherwise, all datasets disseminated through the databases are licensed under the Creative Commons Attribution-Non-Commercial-Share"

This data is provided by the Food and Agriculture Organization Corporate Statistical Database (FAOSTAT) and is based on publicly available GISTEMP [1] data from the National Aeronautics and Space Administration Goddard Institute for Space Studies (NASA GISS).

This dataset provides information on changes in global surface temperature across all countries from 1970 to 2021. The dataset allows for the analysis of temperature trends in different countries and regions, as well as the identification of areas that are particularly vulnerable to temperature shifts. In context of this project, this information can be used to better understand the impacts of climate change on crop yields and inflation in food prices. The temperature measure by the unit of degree Celsius. The data is complete, consistent and relevant. The accuracy of the data is promised by FAOSTAT.

2.3.World Bank Open Data - Free and open access to global development data

- Data Format: CSV

[1] GISS Surface Temperature Analysis: https://data.giss.nasa.gov/gistemp/sources_v4/gistemp.html

[2] CPI – Consumer Price Inflation

- URL: https://api.worldbank.org/v2/en/indicator/FP.CPI.TOTL?downloadformat=csv
- License: **CC-BY 4.0** - The Creative Commons Attribution 4.0 International license allows users to copy, modify and distribute data in any format for any purpose, including commercial use. Users are only obligated to give appropriate credit (attribution) and indicate if they have made any changes, including translations. CC-BY 4.0, with the additional terms below, is the default license for all Datasets produced by the World Bank itself and distributed as open data. More info can be found here.

The World Bank as my data source due to their extensive data on global economic indicators. Their reputation for reliable and comprehensive data makes them an ideal source for my project. This database provides monthly food price inflation estimates for the countries around the globe. The relation between the food price inflation of countries and its dependence on production yields and temperature can be analysed using the data.

The data in its original state was incomplete, which led to cleaning of data by removal of some the countries from the data. I have interpolated the missing values linearly for some data. And since UAE is an important country to analyse in the global food production supply chain I have imputed CPI values of UAE by using domain knowledge using the following formula:

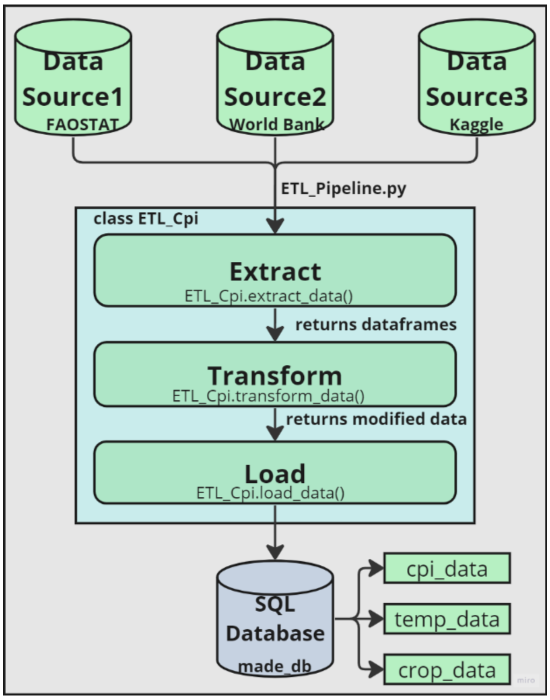
Imputed Value = Mean (all available countries CPI for that year)*Imputed Value Constant_UAE

This is "Imputed Value Constant_UAE" is calculated by taking Inflation of UAE as a percentage of Sum of all the inflation values for that given year, for all the years which have data present for UAE. Then taking a mean of percentage over all the years which have data present for UAE. This turns out to be a constant - 0.9581559454843503

3. Data Pipeline

For extraction of kaggle data, store API credentials of your kaggle account in ./.kaggle folder on your system.

3.1. Architecture of the Data Pipeline

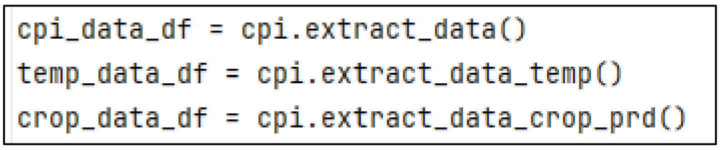


The architecture consists of three separate data sources from FAOSTAT [2], World Bank Open Data and Kaggle. The pipeline can be instantiated by running the external endpoint provided to the user in form of a shell script - **./project/pipeline.sh.** On starting the pipeline, the python script ***ETL_Pipeline.py*** is started which in turn instantiates the ***class ETL_Cpi***.

```
cpi = ETL_Cpi()
```

3.2. Extraction of Data

The extraction of data posed a challenge as the direct link for downloading the CSV data is not available for the World Bank open data. The World Bank CPI [3] data is downloaded in form of a ZIP file and the required CSV data is extracted and stored in data frame. Extraction of data from Kaggle needs the API credential of the user placed in **./.kaggle** directory in the root user. The FAOSTAT data can be easily downloaded from the above mentioned hyperlinks. The method returns list of data frames with crop production data for five regions of world. Since, the data extraction process differs for all the three sources, hence created three separate extraction methods, namely:



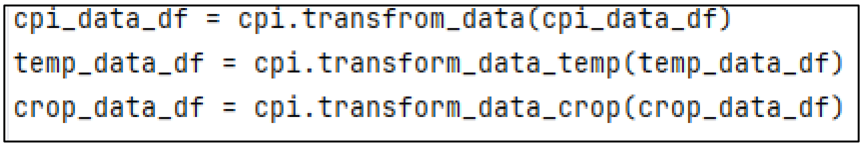
```
cpi_data_df = cpi.extract_data()
temp_data_df = cpi.extract_data_temp()
crop_data_df = cpi.extract_data_crop_prd()
```

[1] GISS Surface Temperature Analysis: https://data.giss.nasa.gov/gistemp/sources_v4/gistemp.html

[2] CPI – Consumer Price Inflation

3.3. Transformation of Data

Data transformation step aimed at making data more complete, consistent and relevant to the problem at hand. Data cleaning is performed using following three methods in the class:



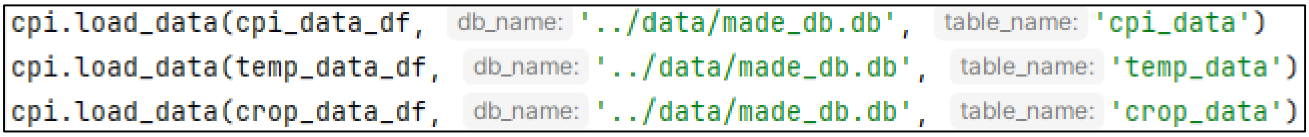
```
cpi_data_df = cpi.transfrom_data(cpi_data_df)
temp_data_df = cpi.transform_data_temp(temp_data_df)
crop_data_df = cpi.transform_data_crop(crop_data_df)
```

- ***cpi.transform_data()***: Drops the columns which were irrelevant to the problem like 'Country Code'. Removes the rows which has all the values as null and then performs linear interpolation on the data frame. Imputes the values of UAE using the domain knowledge using the above mentioned formula due to its geopolitical significance. Return modified data frame.
- ***cpi.transform_crop_data():*** Drops irrelevant columns and iterates through the list of data frames provided by *cpi.extract_crop_prd().* Changed the data types of some columns to strings as per the requirement. Interpolated the data using the 'ffill' method because there were very few NaN values so imputation does not make significant effect on the data. Return modified data frame.
- ***cpi.transform_temp_data():*** Changes the required columns to strings and drops irrelevant columns. Performs linear interpolation over each year and returns the modified data.

The primary challenge at this stage of the pipeline was determining a strategy for managing incomplete data. Correctly handling missing day values by substituting the them based on data distributions was key learning in this step. Also, understanding data quality on the basis of completeness and relevance had a great impact on overall data pipeline.

3.4. Load Data to SQL database tables

Create a connection to the SQL database and save the three data frames in three separate tables named as '*cpi_data*', *temp_data*' and '*crop_data*'. All three tables contain column 'Country Name' as the primary key (PK) and we will try to explore the data further using this data in future to reach an answer to questions (Section 1).



```
cpi.load_data(cpi_data_df,  db_name: '../data/made_db.db',  table_name: 'cpi_data')
cpi.load_data(temp_data_df,  db_name: '../data/made_db.db',  table_name: 'temp_data')
cpi.load_data(crop_data_df,  db_name: '../data/made_db.db',  table_name: 'crop_data')
```

3.5. Error Handling in Pipeline

The pipeline has been secured from run time errors by using try-catch blocks wherever possible and log the errors in the debugger so as to help the user navigate in the case when a run time error occurs.

4. Results, Reflection and Limitations of Pipeline

The result of data pipeline a SQL database with three tables. Each table corresponds to one data source. All the tables have a significant impact on reaching to final answers to the questions posed in section1. The finally saved data is consistent, complete and relevant to the problem. The quality of final data assessed using descriptive statistics.

The SQL table format of the final data is chosen as it is easy to manipulate and derive insights using from. Also, I will be able to use SQL to query the data to get better insights by joining data tables.

4.1. Critical Reflection and Issues

The initial CPI dataset from World Bank was incomplete and imputation in the data make it prone to bias which might change the result. Also, the 'Cost of Living Crisis' does not depend only on CPI and temperature. There are larger set of factors which determine the 'Cost of Living Crisis', which are not considered in this report.

[1] GISS Surface Temperature Analysis: https://data.giss.nasa.gov/gistemp/sources_v4/gistemp.html

[2] CPI – Consumer Price Inflation