Rhodiam Arango

CSCI 455 M01

05-15-2022

My Contributions

My contributions with this project once we had finally picked a topic/area of interest, started with looking for potential datasets/research documents that we could use for our project. One of the datasets I found is a major focus of our project, which is the Our World In Data – CO2 Emissions dataset. I also found other sources that were considered but either weren’t used or ended up being phased out as we made progress throughout the project. I also found several sources relating to machine learning and climate change, to get an idea of the similar use cases that had already existed, to get a better understanding of how we can approach our project.

Afterwards, I mainly worked on the literature review section of our project report, which ended up needing a major revision after feedback from you, Dr. Gu. Once literature review was completed, I took a brief look at the different datasets that we had collected and summarized each dataset, along with screenshots, and explanations as to why these datasets were beneficial/important for our project. With regards to data visualization, I provided visualizations for the PSD Grains dataset, the Berkeley Weather dataset, and the OWID-CO2 dataset.

Text

Description automatically generated

Graphical user interface

Description automatically generated

Next came preprocessing, where I worked specifically on the CO2 dataset, considering it was the largest dataset we had, with 60 attributes, a lot of analysis needed to be done to see which attributes ended up being derivative. I used Correlation heatmaps, along with variance threshold/variance to see which features would end up being removed from the dataset. Variance ended up removing none of the features, since all the features had a high variance measurement.

Text

Description automatically generated

The correlation heatmaps showed that quite a few attributes were correlating with each other, and a huge reason for this was because many of the CO2 dataset’s attributes were derived from a section of the attributes in the dataset. As a result, this led to a large majority of the attributes from being dropped from the dataset, resulting in a final total of 18 attributes.

Graphical user interface, text, chat or text message

Description automatically generated

After I finished the CO2 dataset’s preprocessing, we all worked together to make sure each of our assigned datasets were ready to be integrated together into one database.



Graphical user interface, text, website

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

The CO2 dataset didn’t only consist of countries, but groups of countries, such as the European Union, described as EU-27 or EU-28, for instance. This wasn’t reflected in the Berkeley and PSD datasets, and the iso\_code column was left as a null value, which was going to be important for our database creation. Also considering that these tuples involving the EU, as an example, would be reflected as a sum of the data from the countries that are in those organizations, this data was redundant, and could be reacquired by gathering the countries that are part of the EU in the final dataset, therefore these tuples were dropped.

Graphical user interface

Description automatically generated

Thanks to Jason, we had gathered a list of all the countries that had data in all 3 datasets, so that we wouldn’t be left with any countries missing key data in our final dataset. I used the list obtained to filter the dataset to only have tuples for countries that were in this list.

Graphical user interface, text

Description automatically generated

As a result, the CO2 dataset now had been filtered accordingly, ready for database integration, only containing the countries that all three datasets had in common, and a specific year range that was found in common for all 3 datasets, which resulted in being from 1960 – 2013. It was because of this reason, that the Monthly CLIMAT dataset ended up being phased out, as it only had data starting from 2003, and that would’ve been a much drastic reduction in the quality and quantity of our data. As a result, we ended up using the Berkeley dataset.

One issue that we had run into regarding the PSD Grains dataset was that in general, the dataset was a bit cluttered, due to the number of tuples that were dedicated one year, for one country. For instance, for the year 1960 in the United States, there would be more than 5 tuples dedicated to that year, due to the ‘attribute\_description’ column. As a result of this, we decided to try and change the shape of the dataset, so that each unique value in that column, would become their own column, removing the need for the ‘attribute\_description” column, and having our data be more organized. I managed to find a way to fix this issue, and this made our progress toward the integrated dataset much simpler.

A screenshot of a computer

Description automatically generated with medium confidence

Afterwards, I assisted Jason with regards to the database and ERD design by first gathering all the final attributes together that was guaranteed to be part of the integrated database. This is when we all collaborated for the ERD and database design. During this phase, I noticed that due to our filtering results, the UK no longer had any agricultural data, because according to the PSD Online FAQ, the UK only had reported grains data from 2016 to 2021, due to it splitting from the EU in 2020. Of course, this means that the UK’s grains data would have been dropped from the grains dataset. As a result, the UK would be a country that had only CO2 and temperature data. This revelation impacted the design of our ERD and database, since now a 0 to many relationship needed to be established between a country, and the grains data, to make sure that no issues popped up if it so happens that a country did not report any agricultural data for a specific year.

We had gone through many different database design iterations, due to new discoveries we made about our data, along with new datasets that we implemented into our database, such as the dataset which contained every country’s latitude and longitude. I had to go back and fix an issue regarding the CO2 database, where I had originally imputed all of the missing values with 0, which greatly affected clustering and overall visualization of the dataset. As a result, imputation was now done on the CO2 dataset using KNN, as a more reliable way of imputing potential values based on nearby values, without assuming all of those values are equal to the mean or mode. Jason and I worked on the different implementations of our database, until we concluded on our final database version.

With our integrated dataset, I started testing out KMeans and hierarchical clustering on our dataset, seeing how clusters were formed, and what could be the potentially best clustering result from both clustering methods. Jason and I collaborated on comparing the results we got from the different clusters, and how they looked like. As a result, we stuck with the initial DBSCAN clusters that Jason had gotten, even after comparing with agglomerative clustering, optics, and HDBSCAN.

After some analysis into the DBSCAN clusters, we took a step back and thought about how to incorporate the grain type into our clustering, since the type of grain had a huge impact on the values that some of the attributes had. This is when I applied one-hot encoding to the commodity type attribute, making each potential commodity attribute its own attribute. Afterwards, we did cluster yet again, but for the encoded version of our integrated data, and we ended up with better results across the board, in terms of more balanced clusters on my side, along with overall higher silhouette scores. In the end, we ended up going with 16 clusters from DBSCAN.   
 Once we had the clusters from DBSCAN, I started applying the idea of the CART algorithm, using decision trees to try and interpret what attributes were the major factors behind how some of the clusters were formed. I also did regression analysis on the clusters using different models, along with taking note of the different observations I made about each individual cluster, to extract new information from those clusters, and the data overall.

Finally, we had started discussing some of the conclusions that we could draw from our cluster and regression results and tied it back to our hypothesis. I continued to work with Jason to finish out the project report and the slides and assisted him with the Streamlit demo.