# *PREDICTING CROP YIELD*

ISyE 6414 Regression Analysis, Spring 2023

*Cassidy Gasteiger*
*Raymond Li*
*Suraj Shourie*
*Asadullah Syed*
*Angela Zheng*

*Team 8*

# Table of Contents


*1. Introduction* ..... *2*

*2. Problem Statement* ..... *2*

*3. Data Description* ..... *2*

**Feature Engineering** ..... **3**

**Data Cleaning** ..... **4**

*4. Analysis* ..... *4*

**Model 1 - Baseline Multiple Linear Regression** ..... **4**

**Model 2 - Reduced and Transformed Multiple Linear Regression** ..... **7**

**Model 3 - Regularized Regression** ..... **8**

**Model 4 - Random Forest Regression** ..... **9**

*5. Conclusions and Recommendations* ..... *10*

*6. Appendix* ..... *11*

**6.1 Original Datasets and Variables** ..... **11**

**6.2 Exploratory Data Analysis** ..... **11**

Correlation of Crop Yield Variables ..... 12

Crop Yield ..... 13

All Features Correlation Plot ..... 15

Density Plots ..... 16

**6.3 Model 1 – Baseline Model Residual Analysis Plots** ..... **16**

**6.4 Model 2 – Selected Features and Transformed Y-Variable** ..... **17**

**6.5 Model 3 – Regularized Regression** ..... **18**

**6.6 Model 4 – Random Forest Regression** ..... **20**

## 1. Introduction

To feed a growing world population, food system experts need to be able to accurately predict crop yield based on pre-harvest conditions. Farmers need to know how much they will yield to plan planting, harvesting, and marketing strategies and allocate resources; upstream buyers need to know how much of each crop will be available; and the whole food system relies on adequate crop production to feed livestock, produce packaged goods and processed foods, and feed a growing world population. As a result, the knowledge of expected crop yield is highly valued, especially in a world of constant change and evolution that impacts the ecosystems that we depend upon for food.

## 2. Problem Statement

Our goal is to predict annual crop yield with our analysis. Therefore, we must take into consideration many factors. The most critical variables impacting crop yield are soil fertility, availability of water, climate, and diseases or pests. Diseases or pests are difficult to measure, but pesticide application is a potential proxy to understand if applying more pesticides improves yield.

Our independent variable is the crop yield for ten of the world's most-consumed crops: maize, potatoes, rice, sorghum, soybeans, wheat, cassava, sweet potatoes, plantains, and yams. We predict yield per country for each crop using data from 1990 – 2013 from 98 countries. For each of these countries, we have data of the average rainfall in that country, total tons of pesticides applied, and average temperature.

Because crop yield is a highly studied phenomenon, with many researchers attempting to predict crop yield based on climate and soil quality, we also explored more novel factors to check their correlation to crop yield. Specifically, we explored whether country-level socioeconomic factors like birth rates, land use for agriculture, and life expectancy are correlated with higher crop yields – do wealthier countries tend to have higher yields for the same crops?

Ultimately, our objective was to build a predictive model for crop yield based on precipitation, climate, and pesticide usage that also sheds new light on whether country-level socioeconomic and political factors have any relationship to yield.

## 3. Data Description

We obtained our core yield, precipitation, temperature, and pesticide usage data from a pre-existing Kaggle dataset. This dataset came with 5 files. For this dataset, the pesticide and yield data were extracted from the Food and Agriculture Organization of the United Nations (FAO). The rainfall and temperature data were originally from the World Bank. Our core dataset is a cleaned and merged version of these four variables and has over 27,000 data points that cover the top 10 most consumed crops in the world. Our variables include:

- Area (country) - qualitative
- Item (crop name) - qualitative
- Year - can be qualitative if years are treated as categories or quantitative as numeric data
- Hectogram yield per hectare (Hg/Ha) - quantitative
- Average rainfall by year (mm) - quantitative
- Pesticides used in the country (tons) - quantitative
- Average temperature (Celsius) - quantitative

We used 2013, the final year in the dataset, as our test set to see if it is feasible to accurately predict crop yield in each country for each crop based on rainfall, pesticides, and temperature.

In addition to the variables found in this dataset, we also joined socioeconomic and land use variables obtained from the World Bank to explore relationships between crop yield and country profiles. Feature selection techniques and correlation testing demonstrated which of these factors are highly correlated with crop yield and worth further exploration and inclusion in our regression model. These additional variables include:

- Patent applications, residents – quantitative
- Birth rate, crude (per 1,000 people) – quantitative
- Arable land (% of land area) – quantitative
- Population, female (% of total population) – quantitative
- Land area (sq. km) – quantitative
- Net migration – quantitative
- Population ages 15-64 (% of total population) – quantitative
- Agricultural land (sq. km) – quantitative
- Forest area (sq. km) – quantitative
- Urban population (% of total population) – quantitative
- Labor force– quantitative
- Fertility rate, total (births per woman) – quantitative
- Life expectancy at birth, total (years) – quantitative
- Number of infant deaths – quantitative
- Survival to age 65, female (% of cohort) – quantitative
- Merchandise exports (current US$) – quantitative
- Merchandise imports (current US$)– quantitative

Our data is time dependent, since we have annual data on each of these features for our selected time period, but the values represent an average throughout the year.

## Feature Engineering

Several of the World Bank variables were highly collinear. We feature engineered and transformed the socioeconomic variables to limit collinearity and create new, more explanatory variables. Specifically, we created the following variables:

- Patent applications per 1,000 people (dropped total applications)
- Net migration as % of total population (dropped raw net migration)
- Agricultural land as % of total land area (in place of sq. km.)
- Forest land as % of total land area (in place of sq. km.)
- Balance of trade as merchandise exports – imports, as % of GDP (in place of exports and imports)
- Population in labor force as % of total population (in place of total labor force)

From our environmental variables, we also created a one-year lag variable to test for trends:

- Last year's crop yield for that same country and crop

We dropped percent arable land, life expectancy, and population aged 15-64 to reduce collinearity. This left us with 13 socioeconomic variables, 4 environmental variables, and 2 categorical variables:

| Variable Name | Type | Units |
|---|---|---|
| **Country** | Qualitative | -- |
| **Crop Type** | Qualitative | -- |
| **Year** | Quantitative | -- |
| **Yield** | Quantitative – Y variable | Hg/Ha |
| **1-lag: last year's yield** | Quantitative | Hg/Ha |
| **Average annual rainfall** | Quantitative | mm |
| **Pesticide usage** | Quantitative | Tons |
| **Average temperature** | Quantitative | Celsius |
| **Birth rate** | Quantitative | Per 1,000 people |
| **Patent applications** | Quantitative | Per 1,000 people |
| **Land area** | Quantitative | Sq. km |
| **Urban population** | Quantitative | % of total population |
| **Fertility rate** | Quantitative | Births per woman |
| **Survival to age 65, female** | Quantitative | % of cohort |
| **Mortality rate, infant** | Quantitative | Per 1,000 live births |
| **Balance of trade** | Quantitative | % of GDP |
| **Population in labor force** | Quantitative | % of total population |
| **Agricultural Land** | Quantitative | % of total land area |
| **Forest Land** | Quantitative | % of total land area |
| **Net migration** | Quantitative | % of total population |
| **Patent Applications** | Quantitative | Per 1,000 people |

### Data Cleaning

We merged World Bank data with Crop Yield Data from Kaggle into one csv file. GDP, which we used to feature engineer balance of trade, was the only variable with null values; we imputed for those missing values using the previous year's GDP for that country. We then one-hot encoded our two categorical variables Country and Crop Type.

After this cleaning and preprocessing, we were left with our dataset of 98 countries, 10 crops, and 27,228 data points ready for our analysis. For further information about the dataset, see the Appendix for our exploratory data analysis.

## 4. Analysis

### Model 1 - Baseline Multiple Linear Regression

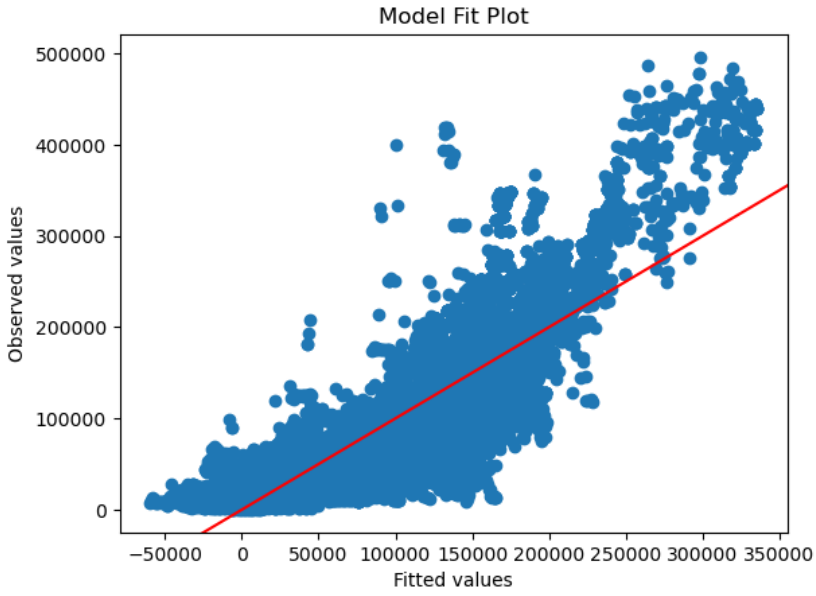
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | hg/ha_yield | R-squared: | 0.758 |
| Model: | OLS | Adj. R-squared: | 0.756 |
| Method: | Least Squares | F-statistic: | 596.5 |
| Date: | Sun, 23 Apr 2023 | Prob (F-statistic): | 0.00 |
| Time: | 21:22:55 | Log-Likelihood: | -2.8410e+05 |
| No. Observations: | 23608 | AIC: | 5.684e+05 |
| Df Residuals: | 23484 | BIC: | 5.694e+05 |
| Df Model: | 123 | | |
| Covariance Type: | nonrobust | | |

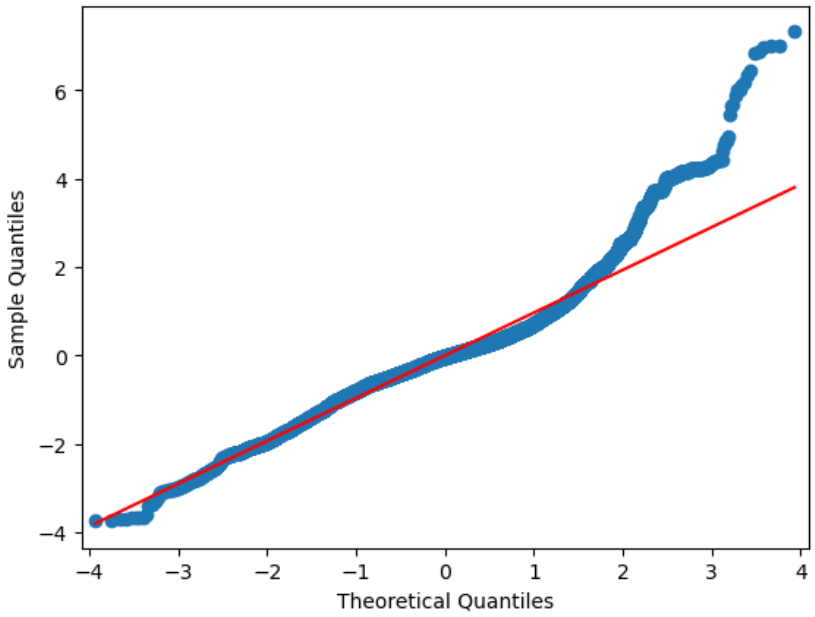
| | coef | std err | t | P>\|t\| | [0.025 |
|---|---|---|---|---|---|
| const | -1.887e+06 | 2.19e+05 | -8.608 | 0.000 | -2.32e+06 |
| Year | 980.0925 | 124.862 | 7.849 | 0.000 | 735.355 |
| average_rain_fall_mm_per_year | -26.1617 | 10.012 | -2.613 | 0.009 | -45.786 |
| pesticides_tonnes | 0.0353 | 0.013 | 2.681 | 0.007 | 0.010 |
| avg_temp | 7.4626 | 152.847 | 0.049 | 0.961 | -292.128 |

For this dataset, we started by building a base multiple linear regression model for the yield and all the quantitative variables. After encoding all qualitative variables, we had a total of 127 features for this model. Looking at the model summary, we can see that we got an $R^2$ of 75.8% and an adjusted $R^2$ of 75.6%. The RSME for this model was 50866 and the MAPE was 83.13%. The F-statistic was very low, so we know that at least some of these values are statistically significant in describing this model. Upon further inspection, we found that there were many features that were not statistically significant which would require further investigation. 55 out of our 127 features in this model had a p-value that was greater than 0.05. The VIF values were high, which indicated multicollinearity may be problematic in this model.

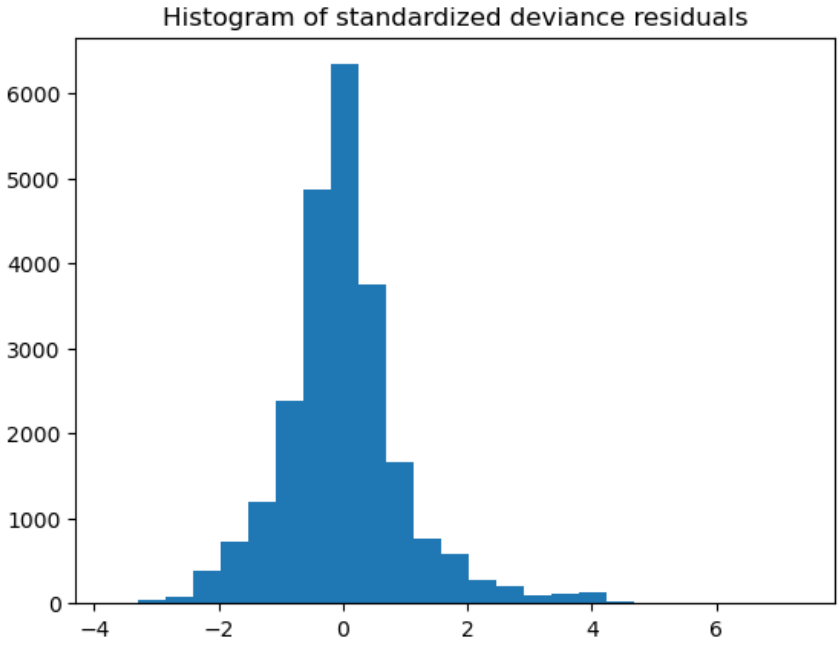
Shown below is our model fit plot.



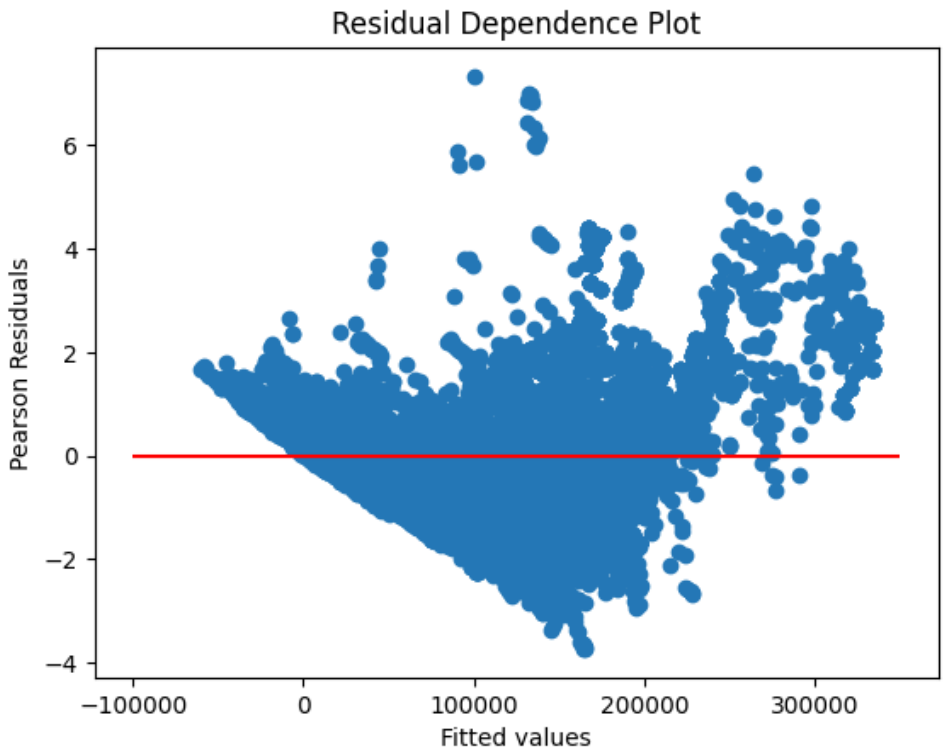

From this graph, it seems that the data points are more evenly distributed around our line towards the center values but are less closely aligned with our linear regression line as values get higher. This can also be seen on our Q-Q plot.




We can tell from our Q-Q plot that values in the higher quantiles are not aligned with our model. This means that the normality assumption is violated for our model because of the outward deviating tails in the plot.




Examining the histogram of residuals, we see the bell-shaped curve without any signs of skewness or heavy tails. These observations imply that the normality assumption seems to hold.



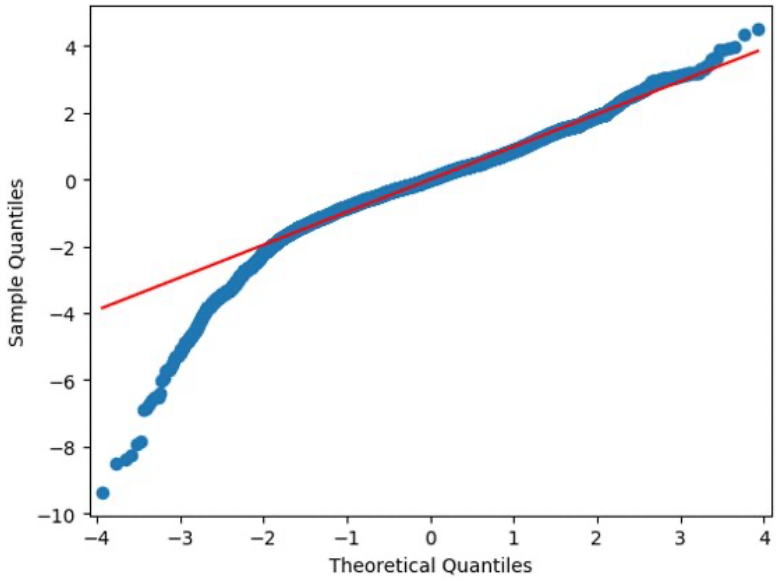

However, our residual dependence plot shows signs of parabolic behavior. This suggested that we should apply a Box-Cox transformation to our y-variable to try to create a more normal fit.

Since we had about 27,000 data points, the Cook's distance is considered high if it is any higher than 1.48 x $10^{-4}$. We observed some potential outliers, with the highest Cook's distance values around 6.59 x $10^{-3}$. This means that these data points required further investigation. We ultimately chose not to exclude them from our model, as they were legitimate data points that represented particularly high or low years of crop yield; crop yield is highly variable, and we wanted to ensure our model captured that variability.

## Model 2 - Reduced and Transformed Multiple Linear Regression

As our next step, we went on to perform a reduced model without the features that were statistically insignificant. Our one-lag time variable was not significant, so we ultimately did not create a time series regression model. We also applied a Box-Cox transformation to our y-variable with $\lambda = .0144$, since we noticed the residual dependence plot showed signs of parabolic behavior.

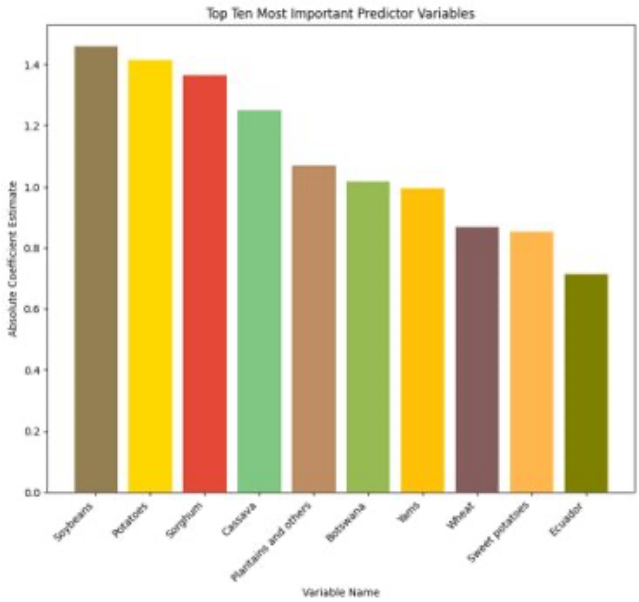
Dropping the statistically insignificant variables reduced our number of features from 98 to 71. Our R-squared was 81.3%, the RSME value was 52720, and MAPE was 44.83%. Our reduced model had increased our R-squared value and decreased the MAPE value compared to the full model. With our reduced model, all our coefficients in this model are statistically significant.




From the Q-Q plot, we can still see that there is deviance in the tail of the plot, but our other residual plots and diagnostic tests revealed a stronger-performing and more normal regression model.

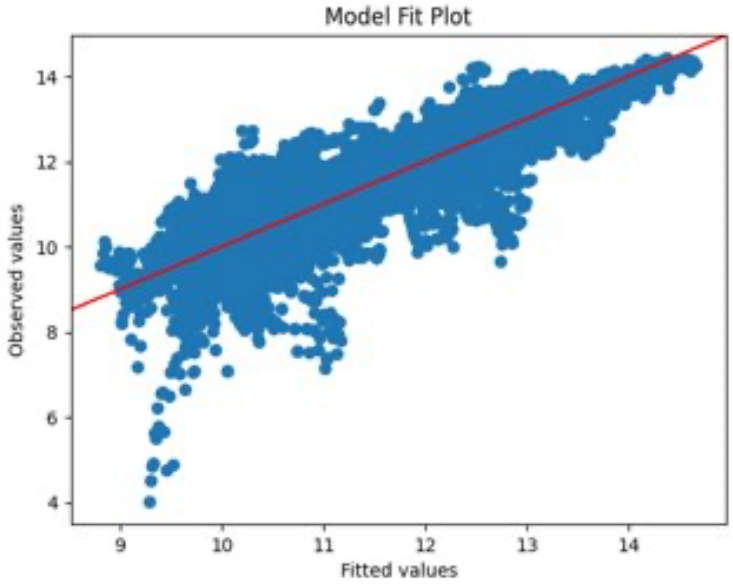
## Model 3 - Regularized Regression

We next tried using Lasso/Ridge regression on all the variables to conduct feature selection. We once again used the same Box-Cox transformation on the y-variable to adjust for abnormal behavior in the residuals. After tuning the hyperparameters using grid search, we found that $\alpha = 0.01$ and L1 regularization weight = 0.1 yielded the best result with a good trade-off between $R^2$ and MAPE. This model had an $R^2$ of 73.1%, RSME of 49808, and MAPE of 39.82%. From this model, we found that the



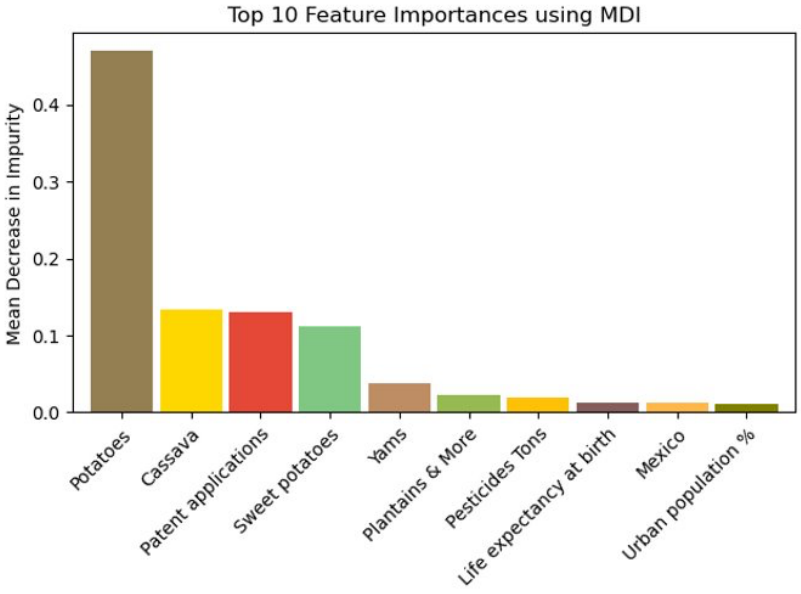

type of crop was the most important feature. Shown in the bar plot, "Top Ten Most Important Predictor Variables", are the 10 most important features described by this model.

Additionally, we can see from this regularized and transformed model that our goodness-of-fit also improved and our residuals showed signs of normal behavior.




## Model 4 - Random Forest Regression

Finally, we built a random forest regression model to test its performance against our regularized model. Despite using hyperparameter tuning with the random forest regression model, we were unable to find a better result than our regularized regression without allowing the leaf sizes to be small and overfitting the data. Our random forest model had an $R^2$ of 71.8%, RSME of 51026, and MAPE of 47.79%. Shown below are the 10 most important features determined by this model.



## 5. Conclusions and Recommendations

| | Model 1: Baseline MLR | Model 2: Reduced MLR | Model 3: Regularized | Model 4: Random Forest |
|---|---|---|---|---|
| **$R^2$** | 75.75% | 81.3% | 73.1% | 71.8% |
| **RMSE** | 50866.34 | 52720 | 49808.41 | 51026.08 |
| **MAPE** | 83.13% | 44.83% | 39.82% | 47.79% |

The regularized regression model with tuned hyperparameters is the most viable model, as it minimizes both RMSE and MAPE. The model can explain about 73% of variability in crop yield by year while also achieving the lowest MAPE score out of all our models.

In terms of the socio-economic factors that we initially set out to explore, there appears to be strong predictive power associated with these unconventional features such as patent applications, life expectancy and urban population, as evidenced by the MDI scores in the random forest model. This suggests that socioeconomic indicators are strongly correlated with crop yield (and our initial exploratory data analysis supports this conclusion). One possible explanation is that wealthier countries have the agricultural resources and investments to produce higher crop yield. We recommend further investigation into these relationships using more advanced models.

# 6. Appendix

## 6.1 Original Datasets and Variables

| Variable Name | Type | Description | Units |
|---|---|---|---|
| Area | Qualitative | Country in question (e.g., USA, UK) --> 98 unique countries included in the dataset | -- |
| Item | Qualitative | Type of Crop (e.g., Maize, Potato) --> 10 unique crop types included in the dataset | -- |
| Year | Quantitative | Year of produce (1990 - 2013) | -- |
| Yield | Quantitative | Yield for each type of crop measured in hectogram yield per hectare | Hg/Ha |
| Average annual rainfall | Quantitative | Annual recorded rainfall for the given country in the given year measured in mm | mm |
| Pesticides | Quantitative | Total Amount of pesticides used in Crops per year measured in tonnes | Tons |
| Average Temperature | Quantitative | Annual Average temperature of the country measured in degrees Celsius | Celsius |

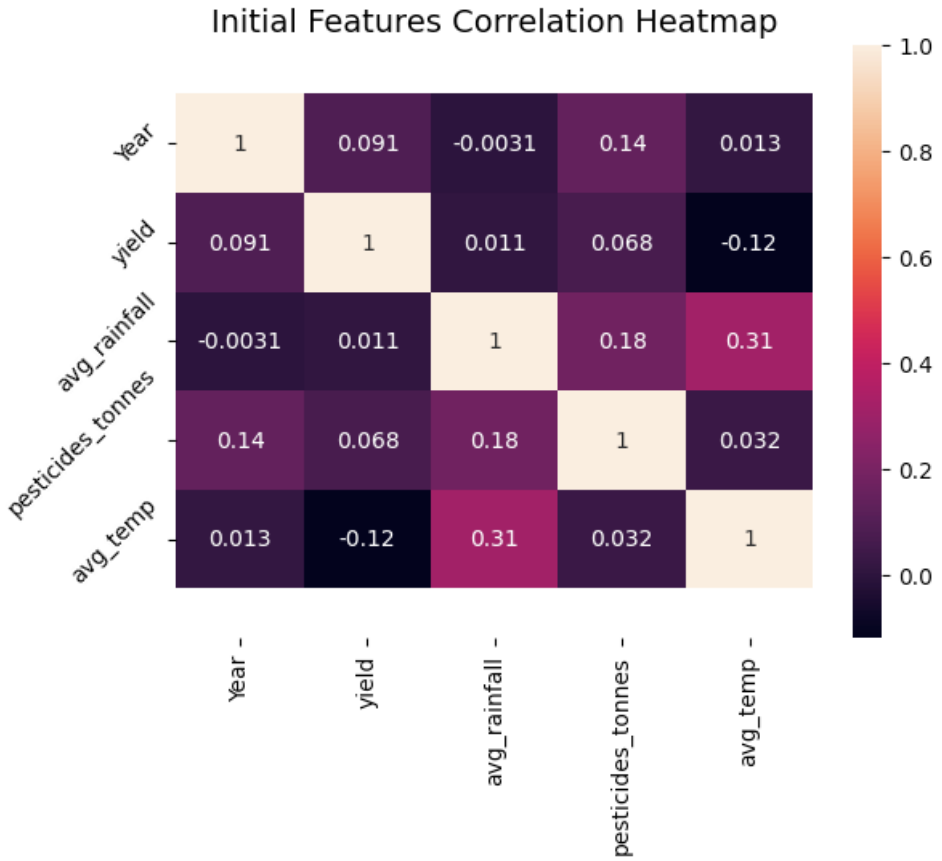
Table: Crop Yield Kaggle Data Variables

| Variable Name | Type | Units |
|---|---|---|
| **Patent applications, residents** | Quantitative | -- |
| **Birth rate, crude** | Quantitative | Per 1,000 people |
| **Arable land** | Quantitative | % of land area |
| **Population, female** | Quantitative | % of total population |
| **Land area** | Quantitative | Sq. km |
| **Net migration** | Quantitative | -- |
| **Population ages 15-64** | Quantitative | % of total population |
| **Agricultural land** | Quantitative | Sq. km |
| **Forest area** | Quantitative | Sq. km |
| **Urban population** | Quantitative | % of total population |
| **Labor force** | Quantitative | -- |
| **Fertility rate, total** | Quantitative | Births per woman |
| **Life expectancy at birth, total** | Quantitative | Years |
| **Number of infant deaths** | Quantitative | -- |
| **Survival to age 65, female** | Quantitative | % of cohort |
| **Merchandise exports** | Quantitative | Current US$ |
| **Merchandise imports** | Quantitative | Current US$ |

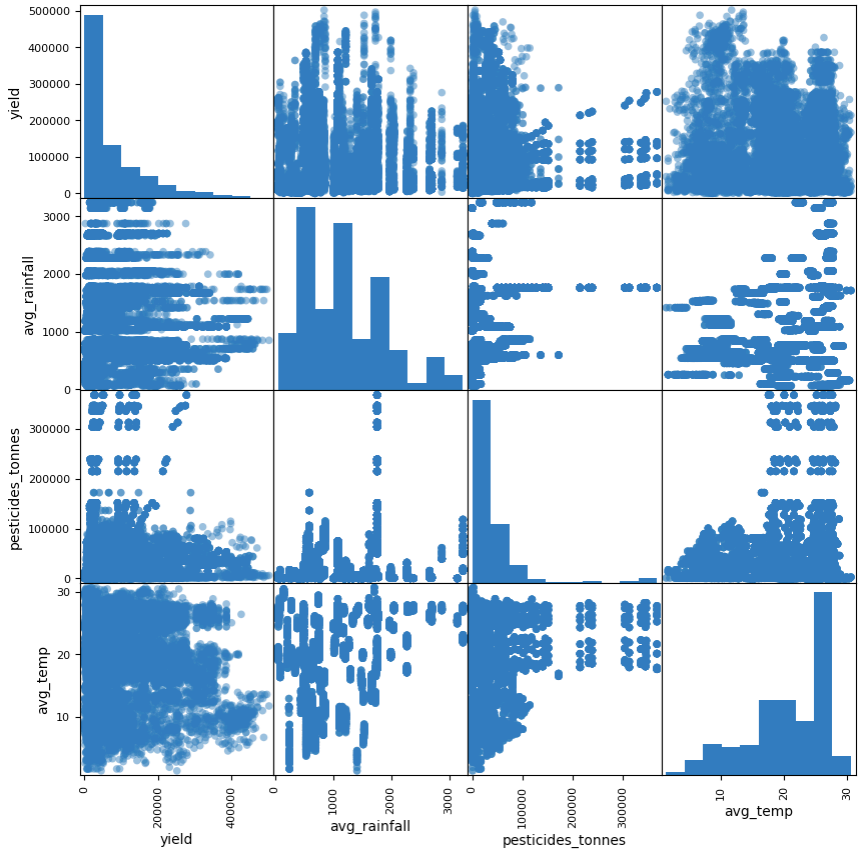
Table: World Bank Variables

## 6.2 Exploratory Data Analysis

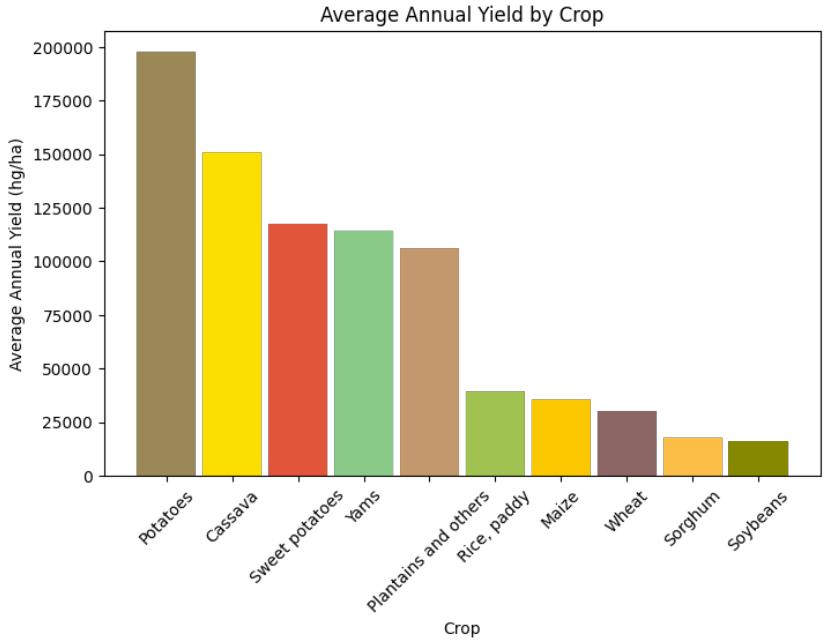
## Correlation of Crop Yield Variables

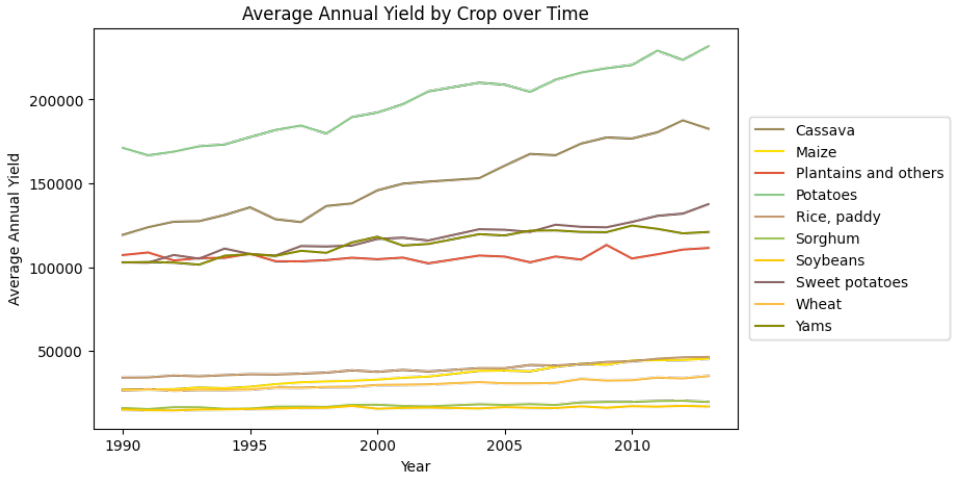


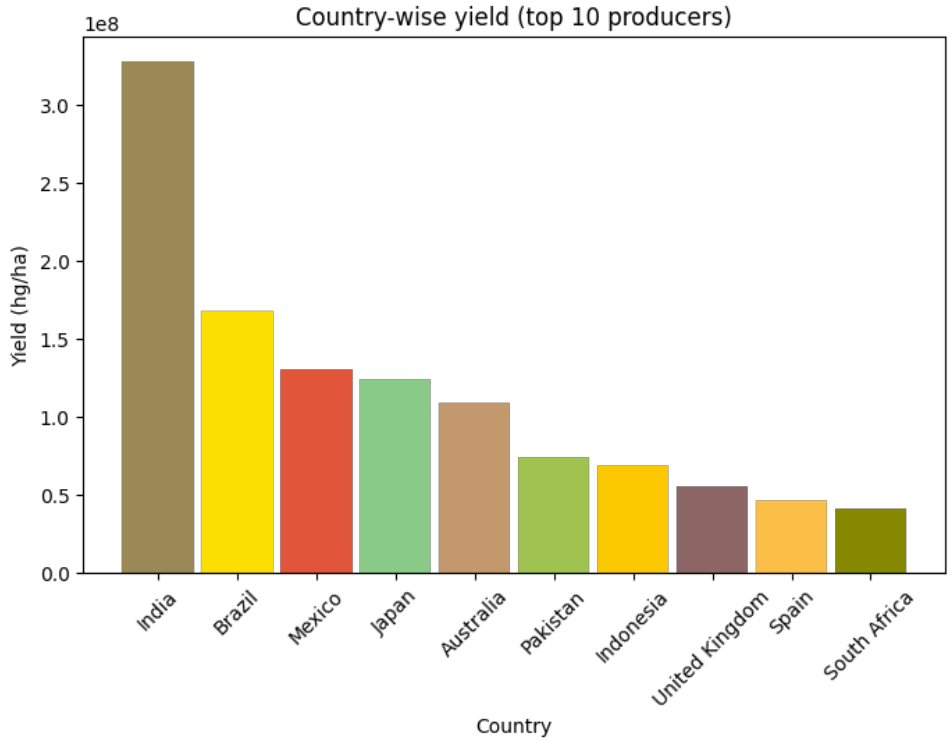
Correlation Plots

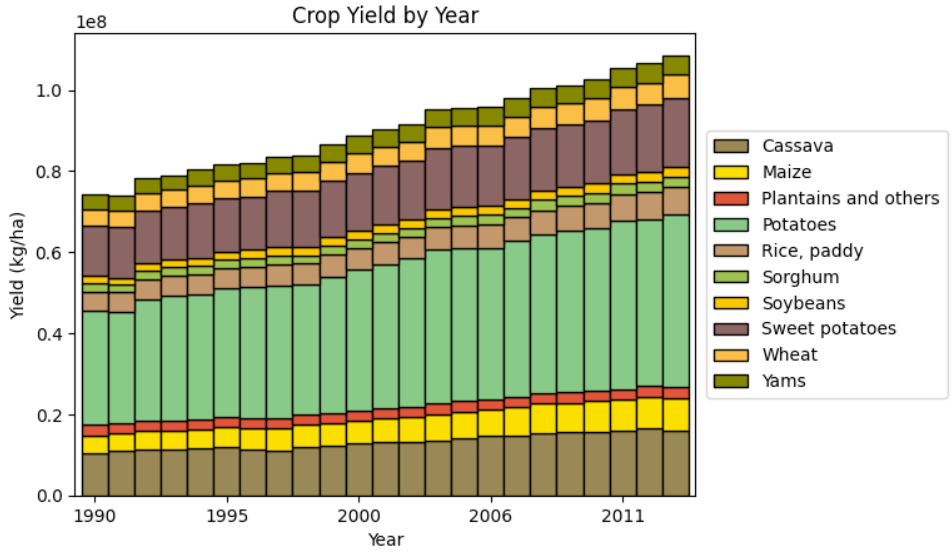



## Crop Yield

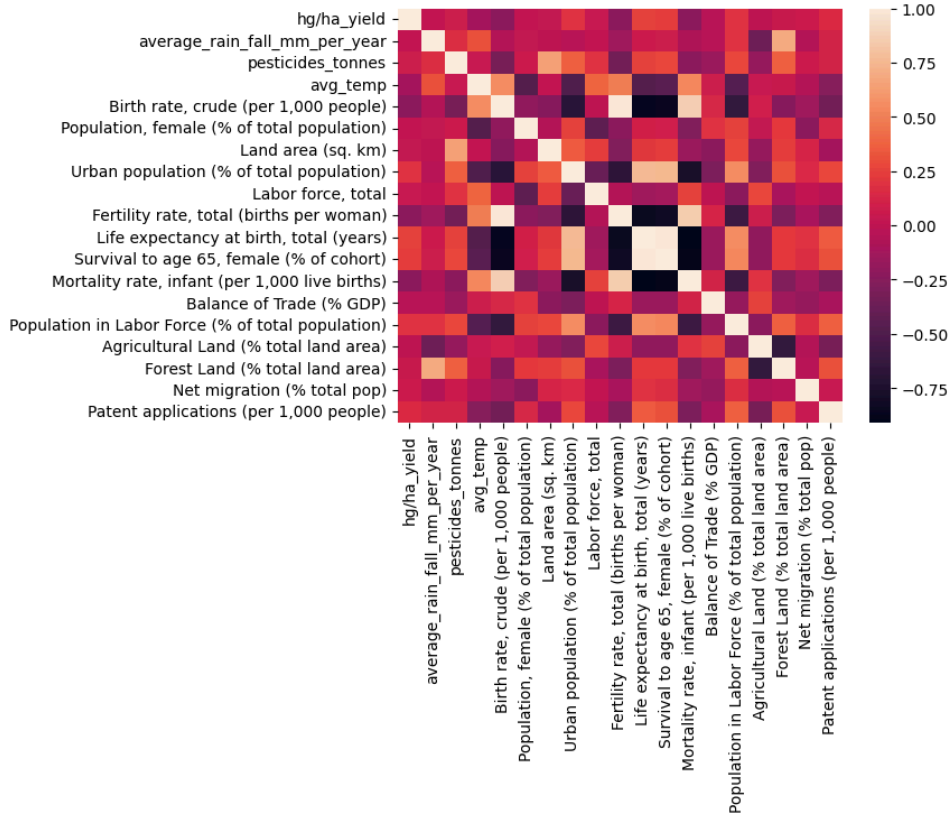



Average Annual Yield by Crop over Time
Average Annual Yield
200000
150000
100000
50000
1990
1995
2000
2005
2010
Year
Cassava
Maize
Plantains and others
Potatoes
Rice, paddy
Sorghum
Soybeans
Sweet potatoes
Wheat
Yams

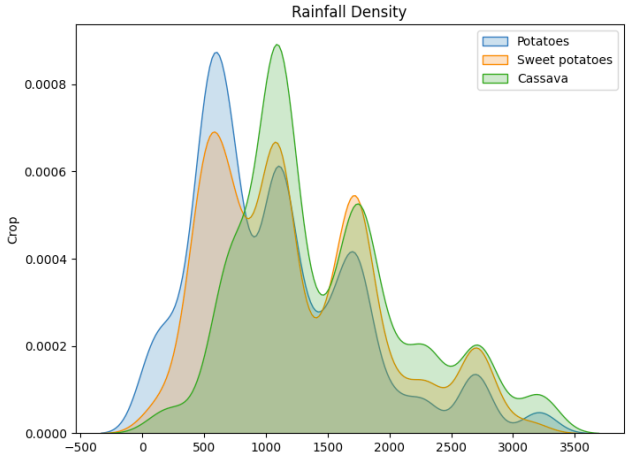

Country-wise yield (top 10 producers)
1e8
Yield (hg/ha)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
India
Brazil
Mexico
Japan
Australia
Pakistan
Indonesia
United Kingdom
Spain
South Africa
Country

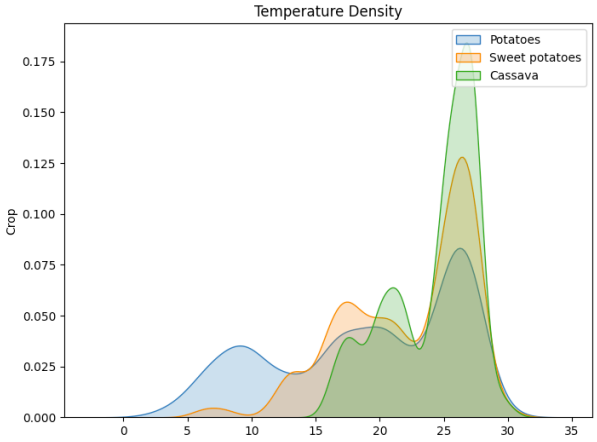


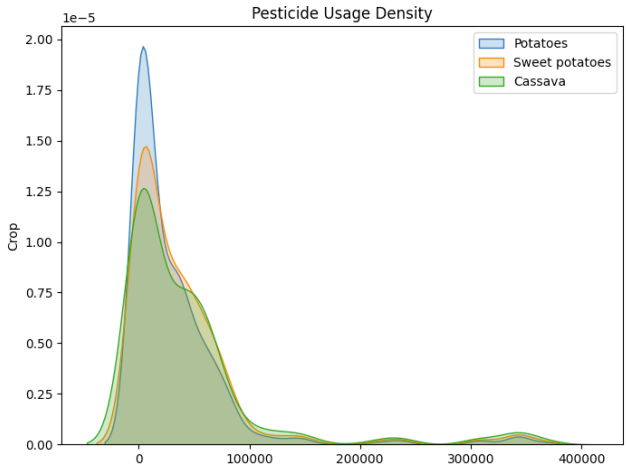

## All Features Correlation Plot



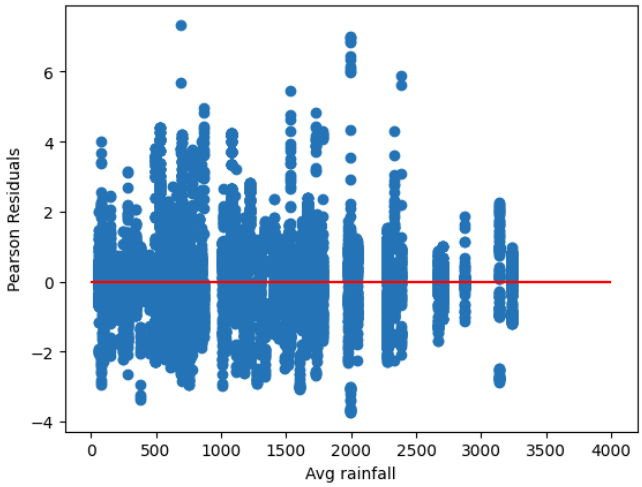
## Density Plots

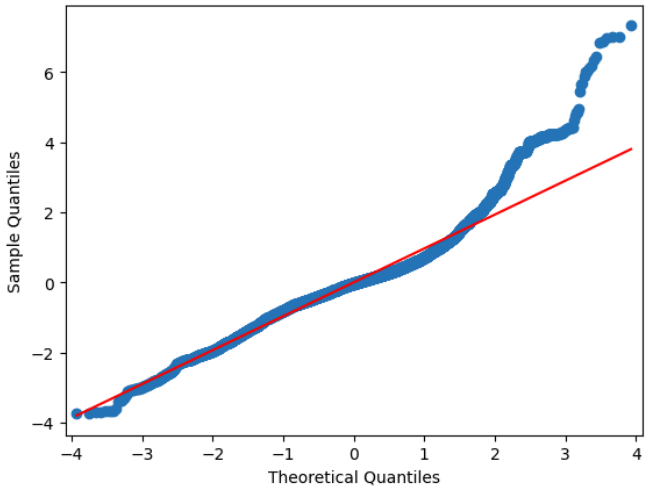


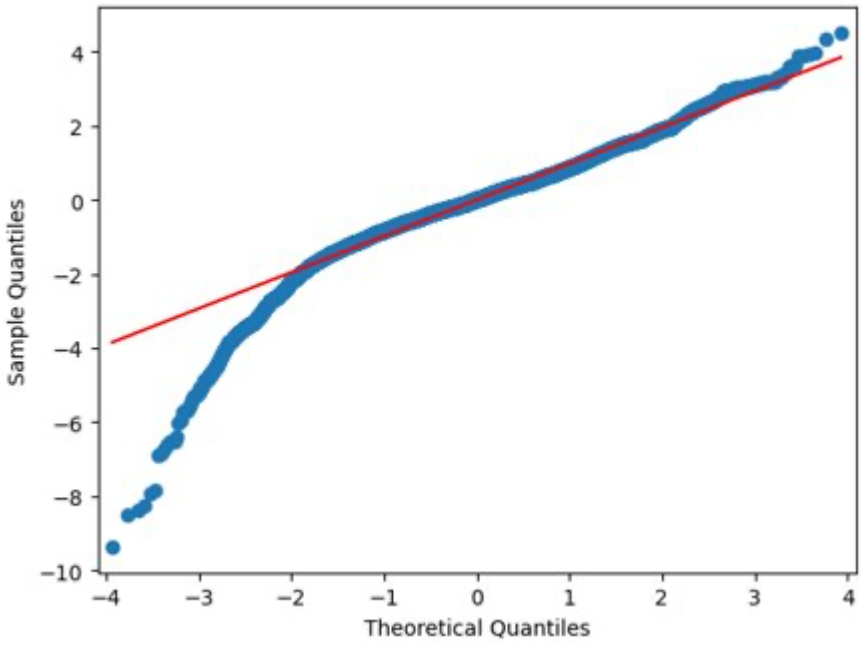




## 6.3 Model 1 – Baseline Model Residual Analysis Plots

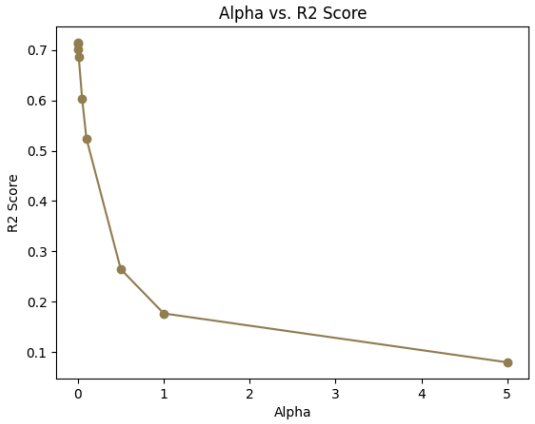


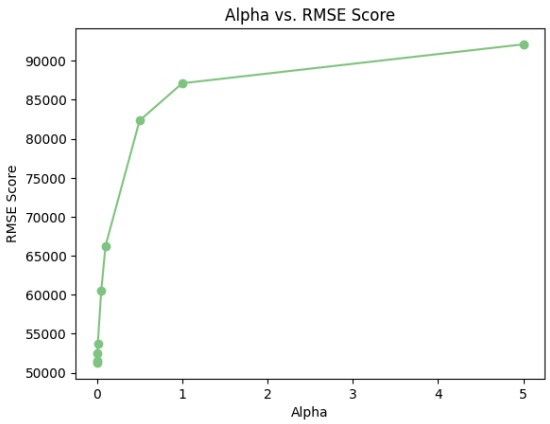



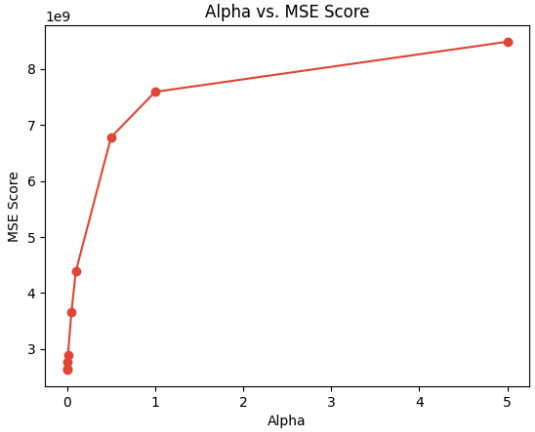

## 6.4 Model 2 – Selected Features and Transformed Y-Variable

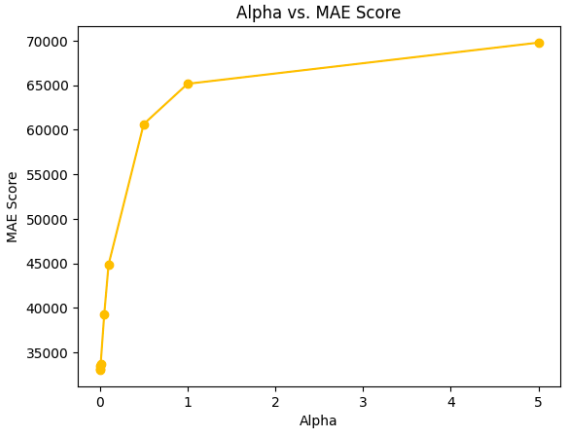


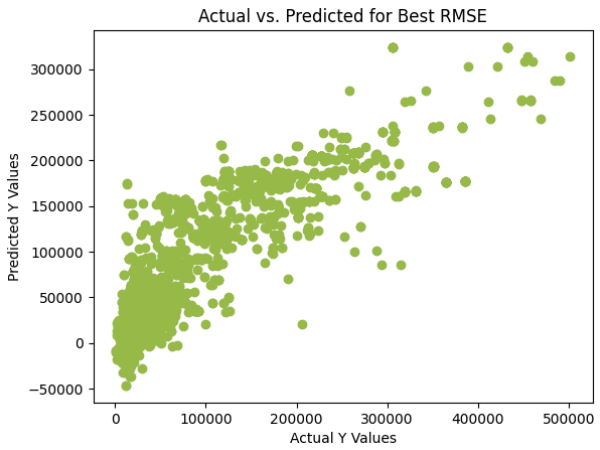
## 6.5 Model 3 – Regularized Regression

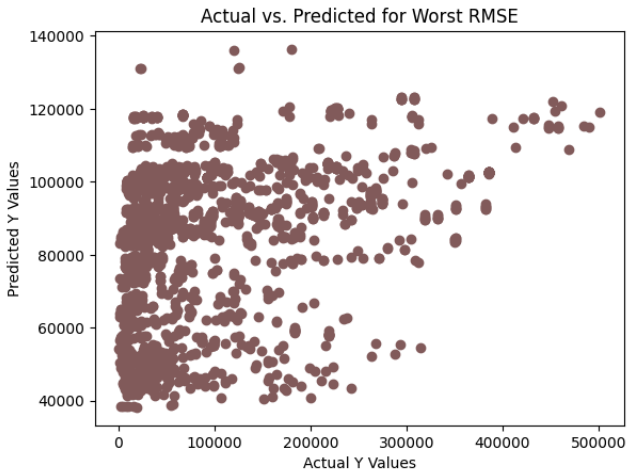


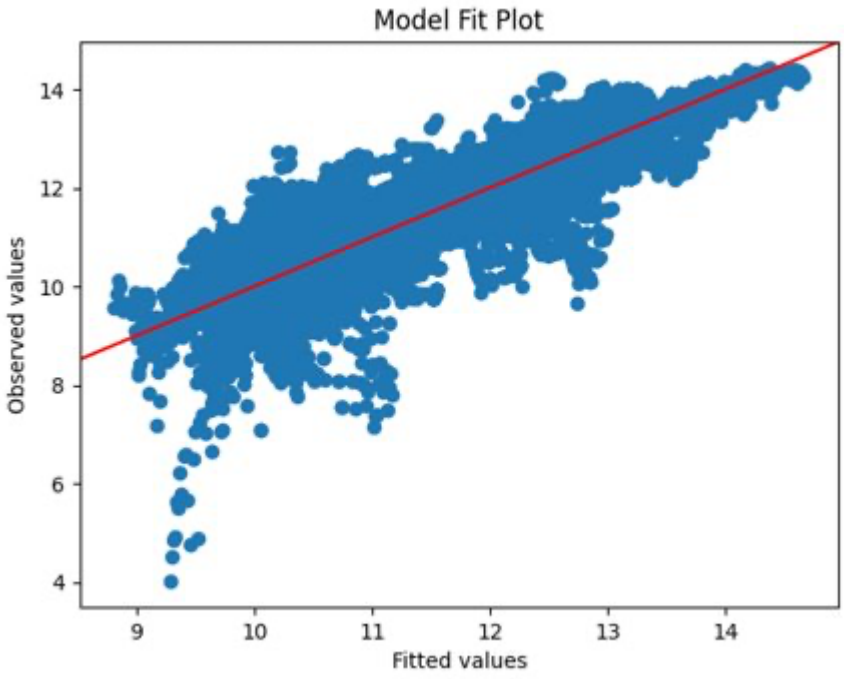



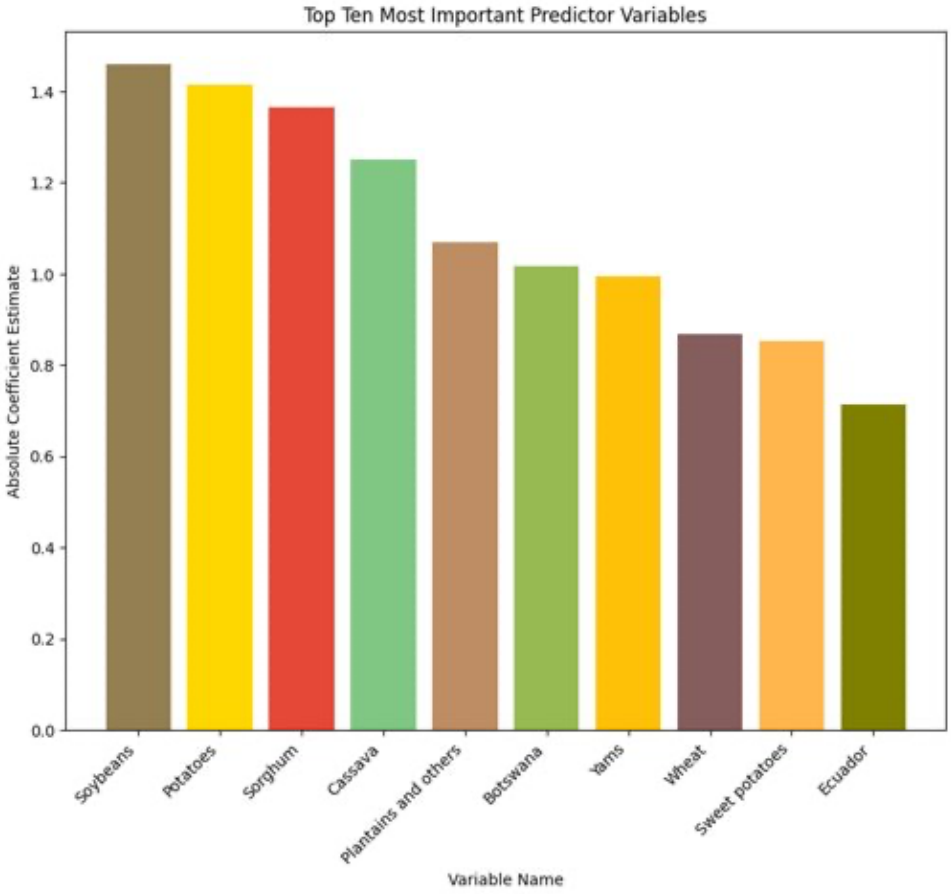




Model Fit Plot
Observed values
Fitted values
14
12
10
8
6
4
9
10
11
12
13
14


Top Ten Most Important Predictor Variables
Absolute Coefficient Estimate
Variable Name
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Soybeans
Potatoes
Sorghum
Cassava
Plantains and others
Botswana
Yams
Wheat
Sweet potatoes
Ecuador

## 6.6 Model 4 – Random Forest Regression