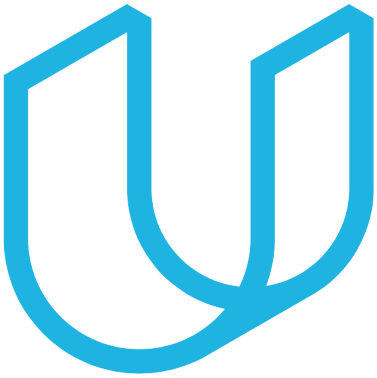
Capstone Project Report



Udacity Machine Learning Engineer Nanodegree

A capstone report for analyzing worldwide food production and population from 1961-2013.

Derek Helms

January 25^{th}, 2021

Table of Contents

I. Definition

a. Project Overview

The current world population is nearly 7.8 billion people, and this number is estimated to rise to around 9.7 billion in the year 2050. This means within the next 30 years, we will need to feed two billion more people without sacrificing the planet (1). This means we will need to double our crop production in order to feed that growing population. Agriculture is one of the greatest contributors of global warming, with farming consuming immense amounts of our water supplies and leaving major pollutants as its byproduct from fertilizer runoff (4).

Two different datasets were used in the analysis, one coming from Kaggle and one coming from FAO (the Food and Agriculture Organization of the United Nations).

The dataset obtained through Kaggle is a Food Balance Sheet that originated from FAO's database, but has been formatted for easy of use, with a total of 63 features and 21477 observations. Food Balance Sheets represent the pattern of a country's food supply during a period of time, in this case it is yearly from 1961 to 2013, and measured in 1000 tonnes (2)

	Area	Item	Element	Unit	Y1961	Y1962	Y1963	Y1964	Y1965	Y1966	...
0	Afghanistan	Wheat and products	Food	1000 tonnes	1928.0	1904.0	1666.0	1950.0	2001.0	1808.0	...
1	Afghanistan	Rice (Milled Equivalent)	Food	1000 tonnes	183.0	183.0	182.0	220.0	220.0	195.0	...
2	Afghanistan	Barley and products	Feed	1000 tonnes	76.0	76.0	76.0	76.0	76.0	75.0	...
3	Afghanistan	Barley and products	Food	1000 tonnes	237.0	237.0	237.0	238.0	238.0	237.0	...
4	Afghanistan	Maize and products	Feed	1000 tonnes	210.0	210.0	214.0	216.0	216.0	216.0	...

The dataset obtained through FAO is a time series dataset that contains the estimated/projected population (for both sexes) in each country from 1961 to 2018. The estimates are based on data from the World Population Prospects and World Urbanization Prospects (3).

	Domain	Area	Element	Item	Year	Unit	Value
0	Annual population	Afghanistan	Total Population - Both sexes	Population - Est. & Proj.	1961	1000 persons	9169.410
1	Annual population	Afghanistan	Total Population - Both sexes	Population - Est. & Proj.	1962	1000 persons	9351.441
2	Annual population	Afghanistan	Total Population - Both sexes	Population - Est. & Proj.	1963	1000 persons	9543.205
3	Annual population	Afghanistan	Total Population - Both sexes	Population - Est. & Proj.	1964	1000 persons	9744.781
4	Annual population	Afghanistan	Total Population - Both sexes	Population - Est. & Proj.	1965	1000 persons	9956.320

The use cases for each dataset is as follows. The Kaggle dataset will be used in analyzing food production, seeing which countries produce the most and if it is for human or livestock consumption. This data will also be used to cluster countries together based on their production levels, ideally separating high and low producing countries. The FAO dataset will be used to compare the global population to the yearly total food production, as well as further analyzing the highest producing countries to see if they have the largest/fastest increasing populations.

b. Problem Statement

With a continuously growing population this leads to the question, how do we supply the necessary amount of food for an increasing world population without sacrificing the climate of our planet? There are many solutions to this question, but the focus of this capstone will be on analyzing the global production of consumables, as well as the ratio of food (human consumption) to feed (livestock consumption) produced by each country. Only 55% of the current world crop production is consumed by humans, with the remaining being fed to livestock. On top of this, nearly 25% of the world's food calories are wasted before they can be consumed (4).

One aspect to explore is which food items are the most produced and the use case for those food items, whether they are for human or animal consumption. This will allow us to understand where a majority of our food is being directed, and as suggested by National Geographic, could a shift in diet lead to more crops being used for human consumption rather than livestock feed. Creating graphs to find insight will allow us to see how population change has correlated to food production, and what percentage of consumables are for humans compared to livestock.

The second aspect to explore is creating a linear model that can estimate the total global production needed for a given global population. This means correlating population and production in order to create a regression model that can predict the necessary production for the 2050 population. These will be compared to scientific articles that determine the necessary growth in production needed, and which model best fits these predictions.

The last aspect of the data to explore is separating the countries into clusters based on their yearly production. Analyzing which countries are responsible for feeding the majority of the population, and how to ensure they can continue to do so. From this, we can also see which countries are not highly producing, and how we can aid them in increasing production if possible.

c. Metrics

To evaluate our graphing and data exploration, we can use the FAO rankings for countries by commodity, which we can then compare to our graphs to see if we have the data correctly labeled. This ranking allows us to see the top 20 country's that produce the most for each food item, as well as the top 20 food items produced by each country from the years 1961 to 2019.

To evaluate our regression, we will use the root mean squared error (RMSE), in order to determine which line is best fitting our data. However, the 2050 estimate will also be used in determining the evaluation of the model in order to avoid overfitting the original data and having a poor future estimate (since we know the production needs to be anywhere for 50-100% increased).

With the data being unlabeled, it will be difficult to ensure that our clusters are properly partitioned. However, we can analyze both the clusters and graphs we create using two separate methods. One metric we can use to evaluate our model will be a silhouette score, which will measure on a scale from [-1,1] how close each point on one cluster is to points in the neighboring clusters. We will want to aim for a silhouette coefficient near +1 to indicate that our samples are far away from neighboring clusters (5).

To evaluate our graphing and data exploration, we can use the FAO rankings for countries by commodity, which we can then compare to our graphs to see if we have the data correctly labeled. This ranking allows us to see the top 20 country's that produce the most for each food item, as well as the top 20 food items produced by each country from the years 1961 to 2019.

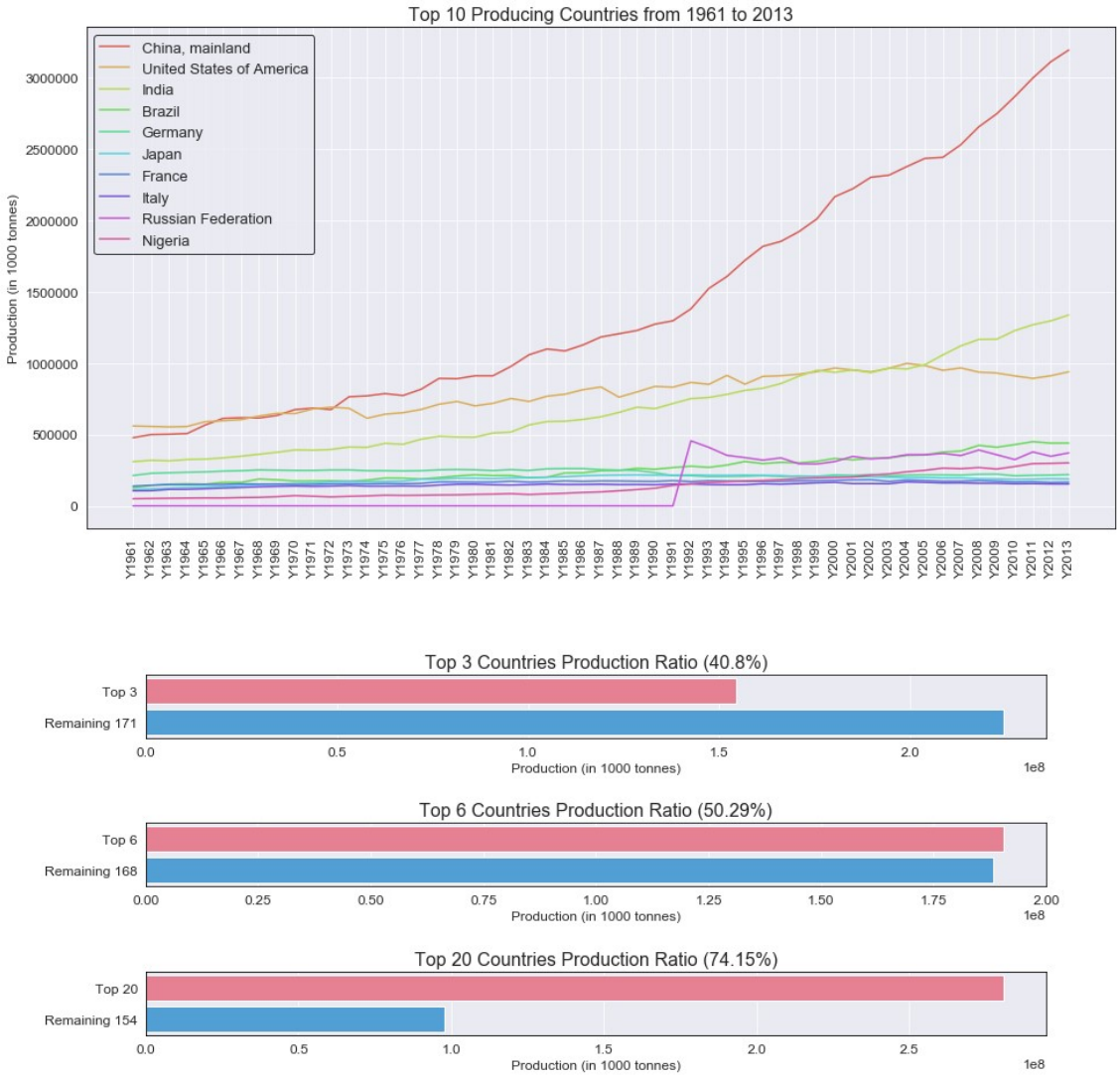
II. Analysis

This is where the majority of the project has taken place, in exploring the data and creating visualizations in order to determine where the food we are producing is going, as well as who is responsible for producing it. The analysis was broken down into two different sections: production and population exploration. In the production section, we analyzed the top producing countries, explored the top produced food items, and finally the use case for the food items (food vs feed). In the population section, we analyzed population vs production, the populations for the top producing countries, and the largest population changes over the last 53 years. A more in-depth description and visualization will be given in the following two subsections.

a. Production Exploration

Top 10 Producing Countries:

To begin the exploration, we wanted to graph the top 10 producing countries for global production from 1961 to 2013. After this, we compared total production for a given number of top countries to the rest of the world to see how much global production each are responsible for.

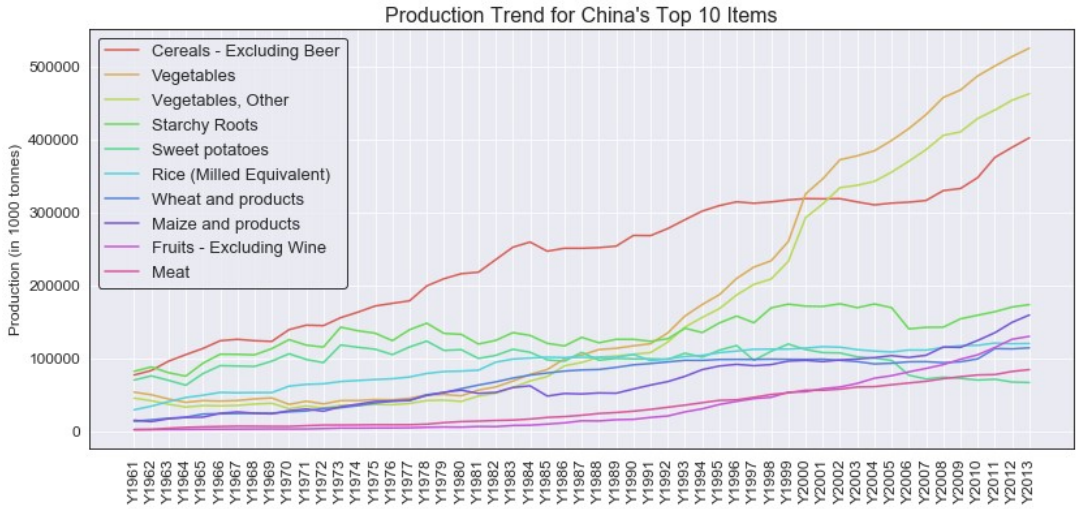


From the line graph, we can see that the top 3 countries have been the top producers over the last 53 years. With China becoming dominant in the year 1972, it looks to be the beginning of an exponential growth pattern. The United States was second to China for a majority of the data, but seems to have plateaued and been passed by India starting in the early 2000's. India, being third highest of the data, has begun growing quickly since the late 1980's and seems to be increasing similar to China. An interesting finding is the Russian Federation spiking from 0 to nearly 500,000,000 tonnes of production from 1991 to 1992.

This was due to the fact that the Russian Federation was found on December 25, 1991 and this was when our data begun with them (no production for Soviet Union was recorded in our data, which was the old name).

Looking at the bar charts, and adding the three top producing countries together, it seems that they are responsible for around 40% of the total global production. Going further, we can see that the top 6 producing countries are responsible for over 50% of the global production, and the top 20 being responsible for 74%. Out of 174 countries, it seems that only a small amount of them are responsible for the vast majority of production (however, these countries populations are also much larger than the others).

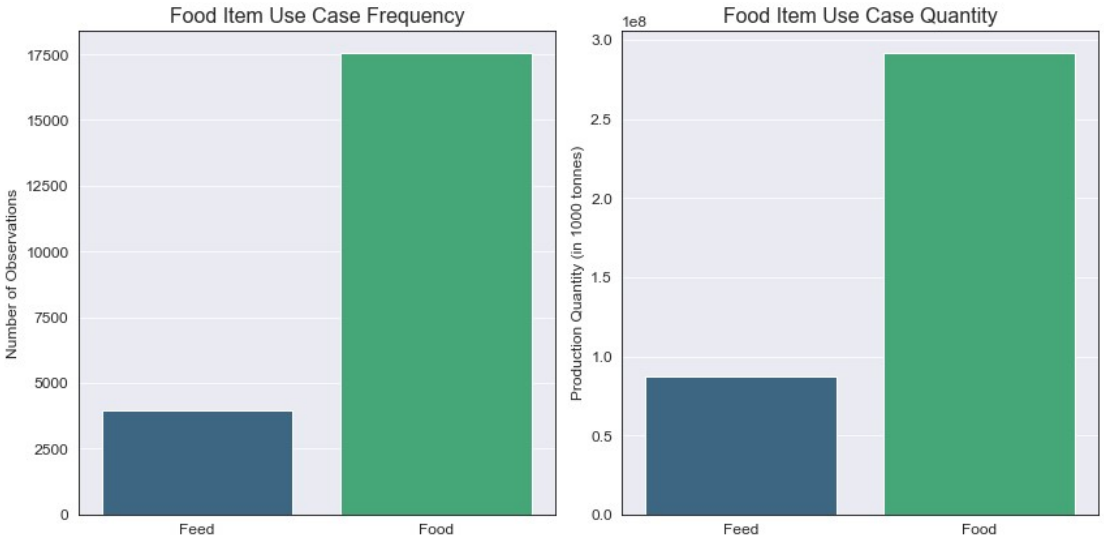
Exploring China's Production:
Being the top producer over the last 53 years (and most likely much before then), we want to further look into the production of the "China, mainland" area. We will want to see what products make up a majority of their production, as well as the yearly breakdown to see any trends in production.



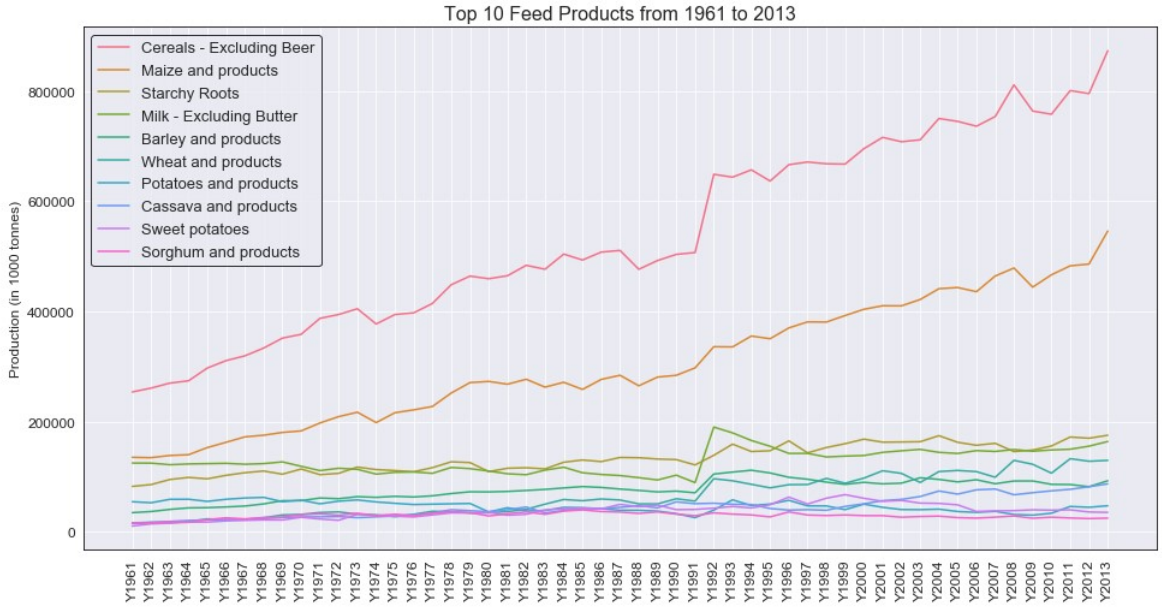
From the graphs, we see that historically *Cereals - Excluding Beer* has been the dominant product for China, but recently there has been a large spike in *Vegetables* and *Vegetables, Other* that has surpassed Cereals starting in the year 2000. However, Cereals seem to be rising again and could become the most produced item if Vegetables plateau. This is important to keep in mind for the next section of the exploration, where we looked at the top produced foot items.

Top Produced Items:

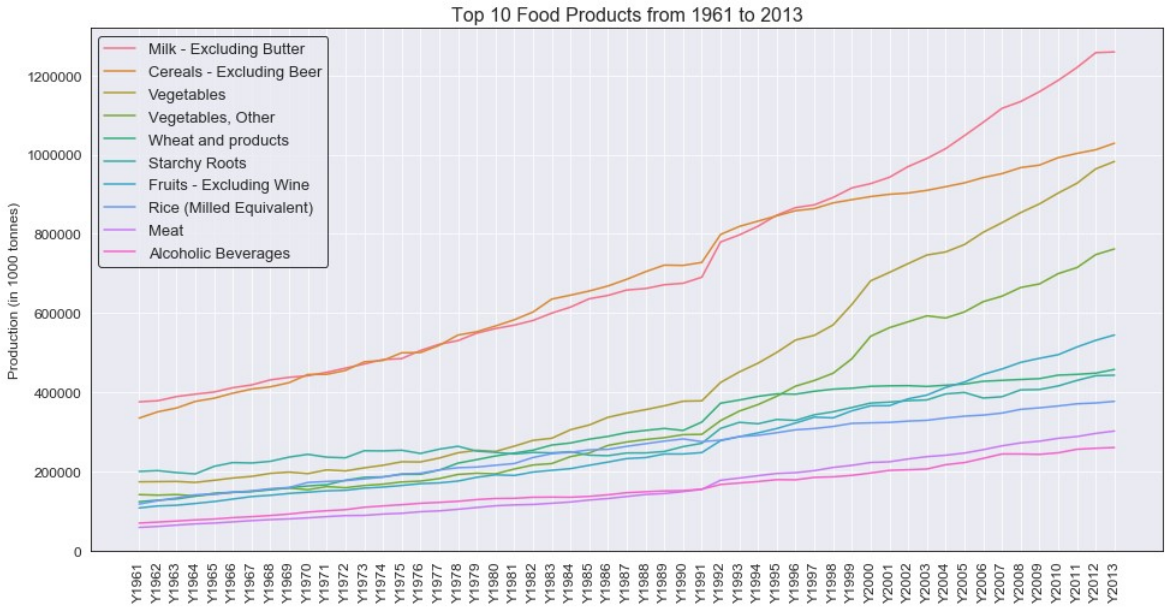
Initially, we wanted to separate the food and feed items from each other and create new datasets for both, in order to see how much is being produced for each use case. Although not all graphs can be included (a more in depth exploration and description can be seen in section 3.3 of the *DataExploration* notebook), we will cover the main findings from the section.



We can see that a majority of the food is being used for human consumption, which is a good finding. There is around 3x more food (by weight) compared to feed, as well the dataset for food being around 4x larger than the feed dataset.



For feed, the top product is "Cereals - Excluding Beer" by a landslide, with "Maize and products" coming in second (but at half the total production). Although no visual is included, there was an analysis to see which countries were the top producers for the top 2 items. It seems that the United States is clearly producing the most items for feed, with China being second to them. They produce the majority of the total, which is expected since they were both in the top 3 global producers.



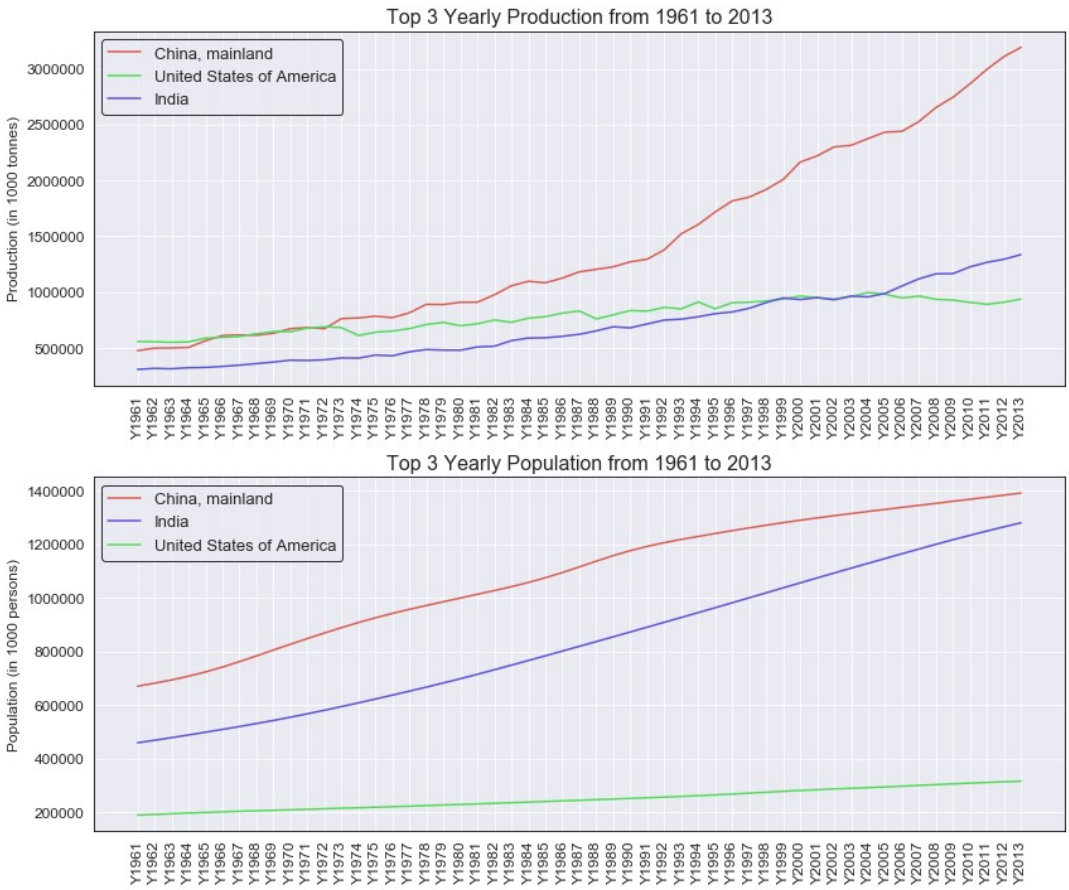
For food, there seemed to be a much closer grouping for production, which "Milk - Excluding Butter" and "Cereals - Excluding Beer" being the top 2 throughout most of the dataset, but "Vegetables" and "Vegetables - Other" seeming to be gaining traction and increasing quickly since the early 1990s. Again, an analysis was done to see who was the top producers for the food items, and it seems that China, India, and the United States were the main leaders for all of them (which is not surprising given our previous findings).

b. Population Exploration

Now that we understand the production, we also want to analyze the population for the countries. We will focus mainly on the top producing countries and their respective population, but before that we want to see any trends between population and production.

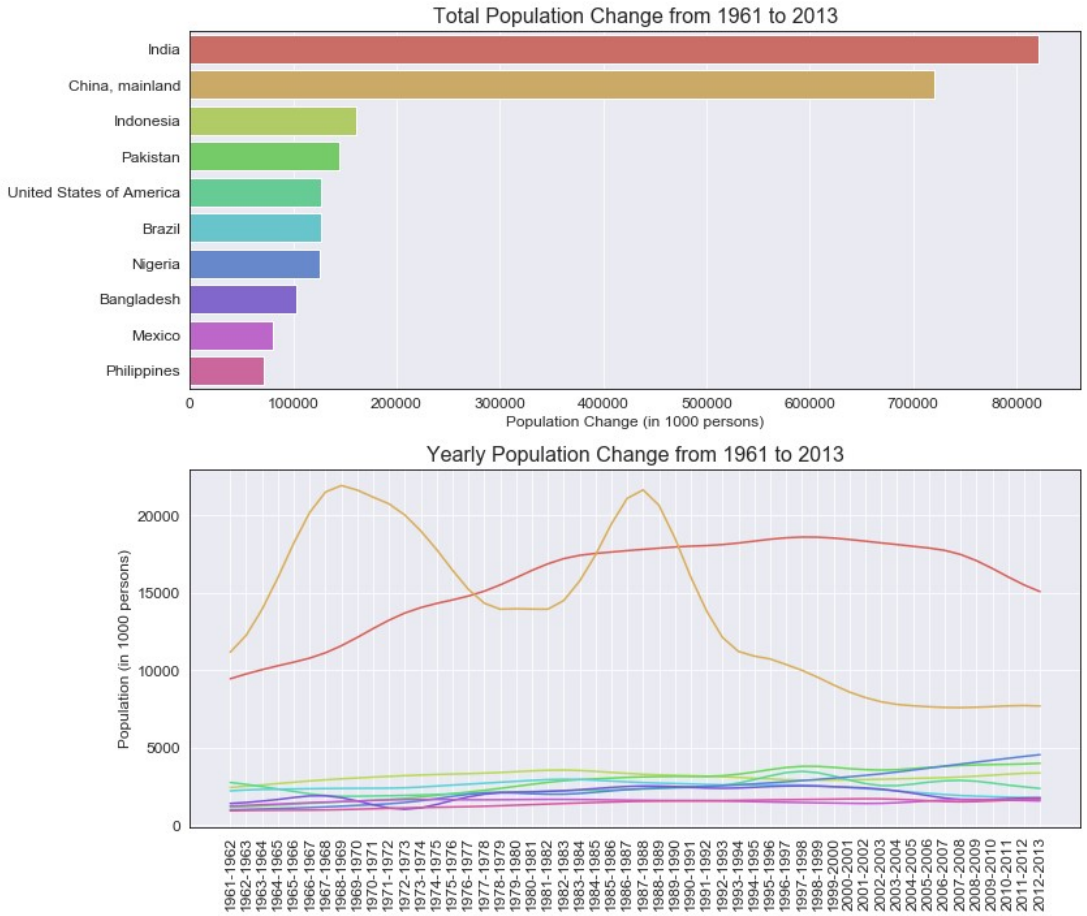
To visualize these depictions, see section *4.1: Population compared to Production* in the *DataExploration* notebook. It seems that population is growing in a linear pattern, with population more than doubling in the last 53 years (1961-2013).

Also, production and population seem the have a linear relationship (although not a completely straight line), which will be useful in the modeling section where we talk about the regression model we built. Now, we will look at the population compared to production for the top 3 countries (since they are responsible for over 40% of the global production).



Looking at the production graph, the United States was the leader in yearly production, but they were quickly passed by China who began to grow rapidly in the 1970's. It also seems that since the early 1980's, India has been increasing and is following a similar pattern as China did in in the late 1960's. Looking at the population graph, we can see that over the past 53 years China has nearly doubled, India has grown almost 3x, and the United States has only grown around 1.5x. This could explain how India has passed the United States in production since the mid 2000's, with the US plateauing in production starting around 2000.

Next, we wanted to see the yearly change for the most increased populations as well as the total population change over the last 53 years. A quick note: the yearly population is the increase from the previous year to the current year, so population is not dropping for negative slope but rather less people are being born or moving to certain countries.



Looking at our bar graph, we can see that again India and China have had the most people either born or immigrated into their countries by a landslide. The United States seems to be around 5th in terms of growth, with around less than 1/5 of the growth as China or India. We can see that India has had quite consistent growth, with a dropping increase in new people. China has fluctuated throughout the years, seeing large growth and small growth, but since the early 1990's it seems there is a decreasing number of new people each year. The remaining countries all have consistent growth, with around 1-5 million new people each year for the most part (including the United States).

Concluding Thoughts:
For this exploration, it is clear that China, India, and the United States are the top producing countries (but also the top populated countries). With around 40% of the global production, it is crucial than we can continue to support these countries in order to ensure survival with an increasing need for food. We also saw that the majority of production is for humans and not livestock, which leads me to think that a shift in diet would not be necessary to supply more food globally.

III. Methodology

a. Data Preprocessing

Two main things were done in the data preprocessing step: reformatting the population data to be similar to the production data layout (years as feature and Area as observations) and dropping China from the production dataset.

Previously, the population data had each year being an observation for each country, but we wanted to have each year be a column and each country be an observation so that we could easily sum across the columns to get total yearly population. A function was written to do this for us (in the *function.py* python module called *format_population_data*) which followed the same layout as the production dataset and filled any missing years with *NaN* values. Below is the original population format, followed by the new format:

	Area	Year	Unit	Value
0	Afghanistan	1961	1000 persons	9169.410
1	Afghanistan	1962	1000 persons	9351.441
2	Afghanistan	1963	1000 persons	9543.205
3	Afghanistan	1964	1000 persons	9744.781
4	Afghanistan	1965	1000 persons	9956.320

	Area	Unit	Y1961	Y1962	Y1963	Y1964	Y1965	Y1966	Y1967	Y1968	...
0	Afghanistan	1000 persons	9169.410	9351.441	9543.205	9744.781	9956.320	10174.836	10399.926	10637.063	...
1	Albania	1000 persons	1685.936	1737.686	1790.573	1843.634	1896.171	1947.830	1998.740	2049.210	...
2	Algeria	1000 persons	11336.339	11619.828	11912.803	12221.675	12550.885	12902.627	13275.026	13663.583	...
3	American Samoa	1000 persons	20.602	21.253	22.034	22.854	23.672	24.462	25.248	25.989	...
4	Andorra	1000 persons	14.375	15.370	16.412	17.469	18.549	19.647	20.758	21.890	...

The next step we took was removing the general China area from the population dataset. This was because it was being double counted, since China was broken down into its regions (whose sum added up to the total China population). Also, the general area China is not in the production dataset (only its regions), so removing this was a key step in ensuring the population data was accurate.

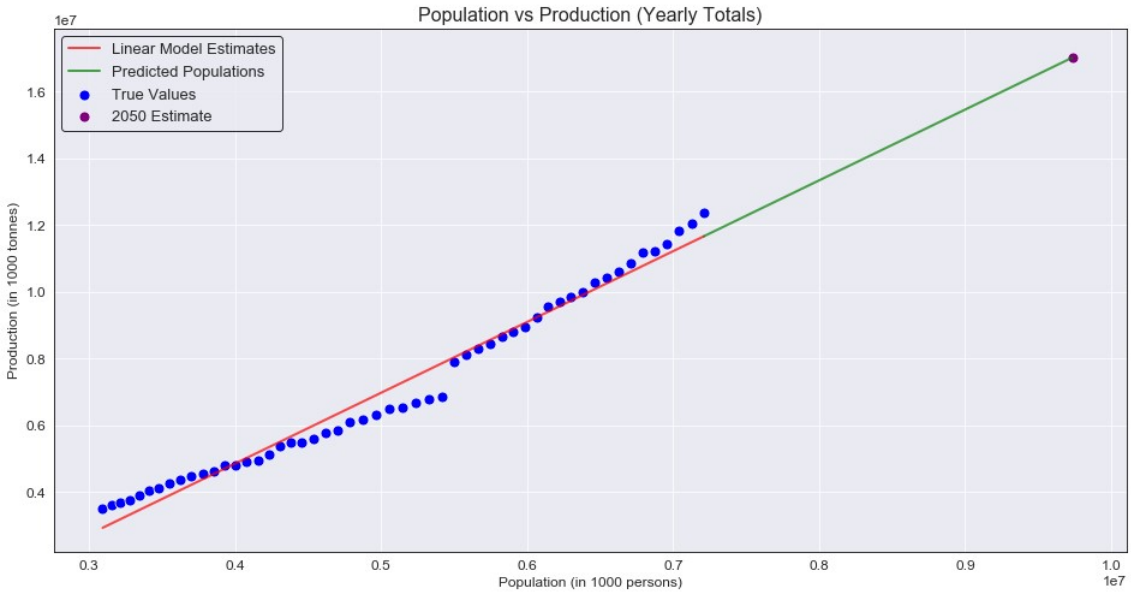
Note that removing China was built into the format population function, and the new formatted population dataset was outputted to the file "FAO_POP_REFORMAT.csv". This was all done in the *DataExploration* notebook in section 2.

b. Implementation

Regression Models:

From our findings in the previous notebook, we noticed that yearly global population and yearly global production have a linear relationship. Using the estimated global population as the independent variable (y) and the global production as the dependent variable (x), we will estimate the total production that is expected given a total population value. This will allow us to see what the necessary production would be to support a given population.

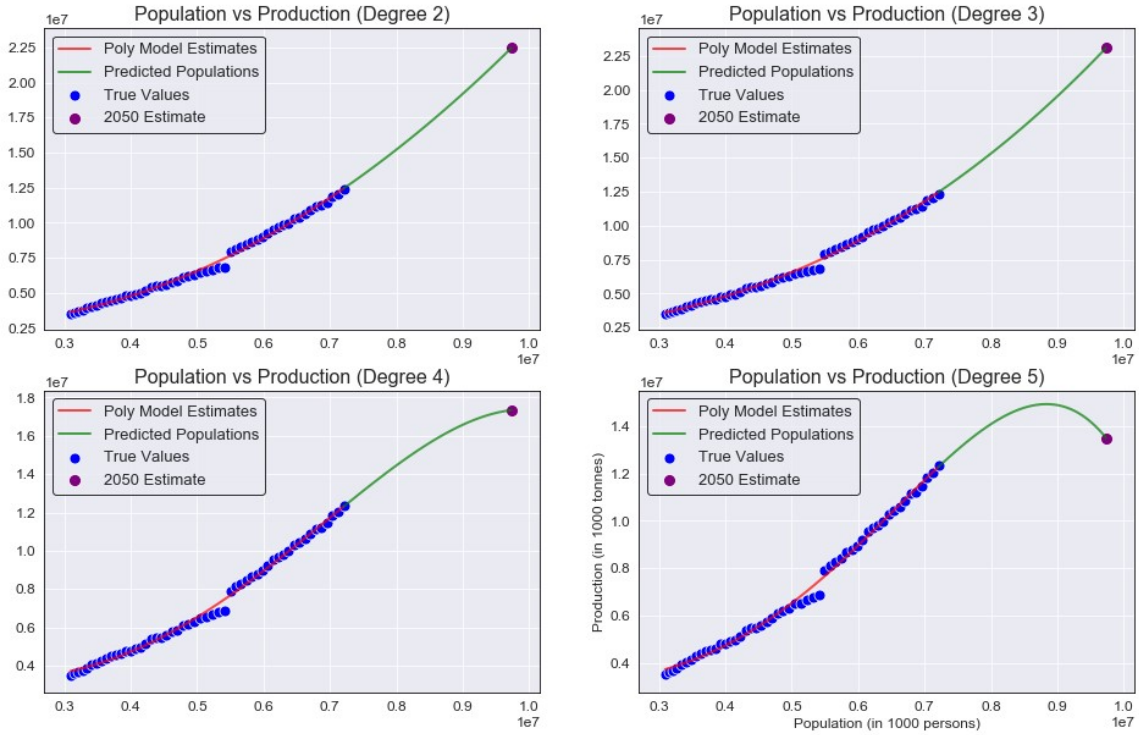
To begin, we want to create a base model using simple linear regression. This will allow us to see how the initial model performs, and how we can improve upon it.



As we can see, the simple linear model does not fit our data very well, with an RMSE of around 405,000. However, its important to note that the data uses numbers in the millions, so the error was expected to be large. With an estimated production amount of 17013250 (in 1000 tonnes), this does not seem very accurate, since most experts say that we need to increase our production anywhere for 50-100% to support to 2050 population. This was a good base model to indicate a base RMSE, but we will try to improve this model. A solution to this problem would be to fit a polynomial regression model that allows the line to bend and is better suited for the data. To do this, we will want to find out which degree minimizes the RMSE for the polynomial model, then use a subset of those to graph the predictions and see which model best represents our data.

By graphing the RMSE and degree, we saw that a degree from 2-5 minimizes the RMSE, so we will want to create a polynomial model for each of these and compare.

	Degree	RMSE	2050 Production Estimation
0	2	179427.41	22463820.97
1	3	182169.96	23084034.55
2	4	171969.60	17341906.93
3	5	169272.95	13500278.99



From the World Resources Institute, we know that we need an increase of around 56% in production from 2010 to 2050. The 2010 global production, from our dataset, was 11,445,072 (in 1000 tonnes). Following this idea, that means we need to produce 17,854,312.32 (in 1000 tonnes) total for the 2050 population. This seems to follow the degree 4 polynomial model, which estimated a production of 17,341,906.93 (in 1000 tonnes).

However, from the United Nations, we have also inferred that global food production must double in order to support the global population. Given that the article was written in 2009, with a global production of 11,211,891 (in 1000 tonnes) for out dataset, this means we need to produce 22,423,782 (in 1000 tonnes) for the year 2050. This idea follows the degree 2 model (almost exactly) with an estimated production of 22,463,820.97 (in 1000 tonnes).

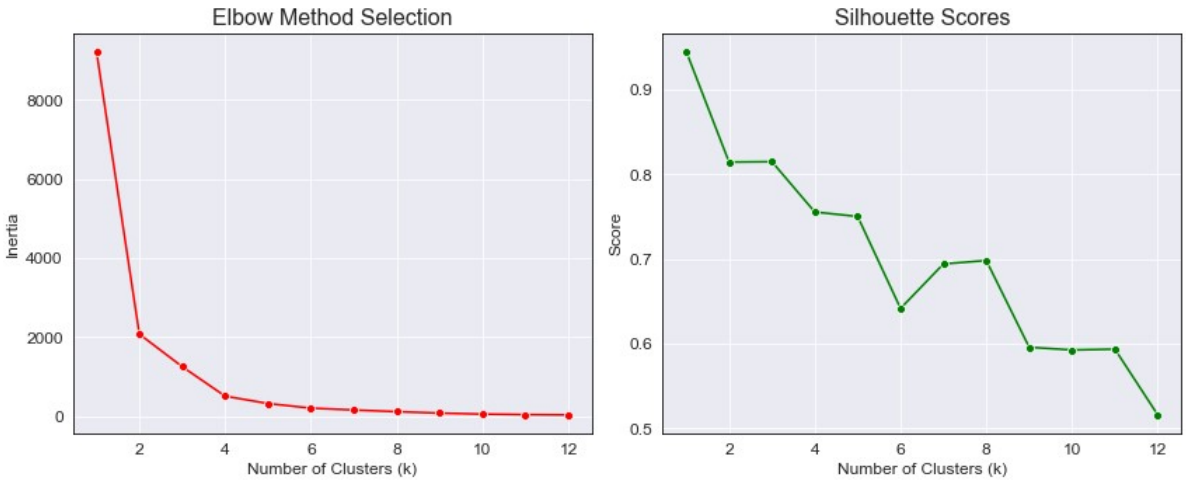
Despite not being the lowest RMSE for the models, these 2 most accurately depict what has been estimated by professionals. Choosing which model is correct will be more of a challenge, since there is no real way to determine what the actual growth will be in the future (global events, pandemics, food supply, etc.). However, these two models are good indicators of a possible future that we could see.

Clustering Model:

Next, we wanted to create a KMeans model to cluster countries together based on their production. We could have left the data how it was and use it for the clustering algorithm, but this resulted in extremely large inertia values (used below). To counter this, I used a z-score to standardize the data, which now is in the range [-0.322, 11.41], and our inertia is much more easily interpreted.

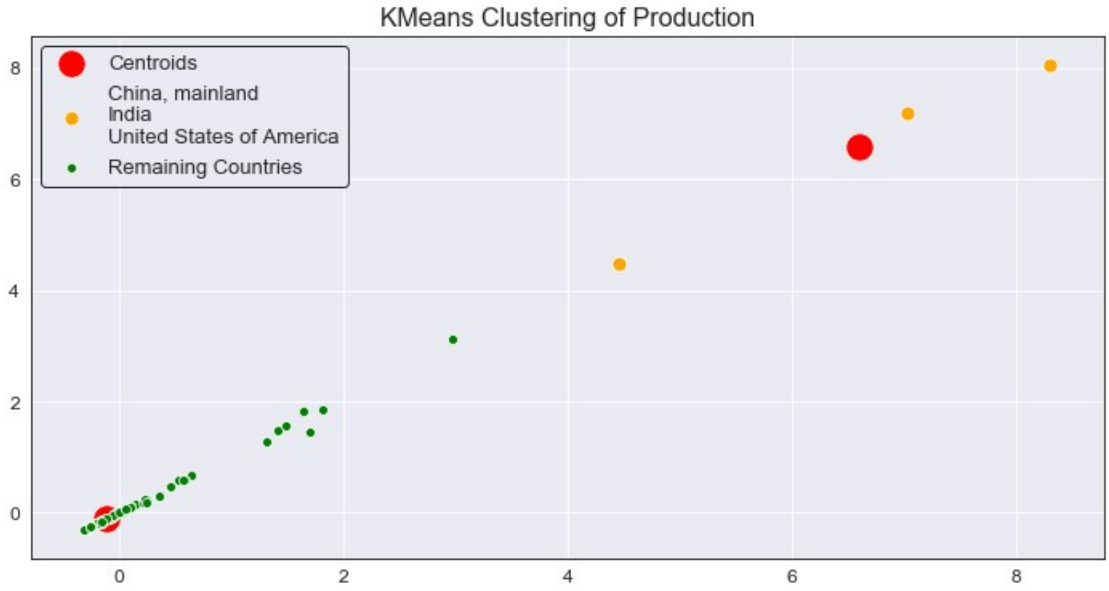
Area	Y1961	Y1962	Y1963	Y1964	Y1965	Y1966	Y1967	Y1968	Y1969	Y1970	...
Afghanistan	-0.163861	-0.169612	-0.179311	-0.169651	-0.166441	-0.175412	-0.164236	-0.167893	-0.167506	-0.187094	...
Albania	-0.283499	-0.284815	-0.290403	-0.292117	-0.286017	-0.284625	-0.286017	-0.287853	-0.287550	-0.285347	...
Algeria	-0.194528	-0.202362	-0.214208	-0.212761	-0.207706	-0.210592	-0.207755	-0.202177	-0.202724	-0.200493	...
Angola	-0.235367	-0.239335	-0.238455	-0.241880	-0.235299	-0.235430	-0.236122	-0.242690	-0.237591	-0.235788	...
Antigua and Barbuda	-0.308334	-0.309689	-0.315263	-0.318648	-0.311076	-0.310305	-0.312131	-0.314956	-0.314628	-0.313608	...

We needed a way to select the number of clusters, which we did by using the elbow method as well as silhouette scores. We want to choose k where the inertia has the steepest slope, but the silhouette score that is closest to 1.



Using the elbow method for the inertia graph, it seems that 2 is the ideal cluster number (but 3 and 4 could also be considered). We also used a silhouette score to determine help determine the number of clusters, which helps determine the distance between the resulting clusters, and we want a value of as close to 1 as possible. Using both of there, it seems that 2 clusters is ideal since this is where the elbow occurs and also has the highest silhouette score (besides 1 which is not considered in this case).

Finally, we created the model to group the countries together based on their production using 2 clusters. This was the initial guess we had, where we would separate the top 3 countries from the remaining since they had a much higher production compared to all the other countries.



As expected, using 2 clusters separated the top 3 producers from the remaining countries. This means the clusters are based on high vs low production, since from the *DataExploration* notebook we saw that the top 3 countries were responsible for over 40% of the global production. The next closest country (Brazil) produced less than 1/3 the amount that the United States or India does and around 1/6 the amount of China. These clusters seem to be accurate in separating our data in correct clusters, and helps us visual how much more the top 3 countries produced compared to the remaining countries.

IV. Results

In conclusion, from our findings in the *DataExploration* notebook, we saw that the top 3 countries are responsible for over 40% of the global production. These included China, India, and the United States, who were all the most densely populated areas and accounted for around 1/3 of the total global population. We saw that the majority of production and accessible goods were for human consumption rather than livestock, which hinted at the idea that a changed of diet might not be necessary to support an increasing population. The main crops for both food and feed has been consistent throughout the dataset, but since the early 1990s there has been a large increase in the top 3 items both both consumable type (most likely due to the increasing population). We saw that China and India have the largest population increases (both yearly and over the past 53 years) by a landslide, with most other top countries having less than half the growth.

Continuous support for the success of the top 3 countries (and all others as well) will be crucial in ensuring that we can meet global production necessary for the 2050 population. To estimate this need, let's discuss the concluding findings from the regression model as well as the clustering algorithm.

The findings from the regression models indicated that one of two scenarios will occur: either a increase of around 56% in global production or an increase of 100% in order to meet the needs of 9.8 billion people in 2050. For the first scenario, the degree 4 polynomial regression model correctly mapped this outcome, estimating the needed production within 500,000 (in 1000 tonnes) based on the World Resources Institute findings. The second scenario was mapped by the degree 2 polynomial regression model within 40,000 (in 1000 tonnes) units of the United Nations estimated value. Choosing between the models depends on population demand, and will need a professional to determine which correctly estimates the current trends, but successfully works as a solution to our question: what is the necessary global production for the 2050 population?

Finally, the findings from the clustering algorithm closely followed what we had estimated in the beginning. We had believed that dividing the countries based on their production would result in high vs low, which was correct since the model grouped the top 3 producers together, and the remaining 171 in another cluster. Using the elbow method and silhouette scored, we maximized the score and chose the correct number of cluster than fit our model. Graphing our findings, we saw that the centroids worked well for the given model, and that there were tight grouping with the lower producing countries compared to the higher ones.

Overall, the project was a success. It was eye opening to see how machine learning and sustainability could be intertwined, and used in estimating the survival of the future. This project has sparked a new drive within me, and will lead me to continue my exploration of this topic outside of this project.

V. Resources

[1] Oppenheim, Dor. "Who Eats the Food We Grow?" *Kaggle*, 30 Nov. 2017, www.kaggle.com/dorbicycle/world-foodfeed-production.

[2] "Food Balances (Old Methodology and Population)." *FAOSTAT*, Food and Agriculture Organization of the United Nations, 12 Dec. 2017, www.fao.org/faostat/en/#data/FBSH.

[3] "Annual Population." *FAOSTAT*, Food and Agriculture Organization of the United Nations, 16 Dec. 2019, www.fao.org/faostat/en/#data/OA.

[4] Foley, Jonathan. "A Five-Step Plan to Feed the World." *Feeding 9 Billion - National Geographic*, www.nationalgeographic.com/foodfeatures/feeding-9-billion/.

[5] "Selecting the Number of Clusters with Silhouette Analysis on KMeans Clustering." *Scikit*, scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html.

[6] Kiersz, Andy. "The World Could Have Another Billion People in Thirteen Years." *Business Insider*, Business Insider, 29 July 2015, www.businessinsider.com/un-world-population-projections-2015-7.