Data Mining Analysis on Climate Change and Global Grain Supply

**─**

Jason Ortiz, Rhodiam Arango, Daniel Cohen

Computer Science Senior Project

CSCI 455 - M01

NYIT Manhattan Campus

Advisor: Dr. Helen Gu

# Table of Contents

### Abstract 3

### Introduction 3

### Background 3

### Objectives 4

### Timeline 4

### Literature Review 7

### Dataset Exploration/Analysis 10

Data Descriptions 10

Data Visualizations 14

### Materials 18

### Methods & Design 19

Machine Learning and Data Mining Techniques 19

Predictive/Forecasting Models 19

Clustering Models 19

Decision Trees 20

Regression Models 20

### Data Preprocessing (Individual) 21

### Database Design 29

### Integration 30

### Data Preprocessing (Integrated) 31

### Clustering 31

### Decision Trees 32

### Decision Trees 32

### Results 33

Clustering 33

Decision Trees 33

Regression Models 34

### Conclusions 35

### Challenges 35

### Contributions 37

### References 40

Data Sources 40

Research Sources 40

### 

# Abstract

Climate Change has continued to change Earth’s environment, as a result of our actions over decades, and is still ongoing. We’re aware of the impacts that have already occurred as a result of Climate Change, but we need to look to the future, and data mining analysis can be helpful in analyzing how climate change will have an impact on the global grain supply chain now in our interconnected world, and into the future if we do not make drastic changes with regards to our co2 emissions.

# Introduction

In general, we already know that climate change exists, and that increasing carbon dioxide levels are a major part of that as we continue to burn fossil fuels every day. However, we need to look at how our world that’s being changed by climate change will impact one of the major food sources that we have, which is grains. We aren’t simply talking about one country however. We are talking about a world that’s interconnected, thanks to trade between countries, and how the affected production of even one staple crop can affect our global food supply for years or even decades to come if we do not take action to reduce our carbon footprint. As the human population continues to rise, food production will need to increase, and our reliance on developed countries responsible for the majority of grain production will increase further. Data mining analysis can be used on real world data to look at the existing trends that exist in our world, with regards to CO2 emissions and the global grain food chain, and what’ll be at risk if we don’t find a sustainable way to advance in technology and advance as a society of humans, without harming our world even more than we have.

## Background

We have a good understanding with regards to how the rising carbon dioxide levels will affect crops. According to a new NASA study, the increased carbon dioxide levels in the atmosphere can affect crops in 2 ways. One of these ways is by boosting crop yields through an increased rate in photosynthesis, which will result in higher growth. The other way is by increasing a plant’s water-use efficiency (Hille 2016).

NASA has also made projections about how Wheat and Corn yields and production will be affected by Climate Change as soon as 2030, going on into the future. NASA projected that Maize, aka Corn yields are projected to decline 24%, whole wheat could see a 17% increase in yields (Gray 2021). NASA’s climate and agricultural models helped scientists find that this is due to increases in temperatures, along with rainfall patterns shifting, along with the increasing carbon dioxide concentration in the atmosphere (Gray 2021). This highlights the idea that carbon dioxide emissions are not going to affect crops in the same way. There are crops that might benefit from the effects of climate change, but at the potential cost of detrimental effects to other crops.

With the idea of how climate change, spurred on by increasing carbon dioxide emissions can affect certain crops, we need to see what patterns currently exist in the global grain supply chain, to see what countries are at great risk of food shortages, should this supply chain decrease in efficiency due to crops not providing enough for the global population. We also need to look at which countries are taking on the burden of the supply, and how reliant the global grain supply is on those countries.

# Objectives

1. Use Machine Learning & Data Mining models and techniques to analyze and visualize very large datasets to find new patterns and reinforce expected patterns regarding Climate Change and Global agricultural data.
2. Using the reinforced patterns, alongside the new patterns that have been discovered within the data, make some conclusions on how climate change will impact the global grain supply.

# Timeline

## 02/23/2022 - 03/02/2022

We started to analyze and visualize various weather and agriculture datasets, including both historical data as well as predictions made by previous researchers. We will take a closer look at which features will be important throughout the project through correlation analysis, as well as models used by previous researchers.

## 03/02/2022 - 03/09/2022

Dataset exploration and analysis, including visualizations, and choose which datasets are most relevant to our project. We also need to establish a proper timeline for the entire project, as well as establish everybody’s contributions.

## 03/09/2022 - 03/16/2022

We are going to finalize the preprocessing stage of our chosen datasets, along with doing feature selection, along with providing reasoning for any removed features from any data sets. We are also going to look and choose some of the most relevant algorithms for our project.

## 03/16/2022 - 03/30/2022

Making sure that all of our datasets are capable of being integrated within each other, so that data can remain consistent. Once all the data has been prepared, we use ERD diagrams and plan out how to integrate our data into a SQL database, which is what we’re going to use our algorithms and machine learning approaches on.

## 03/30/2022 - 04/06/2022

Our first iteration of integrating the prepared datasets into a SQL database. We had run into a few issues while attempting to create our integrated database because of how our structure for the database was determined. We needed to make sure countries were able to have missing agricultural data for a year where they may have reported CO2 emissions data, but no agricultural data. We ran into a similar issue with the CO2 emissions data, so we needed to have the Berkeley dataset provide the data for the year in which the data was recorded, in order to allow CO2 and Grains data to be null.

## 04/06/2022 - 04/13/2022

Insertion of our datasets into the database gave rise to a few issues because of how the foreign key restraints worked for our dataset, which eventually was resolved. Now that the database had been successfully integrated through Python, we had to start making decisions about what models we were going to be using. We had considered using Regression models at first, along with neural networks, but we weren’t clear on our next steps, which ended up delaying our progress.

## 04/13/2022 - 04/20/2022

Regarding the values of our dataset, we had to take a step back and deal with how the values of the CO2 dataset were originally imputed. At first they had been imputed simply by replacing 0, which was a messy way of dealing with real world missing data. In the end, the dataset’s missing values were imputed by KNN, to determine the most probable value to replace the missing value, which was a better option than simply replacing using mean or mode. We had also brought Latitude and Longitude into our database, which ended up leading to a redesign of our database design/structure. Afterwards, we did some testing with KMeans and DBSCAN clustering methods.

## 04/20/2022 - 04/27/2022

We adjusted our clustering procedure, by normalizing the numeric attributes that could be normalized, in order to keep our attributes on a common scale, to keep clustering cleaner. At first, clustering was tested using subsets of the data, but then was applied to the entire dataset. We had also considered trying to do Association Rule Mining, where we had hit a dead end with regards to the binning methods, and the warnings that would come up as a result. Agglomerative clustering was another method that was also being tested with the entire dataset. One of the biggest issues we faced with regards to clustering at this stage was trying to interpret the actual cluster results.

## 04/27/2022 - 05/04/2022

This week, we had to come together and make progress on our clustering, being able to draw some observations from them and the data in general, and decide what clustering methods have been good for our data, bad for our data, and ultimately decide which ones we were going to use. Clustering has been expanded by using methods such as HDBSCAN and OPTICS. We used PCA in order to reduce our high dimensional data into low dimensional data, and performed clustering with the Principal Components. We made different comparisons and observations regarding the quality of the clusters, especially regarding the number of clusters we would be using. We had finally determined that we were using DBSCAN clustering with 10 Clusters

## 05/04/2022 - 05/10/2022

For this final week, we are coming together to finish the project by using decision trees on our clusters to see what attributes might’ve been most impactful when it comes to splitting the data into labels/clusters. We’re also working on regression analysis of our clusters. We will make our conclusions based on the observations we’ve made from the data and our clusters.

# Literature Review

***CO2 Emissions: Our World in Data***

[CO2 emissions - Our World in Data](https://ourworldindata.org/co2-emissions)

This source provides detailed information and data on a country’s CO2 emissions, and presents detailed metrics in the data set. Some of these metrics include annual emissions, average emissions per person, how much CO2 has been emitted over time. The data even looks at CO2 emission due to coal, oil, gas, flaring, or cement production, etc. Overall, this source is very detailed with regards to CO2 emissions, and the source of those emissions, which is very positive for our project, and definitely provides a very useful data source. Something missing from this source specifically, which would be a huge help for our project, is a metric showing at the very least, an average temperature measurement every year. Looking at the change in temperature over a period of time is crucial for our machine learning project. Despite this however, this source still remains very beneficial because looking at the CO2 emissions for individual countries can give us an idea as to which countries will see a greater temperature change, due to the correlation between CO2 emissions and climate change, and overall temperature change. Aside from the importance of the data set and the different visualizations provided, there is also information provided on the importance of the sources behind the CO2 emission, like Greenhouse Gasses, and coal, oil, gas, etc. This source overall provides useful information not only for the application of our project, but from an informational standpoint, so that we know more about the topic that we are focusing on.

***Predicting Weather Temperature Change using Machine Learning Models***

<https://medium.com/swlh/predicting-weather-temperature-change-using-machine-learning-models-4f98c8983d08>

This source focuses on the use case of using Machine Learning models in order to try and predict Weather Temperature change. The source goes through an in-depth explanation of how Machine learning is used for this topic/use case. It goes through the different steps that are used in machine learning, such as getting the data set, preprocessing the data, using visualization, and then applying a machine learning algorithm. This source also provides another data set which is going to be useful for our project because we’re trying to predict global temperatures, but are focusing specifically on the agricultural impacts, so we’re expanding our project beyond just trying to predict future temperatures. Overall, this is a useful source not only because of the data set, but because we can take a look at the use case, and how machine learning was used here, to see whether we could improve upon their approach to try and get more accurate results.

***13 Machine Learning Use Cases for Climate Change***

<https://vitalflux.com/machine-learning-use-cases-climate-change/>

This source looks at several different use cases where Machine Learning focuses on the topic of Climate Change, ranging from estimating CO2 emissions to predicting extreme precipitation. It even describes the use case of predicting climate change impacts on crop yields, which is very similar to what we are attempting to do within our project. While it’s good that this source describes the different use cases for Machine Learning and Climate Change, they don’t really go into detail about how Machine Learning is actually used here. It’s more of a brief explanation as to where Machine Learning is used, instead of how, which doesn’t really help us, since we already know what we’re trying to focus on with our project.

***Global Climate Change Impact on Crops Expected Within 10 years, NASA Study Finds***

<https://climate.nasa.gov/news/3124/global-climate-change-impact-on-crops-expected-within-10-years-nasa-study-finds/>

This source from NASA explains the results that were taken from a NASA study, which found that Climate change could potentially impact maize production as early as 2030. The study found that Maize crop yields are projected to decline 24% while wheat production could actually grow by 17%. The source explains that scientists found the changes in crop yields were mainly due to “projected increases in temperature, shifts in rainfall patterns, and elevated surface carbon dioxide concentrations from human-caused greenhouse gas emissions''(NASA). This source is very useful because it describes the results from a study that is similar to ours, where we are trying to look at the increase in temperature and carbon dioxide emissions to see what kind of agricultural impact it would have. Since this source gave a small idea as to what to expect from our models, we have a sort of baseline to see whether our results/predictions are in some way similar to what NASA had predicted, especially since this study was very recently done, only a few months ago.

***Food and Agriculture Organization Macroeconomy Map*** <https://www.fao.org/3/i2490e/i2490e01c.pdf>

This source, which is the Food and Agriculture Organization of the United Nations, gives detailed descriptions and diagrams/graphs on a country’s agricultural scope. One map, for example, shows every country’s share of agriculture in GDP, which was taken from 2009. This shows us which countries are more agricultural than others, which gives us an idea of which countries could potentially be impacted due to climate change conditions. There is also data provided which shows a country’s population total and growth, along with agricultural population. This helps supplement the previous map in giving us an idea as to how much of a particular country is agriculture based. This source is very useful because as we look at the results that we get from our Machine learning models, this source will help us get an idea of what regions will be hit harder from an agricultural standpoint. The source’s multiple data sections and maps provide us with in-depth information on each country, allowing us to come to more accurate conclusions for our project.

***Rising Carbon Dioxide Levels Will Help and Hurt Crops***

[Rising Carbon Dioxide Levels Will Help and Hurt Crops | NASA](https://www.nasa.gov/feature/goddard/2016/nasa-study-rising-carbon-dioxide-levels-will-help-and-hurt-crops)

This source looks specifically at the impacts that rising carbon dioxide levels will have on crops, which relates heavily to one of the data sets that we have, which looks at carbon dioxide emissions from each country. This source, in particular, mentions how higher concentrations of carbon dioxide in the atmosphere boosts crop yields by increasing the rate of photosynthesis, which requires carbon dioxide in the first place. Higher concentrations of carbon dioxide in the atmosphere also allows crops to lose less water, due to the impact it has on crops’ transpiration. This is an important source of information to look at, because while we are trying to predict global temperatures, we have to keep in mind that temperatures are not the only factor that affects crop yields. We have to take into account that the rising carbon dioxide emissions provide some benefit to the crops, but the question is, does the benefit of rising carbon emissions outweigh the harm of the overall global temperature increase. This is why this research source is important to keep in mind as we work with temperature and carbon dioxide datasets.

# Dataset Exploration/Analysis

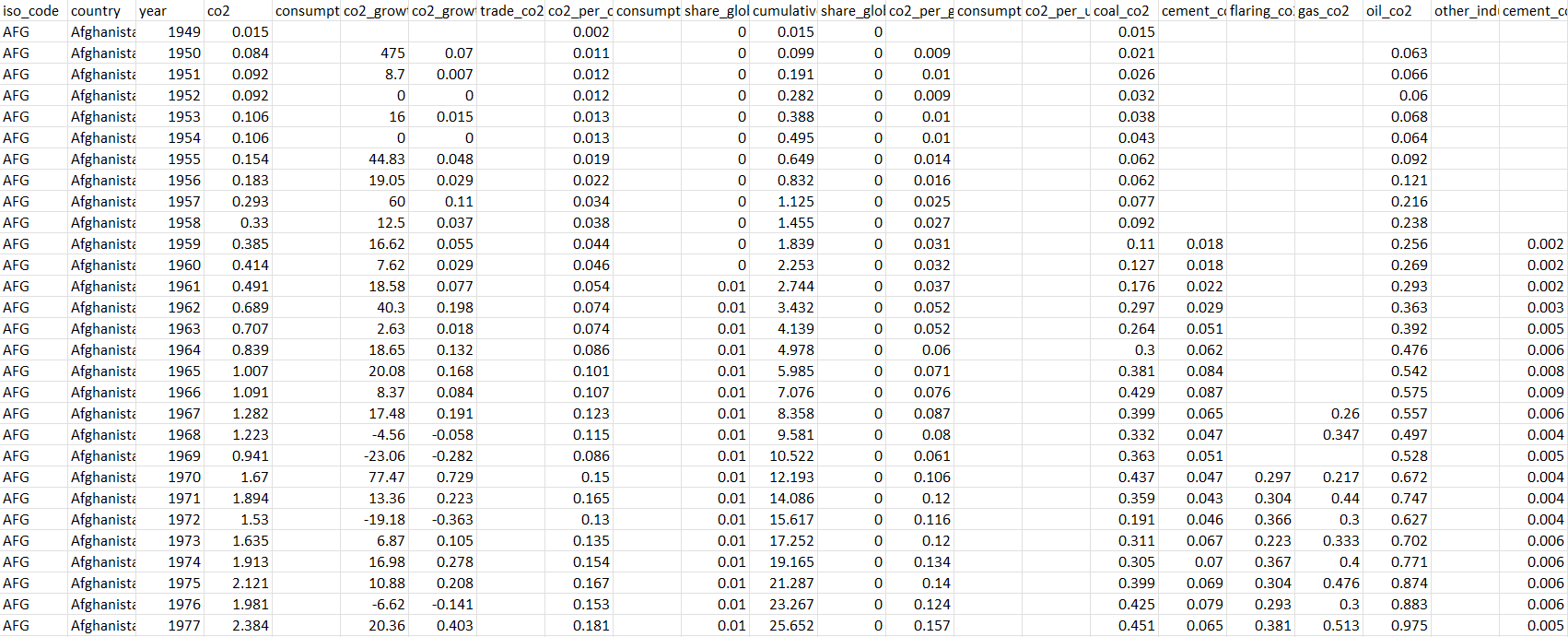
## Dataset Descriptions

**Our World In Data - CO2 Emissions**

**Dataset Source:** [**co2-data/owid-co2-codebook.csv at master · owid/co2-data · GitHub**](https://github.com/owid/co2-data/blob/master/owid-co2-codebook.csv)

This data set describes the carbon dioxide emissions of every country in the world, throughout the history of the country. Some countries have more reported data than others depending on the number of years where carbon dioxide emissions have been reported. Carbon dioxide emission metrics in this data set is measured based on different sources, such as greenhouse gasses, flaring, coal, oil, gas, etc. The absolute/percentage growth of CO2 emissions is also measured, along with production/consumption based carbon dioxide emissions. Overall, this data is incredibly detailed with regards to a country’s carbon dioxide footprint.

This dataset is beneficial for our project because since we’re focusing on the impacts of climate change on regional agriculture, CO2 is a very important factor to focus on, especially considering the close relationship between CO2, Agriculture, and Temperature.

****

This data set has a total of 60 attributes, which are mainly numerical/interval, and there are a total of 25,192 objects.

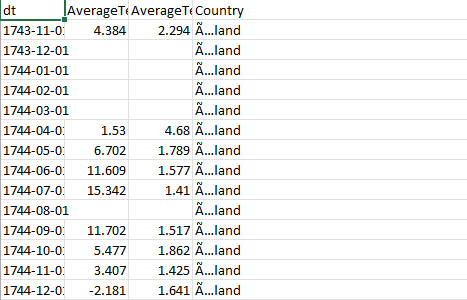
**Climate Change Earth Surface Temperature Data**

**Data Set Source:** [**Climate Change: Earth Surface Temperature Data | Kaggle**](https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data)

This data set describes the average land temperatures in a given area at a certain time. Specific coordinates of the location is given within the data set, along with the average land temperature, and the level of uncertainty with regards to the average land temperature, which could be considered the standard deviation. Since the data is split among 5 different csv files, you can choose the level of detail that you are looking for when it comes to the average land temperatures.

This data set is beneficial for our project because looking at the trend of the earth’s surface temperature over time allows us to be able to create an algorithm to try and predict future surface temperatures. With those predictions, we’ll be able to come up with a more detailed analysis regarding the impact that climate change can have on regional agriculture.

**Dt- refers to the specific Date-Time**

****

This data set has 4-9 attributes, which are mainly interval attributes, depending on what focus is being utilized in this data set. The data set is split into 5 different levels, based on either city, country, major city, state, and general global temperatures since the 1750s. As a result, the number of objects ranges from approximately 2100 to approximately over 1 million objects.

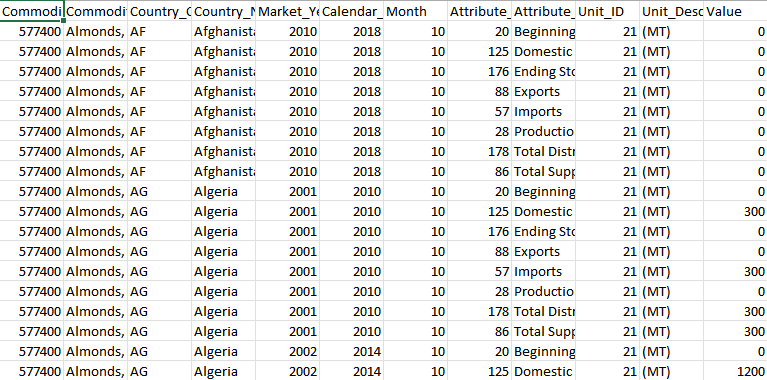
**USDA Foreign Agriculture Service PSD All Commodities Data/ Grains Data**

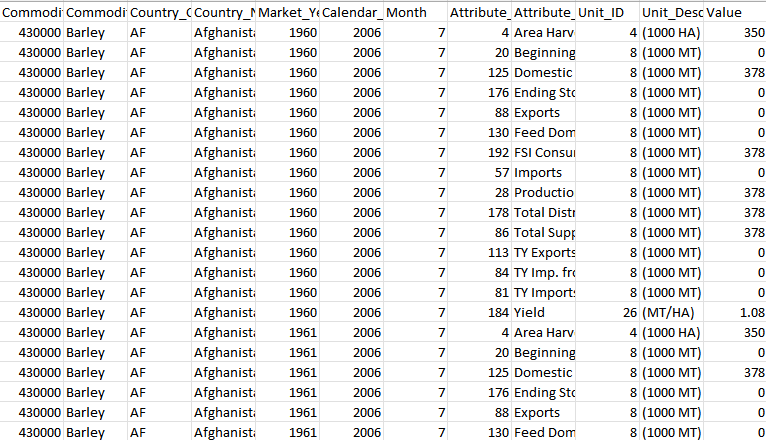
**Data Sets:**

<https://apps.fas.usda.gov/psdonline/downloads/psd_alldata_csv.zip>

<https://apps.fas.usda.gov/psdonline/downloads/psd_grains_pulses_csv.zip>

The first dataset looks at the market and trade data of all commodities around the world. The data set focuses on a country’s market and trade data, depending on the market year, the commodity in question, and what type of market data (Exports, Imports, etc.), along with the value which when multiplied by the Unit\_Description will give the total value of that object.

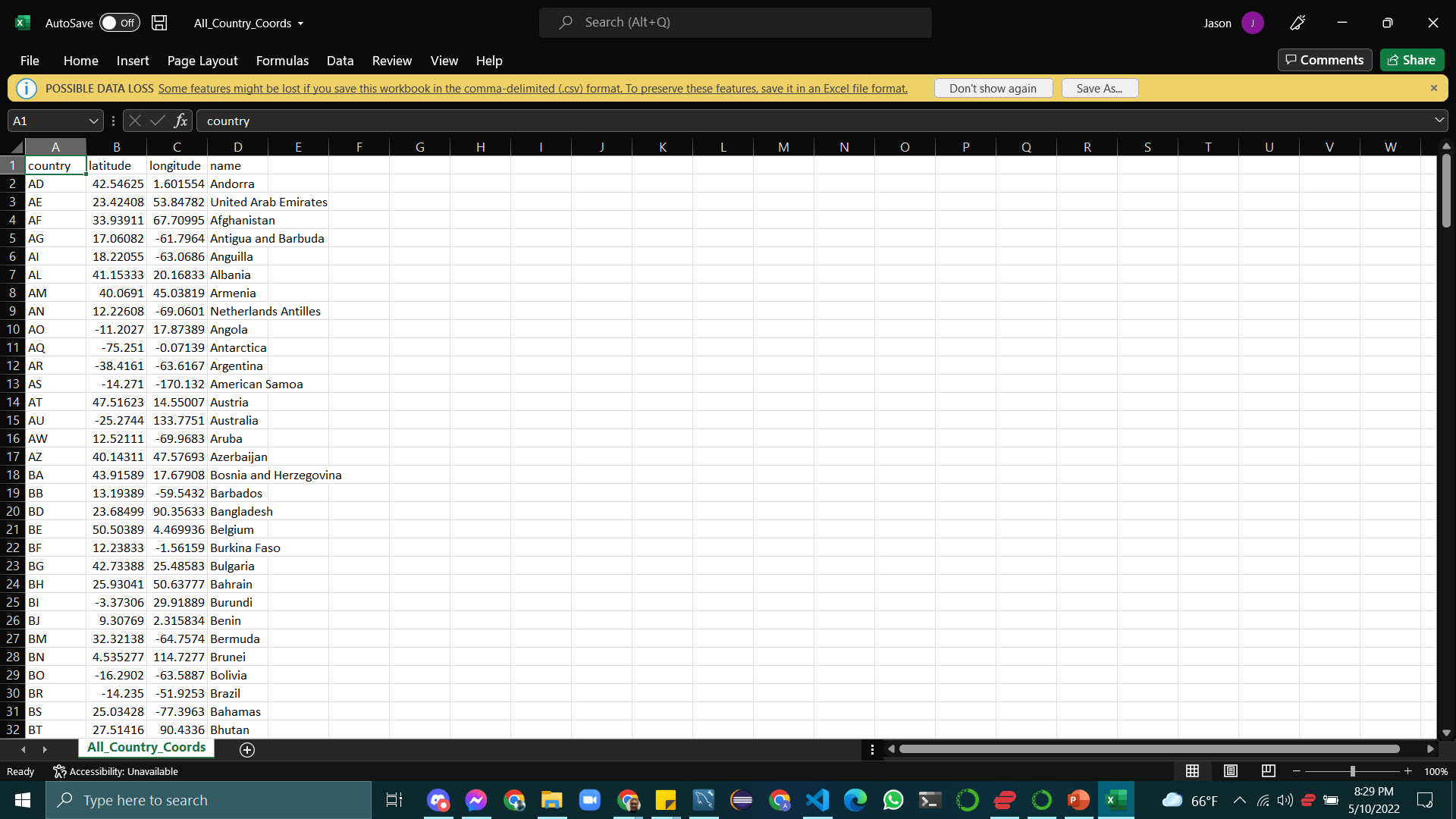


The second dataset looks at the market and trade data of different grains around the world. The data set focuses on a country’s market and trade data, depending on the market year, the grain in question, and what type of market data (Exports, Imports, etc.), along with the value. 

These 2 datasets have a total of 12 attributes, which consists of a combination of both Nominal and Interval attributes. The Grains Dataset contains approximately 535,000 objects, while the “All Commodities” dataset contains over a million objects.

**Latitude/Longitude Data for Countries**

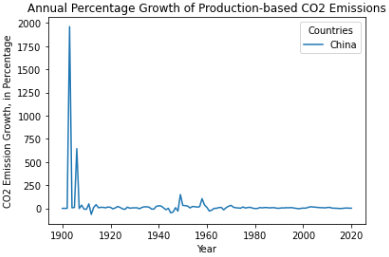
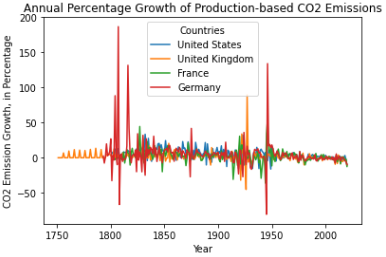
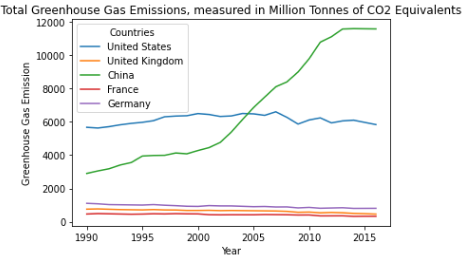
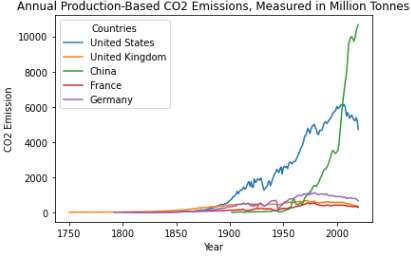
[**countries.csv | Dataset Publishing Language | Google Developers**](https://developers.google.com/public-data/docs/canonical/countries_csv)

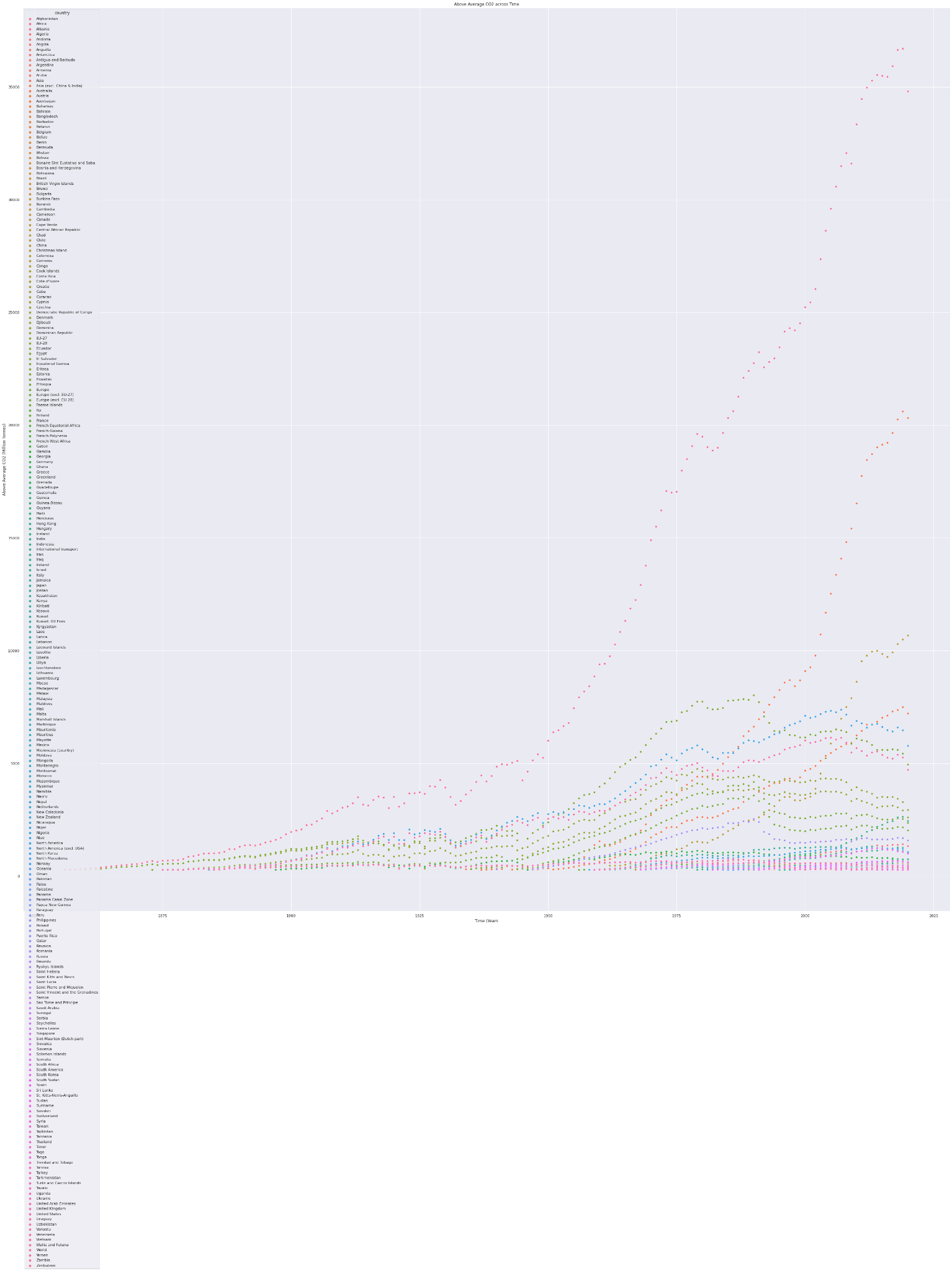


The countries.csv dataset contains the Longitude and Latitude of 245 Countries, of which we only required 136 Countries. Not all countries from this dataset have corresponding data in all of the datasets. One more note is that some of the countries in this dataset are not true countries and rather territories of other countries, like Puerto Rico, for example.

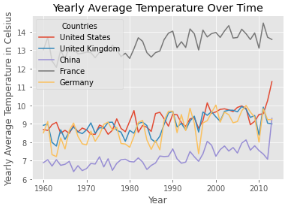
## Data Visualization

**Our World In Data - CO2 Emissions Line Chart Visualizations**

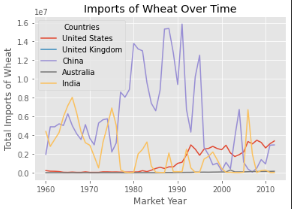


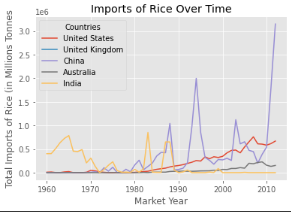
**Our World In Data - Above Average CO2 Emissions Over Time, per Country**

Berkeley Temperature Data Visualization



**PSD Grains Visualization**

****

****

# Materials

* Integration/Clustering Tools:
  + Python has many useful libraries to use for data analysis
    - Pandas
    - NumPy
    - SKLearn
    - Streamlit
    - HDBSCAN
  + MySQL allowed for us to create the Relational Database that we will use to integrate & store our data for queries
* Visualization Tools:
  + Python libraries such as:
    - Seaborn
    - Matplotlib
    - Streamlit
    - Pandas Profiling
* Communication Tools:
  + Google Drive:
    - Sharing files
    - Real Time Collaboration
  + Discord:
    - Group chat feature with collaboration tools (such as pinning important messages and/or links)
  + Draw.io:
    - Useful for modeling out the data into an Entity Relationship Diagram

# Methods & Design

## Machine Learning and Data Mining Techniques

* Various visualization techniques and tools for data analysis and visualization.
* Avoiding the “Curse of Dimensionality” by using techniques such as feature selection and Principal Component Analysis.

## Predictive/Forecasting Models

With the datasets that we chose consisting of unlabeled data, we decided that the models we would focus on would be predictive models; focusing on future temperatures, grain yield, co2 emission, etc.. These models will be used to analyze and visualize the datasets we are working with. We chose these with the hope of discovering patterns from the data, as well as predicting future instances of the data and the effects those predictions will have. The models used for predictions will be checked for accuracy/error and will be tuned and compared with other model’s predictions. These models will consist of different types of models including: Regression & Clustering. These models will be further explained below with how they work and how they are relevant to our data.

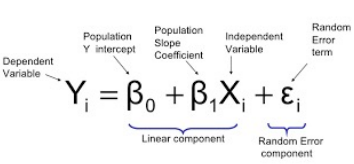
## Clustering Models

* KMeans
  + A clustering method that will separate the dataset into K clusters, which is provided by the user. Objects within a cluster share similarities, and are dissimilar to objects within another cluster.
* DBSCAN
  + A density based clustering method that uses both the Euclidean distance metric as well as two parameters Epsilon, radius around a point that determines its ‘neighborhood’, and min\_samples, which determines the minimum number of points necessary to be considered a core point. All points in a core points’ epsilon radius are then considered a part of that cluster. If another point that was added to the cluster is labeled as a core point it can ‘hop’ across dense clusters to find shapes other clustering methods would have missed.
* HDBSCAN
  + An extension of the DBSCAN clustering method, by transforming/converting it into a hierarchical clustering algorithm. Performs DBSCAN over varying epsilon values in order to find the most stable clustering
* OPTICS
  + Stands for Ordering Points to Identify Cluster Structure. Draws inspiration from the DBSCAN clustering algorithm, but uses 2 more parameters than DBSCAN. This clustering algorithm is different by visualizing how clusters would be formed.
* Agglomerative Clustering
  + A bottom-up clustering method, where each data point is its own individual cluster, and these clusters will start mixing into bigger clusters until it becomes one giant cluster. One can choose the number of clusters they want from Agglomerative clustering by choosing the cutoff point through a dendrogram.

## Decision Tree Models

* Decision Tree Classifier
  + A supervised learning method, specifically a tree-based classifier, which tries to predict class labels for data, based on the true labels of a dataset. It splits the data based on specific conditions for a selected attribute, and continues until no more splits can be made, or a specified max depth is reached.
* Decision Tree Regressor
  + Similar to a Decision Tree Classifier, but uses a different criteria compared to the classifier tree. Instead of trying to fit the data into a class label, however, it tries to predict the target value based on the other attributes.
* Random Tree Regressor
  + Follows the same concept as the Decision Tree Regressor, but is meant to try and increase accuracy of the model. The goal is to generate

## Regression Models

* OLS (Ordinary Least Squares)
  + A simple or multiple linear regression model that assumes no strong correlation between any two independent variables.
  + First OLS Assumption - Linearity: must be a linear regression/must be a linear equation (Not exponential or logarithmic)
  + Second OLS Assumption - Endogeneity - Assumes no relationship between errors and the independent variables (omitted variable bias).
  + Third OLS Assumption - Normality - Error term must be normally distributed, have an expected value of zero on the mean of error terms.
  + Fourth OLS Assumption - No Autocorrelation - errors are assumed to be uncorrelated.
  + Fifth OLS Assumption - No Multicollinearity - Multiple variables cannot have a high correlation.
* Simple Linear Regression Model**:**
  + Will be used to analyze the relationship between features in the clusters formed. Each feature played a different role in the creation of the cluster so we experimented with using those features to further analyze their relationships.
  + Will be used to predict future temperatures, CO2 emissions, and grain crop yield
  + Can be used as a simple regression model to compare with other model’s predictions.
    - Will calculate/validate with the R2 metric.



* Multiple Linear Regression Model**:** 
  + We will be using multiple linear regression to use multiple attributes to predict the output, ideally creating a more accurate model.
  + We will predict the same values and compare the accuracy with the other models, such as simple linear regression.

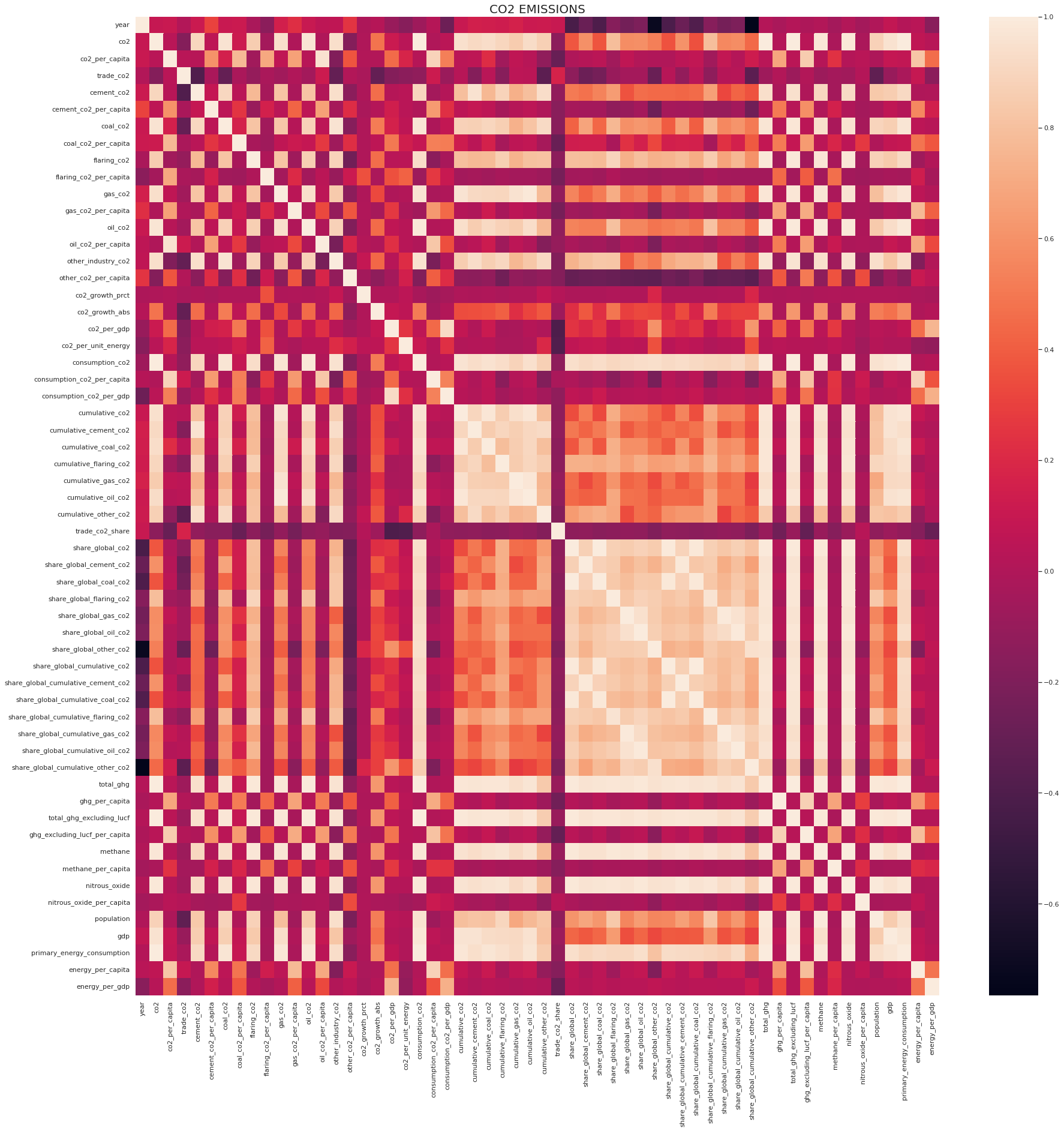
# Data Preprocessing (Individual Datasets)

**Our World In Data - CO2 Emissions**

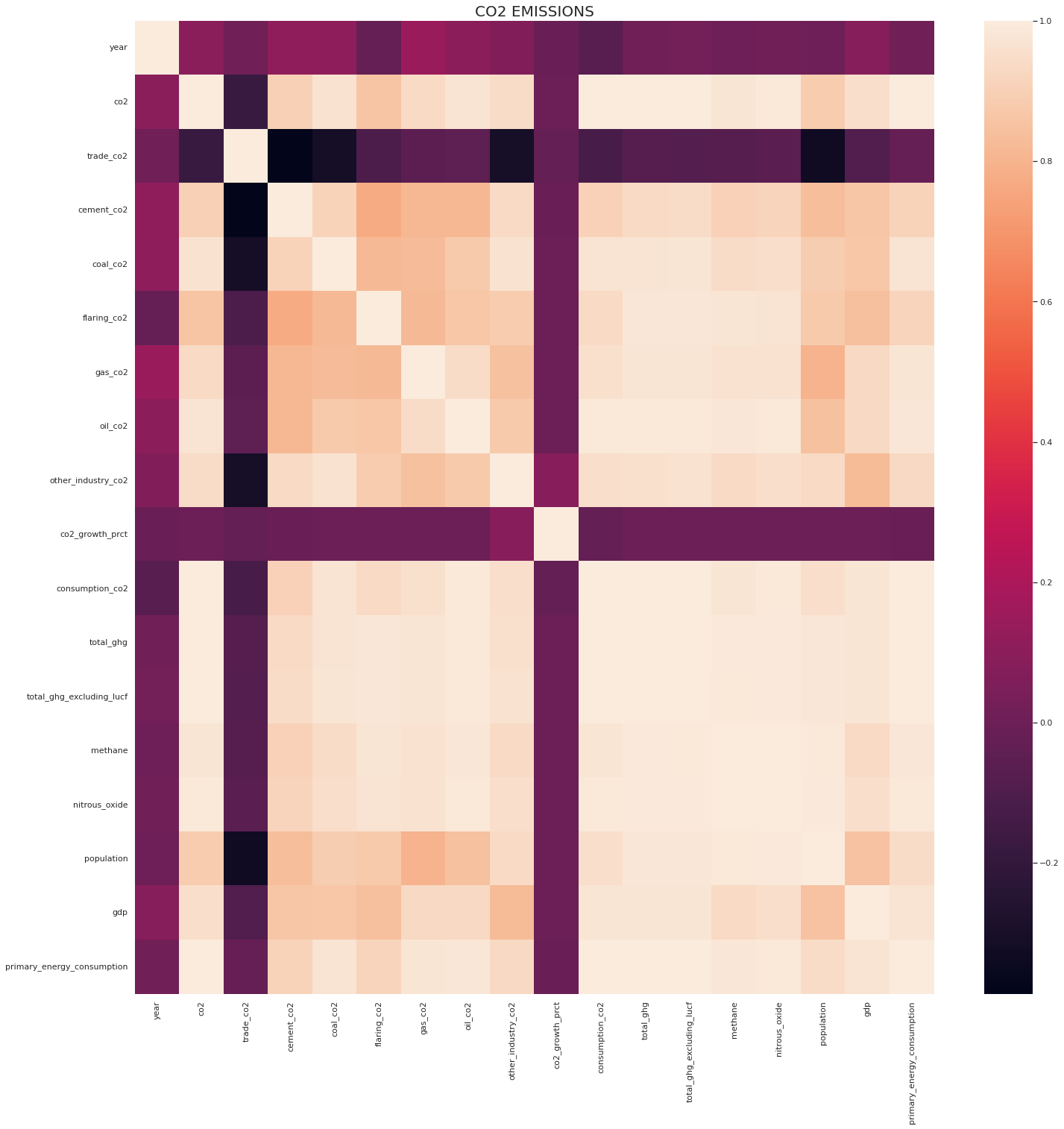
**Dataset Source:** [**co2-data/owid-co2-codebook.csv at master · owid/co2-data · GitHub**](https://github.com/owid/co2-data/blob/master/owid-co2-codebook.csv)

This dataset contains a total of 60 different attributes, which will make data mining a lot more difficult to get clear accurate results, so some of these attributes which are irrelevant to our focus/project need to be removed. A lot of the attributes here had high correlations to each other, so careful analysis needed to be done to remove irrelevant/unnecessary attributes.

Below is a representation of the correlation heatmap for the entire dataset with its 60 attributes. The correlation heatmap shows how each attribute correlates with every other attribute, to determine which attributes can be safely removed without impacting the quality and accuracy of the data.



Before using the correlation heatmap, however, a majority of these attributes can already be removed because they are redundant, simply using calculations from data already existing in the dataset. Attributes that have labels such as ‘per\_capita’, ‘cumulative’, ‘share’, and ‘per\_gdp’ are a result of calculations through multiple already existing attributes, in order to create a new attribute. In this dataset, 40 of these attributes fall under this category. As a result of this, they were removed from the dataset, since they do not actually offer any new information. Based on the correlation heatmap as well, these attributes heavily correlate with each other, which helps justify the reasoning behind removing these attributes.



This is the correlation heatmap of the end result, when removing all of those attributes mentioned previously, leaving us with 18 numerical attributes, and 2 categorical attributes. The ‘co2\_growth\_abs’ was also removed from the dataset, since this was another representation of the ‘co2\_growth\_prct’ attribute, and was redundant. Looking at the new correlation heatmap, we can see that a lot of these attributes heavily correlate with each other, which make sense given the context. A majority of these attributes consist of co2 emissions, specifically based on the source which is causing the emission. All of the emissions, specifically based on source, are added together to create the ‘co2’ attribute, and we can see that the ‘co2’ attribute heavily correlates with every single attribute in the dataset. As a result, we can remove the co2 attribute from our machine learning process, since it is another redundant attribute which is simply a sum of multiple attributes. We cannot remove the attributes specified by source, however, because there is no way to acquire this information from the ‘co2’ attribute. Other attributes such as ‘gdp’, ‘population’, ’primary\_energy\_consumption’, ‘total\_ghg’, and etc. provide additional information on the climate change impact that a country has on the planet.

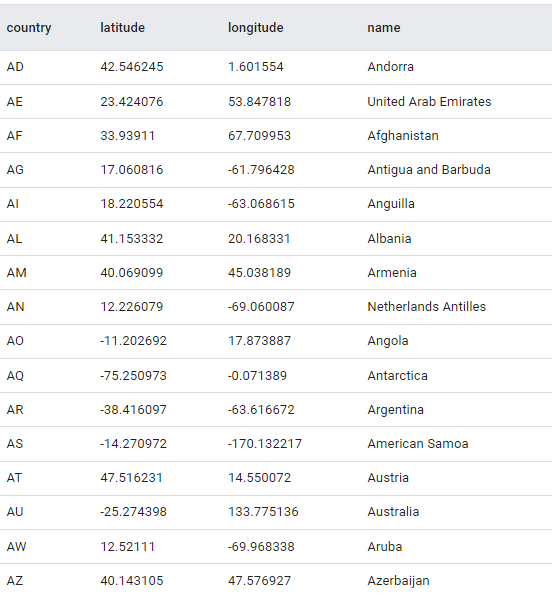
Eventually, the ‘co2\_growth\_prct’ attribute would also end up being removed from the dataset, because it was derived from the ‘co2’ attribute, which was derived from multiple attributes.

As a result, our dataset remains with 18 attributes now, with 2 categorical attributes, and 16 numeric attributes.

**Latitude/Longitude Data for Countries**

[**countries.csv | Dataset Publishing Language | Google Developers**](https://developers.google.com/public-data/docs/canonical/countries_csv)

This dataset contains the latitude and longitude coordinates of every country in the world. No real preprocessing needed to be done on this dataset, aside from the preparations that come with integration for the datasets. This dataset has 4 attributes, one showing the country name, 2 attributes for latitude and longitude respectively, and the final attribute showing the country abbreviation.



**Berkeley Earth Surface Temperature**

**Dataset Source:** [**https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data**](https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data)

This data set was collected from multiple sources and is broken down into multiple different data types of differing specificity in location, eg. GlobalLandTemperaturesByCity, GlobalLandTemperaturesByState, GlobalLandTemperaturesByCountry, etc,. The GlobalLandTemperaturesByCountry Dataset, which is the dataset that we chose to focus on, has 4 attributes; dt (Date), Country, AverageTemperature OC, & AverageTemperatureUncertainty. When initial attempts were made to plan out how the datasets would be integrated it was decided that to keep consistency between all datasets we would use Country and Year since all three datasets both had countries that they all share and a common timespan range, which helped us move forward with our datasets. Using the countries and Years shared between the three datasets allowed us to look at the GlobalLandTemperaturesByCountry under a different light as now what was needed was to filter the data to fit within the chosen timespan (from 1960-2012) and unique countries (136 shared countries between the three datasets).

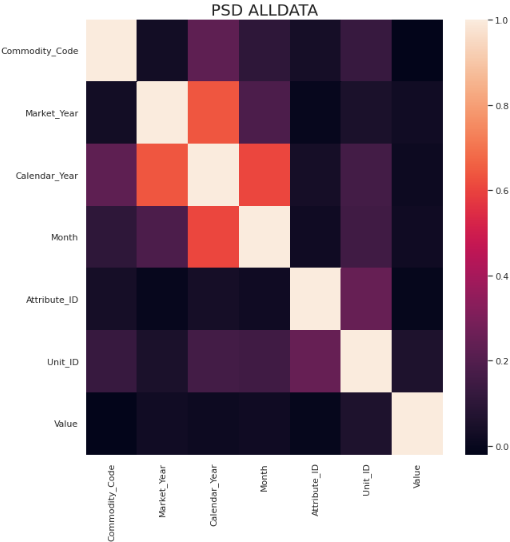
The four data sets that relate to global land temperatures by city, etc. are probably going to keep all attributes except Latitude and Longitude, which are simply another representation of the location where the temperatures are related to.

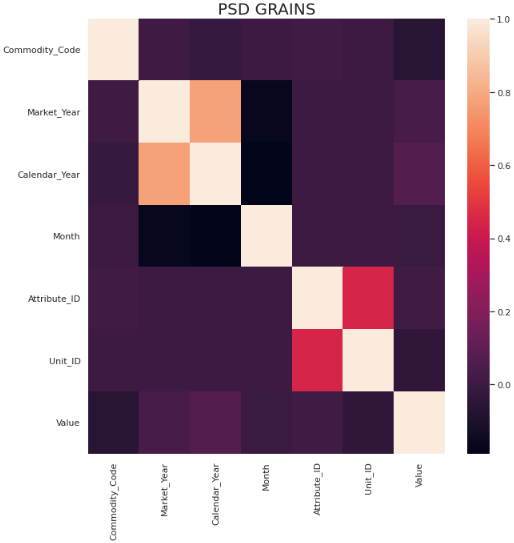
For the main Global Temperatures data set, while all of these attributes are relevant, attributes relating to “LandAndOcean” are going to be removed because that measurement is taking into account the temperature of the oceans around the world. Our focus is on regional agriculture, which isn’t going to have ocean temperatures as a major factor.

USDA Foreign Agriculture Service PSD All Commodities/All Grains Data

**Data Set:** <https://apps.fas.usda.gov/psdonline/downloads/psd_alldata_csv.zip>

**Data Set:** <https://apps.fas.usda.gov/psdonline/downloads/psd_grains_pulses_csv.zip>





Based on the correlation heatmaps of both the Grains and All Commodities datasets, there isn’t really any major correlation between the attributes, aside from attribute id and unit id, which are heavily linked, and market year/calendar year.

The shape of the dataset was changed, so that the values of the ‘Attribute\_Description’ column would become attributes themselves. This allowed us to remove the ‘Attribute\_Description’ column, since it was now redundant.

Other attributes such as ‘Unit\_ID’, ‘Unit\_Description’, and ‘Attribute\_ID’ were removed, since they weren’t necessary, since each agricultural attribute in the dataset was an actual attribute.

Attributes ‘Total\_Supply’, ‘Total\_Distribution’, ‘Yield’, and ‘Domestic\_Consumption’ were removed from the dataset, because they are derived from multiple attributes within the dataset.

Attributes ‘Calendar\_Year’ and ‘Month’ were removed from the dataset early on because these 2 attributes were relating to when this dataset was updated by the USDA, which is irrelevant to any of the actual agricultural data.

# Database Design

Initially there was no clear plan as to what entities were going to ultimately be used for our database, however through taking a step back and fully understanding the data that we collected we were able to agree on the direction to take. It was decided, due to the nature of our data, that we would have two main entities, Countries and Grains. Due to the fact that all the datasets had two overlapping attributes throughout, which were Time and Country, we were able to see the connection between the datasets and their attributes clearer. It became clear that each instance of a country/year represented a record of data, each country having its respective attributes, with each country growing and maintaining zero to many types of grains.

Our initial relational database design consisted of 2 tables. One table would be called Countries, and hold information from both the CO2 dataset and the Weather dataset, and the other table would be called Grains, and hold simply data from the Grains dataset. We created a Record\_ID attribute in the Countries table so that we would be able to identify every unique pair of Country and Year, and we had a foreign key that would correspond to the Grains table as well. However, because of the fact that we were essentially trying to insert 2 data frames into one table of the database at once, by separating the columns and making them separate dataframes, we ran into all of these issues. As a result, we needed to take a step back and restructure our database.

The second iteration of the database now consists of 3 tables. One table is called Weather, which has the attributes of the Weather dataset, along with a Record\_ID attribute. This became the Parent table, because we realized that the Weather dataset had the most complete representation of data among all 3 datasets. In other words, all of the countries in this dataset started reporting in 1960. The second table would be Countries, which held all of the co2 attributes, along with a record\_Id, which is used as the foreign key for the Weather table. Countries\_Record\_ID would be used solely as the primary key for the Countries table, in the same way as the Weather table uses Record\_ID. The final table was then Grains, which has all of the PSD attributes, along with the same Record\_ID foreign key as Countries, and a Grains\_Record\_ID, which serves as the primary key for the Grains table.

The final iteration of the database consists of 3 tables, similar to the previous iteration. The major difference between these 2 iterations is that the attributes from the Berkeley dataset were put with the same attributes from the Our World In Data - CO2 Emissions dataset, to form the Countries table. The Grains table remained the same, while there would be another table called CountryLocation, which holds the name of the country, along with its latitude and longitude coordinates.

# Integration

One of the main priorities after preprocessing all of our individual datasets, was to integrate all of them into a SQL database. In order to do this, however, we needed to make sure that our datasets shared some common ground, so that we could minimize as many null rows as possible, without losing a sizable portion of quality within the data. This led to us gathering a list of common countries within all 3 datasets, along with a common range of values for the year, since the starting year for data reporting varied greatly. As a result, we got a total of 136 countries between all 3 datasets, and our data ranges from 1960 - 2012. One issue that was discovered in the PSD dataset, was that the United Kingdom had no data reported, because due to its separation from the European Union, the UK had historical data starting from 2016 onwards reported, and this fell outside of our year range. As a result, we needed to shape our database to allow for the possibility of having years with no agricultural data, even if the country reported CO2 or Temperature data.

Another issue that was found during the algorithm creation, in order to insert our data into the database, was that a few countries hadn’t started reporting in 1960, but later on, and based on how we were planning our algorithm, this would become a huge issue. As a result, the 5 countries that were the issue were separated from the co2 dataset, and placed into its own csv file, so that those rows would be integrated last, without obstructing our main algorithm.

While trying to integrate our datasets and insert them into the appropriate tables, we ran into a number of issues due to how we were initially trying to insert our data into the tables. Our main issue is that we were trying to insert data, by taking each column of a dataframe, for example, and then creating an entire dataframe for data from that specific column. We would essentially try to iterate through all of the column dataframes for one dataset, adding one value at a time to the table rather than one row at a time, which led to a number of DataErrors, InsertErrors, and ProgrammingErrors from Python. As a result of this, we needed to switch gears with regards to our database design. This is what led to the current/final design of our database.

With our revised database, we then started to insert each dataframe into their respective tables, making sure that the record\_id column from the Weather table would remain consistent throughout the Countries and Grains table. After all of the tables had their data inserted, MySQL was used to generate a csv file of the integrated database using SELECT and JOIN statements. Using that integrated csv file, along with the individual csv files from the individual datasets, we can now compare the rows between the integrated data, and the individual datasets, to make sure that the data for each country and year in the integrated database matches the data from the individual datasets.

# Data Preprocessing (Integrated)

With all of the different datasets now integrated into one relational database, we did some additional preprocessing on our integrated data in order to prepare for using some of the different clustering models. First, due to the large number of attributes using different units, we had the data scaled to a range of 0 to 1, which helps explain visualizations. This also ensures that clustering can be used properly with all of the selected attributes. Now our data was ready to be clustered, but we took an additional step, and applied Principal Component Analysis in order to reduce the number of features in our dataset, hoping that it would lead to better clustering overall. We also took the categorical attribute “Commodity\_Description”, which detailed what grain the agricultural data was referring to, and did one-hot encoding, to take those values and make them attributes.

# Clustering

Once integration of our database had been finished, we moved on to using different clustering models/methods in order to determine which clustering method would end up creating the best quality clusters that represented our data. The methods we used were KMeans, KModes at one point, which ended up not being used anymore since we stuck to clustering mainly numerical data, DBSCAN, HDBSCAN, OPTICS, and Agglomerative Clustering, with different iterations based on the linkage methods. We had gotten all of our models into one python notebook, and made observations about which clustering methods had resulted in clusters that weren’t well defined, and wouldn’t be useful for drawing conclusions/observations from the data/clusters. We had settled on DBSCAN providing the highest quality of clusters based on how they were more well defined, compared to the other models. Once we had decided on the clustering method, we had to decide on the number of clusters that would be best. We used silhouette score as a heuristic measure for the quality of our clusters, as well as noting observations about the differences and similarities between the clusters.

# Decision Trees

We had gone back and forth with regards to what clusters we were going to be using as we reached the end of the deadline for our project. We decided to stick with clustering without PCA, through the DBSCAN model, which led to the creation of 16 clusters, with a silhouette score of .75. This model produced one of the best models with regard to noise data, silhouette score, and observations of the clusters. Now that we have our clusters, we moved onto classification, by using our clusters as the target variable, in order to see what attributes might’ve had the biggest impact on how the clusters of our data were formed. We used the default decision tree classifier from SKLearn in order to analyze how the data might’ve been clustered, and why. We used decision tree regressors in order to try and determine what attributes have the highest impact when it comes to trying to predict a target variable/attribute from our dataset, such as Production of grains, for example. We also tried using Random Forests for both classifiers and regressors, as a way of improving model performance. We ended up not using the random forests for classifiers, since the main purpose of the classifier is to try and emulate our cluster results as best as possible.

# Regression

We used different regression models on the different clusters, along with the correlation heatmap, to see what useful patterns we could observe from the different clusters. Because of the nature of the correlation heatmap, linear regression was a major part of this process, trying to explain the relationship between some of the attributes, such as Production and Imports of certain grains, the GDP of a country and its grain production. We attempted another model like OLS, delivering very similar results to SKLearn’s Linear Regression model. We did also use regular multiple linear regression by using RFE feature selection, which resulted in very high r-squared values, due to the nature of the r-squared values when it comes to multiple independent variables. Overall, it was difficult trying to assess the clusters using multiple linear regression because if we had more than 2 independent variables, we couldn’t necessarily visualize that trend on a plot, to see what the shape of the cluster looked like with those variables.

# Results

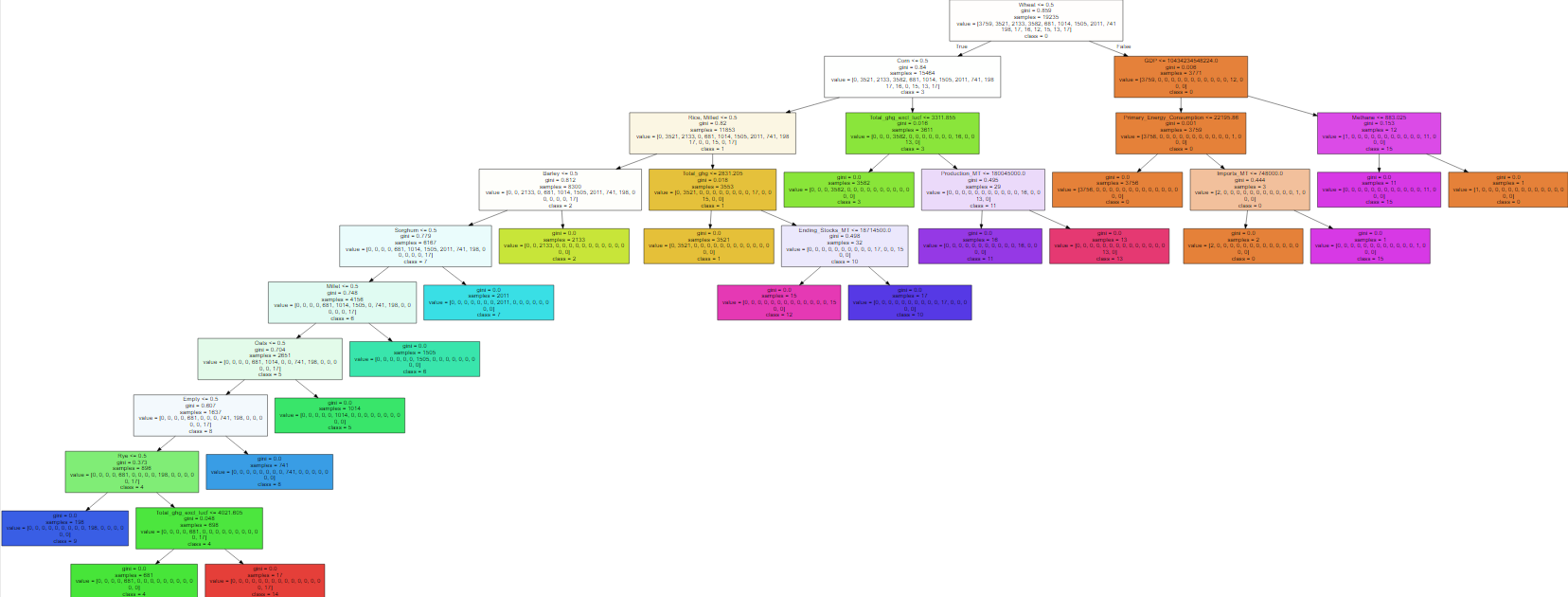
## Clustering

With our finalized DBSCAN clusters, using the full 25 attributes from the dataset, along with some finetuning, we got a total of 16 clusters from the dataset. We had attempted clustering beforehand, without the one-hot encoded attributes, which led to the creation of one mega cluster, one regular cluster, and the rest of the clusters holding very little data, which made it difficult to get any useful results. The mega cluster consisted around 80% of the original dataset, which doesn’t help with regards to analysis. A lot of the mini clusters, on the other hand, consisted of a very small minority of the dataset, which makes it difficult to make overall observations about the entire dataset.

With the one-hot encoded clusters, however, the clusters ended up being much more balanced, having some clusters hold 1000, 2000, 3000 and 5000 values, and a chunk of the clusters holding a few data points. Around half of the clusters were formed based on the grain type, and the rest took subsets of specific grain data from individual countries. Due to the nature of our clusters now, we can analyze the differences between the grain types, and how climate change could affect not simply grain production, but certain types of grains as well, and then make conclusions about how it would affect the global grain supply chain.

## Decision Trees

Our default decision tree used a 70-30% train/test split, along with the cluster labels obtained from the DBSCAN clustering method, in order to determine what attributes would/could be used as the split when determining different labels/clusters. In general, one hot encoding isn’t recommended with decision trees, as it can reduce model accuracy. However, considering that we aren’t using decision trees for actual classification purposes, and the fact that our labels aren’t true labels, it serves as a good visualization to see some of the deciding factors behind the prediction of the clusters. As expected, 9 of the splitting attributes were based on the one-hot encoded attributes, so it would determine a label based on whether the data was pertaining to a specific grain. For example, if the splitting attribute was Wheat <= 0.5, if this was true, then it would apply to all data that doesn’t include Wheats, if it was false, then it would separate wheat data from the entire dataset. What’s useful about having the clusters/labels determined by the grain type, is that it allows us to see how the different grains differ between each other when it comes to production, exports and imports, and how the grain overall impacts the global food cycle.



## 

## Regression

Part of the Regression process included the usage of Decision Tree Regressors/Random Forest Regressors, along with other regression models, such as SKLearn’s Linear Regression model, and Ordinary Least Squares. With the clusters that didn’t use One-Hot encoding, using these models on the clusters yielded mixed results when trying to look at the relation between Grain Production and Grain exports, where some regression models had a high coefficient of determination on some clusters, and other clusters simply had poor R-squared scores. Considering the steps that were taken in preprocessing, such as the imputation of missing values through KNN, it makes sense that some of these clusters might yield poor R-squared scores compared to others.

With the clusters that were formed because of one-hot encoded attributes, we can look at regression analysis in a different way because of how our clusters are structured. For example, Cluster 3 consists of corn data from 119 countries, and with regression analysis, we can see that there is a positive correlation between the production of corn, and the export of corn.

Cluster 8 is the only cluster that differs with analysis, because this cluster consists of the data from countries that didn’t report any agricultural data, so we cannot make any concrete conclusions/observations between grains and climate change. However, we can look at how these countries have contributed to climate change through their co2 emissions, and whether they are importers or exporters of co2.

# Challenges

One of the major challenges of our project was the lack of communication that we had in the beginning of the project. While we did maintain communication through chat, we didn’t really meet up in-person or through calls, which made real-time communication pretty much impossible, and as a result, hindered our progress. Along with a not so clear understanding of what we were exactly trying to do with our project, such as doing regression analysis on the dataset entirely, without clustering, or applying any other models, set us back a bit. Code-wise, we ran into some issues trying to implement models, which ended up taking more time than we anticipated, along with major changes that we have made to our database design until we had reached our final iteration of the database. Through learning from our mistakes, and starting to communicate more thoroughly as a team, we started making progress at a better pace, and whenever we ran into issues, we were able to resolve them much quicker than before.

# Conclusions

As the global temperature continues to rise as a result of rising CO2 emissions, grain output will slowly increase, due to the increased CO2 concentration in the atmosphere, and especially due to the advances in technology. This increase in temperature & CO2 will at first increase the outputs for certain grains, but others will end up decreasing due to the changes in environmental conditions, making the land no longer suitable for growing certain crops.

Rice as a grain remains supplied to the world thanks to countries such as India, Pakistan, Thailand, Vietnam, and The United States. With China as the biggest importer of rice, at a population of 1.4 billion people, and still rising, along with other countries such as Indonesia, The Philippines, Bangladesh, and Japan, they remain heavily dependent on a steady supply of rice. In the last 12 years, 99 countries have imported over 330 million MT of rice, and 44 countries have exported 390 million MT of rice, which shows how significant of a grain it is in the global grain supply.

One pattern of the rice data was in contrast with our hypothesis, was that we believed countries with a 0 milling rate, countries that aren’t capable of milling threshed rice, would be exporting a lot of that threshed rice to other countries that would be able to mill it, increasing overall rice production globally. However this wasn’t the case, and simply, countries that do not have milling capabilities, do not produce any quantities of threshed rice. They only really export either rice that they have stockpiled, or broken rice from the fields. This puts further strain and responsibility on the countries that are producing and milling rice, since they are going to be the only noticeable contributors to the rice supply.

Corn is a major staple for different foods throughout the world, with China and the United States as the top producers and exporters of Corn. The biggest importers of Corn are Egypt, Japan, Mexico, Taiwan, and Russia at one point. In terms of the global supply chain, however, in the last 12 years, 105 countries have imported 900 million MT of corn all around the world, and 75 countries have exported over a billion MT of corn. This shows how Corn is one of the most dominant crops in the global supply chain, and overall a crop that at least over a 100 countries are reliant on.

Other grains like Oats and Sorghum do not have as significant of an impact as Wheat, Corn, and Rice, but still has a considerable spot in the supply chain, consisting of over 20 million metric tons both imported and exported for Oats, and over 70 million metric tons imported and 80 million metric tons exported for Sorghum. Rye by far takes the smallest portion of the global grain supply chain, having less than 10 million metric tons imported and exported, and Millet is a grain that isn’t imported or exported, mainly used for sustained farming, and in-country use.

Wheat is yet another major staple grain, heavily exported by Russia, The United States, Canada, Australia, and Ukraine. Considering how dominant it is in the global grain supply, there is a lot of pressure on these countries to make sure that they can provide a constant supply of wheat to countries that are importing it, such as India, Egypt, China, and Brazil. This pressure is further realized when noting that from 2000 to 2012, 1.3 billion MT of wheat was exported and imported.

Another important thing to note is that the global supply chain shifts over time, where countries that may have been major importers of certain grains, may become exporters of those grains, like Russia, who used to be a major importer of wheat, and is now becoming one of the leading producers and exporters.

The global grain supply chain is constantly evolving, with countries taking new roles as importers or exporters/producers of certain grains. Especially with the human population continuing to grow over time, supply and demand is going to increase along with it, and at the current rate that the world is heading towards, with increasing CO2 emissions and temperatures, the global grain supply chain is going to continue being put under increasing pressure, specifically on the countries that are responsible for the majority of the grain supply, due to the stark contrast that exists between these countries, and those that aren’t technologically advanced enough, or financially well off enough in order to assist in grain production. Climate change has the potential to change the way we grow our grains, where we need to grow our grains, and whether we’ll even be able to grow our grains the same, as the world’s climate changes more.

Overall, looking at the early initial hypothesis, it seems like we were correct with regards to how countries that are developing, not as technologically advanced, or not very strong financially, in terms of GDP, are heavily reliant on the major producers of grains. In 2012, The United States had the highest GDP in the world at around 15.84 trillion dollars. Only 8 other countries had at least a tenth of that in the same year, and 20 countries in the world had at least a twentieth of that. Meanwhile, 87 countries of the 136 had less than a hundredth of the United State’s GDP in that year. Considering the strong positive relationship between a country’s GDP and grain production, this helps explain how it’s possible for countries with high GDPs to be the cornerstone of grain production, leaving many countries with little wealth in comparison to rely on these countries. As a result, they become reliant on the countries that are top producers with these different grains, and puts them at great risk of food shortages, if climate change gets to the point where the grain supply is no longer as abundant as it used to be.

# Contributions

## Jason Ortiz

* Collected various resources to try and assist Daniel in catching up with Python and the libraries we planned on using throughout the project. (Python basics and Python for Data Science).
* Found a weather dataset that we had to scrap due to not going far back enough, only until 2003. As such we decided to go with the Berkley dataset.
* After finalizing our datasets we started the EDA process, analyzing and visualizing the weather and grains datasets.
* After researching for models for us to later use Rhodiam and myself started preprocessing to get the data in the correct format for insertion into the MySQL relational database we planned to create. The database would help us both query and integrate the tables through joining with the shared features in all three datasets, being Country & Year. I had realized that since all three datasets had Country & Year, later 4 when we included a dataset containing Longitude and Latitude of the countries we were analyzing.
* I started to create the ERD & MySQL database for integrating the datasets. After my initial design I showed the group where Rhodiam had suggestions that were then implemented and iterated on to make sure our data would be consistent and no problems would arise when inserting the data. After finalizing the Database design I started to implement it in the MySQL workbench and testing the insertion of dummy data.
* After completing the creation of the database I started to research and test methods for connecting our MySQL database to Python in order to use Python to insert and Query the data from the integrated database. This was completed through the use of the mysql-connector-python library. After initial tests I created a Notebook showing how to use the library to showcase to the team in our following meeting. This allowed us to successfully join our datasets together and query them in Python, further allowing us to save useful queries in a .CSV file for future analyses and communications.
* I moved onto researching and implementing DBSCAN, HDBSCAN, & OPTICS clustering models. I created functions to streamline the testing of the various clustering models along with their respective visualizations (scatter plots with points colored according to cluster labels), along with printing the number of clusters formed and the Silhouette Score.
* I then worked with Rhodiam to make similar functions for KMeans and the various versions of Agglomerative Hierarchical Clustering algorithms (SINGLE/DOUBLE Link, Average & Ward). This made testing and comparing models faster and more efficient. We would then select the number of clusters we thought was appropriate and tested all Clustering algorithms for that number of clusters, taking into consideration the Silhouette Scores of each model with that respective number of clusters. This led us to decide on the DBSCAN algorithm for the model we would use for further analysis.
* I started to work on learning and applying the Streamlit web app framework for our project, with the goal being to give us something to showcase our project. Daniel was to help with this but couldn’t get Streamlit working on his PC so it was left to me. I struggled initially but was able to get a basic web application created as I had only learned of Streamlit slightly more than a week before the Presentation date. The goal was to create an application that would showcase the process we went through while also allowing the user to interact with the model parameters to see how drastic the difference in results would be (eg. EDA through Pandas\_Profiling. eg. visualizing changed in clustering results using PCA for visualization of colored/labeled scatterplots using widgets).
* Continuously checked back with Rhodiam as he worked on the regression analysis of the final clusters created from DBSCAN, learning from the findings he made as well as offering suggestions for moving forward.
* Worked on getting the slides and report prepared and finished with Rhodiam.

## Daniel Cohen

* Did initial preprocessing to the Grains\_PSD dataset. Lack of experience led to getting assistance as a team effort to finish.
* Researched various regression models used in similar projects to identify what could be applied to this project.
* Worked with Jason on the setup of ERD.
* Attempted to utilize association rule mining, but dropped it due to difficulty binning the dataframe.
* Did initial mapping of latitude and longitude on world map as the groundwork for visualization, which was dropped once Streamlit app was discovered and utilized for cleaner and easier visualization.

## Rhodiam Arango

* Found one of the main datasets - Our World In CO2 Emissions, and applied imputation to missing values for the CO2 dataset. Changed structure of the Grains\_PSD dataset in order to integrate datasets.
* Visualizations for CO2 Dataset
* Looked at similar use cases to get a better understanding of the project
* Worked with Jason throughout the different Database designs, and integration methods for our dataset.
* Used clustering models such as KMeans and Agglomerative Clustering
* Used decision trees to try and visualize what attributes might’ve been the main cause for the clusters to be formed, what attributes were the main factors.
* Used regression analysis on the different clusters, and made observations based off those clusters in order to be able to draw conclusions.

# References

# Dataset Sources

<https://github.com/owid/co2-data>

[co2-data/owid-co2-codebook.csv at master · owid/co2-data (github.com)](https://github.com/owid/co2-data/blob/master/owid-co2-codebook.csv) - Codebook for dataset

<https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data>

<https://apps.fas.usda.gov/psdonline/downloads/psd_alldata_csv.zip> - Full Agriculture

<https://apps.fas.usda.gov/psdonline/downloads/psd_grains_pulses_csv.zip> - Grains Agriculture

[countries.csv | Dataset Publishing Language | Google Developers](https://developers.google.com/public-data/docs/canonical/countries_csv) - Latitude/Longitude Data

# 

## Research Sources

Eze, U. (2020, September 6). *Predicting weather temperature change using machine learning models*. Medium. Retrieved May 10, 2022, from https://medium.com/swlh/predicting-weather-temperature-change-using-machine-learning-models-4f98c8983d08

Food and Agriculture Organization of the United Nations. (n.d.). Macroeconomy. Retrieved from https://www.fao.org/3/i2490e/i2490e01c.pdf

Gray, E. (2021, December 2). *Global climate change impact on crops expected within 10 years, NASA study finds – climate change: Vital signs of the planet*. NASA. Retrieved May 10, 2022, from https://climate.nasa.gov/news/3124/global-climate-change-impact-on-crops-expected-within-10-years-nasa-study-finds/

Hille, K. (2016, May 3). *Rising carbon dioxide levels will help and hurt crops*. NASA. Retrieved May 10, 2022, from https://www.nasa.gov/feature/goddard/2016/nasa-study-rising-carbon-dioxide-levels-will-help-and-hurt-crops

Kumar, V. (n.d.). *What is K-Means Algorithm and how it works*. Towards Machine Learning. Retrieved May 10, 2022, from https://towardsmachinelearning.org/k-means/?msclkid=683a1a1ace7d11ec8481b4641adc09a4

KumarI, A. (2022, March 10). *Machine learning use cases for climate change*. Data Analytics. Retrieved May 10, 2022, from https://vitalflux.com/machine-learning-use-cases-climate-change/

Ritchie, H., Roser, M., & Rosado, P. (2020, May 11). *CO2 emissions*. Our World in Data. Retrieved May 10, 2022, from https://ourworldindata.org/co2-emissions

USDA. (n.d.). PSD Online. Retrieved May 10, 2022, from https://apps.fas.usda.gov/psdonline/