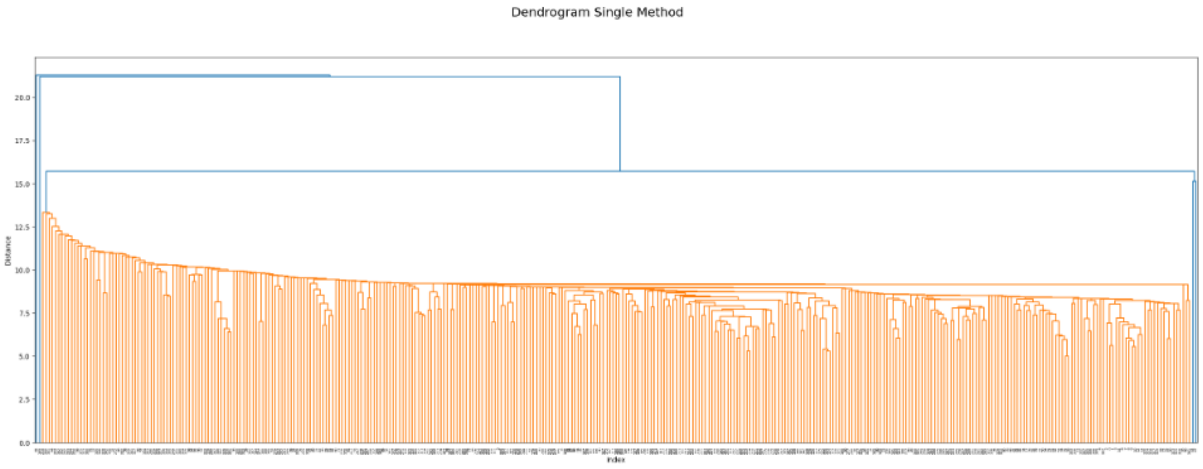
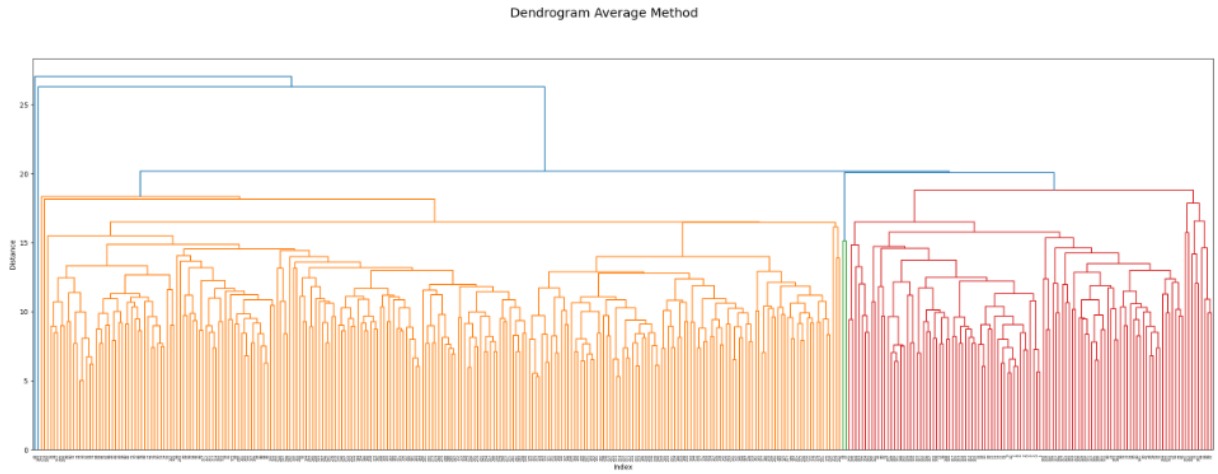
Dendrograms for all weather stations in 1990

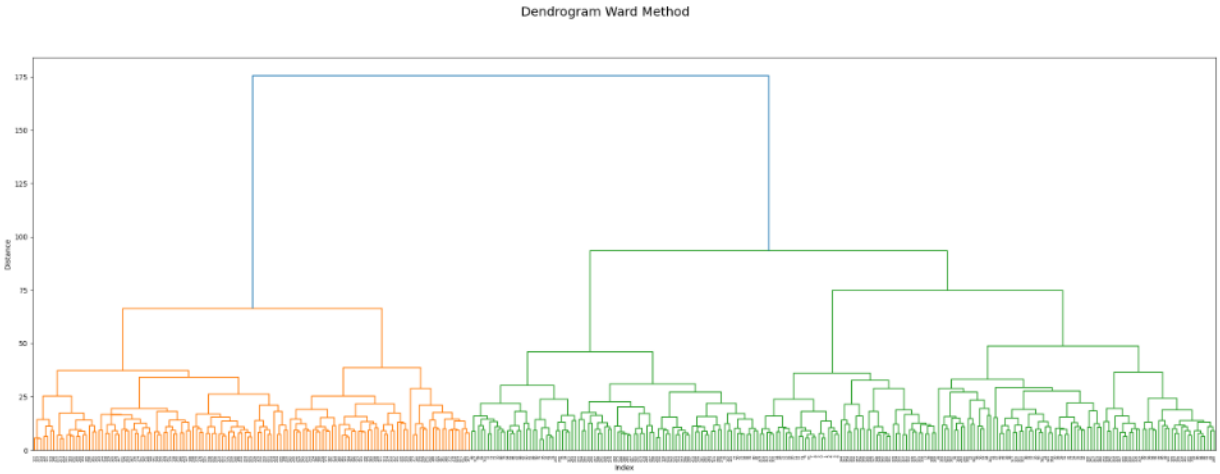


Single method: this method looks at the two closest members of each cluster. Given the unform orange, most weather stations in 1990 are linked together in long, chain-like structures, making it hard to pull useful insights. Moving farther down, there are groupings at a finer scale, indicating that between the stations there are shared characteristics like similar temperature ranges, wind patterns, or precipitation levels, to name a few.

Complete method: This method uses the distance between the farthest members of each cluster. Readily apparent are the five colored clusters that are easily distinguishable from each other, showing several broad groupings at the higher levels. In this case, this method indicates that certain weather stations are exhibiting relatively unique climate profiles compared to the others, which leads to the separation at higher levels. Also, the cluster sizes are balanced, indicating each cluster housing multiple stations.

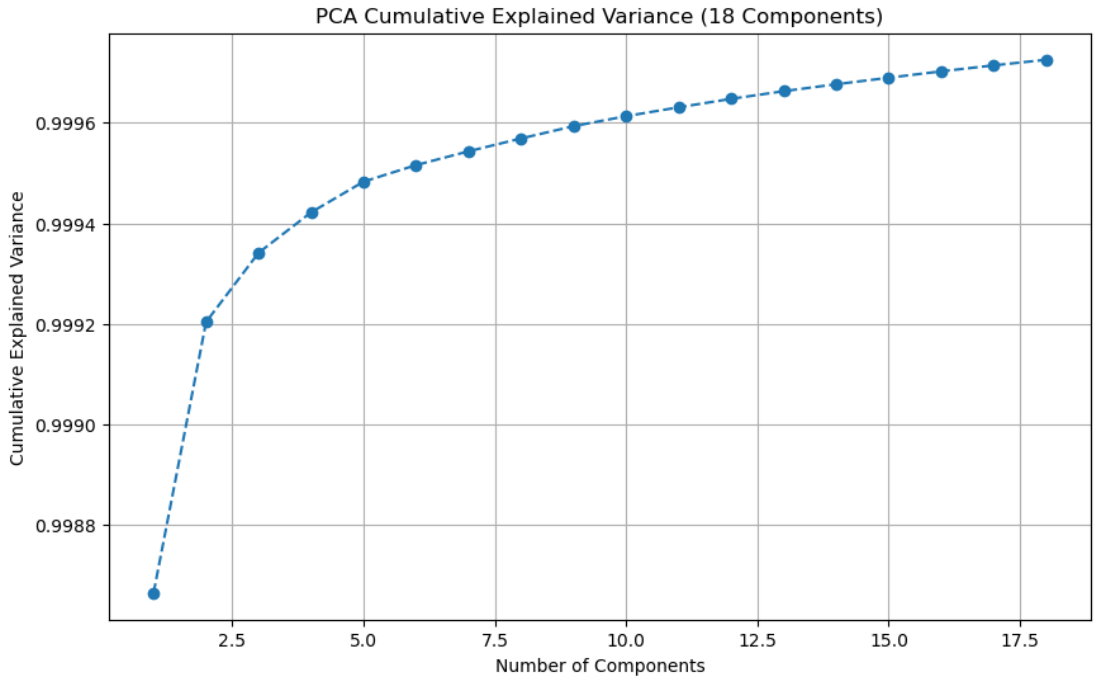


Average method: This method takes the distance between the average of each cluster. We can see about three distinguishable clusters here, with one being extremely compact (perhaps one unique station) and two larger, more encompassing clusters. This method is clearly finding some higher-level similarities between larger groups of stations.



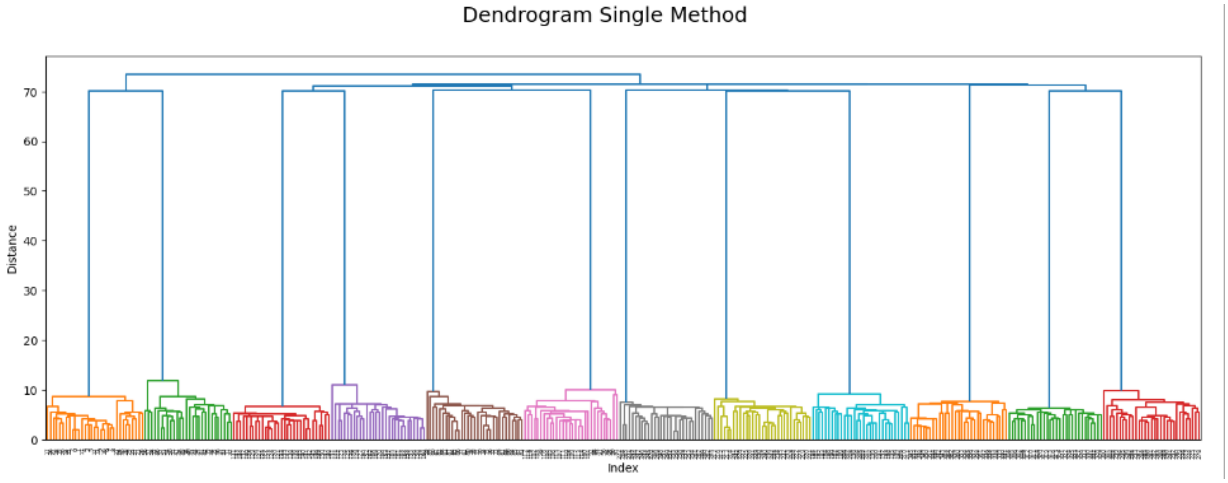
Ward method: This method aims to reduce the variance within each cluster. We see two well-defined clusters here. However, within each cluster we see higher-level separations occurring, indicating significant differences. Perhaps this indicates the clear division of pleasant weather labels. Out of all the methods this one shows us clear subgroupings within clusters, indicating the presence of climate similarities between different locations.

PCA- How many components to use?

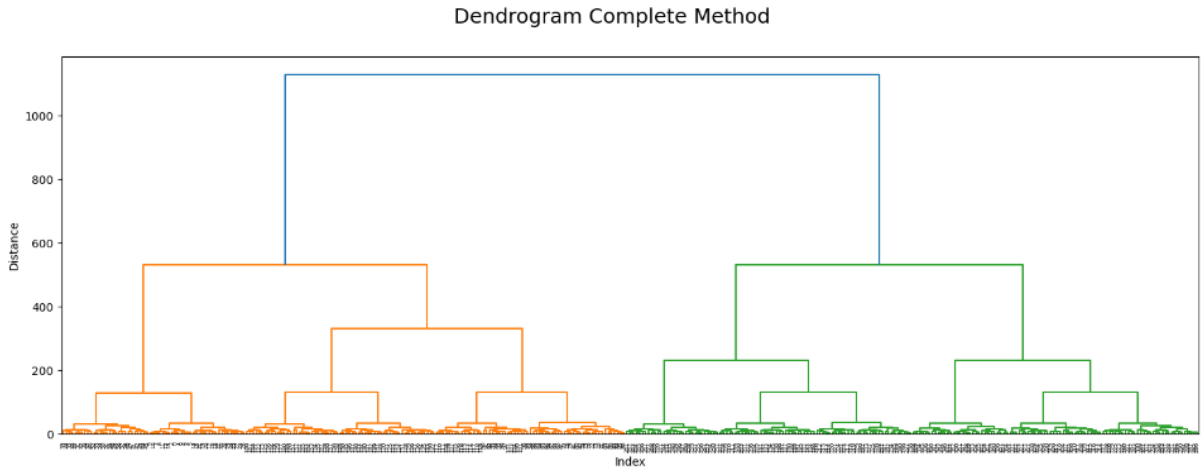


I decided to run an initial PCA with 18 components (for the 18 weather stations) and plotted the cumulative explained variance to determine how many components would capture a high percentage of the total variance. After looking at this graph, I settled on using **10 components** for my PCA.

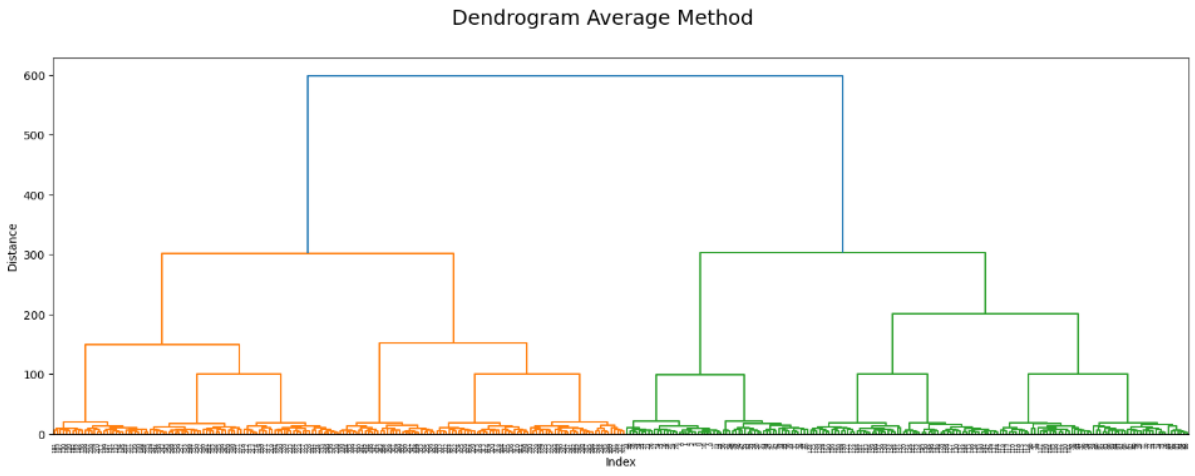
Dendrograms Post-PCA



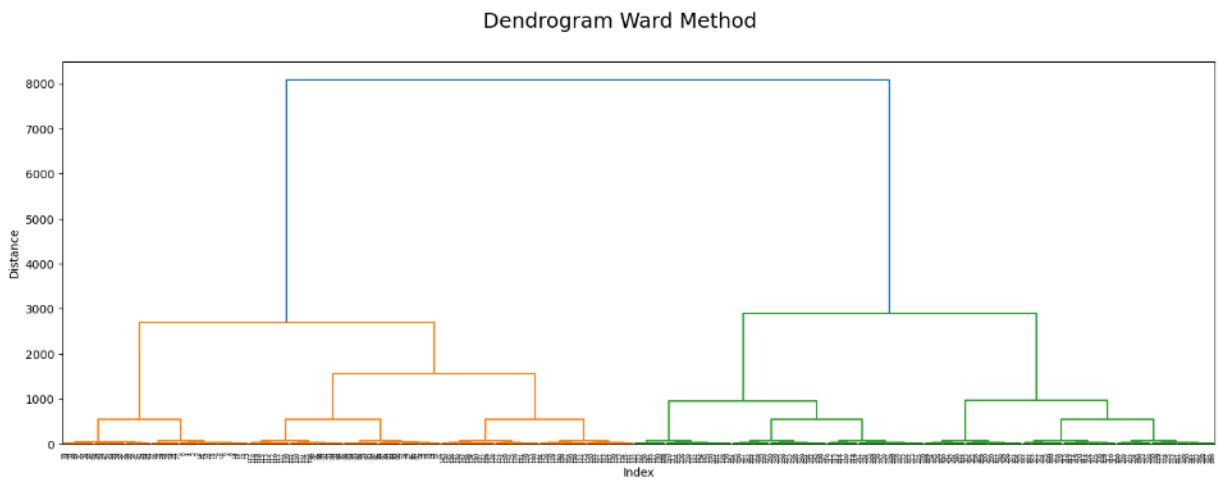
Single: Compared to the initial single method pre-reduction, this shows a number of clear groupings at a high level. Reducing dimensions may have also helped to reduce the noise and highlight essential clustering patterns. Unlike the first round, this shows no evidence of long chains forming. The higher-level clusters that form may reflect similarities between weather stations.



Complete: There's a clear split at a high level that divides the data into two clear clusters, telling us this method has identified two broad groups, which is a clear fit for the pleasant weather labels: 0 or 1 (unpleasant versus pleasant). Going farther down we see the formation of several, smaller sub-clusters, a further division of locations with more specific similarities (similar climatic patterns, latitudes, etc).



Average: This looks very similar to the complete method above- there's a sharp division into two main groups of clusters, albeit at a slightly lower distance. In this situation, the larger the distance at the point of separation, the more dissimilar the two groups are. So in this case this method tells us that the two distinct groups are not quite as distinct as the complete method says above.



Ward: While this one looks similar to the previous two, there is a fundamental difference: the split into two major clusters occurs much higher (an order of magnitude higher than average, and several times higher than complete), suggesting a much stronger and stark difference between the two clusters. Again this mirrors what we would expect to see in the pleasant versus unpleasant weather days. Within each cluster we see a fairly even balance of sub-clusters, suggesting the Ward method has managed to identify well-defined groups with high internal consistency.

Reduced versus non-reduced: Overall the PCA-reduced dataset consistently reveals two main clusters, indicating we're on the right track for what we are looking for. The subdivisions that form beyond that can tell us more info, like coastal versus inland stations, however it could be along a different distinction as well (a division along latitudes).