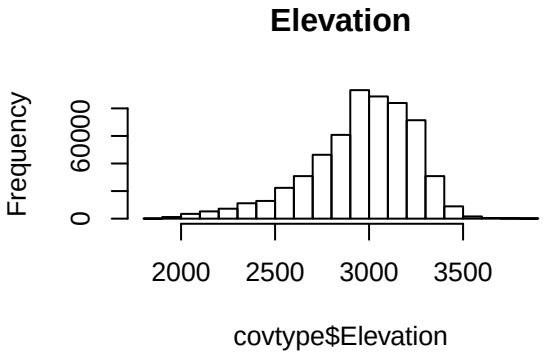
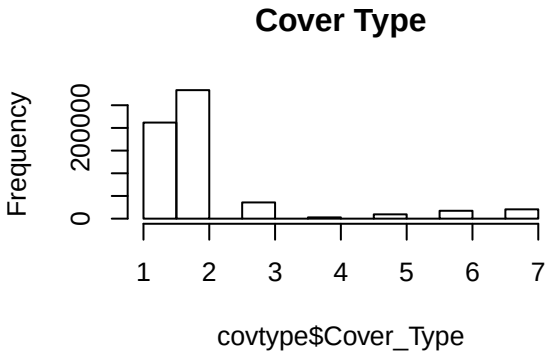
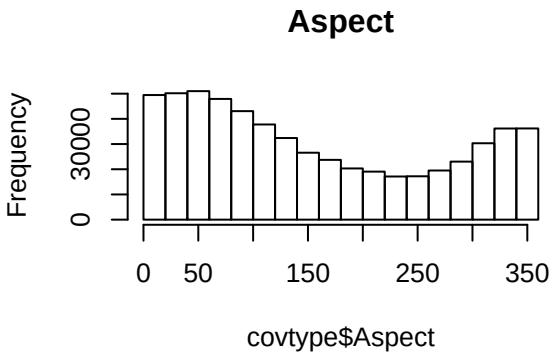
Forest Cover

Due to the relatively large number of columns, I will omit showing the header and summary and provide a quick textual summary. The data set consists of 581'012 observations whereas one observation consists of 55 variables. Ten of these variables are continuous and the rest of them categorical. There are three categorical attributes: Wilderness_Area, Soil_Type and Cover_Type. Soil_Type is responsible for most columns, as each type is represented as a column of a one-hot vector fashion, summing up to 40 columns. Wilderness_Area makes up for four columns and Cover_Type for one categorical column with values from 1-7. Followingly we will have a closer look at the continuous variables to potentially discover outliers or any other peculiarities. The first quantitative variable is Elevation with values from 1859-3858 and it is given in meters. Following is Aspect which is given in azimuth with a range from 0 to 360, describing where the surface is facing. The Slope describes the steepness and the values in our case range from 0 to 66.0. The following 4 variables describe the distance of the observation from either Hydrology, meaning any kind of water that is on the surface, Roadways or Firepoints. All of the values are given in meters and there are no obvious outliers. There are however negative values in the Vertical_Distance_To_Hydrology, which can be explained by the fact that it makes a difference if the water is below or above the observation. The Hillshade variables give an hillshade index from 0 to 255. Along with the data set, there is some background information about the Wilderness_Area and the Cover_Type that might be helpful for further analysis. There are 4 areas, namely Rawah (a1), Neota (a2), Comanche Peak (a3) and Cache La Poudre (a4). It is assumed that Neota would have the highest elevation, whereas Rawah and Comanche Peak would be about average and Cache La Poudre would be the lowest in elevation. Furthermore it is said that (a1) and (a3) would mostly have lodgepole pine (t2), followed by spruce/fir (t1) and aspen (t5). (a2) is predicted to have mostly (t1) and (a4) would then consist of the remaining Ponderosa pine (t3), cottonwood/willow (t4), Douglas-fir (t6). The goal of this project is to detect some kind of clustering structure that would resemble the clustering for Cover_Type.

From the boxplots, that are not displayed here, we are able to see that there are outliers, as they pass the Q1 – 1.5×IQR or Q3 + 1.5×IQR threshold. However, the values do not appear to be measuring errors, as they all seem physically possible, if considered one by one. Additionally, it is highly likely that there are multiple underlying distributions, considering the different areas and elevation levels. The boxplot cannot cover for this fact and thus, a lot of values are considered to be outliers. That is why I decide not to remove any of the outlier values. Furthermore, I will plot distributions of variables which will underline my decision to keep the outlier values. The distributions will also bring insights for the Cover_Type, which will be helpful for the clustering. Afterwards we will also look at some correlations between the continuous variables.

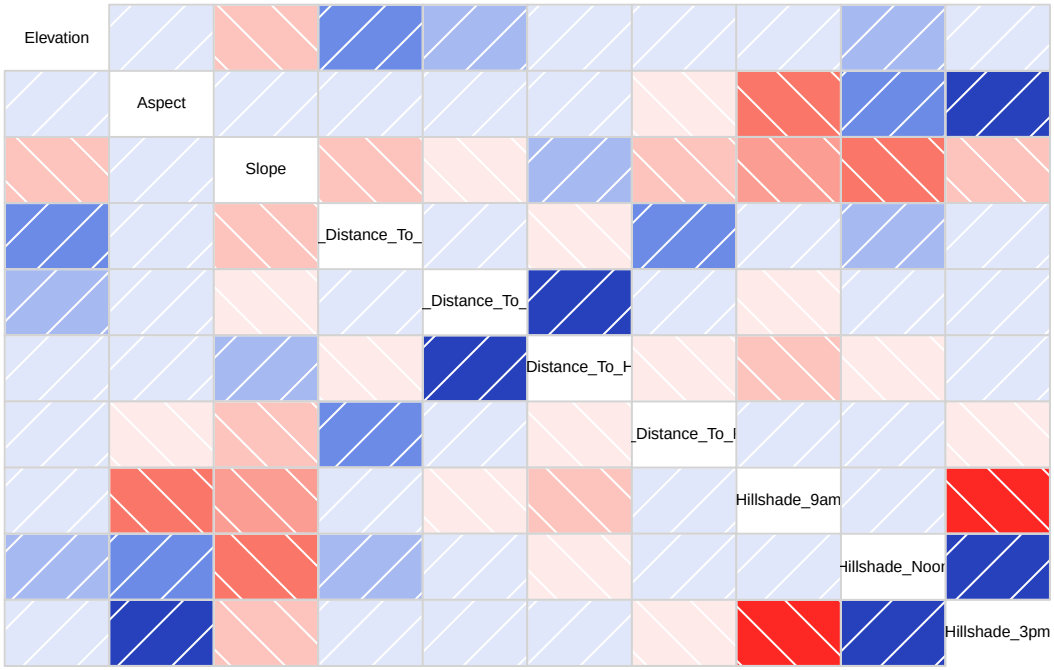






I show Elevation for general information about the data set, such that we know what kinds of altitudes this data set is about. Aspect is an interesting distribtion because it ranges from 0 to 360 degrees, so we can imagine the distribution in a circular fashion, where after 360 it will start at 0 again. What looks like two distributions is actually one distribution about the 0/360 value. From the Cover_Type we can also see that (t1) and (t2) are predominant. Thus from the clustering we should expect 2 bigger clusters and 5 smaller clusters. There will be a multitude of clustering algorithms tested. From the centroid-based family I will use kmeans and kproto. kproto is a variant of kmeans that can be used for data that consists both of continuous and categorical values. From our background knowledge we know that Wilderness_Area might have a big influence on the Cover_Type. That is why I will include Wilderness_Area in a special data set on which I will apply kproto. From the connectivity-based family I will be looking at the top-down approach, which in R is called by the diana() function, which stands for Divisive Analysis. Also, I will be looking at the bottom-up approach, which is implemented either in the agnes() or hclust() function. I decided to investigate these two families as they were part of our lecture. Before continuing, I will extract two new datasets from the old one. For the kmeans and hierarchical clustering analysis, I will exclude the categorical values, as they do not work well with the concept of distance if the matrix is mixed with continous and categorical values. In this dataset, we will also omit the missing values, as missing values do not seem to add value in our case and we have sufficient data after omitting. Most importantly we also scale the values into the same relative range. We do this because kmeans is sensitive to using different scales in the continous variables. For example, a distance of 60 meters would be more influential than a difference of 50 degrees of slope, which intuitively does not make sense. As previously mentioned, there will also be a second data with the Wilderness_Area columns to see if including this variable will bring better clustering. In both data sets we are excluding the Soil_Type data, as the background knowledge does not seem to imply any important correlation. It also allows us to keep the complexity of the solution relatively low. Due to a lack of computing power, we will also not work with the whole data set, as it would take too much time, at least for the 'agnes' clustering.

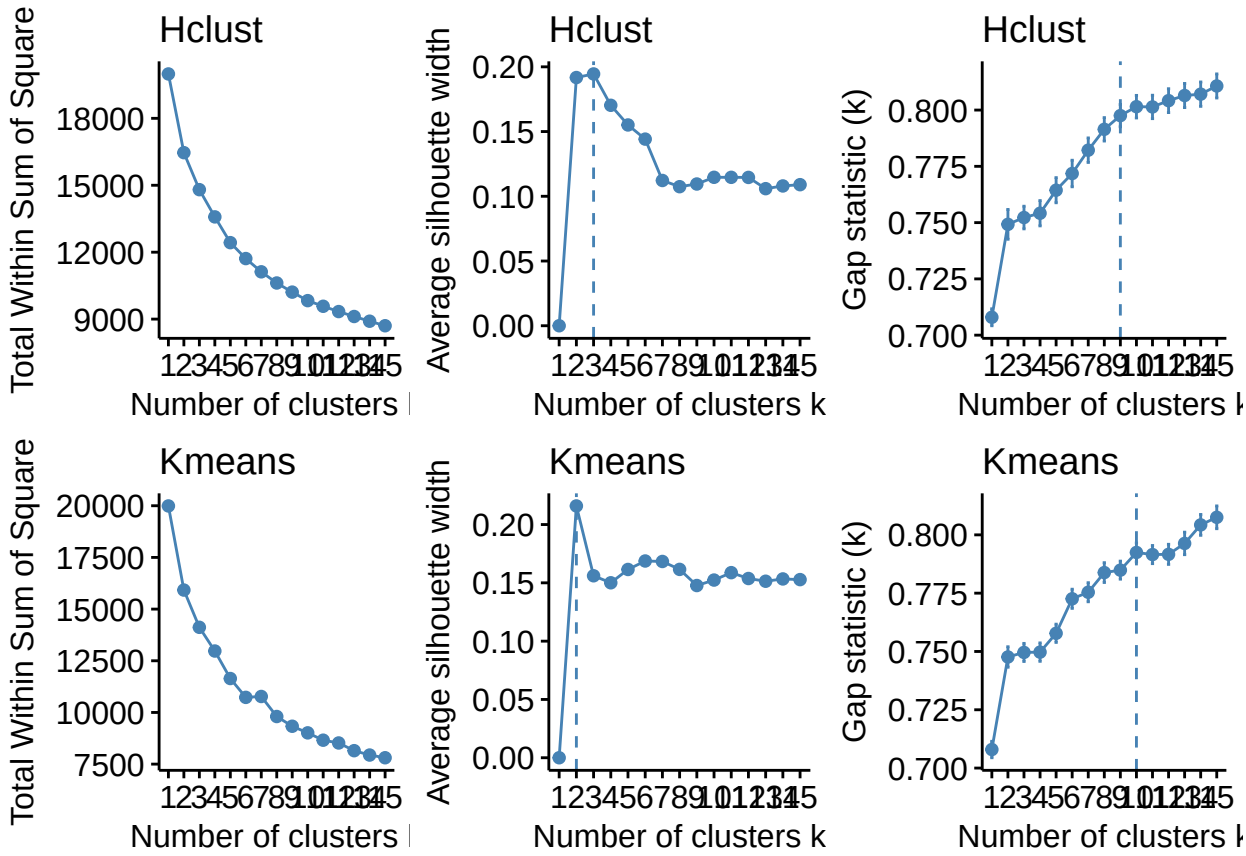
There is not much additional information to be gained from the correlation matrix of the continuous variables. The strong correlations are where they are expected.



For the further analysis, we need the distance measures between the individual points. They are required by the hierarchical clustering algorithms. I choose the euclidean distance because that allows a better comparison with the kmeans algorithm, which by definition uses the euclidean distance. There will be further analysis where we change the distance measure to the Manhattan distance. Despite that we know that there should be 7 clusters, as there are 7 cover types, we will try to find the approximately best number of clusters from the Akaike and Bayesian Information Criterion.

With a sample size of n = 10'000 and trying k = 3 to k = 100, the most optimal k according to the AIC appears to be the maximum value defined. For the BIC it appeared to be 51. This difference might originate from the fact that BIC appears to be more punishing for using additional clusters. The values applied above will be for a smaller sample size.

The next graphs are displaying 3 different ways of finding the right amount of clusters, namely the Elbow Method, the Average Silhouette Method and the Gap Statistic Method. We will be applying them to the kmeans, as a representative of the centroid-based family and hclust as a representative of the connectivity-based family. The Elbow Method tries to minimize the intra-cluster variance, also known as the within-cluster sum of square. Most optimally there would be some elbow-like behaviour in the graph. Usually the place were the 'elbow' bends is the most optimal number of clusters. The Average Silhouette Method looks at the average silhouette width of the clusters. Intuitively, if there is a high average silhouette width, there is a good clustering, where narrower silhouettes rather suggest clusters that do not generalize well. Third is the Gap Statistic Method, which again compares the different intra-cluster variations. It does so in a more intricate way however. In theory, we should be looking to maximize the Gap statistic and as we can see, we should be taking the k=15 or even more, as the trend seems to be rising. However, we choose the k after which the trend seems to slow down drastically, which in our case is at the k=6 point. In the two other methods we can also see that at k=6 there is usually a significant jump followed by a flattening change.



There are multiple methods to compute the bottom-up hierarchical clustering. That is why first we need to investigate which method performs best, which turns out to be the "ward" method.

```
##   average    single  complete      ward
## 0.8644065 0.7309931 0.9181611 0.9881532
```

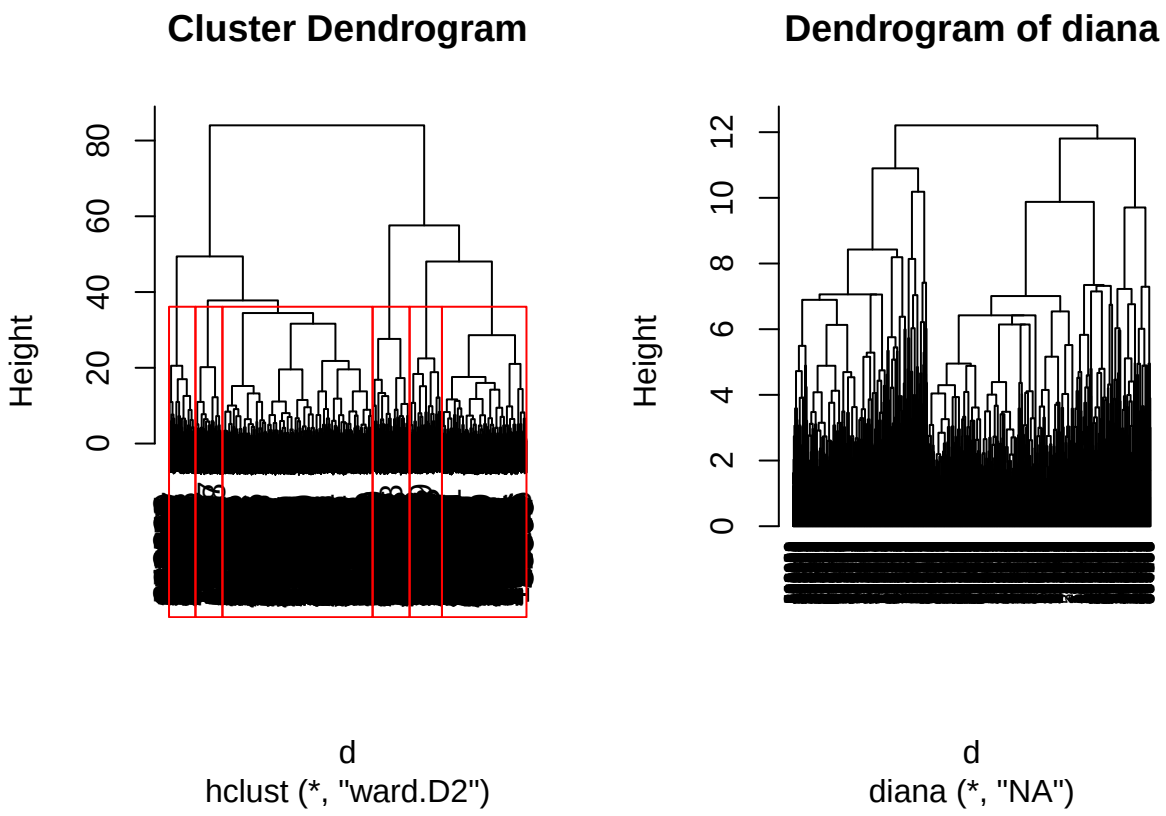
Next we will compare the bottum-up and the top-down approach. We can do see both clusterings in the dendrograms. Additionally we can compare the agglomerative and the divise coefficient of the two algorithms. They indicate how much clustering structure was found. The results suggest that the bottom-up approach has found more clustering structure than the the top-down approach. However, this might not come from their direction of clustering, but the fact that the agglomerative is using the "ward" method. The high values indicate that both algorithms are able to find good clustering structures, as 1 would be the maximum value.

```
## [1] "Agnes"
```

```
## [1] 0.9881532
```

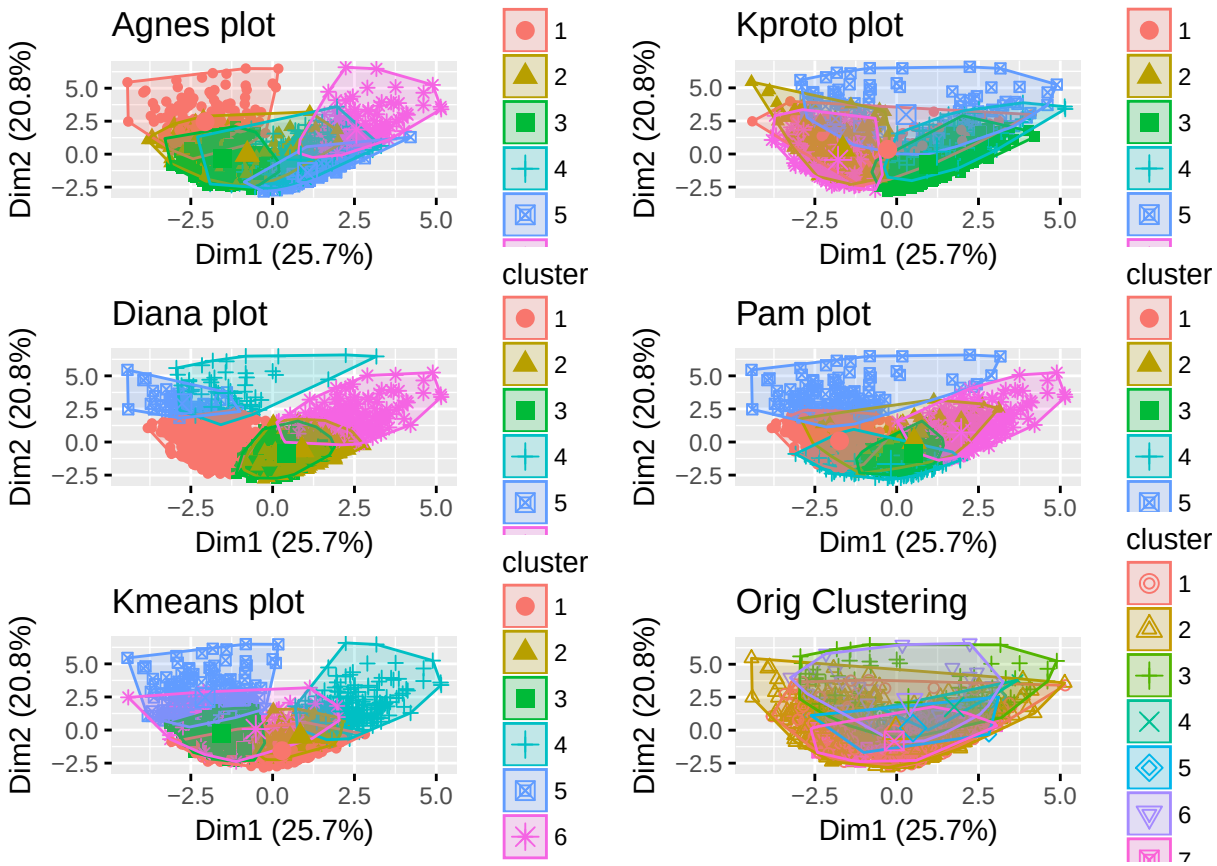
```
## [1] "Diana"
```

```
## [1] 0.9059308
```



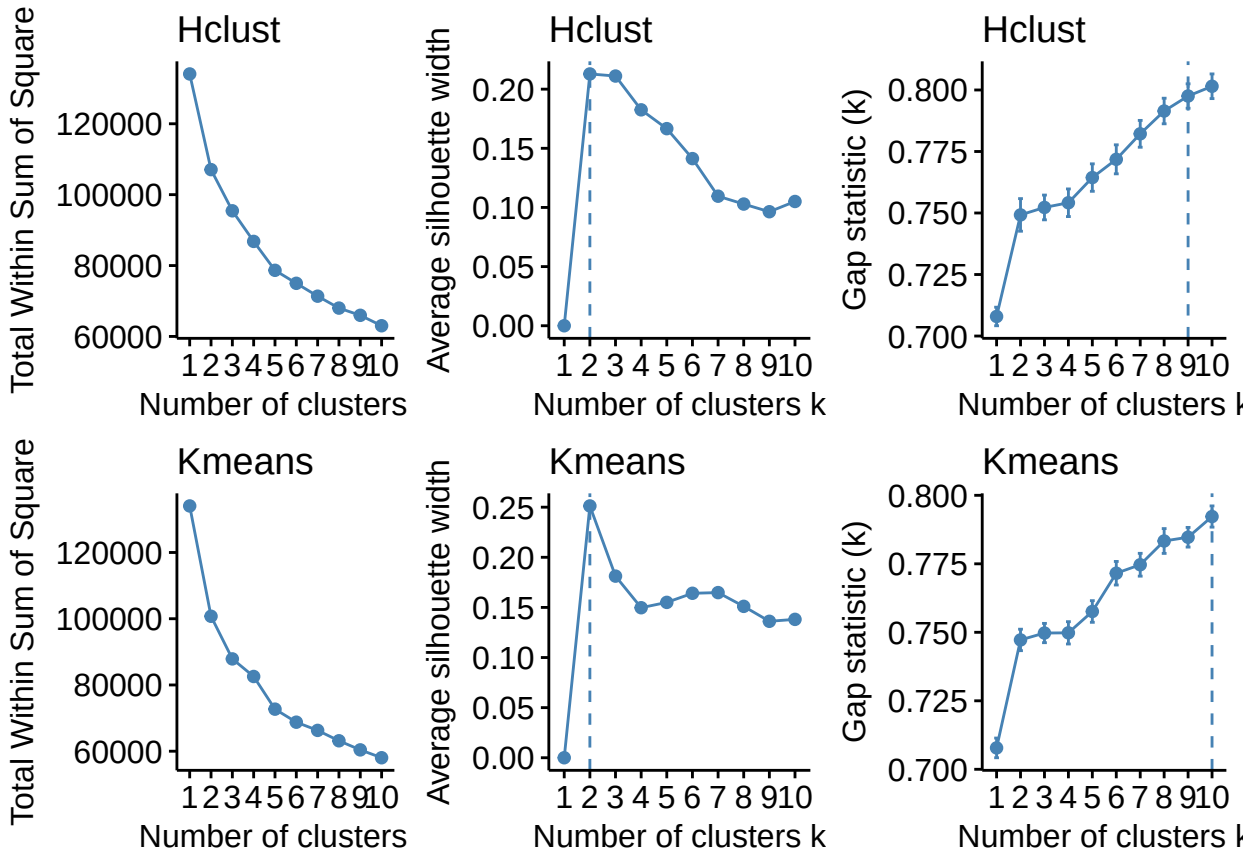

```
## Estimated lambda: 3.392893
```

The next few plots will display the different clusterings acquired by our algorithms. The axes that are being displayed are chosen based on a principal component analysis. One should note the 'Orig Clustering' plot where I show the clustering if we used the clustering by the original Cover_Type values. As we can see, the original clusterings do not spread over the two variables with the most variance but they seem to overlap. This makes it very hard for the clustering algorithms to detect these clusters, as they try to minimize the intracluster distance and thus the variance. From a visual perspective, the bottom-up hierarchical approach and kmeans appear to create similar results, as well as kproto and Pam, where Pam is a kmedoid algorithm that chooses data points as cluster centers. The top-down approach captures the cluster in the bottom left quite well, however, this might just be a coincidence. Additionally it also appears to be a coincidence that the methods used to choose the amount of clusters seem to line up with the number of different cover types at k=6 / k=7. One could have assumed that the clustering structure that was detected was indeed connected to the Cover_Type. The plot by kproto might suggest that the addition of the Wilderness_Area variable indeed increased the intrinsic clustering structure related to the Cover_Type.



Since we were not able to find a clustering that would indicate a natural clustering of Cover_Type, we need to approach it differently. Let us therefore look at a different distance measure, namely the Manhattan distance.

This time there appears to be no decisive best number of clusters in the range of 1 - 10, so we are encountering a similar situation as we did before with the AIC and BIC. As before, there appears to be a spike at k=2 for the Silhouette Method. Let us investigate k=2 a bit more.



As one can see, k=2 with the Manhattan distance achieves better coefficients for the hierarchical clustering than k=6 with the Euclidean distance.

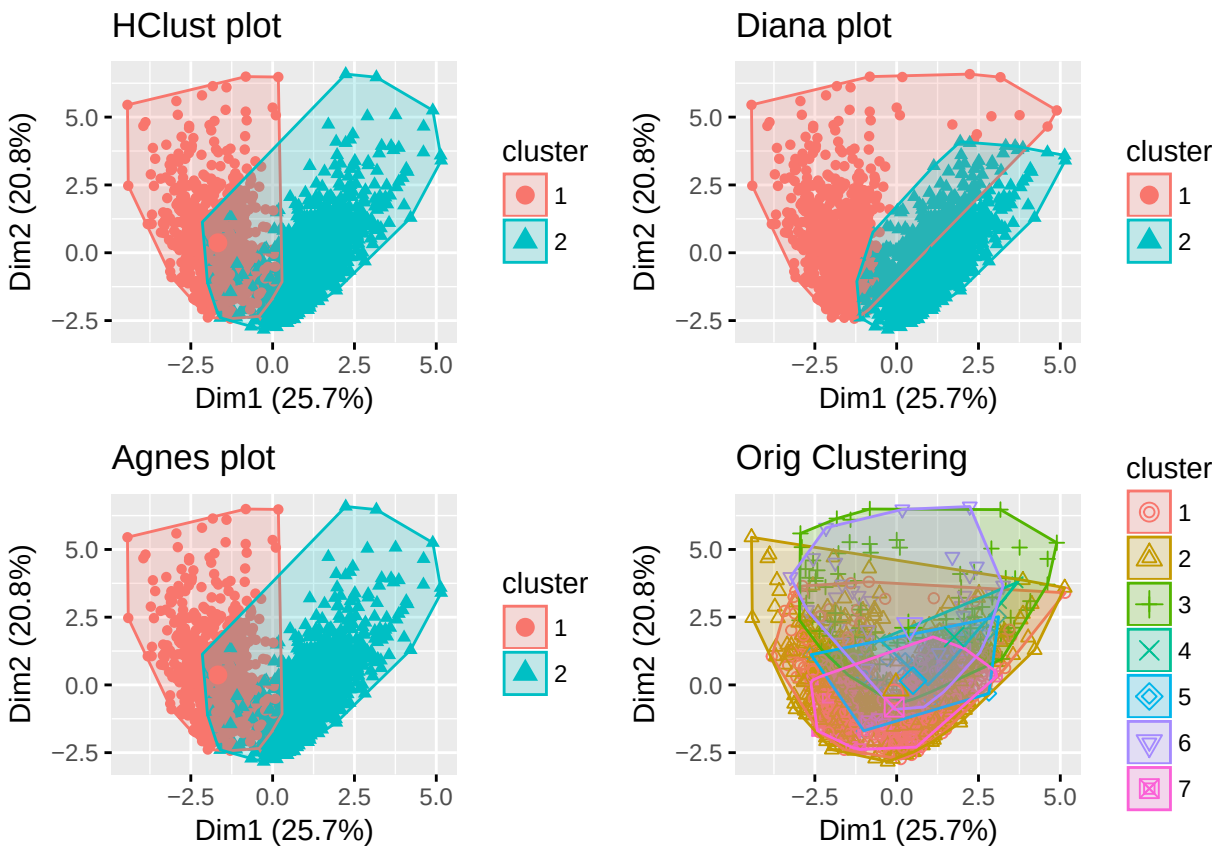
```
## [1] "Agnes"
```

```
## [1] 0.9903049
```

```
## [1] "Diana"
```

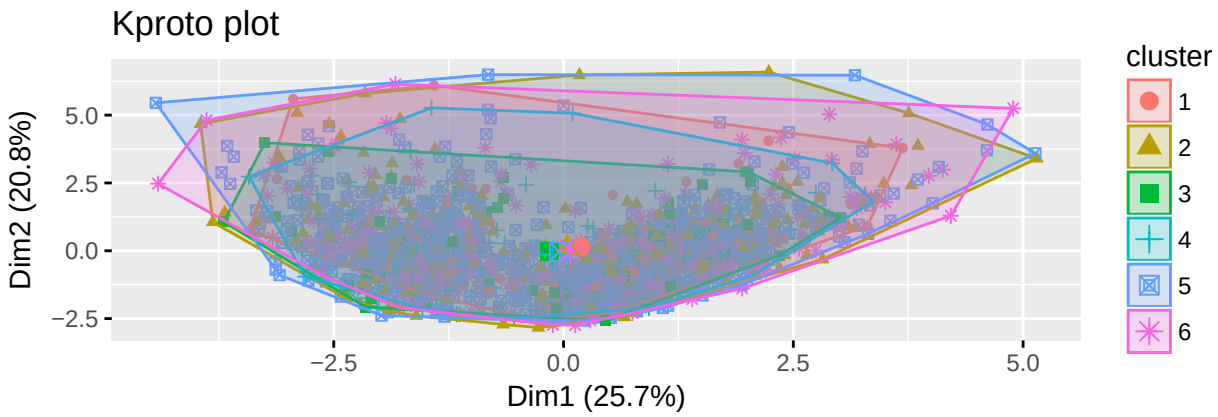
```
## [1] 0.9203888
```

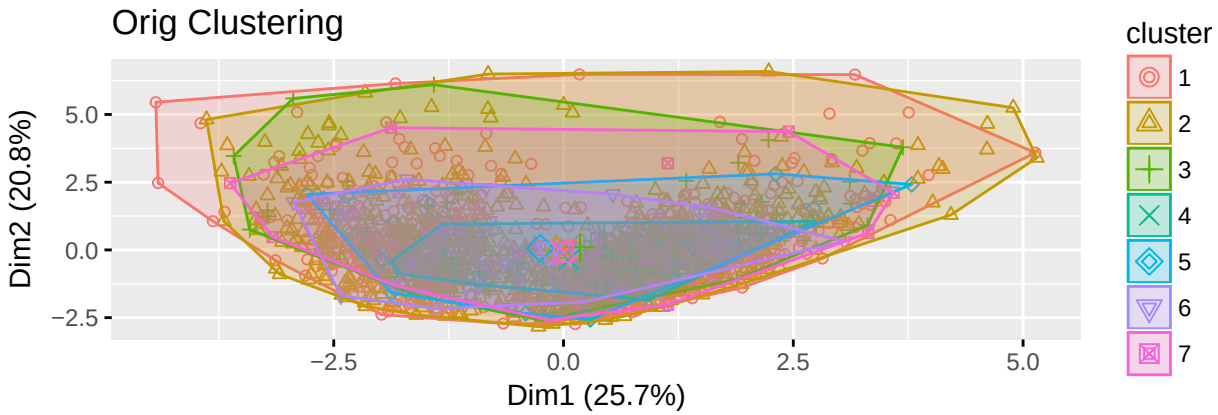
Obviously, the two clusters are again not resembling in any way the clustering for Cover_Type. However, it appears that there seems to be some natural division into two clusters, which the Silhouette Method detected. The two groups are nearly linearly separable. Conclusively one can say that in the variables chosen, there appears to be no clustering structure related to the Cover_Type. However, given that I excluded the Soil_Type in the beginning, there might be some clustering structure added through the Soil_Type variable in connection to the cover type. We already saw in the Euclidean case, that by including the Wilderness_Area variable, we were able to bring the cluster structure closer to what would be the original cluster. Because of these two findings, I will continue this analysis by including the Soil_Type variables and using kproto.



As we can see, the clusters found by kproto resemble much more the clustering based on now that we included the Soil_Type. The centroids are close to one another.

```
## Estimated lambda: 14.44987
```





The following table indicates that the results are not as close as they appear. For example the kproto clustering is assigning the observations that would be part of either (t1) or (t2) across all of its clusters. This is suboptimal as we would expect that there are 2 mayor types and not 4, since the distribution of Cover_Type suggest so.

```
##
##        1   2   3   4   5   6   7
##   1    0   9  85  11   0  23   0
##   2  143 227   0   0   5   0   2
##   3    5 103   0   0   0   0   0
##   4   71  99   0   0   0   0   1
##   5  260 370  58   0  15  27  54
##   6  230 180   0   0   7   0  15
```

Citations

https://cran.r-project.org/web/packages/ClusterR/vignettes/the_clusterR_package.html

https://archive.ics.uci.edu/ml/datasets/covertype

https://uc-r.github.io/kmeans_clustering

http://uc-r.github.io/hc_clustering