# IS525: Database Warehousing and Business Intelligence
# Project: Final Report

Group 10:
Iishi Patel (iiship2)
Shreyasi Biswas (sb69)
Shruti Ravichandran (sr67)
Yashas Prashanth (yashasp2)

## Indicators and Impact of Climate Change

## Introduction

This project aims to analyze the impact of climate change on the world by combining emissions data, temperature data, and food production data. The project uses publicly available datasets - year-on-year emissions data by country from FAOSTAT, time–series average temperature anomaly data collected globally from Berkeley data, to establish the effect of climate change in the agriculture industry using food production data also obtained from FAOSTAT.

The data was extracted from these various sources, cleaned and wrangled using Python for standardization. The datasets underwent ETL (Extract, Transform, Load) using Talend. Once the final datasets were prepared, they were used to create interactive dashboards and reports on Tableau that showcase the findings in a user-friendly and easy-to-understand format. The visualizations include charts, maps, and graphs that allow users to explore the data and gain insights.

The audience for this data analysis and visualization are stakeholders capable of driving decisions that can potentially curb the impact of climate change, and stakeholders in the food and agriculture industry.

## Dataset

### FAOSTAT data: Food Production Data and Emissions Data

FAOSTAT is a comprehensive database maintained by the Food and Agriculture Organization of the United Nations (FAO). It contains statistical information on a wide range of topics related to food and agriculture, including crop production, livestock, fisheries, forestry, land use, and population. The data in FAOSTAT are obtained from various sources, including national statistical offices, international organizations, and research institutions. For the scope of our project, we focus on country-wise and item-wise annual food production data and global annual emissions data.

**Berkeley Earth data: Global Average Surface Temperature**

Berkeley Earth is a non-profit organization that works on the compilation and analysis of data related to the environment. The data collected by Berkeley Earth includes temperature data and air pollution data. The dataset used for the purpose of this analysis is the land-surface and ocean temperature data, with the ocean temperature data measured being the air temperature above sea ice for locations with sea ice. The temperature measures are in Celsius and denote the anomalies relative to the Jan 1951-Dec 1980 average temperature, and uncertainties are the 95% confidence interval.

## Data Pipeline Model

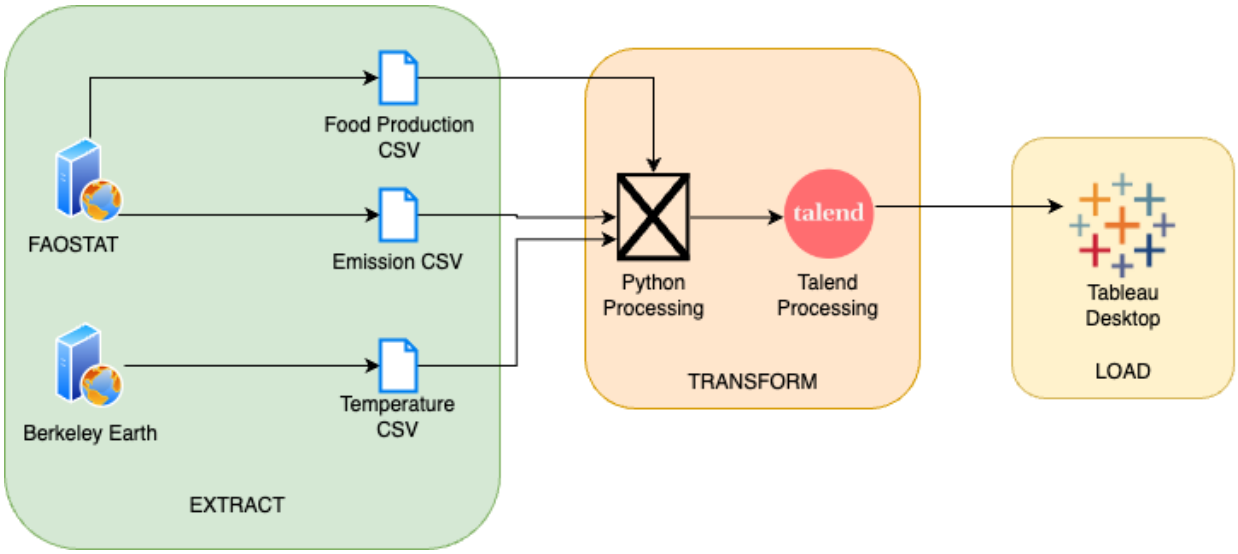


Fig 1. Data Pipeline Diagram

The data pipeline used extract, transform, and load (ETL) to create insights and visualizations:

- **Extract:** This step involves retrieving data from multiple sources. In this case, the data is being extracted from FAOSTAT and Berkeley Earth, which are databases that contain information about food production, emissions, and temperature.
- **Transform:** Once the data has been extracted, it needs to be cleaned and formatted to make it ready for analysis. The transformation process can involve several steps, including removing duplicates, merging datasets, and converting data types. In this case, pandas in Python is being used for data formatting, and Talend ETL tool is being used for combining datasets.
- **Load:** After the data has been transformed, it is loaded into a data visualization tool such as Tableau. This step involves using our transformed CSVs and creating the appropriate data visualizations to communicate insights.

## Python Processing

### Food Production Data

The food production dataset contains the columns - area code and area of the region, item code and food item names, and the quantity of food production for each year from 1961 to the present. Python is used for pre-processing, by reading the CSV dataset using Pandas and pivoting down the year-wise columns into rows, and creating a column for item-wise food production which is grouped by item and country-wise food production which is grouped by country. The code for this Python preprocessing is documented in data_cleaning_food_production.ipynb.

### Emissions Data

The emissions dataset contains the columns - area code and area of the region, item code and item contributing to the emissions, the emission type (CO2, N2O etc), and the quantity of the emission for each year from 1961 to present. Python is used for pre-processing, by reading the CSV dataset using Pandas and pivoting down the year-wise columns into rows, and creating a column for every emission type, by aggregating over year and area. A new Pandas dataframe is used to store the resulting data. The code for this Python preprocessing is documented in data_cleaning_emissions_and_temperature.ipynb.

### Temperature Data

The temperature dataset contains the columns - year, and annual anomaly, annual uncertainty, five-year anomaly, five-year uncertainty of the land and ocean temperature above and below sea ice. The data is present in the form of a text file. Consequently, Python is used to open and read the data from the text file and store the annual anomaly and annual uncertainty of the land temperature and ocean temperature above sea ice, in a Pandas data frame. The code for this Python preprocessing is documented in data_cleaning_emissions_and_temperature.ipynb.

## Talend ETL

In order to be able to visualize some sort of correlation between food production and emissions, the datasets were joined together and exported. Both Emissions and Food Production data are reported at a year and country level, which became the key for combining the two datasets. Both datasets were first loaded into Talend using the tInputDelimited component. Their data types were standardized. Tmap component was used to map all the columns from Food Production

data as well as Emissions data. Finally, the combined file joined on Year and Area was saved as a csv file and used to make visualizations in Tableau.

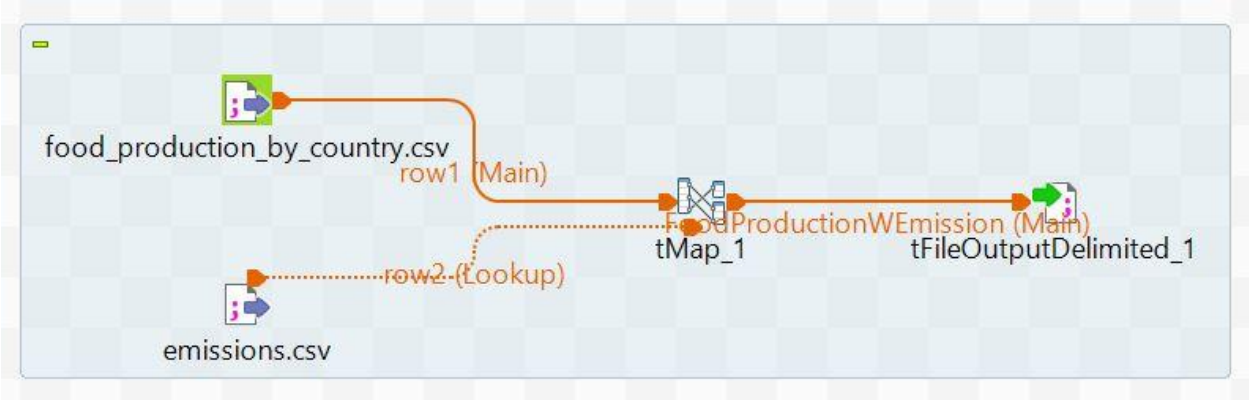


Fig 2. Talend Job

## Tableau Dashboard

Two tableau dashboards were created to understand if there is any visible impact of temperature or CO2, CH4 and NO2 emissions on food production. The first dashboard shows the temperature and emission trends from 1990 to 2020 across various countries in the world. The first view is a world map showing the top and bottom n countries by CO2, CH4 and NO2 emissions. Users can use the *Select number of countries* parameter field to mention the value of n, based on which n top and n bottom countries will automatically be updated. The second view shows the increase / decrease in emissions from 1990 to 2020, meant to show the jump in CO2, CH4 and NO2 emissions in 30 years. The third view is similar, except is shows the year on year trend for CO2, CH4 and NO2 emissions. Users have the flexibility to filter the second and third view for a specific country or for all countries. The fourth view shows the temperature anomaly trend over the years.

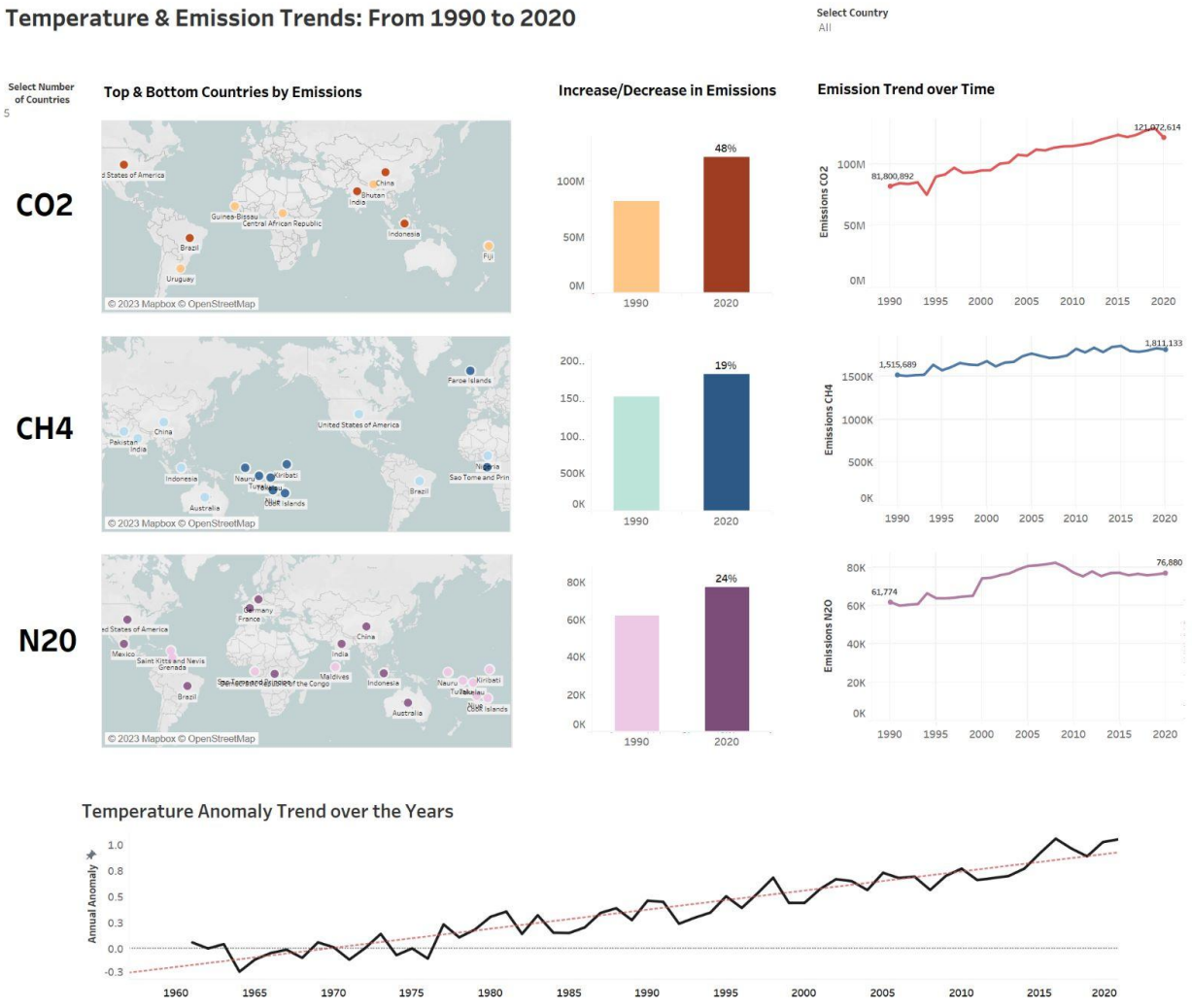


Fig 3. Dashboard 1

The second dashboard shows the food production trends, as well as the relationship between emissions and temperature data each with food production. There are two sets of views in this dashboard. The first set shows the top n items produced over the years globally, and the top n countries in terms of food production, where the value of n can be selected using a filter. The second set visualizes the correlation between the emissions data and food production data, and the temperature data and the food production data using scatterplots.

## Food Production Trends: From 1990 to 2020

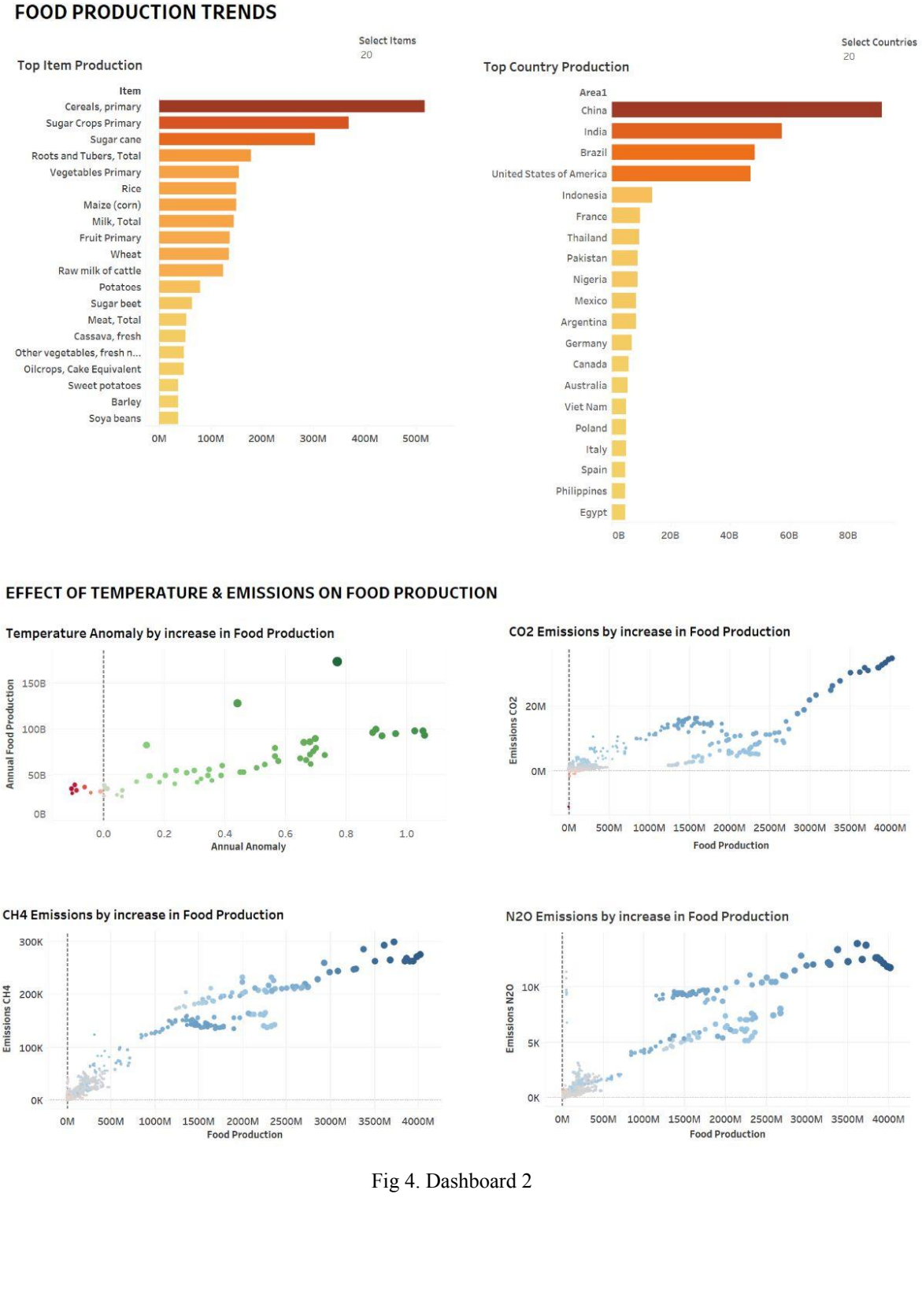


Fig 4. Dashboard 2

From the various visualizations in the two dashboards, it is evident that the average land and ocean surface temperature anomaly has increased over the years. Moreover, the quantity of the different emission types have also increased over time, with CO2 emissions witnessing the highest increase of 48% from 1990 to 2020. Increase in emissions along with the indication of the average surface temperature anomalies increase over time demonstrate the occurrence of climate change. In addition, food production data trends also show an increase in food production over time. Visualizing the correlation of the food production data and temperature data, and food production data and emission data does not reveal a clear trend of the negative impact on food production, and indicate a weak trend of increase in either of these factors also showing an increase in the amount of food production. Though temperature extremities and other indicators of climate change logically indicate decrease in food quality and quantity of food produced, this trend is not evident from the overall correlation between these factors of food production. A positive correlation trend can be explained due to the interplay of other variables like population and technological advancement over time, which has caused increase in food production to meet the growing demand. However, pockets of low food production in areas with extreme emissions and temperature might still be present.

## References

Berkeley Earth. (2023). Data Overview. *Berkeley Earth*. https://berkeleyearth.org/data/

Food and Agricultural Organization of the United Nations. (2023). *FAOSTAT*. FAOSTAT. https://www.fao.org/faostat/en/#home