**05-04-2022**

**Rhodiam -** Tested out regression models such as OLS, using the Principal Components, on the entire data set, and on intracluster data. Results led to changing course and focusing on CART algorithm, and use of decision trees.

**Jason -** 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.

**Daniel -**

**05-05-2022**

**Rhodiam -** Using Decision Tree Classifiers with and without PCA attributes, and the clusters as the target variable in order to test how accurate/valid a default decision tree classifier can be used to test out what PCA components/attributes might have the most influence on the clusters/splits

**Jason -** 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.

**Daniel -** Laid down the framework for the final presentation’s slides.

**05-06-2022**

**Rhodiam -** Using Decision Tree Regressors to look at what variables end up being the most important when it comes to predicting a specified variable

**Jason -** 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.

**Daniel -**

**05-07-2022**

**Rhodiam -** Started Regression Analysis with clusters, along with use of Random Forest Regressors, worked on sections of the Project Report.

**Jason -** 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.

**Daniel -**

**05-08-2022**

**Rhodiam -** Applied One-Hot Encoding to the Commodity\_Description attribute of the dataset, to have features that details what row pertains to what grain, in order to be able to utilize that feature in clustering, leading to the formation of new clusters with DBSCAN.

**Jason -** 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.

**Daniel -**

**05-09-2022**

**Rhodiam -** Running different regression models, and making different observations about the different regression models, along with the clusters and the trends based on certain variables

**Jason -** 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.

**Daniel -**

**05-10-2022**

**Rhodiam -** Finishing up the slides and working with Jason to come up with some of the conclusions that we found with our regression analysis. Tying those conclusions to the hypothesis, and what we learned overall about our data and topic.

**Jason -** Myself and Rhodiam worked to finish the slides and the report. We analyzed our findings once more and compared then with our initial hypotheses in order to come to a conclusion.

**Daniel -**