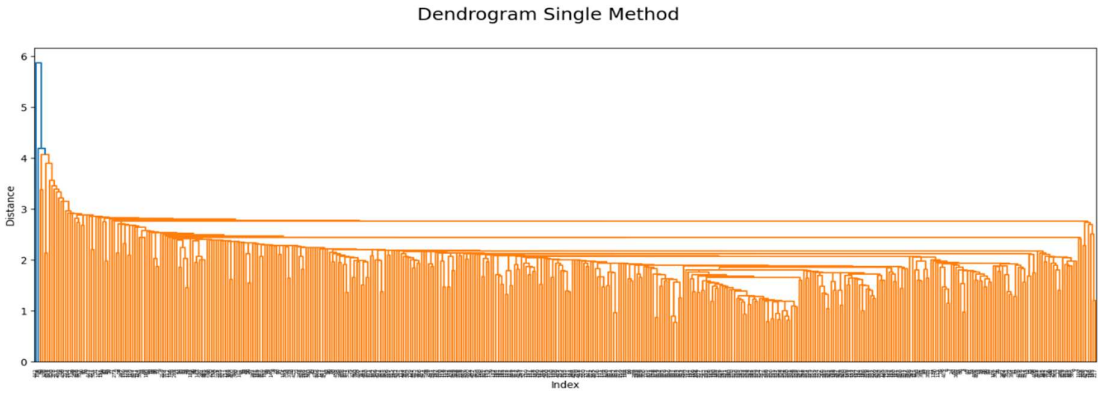
Exercise 2.1: Unsupervised Learning Algorithms

Timothy Aluko

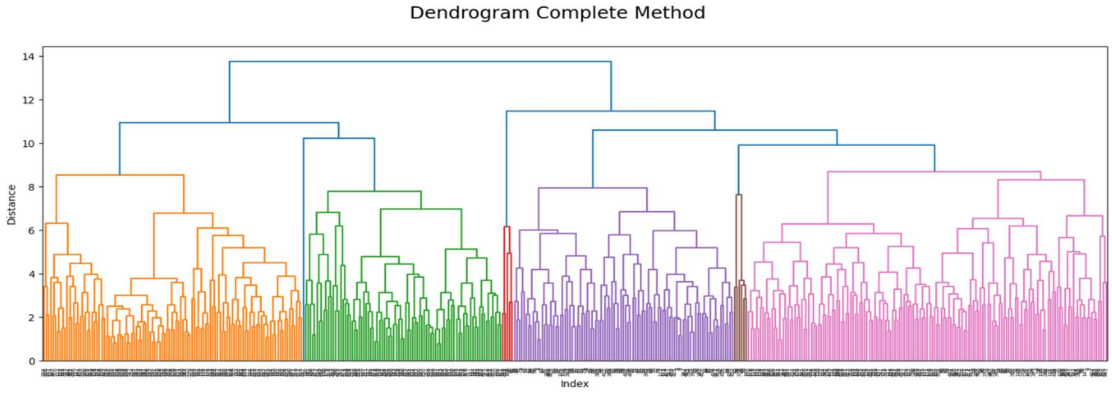
Overview:

To comprehend the definition of pleasant weather at various weather stations, the exercise employed an unsupervised learning machine to examine ClimateWins weather data, identifying clusters and trends. This process aims to identify any significant correlations between the weather conditions at different stations by comparing the clustered data to a predetermined "pleasant weather" standard. The idea is to reveal weather patterns, parallels, and deviations between stations and provide insights into local weather traits by improving ClimateWins' weather and climate data trends.

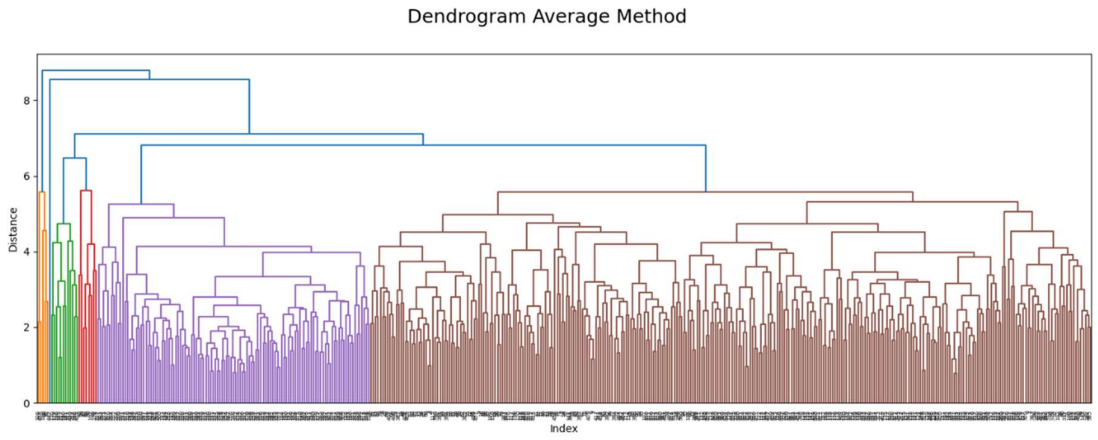
1. Dendrograms comparing Madrid and Belgrade in 2010



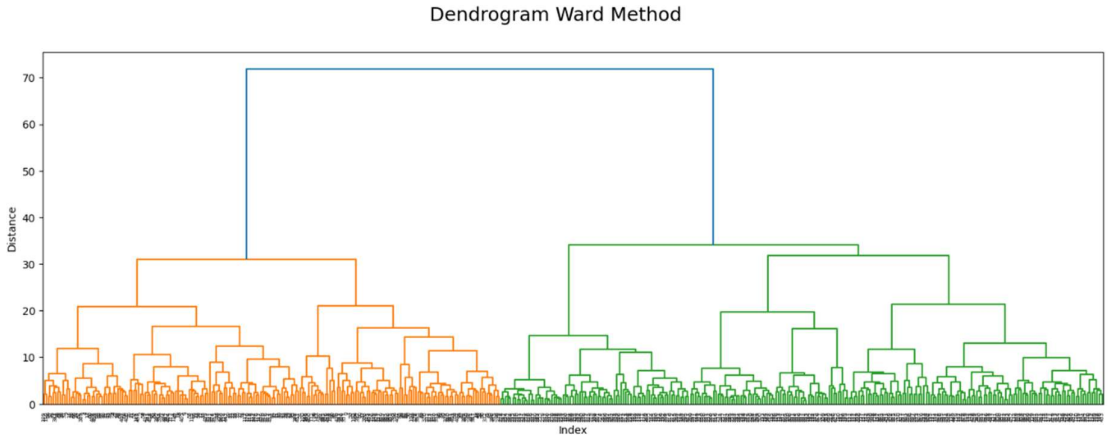
Single method: This analysis of the two closest members of each cluster indicates that the orange represents the majority of 2010 days. This analysis does not yield any significant findings. The small group on the left was likely regarded as an outlier.



Complete method: By examining the distance between the furthest members of each cluster, which is more evenly divided into six categories, this presents a more uniform distribution for the year. To understand these clusters' trends, this may be worth investigating.

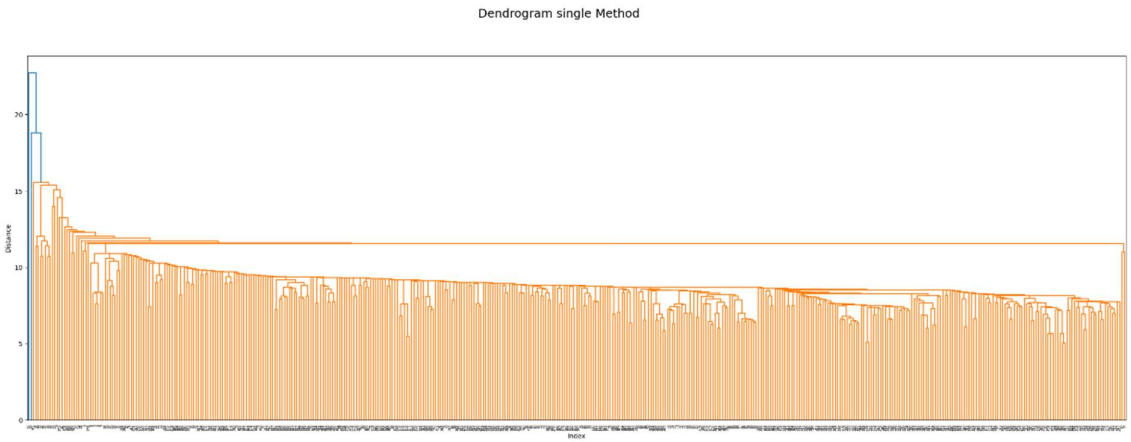


Average method: It divides the year into five groups. Given that this approach examines the distance between the average of each cluster, it may be possible to check if the weather changes within the broader context of the 2010 data.

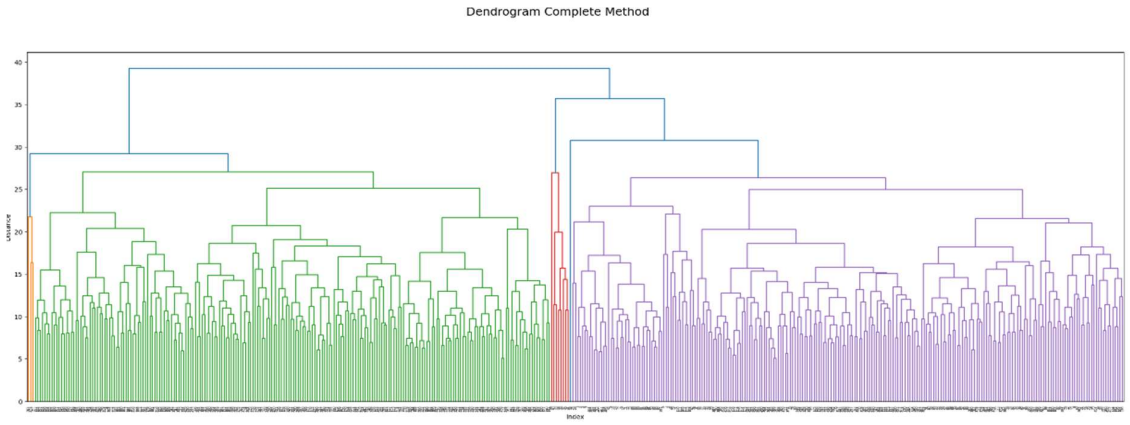


The Ward method examines the distance between two clusters and aims to minimize the variance between them.

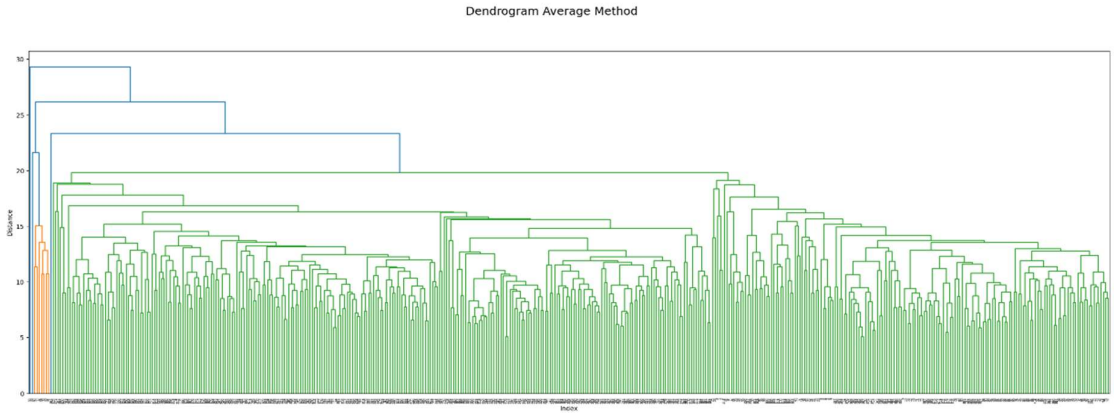
2. Dendrograms comparing all stations in 2010



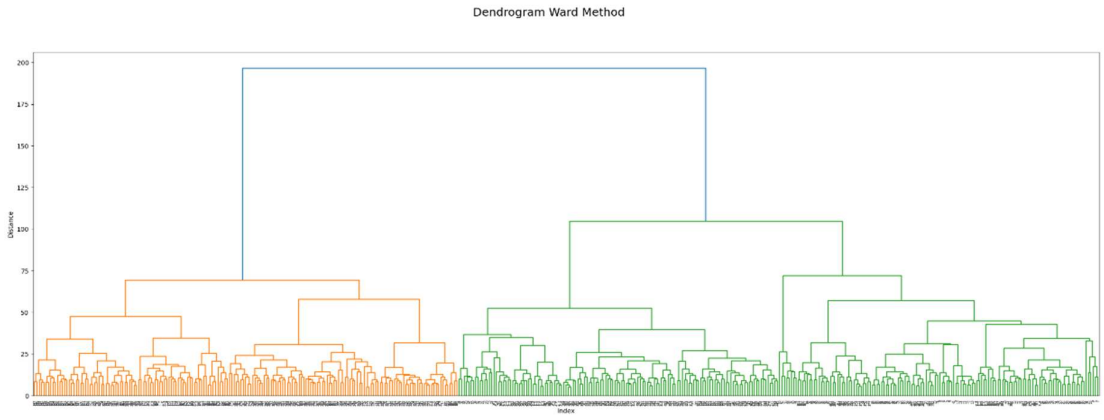
Single method: All stations and days are lumped into one category, which is not useful.



Complete method: This is divided into four categories: two large and two smalls. Almost equally, it divides into two clusters, each smaller than the other.

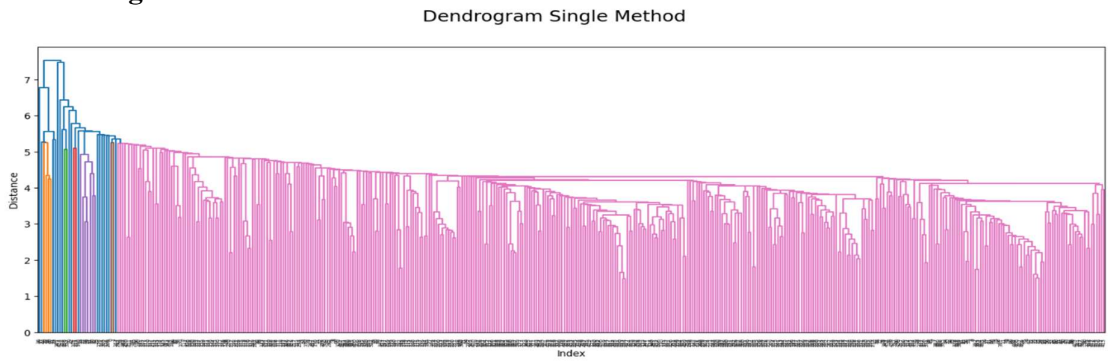


Average Method: The data has clustered into two groups, likely due to overfitting in the green category, and does not appear to provide meaningful insight. This could be a result of having a split between average temperatures within the 18 weather stations (it might be worth exploring if this is due to locations that have hotter vs. cooler temps).

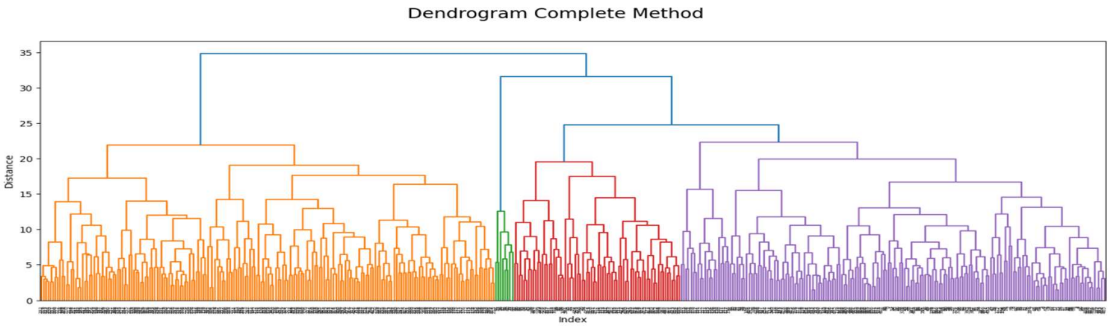


Ward Method: This split all locations into two groups for 2010. Again, this could be because of how their position affected average temperatures in general.

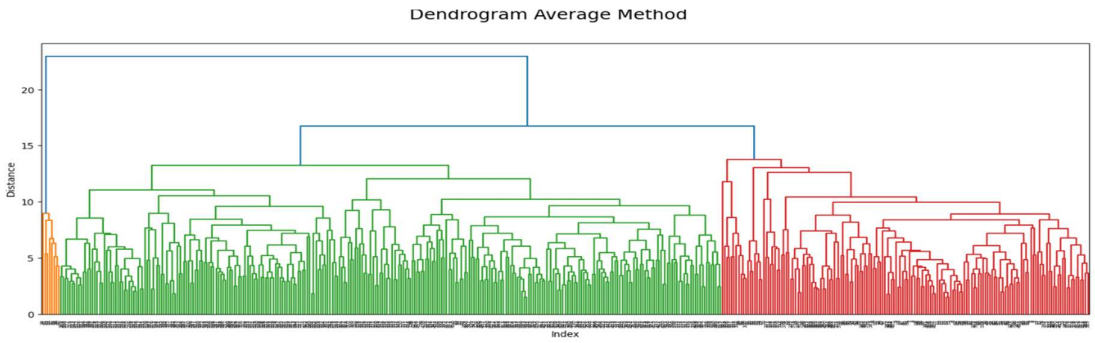
3. Dendrograms with reduced data for 2010



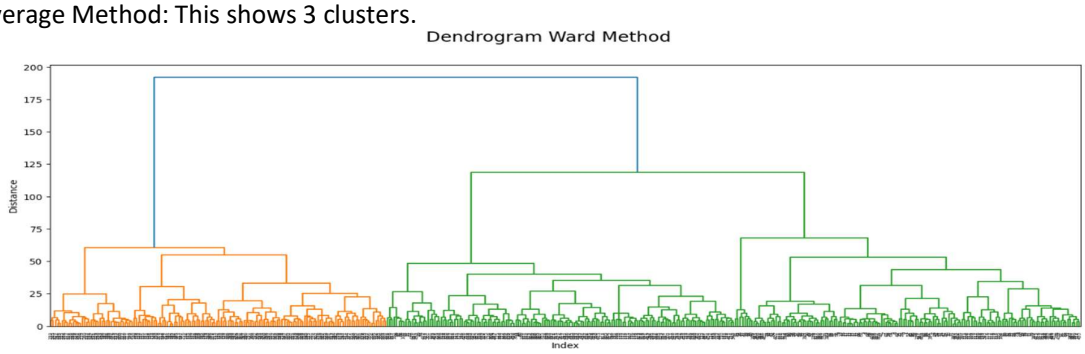
Single Method: shows more categorizations, but mostly just one; is not likely to be completely useful.



Complete Method: That's a nice mix of clusters. Location or time of year could cause this.



Average Method: This shows 3 clusters.



Ward Method: This shows 2 clusters.

The Complete Method offers a more detailed visual representation of data patterns by comparing the maximum distances between clusters. This approach clusters the data, including both unscaled and reduced forms, into more evenly distributed clusters, offering a balanced view of relationships within the dataset. Based on this information, it is ideal to apply unsupervised algorithms for future analysis.