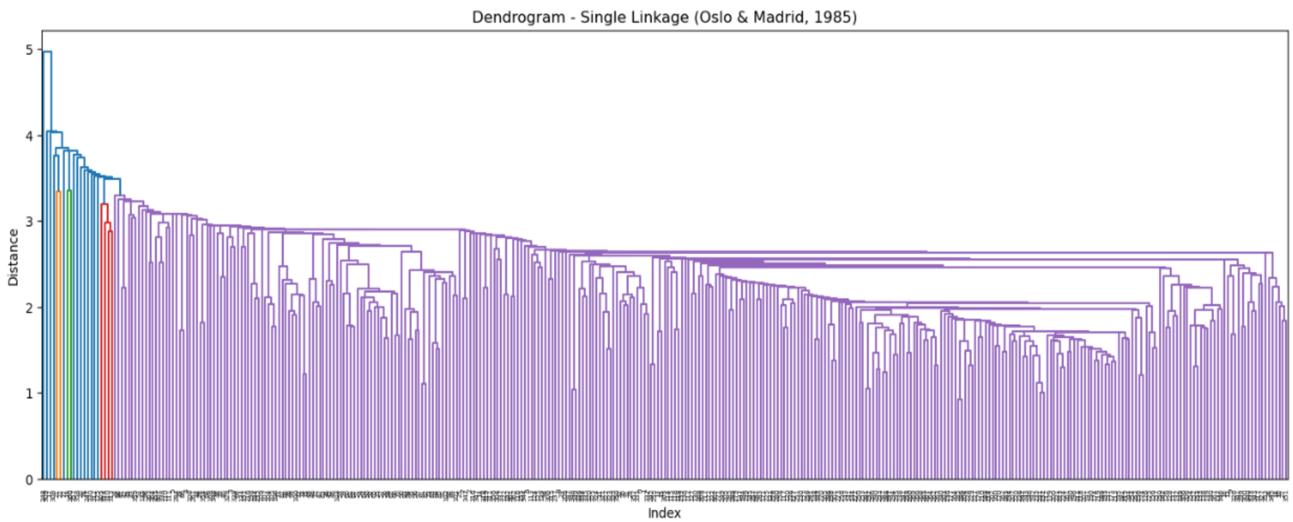
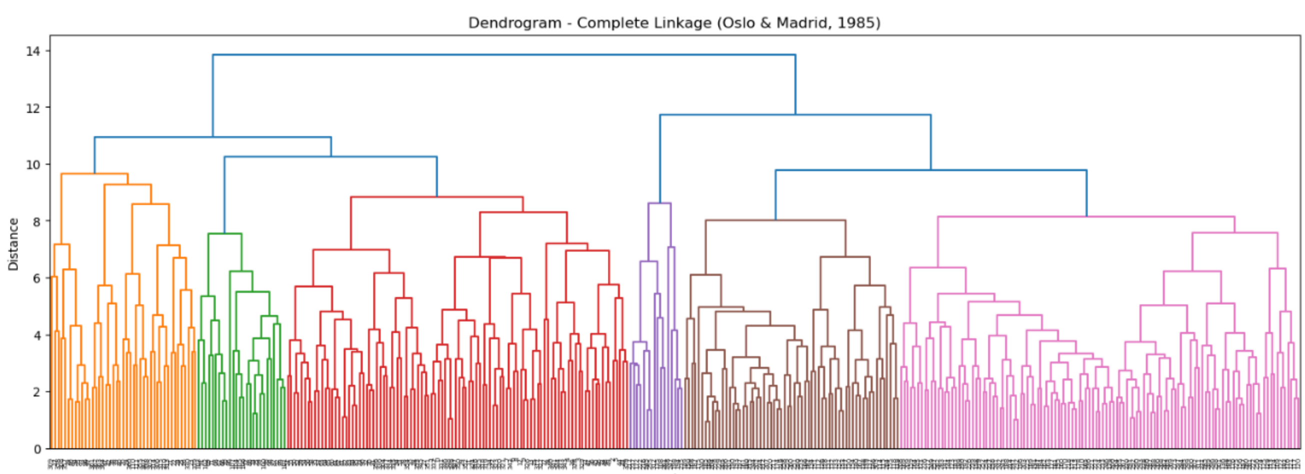
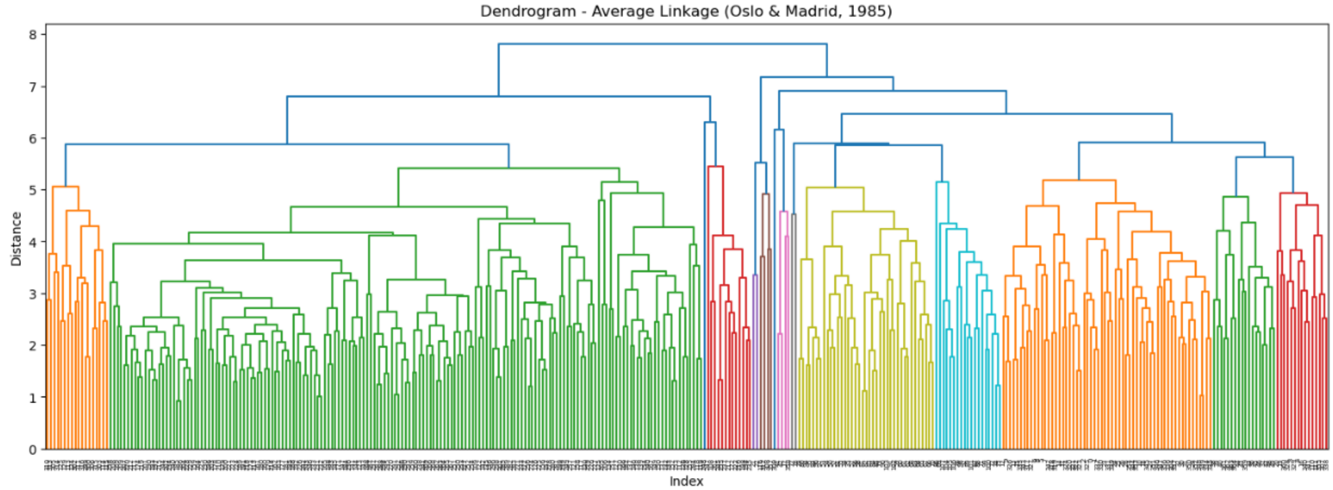
Isaac Contreras

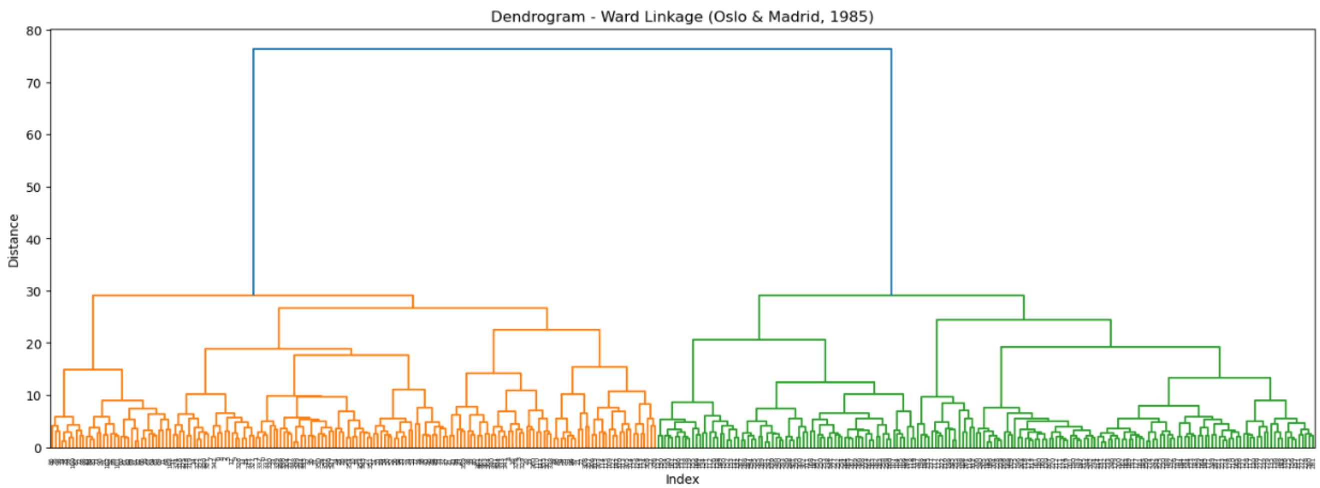
Real-World Applications of Machine Learning

Dendrograms comparing Oslo and Madrid in 1985:





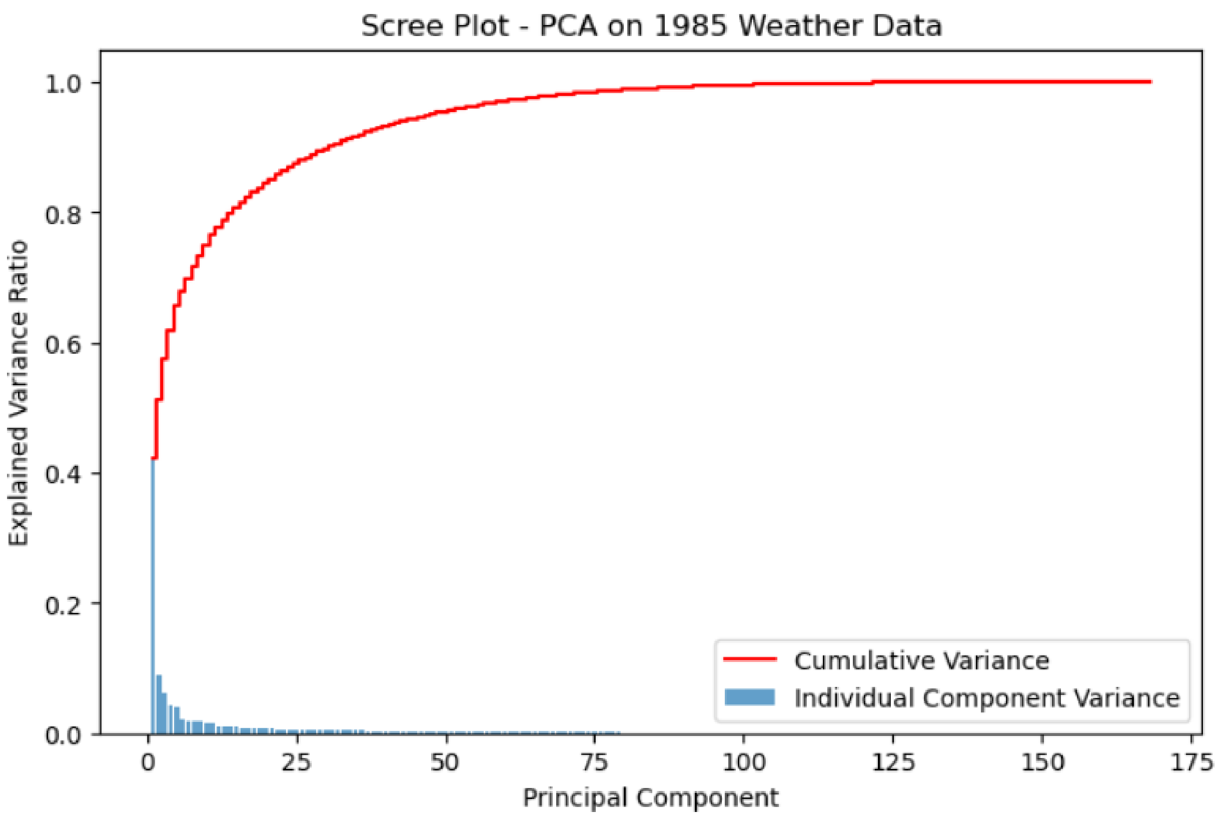




The **single** and **average** linkage dendrograms show that nearly all the data points fell into one big cluster, which isn't surprising. Single linkage often causes a "chain-like" effect where points are added one by one, and average linkage merges groups easily when distances aren't too large. As a result, Oslo and Madrid's 1985 data looked uniform at the chosen cut level.

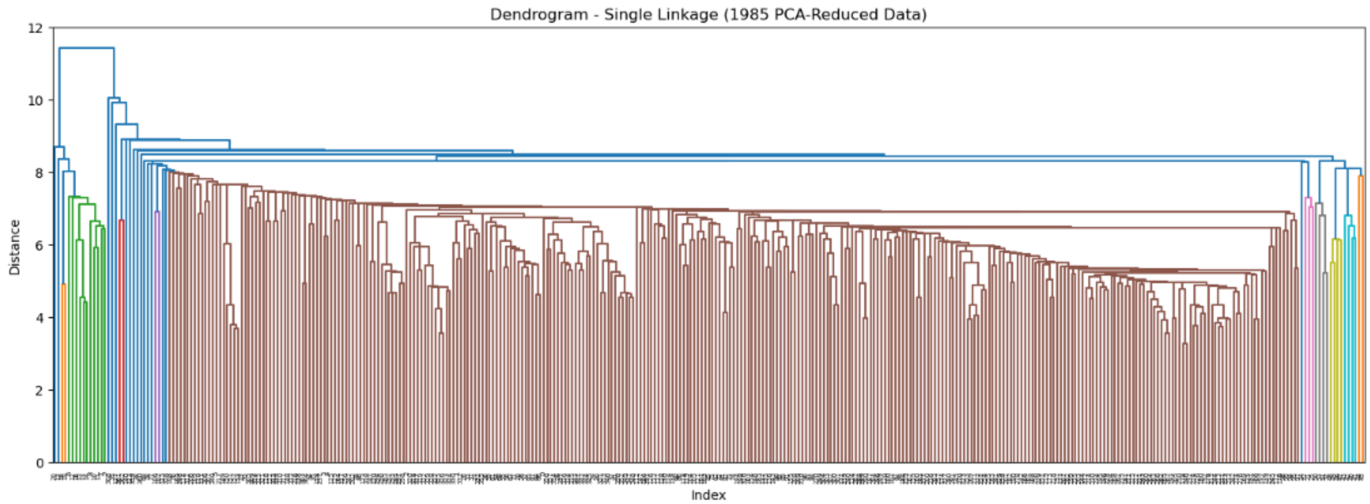
By contrast, **complete** linkage identified five distinct groups, showing a more moderate separation among daily weather observations. **Ward's** method went even further, splitting the data into 22 clusters. This highlights the impact that different linkage methods and cut thresholds can have on hierarchical clustering results, emphasizing the need to try multiple approaches to find the most revealing breakdown of the dataset.

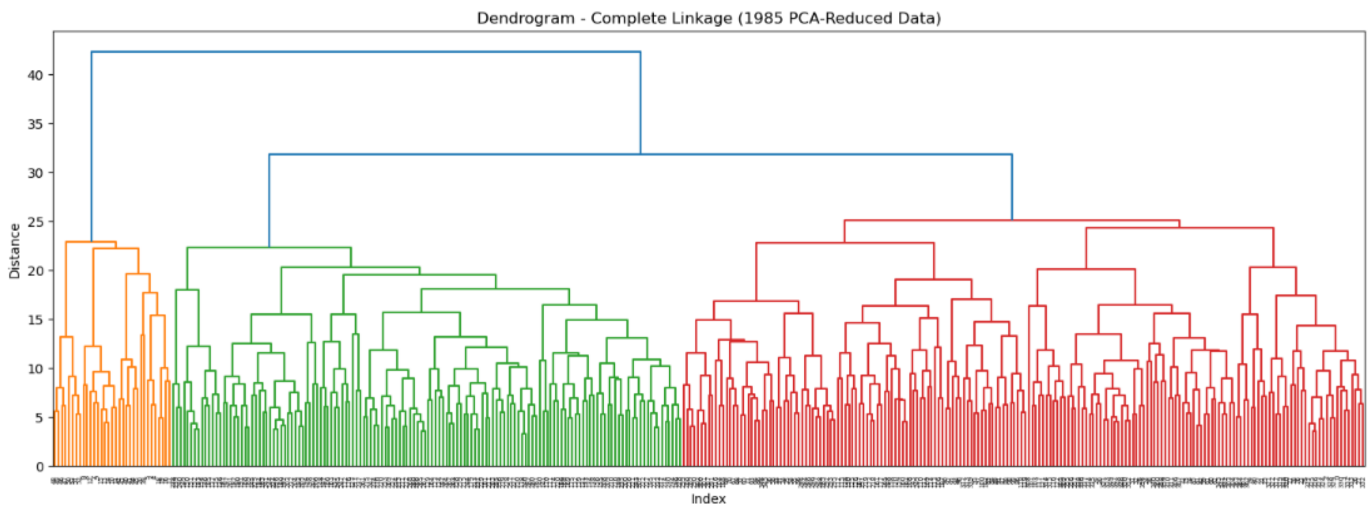
Although the data already gave us some insight, the high dimensionality can obscure clear patterns in hierarchical clustering. By **applying PCA to reduce the dataset's dimensions**, we can remove redundant information and highlight the most essential features, often yielding sharper, more interpretable dendrograms. This reduction helps focus on the core structure of the data and can improve alignment with meaningful patterns—such as "pleasant" versus "unpleasant" days.



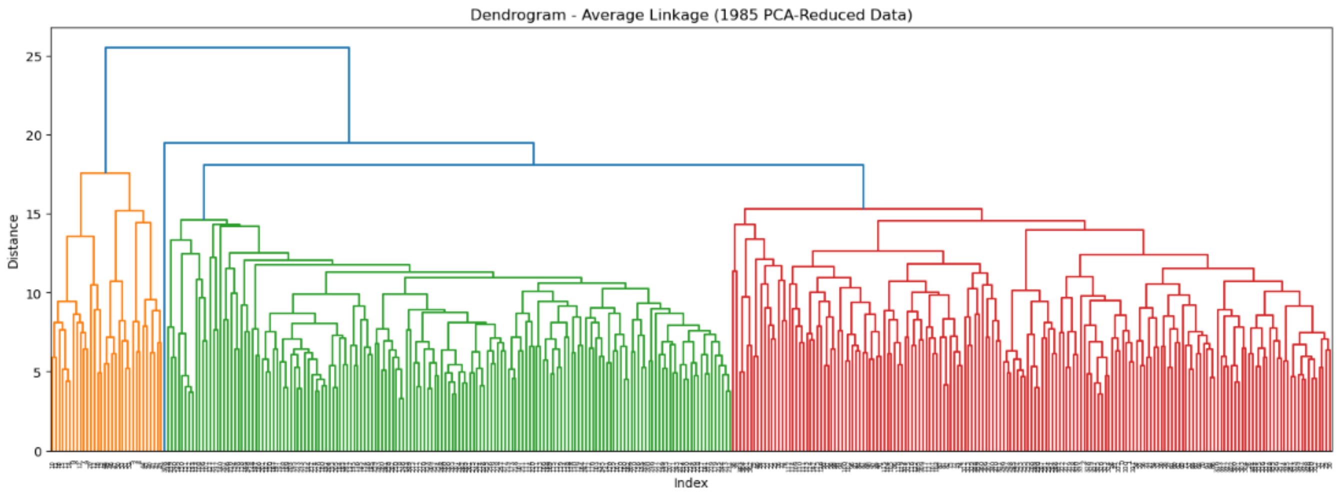
After generating the scree plot, reducing the dataset to around 20 principal components seemed appropriate, based on the point at which additional components yielded only minimal variance. The code snippet demonstrates how to apply PCA (pca = PCA(n_components=20)) and transform the data into a new DataFrame with columns PC1 through PC20. This process strips away noise and repetitive information, leaving a leaner feature set that typically produces more clear-cut hierarchies in dendrograms.

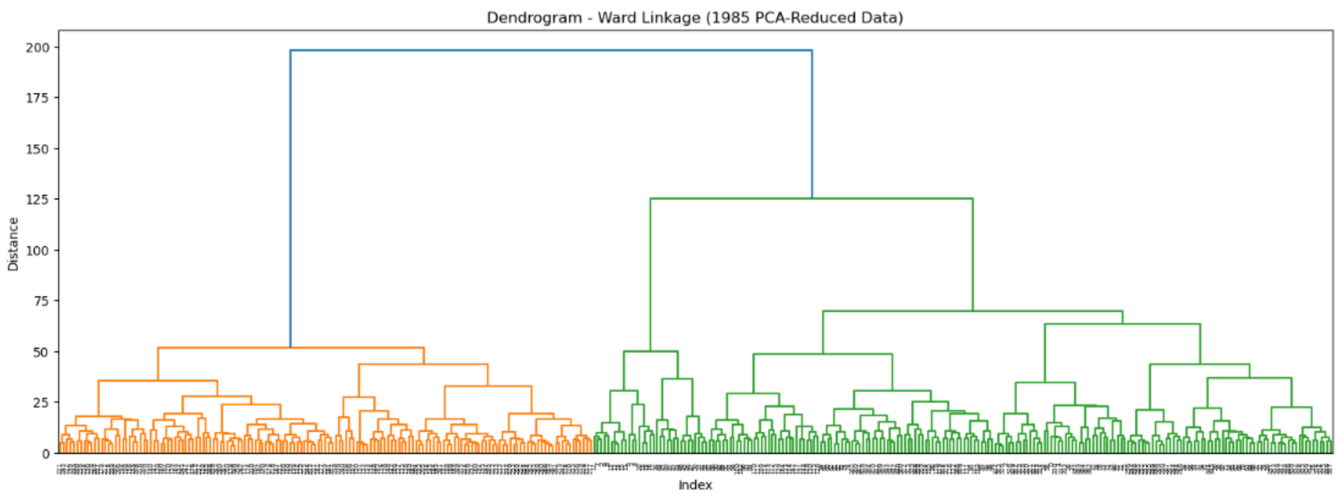
Dendrograms comparing Oslo and Madrid in 1985 with data's dimensionality reduced (PCA):





Comparing dendrograms before and after PCA shows that reducing the dataset's dimensionality sharpened the clusters. Looking at the single linkage dendrograms, we still see a chain-like structure both before and after PCA, so the reduction didn't resolve its "chaining" problem. However, for complete linkage, PCA makes the clusters stand out more clearly, indicating that removing redundant features helps reveal more distinct groupings in the data.





Both **average** and **ward** linkage dendrograms appear more structured and distinct after PCA, whereas the raw data had more noise and less clear cluster boundaries. By reducing the dimensions, average linkage splits the data into more balanced subgroups, and ward linkage reveals larger distance gaps between clusters. This suggests that PCA helps highlight the natural separation in the weather observations, making the final clusters easier to interpret.