Jason Ortiz

CSCI 455 - M01

I initially was going to try and keep anonymity when talking about group mates however I realized it wouldn’t do anyone any good. I tried to mention instances of poor communication when applicable, I can remove all of those upon request, however I wanted to include them for some context for you since I don’t feel comfortable talking about group members like this during the meetings. I will try to add pictures/snapshots when possible.

# **Dataset Searching & Initial (Lack of) Planning**

As we formed a group later than intended we did not have a clear understanding of a goal that we wanted to accomplish, and as such we just had a general idea of wanting to focus on Climate Change, more specifically Global warming. Later we refined this to analyzing and possibly predicting how Climate Change would affect certain countries more than others. This led us to partition the searching of datasets between the three of us, with myself in charge of looking for a Weather dataset, Rhodiam looked for co2 datasets, and Daniel was just expected to use the resources supplied to get more familiar with the Python language and basic machine learning data mining concepts, reinforcing what he is learning in his Intro to Artificial Intelligence. I eventually narrowed it down to two potential datasets, the CLIMAT Weather dataset and the Berkeley Earth Climate dataset. At first we chose the CLIMAT Weather dataset for us to focus on as it contained potentially useful attributes such as precipitation and location. After finding the CLIMAT weather dataset, I also had to search for the agricultural dataset that we would use, since the group mate responsible for researching and finding an agricultural dataset did not do so, and as such the responsibility fell unto me to make sure we did not continue to keep falling behind. Ultimately after confirming with the team the PSD Agricultural dataset was chosen, or more specifically the Grains subset of the PSD ALL DATA Agricultural dataset, as grains have been the most important food group in human society for most of written history and even before. This was all during a period of difficulty in team communication and coordination, as even through the use of Discord and its pinging feature Daniel would either stay silent or respond with a very simple answer that made getting any concrete group direction decisions difficult to make. I have attempted to share the leadership role with Rhodiam, however it generally falls to me to try and get the team to come together as a unit to get a substantial amount of work done, however I would like to mention that Rhodiam has assisted with any questions and has done any tasks asked of his assistance with. I have tried to make myself available to assist with any potential issues from anyone to ensure nobody gets stuck, however that requires the person who is stuck to be forward about it.

# **Data Cleaning & Preprocessing**

Throughout the datasets cleaning and preprocessing steps, I was once again in charge of handling the CLIMAT dataset, which at the time it was still unknown that the dataset didn’t meet our not yet definitively chosen criteria. Since we did not have the concept of filter years or filter countries in mind yet, since we were behind from day 1 and were trying to rush into catching up to where we were supposed to be, the preprocessing steps were simple in nature. The cleaning involved removing null values, since when dealing with a continuous value like temperature it is not as simple as replacing all null values with a default value. Then correlation analysis was done to find the correlation between the attributes in the dataset to help back up the decisions to emphasize or deemphasize certain attributes. After doing these steps some efforts were made to visualize the data however my attention was directed towards attempting to assist Daniel with the Agriculture dataset since he was having problems analyzing . It was during this initial preprocessing phase that certain attributes in the agriculture dataset were not in the proper structure that would allow for easy data retrieval nor integration and it was taken note of for future use. The problem was the ‘Attribute\_ID’/’Attribute\_Description’ features and their corresponding attributes weren’t separated into separate attributes but instead were grouped into a multi-valued attribute that also changed other attributes’ values depending on those and it was a bit confusing at first. The issue was documented and later passed onto Rhodiam who was able to figure out how to reshape the DataFrame to better represent the data stored inside the Grains dataset.

# **Models & Algorithms**

While I was working on both the weather and agricultural datasets, Rhodiam was working on the co2 dataset (which had the most attributes), and Daniel was in charge of researching models for two weeks while using the continuing to use the resources supplied to get more familiar with the Python language and basic machine learning data mining concepts. When it came time for our next meeting we were again informed that Daniel had issues with researching models due to his unfamiliarity with machine learning or data mining models and their concepts. Afterwards we switched and I suggested Rhodiam and Daniel work on cleaning and preprocessing the remaining datasets together as a pair, with the idea being any questions Daniel had along the way he could ask Rhodiam who was working on a similar task. Unfortunately what ended up happening was they partitioned the work between them and so no issues were known about on Daniel’s side until it was too late for myself or Rhodiam to assist with before our next meeting. Regardless, during that time I took over the model research and throughout this research I asked the group to start thinking about what it was we actually wanted as an output from the models we would be working on later down the road. We came to the agreement that we wanted to attempt to create predictive models to predict values of attributes such as crop production or temperature based on other descriptive attributes such as CO2 emissions, etc.. This allowed me to delve deeper into a type of model, and as such find several including the simple linear regression, Least Squares Regression,Ordinary Least Squares, and finally looked more into the feasible generalized least square (FGLS) model found by Daniel the week prior. These, along with multilinear & polynomial linear regression are the models we will be focusing on for making predictions. During my research on potential modes we could use I noticed that “Time Series Analysis” and it’s respective models are very commonly applied and potentially very accurate models may be applicable to our datasets due to the common attribute of Time in the form of years shared between them. This ultimately led me to do research on Time Series analysis and it’s respective models and algorithms, however ultimately it was realized/decided by the group that Time Series models were not applicable to our data due to the necessary attribute of data that only one of our datasets had, Berkeley Earth. This was slightly frustrating as I thought I found potential models for prediction, however I realized that Descriptive models were also important to keep in mind for more models research. Due to time constraints and multiple issues arising on both the PSD Grains and the OWID CO2 datasets had by Rhodiam & Daniel the group decided to virtually meet to discuss the issues and potential solutions. Seeing the issues that were arising, I recommended that we should take a step back and look at our project and data as a whole, then go back to examine, extract, and transform the data into the (at the time decided) direction we as a team wanted to move towards. This allowed us to break down our data into their Entities, Attributes, and their relationships mentioned in the next section.

**Data Modeling**

Initially there was no clear plan as to what entities were going to ultimately be used for our database, as initially my idea of putting the data into a database (or a data warehouse depending on the nature of our data) seemed to be getting outvoted in favor of keeping the data in CSV files. I foresaw issues down the line and over spring break, through making the team take a step back and fully understand the data that we collected we were able to agree on the direction to take. During this Spring Break the communication between team members improved, and it allowed us to work better as a team than previous weeks.

Since we had not put the initial proper planning when choosing the Entities and Relationships that would be useful/important to use we did not have a clear picture of the data we were working with. This caused problems down the line as only after various data cleaning and preprocessing steps did we realize that in order to integrate the data together they had to be put together under Entity-Relationship Diagram(ERD) modeling . As such we broke down all of the datasets into their entities, attributes, and their relationships between each other and 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 in the specified year.

This forced us to consider the time range and countries in the datasets that we had chosen, which is where our first big issue arose, the original CLIMAT Dataset chosen for weather, more specifically temperature data, did not include years before 2003. This recent beginning year for our timespan was too recent for our likings and it forced us to drop way more data than we were comfortable doing for fears of it affecting our model accuracy down the line. After coming to this realization it was clear that the next step, for me at least, would be to look for another weather dataset that fit the current idea we had for the ERD. After some research I found the weather dataset that we chose to use as the final weather dataset, the Berkeley Earth Earth Surface Temperature, which was a candidate dataset from the beginning of our dataset research that fit our timespan with years to spare. With this dataset in hand I made a simple program to extract the unique countries in each dataset, put each of them into respective lists, and compare the lists to retrieve only the unique countries that were common between all three datasets. This was to ensure data consistency down the line.   
 From this point I was ready to start planning and designing the ERD iteratively, as I did for my DataBase management course, with each attempt getting closer to 3NF until the database architecture was at 3NF to ensure that no Anomalies would occur when attempting to insert data into the database. I would communicate with my team with any potential ideas regarding the architecture of the database. This includes normalization steps, as well as initial ideas as to how we would integrate the data. During this stage of the project each team member assisted in their own way. Rhodiam assisted by first listing all attributes we decided to use, then he helped identify and remove any derived attributes (since they can be found at the time of the Query). Most importantly he discovered that the country ‘United Kingdom’ did not have any crops grown in our time frame and was as such grain information on from the grains dataset was missing however the ‘United Kingdom’ was a country that we had other information on which meant that the relationship between Country & Grains changed from a 1-N to a 0-N relationship, which changed the design of the database and would have gone unnoticed if not for that discovery. Daniel used the attributes list that Rhodiam had supplied and wrote out which Attributes corresponded to which Entities to help me with my step of making ERD designs and normalizing those designs.

After planning/designing I created the current database design, and a second one that is very similar however approaches a problem differently and as such we wanted to make sure our design would be functional before attempting to migrate all of our data into MySQL.

I have also gone through the process of creating the MySQL database and tables, and now the next step is to further test the database by inserting a few objects & their attributes values. After this testing has shown the results we plan to use the MySQL library within Python to access the existing MySQL databases and insert the data from the CSV files into the database. It was decided that this would allow us to create the database and its tables in its proper IDE, then use Python, and its libraries such as Pandas, to insert the data. We decided this would be better in the long run as it would allow us to store & query our data in SQL.

**As of 04/03/2022**

We went into the week with the idea of fixing the Database and inserting the data into the Database, successfully integrating the three datasets.

During last week's meeting, Wednesday 03/31/2022, Rhodiam mentioned that he found a flaw in the design for our database and proposed a solution to an unforeseen problem. After the meeting we spoke further about the potential solution and I started to use different test cases to test the new design. This includes a very small subset of Countries, only 3, and a small subset of their respective Grain data into the Countries & Grains tables respectively. This was followed by attempting to insert more than 1 years worth of data for the chosen Countries, as well as attempting to insert more than 1 type of Grain per year for the chosen countries. This was to ensure the design and architecture chosen for our database would accurately represent the entities, their attributes, and their relationships. It also ensured that when attempting to apply the chosen models we could query data in a way that is efficient, both in terms of data quality and integrity, but also more time efficient.

Due to working on other projects for other classes, as well as studying for midterms the days immediately after last week's meeting, Wednesday 03/31/2022, there was little communication between the group until that Friday. On Friday I reached out to the group members letting them know I had tested the data, as mentioned on Wednesday however it seemed to go back to how things were before Spring break and we couldn’t get all three team members together. It was particularly difficult to get hold of Daniel, who would rarely say anything in the chat, and only responded when directly shouted out at using discords @message feature, even then he would choose what to respond to with simple responses.

After confirming that the Database worked as intended, by using JOINs to connect the required data from both tables, successfully integrating the data in the intended way.   
  
After this I started to learn the syntax and functions of the library that would be used to integrate our Python Notebooks to the MySQL database we created. While following along with tutorials and reading the documentation I created a tutorial on how to connect to the MySQL database we created through Python, and shared it with the group since they had not looked into the library we decided to use. After assisting Rhodiam with issues using the tutorial created due to using Google Colab to code in python, and due to its cloud nature could not connect to the MySQL on the local machine. I recommended him to download JupyterLab as I had previously been using Jupyter Notebooks and recently had upgraded to using JupyterLab, which allowed me to connect to the MySQL Server/Database on my local machine. This solution worked for Rhodiam, however when I asked Daniel if he had the same problem while trying to assist Rhodiam he said he hadn’t had a chance to look at the content that was worked on previously and he reassured me that he would attempt later that night, however it seems that he had not. The lack of communication has made planning any larger tasks more difficult since it is mainly Rhodiam and myself communicating ideas with each other. I have attempted to assure Daniel that if he has trouble with anything he can come to us for help, however if it is difficult to get a hold of we cannot assist him as we don’t know what he needs.

While thinking about how we would insert the records of data from the CSV files we had into the MySQL database we created I realized we would need to create an algorithm to automatically fill in the data for us, as inserting the vast amount of data by hand was absolutely not an option like in previous projects. When brainstorming how to tackle the problem I brought up the idea of writing pseudocode for the algorithms, to help plan the process of insertion. To start I thought it would be better to split the algorithms apart, as the Countries and Grains tables will have different amounts of rows and attributes, and if combined into a single algorithm would have made it too difficult to understand, potentially causing issues during implementation.

Further thinking about how to solve the problem, I first thought about it with the Brute Force technique in mind, which would be to loop through the data iteratively with potentially nested loops to insert each row of data from each required variable iteratively. This process helped me figure out what the problem really was, by picturing the problem as a black box with an input and output, and it led to me thinking of the problem in a different way. By having a better understanding of the problem, I came to the realization that the Decrease-By-One & Conquer technique could potentially be used as the dominant technique to be used by the group. Because of the problem’s nature, doing the same tasks over and over for multiple inserts, I thought that it might work better than an attempt at brute forcing it with loops and potentially creating an algorithm that takes way too long. After further research I found that it could be a technique used in conjunction with the libraries being used. With this in mind I started to type out a pseudocode for a recursive insertion algorithm. Recursion should allow us to more elegantly solve the problem, allowing the solution to be more easily explainable and reproducible.

**As of 04/05/2022**

Upon attempting to draw up some idea for pseudocode for our insertion algorithm(s) it was realized that we were going to run into trouble due to the fact that it was earlier falsely assumed that both Berkley Weather and OWID CO2 datasets had the same amount of rows of data. I mistakenly thought that they both had data for all countries going back to 1960 and did future calculations based on the idea that both datasets relied on the same two attributes of Country & Year and have the same number of rows (136 countries \* 53 Years = 7208 rows). Berkeley Weather dataset has the full amount of data however the same could not be said for the OWID CO2 dataset, and due to this it set us back to the drawing board for database design, this time with the knowledge of last time’s mistakes. We eventually found the mistake however it was a setback that could have been avoided if I was more careful and communicated better my wrong assumptions about the data. Communication has been difficult within the group, making it difficult to bounce ideas off each other to better understand and solve the problems at hand, as well as planning out the week's goals and effectively partitioning work. We have decided to break the data into three main tables, one for each database, and integrate them together using a record table which stores the different country/year pairs. Rhodiam assisted with making edits to the tables and sorting the data for insertion while I did research on how to use python variables, such as lists, along with the MySQL Python Connector library we are using, to see if that changed how we went about planning the algorithm. It did indeed and we are continuing to make adjustments to our course of action. As of today I reached out to Daniel again and he said he was about to look at the stuff Rhodiam and myself had been working on during the weekend to catch up to what we did. He asked how he could add to what we have done, however as he had not yet looked at and understood the steps prior, he had trouble understanding the problems that we were trying to solve. I asked him to just catch up for the night and to please better communicate and work harder next week as it was the night before the meeting and he was behind.

**As of 04/13/2022**

Rhodiam believes he found a potential new design that allows us to successfully insert the data into the databases, however as of before the meeting more extensive testing has not been done on the new design. We thoroughly tested inserting data on both MySQL and Python to ensure that there would be no more database architecture issues, allowing us to more comfortably move forward with the rest of the project.

Our next steps were to finish deciding models and methods used for **Prediction** and **Description**, including but not limited to:

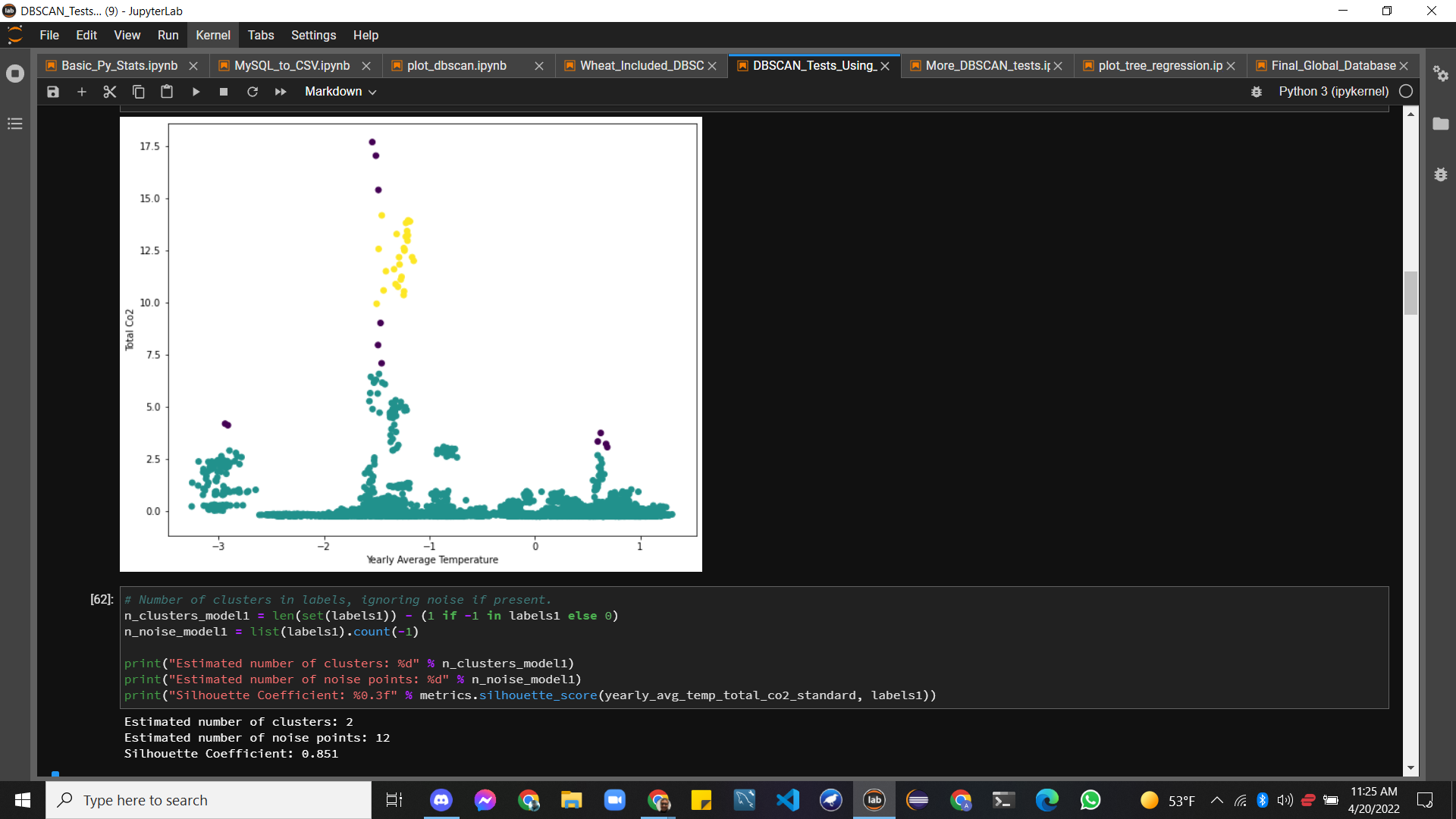
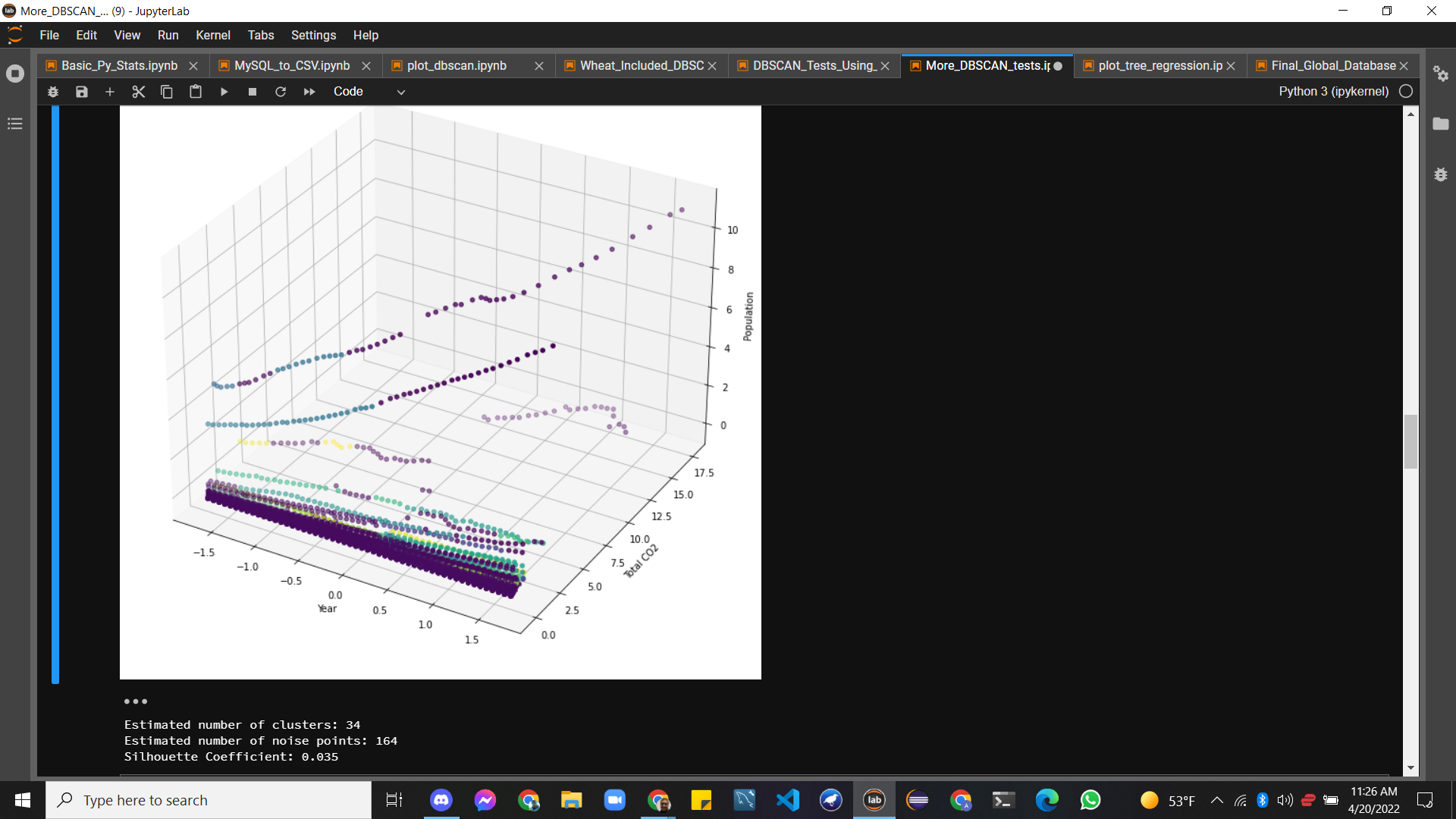
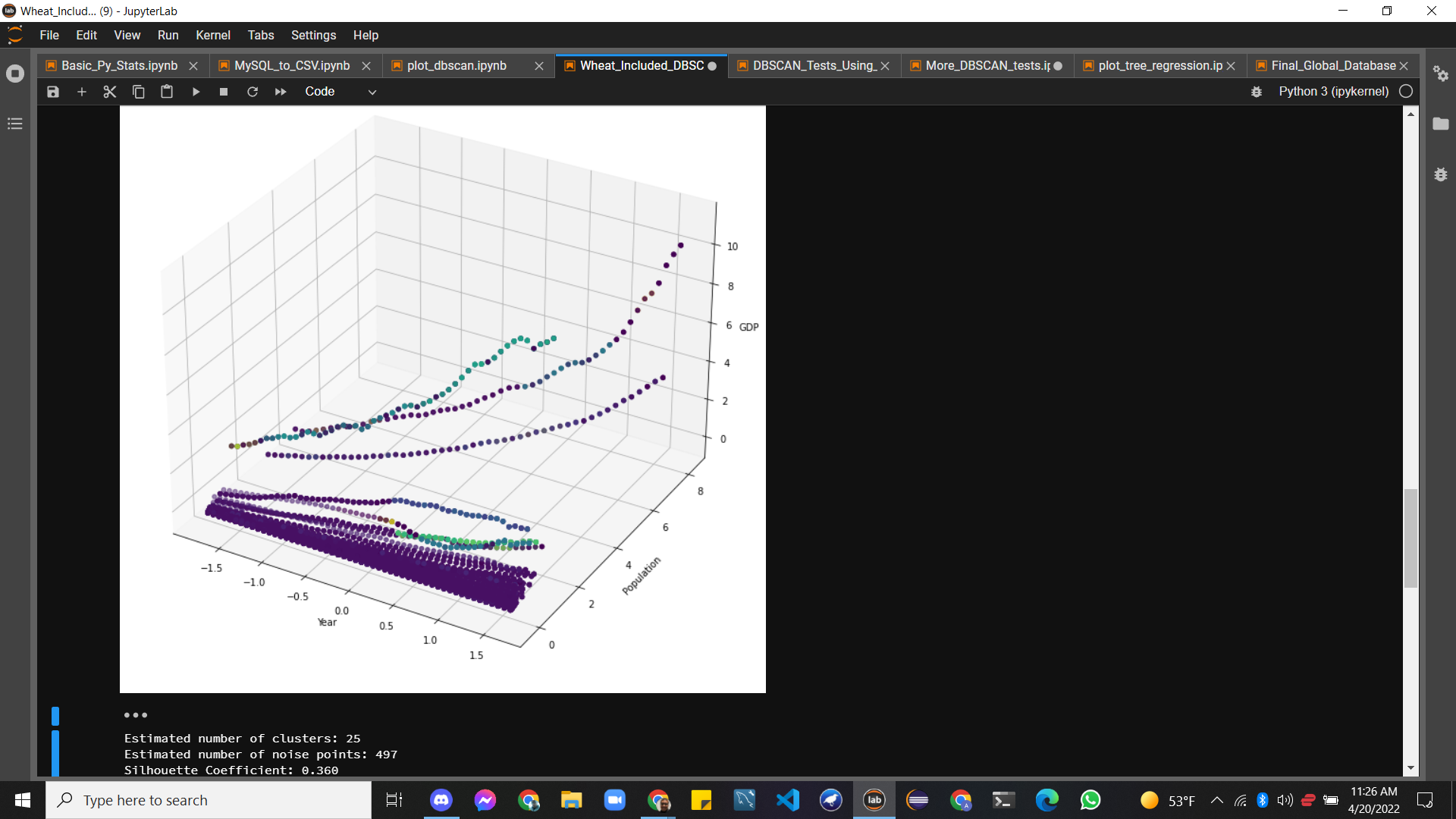
* Clustering
  + K-Means
  + DBSCAN - Density-based spatial clustering of applications with noise
  + HDBSCAN
  + Hierarchical (agglomerative)

**As of 04/20/2022**

Since the last meeting, 04/13/2022, I have finished relaying to the team which clustering models we can use and compare with their Silhouette Coefficients for validation of findings. I have begun implementing the DBSCAN model with various changed, included but not limited to:

* Seeing differences between Default DBSCAN parameters and changed parameters
* Exchanging different Features to see importance of features on clustering
* Exchanging amount of features used to see effect of curse of dimensionality has on DBSCAN, due to being based on Euclidean distance

Examples with Scatter Plots, Number of clusters generated, Number of noise points generated, and Silhouette Coefficient



**As of 04/27/2022**

Since the last meeting, 04/20/2022, I had discovered that PCA could help us better visualize the clusters and help our analyses. With this knowledge I worked towards implementing PCA with our data, inputting the scaled Numeric Columns and receiving/relaying the findings of the Principal component analysis to the rest of the team. This way we were able to discover a method of seeing which features played the biggest role, with Rhodiam adding the principal components to the scaled dataset and creating a Correlation Heatmap to do so. This allowed us to analyze the features with the highest variance AND plot the clusters in a Lesser Dimensional Space, due to the Dimensionality Reduction of PCA. This initially worked when trying to understand the creation of our clusters, thus understanding the characteristics of each cluster in more detail. We still noticed an abnormally large variance in the Milling\_Rate\_MT that massively affected the clustering models, which we found through Principal component analysis. This allowed us to go back and realize that Milling\_Rate\_MT is actually zero for some countries that grow rice. This led us to the realization that there are countries that have the arable land and proper climate to grow rice, but not the technological advancements to process their own rice. This leads to an overreliance on the countries that can process their food as the rice needs to be

**05-04-2022**

Worked with Rhodiam to further interpret and finalize the clusters through the use of CART (DecisionTreesClassifier in SKLearn). We determined that PCA would solely be used for cluster visualization and not for the CART interpretation process.

**05-05-2022**

Worked with Daniel getting started on the slides before moving onto continuing working with Rhodiam to interpret exactly how our current clusters were made. We had realized that interpretability and understandability are 2 extremely important aspects of clustering, not the Silhouette score that had previously been our deciding factor for choosing clusters.

**05-06-2022**

Started working on the Streamlit EDA and ML web app to showcase our project. Initially I hit a roadblock in that our data is large and thus the web app takes slightly longer than preferred to load so I looked into caching options.

**05-07-2022**

Started working on the Streamlit EDA and ML web app to showcase our project. Initially I hit a roadblock in that our data is large and thus the web app takes slightly longer than preferred to load so I looked into caching options.

**05-07-2022**

Working alongside Rhodiam we were able to attempt using Random Forest Regressors for increasing accuracy from our Decision Tree Regressor (CART). While we got significantly higher accuracy it was quickly obvious that the Forests were going to be difficult to interpret/explain. This was due to the fact that we were less concerned with accurately predicting and outputting predictions, but on interpretation of our data and the overall implications of our project. I continued working on the Streamlit demo, adding an EDA page and moving onto trying to implement some ML models we were using, a page for clusters and a page for regression, perhaps one just for dashboard/visualizations.

**05-07-2022**

Something that we had initially done was clustering on the Dataset without Encoding our categorical data (we were only initially going to remove country as One-Hot-Encoding would have added 136 features for Untrt alone. We wanted to see what would happen without encoding, and in the process we had discovered interesting findings about rice, such as certain countries may have a Rough\_Production but 0 Milling\_Rate. This indicated that some countries were capable of growing rice but didn’t have the technology to efficiently process the rice that they grew. We ultimately decided to encode the One-Hot-Encode the Commodity\_Description in order to better separate our clusters. This was how we decided our final clusters, moving on with our final clusters for regression analysis. I also continued working on the Streamlit App, adding more pages and visualizations.

**05-08-2022**

I worked closely with Rhodiam to run and optimize different regression models, such as Linear Regression, Multiple Regression, as well as Decision Tree Regression (CART) and Random Forest Regression. We noticed that with Multiple Regression that the R2 metric could not be used directly such as with Simple Linear Regression, due to the fact that for every additional feature added the R2 metric would increase to 100% accuracy. We immediately realized that was wrong and went about implementing an adjusted R2 that Multiple Regression could be measured with.