Capstone Three – Final Project Report

Problem Statement:

Nitrogen is ubiquitous in the environment. Naturally present in air and water, the total volume of nitrogen in the environment is increased through the introduction of elements such as fertilizers. Excessive nitrogen poses a hazard to both wildlife and humans with the introduction of large algae blooms and increased secondary bacteria and toxins. According to the United States Environmental Protection Agency (EPA) “Nutrient pollution is one of the America’s most widespread, costly and challenging environmental problems, and is caused by excess nitrogen and phosphorus in the air and water.” As a result, using materials that knowingly introduce hazardous elements should be accomplished efficiently.

This project will focus on the efficiency of various countries in the Americas. A total of seven crops were randomly selected from Food and Agriculture Organization of the United Nations database as well as Nitrogen and Phosphate production.

Data Wrangling:

The Food and Agriculture Organization of the United Nations (FAOSTAT) provides information on area harvested, yield, and production of various crops in multiple countries. Values are determined through various means and flagged for clarification. Data selection is initially limited to a ten-year period from 2005 to 2015. Additional selection was limited geographically to countries within the America and a selection of seven crops: Almonds, Apples, Apricots, Avocados, Bananas, Blueberries, and Cherries.

Individual crop data was collected from FAOSTAT through their download selection process as individual csv files. Each file contained multiple columns detailing the domain, area, element, item, year, unit, and value. A flag column was also included however for initial analysis this column is omitted without inspection.

The standard csv file setup initially stacked the 3 unique values of the element column, Area harvested, Yield, and Production, within the same column. For cleaning and modeling purposes these 3 values would need to be in separate columns. Dividing the csv files into individual data frames was accomplished through filtering on the values of the element column. Once divided, the columns needed to be merged into a single crop data frame. Element data frames were merged on the Area code and Year columns to ensure consistency between the countries and years.

	Area Code	Year	Almond_Area	Almond_Yield	Almond_Prod	Apples_Area	Apples_Yield	Apples_Prod	Apricots_Area	Apricots_Yield	...
0	9	2005	257.0	19027.0	489.0	43723.0	275875.0	1206210.0	2305.0	110638.0	...
1	9	2006	260.0	18846.0	490.0	40173.0	273816.0	1100000.0	2300.0	109565.0	...
2	9	2007	270.0	18519.0	500.0	36222.0	276075.0	1000000.0	2400.0	106250.0	...
3	9	2008	283.0	19011.0	538.0	34021.0	279239.0	950000.0	2479.0	105518.0	...
4	9	2009	300.0	19233.0	577.0	31657.0	282402.0	894000.0	2380.0	110941.0	...

5 rows × 25 columns

This action resulted in repetitive columns during each merge. The repetitive columns were removed from the data frame after each merge to avoid later confusion. Crop data frames were then condensed to area, item, and year information as well as the area harvested, yield, and production values in

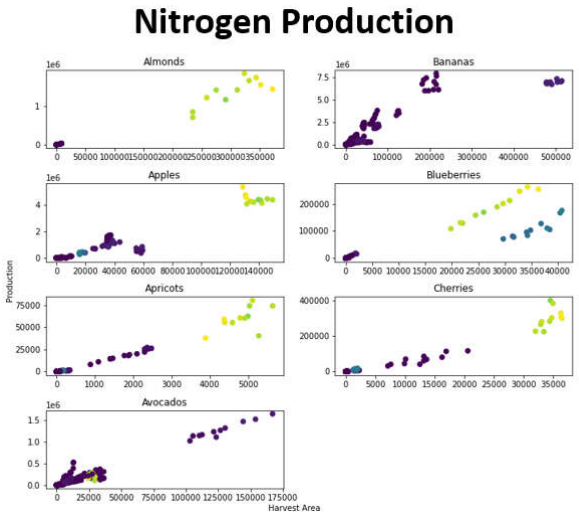
separate columns. The process was then repeated for all additional crop csv files. Nitrogen and Phosphate csv files having only a single element value did not require such extensive cleaning. For these pollutants, the appropriate columns were dropped or renamed to match with the crop data frame column names.

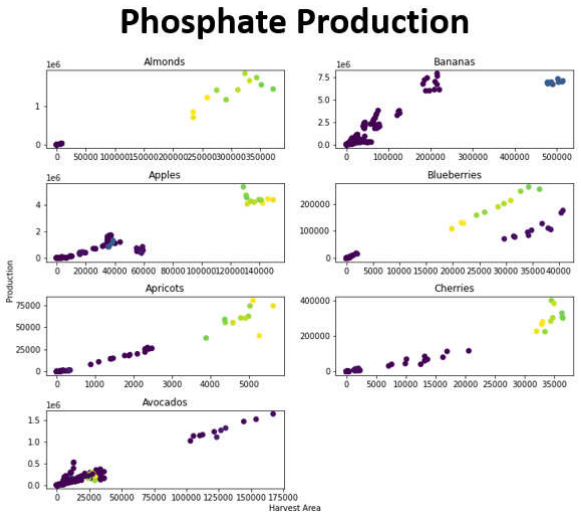
Exploratory Data Analysis:

Data exploration began once all crop information was combined into a single data frame. As not all countries listed were presumed to grow each crop null values were expected throughout the data frame. This assumption was confirmed on checking null values in the data with almonds, apricots, and blueberries being the least represented within the data frame. A total of 40 countries were kept after the merging of the various crops. The minimum and maximum values of nitrogen were documented, and a total list of countries used was recorded.

Analysis:

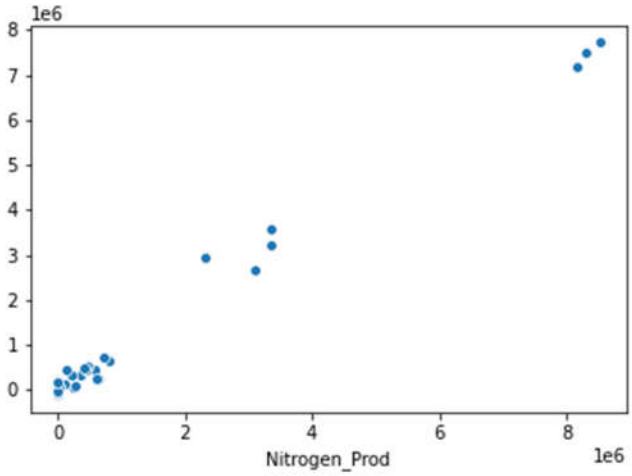
From inspection the data appears to have some linear correlation for the harvest area and production of each crop. In addition, each apparent linear correlation contains a noticeable gap between countries and years with low harvest area vs countries and years with high harvest area. Overlaying the nitrogen and phosphate production information on these graphs highlight reporting gaps in the data for both pollutants. Bananas have no nitrogen values reported and only phosphate values. Additional research on banana farming and production would be necessary to determine if nitrogen is included in banana growing and production.





Results:

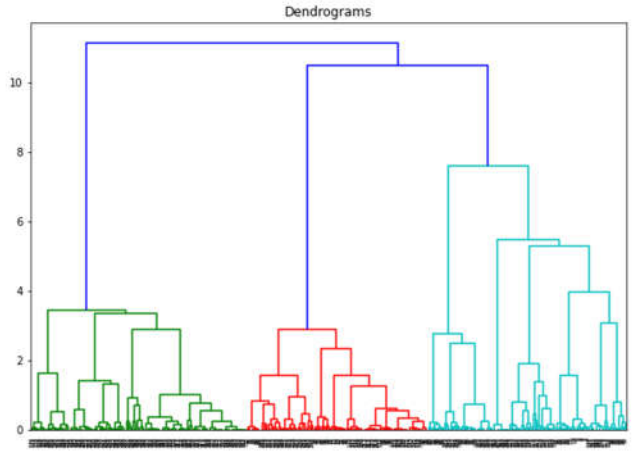
Processing the information through linear regression resulted in high values mean squared error as well as the coefficient of determination. High value in the coefficient of determination indicates that the data fits the linear regression well. The figure below contains the test values compared to the prediction values. Significant gaps between the values are observed in this result as in the initial data set.

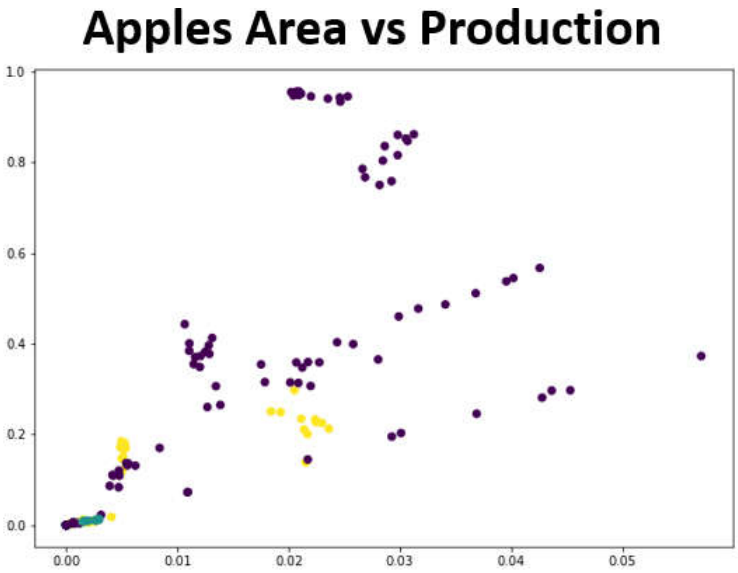


However, the high value mean squared error indicates a high level of error in the individual data points.

Mean Squared Error:	48538039276
Coefficient of determination:	0.98

Other methods attempted on the data set included principal component analysis, hierarchical grouping and DBSCAN. From these hierarchical groupings provided the best result when applied to the data set.





From the modeling results a linear regression provides the best methodology for determining the level of pollutant. It should be noted that this method maintains a high level of error for individual data points. The suspected source of high error is in the value range between the different crops. Normalization of the values or running individual linear regressions on each crop may result in lower error values.