1 Question

How will projected climate changes impact agricultural yields across different regions globally?

2 Data Sources

2.1 Data Source 1

Metadata URL: Crop irrigated data set Meta data
Data URL: Crop irrigated data set
Description: This dataset provides statistics about corn production held across the country of USA, which was not irrigated in Bushels per acre from 2005- 2022.

2.2 Data Source 2

Metadata URL: Crop not irrigated data set Meta data
Data URL: Crop not irrigated data set
Description: This dataset provides statistics about corn production held across the country of USA, which was irrigated in Bushels per acre 2005- 2022.

2.3 Data Source 3

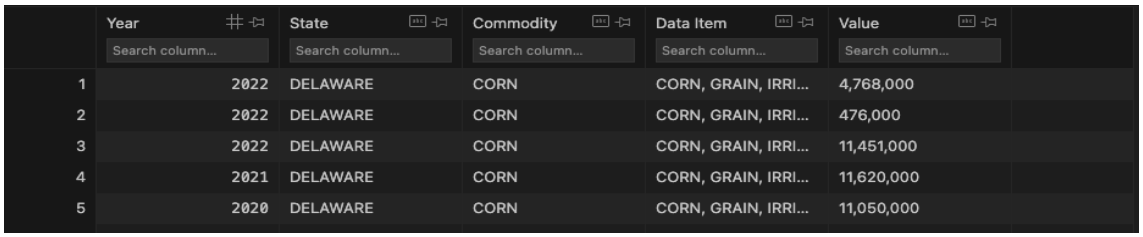
Metadata URL: Climate change Meta data
Data URL: Climate change data set
Description: This Data set gives a comprehensive report about climate change In USA between years 2004 - 2022.

3 Data Quality and Structure

Climate change dataset: This Data set gives a comprehensive overview of the climate change using parameters like Co2 emission, temperature, precipitation and wind speed and has the information about different countries.

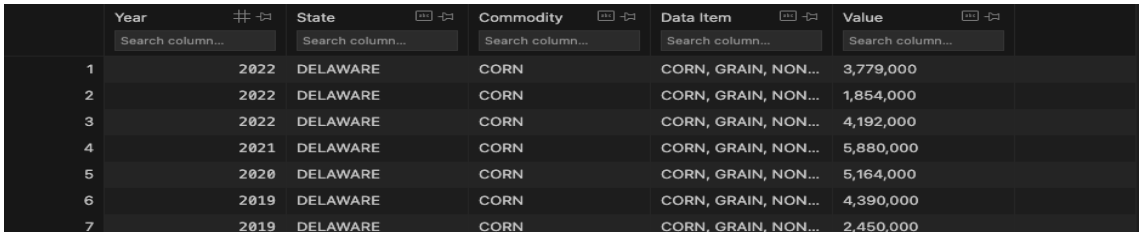
	Location	Country	Temperature	CO2 Emissions	Sea Level Rise	Precipitation	Humidity	Wind Speed	Year
	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...	Search column...
1	Garciatown	United States of Am...	13.504183317778...	354.39046690733...	-0.484336288143...	10.686071221106...	35.82337264062079	31.527323641426...	2000
2	Ryanmouth	United States of Am...	20.597688570452...	403.02910342156...	-0.489247112442...	70.12808267000979	38.63026210673013	25.181792780264...	2001
3	Wrightfurt	United States of Am...	20.559158338562...	384.6233096593341	-0.473106348178...	76.97812581643406	74.45045150867877	38.317539257428...	2002
4	Davisville	United States of Am...	17.573796063255...	342.6103138235293	-0.582807415762...	86.1978426989311	28.08530268701912	7.541851512570696	2003
5	South Christopherfurt	United States of Am...	24.61182598654026	402.0164215013944	-0.854038855847...	97.39930485051774	1.4108493967769...	0.4843512156384...	2005
6	Hollandbury	United States of Am...	23.203795793054...	492.149319872971	-0.436715454471...	64.96549506928099	16.72782840418585	6.373826914708836	2005

Corn irrigation data set: This data set provides information about crop which are irrigated from year 2004 - 2023, and has total values as well as the location of the irrigation done across the country, which helps to plot a trend.



	Year	State	Commodity	Data Item	Value
	Search column...	Search column...	Search column...	Search column...	Search column...
1	2022	DELAWARE	CORN	CORN, GRAIN, IRRI...	4,768,000
2	2022	DELAWARE	CORN	CORN, GRAIN, IRRI...	476,000
3	2022	DELAWARE	CORN	CORN, GRAIN, IRRI...	11,451,000
4	2021	DELAWARE	CORN	CORN, GRAIN, IRRI...	11,620,000
5	2020	DELAWARE	CORN	CORN, GRAIN, IRRI...	11,050,000

Corn not irrigation data set: This data set provides information about crop which are not irrigated from year 2004 - 2023, and has total values as well as the location of the irrigation done across the country, which helps to plot a trend.



	Year	State	Commodity	Data Item	Value
	Search column...	Search column...	Search column...	Search column...	Search column...
1	2022	DELAWARE	CORN	CORN, GRAIN, NON...	3,779,000
2	2022	DELAWARE	CORN	CORN, GRAIN, NON...	1,854,000
3	2022	DELAWARE	CORN	CORN, GRAIN, NON...	4,192,000
4	2021	DELAWARE	CORN	CORN, GRAIN, NON...	5,880,000
5	2020	DELAWARE	CORN	CORN, GRAIN, NON...	5,164,000
6	2019	DELAWARE	CORN	CORN, GRAIN, NON...	4,390,000
7	2019	DELAWARE	CORN	CORN, GRAIN, NON...	2,450,000

4 Licenses and Permissions

The crop irrigated and not irrigated data set are under open data license and public domain as it is fetched from a government site which is United States Department of Agriculture as it is a survey data. For the third data set it is mentioned under CC0: Public Domain as mentioned on the kaggle site by the owner.

5 Data Pipeline

5.1 Technologies used

- Python for cleaning and processing data.
- Packages inside python which were used are Numpy, pandas, Sql alchamy, searborn and matplotlib.
- Sqlite for Database storage
- Bash to run the pipeline and calling python files

5.2 Pipeline description

- The pipeline is triggered using a bash file, which in turn calls 2 files.

- The first Python, which is called the pipeline, is responsible for data downloading and processing.
- The data sets are being downloaded from Kaggle and the United States Department of Agriculture website.
- The Kaggle data set is then unzipped and then stored in the data directory.
- The data sets fetched from USDA are CSV files, so are straight away stored in the data directory.
- For processing of Data, any empty values are removed and all the unnecessary columns are removed which are not required
- Then all the files are stored in a Sqlite file with different table names.
- During the analysis of data, the data is fetched from the sqlite tables and analysed

5.3 Data processing steps

- Removing all the rows which has empty entries in rows.
- Droping columns like Location, Geo Level, Week ending, State ANsi etc.
- Filtering the Data in crops data set to just have data from 2005- 2022.
- Only storing the values for climate change related to just United States of America.

6 Result and Limitations

6.1 Results

- After processing the data is stored in a SQLite database which is a relational database.
- Fetching and reusing of data is easier if the data is stored in a tabular format

6.2 Limitations

- The data set for climate change is still not sufficient for better analysis.
- There are lot of inconsistencies for looking at a correlation between the climate change and crop yield