Prepared by: Zahid Niloy

Overview

In this analysis, unsupervised machine learning techniques were applied to historical weather data from the 1980s to identify potential patterns or clusters in temperature and related variables. The goal was to help ClimateWins better understand regional or seasonal groupings that may align with 'pleasant weather' conditions.

Methods Summary

1. Data Preparation:

- Weather data from the 1980s was extracted and scaled using StandardScaler.
- Dimensionality was later reduced using PCA (30 components) to preserve over 90% of variance.

2. Clustering Algorithm:

- Hierarchical clustering was applied using four linkage methods: Single, Complete, Average, and Ward.
- Both full and PCA-reduced data were tested for comparative analysis.

3. Visualization:

- Dendrograms were generated for each linkage method (with and without PCA).
- The number of clusters was estimated and compared across methods.
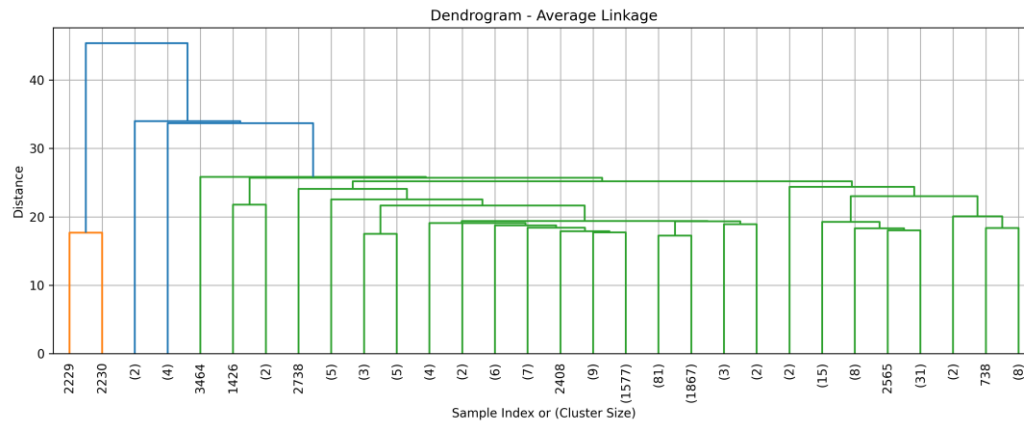
Dendrogram Results (Full Dataset)

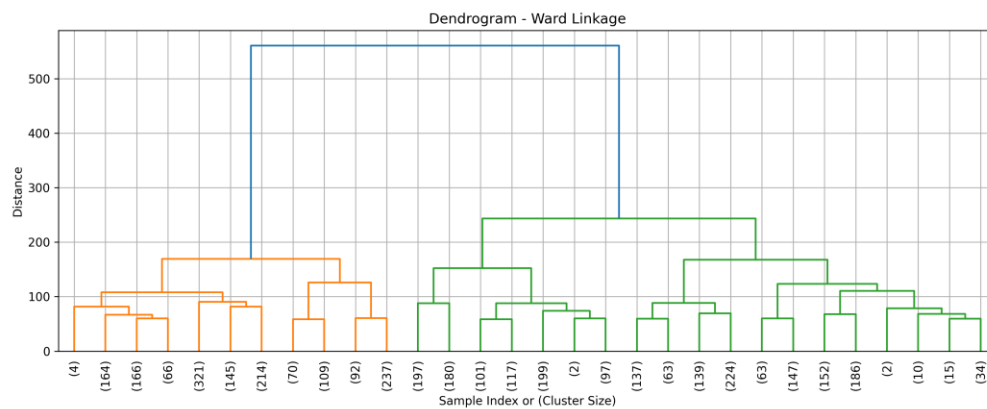• Single Linkage: 3,639 clusters – highly fragmented due to chaining.
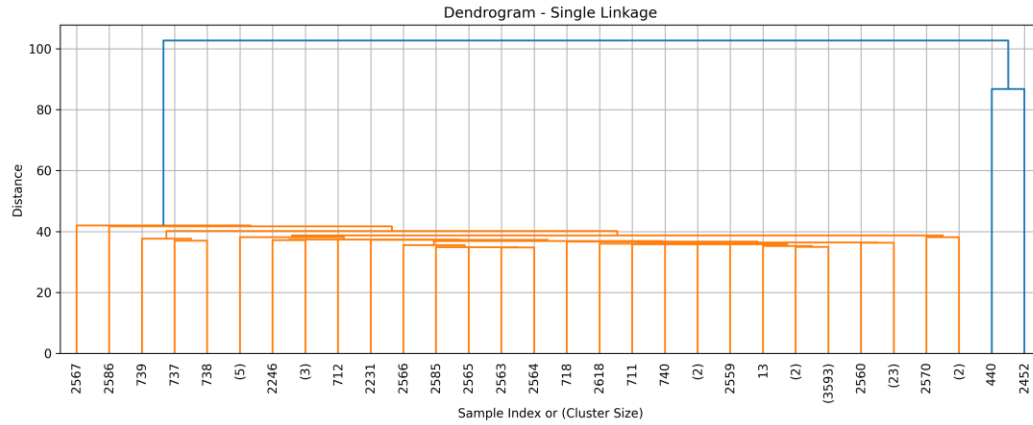
• Complete Linkage: 259 clusters – more compact clusters, better structure.
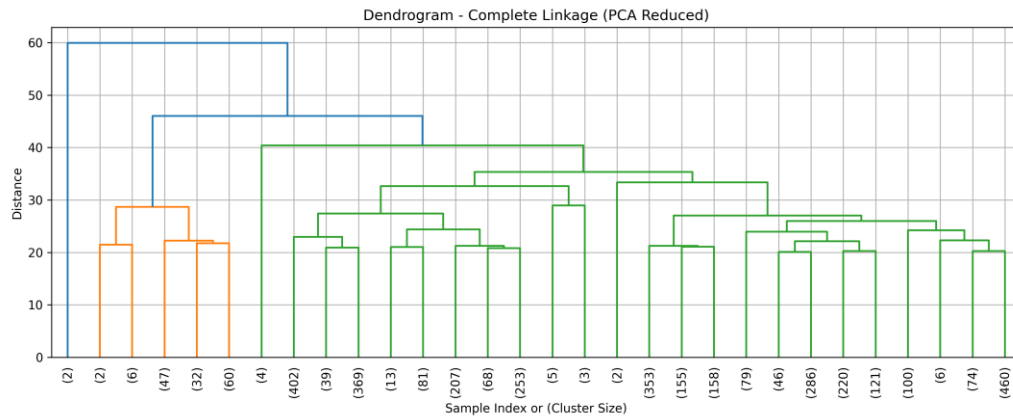
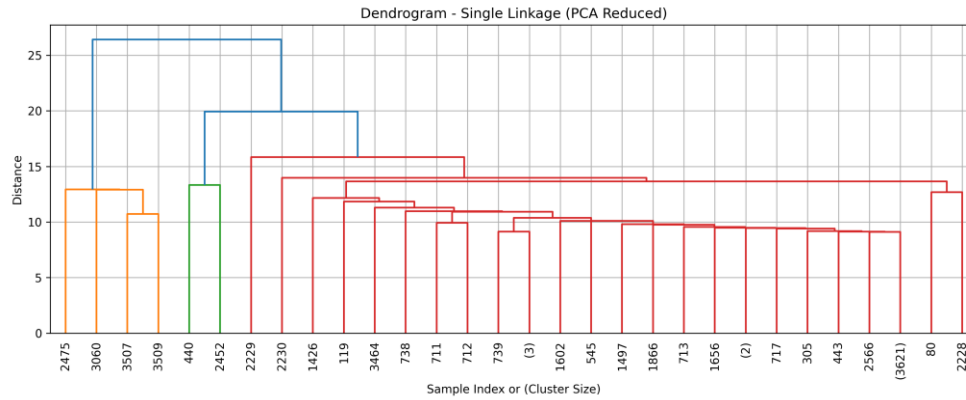• Average Linkage: 72 clusters – balanced structure and interpretability.

Dendrogram - Average Linkage

• Ward Linkage: 565 clusters – better variance minimization but still fragmented.
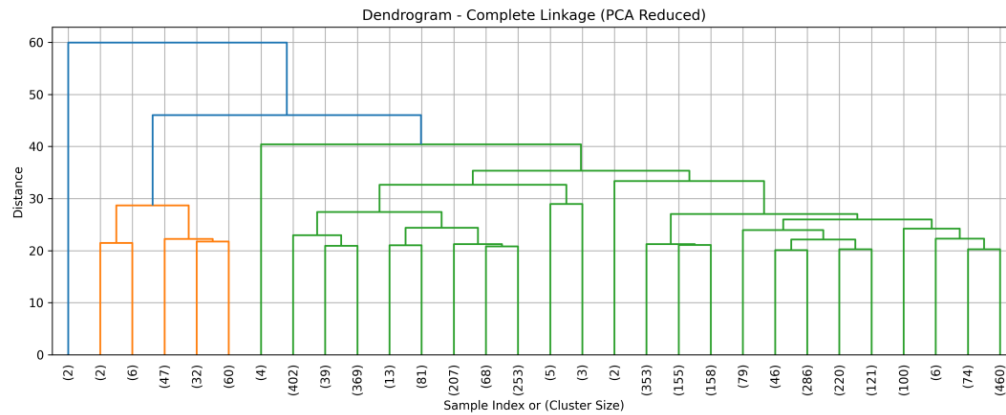

Dendrogram - Ward Linkage
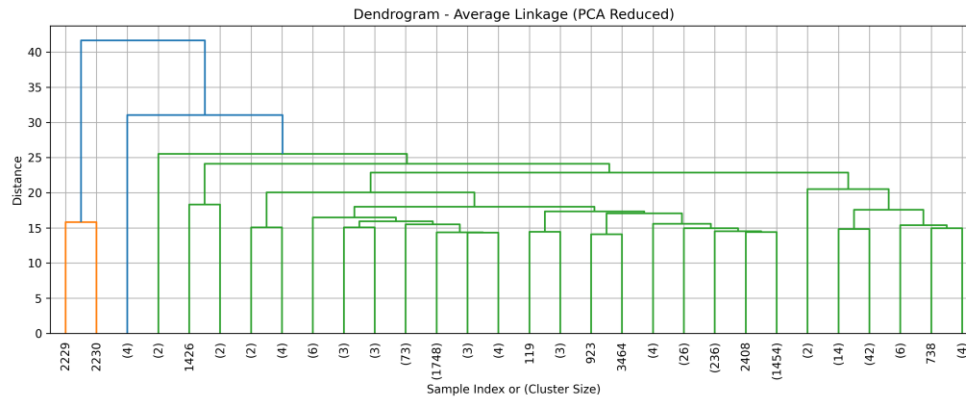
Dendrogram Results (PCA-Reduced Dataset)

• Single Linkage: 4 clusters – major improvement after PCA.
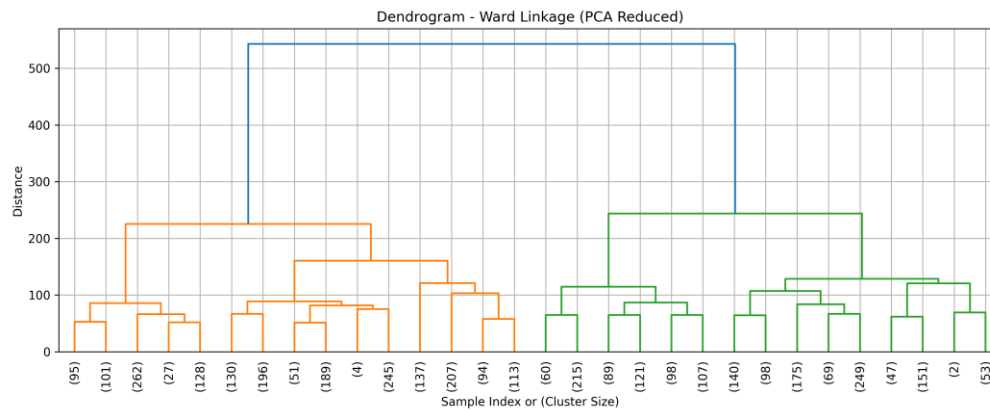

• Complete Linkage: 114 clusters – sharper separations than full data.

• Average Linkage: 21 clusters – best interpretability.

Dendrogram - Average Linkage (PCA Reduced)

• Ward Linkage: 367 clusters – cleaner clusters, but slightly dense.

Dendrogram - Ward Linkage (PCA Reduced)

Insights & Conclusions

- PCA significantly improved clustering quality by consolidating similar patterns.

- Average Linkage (PCA) gave the most manageable and meaningful clusters (21).

- Ward Linkage (PCA) remained effective for identifying compact groups.

- Single Linkage, both before and after PCA, was ineffective for this weather data.

- Scaling the data and reducing dimensionality proved essential for clean, interpretable clustering results.

Recommendations

- Use PCA before clustering for future climate pattern analysis.

- Prefer Ward or Average linkage depending on the desired cluster compactness.

- Compare resulting clusters to 'pleasant weather' labels to validate grouping quality.

- Extend analysis across longer timeframes (e.g., full decades) for historical trend discovery.

- Establish a pipeline: Scaling → PCA → Ward Clustering for ongoing ClimateWins analysis.