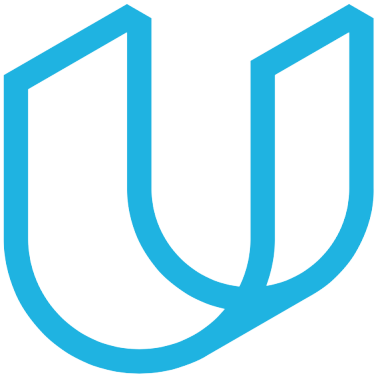
Capstone Project Proposal



Udacity Machine Learning Engineer Nanodegree

A capstone proposal for analyzing worldwide food production from the years 1961-2013.

Derek Helms

January 13^{th}, 2021

Table of Contents

I. Domain Background

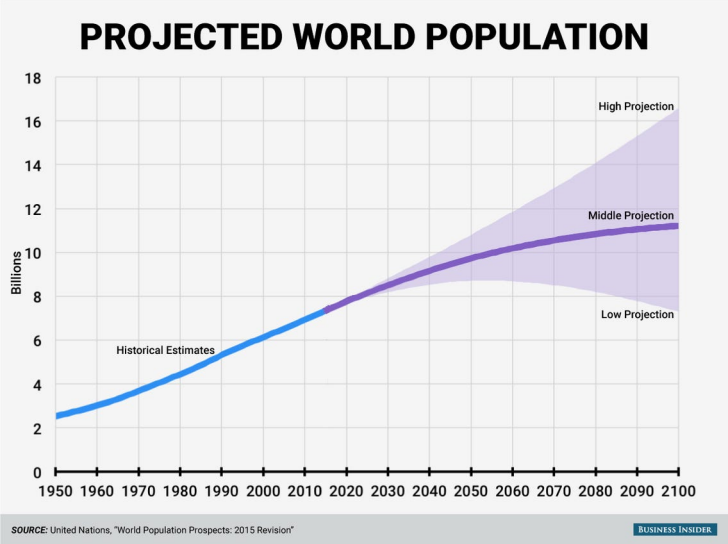


Image obtained from source (6).

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). So how do we increase food supply without destroying our environment?

The personal motivation for investigating this problem is my interest in applying machine learning to the field of agriculture. With an increasing need for sustainability, applying machine learning techniques to predict yield amounts or for detecting crop diseases has been a key factor in my motivation behind studying machine learning.

II. 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).

III. Datasets and Inputs

Two different datasets will be 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, labeled *FAO_FOOD_STAT.csv*, 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). Important features that this file contains are the country, item, element, and yearly production:

	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	...

- Area: 174 unique country names from around the world.
- Item: 115 unique food items being globally produced.
- Element: food - human consumable food available at a given time.
  feed - livestock consumable food available at a given time.
- Y1961 - Y2013: item available at time of measurement per year.

The dataset obtained through FAO, labeled *FAO_POP.csv*, 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). An important note is that the population dataset contains countries and later years that are not included in the Food Balance Sheet dataset, but will be left in to account for global population and future estimates. Some important features to note within the dataset:

	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

- Area: 245 unique country names (71 more than first dataset).
- Year: spans from 1961 to 2018 (5 years more than first dataset).
- Unit: population estimations are counted in units of 1000 persons.
- Value: estimate population count for each year.

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.

IV. Solution Statement

As initially stated above, the problem we face is how do we produce the necessary amount of food for an exponentially increasing population without sacrificing our own planet.

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.

Another 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.

V. Benchmark Model

In general, the majority of this capstone will be in exploring the data and finding insights about food production through graphing and statistical methods. Creating plots for top producing country's, top produced food items, and population changes will be the main focus of the work. A large part of this will be done using Pandas, Matplotlib, and Seaborn in order to create visuals that depict trends within the two datasets.

We will also create a KMeans clustering model to group our data based on food production, partitioning the 174 original country's into k clusters. We will want to use the elbow method in selecting our number of clusters, with an initial guess of using 2 to 3 cluster to separate countries based on high and low production (and possibly medium).

Another idea to explore will be to create a linear model that can estimate the total food production for a given year, with total population being the independent variable and total food production being the dependent varaible. This will allow us to see what the necessary production would be to support a given population number (test using the 2050 population estimate of 9.7 million).

VI. Evaluation Metrics

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.

VII. Project Design

1. Obtaining Data and Loading:
We will need to download data from two different sources, ensure both csv files are UTF-8 encoded, and important them to the workspace.

2. Data Exploration:
This is where a vast majority of the work will take place. This step includes creating graphs from subsections and groupings of the data, as well as comparing the two datasets to find insight about population and production correlation.

3. Data Preprocessing:
We will want to copy the original Food Balance Sheet dataset and format it in a way that can be used in our clustering algorithm. We will want to cluster based on the yearly production for each country, so dropping and filling columns necessary for that will be done in this step.

4. Modeling:
Here is where we will define a KMeans algorithm that will be used on the data created from step 3. We will use the elbow method in selecting our k value, as well as exploring the results from our model. A possible subsection for this step would to be to explore predicting the food production for future years based on population and total yearly production using a linear model.

5. Conclusion:
This is where we will recap that major finding throughout the notebook and correlate it to real world problems. Any insight about the future or predictions will also be given here.

VIII. 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.