**FOOD REQUIREMENT ANALYSIS IN AN AREA**

**Using K-Means Clustering**

**Developed by: Bindu D.P, Kavita, Pooja,**

**Sanskriti Kumar**

**Smart Bridge-Remote Summer Internship Program**

**1. INTRODUCTION**

There is no single way to measure food security, the concept itself being rather elusive. Analysis of food security by WFP generally uses food consumption as the entry point. Food consumption measured in kilocalories is the gold standard for measuring consumption, and often considered to be one of the gold standards for food security- but the collection of detailed food intake data is difficult and time consuming.

WFP’s goal is to have a standard food consumption data collection instrument and analysis approach that is flexible enough for different needs and contexts, while standard enough to have equally applicable analysis techniques and equally interpretable results, and also one that can be implemented in the field in a reasonable data collection and analysis timeframe. There are several alternative ways to collect and analyze food consumption information using indicators that are proxy for actual caloric intake and diet quality.

Such proxies generally include information on dietary diversity, sometimes with the addition of food frequency. WFP has adopted this data collection tool-measuring dietary diversity and food frequency - because several different indicators built on this sort of data have proven to be strong proxies for food intake and food security.

Analysis of dietary diversity and food frequency can be done in several ways, each with its own specific aims - looking at consumption from different angles, and with different strengths and weaknesses. Building composite scores which measure food frequency and/or dietary diversity is one of the more explored and tested methodologies. Well defined examples include the FANTA dietary diversity score and the DHS Food groups indicator. There are several other indicators found throughout the literature.

WFP has taken a direction of food consumption measurement tailored to its own information needs. To further harmonize WFP’s data analysis, standard methodologies have been introduced to analyze this food consumption data.

**1.1 OVERVIEW**

**Clustering** is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific. Clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples. We’ll cover here clustering based on features. Clustering is used in market segmentation; where we try to find customers that are similar to each other whether in terms of behaviours or attributes, image segmentation/compression; where we try to group similar regions together, document clustering based on topics, etc. Unlike supervised learning, clustering is considered an unsupervised learning method since we don’t have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. We only want to try to investigate the structure of the data by grouping the data points into distinct subgroups.

In this post, we will cover only **Means** which is considered as one of the most used clustering algorithms due to its simplicity. Our dataset was meticulously gathered, organized and published by the Food and Agriculture Organization of the United Nations. The main aim of this project is to perform in-depth EDA of the data and figure out the annual produce of different countries, analyse lead producers of food. We also perform clustering algorithms to classify countries based on productivity scale. Finally, it will be integrated to web based application .

**1.2 PURPOSE**

Our aim from the project is to make use of TensorFlow, scikit and flask libraries from python to extract the libraries for machine learning for the food requirement analysis. Then, to make out which country is the highest producer , for that we have used K-Means clustering method.

**2.LITERATURE SURVEY**

K-Means clustering is a fast, robust, and simple algorithm that gives reliable results when data sets are distinct or well separated from each other in a linear fashion. It is best used when the number of cluster centers, is specified due to a well-defined list of types shown in the data.

**2.1 Existing problem**

In general, Our world population is expected to grow from 7.3 billion today to 9.7 billion by 2050. Finding solutions for feeding the growing world population has become a hot topic for food and agriculture organizations, entrepreneurs and philanthropists. These solutions range from changing the way we grow our food to changing the way we eat. To make things harder, the world's climate is changing and it is both affecting and affected by the way we grow our food – agriculture. Hence, it is necessary that we analyze the food production and act faster rather than repenting later.

**2.2 Proposed Solution**

The main objective of this project is to help the agriculturist who are suffering from the crop loss and also those who are facing the food shortage in their particular area. The main aim of this project is to perform in-depth EDA of the data and figure out the annual produce of different countries, analyse lead producers of food. We also perform clustering algorithms to classify countries based on productivity scale.

**3. THEORETICAL ANALYSIS**

Cluster analysis is a generic name for a wide range of exploratory multivariate statistical procedures which aim at identifying homogeneous groups within a data set. This chapter focuses on the three most commonly used techniques in food science and technology: hierarchical cluster analysis; k-means and fuzzy clustering. It then discusses the theory behind the methods and presents examples of application and implementation in R free statistical software. The most popular type of cluster analysis in food science and technology is agglomerative hierarchical cluster analysis, which aims at identifying a series of clusters within a nested structure. K-means is the most popular partitioning clustering method. It aims at partitioning the data set into k different clusters of greatest possible difference between them.

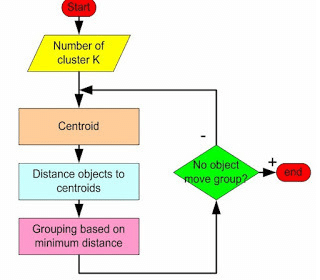
**Libraries Used:**

* **Numpy:** Numpy stands for ‘Numerical Python’ or ‘Numeric Python’.

 It is an open source module of Python which provides fast mathematical computation on arrays and matrices.

* **Pandas:**  Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools.
* **Matplotlib : matplotlib**.**pyplot** is a collection of command style functions that make **matplotlib** work like MATLAB.
* **Seaborn: Seaborn** is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics

**3.1 FLOW CHART**



**3.2 SOFTWARE DESIGNING**

● Jupyter Notebook Environment

● Spyder Ide

●Machine Learning Algorithm (K-Means Clustering)

● Python

● HTML

● Flask

We developed the Food requirement analysis by using the Python language which is an interpreted and high level programming language and using the Machine Learning algorithms. For coding we used the Jupyter Notebook environment of the Anaconda distributions and the Spyder, it is an integrated scientific programming in the python language. For food requirement analysis we used the Flask. It is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions, and a scripting language to create a web page is HTML by creating the templates to use in the functions of the Flask and HTML.

**4.EXPERIMENTAL ANALYSIS**

In this project, the dataset we downloaded is from some dataset available on kaggle. It contains Area code - Country name abbreviation Area - County name Item - Food item Element - Food or Feed Latitude - geographic coordinate that specifies the north–south position of a point on the Earth's surface Longitude - geographic coordinate that specifies the east-west position of a point on the Earth's surface Production per year - Amount of food item produced in 1000 tonnes.

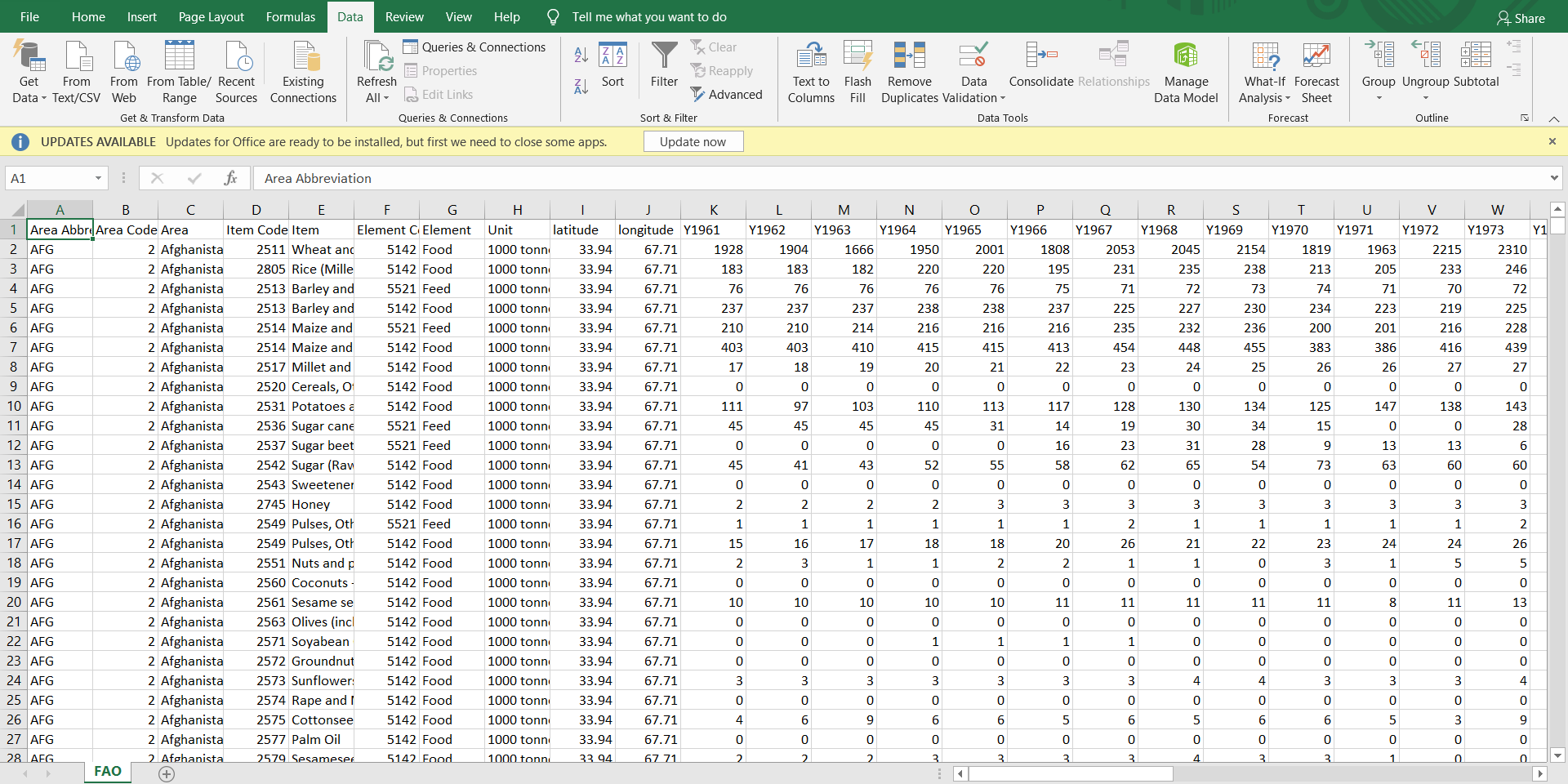
****

Fig 1. Dataset FAO.csv

**5.RESULT**

We have imported libraries and also the dataset. Then we check for the null values. If there are null values , we will drop those null values by using the command. Then we have Plot for annual produce of different countries with quantity in y-axis and years in x-axis.

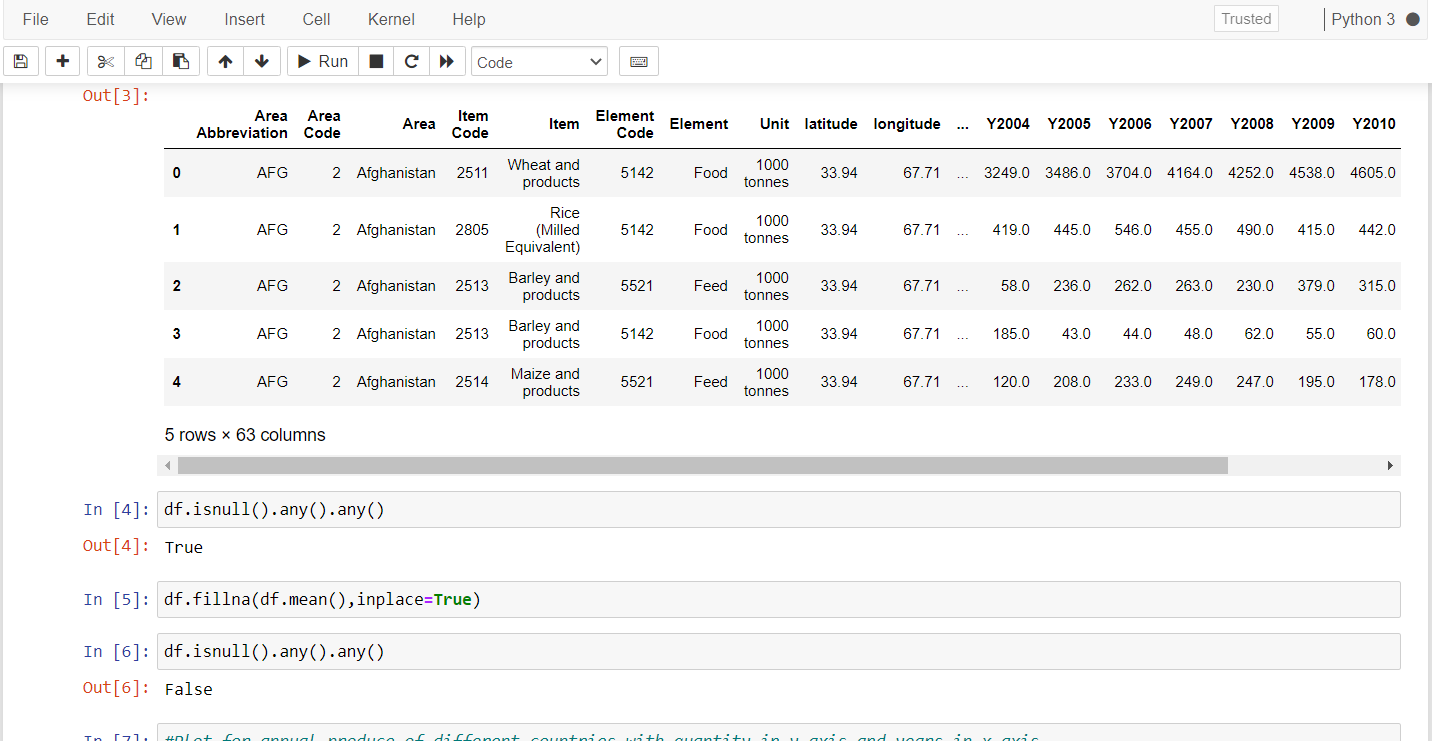


Fig 2. Command to drop the null values in the dataset

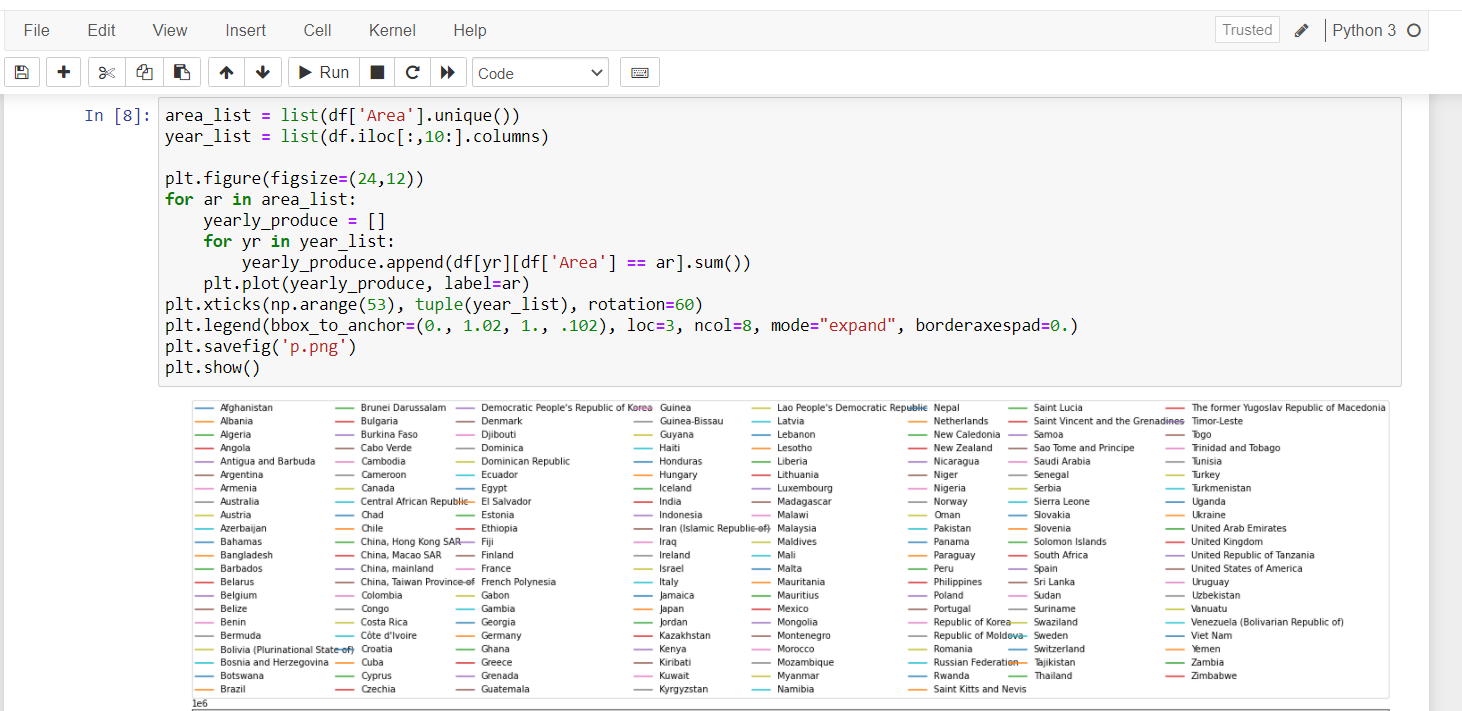


Fig 3. Plot for annual produce of different countries with quantity in y-axis and years in x-axis.

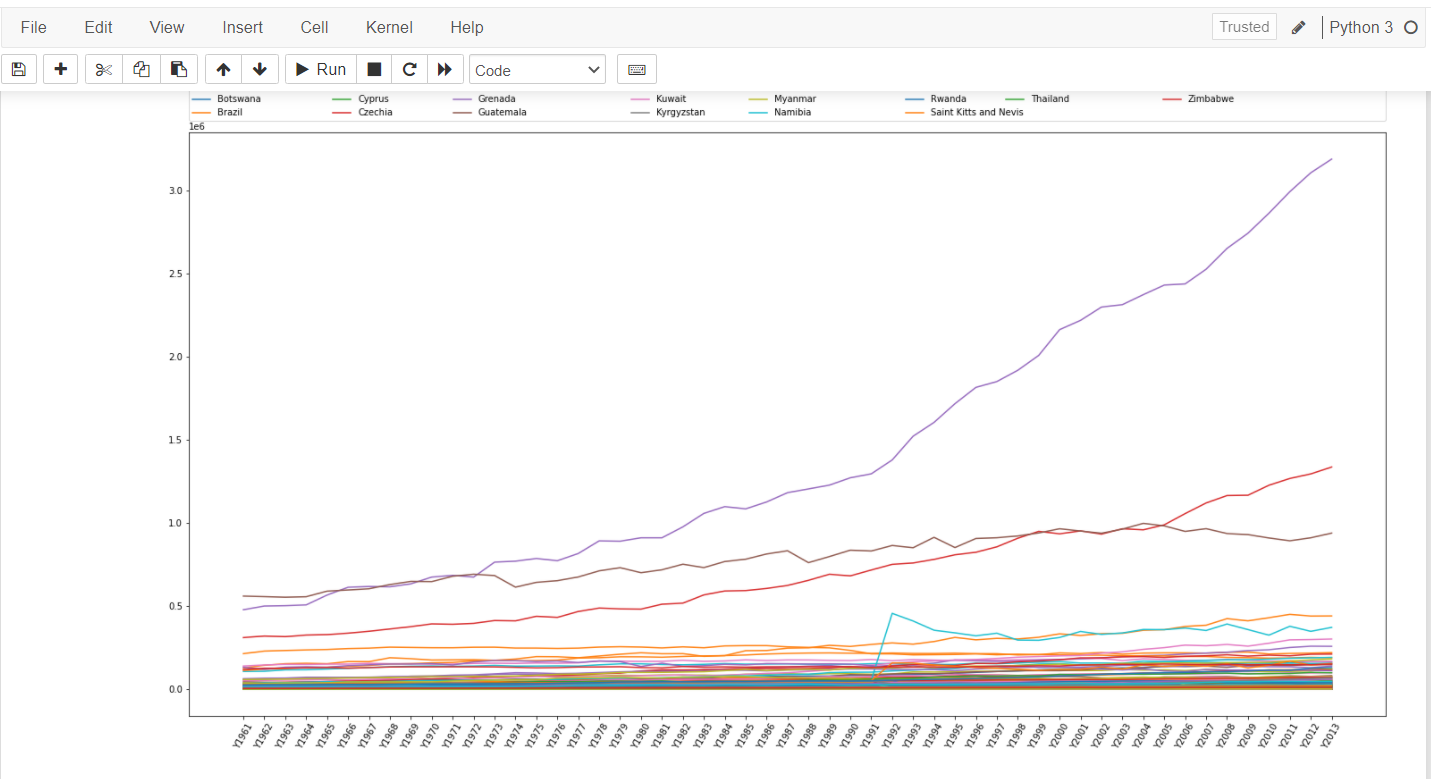


Fig 4. ‘p.png’ diagram

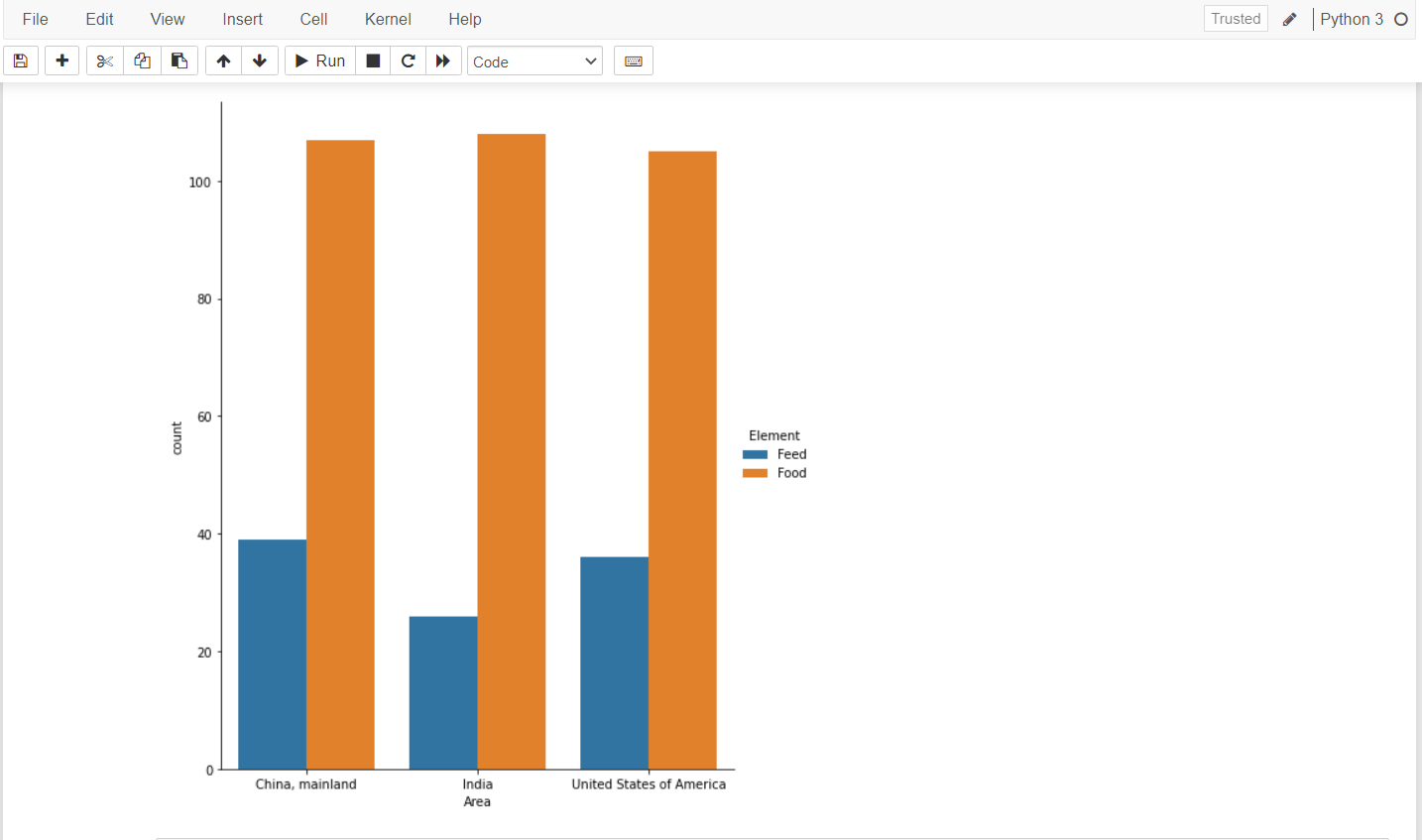
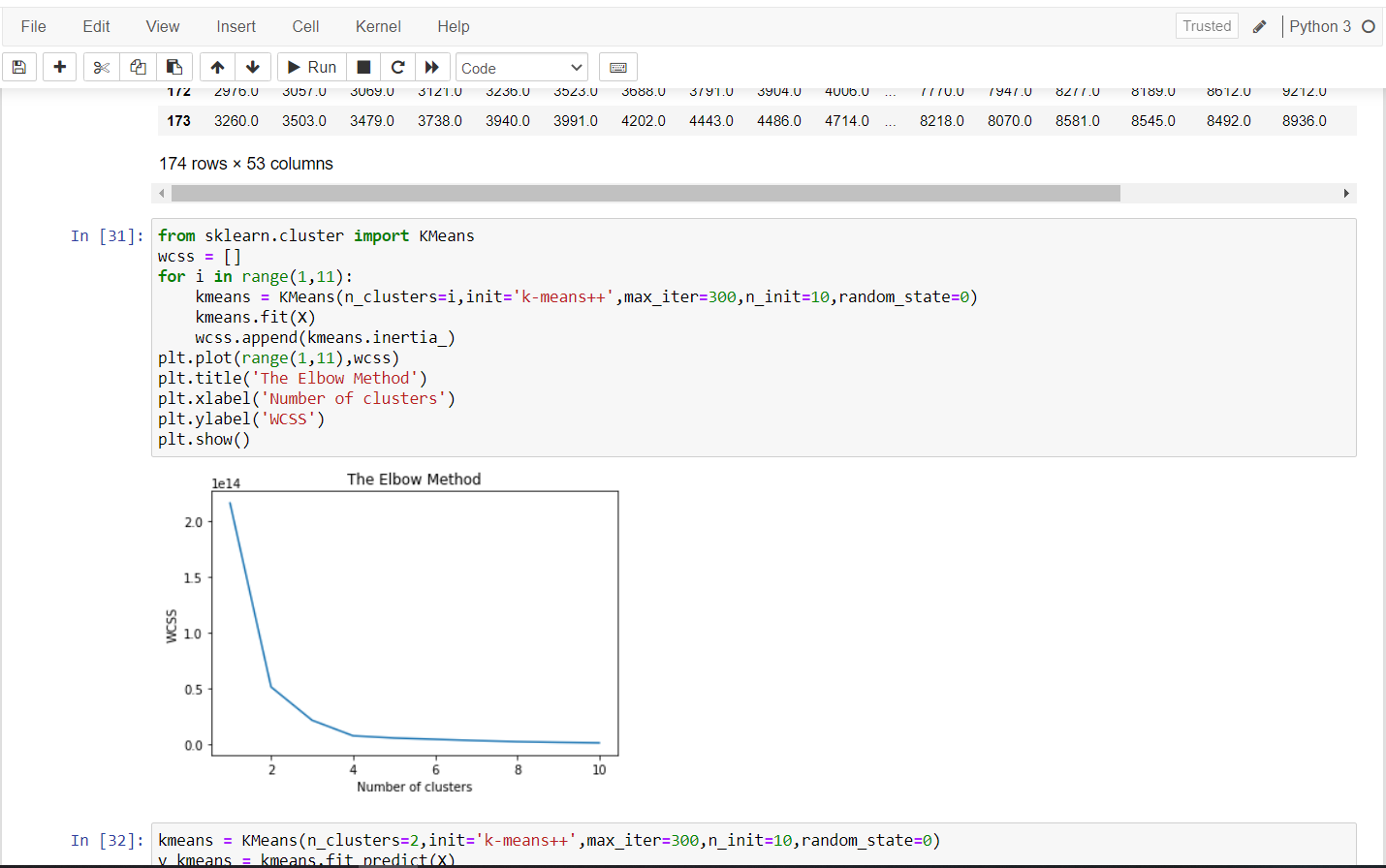


Fig 5. Food and feed plot for the largest producers(India, USA, China)



**6.CONCLUSION**

Clustering is techniques by which large dataset are dividing into small data collections that are called clusters. There are number Of algorithms that work well for clustering the data that can divide a dataset in clusters. Survey on k-means clustering algorithm proposes different advantage and disadvantage in different k-means application algorithm. It clearly shows that clustering can be used effectively to extract the knowledge in precision agriculture field. Traditional k-mean method is combined with Hierarchical algorithm for the centroid selection.

**7. BIBLIOGRAPHY**

* <https://www.kaggle.com/dorbicycle/world-foodfeed-production> for dataset.