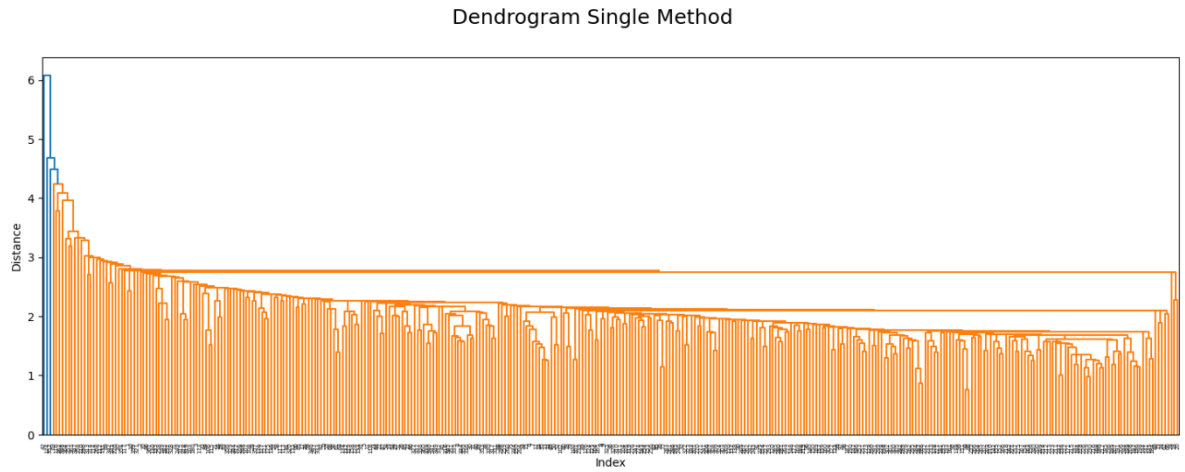
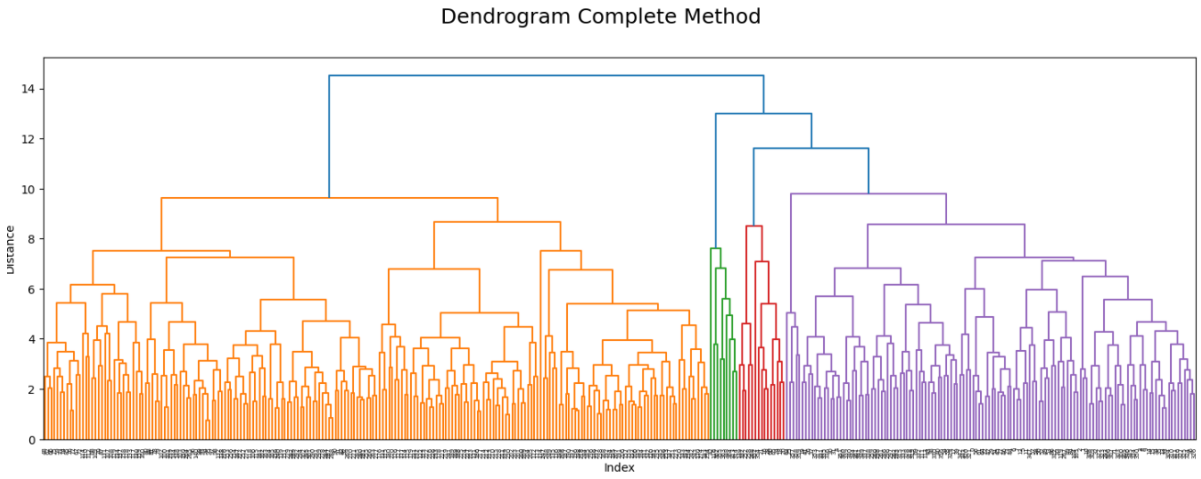
Unsupervised Machine Learning with Python: Dendrograms and PCA

Dendrograms: Madrid and Heathrow 1995

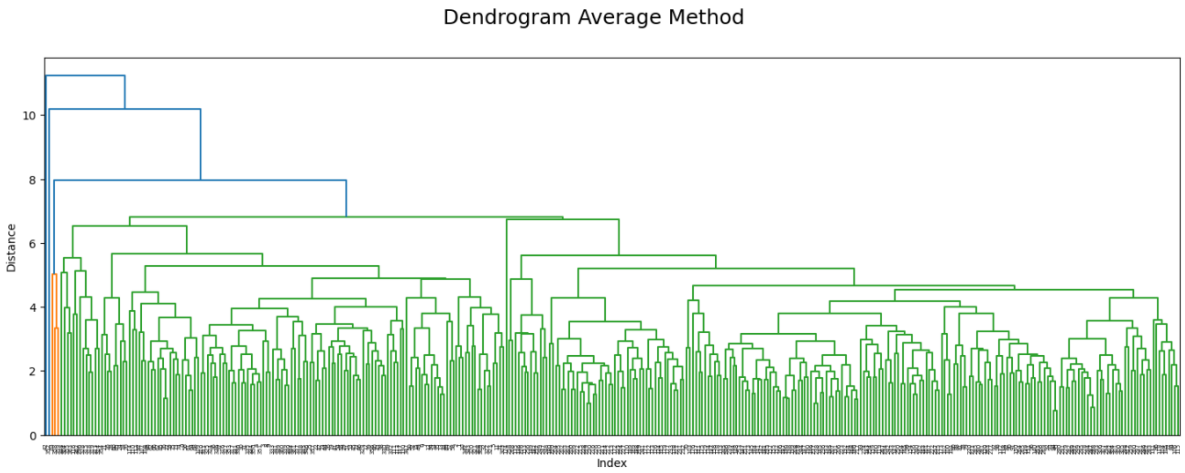
Single Method: The distance between the two closest members of each cluster to determine which group they should belong to.



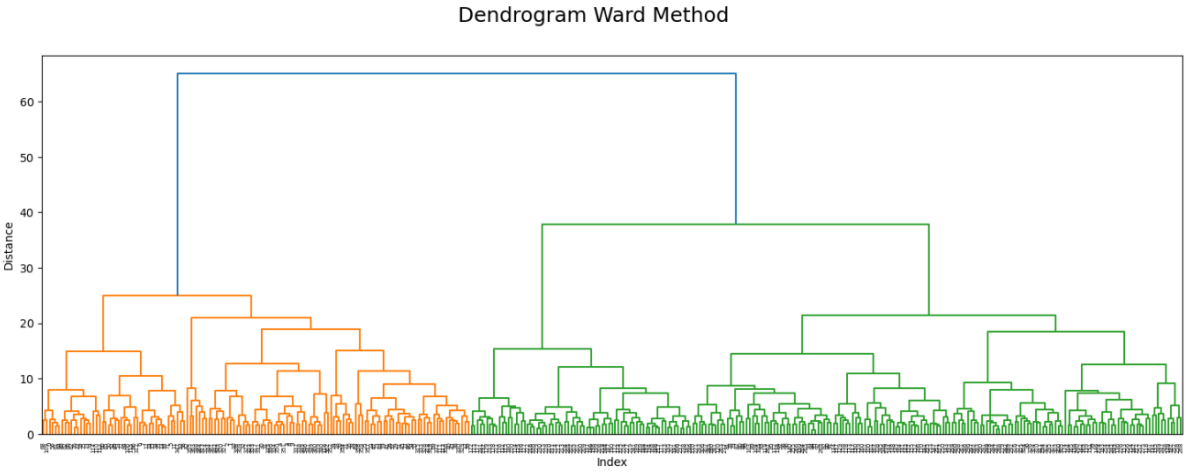
Complete Method: The distance between the two farthest members of each cluster to determine which group they should belong to.



Average Method: The distance between the average of the members of each cluster to determine which group they should belong to.



Ward Method: This uses a metric called the minimum increase of sum of squares (MISSQ) to find the distance between two clusters. It attempts to minimize the variance between the two clusters.



Observations:

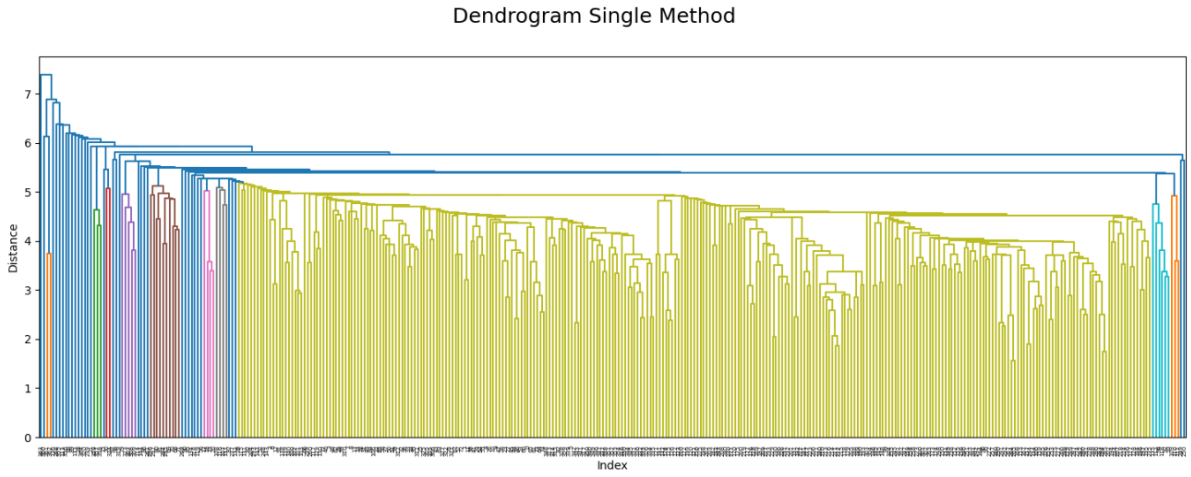
- **Single Method** – It has over generalized most of the data creating one huge cluster that contains most data points and a small cluster. This is probably separating the data into outliers and the rest of the data.
- **Complete Method** – Has created 4 main clusters 2 big and 2 small, this could be seasonality based (4 seasons) or even warm seasons (spring, summer) and cold seasons (autumn and winter) with their respective pleasant weather and unpleasant weather
- **Average Method** – Similar to the Single Method it has overgeneralized most of the data creating one big cluster and 2 very small clusters.

- **Ward Method –** It has created the most amount of subcategories but 2 main clusters probably accounting for seasonality or geographic location (Madrid and Heathrow)

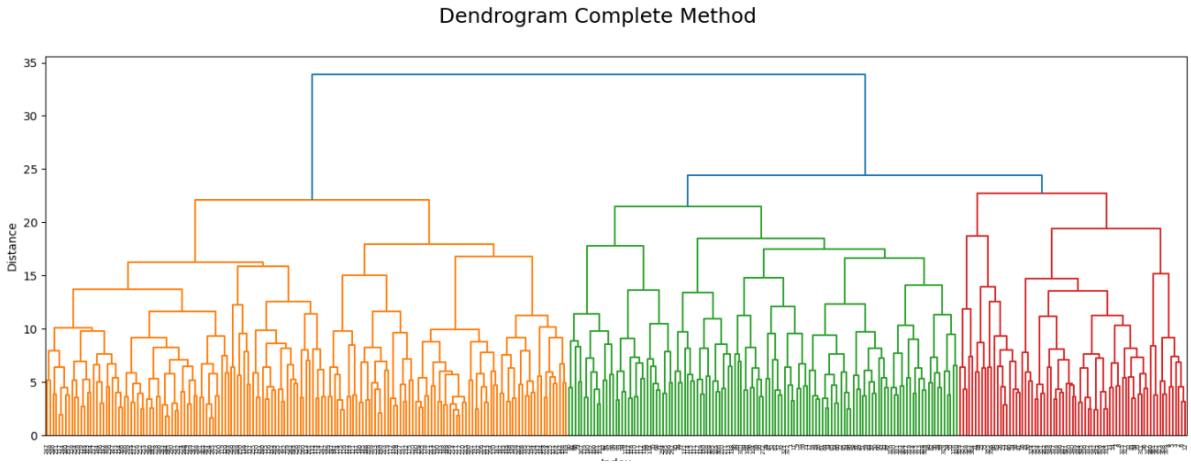
Each Method is Unique and gives varied insights, given the seasonality elemental and geospacial characteristics in my opinion the Complete or the Ward Method are best suited to our analysis.

Dendrograms: All Stations with Reduced Data

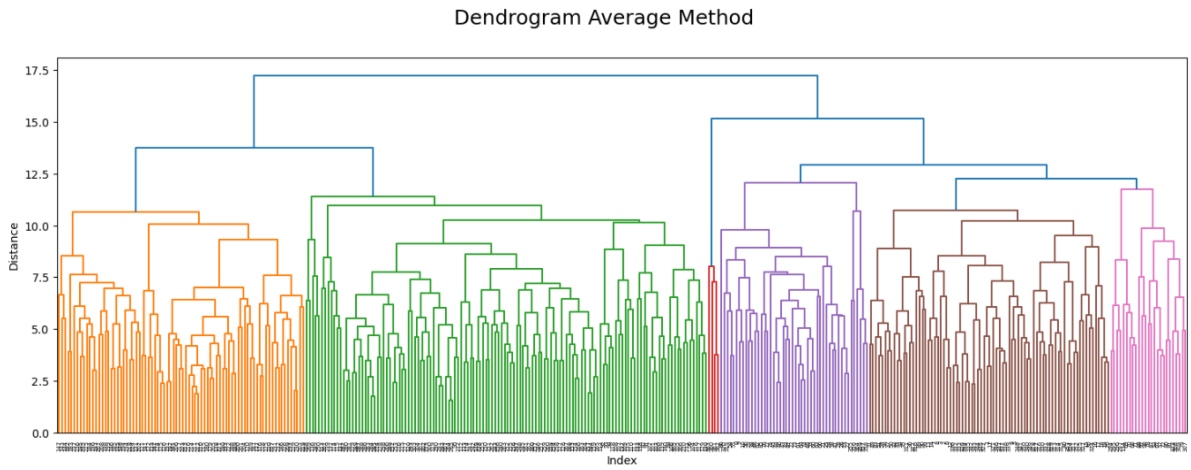
Single Method: The distance between the two closest members of each cluster to determine which group they should belong to.



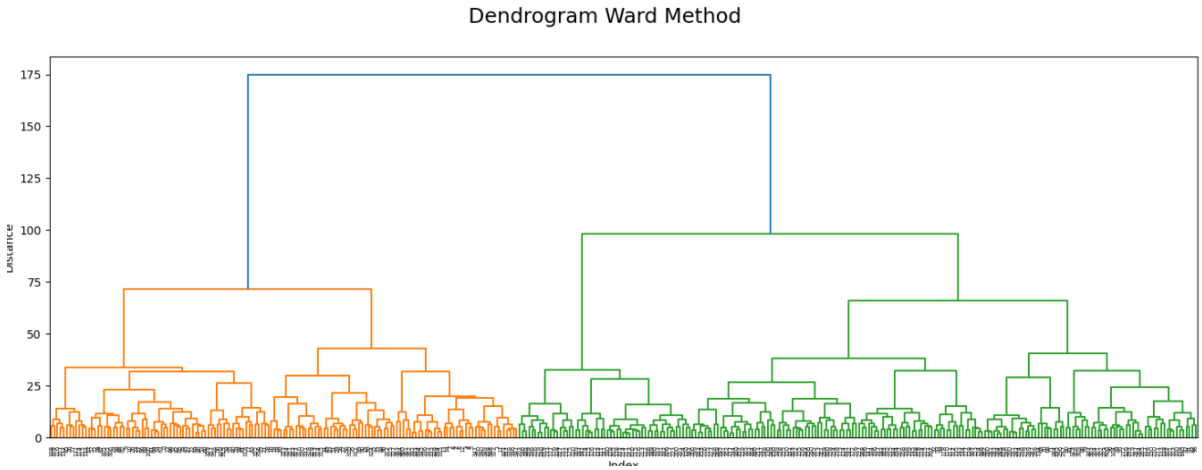
Complete Method: The distance between the two farthest members of each cluster to determine which group they should belong to.



Average Method: The distance between the average of the members of each cluster to determine which group they should belong to.



Ward Method: This uses a metric called the minimum increase of sum of squares (MISSQ) to find the distance between two clusters. It attempts to minimize the variance between the two clusters.



Observations:

- **Single Method –** Has Created 9 clusters one for each PCA column however most of the data resides in 2 clusters (Navy and Gold) hard to assume a lot from this.
- **Complete Method –** Good clustering Separation with 3 main clusters and no overlap this is probably due to location or seasonality.
- **Average Method –** Has created 6 distinct clusters this could be due to a combination of location and weather conditions.
- **Ward Method –** Has created to main clusters again categorizing the data in a similar way to the un-reduced data. This could likely be Pleasant and Unpleasant weather categories.