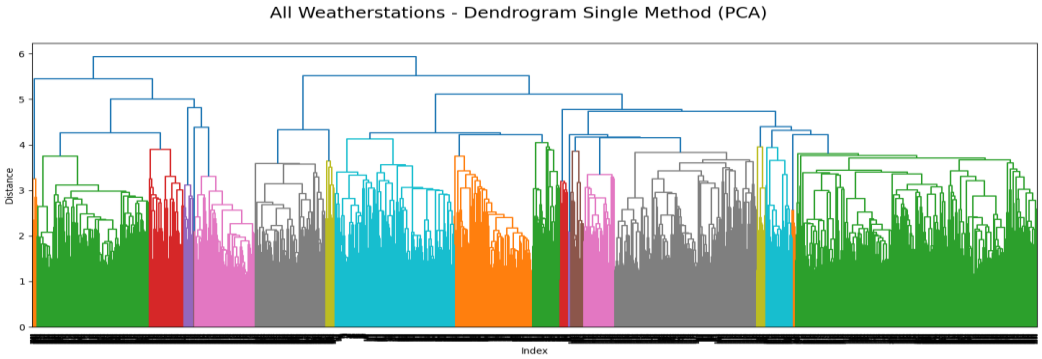
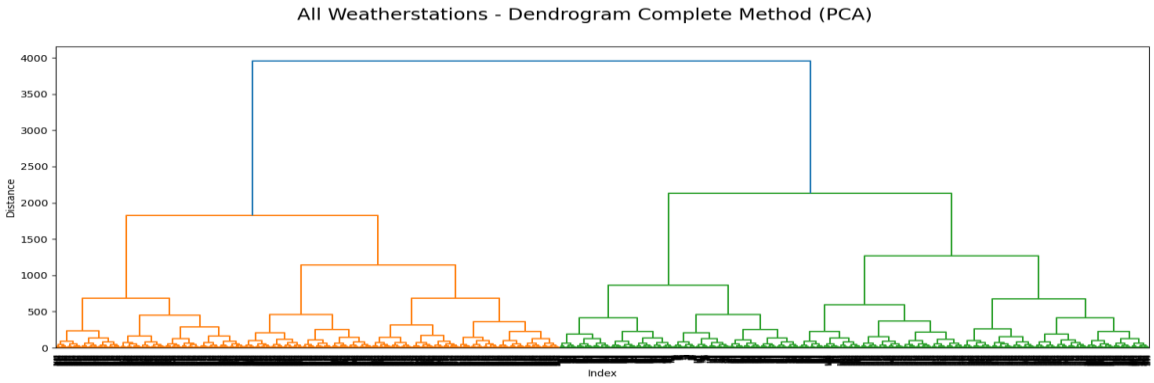
2.1: UNSUPERVISED LEARNING ALGORITHMS



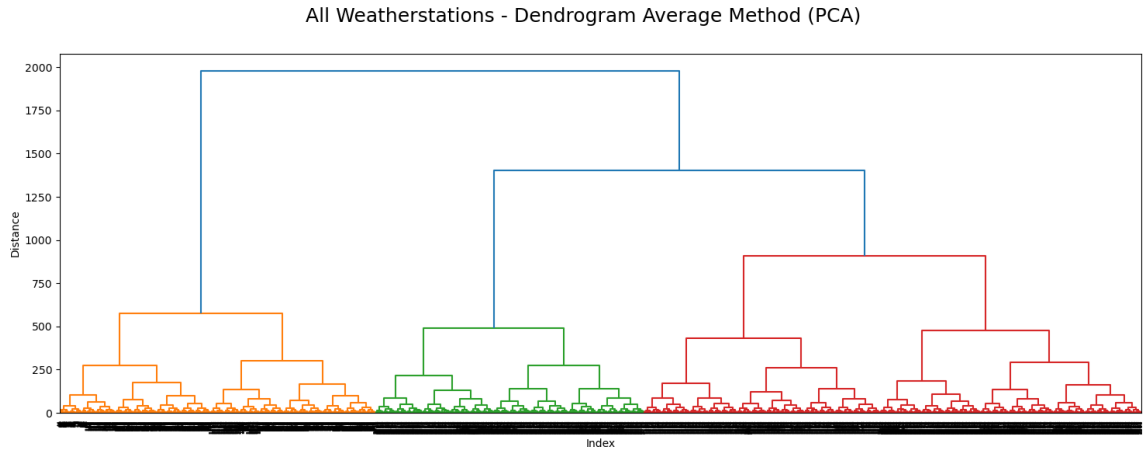
Single Method:

- **What it does**: Single linkage clustering focuses on the minimum distance between clusters.
- **Insights for Weather Stations**:
    - Clusters tend to form long chains rather than distinct groups.
    - **Outcome**: The single method fails to create meaningful separations among weather stations, which might obscure insights into patterns like regional similarities or changes over decades.
    - **Relevance**: Unsuitable for ClimateWins' goal of uncovering interpretable associations.



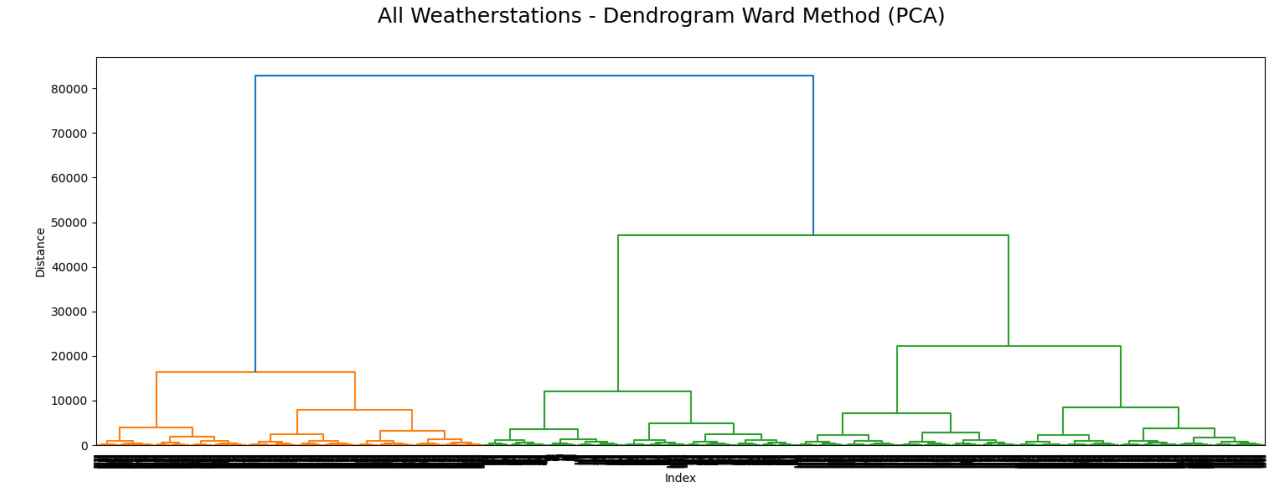
Complete Method:

- **What it does**: Considers the maximum distance between data points in different clusters to minimize variance.
- **Insights for Weather Stations**:
    - Weather stations with similar patterns (e.g., stations in comparable climatic zones like northern vs. southern Europe) form distinct clusters.
    - **Outcome**: This method is effective in grouping data, especially after scaling and PCA, making it possible to detect regions with common weather trends.
    - **Relevance**: Highlights weather stations that share extreme or similar trends over the decades, useful for comparative analysis.



Average Method:

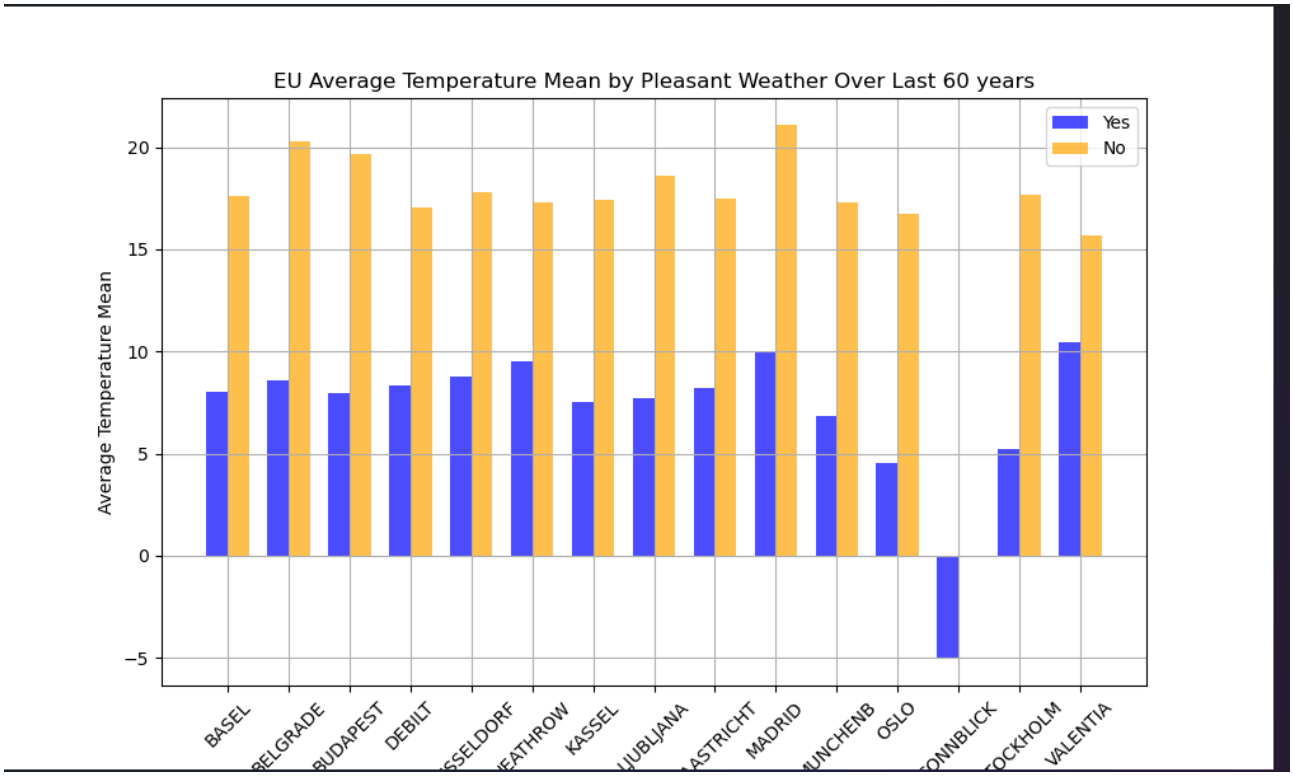
- **What it does**: Averages the distances between clusters to balance chaining and overly tight clusters.
- **Insights for Weather Stations**:
    - Weather stations cluster into roughly two main groups. These may correlate to broader climatic classifications (e.g., temperate vs. Mediterranean climates or urban vs. rural stations).
    - **Outcome**: Provides insights into macro-level patterns like continental vs. coastal weather trends but less granularity compared to Ward's method.
    - **Relevance**: Useful for high-level trend analysis but may lack finer details.



Ward Method:

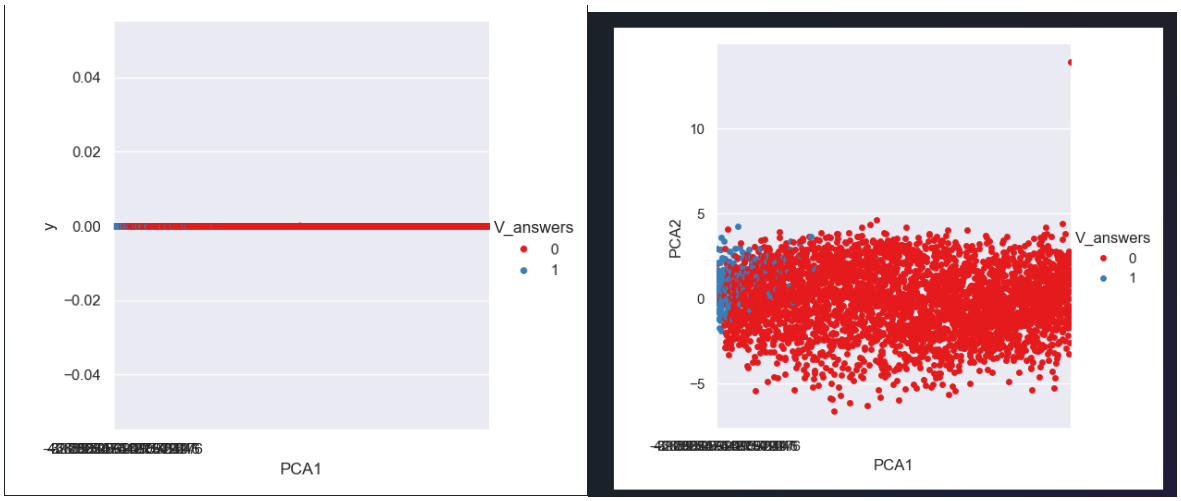
- **What it does**: Minimizes within-cluster variance, creating tightly packed, distinct clusters.
- **Insights for Weather Stations**:

- Weather stations group into 3-4 meaningful clusters, likely reflecting significant associations such as:
    1. **Regional Clusters**: Stations in similar latitudes or affected by similar atmospheric conditions.
    2. **Temporal Clusters**: Stations showing comparable trends (e.g., increasing temperatures or extreme weather events).
    3. **Outliers**: Isolated stations (e.g., island stations like Valentia) that don't conform to broader trends.
- **Outcome**: Most optimal clustering method for detecting associations in weather patterns among stations.
- **Relevance**: Best for ClimateWins to identify actionable insights like regional vulnerability or resilience to climate change.

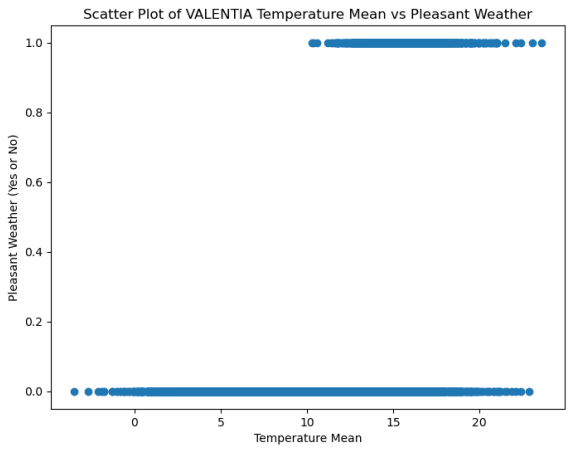


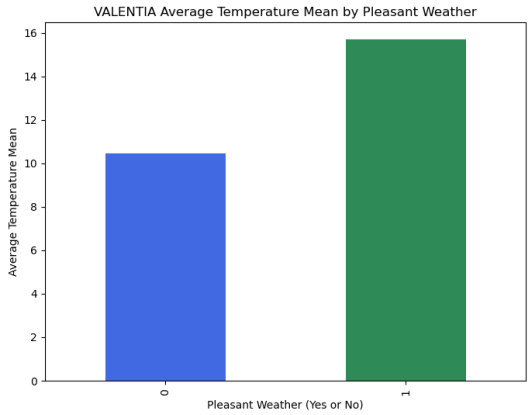
Cat Plot Using PCA:

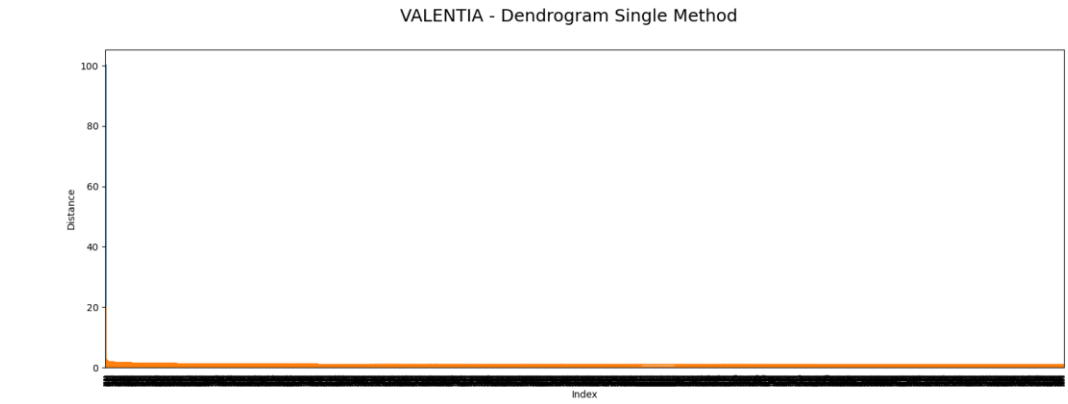
- Description: Displays dimensionality reduction outcomes for the Valentia weather station data, categorized by its "pleasant weather" outcomes.
- Interpretation: The plot revealed that in the last decade, there were no "pleasant weather" outcomes at the Valentia weather station, which is concerning for climate trend analysis.

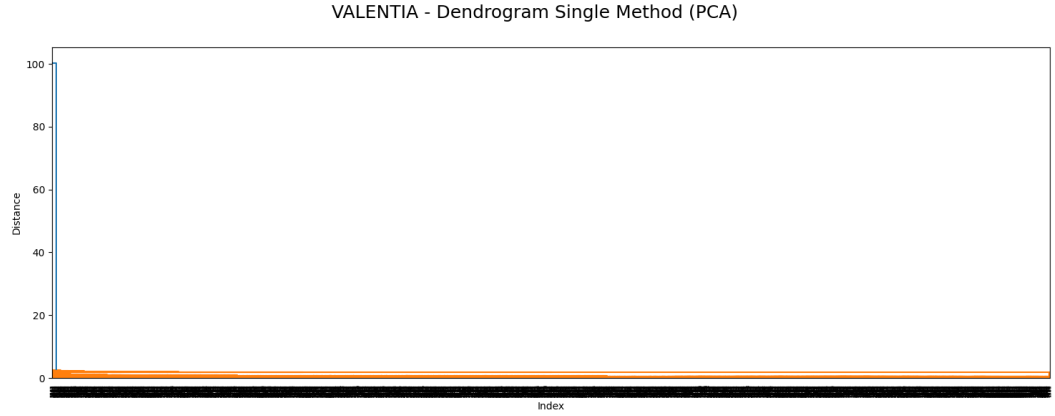


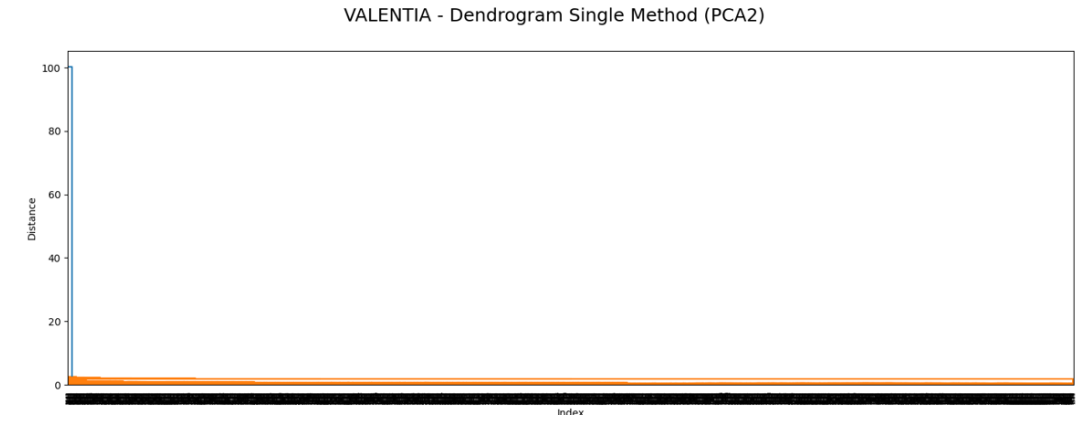
- As we can see, in the last decade, VALENTIA has had no pleasant weather outcomes. *[Yes = 1, No = 0]*





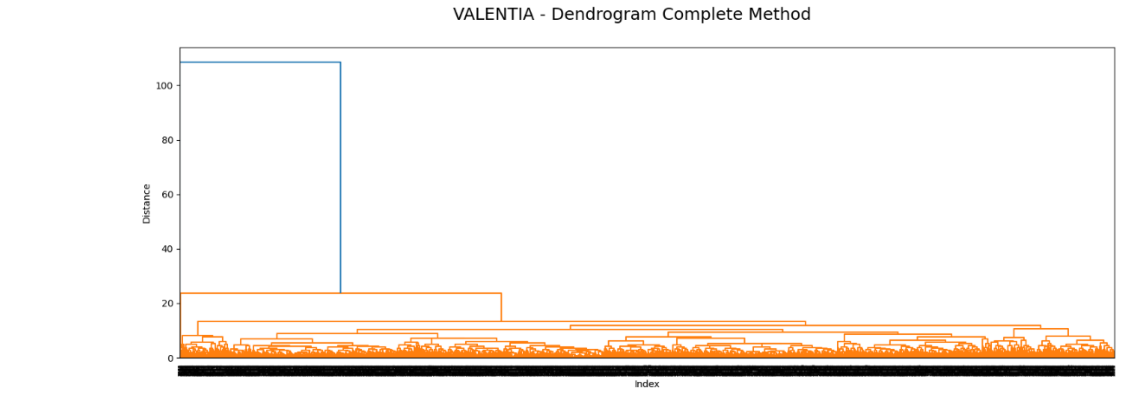


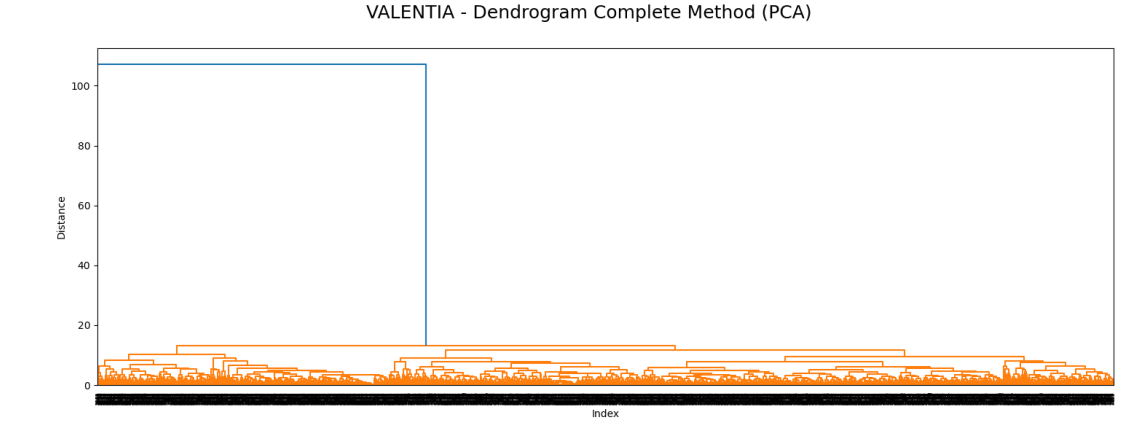


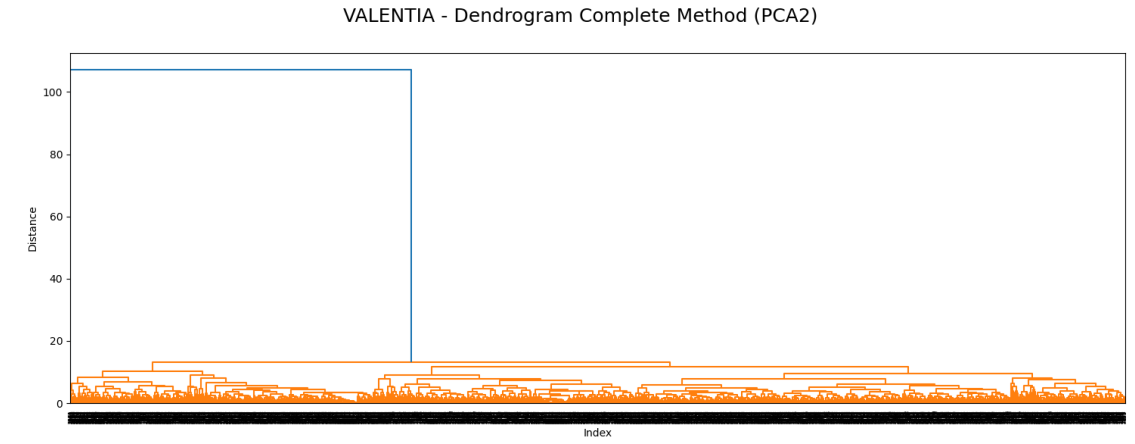


Single Method Dendrogram:

- **Observation**: Valentia does not form a meaningful cluster with other weather stations.
- **Reason**: The single method's chaining nature makes it hard to detect meaningful connections, especially for isolated stations like Valentia, which experiences a specific Atlantic-influenced climate.
- **Conclusion**: This method is ineffective for interpreting Valentia's data.

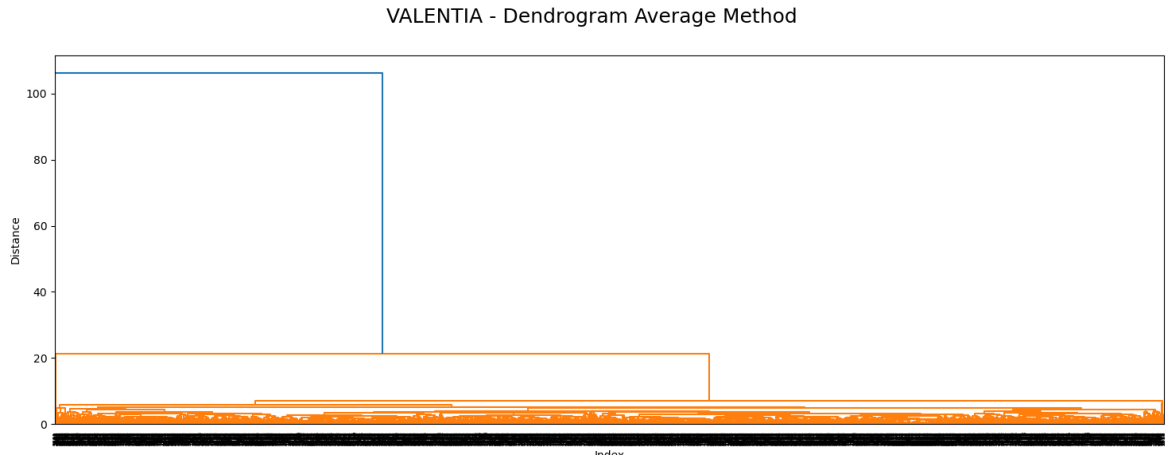


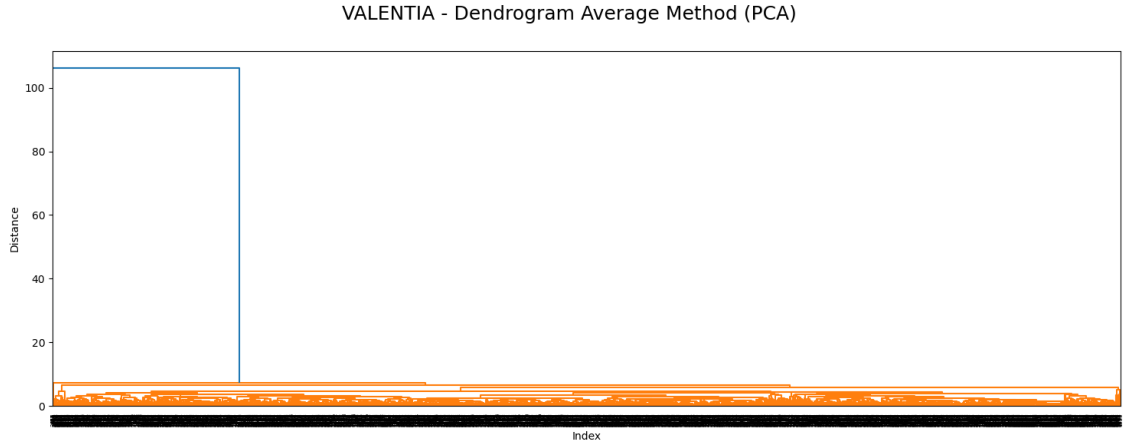


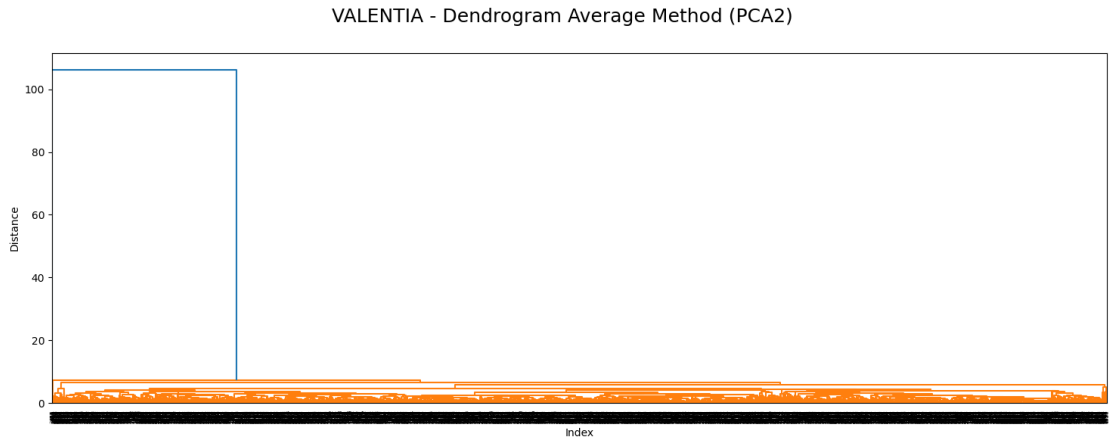


Complete Method Dendrogram:

- **Observation**: Valentia begins to cluster with stations that might share similar proximity to coastal regions or Atlantic influences.
- **Reason**: Complete linkage considers maximum distances and groups stations with overarching similarities (e.g., oceanic climates or comparable temperature ranges).
- **Conclusion**: Provides moderate insight but lacks granularity compared to Ward's method.

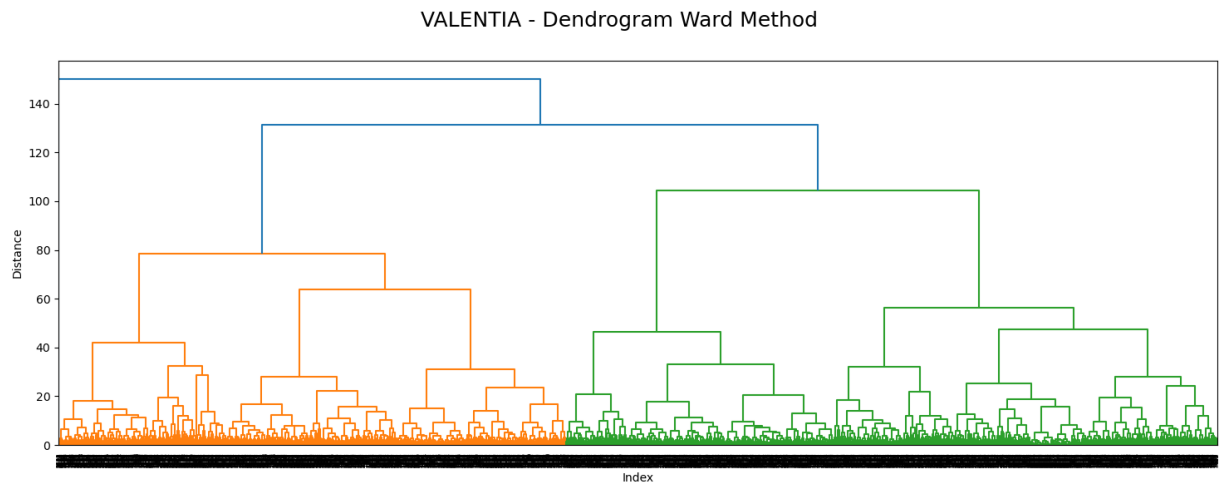


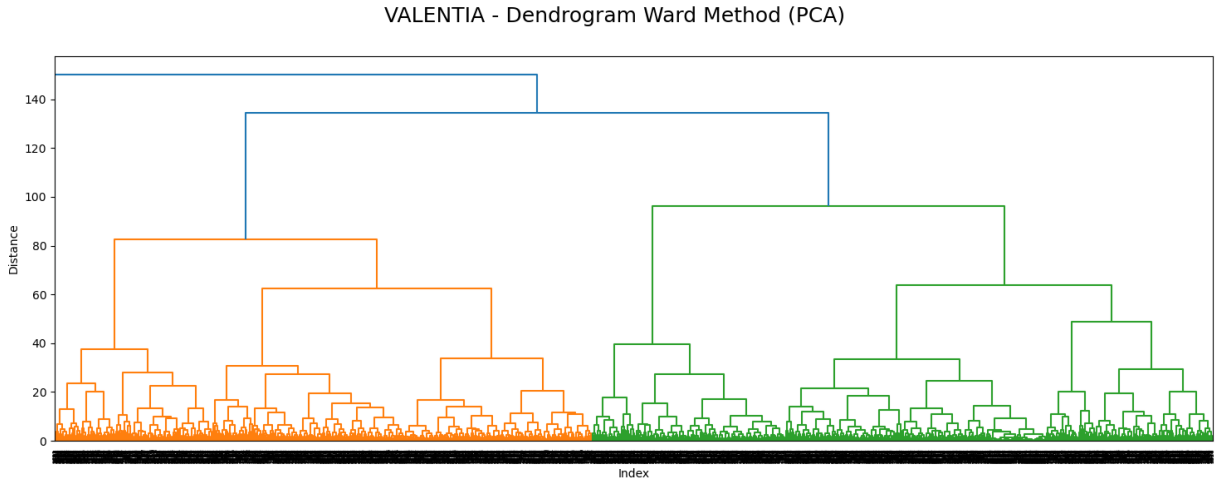


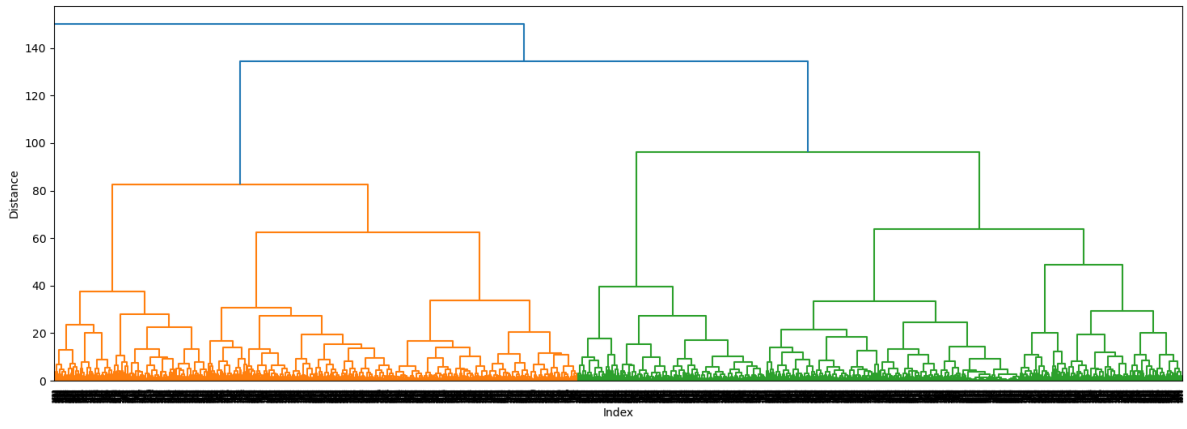


Average Method Dendrogram:

- **Observation**: Valentia clusters with stations experiencing similar weather patterns, likely coastal or temperate zones.
- **Reason**: Average linkage smooths out extremes, highlighting Valentia's connection to broader climate categories.
- **Conclusion**: Helps identify macro-level trends but is less effective at uncovering localized patterns..







Ward Method Dendrogram:

- **Observation**: Valentia emerges as a distinct or semi-distinct cluster.
- **Reason**: Ward's method minimizes within-cluster variance, making it easier to spot Valentia's divergence from mainland and continental stations.

- **Key Insight**: Valentia's unique climatic conditions (frequent rainfall, mild temperatures, Atlantic exposure) make it stand out.
- **Conclusion**: This method offers the best insights into Valentia's role as a climatic outlier and its broader implications for Atlantic weather pattern

Answers to Key Questions:

1. **Clustering Insights**:
    - **Ward Method**: Produced the most distinguishable and interpretable clusters, making it the best choice for ClimateWins.
    - **Average and Complete Methods**: Delivered similar but slightly less effective results compared to Ward. Both are viable secondary choices.
    - **Single Method**: Ineffective due to its chaining nature, making it unsuitable for analyzing weather data.
2. **Pleasant Weather Labels and Clusters**:
    - The reduced dataset clustered using Ward's method showed alignment with the pleasant weather outcomes, suggesting this method can effectively group weather data relevant to ClimateWins' focus.
3. **Scaled vs. Unscaled Data**:
    - Scaled data produced better results for clustering, especially with the Ward and Complete methods, highlighting the importance of scaling when working with weather data.
4. **Dimensionality Reduction via PCA**:
    - Reducing the dataset to 1, 2, or 5 components via PCA yielded comparable clustering results. Dimensionality reduction simplifies the dataset without significantly compromising clustering quality.
5. **Challenges Identified**:
    - Overlapping data label points.
    - PCA categorization by pleasant weather outcomes sometimes resulted in inconclusive clusters.

Final Recommendation:

Ward's method with scaled data and PCA-reduced datasets is the most effective approach for ClimateWins to explore its weather data. This method provides meaningful clusters that align well.