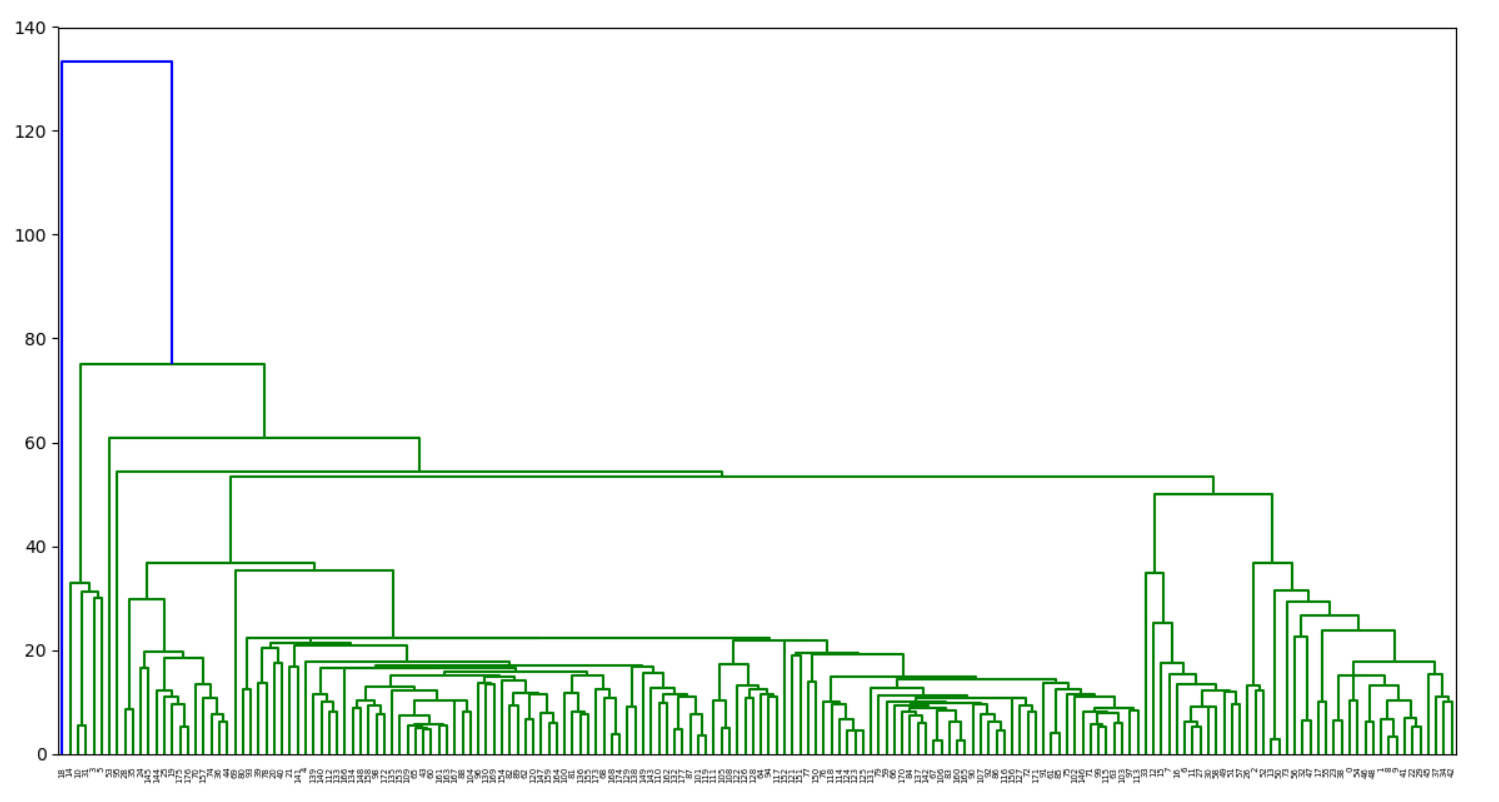
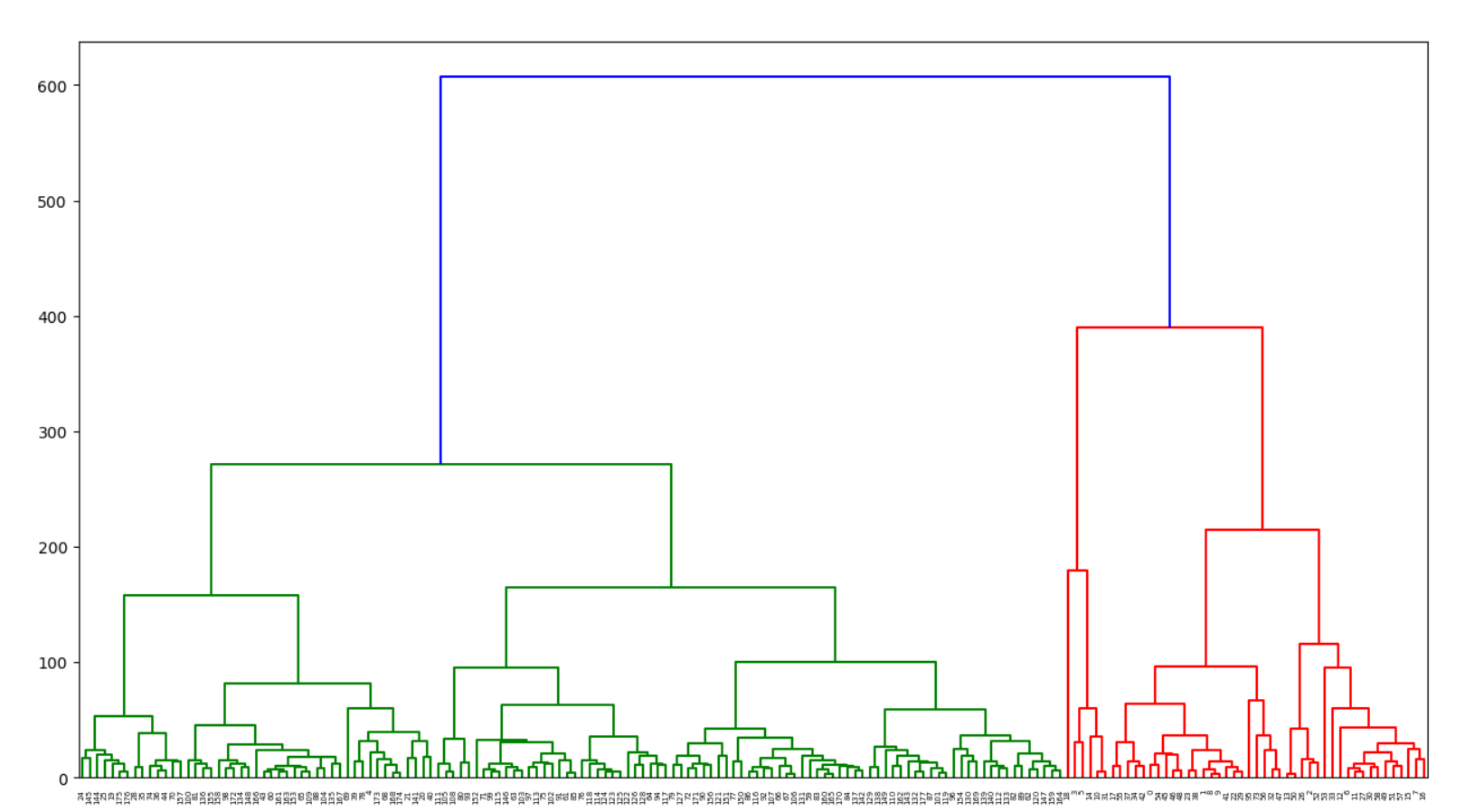


Complete link-using max distance



Single link-using min distance



Average

1. Difference between complete, single link and average

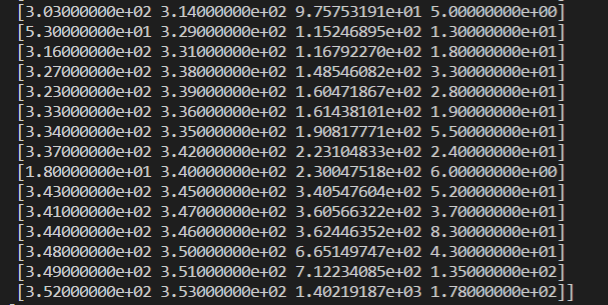
If we compare the complete and single link, we can see that complete link has a bigger range (from 0 to 1400) compare to single link (from 0 to 140) as complete link is using the maximum distance; while single link is using the minimum distance.

The bar chart generated by single link is long, scraggly clusters compare to complete link.

Therefore, complete and average link need more comparison as the dendrogram is more spread out, evenly spread. Single link is more dense than complete and average link, so it requires less comparison.

1. Distance of which cluster merge occurs and possible pattern

Complete link:



This is the last merges of the complete link. We can notice that the 3rd column of the last row, has a distance of 1402, which is a very big distance, and it jumps from 712, this indicates that 2 clusters shouldn’t be merged were combined into 1. We can see that the 4th column first and second row has 43 and 135 observations and the sum of them formed the last row, which is 178 observations.

Single link:



Also, there is a big jump from 75 to 133, indicates that the clusters that 2 clusters shouldn’t be merged were combined into 1. Noted that the last column, from 177 points to 178 points, in this case, just 1 point (18) had been merged with the 177 points cluster, indicates this point may be outliers as it has the biggest distance among all other points and has been merged the last. Therefore, this can show that single link is easier to detect outliers.

1. (a) Compare with K-means clustering solution accuracy score comparison

|  |  |
| --- | --- |
| Complete | 0.0 to 0.51 |
| Single | 0.0 to 0.41 |
| Average | 0.0 to 0.42 |

Based on the dendrogram, the complete and average is less likely to have a same converging point after running few times as they have a more obvious cluster separated. Single is more likely to have a converges point as the number of cluster based on observation is less.

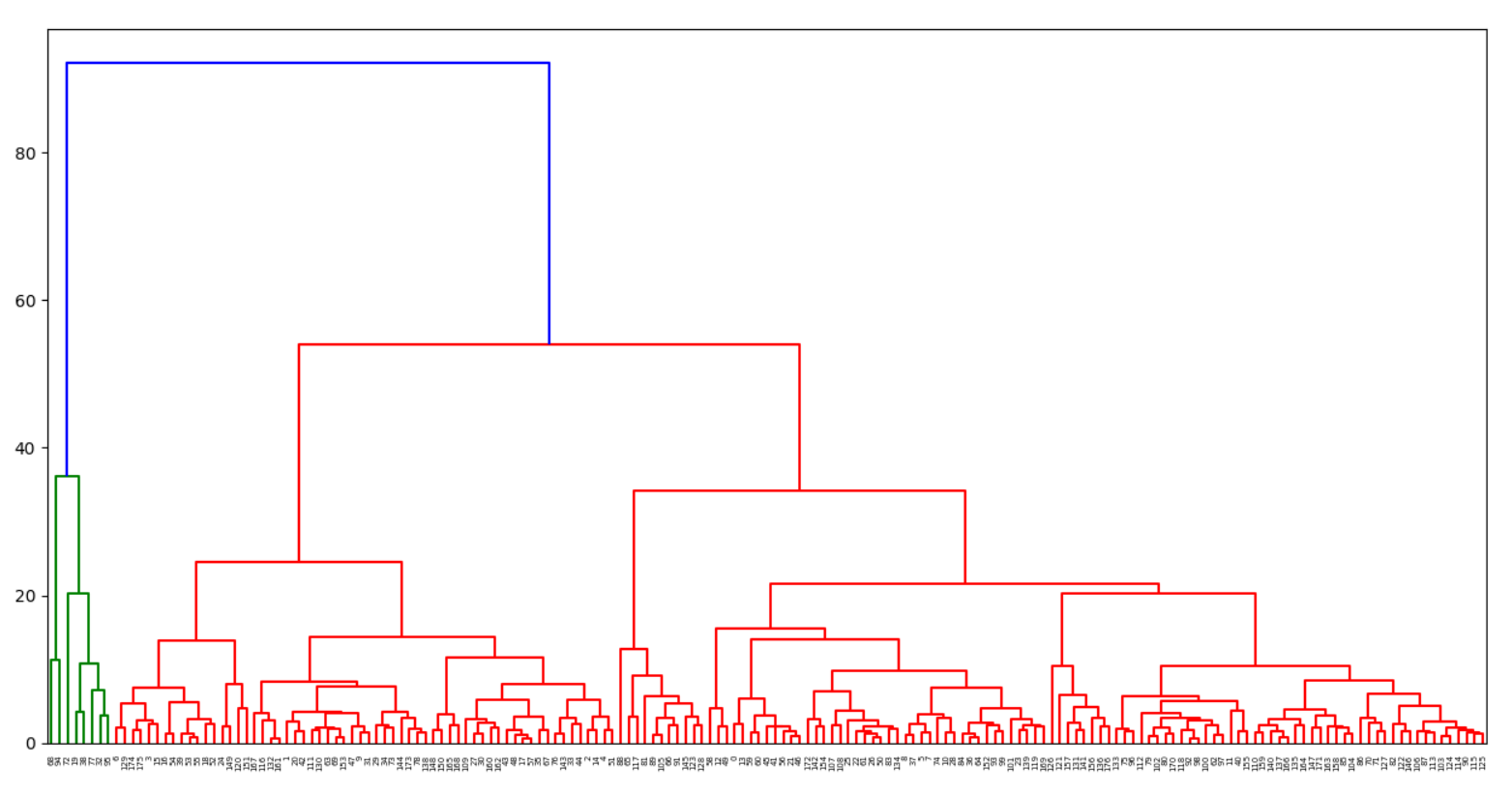
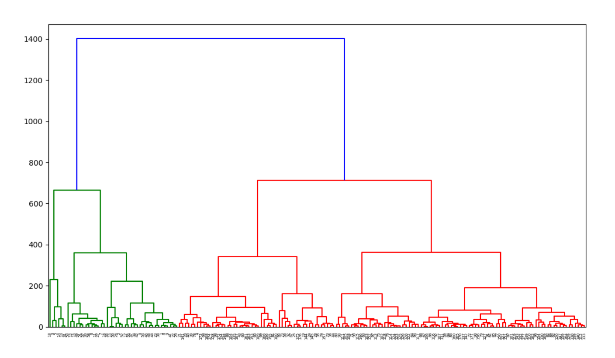
(b) Compare with actual groupings accuracy score comparison

|  |  |
| --- | --- |
| Complete | 0.47752808988764045 |
| Single | 0.42696629213483145 |
| Average | 0.07303370786516854 |

We can see that the result of comparing with actual groupings is constant, because the value is constant, it will not fluctuate like K-means. Single has a lower accuracy score because it accepts the outlier. Since data set has a wide spread, the average value becomes inaccurate when use to determine the clustering solution. Complete has the highest accuracy score

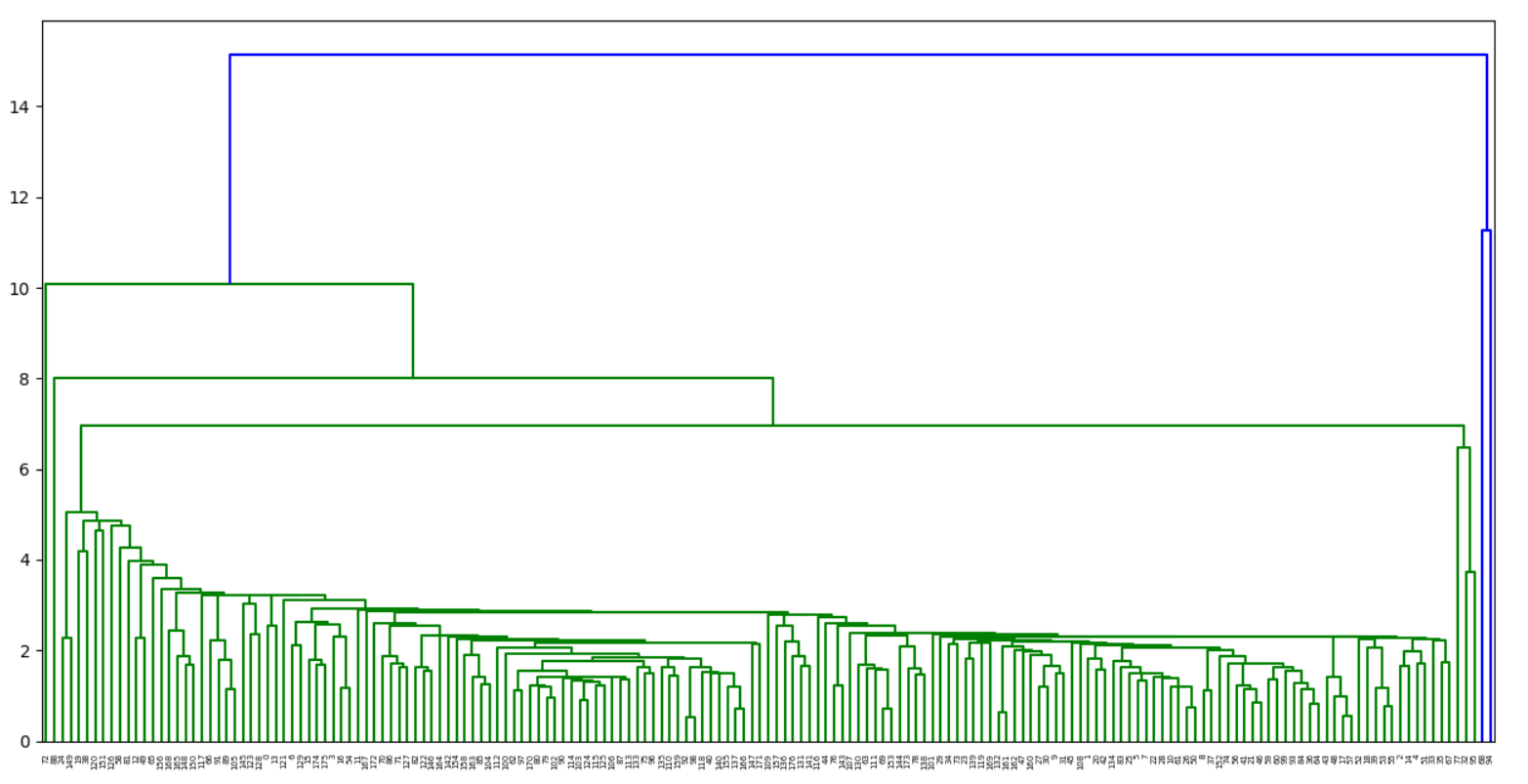
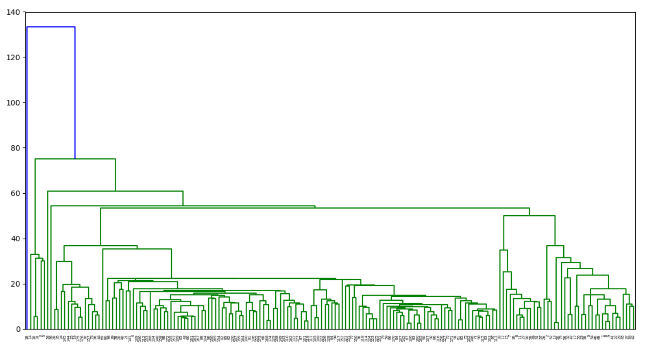
1. Compare the subset with original

All subset is using the first 6 attributes.



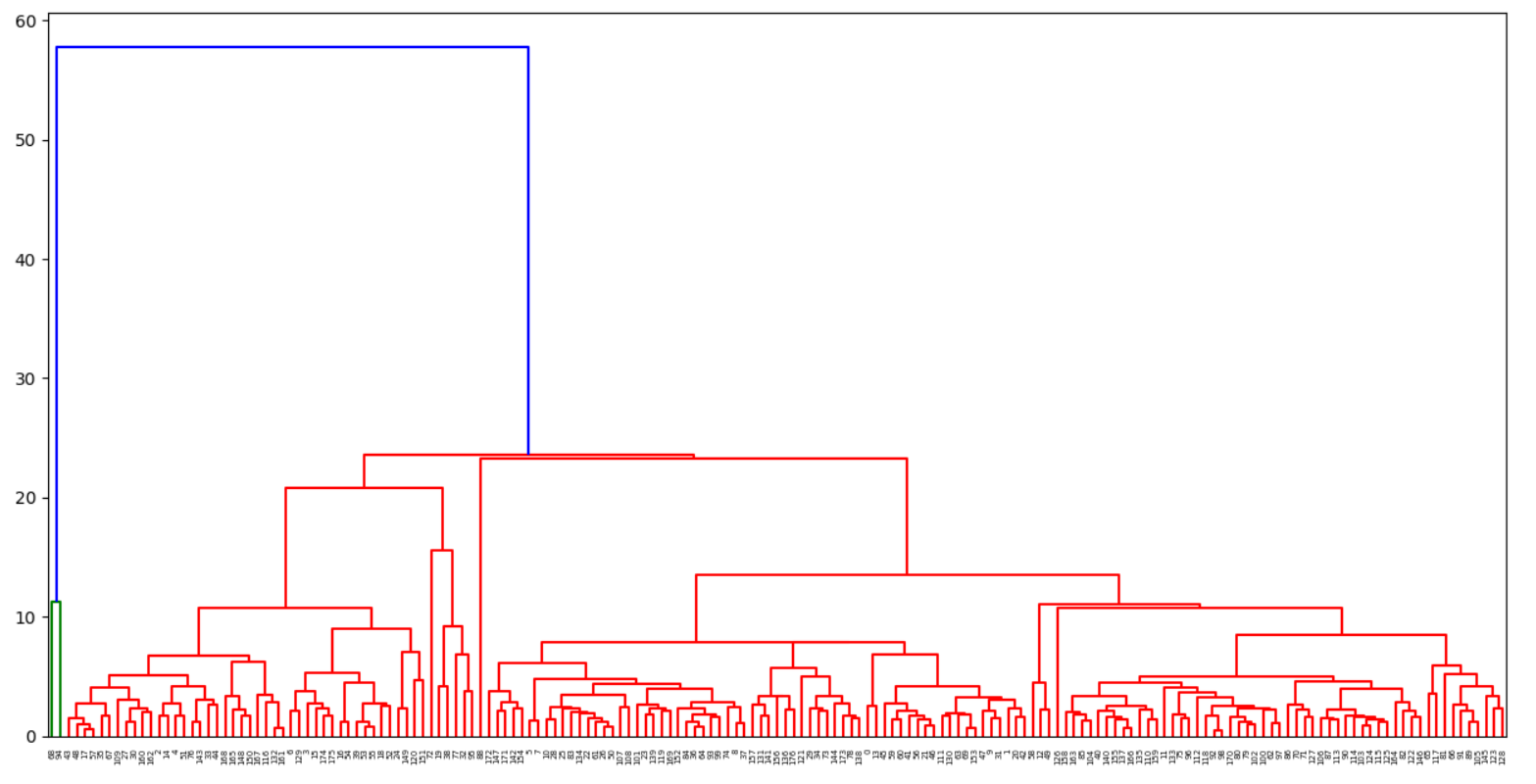
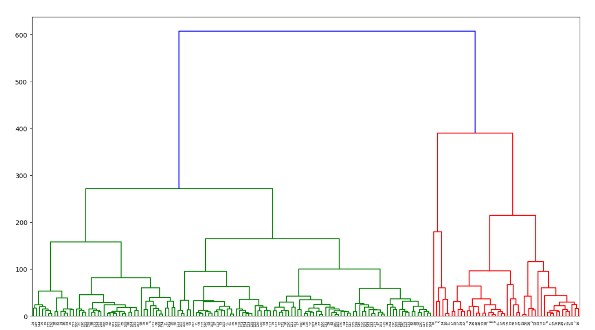
Subset of complete link

We can see that by eliminating almost of half the attributes, the distance differences significantly dropped form range 1400 to 90. We can still see the trend of the subset is quite similar with the original. However, some of the points which should be in the green cluster has been merged to be red cluster. This may be related to the eliminated of some of the important attributes which determine it is the green cluster.



Subset of single link

Also, the same case for single link, the distance range dropped from 140 to only 15. The result is very similar compare means that the data attributes does not affect.



Subset of average

The average case is similar with the complete link. Some of the points which should be in the green cluster has been merged to be red cluster.