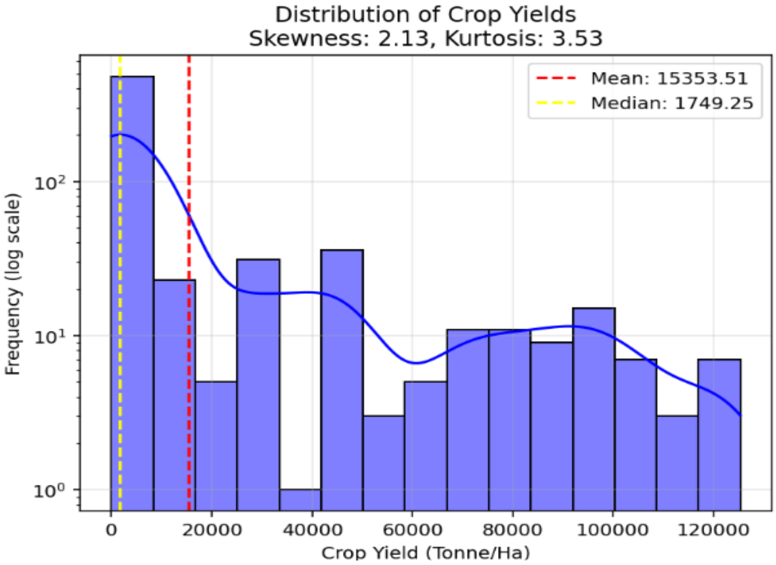
Title : Crop Yield Analysis Report
Name : Arpita Thokal
Student ID : 24005907
Github Link : https://github.com/arpitathokal/Fitting_and_Clustering

Introduction

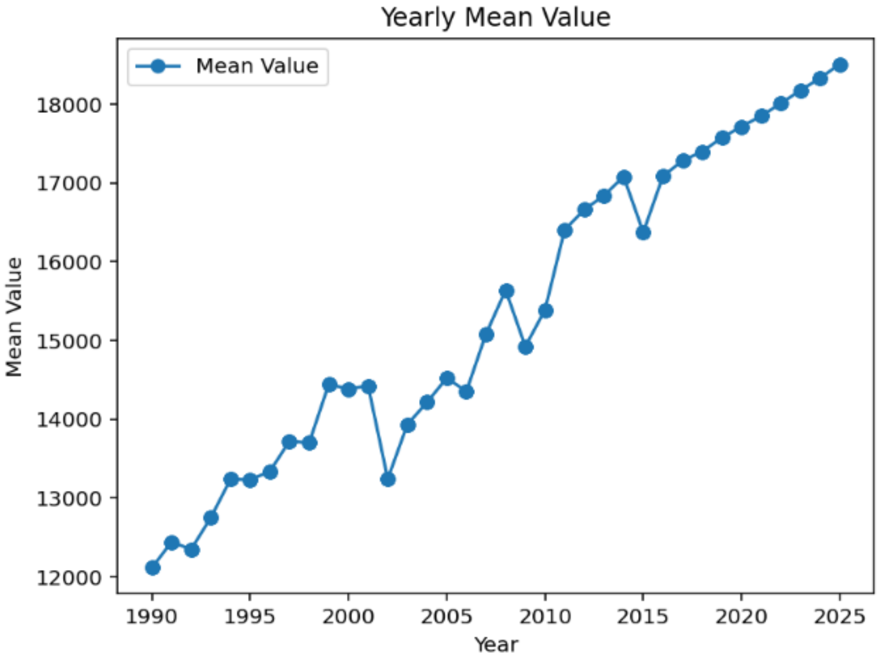
This report focuses on changes in crop yield statistics across various regions, from 1990 to 2025, highlighting distribution patterns, variability, and growth trends across years. The report has meaningful Insights based on the aggregated data and six visualizations, a histogram of crop yield distributions, boxplot reveal Crop-based clusters and locations-based clusters, line plot fitted using logistic fit to the data to interpret growth trends over time and an elbow method plot showing optimal clustering which uses K means algorithm. The key metrics used for analysis are, Year and Value columns where Crop yield measured in Tonnes per Hectare and location, crops columns are encoded using one hot encoding. In addition, by analysing the historical trend we can predict the future yield which provides valuable insights to support informed decision-making for policymakers, resource planners, and key stakeholders in the agricultural sector.

1.Distribution of Crop Yields (Histogram)



The Histogram depicts the distribution of crop yield between 0 to 125000 (tonne/Ha) with a skewness score of 2.13, the crop yield distribution shows a large right-skewness, meaning that the majority of yields cluster around lower values, while notable high-yield outliers provide a lengthy tail. A leptokurtic pattern, which is distinguished by heavier tails and a more peaked distribution than a normal curve, is highlighted by the kurtosis of 3.53. This plot helps us to know the frequency of crop yield where the value between 0 to 10000 has highest frequency log scale.

2. Line plot

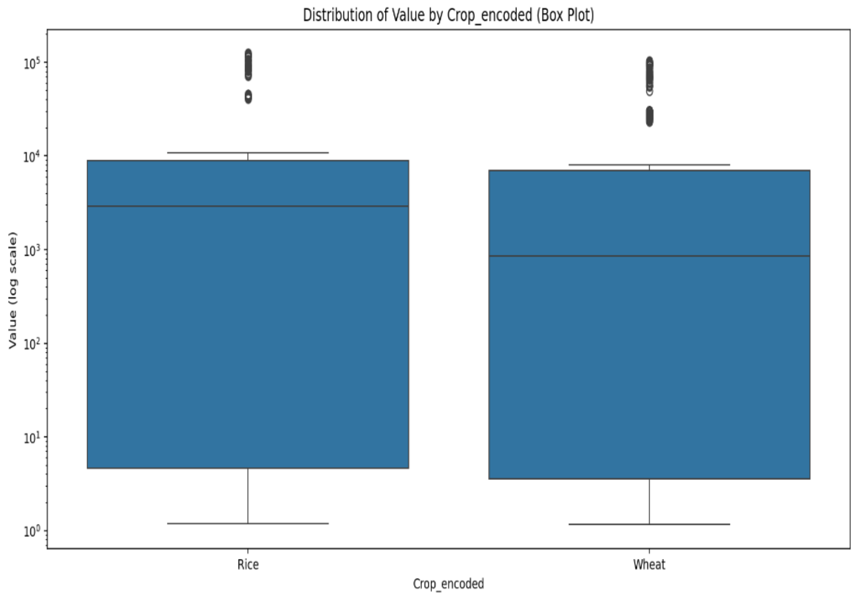


The line plot displays the yearly mean value from 1990 to 2025. It has a steady upward trend starting from around 9000 to 16000.

The overall visualization suggests a significant increase in values over the 35-year period, with a particularly marked transition in the latter half of the timeline.

The increase in yield over the years help us to decide to use these two columns for fitting.

3.Box Plot



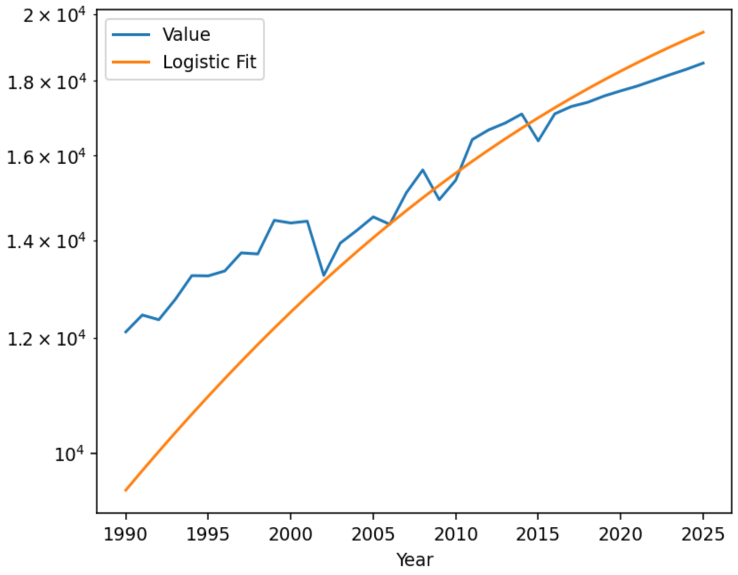
The1st box plots display the distribution of values for two crops - Rice and Wheat - with both showing almost similar median values but some outliers at the higher ranges.

It has yield less than 20000 tonne/Ha. It emphasizes the consistency of median crop yields, which are continuously at about 10^3 tons per hectare, using a logarithmic scale.

The interquartile range (IQR) for both crops is quite large, indicating significant variability in the data. Notably, there are several outliers above the upper whisker for both crops, suggesting some values are much higher than most of the data points.

To plot this, we have used one hot encoding to convert the categorical data to numeric and is helpful for comprehending variations in crop value distributions. There is one more box plot for Locations.

4. Growth Trend and Logistic Fit



The line-fit graph illustrates the evolution of agricultural yields from 1990 to 2025, featuring both actual yield data (represented by the blue line) and a logistic fit (the orange line).

Beginning at about 12,000 tonnes per hectare in 1990 and gradually rising to about 14,500 tonnes per hectare by 2000, the actual yield data demonstrates a definite increase trend then slight decreases and then decelerating growth phase is observed.

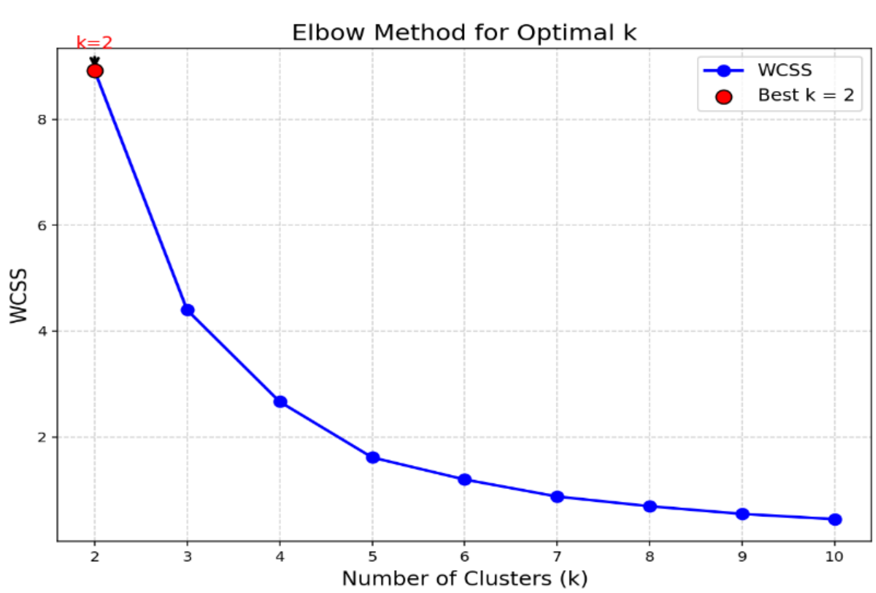
To perform logistic fit, following formula is used

f = n0 / (1 + np.exp(-g * (t - t0)))

this appears to be the perfect fit as during 2005-2015 period where the curves closely overlap and accurately models the slowing growth rate as S shaped.

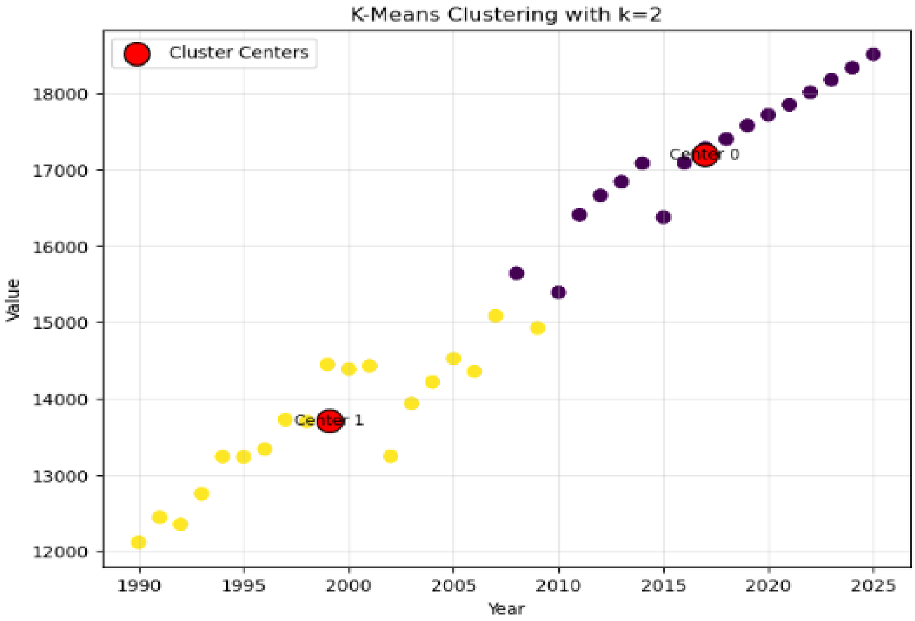
5. Optimal Clustering Analysis (Elbow Plot)

The elbow method analysis for K-means clustering displays the WCSS (Within-Cluster Sum of Squares) values against the number of clusters (k) between 2 and 10. The WCSS drops dramatically from k=2 to k=3 (from approximately 9 to 4.5) but shows diminishing returns, thereafter, making k=2 the most efficient choice for clustering the yield data.



This implies that two clusters, which most likely represent low-performing and high-performing yield categories, can be effectively formed from the agricultural yield data.

6. K means Clustering



This Plot visualise two separate clusters with purple and yellow dots. This is basically dividing the dataset into each group's centroid which is shown by red cluster centre. Early years (1990-2005) are represented by the lower cluster (yellow). Later years (2005–2025) are displayed in the upper cluster (purple). Yellow cluster seems to grow steadily and Purple cluster Growth is slowing down as we approach 2025.

Conclusion

In conclusion, the analysis provides valuable insights into historical yield trends and future growth potential Opportunities. This analysis combines statistical modelling (logistic fit), validation (elbow method), and pattern recognition (clustering) to understand how values have changed over time and what we might expect in the future. Stakeholders may create efficient plans to increase crop yields or may predict the crop yields to accomplish sustainable agricultural development by utilizing the ideas in this research.