1. Introduction

Objective:

To investigate trends, distribution, and clustering of barley yields in Saskatchewan from 2000 to 2023, with a comparative analysis against wheat yields.

Datasets:

- RM-level yield data (rm_yield_00_23_major_crops.csv)
- GIS data for spatial analysis (RuralMunicipality.shp)
- Comprehensive annual yield data (rm-yields-data.csv)

Analytical Techniques:

- Outlier detection and removal
- Trend analysis
- K-Means clustering
- Comparative analysis with wheat yields

2. Theory and Definitions

Machine Learning Algorithm: K-Means Clustering

K-Means is a type of unsupervised learning algorithm used to partition a dataset into a set of k groups (clusters).

Purpose:

To classify data points into clusters based on their features, ensuring that data points within each cluster are similar to each other and dissimilar to those in other clusters.

- How it works:

1. Select k initial centroids randomly.

2. Assign each data point to the nearest centroid, forming k clusters.

3. Calculate the new centroids as the mean of the data points in each cluster.

4. Repeat steps 2 and 3 until convergence (centroids do not change).

Clustering

- Definition: Clustering is a technique used to group a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups.
- Purpose: To identify natural groupings within data, which can provide insights into the underlying structure of the data.

Elbow Method

The elbow method is a technique used to determine the optimal number of clusters in K-Means clustering.

How it works:

Plot the sum of squared distances (distortions) from each point to its assigned centroid as a function of the number of clusters. The optimal number of clusters is typically at the "elbow" point where the distortion starts decreasing more slowly.

3. Data Preprocessing and Cleaning

Initial Data Cleaning

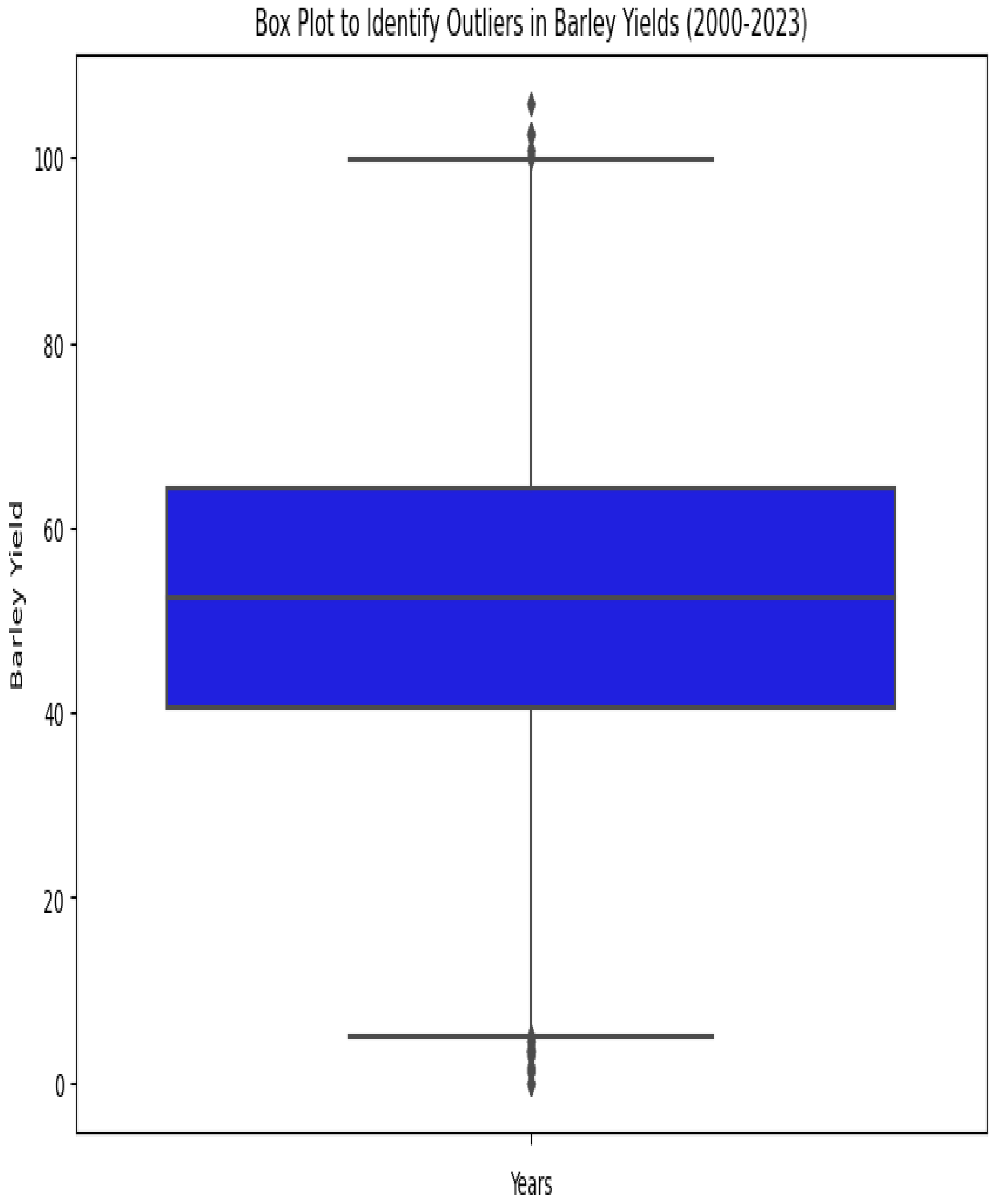
Steps:

- Load and inspect the datasets.
- Convert RM columns to strings for consistency.
- Clean RM columns by stripping extra spaces and ensuring consistent capitalization.

Outlier Detection and Removal

Box Plot Analysis:

- Identify outliers using an initial box plot.
- Outliers can skew the results, making it essential to identify and remove them for a reliable analysis.



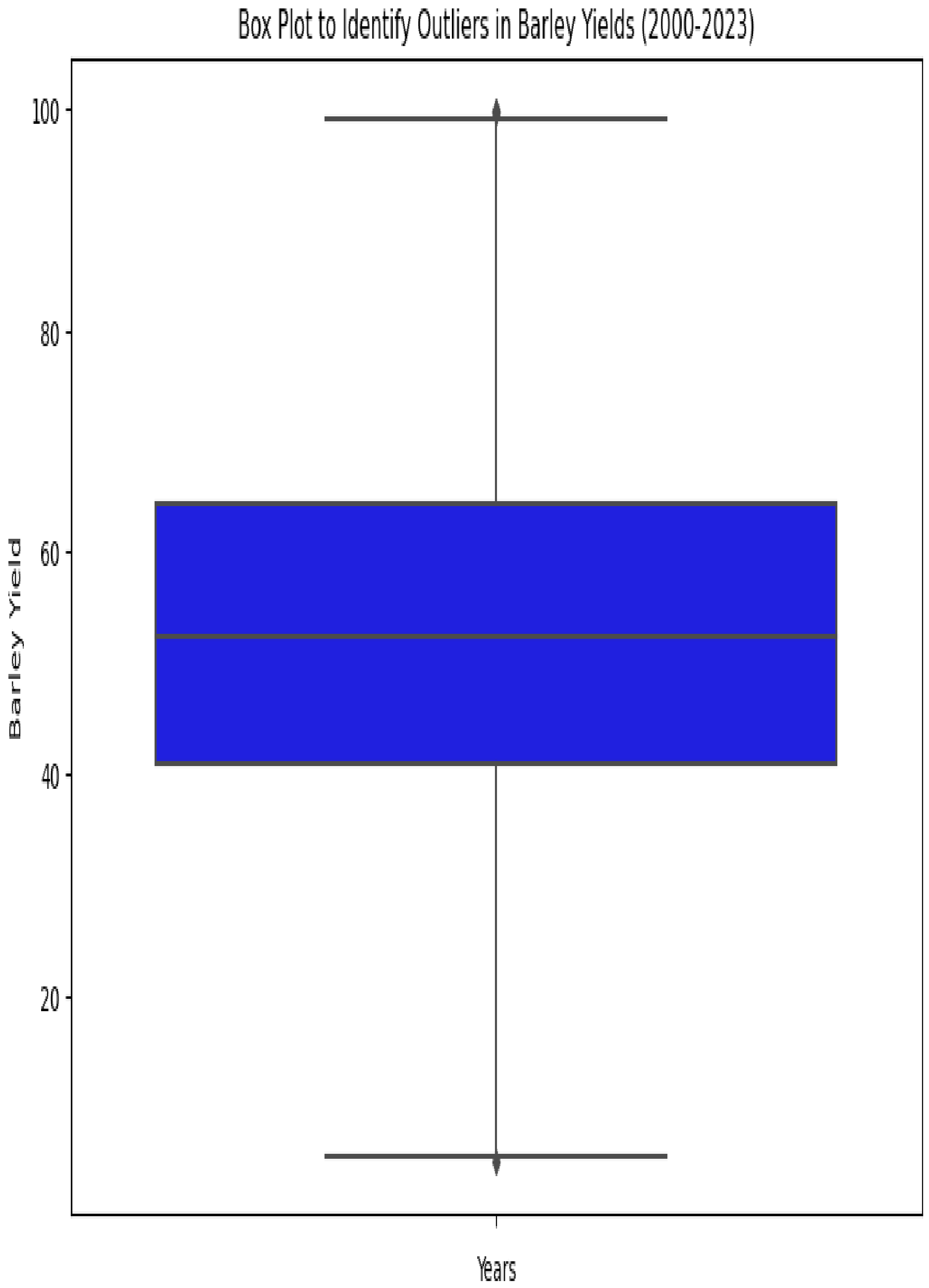
Explanation:

- The initial box plot helps to identify extreme values in the dataset.
- Outliers are points that lie significantly outside the interquartile range.

Outlier Removal

Explanation:

- Remove identified outliers to clean the data.
- This ensures that subsequent analyses are based on a consistent and accurate dataset.



Explanation:

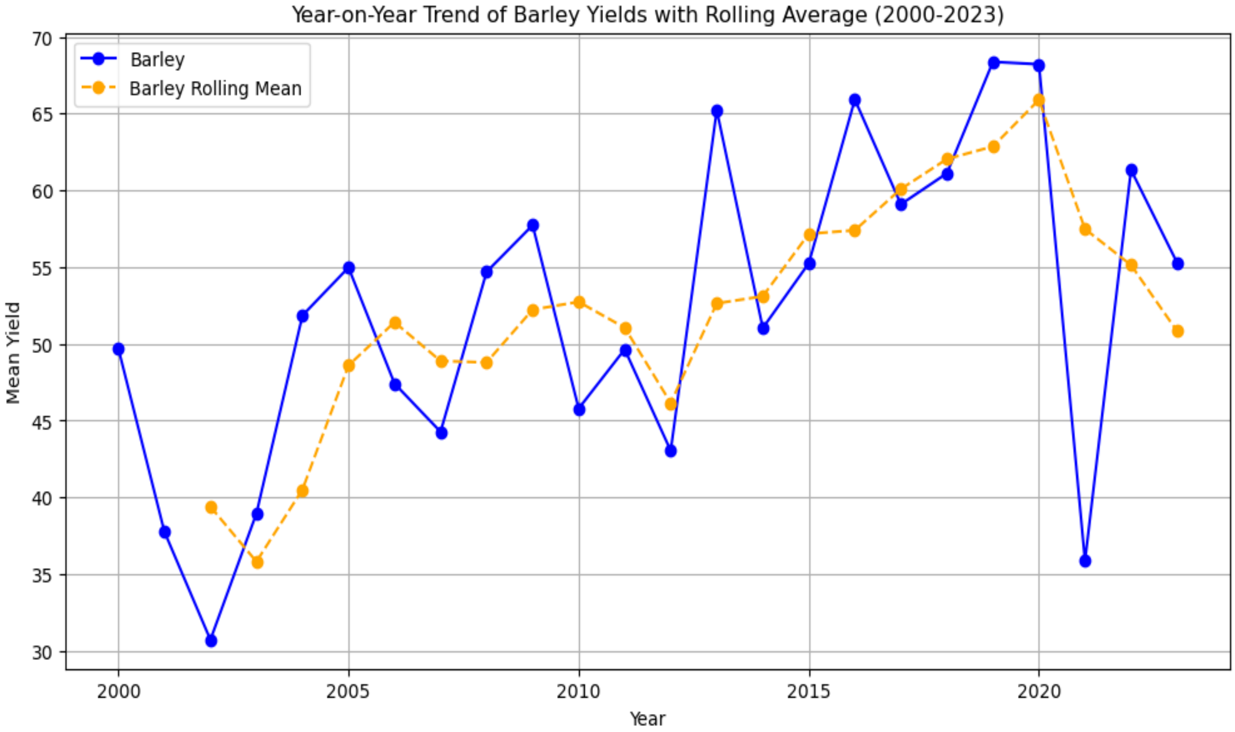
- The post-removal box plot shows a more consistent dataset without extreme values.
- This step is crucial to ensure the integrity of subsequent analyses.

4. Trend Analysis

Year-on-Year Trend Analysis

Steps:

- Calculate the annual average yields for barley.
- Plot the trends to visualize long-term performance.
- Purpose:
- To identify long-term trends in barley yields over the years.



Explanation:

- The trend plot reveals changes in barley yields from 2000 to 2023.
- This visualization helps in understanding the overall performance and identifying any significant fluctuations.

5. K-Means Clustering Analysis

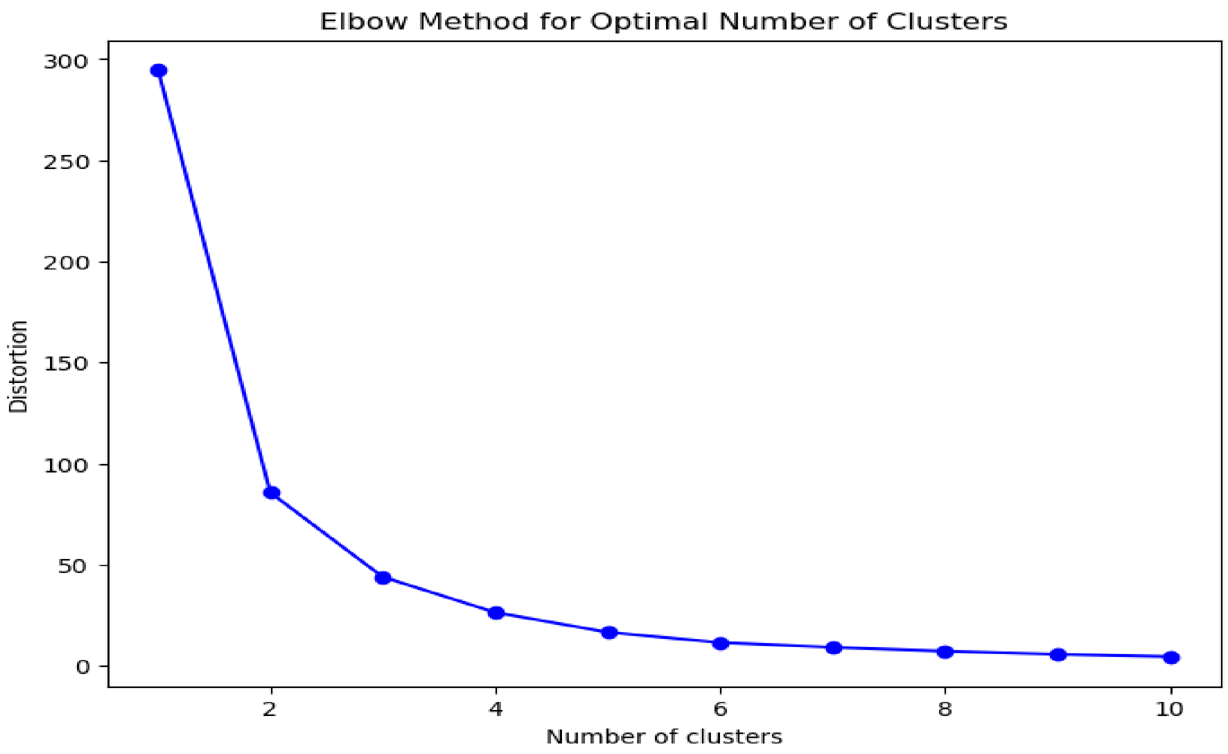
Elbow Method for Optimal Clusters

Steps:

- Use the elbow method to determine the optimal number of clusters.
- Plot distortions against the number of clusters to identify the "elbow" point.

Purpose:

To ensure that the chosen number of clusters accurately represents the data structure.



Explanation:

- The plot shows the distortions for different numbers of clusters.
- The optimal number of clusters is identified at the elbow point.

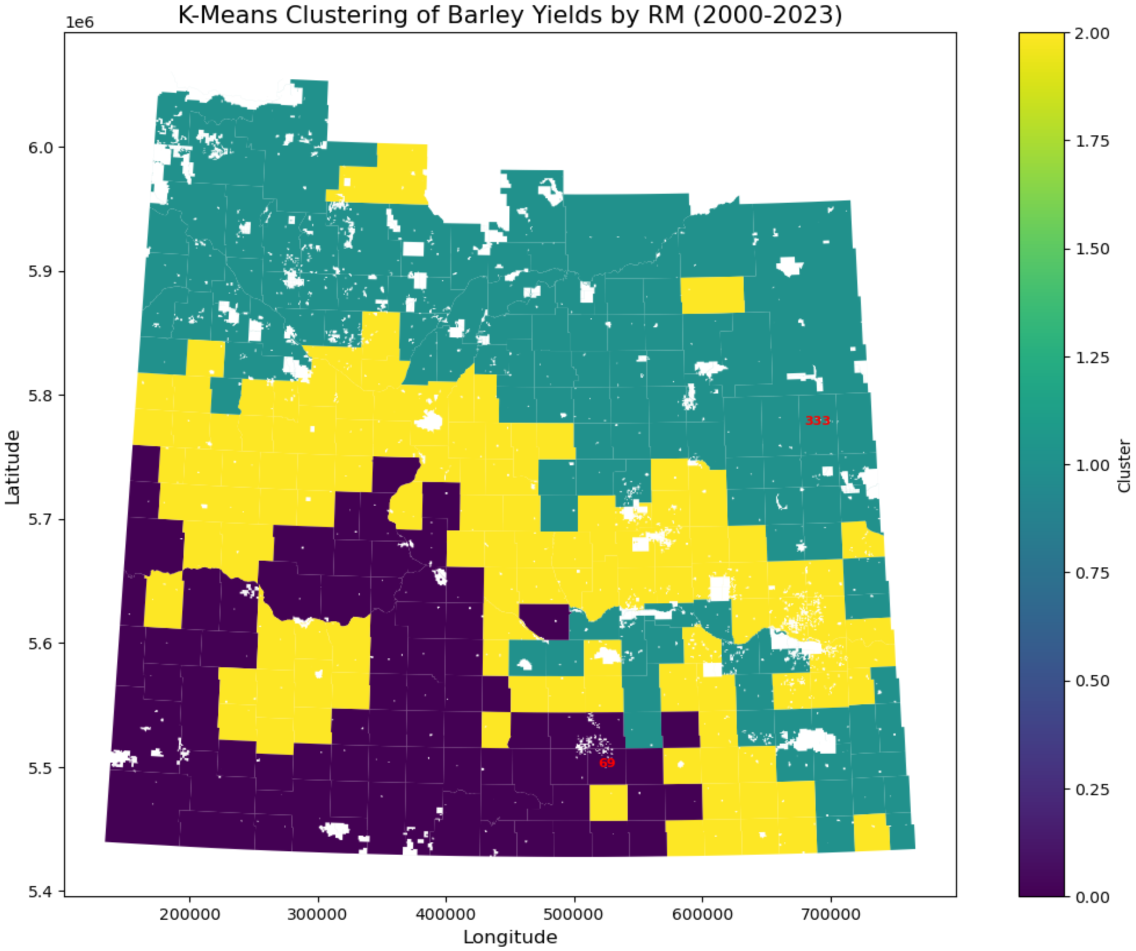
K-Means Clustering

Steps:

- Apply K-Means clustering using the optimal number of clusters.
- Visualize the clusters on a map to identify regions with similar yield patterns.

Purpose:

To identify distinct groups of RMs with similar barley yields, providing insights for targeted agricultural strategies.



Clusters are based on the average barley yields from 2000 to 2023.
Annotations highlight regions with the highest and lowest yields.

Explanation:

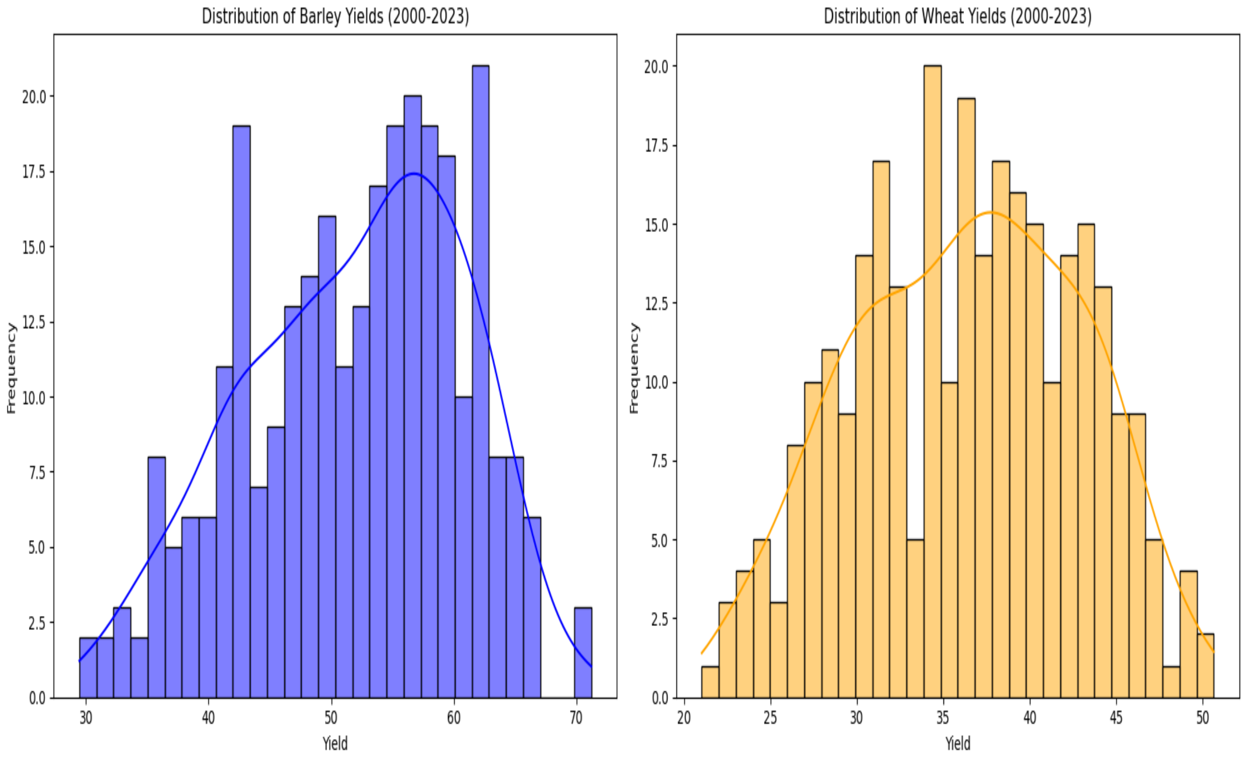
- The map visualization shows the spatial distribution of clusters.
- Clusters are based on the average barley yields from 2000 to 2023.
- Annotations highlight regions with the highest and lowest yields.

6. Comparative Analysis with Wheat Yields

Distribution Comparison

Steps:

- Plot histograms for barley and wheat yields side by side.
- Analyze the distribution patterns of both crops.
- Purpose:
- To understand the variability and central tendencies of barley and wheat yields.



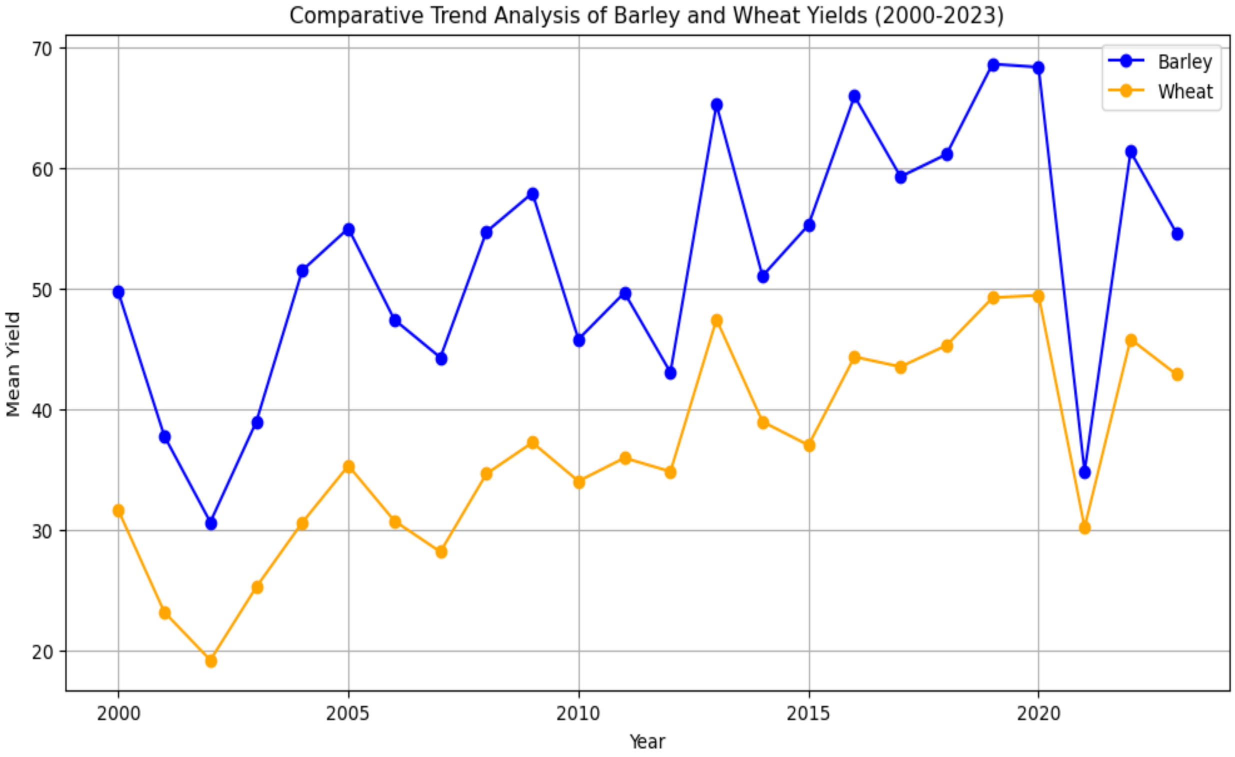
Explanation:

- The histograms reveal the distribution patterns of barley and wheat yields.
- Barley yields show a wider range and higher variability compared to wheat yields.

Trend Comparison

Steps:

- Calculate the annual average yields for both barley and wheat.
- Plot the trends to compare the performance of the two crops.
- Purpose:
- To identify any significant differences or similarities in the yield trends of barley and wheat.



Explanation:

- The trend plot shows the annual average yields of barley and wheat from 2000 to 2023.
- This comparison helps in understanding how the performance of the two crops differs over time.

7. Conclusion and Future Directions

Key Findings:

- Barley yields have shown significant variation across different RMs and years.
- Clustering analysis revealed distinct groups of RMs with similar yield patterns.
- Comparative analysis with wheat showed different yield trends and distributions.
- Identification and removal of outliers improved the reliability of the analysis.

Future Directions:

- Further research could incorporate additional variables such as weather data, soil health metrics, and economic factors to enhance yield predictions.
- Extending the analysis to other crops and regions could provide broader agricultural insights.
- Investigating the impact of specific agricultural practices and policies on yield performance.

Insights:

- Understanding yield patterns and variability can help in developing targeted agricultural strategies.
- Identifying regions with high and low yields can inform resource allocation and policy decisions.
- Comparative analysis with other crops can provide a more comprehensive understanding of agricultural performance.