

# PROJECT REPORT

**Analyzing impact of Climate change on Crop production and CO2 emission in the USA over the period 1990 to 2015.**

## INTRODUCTION :

Climate change is a major problem because it affects our environment, weather patterns, ecosystems, and human society, leading to severe consequences such as rising temperatures, extreme weather events, and sea-level rise, Co2 Emission. I chose these datasets as this will help to analyze the history data of Co2 emission and crop production from 1990 to 2015. The results will give insights into how climate change impacted crop production and co2 emission over the period of time.

## DATASOURCES :

### 1. Crop Production & Climate Change [Link](#)

**Description :** This dataset from kaggle provides data on crop yields, harvested areas, and production quantities for wheat, maize, rice, and soybeans. Crop yields are the harvested production per unit of harvested area for crop products. Crop production is measured in tonnes per hectare. This dataset includes information on crop production from 2010-2016

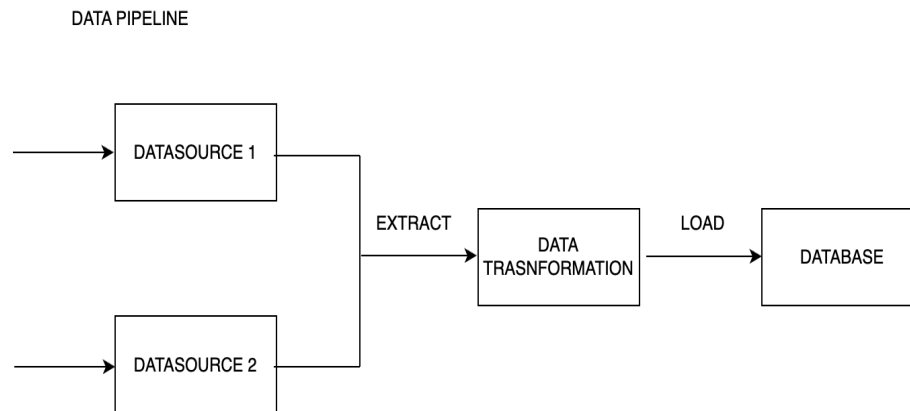
### 2. Data source 2 : CO2 Emissions [Link](#)

**Description :** The CO2 emissions dataset from kaggle provides a comprehensive overview of the amount of CO2 emitted by each country. The dataset includes information on CO2 emissions by country from 1960 to 2023. It covers all countries in the world.

## LICENSES :

Both the datasets are free to share, adapt, transform. I will follow their obligations by giving credits to the author by mentioning the link of the datasets for license and more details.

## DATA PIPELINE :



In this project, a pipeline is created using the Python programming language. The pipeline follows the processes listed below.

### 1. EXTRACTION :

This step includes extracting the data from the source. We have a CSV dataset from Kaggle. The python code will download the datasets into a temporary directory using api authentication, and it will then load them into a pandas dataframe.

### 2. TRANSFORMATION :

The transformation of data involves several critical processes to ensure it is clean, filtered, and consistent. In this project, data cleaning encompassed renaming dataset columns for clarity, and converting time and year fields to a consistent datetime format as a best practice. Data for the USA region from 1990 to 2015 was filtered. Using group by, I calculated the year's averages and combined the datasets into a single table based on the date.

### 3. LOAD :

This is the final step of the pipeline, which involves storing the data in a suitable format within the database.

### **PROBLEMS ENCOUNTERED :**

- 1) Initially, I saved Kaggle datasets to GitHub and accessed them via direct paths in my code. Later, I discovered API authentication and streamlined the process by loading data directly into a pandas DataFrame from the web.
- 2) After attending a lecture on good programming practices, I learned and implemented several best practices in my code.
- 3) Date format inconsistency.

### **RESULTS :**

- Output of the data contains years from 1990 to 2015 with average Co2 emission and Crop production for the USA.
- Output data is structured and can be easily analyzed and it is of good quality as it satisfies all conditions.
- I have stored the data in a sqlite database in a sqlite file as it eliminates the need of complex configuration.

### **LIMITATIONS:**

- The Crop production dataset is limited and does not include the most recent data.