Clustering and Fitting Presented By

NICOLAS NIMACH NKENGFACK

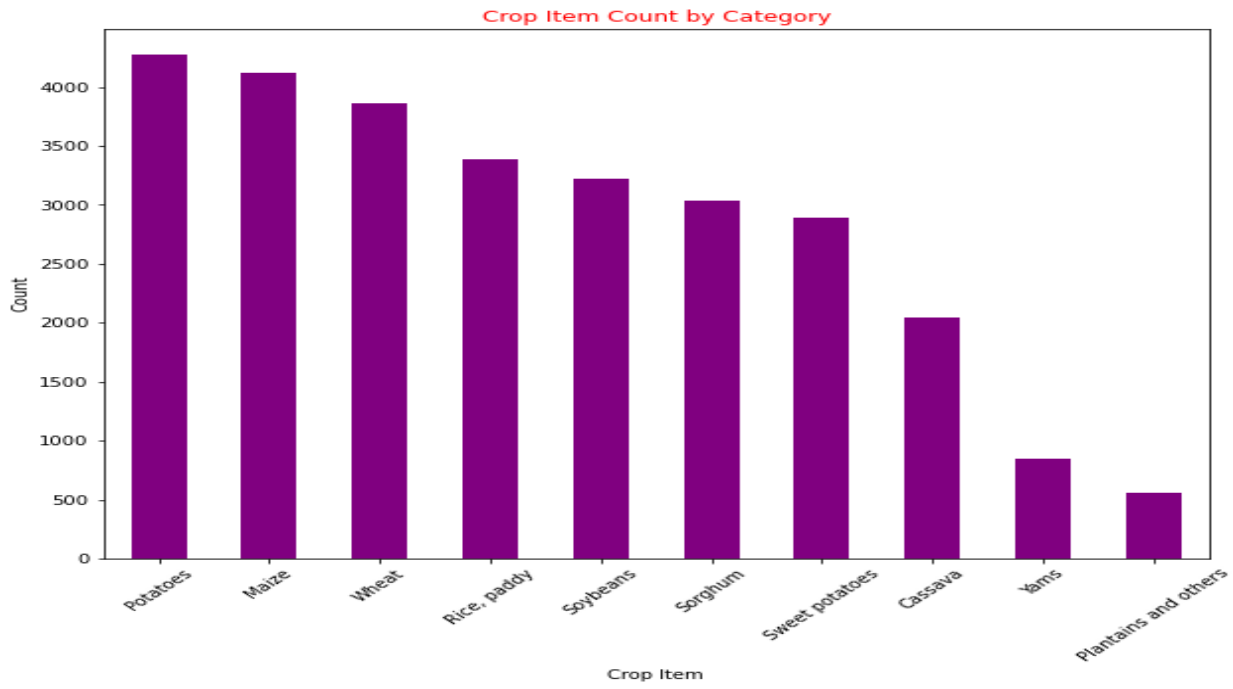
ID:23087514

https://github.com/Nkengfack/Applied-Data-Science-Assignment-.git

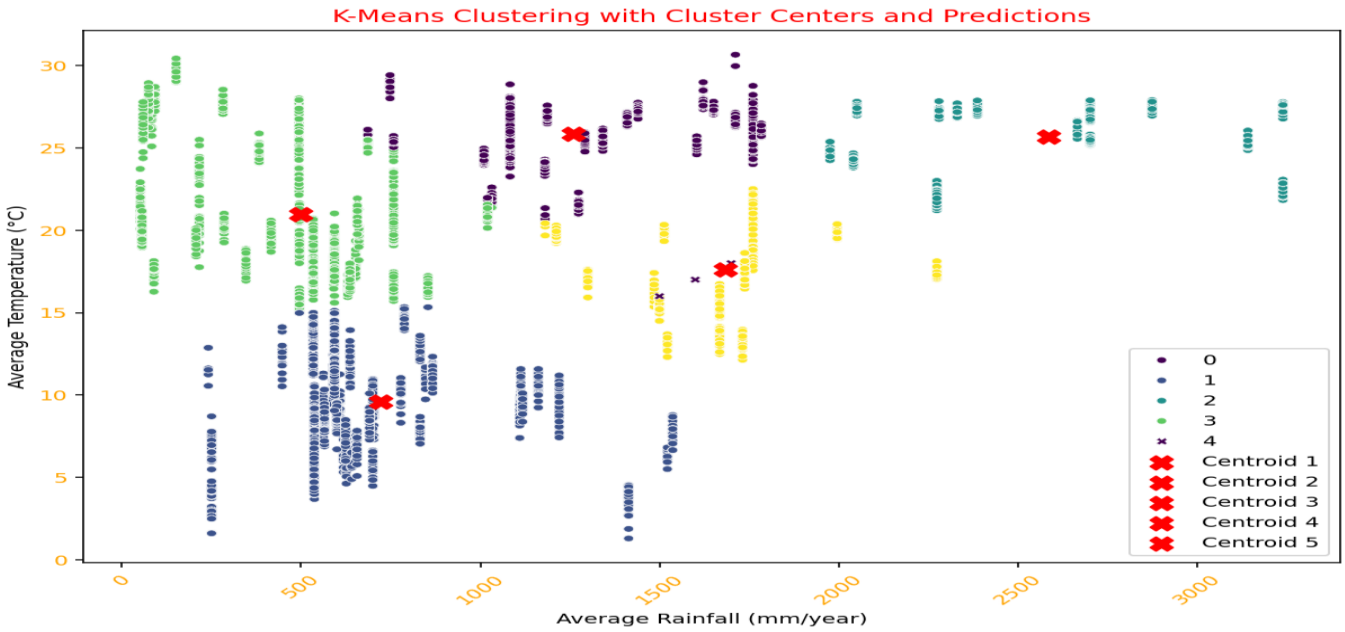
Introduction

This presentation is drawn from a dataset from the Kaggle website showing statistics of various countries and their crop produce in different years, hg/ha yield, average rainfall in mm per year, average temperature and pesticides, I will use a scatter plot, histogram, heatmap, elbow.

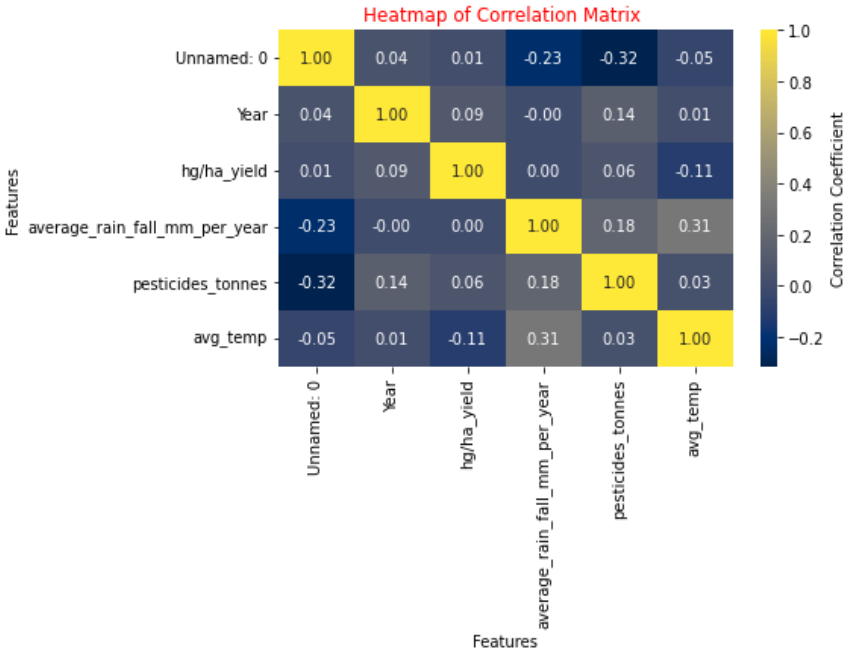
Crop item count by category: The chart displays the distribution of different crop items, with Potatoes having the highest count of above 4000 in various years of different countries and Plantains and others having the lowest below 500. Dominance of Potatoes: Potatoes have the highest count, indicating they are the most frequently occurring item in the dataset. Gradual Decline: As we move from Potatoes to Plantains and others, there is a general trend of decreasing count. This suggests a decreasing frequency of these items in the over the years.



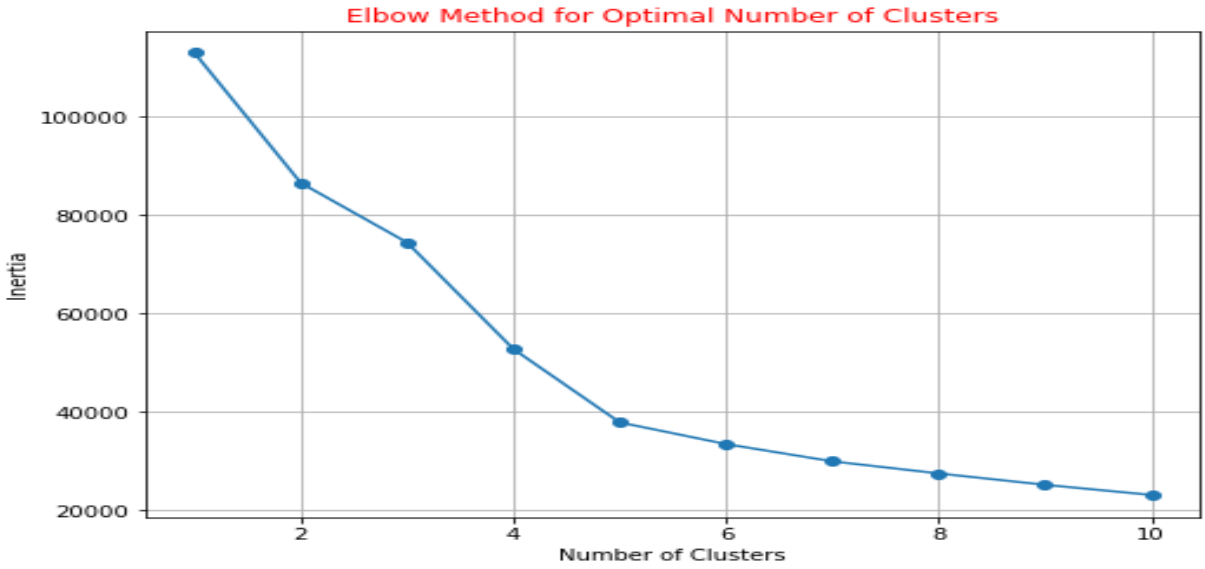
The K mean clustering scatter plot with predictions: Depicts the average temperature and rainfall (mm/year) of various crops over a span of years. The blue clustering in the lower left corner, indicates that crops thrive in regions with low average temperature from (1C to 15 C) such as maize and low average rainfall, green to the left region, crops prefer high to temperature and moderate rainfall. Yellow cluster is positioned towards the centre, indicating that these crops are adaptable to a range of temperature and rainfall conditions. sky-blue cluster is situated in the upper-right region, suggesting that these crops prefer high temperature and higher rainfall. While the c purple show at temp of 30 C and 1500mm crop adapted. The prediction points are red x



Correlation Heatmap: it shows a visual representation of how year, yield, rain, pesticide and temperature correlate to each other in the dataset. The colour yellow (1.00) shows the highest correlation individually and also, average rainfall turn to increase as temperature goes high (0.31) as in the case of Angola from the dataset. A zero correlation between yield and rainfall show in some countries both factors are parallel with impact against another



Elbow Method for Optimal Number of clusters: it depicts the number of clusters and inertia. The inertia of 100000 show a poor clustering of crops with respect to factors affecting them like rainfall, temperature, pesticide in various countries like Brazil, Algeria etc. There is a rapid decline indicates that each additional cluster significantly reduces the variability within the clusters. At the elbow point 4, the curve starts to flatten out indicating the rate of decrease in inertia pointing out a good clustering quality, which indicate some countries could actually have same temperature leading to a proportionate number of crops. The clutter 10 show that we could barely identify clusters for example the clustering of countries.



Yield analysis and prediction with uncertainty: The exponential fit data suggests a general upward trend in yields over time, with the rate of increase accelerating as time gears up yield could exponentially pass 10000 in 2020