ML – Unsupervised Algorithms

Hierarchical Clustering and Dimensionality Reduction

ClimateWins wants to see how data is clustered by an unsupervised learning algorithm compared with its database of "pleasant weather." Are there any connections that stand out among the weather stations from which data has been collected?

Data Preparation

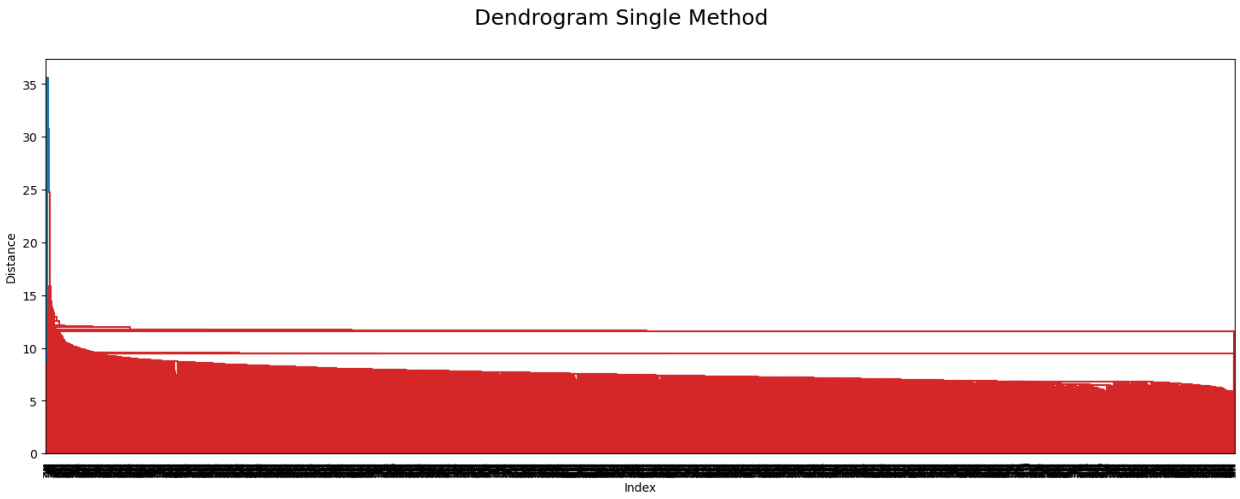
- Imported unscaled data set and answers data set
- Dropped the columns not present in the answers data set to provide consistency
- Reduce data set to the decade from 1990 to 2000
- Dropped the date and month columns
- Scaled the data set

Note: a quick look at the value counts for the answers data set shows that there are usually more "0" (unpleasant) days than "1" (pleasant) days. The ratio is about 7 to 3. That means we will be looking for two clusters of unequal size.

Hierarchical Clustering Methods

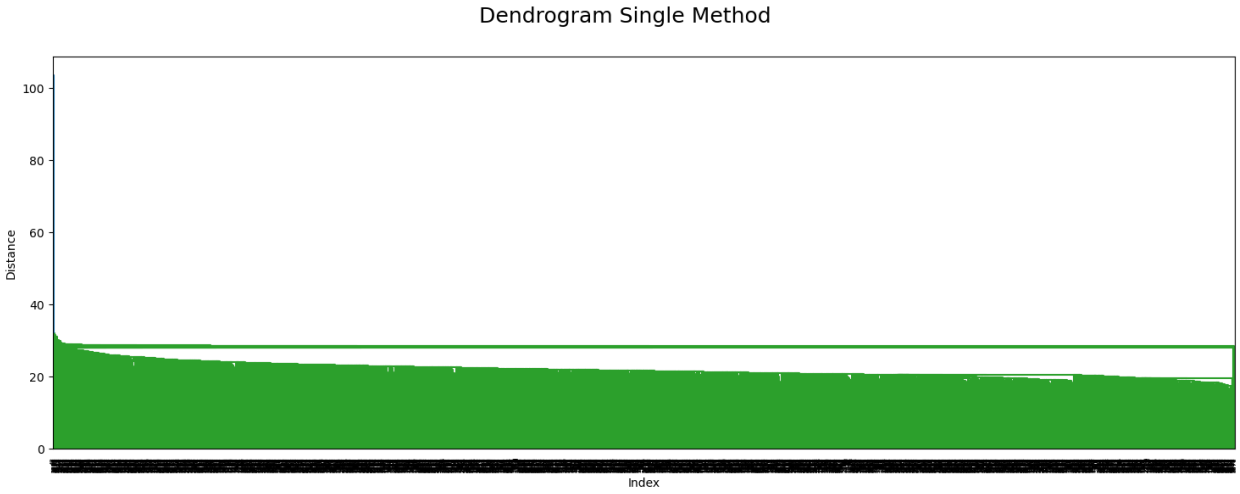
- Single – groups by distance of two closest members of cluster
- Complete – groups by distance of two farthest members of cluster
- Average – groups by distance of average of members of cluster
- Ward – uses MISSQ to minimize variance between clusters

Dendrograms on All Weather Stations from 1990 to 2000



Single Method on Scaled Data:

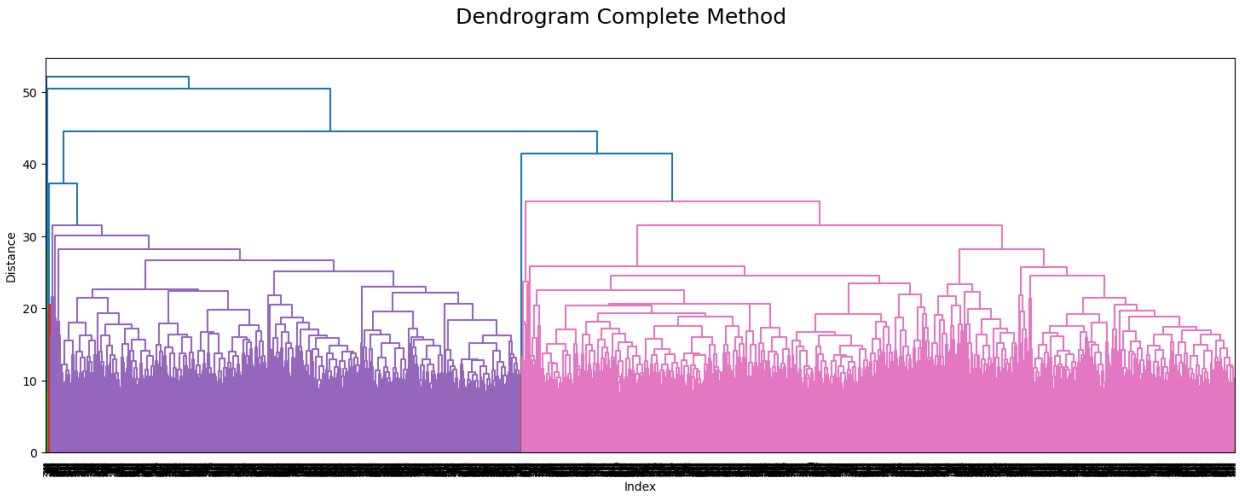
- One cluster found
- No discernable location where we could “cut” the tree
- Could be that all points are too close for algorithm to separate into clusters



Single Method on Unscaled Data:

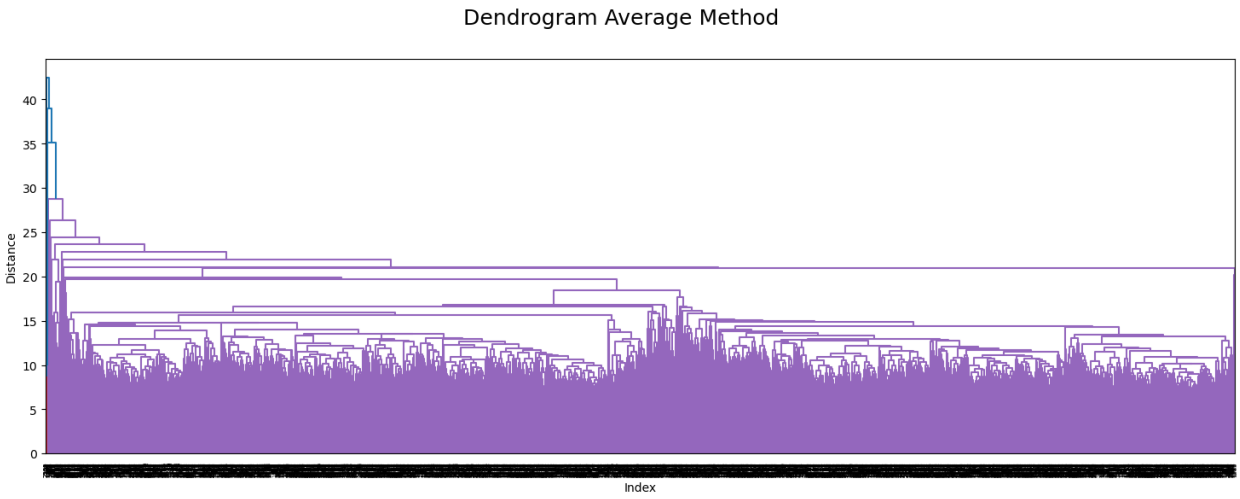
- Only one cluster

- No improvement on the result of the dendrogram for scaled data
- Method not suitable for clustering



Complete Method on Scaled Data

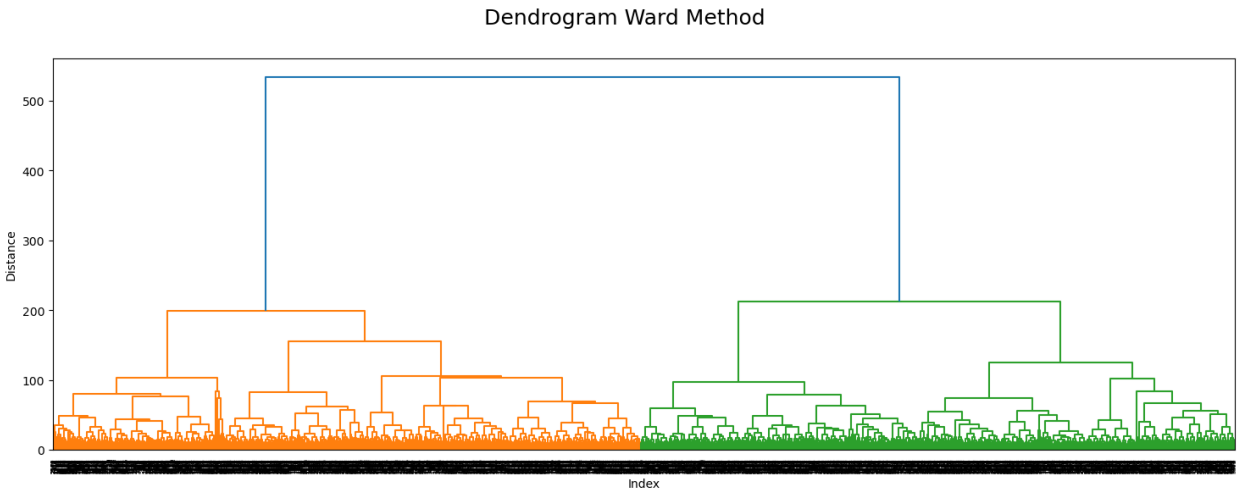
- Two discernable clusters
- Slim cluster on the left side of dendrogram, may be third cluster?
- Division of large clusters seem to match the answers groups



Average Method on Scaled Data:

- One cluster

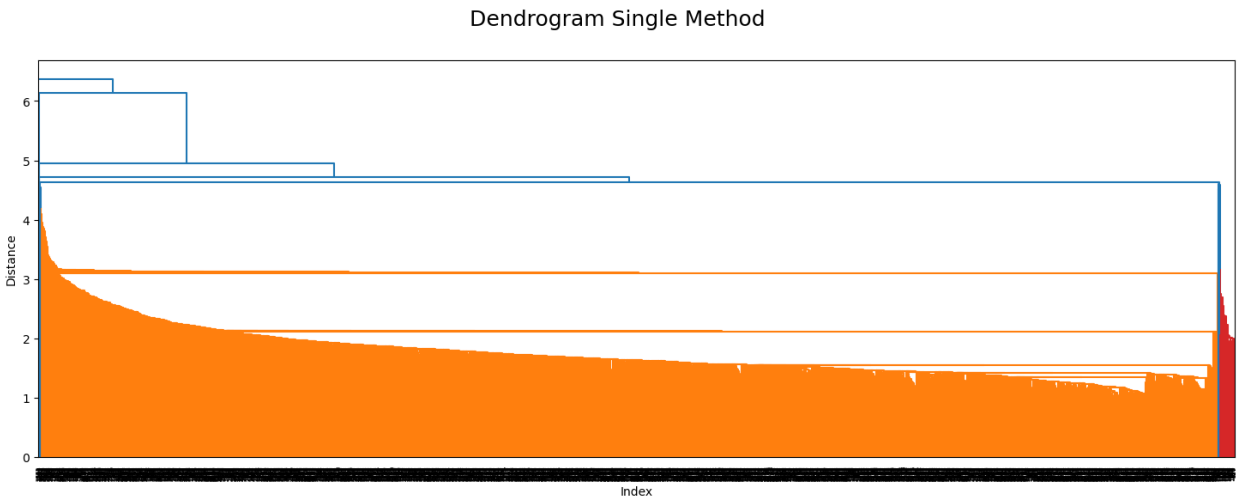
- Hierarchy visible (as opposed to single method) but no visible place to cut the tree.
- Average of distances may be all too close for the algorithm to separate into smaller groups.



Ward Method on Scaled Data

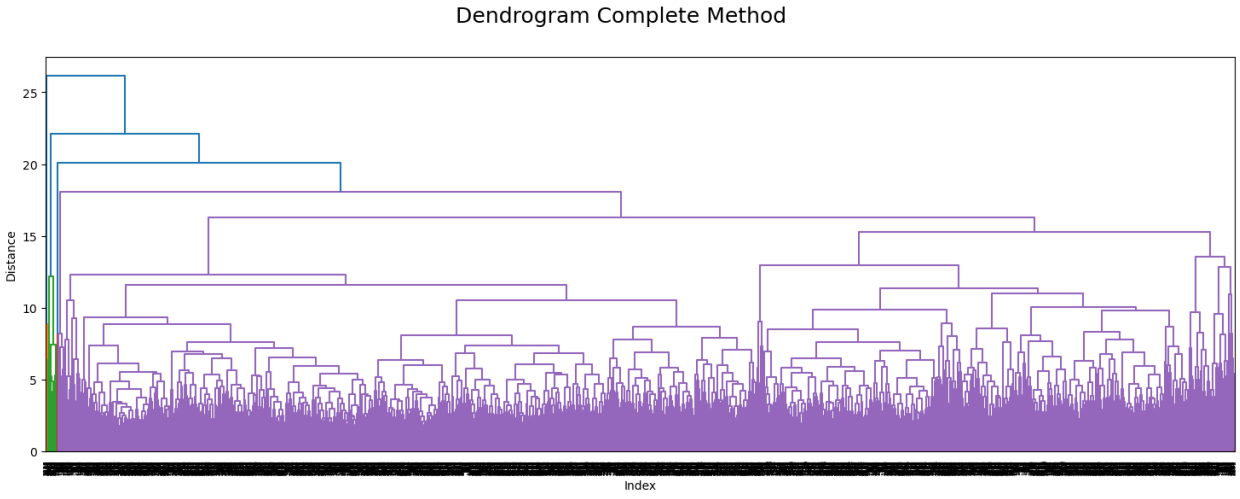
- Two equal clusters visible
- As the ratio of the clusters don't match the 'pleasant' data set groups, this ward method may be clustering the data on the winter/ summer seasons, which show more even distribution.

Dendrograms on Basel and Madrid Stations (only scaled data)



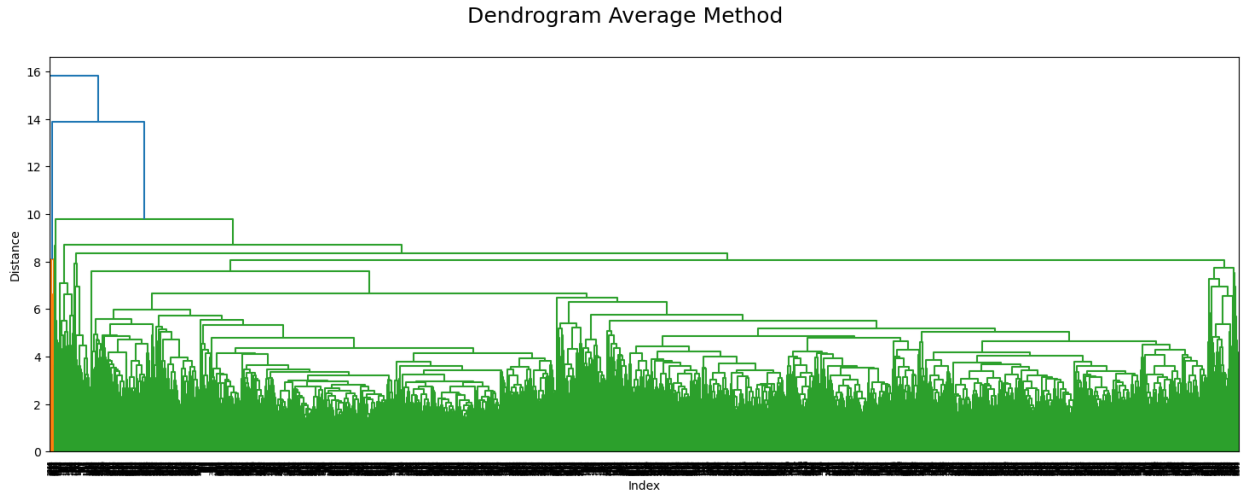
Single Method

- Two clusters
- One cluster is significantly larger than the other, which does not match what we know of the answers.
- The small cluster may be the outliers.



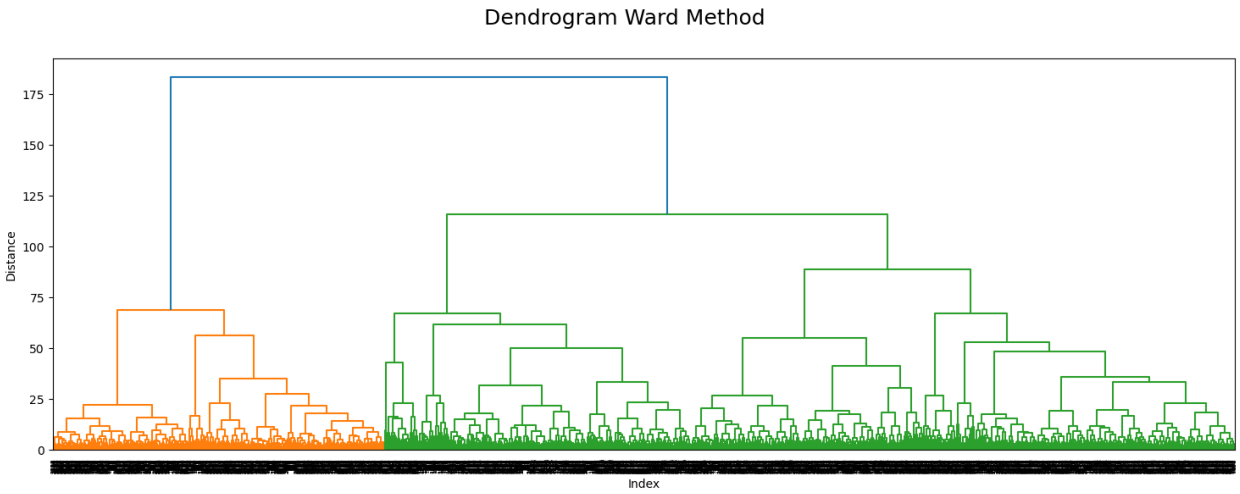
Complete Method

- Mainly one cluster
- Second cluster is very small, again perhaps due to outliers.
- Possible to “cut” the tree at an earlier point to make two clusters



Average Method

- One main cluster
- Very small and insignificant second cluster
- No clear cutting point for the main cluster



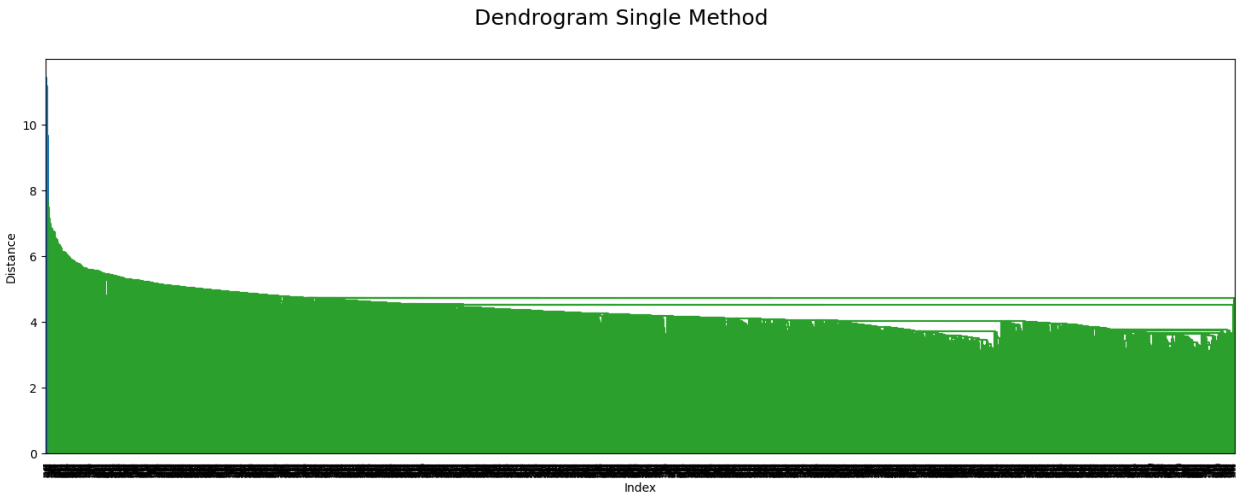
Ward Method

- Two discernible clusters
- Three clusters also possible if cut
- Proportion matches the 70/30 of the ‘pleasant’ answers data set.
- If cut, the third possible cluster may be formed due to how ‘unpleasant’ weather would be defined by differing geographical regions.

Dimensionality Reduction – PCA

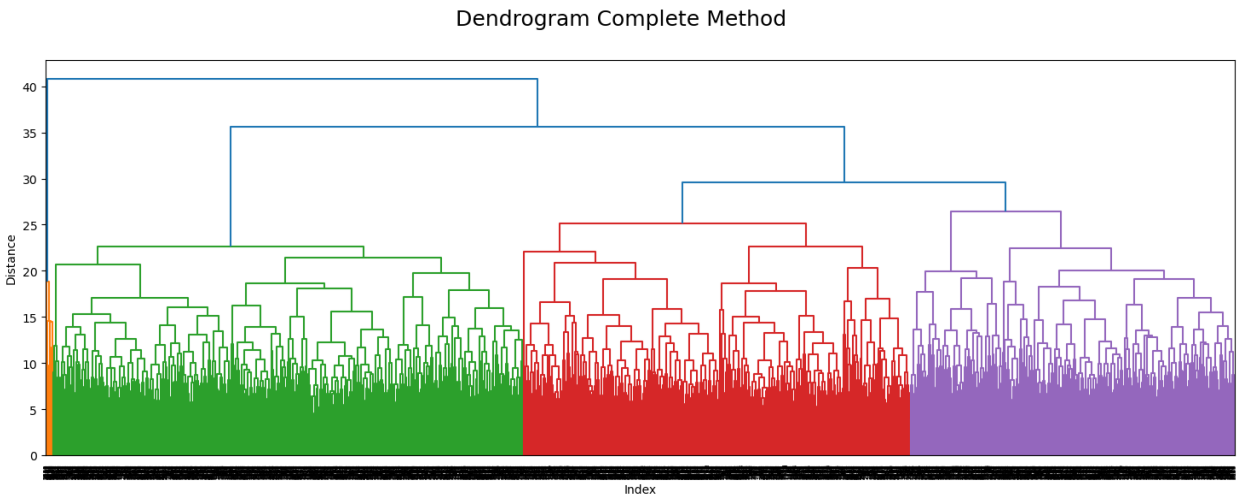
- Reduced data to 15 dimensions (number of weather stations).
- New reduced dataset exported as new csv file.

Hierarchical Clustering on reduced data



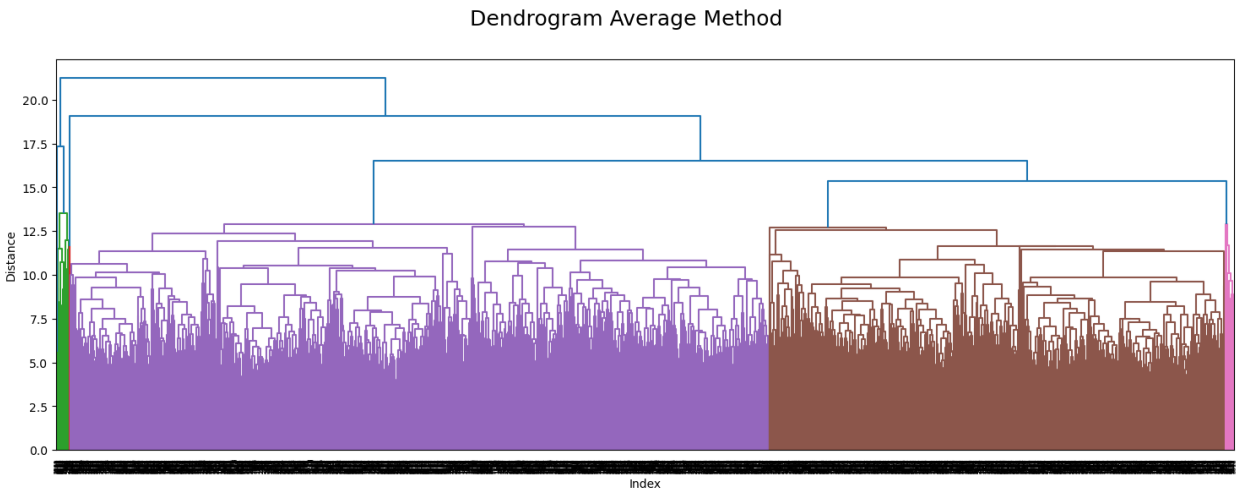
Single Method

- No difference to the results of single method on full data – one cluster



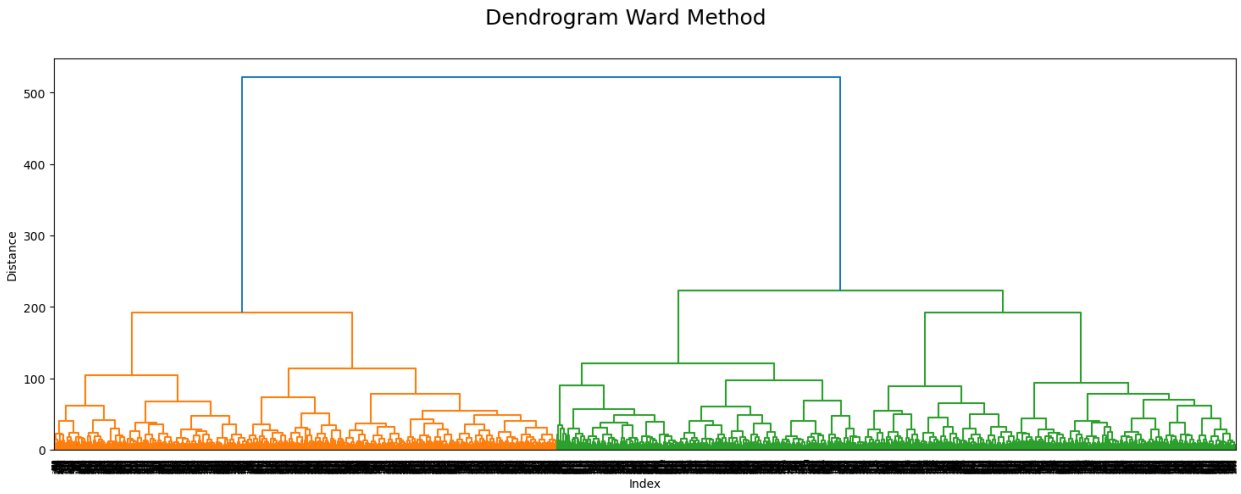
Complete Method

- Three main clusters, one very small one toward the left
- Does not match answers data set
- Division of clusters may be based on regionality



Average Method

- Two large clusters and two very small ones on the edges
- Divided into two groups like in the answers
- The extra small clusters may indicate outliers



Ward Method

- Clean separation into two clusters
- Proportion could match, as one group is slightly larger.

Conclusions and Insights

- The single method works poorly for this dataset; all data points are chained together and form one large cluster.
- In all three data set analyses (all stations, only two stations, and reduced) the **ward method** seems to be most successful in clustering into the two groups that match the answers data set.
- Reducing the dimensions of the original data set increases the likelihood of the algorithm detecting clusters (more clusters in the complete/ average methods).
- If we are looking for clustering that matches our answers data set, the closest result would be the ward method on only two weather stations.