Prove 10 - Clustering

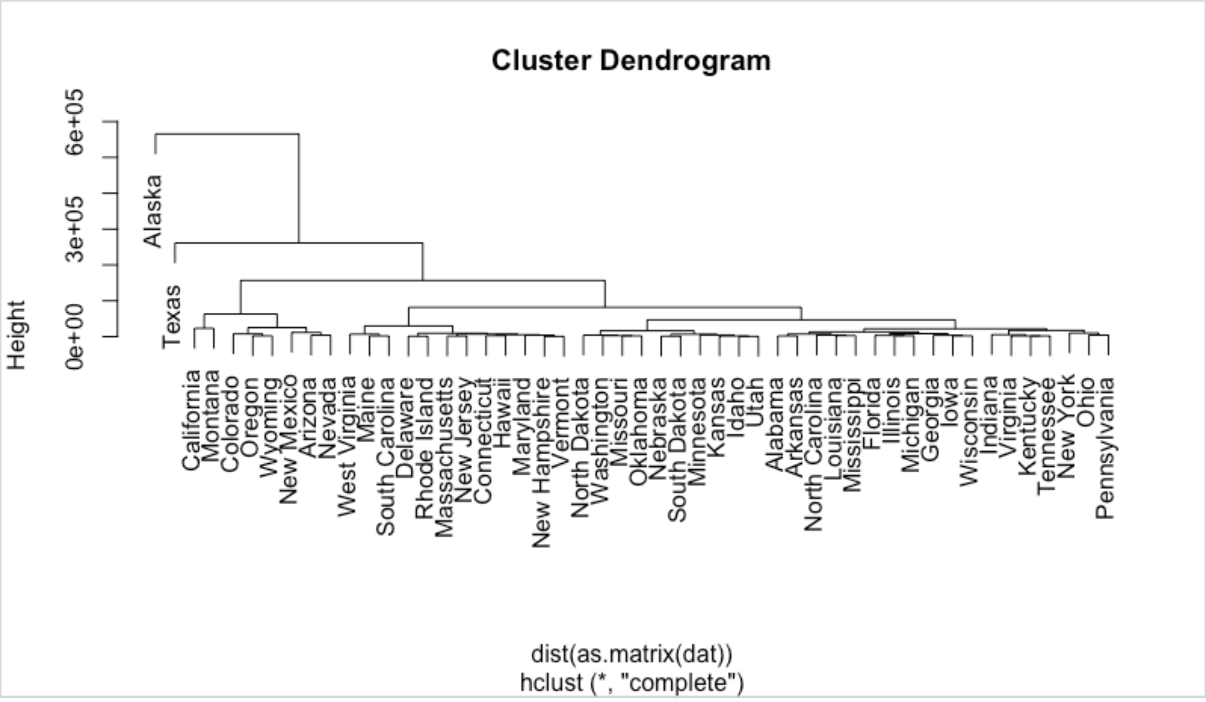
Etienne Beaulac

March 17, 2018

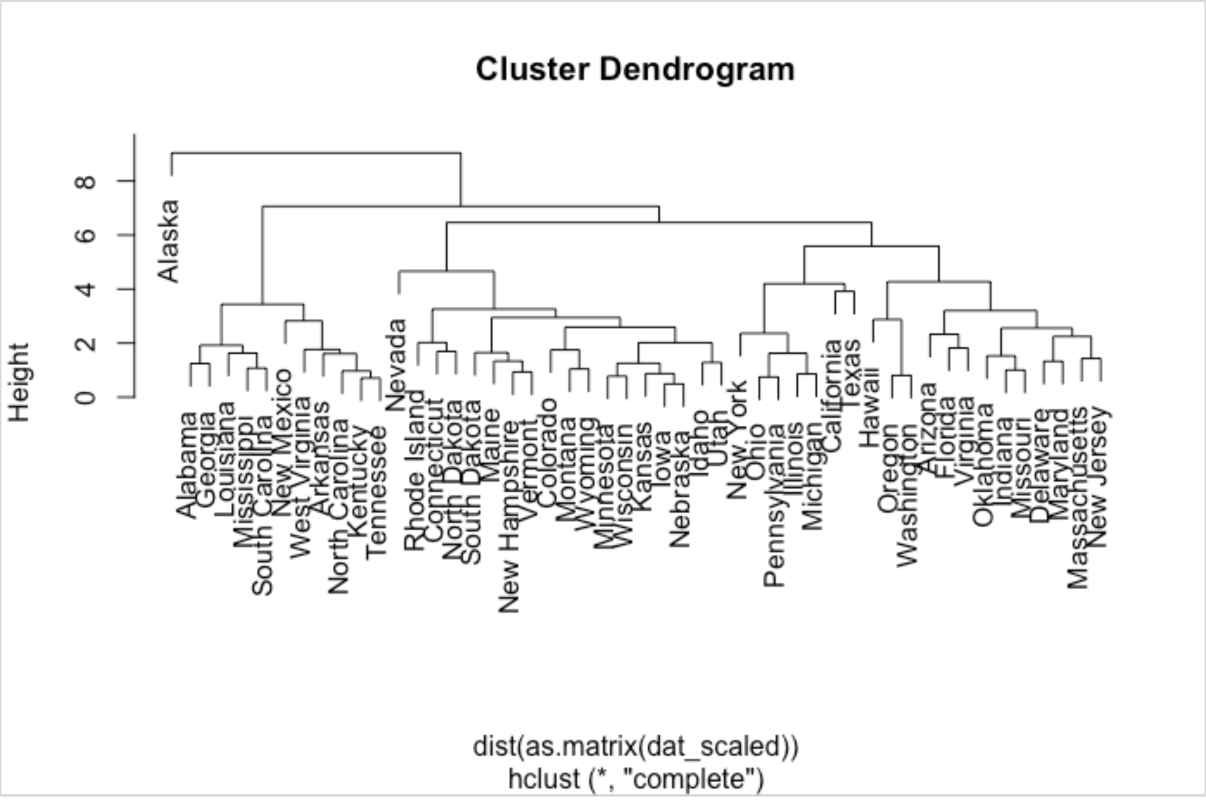
AGGLOMERATIVE HIERARCHICAL CLUSTERING

Differences between normalized and non-normalized

Not Normalized:

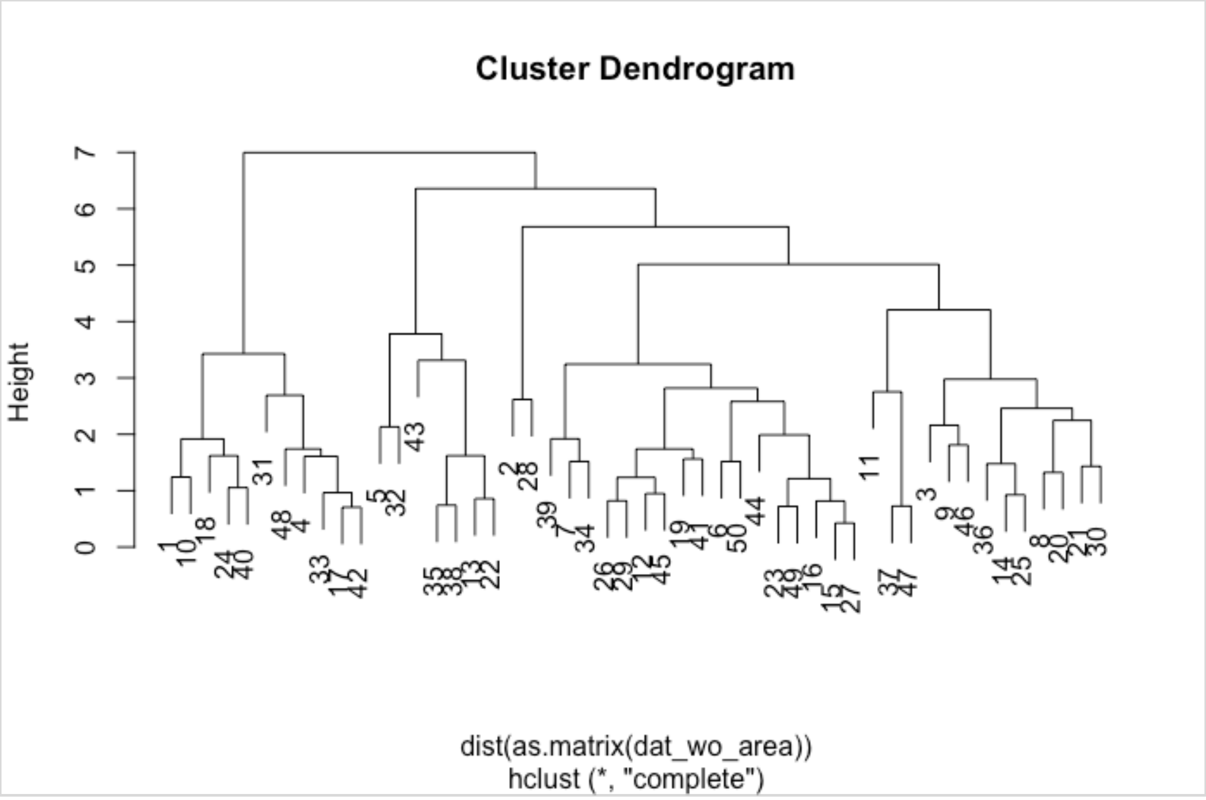


Normalized:



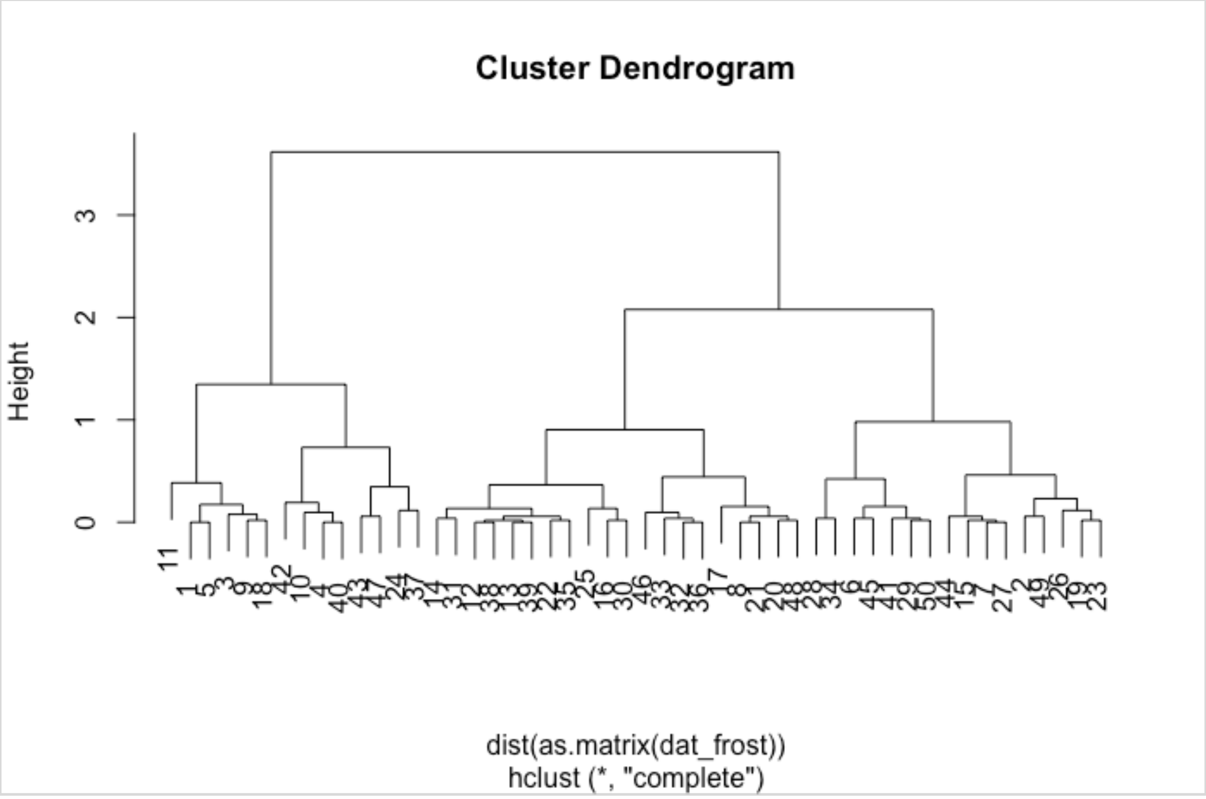
I think the main difference I see is that there were much clearer groupings in the normalized version of the dendrogram. Texas actually belongs somewhere, close to California, instead of being in the outskirts with Alaska. It just looks like a much more accurate representation of the data.

Without "Area" attribute



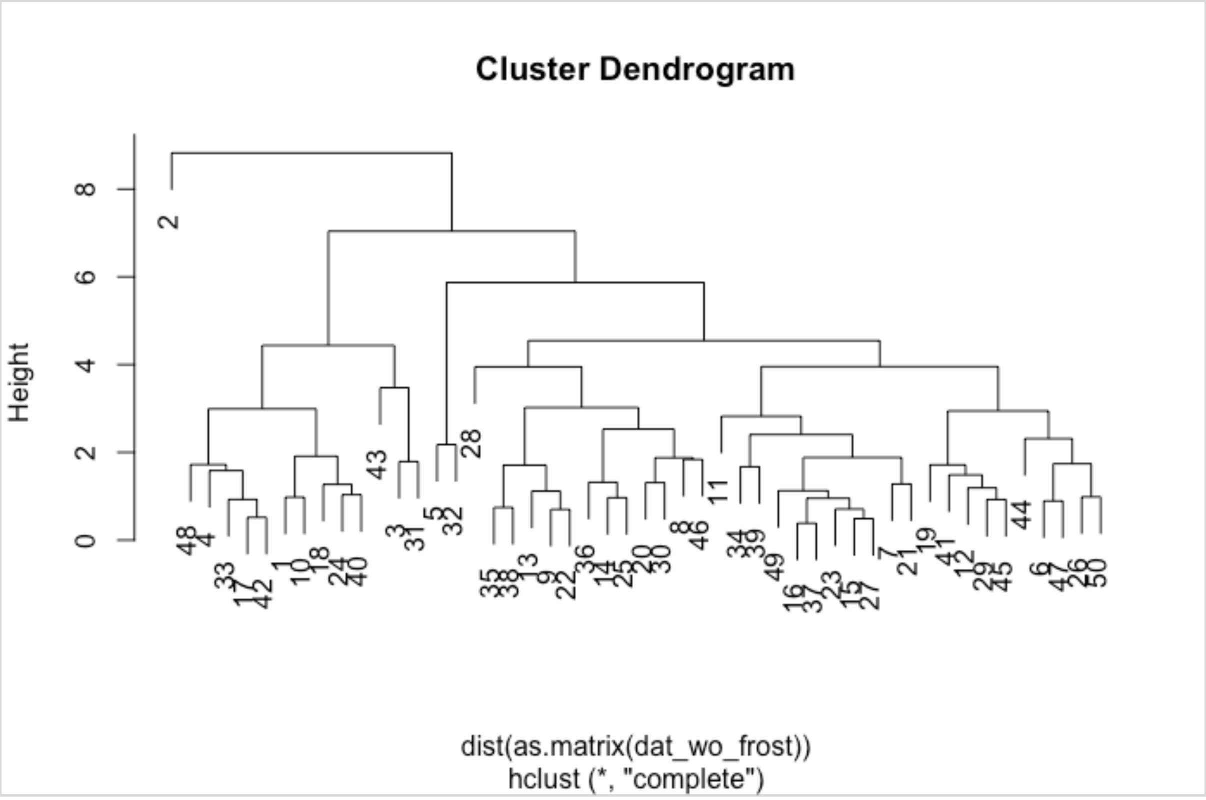
So, I'm not sure where my labels went with this plot. I guess it could be the effect of removing the Area attribute. However, it looks like the dendrogram is very similar to the scaled dendrogram. I'm guessing that it would be similar to the not-scaled dendrogram had I used the not-scaled data for it.

Only "Frost" attribute



This dendrogram is a little different from the scaled one. Of course, only having one attribute has a pretty big effect on the outcome. However, you can see that there is a similarity. I believe that means that the "Frost" attribute has a strong influence on the outcome of the clustering. I will created a dendrogram with everything but Frost.

Everything but "Frost"

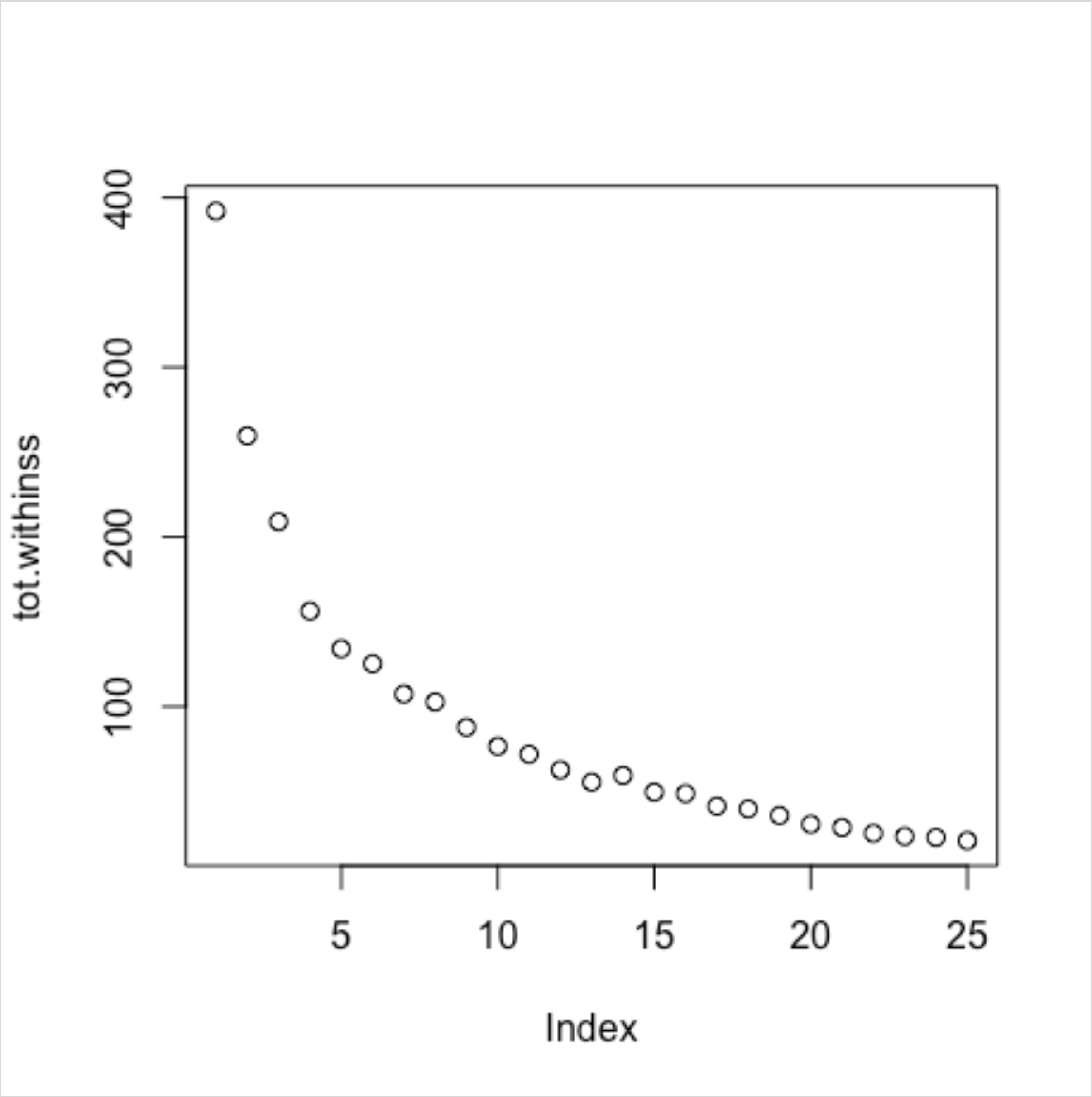


So, if we remove the Frost attribute we lose a little bit of detail that helps the groups from accurately. However, we can see that Alaska is in the same position it was before removing Frost.

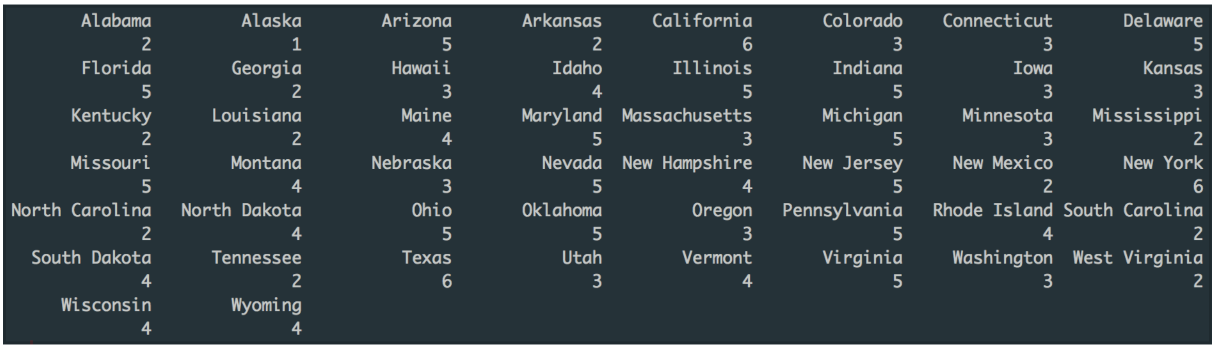
USING K-MEANS

1. Note the size of each cluster and the mean values. Do you have any insight into why they were divided this way?
   I think they were divided this way because there were some outliers such as Alaska and Texas. Had they not been there, the first and second cluster would have fought a little more for the data that could belong to both groups.

Total sum of squares plot

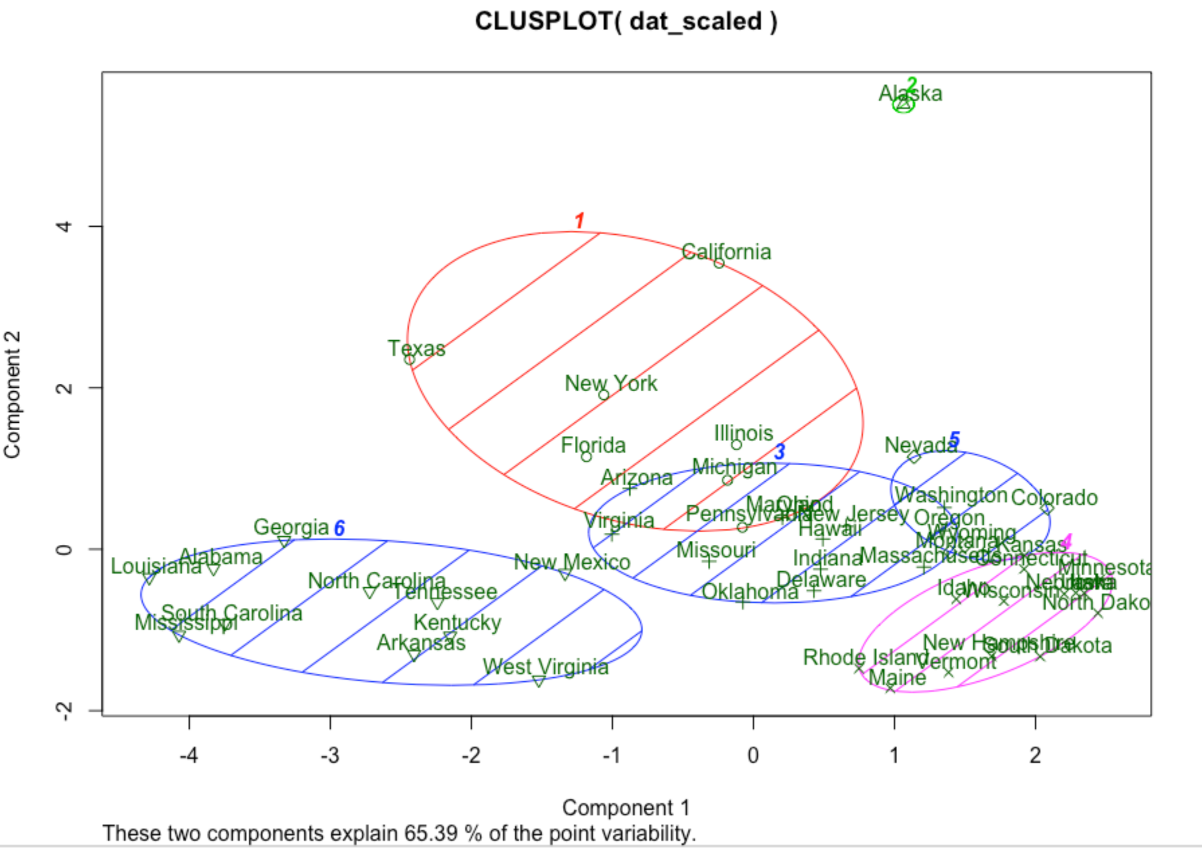


Groups according to states for k of 6.



```
       Alabama         Alaska        Arizona       Arkansas     California       Colorado    Connecticut       Delaware 
             2              1              5              2              6              3              3              5 
       Florida        Georgia         Hawaii          Idaho       Illinois        Indiana           Iowa         Kansas 
             5              2              3              4              5              5              3              3 
      Kentucky      Louisiana          Maine       Maryland  Massachusetts       Michigan      Minnesota    Mississippi 
             2              2              4              5              3              5              3              2 
      Missouri        Montana       Nebraska         Nevada  New Hampshire     New Jersey     New Mexico       New York 
             5              4              3              5              4              5              2              6 
North Carolina   North Dakota           Ohio       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
             2              4              5              5              3              5              4              2 
  South Dakota      Tennessee          Texas           Utah        Vermont       Virginia     Washington  West Virginia 
             4              2              6              3              4              5              3              2 
     Wisconsin        Wyoming 
             4              4 
```

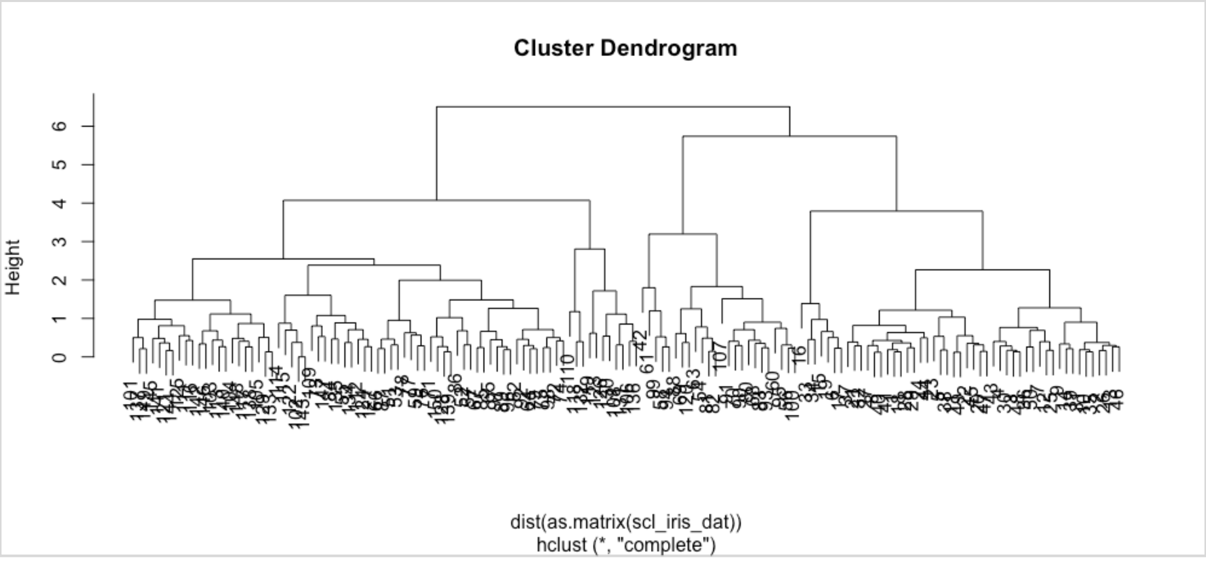
2D Plot of clusters



I can see that the states are clustered by type as we see them normally. For example, you've got "Power-States" such as California, Texas, New York, and Florida all grouped into one. You've got cluster 6 that are all the states that are around the same types of Southern States. Of course, being Canadian, I'm not super familiar with the United States and I would need better background information to have greater understanding of the groupings.

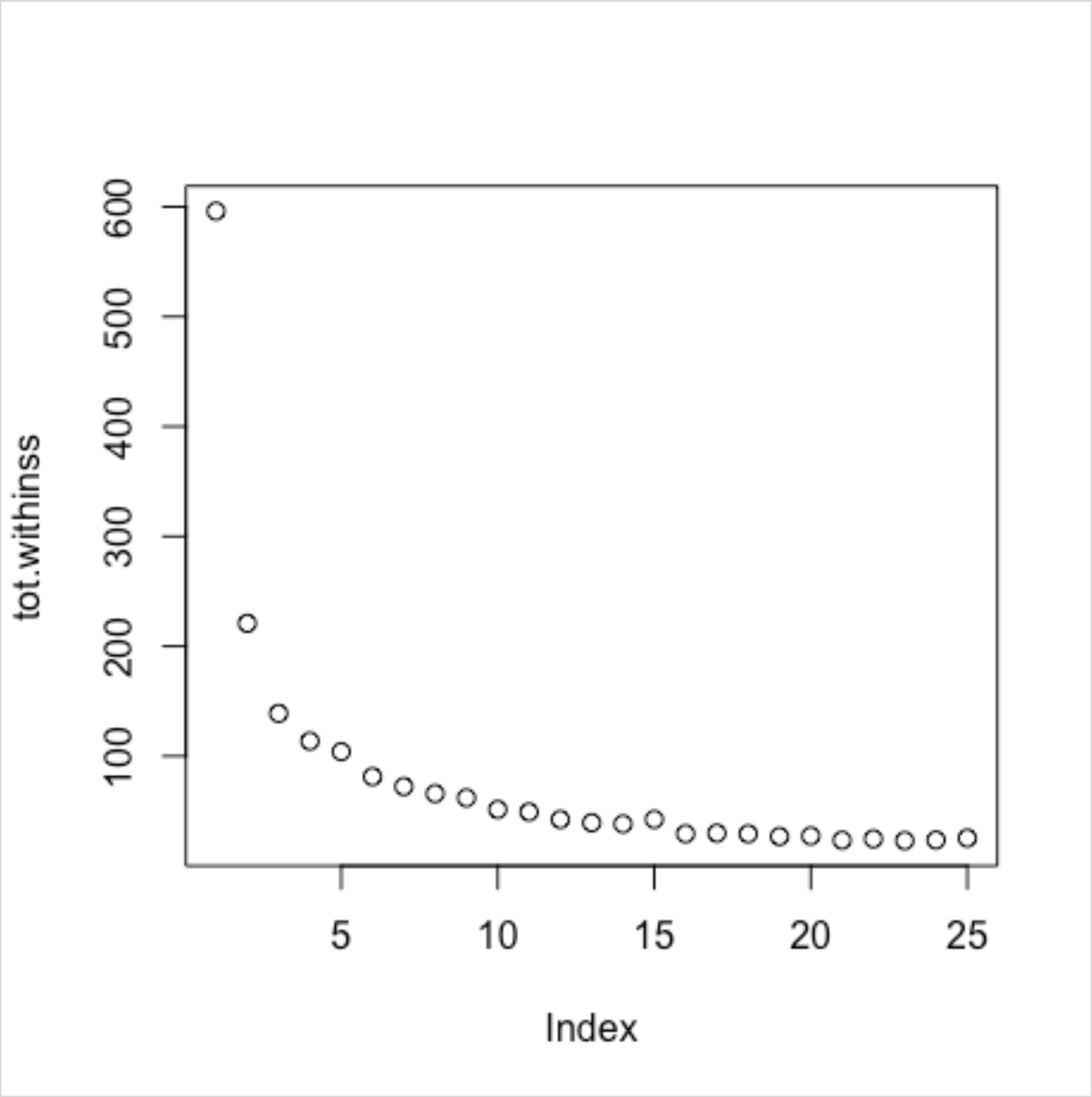
Extra - Iris

Here, I am repeating the process for the Iris dataset.



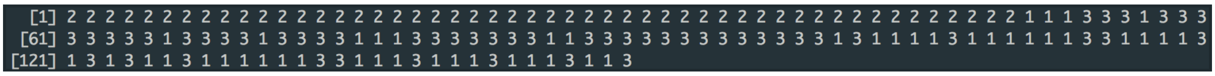
Of course, since there are no row names, it’s a little difficult to properly understand what is going on. However, at height 3-4, it looks like there are 3 distinct groups (which corresponds to the iris dataset).

Sum of squares plot

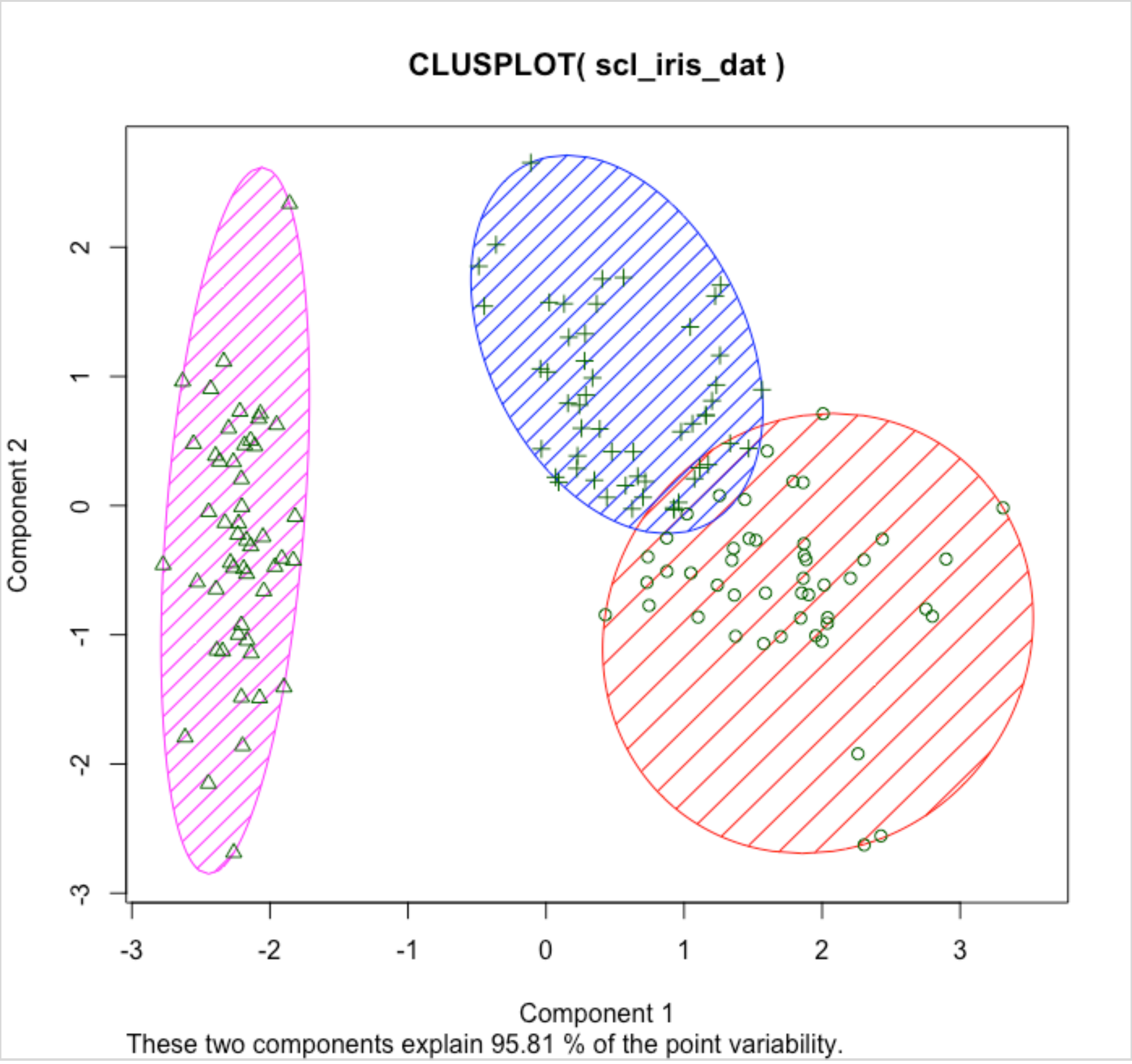


Of course, since I know that there are only 3 types of Irises in the dataset, I'll be choosing k=3.

Groups according to k of 3



```
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 3 3 3 1 3 3 3
 [61] 3 3 3 3 3 1 3 3 3 3 1 3 3 3 3 1 1 1 3 3 3 3 3 3 3 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 1 1 1 3 1 1 1 1 1 1 3 3 1 1 1 1 3
[121] 1 3 1 3 1 1 3 1 1 1 1 1 1 3 3 1 1 1 3 1 1 1 3 1 1 1 3 1 1 3
```



So, as you can see from this graph, there is one specific type of Iris that is very easy to distinguish from the rest, whereas the other two sort-of melt together at some point. This is where the errors come from when doing supervised learning on the Iris dataset.

To finish off…

A) Some attempt was made

B) Developing, but significantly deficient

C) Slightly deficient, but still mostly adequate

D) Meets requirements

E) Shows creativity and excels above and beyond requirements

I think I went beyond the requirements by:

- Doing a dendrogram of everything but the Frost attribute
- Going through the whole process with the Iris dataset