Applied Multivariate Analysis

Clustering Project

The American University In Cairo

Beshoy M. Botros 900226446

Ahmed M. Monir 900226418

Spring 2025

27.04.2025

Abstract

The primary goal is to group the wine samples into natural clusters that correspond, as closely as possible, to their true underlying types, without using any label information. We experiment with a wide range of distance metrics and linkage methods for hierarchical clustering and compare the results to k-means clustering. Each clustering configuration is evaluated using the R^2 statistic, which measures the proportion of variance explained by the clusters, as well as the compactness and balance of the resulting cluster sizes.

Data Source

Our data is about wine types, and it includes attributes such as:

- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

Here is our link for the data.

Methodology

We used two methods in cluster analysis, which are:

- K-means
- The Heirarichal

In order to apply these methods, we have to get the distances between clusters and between observations. However, there are many methods to get these distances. So, we wanted to use all possible combinations between them.

Distances between observations Used: Euclidean - Manhattan - Canberra - Minkowski

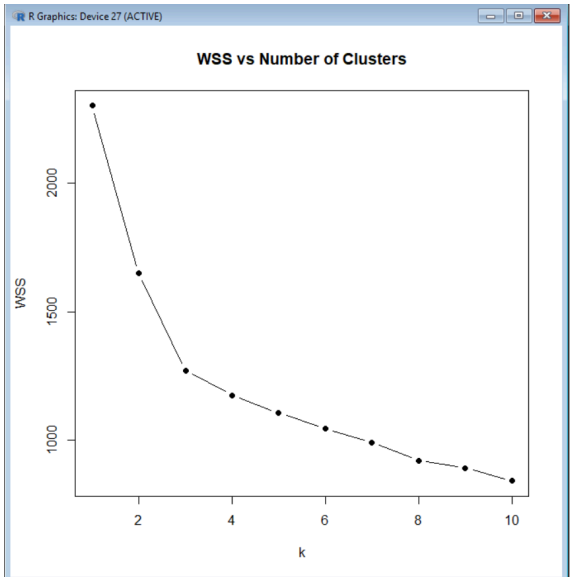
Distances between clusters Used: "ward", "single", "complete", "average", "centroid", "median", "mcquitty"

RESULTS

1. K-means

The output of the K-means clustering algorithm was as follows:

The graph between the number of clusters and WSS shows that the number of clusters that minimizes the within sum of squares of the clusters is k = 3 according to the elbow method.



According to the K-means clustering, these are three groups of wine with sizes 62, 51, and 65 respectively, as shown in the output.

```
K-means cluster sizes:
> print(table(kmc$cluster))

 1  2  3
62 51 65
>
```

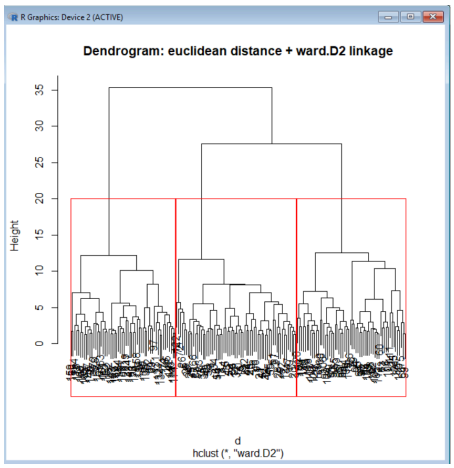
To assess this method, we calculated the R^2 ratio, which is WSS/TSS. This ratio was 0.447, which is classified to be a moderate clustering algorithm.

```
> cat("K-means R²:", km_r2, "\n")
K-means R²: 0.4477405
>
```

2. Hierarchical Method

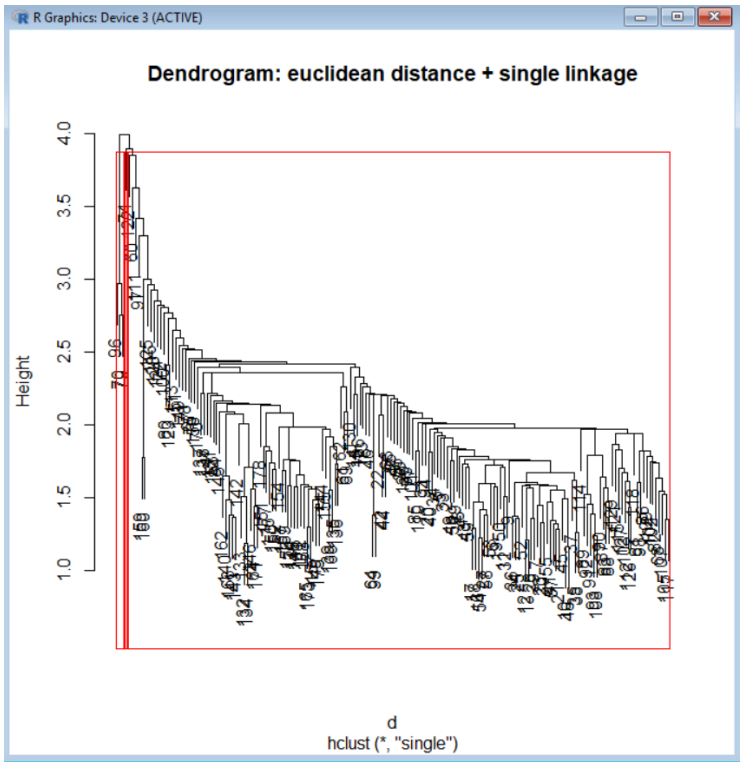
As mentioned above, there are many combinations of methods to be tested. For each combination, the code generates a dendogram that demonstrates the clustering analysis done using the this combination and the code output that demonstrates the size of the clusters, its R^2, and the centers of the clusters. At the end, we calculated the R^2 for each combination to get the best one of them and compare it with the K-means clustering Algorithm.

a) Euclidean Distance + Ward's Method



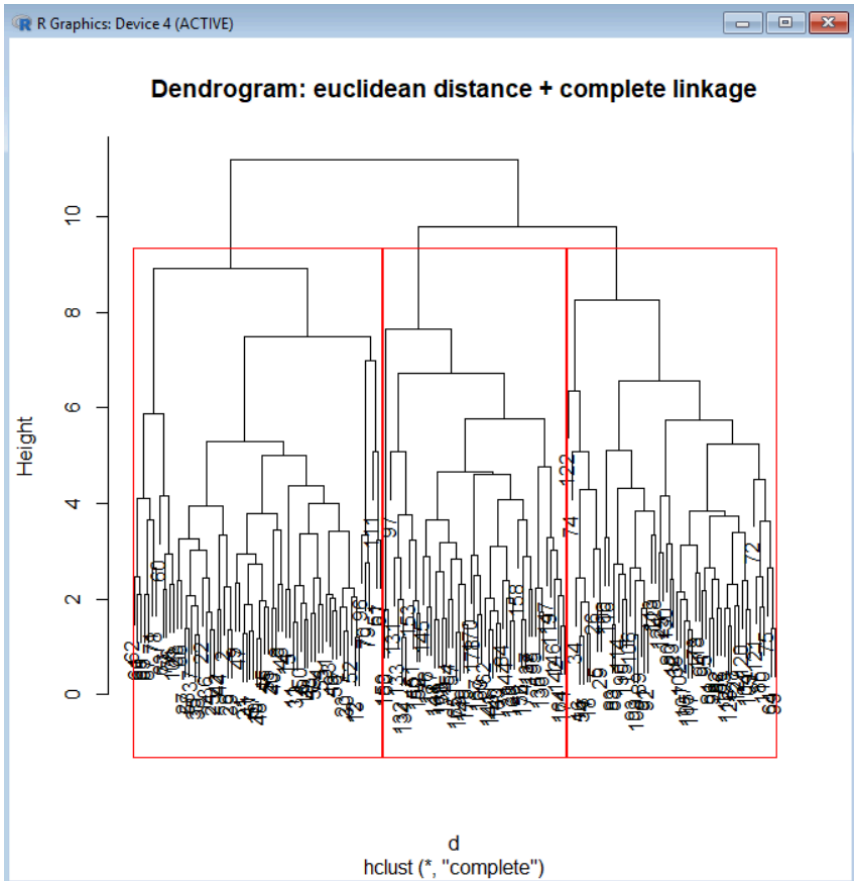
```
=== Analyzing with: euclidean distance and ward.D2 linkage ===
Cluster sizes:
clusters
 1  2  3
64 58 56
R-squared: 0.4360204
Cluster centers:
      Alcohol Malic_Acid        Ash Ash_Alcanity    Magnesium Total_Phenols
1  0.82357590 -0.3279323  0.3521424   -0.5889448  0.449128960     0.8866129
2 -0.98130872 -0.3557125 -0.5516594    0.2137036 -0.503403269    -0.0514201
3  0.07512586  0.7431963  0.1689131    0.4517439  0.008091718    -0.9600154
   Flavanoids Nonflavanoid_Phenols Proanthocyanins Color_Intensity        Hue
1  0.98153808          -0.56856797       0.5542434       0.1690517  0.4985608
2  0.05890101          -0.05230157       0.1671204      -0.9329070  0.4486584
3 -1.18276242           0.70396144      -0.8065100       0.7730232 -1.0344657
       OD280    Proline
1  0.7700838  1.0452391
2  0.3529005 -0.7794741
3 -1.2455998 -0.3872465
```

b) Euclidean Distance + Single Linkage



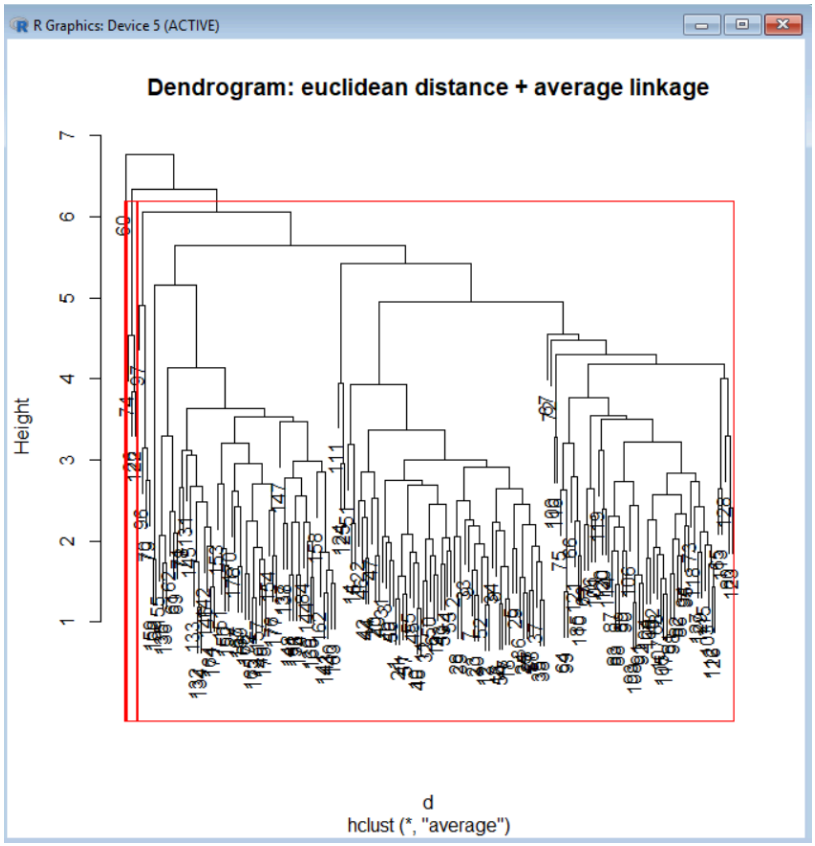
```
=== Analyzing with: euclidean distance and single linkage ===
Cluster sizes:
clusters
  1   2   3
174   3   1
R-squared: 0.04435988
Cluster centers:
      Alcohol Malic_Acid         Ash Ash_Alcanity   Magnesium Total_Phenols
1  0.01417601  0.0204513  0.02023777 -0.004509201 -0.07606525  -0.003393545
2 -0.81784874 -0.9873507 -1.45747738 -0.787012118  3.49554709  -0.338385670
3 -0.01307912 -0.5964737  0.85105976  3.145637249  2.74871152   1.605633868
     Flavanoids Nonflavanoid_Phenols Proanthocyanins Color_Intensity
1 -0.0009949364            0.01973762     -0.04153415      0.01991255
2 -0.2295316308           -0.73805920      2.19402174     -0.90933129
3  0.8617138240           -1.22016759      0.64487710     -0.73679001
          Hue        OD280      Proline
1 -0.02464588 -0.008606708 -0.007344375
2  0.91532482  0.082134356  0.173934926
3  1.54240778  1.251164200  0.756116514
```

c) Euclidean Distance + Complete Linkage



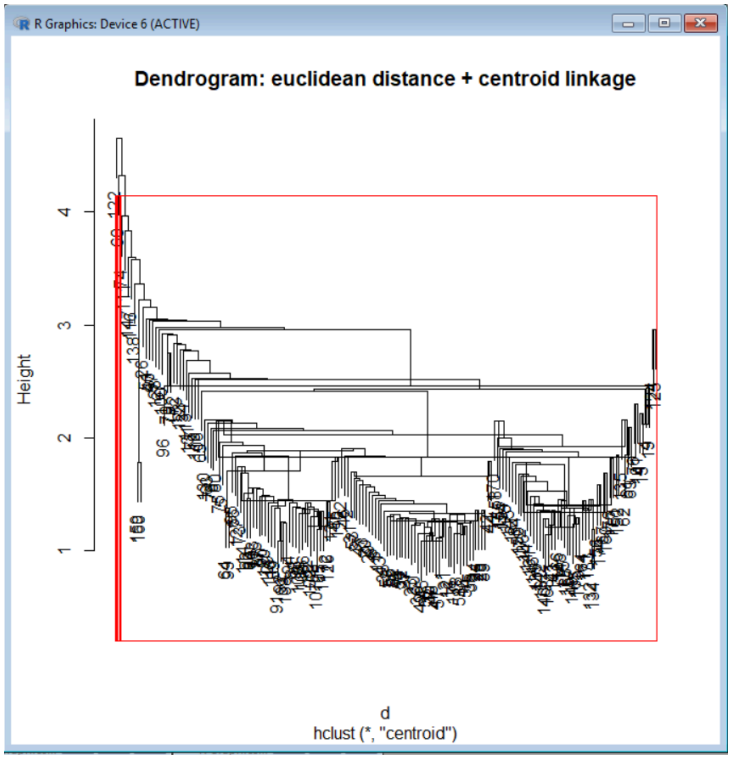
```
=== Analyzing with: euclidean distance and complete linkage ===
Cluster sizes:
clusters
 1  2  3
69 58 51
R-squared: 0.3589266
Cluster centers:
     Alcohol Malic_Acid         Ash Ash_Alcanity   Magnesium Total_Phenols
1  0.4994527 -0.3915001 -0.22132706   -0.8152203  0.36817315     0.5573222
2 -0.7211461 -0.3101839  0.05417453    0.4857813 -0.41286555     0.2072621
3  0.1443968  0.8824348  0.23783225    0.5504879 -0.02858324    -0.9897340
  Flavanoids Nonflavanoid_Phenols Proanthocyanins Color_Intensity        Hue
1  0.6404468          -0.52378880       0.4228111     -0.02905821  0.4908867
2  0.3109128           0.04051815       0.1523600     -0.76683600  0.4396067
3 -1.2200739           0.66257597      -0.7453106      0.91140205 -1.1640857
       OD280    Proline
1  0.5046753  0.7551500
2  0.5165744 -0.5509998
3 -1.2702727 -0.3950464
```

d) Euclidean Distance + Average Linkage



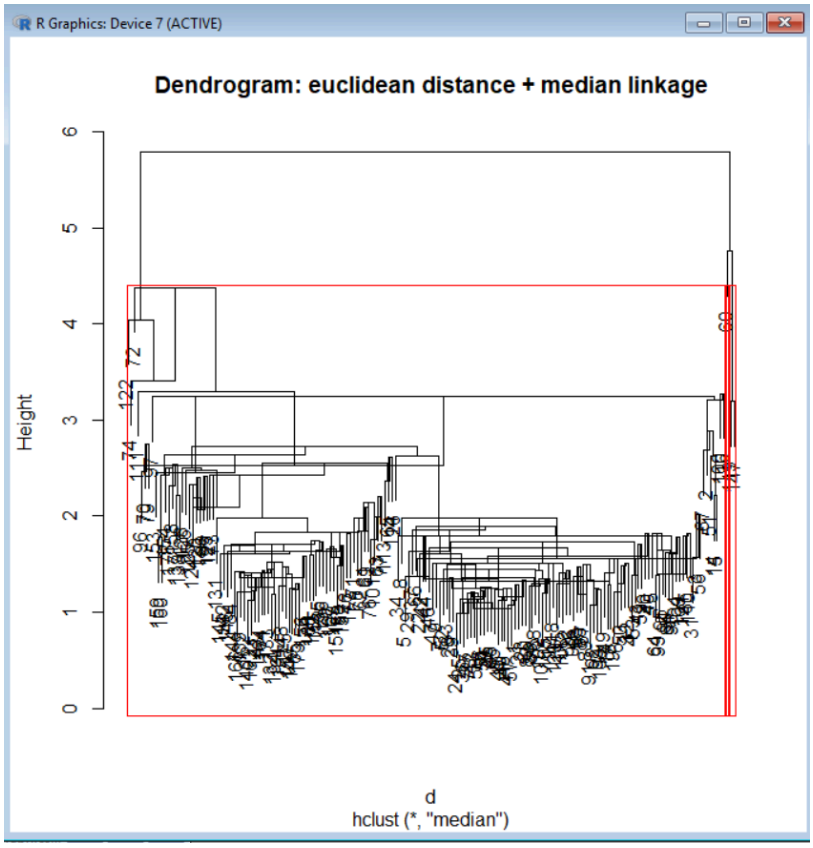
```
=== Analyzing with: euclidean distance and average linkage ===
Cluster sizes:
clusters
  1   2   3
174   3   1
R-squared: 0.04404431
Cluster centers:
      Alcohol  Malic_Acid         Ash Ash_Alcanity   Magnesium Total_Phenols
1  0.01438839  0.01355771 -0.01977408  -0.02774164 -0.02858324   -0.01753521
2 -0.57559666 -0.36970543  2.36983419   2.49684992  1.93186012    1.18487348
3 -0.77678907 -1.24992453 -3.66881295  -2.66350471 -0.82209603   -0.50349418
   Flavanoids Nonflavanoid_Phenols Proanthocyanins Color_Intensity         Hue
1 -0.01785326         0.0008042404     0.001943917      0.01326872 -0.01483985
2  1.52246795         0.1725899899     0.569167024     -0.32269096  0.72574161
3 -1.46093705        -0.6577077995    -2.045742545     -1.34068448  0.40490846
        OD280      Proline
1 -0.01427297  0.003423247
2  1.19952031  0.041620929
3 -1.11506488 -0.720507695
```

e) Euclidean Distance + Centroid Linkage



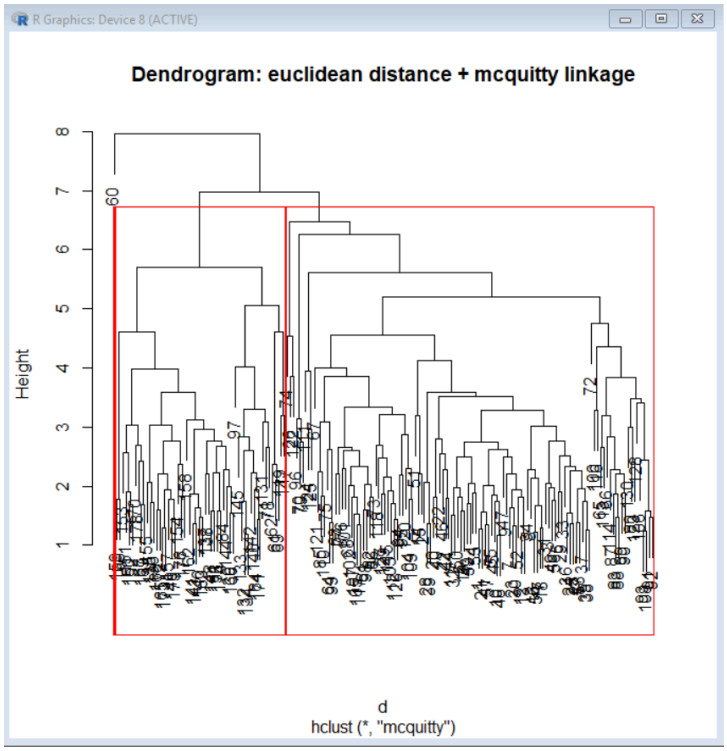
```
=== Analyzing with: euclidean distance and centroid linkage ===
Cluster sizes:
clusters
  1   2   3
176   1   1
R-squared: 0.03131278
Cluster centers:
      Alcohol   Malic_Acid          Ash  Ash_Alcanity    Magnesium
1  0.01449618  0.008558215  0.002962308 -0.0001873414 -0.002990334
2 -0.77678907 -1.249924533 -3.668812953 -2.6635047091 -0.822096033
3 -1.77453915 -0.256321277  3.147446700  2.6964767879  1.348394832
  Total_Phenols   Flavanoids Nonflavanoid_Phenols Proanthocyanins
1  -0.005172732 -0.009052722         -0.001200346     0.008852895
2  -0.503494178 -1.460937052         -0.657707799    -2.045742545
3   1.413894955  3.054216160          0.868968781     0.487633095
  Color_Intensity         Hue        OD280      Proline
1     0.005309026 -0.00161828 -0.002293799  0.009179942
2    -1.340684477  0.40490846 -1.115064881 -0.720507695
3     0.406295932 -0.12009122  1.518773441 -0.895162171
```

f) Euclidean Distance + Median Linkage



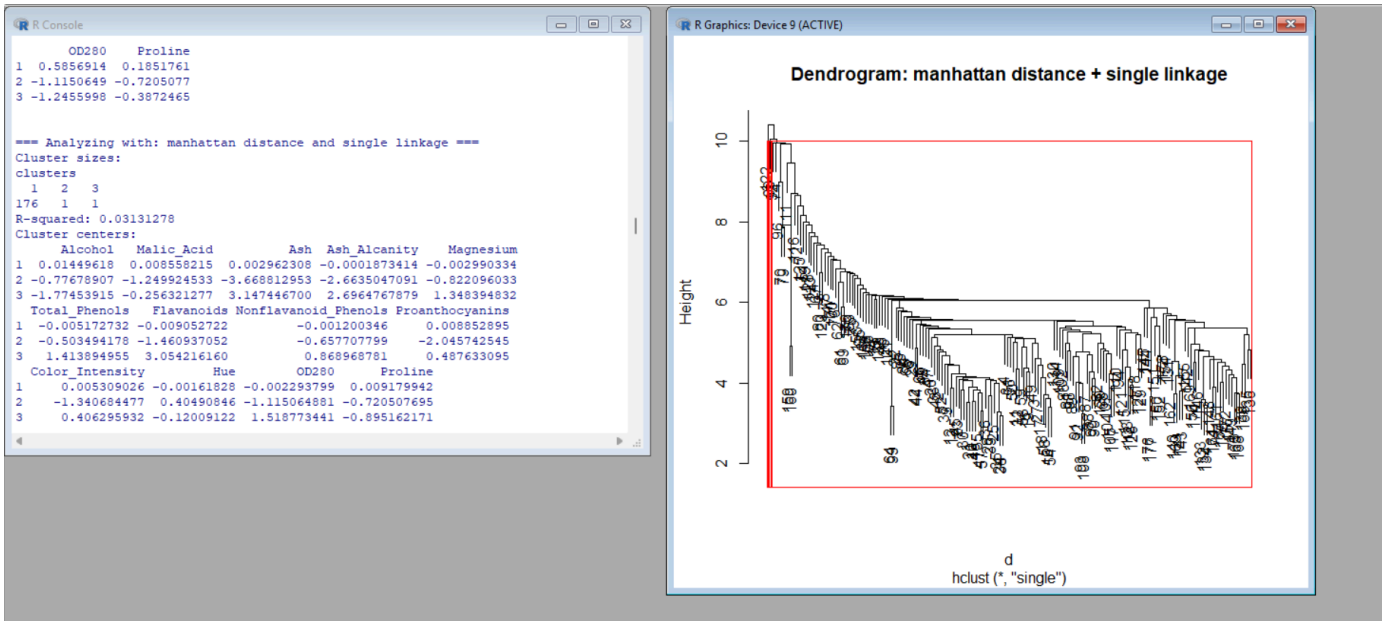
```
=== Analyzing with: euclidean distance and median linkage ===
Cluster sizes:
clusters
  1   2   3
175   1   2
R-squared: 0.03024449
Cluster centers:
        Alcohol  Malic_Acid         Ash Ash_Alcanity   Magnesium Total_Phenols
1 -0.0001277266 -0.01228106  0.03185887   0.02033599  0.02049452     0.0209574
2 -0.7767890652 -1.24992453 -3.66881295  -2.66350471 -0.82209603    -0.5034942
3  0.3995706069  1.69955540 -0.95324427  -0.44764644 -1.38222271    -1.5820256
   Flavanoids Nonflavanoid_Phenols Proanthocyanins Color_Intensity        Hue
1  0.02247027         -0.001122082      0.02838076      0.01213771  0.0135587
2 -1.46093705         -0.657707799     -2.04574255     -1.34068448  0.4049085
3 -1.23567996          0.427036087     -1.46044541     -0.39170747 -1.3888405
        OD280     Proline
1  0.02064459  0.01694248
2 -1.11506488 -0.72050769
3 -1.24886950 -1.12221299
```

g) Euclidean Distance + Mcquitty linkage

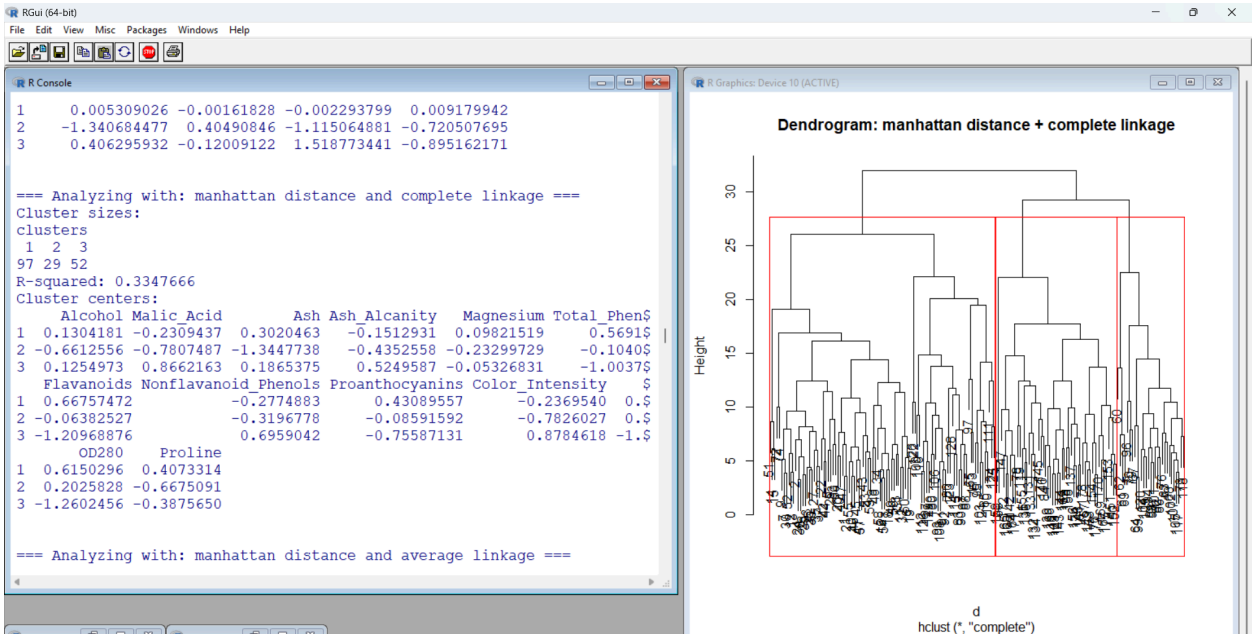


```
=== Analyzing with: euclidean distance and mcquitty linkage ===
Cluster sizes:
clusters
  1   2   3
121   1  56
R-squared: 0.2867781
Cluster centers:
      Alcohol Malic_Acid         Ash Ash_Alcanity    Magnesium Total_Phenols
1 -0.02834925 -0.3336287 -0.04785388   -0.1870591  0.003049255     0.4484658
2 -0.77678907 -1.2499245 -3.66881295   -2.6635047 -0.822096033    -0.5034942
3  0.07512586  0.7431963  0.16891307    0.4517439  0.008091718    -0.9600154
  Flavanoids Nonflavanoid_Phenols Proanthocyanins Color_Intensity        Hue
1   0.559468            -0.3203647       0.3901678      -0.3466827  0.4754146
2  -1.460937            -0.6577078      -2.0457425      -1.3406845  0.4049085
3  -1.182762             0.7039614      -0.8065100       0.7730232 -1.0344657
       OD280    Proline
1  0.5856914  0.1851761
2 -1.1150649 -0.7205077
3 -1.2455998 -0.3872465
```

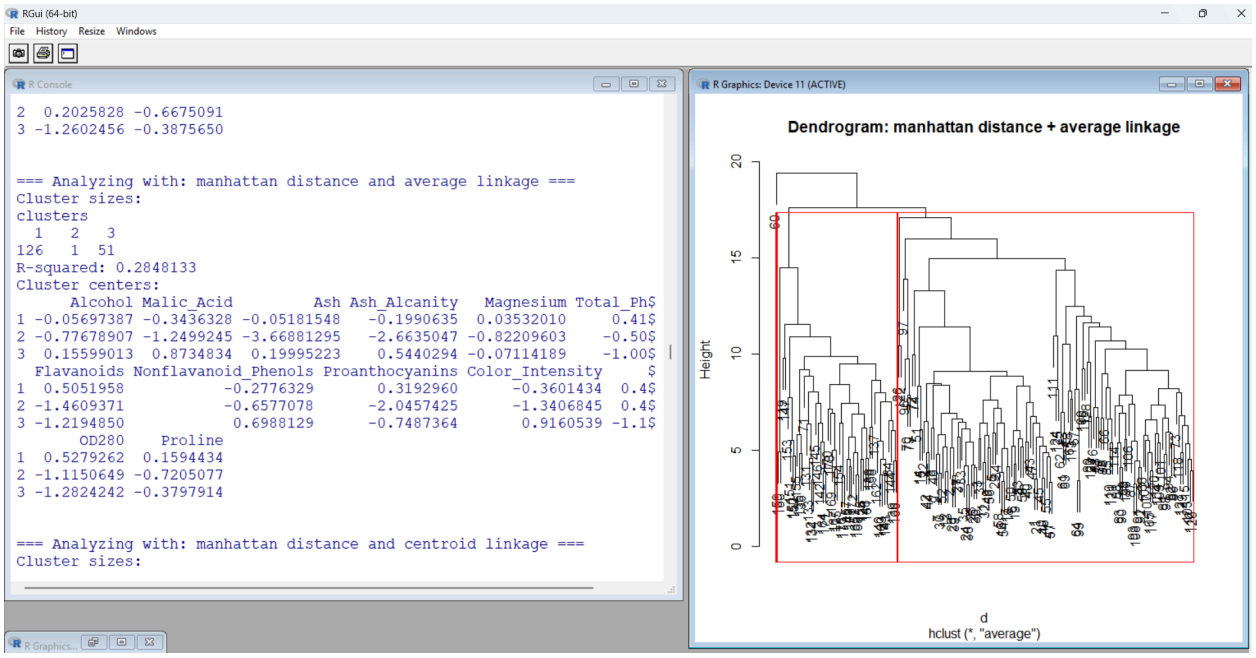
h) Manhattan Distance + Single Linkage



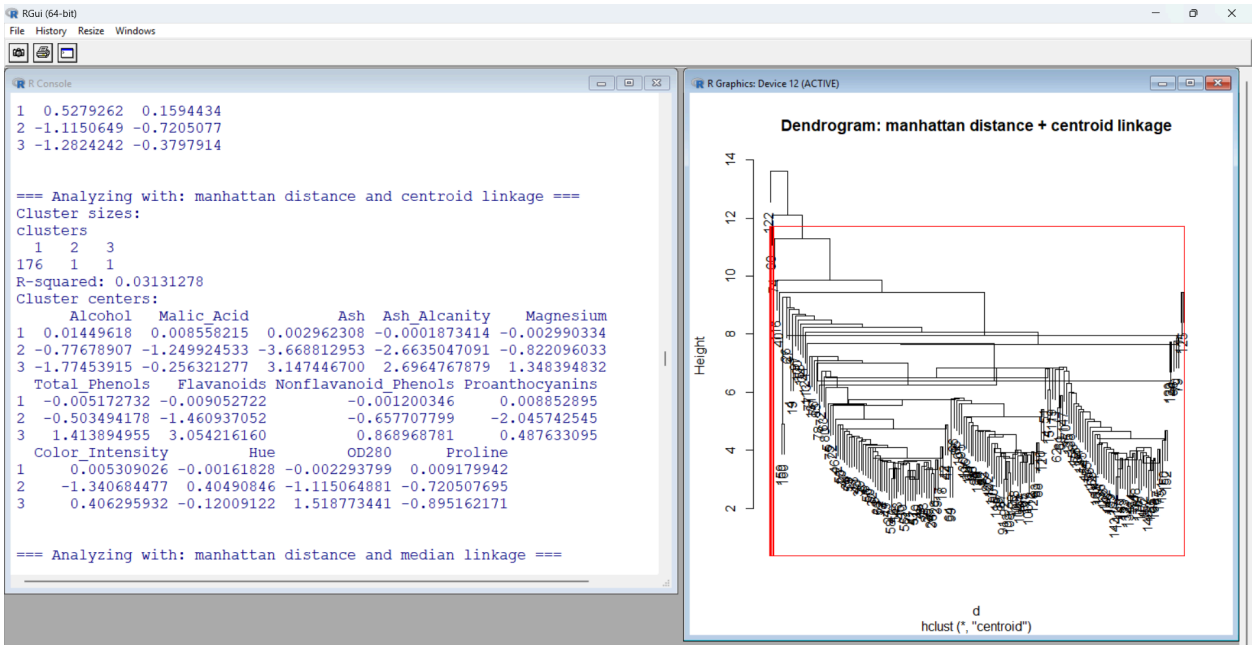
I) Manhattan + Complete Linkage



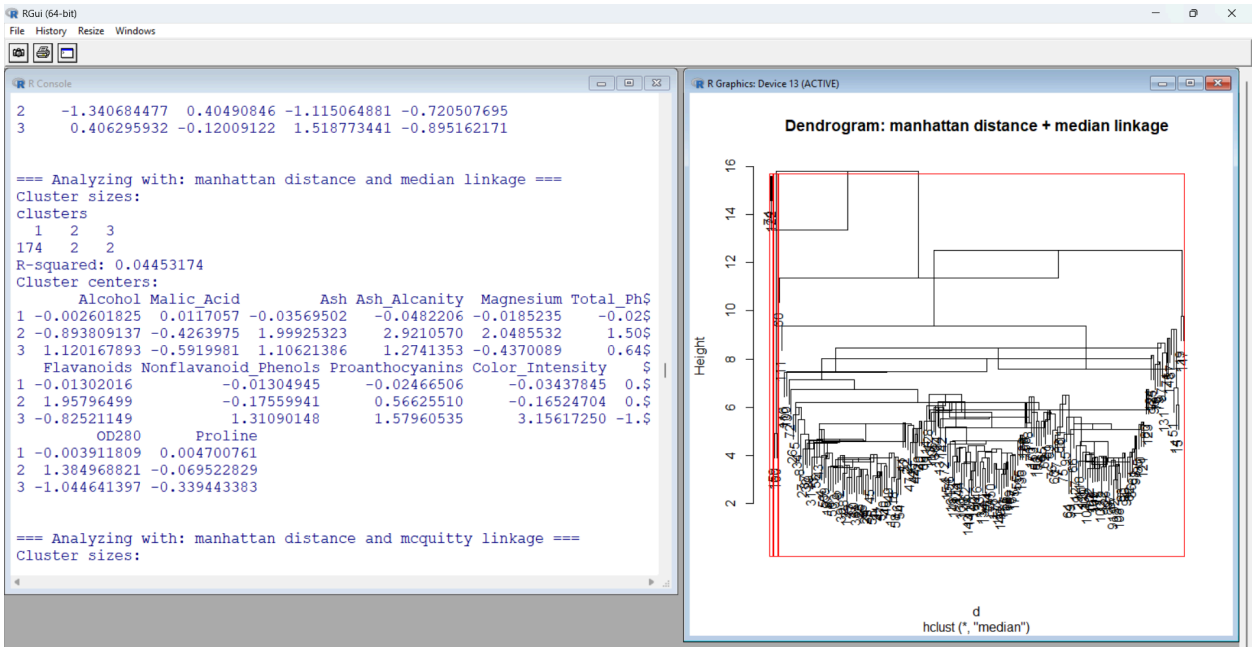
j) Manhattan + Average Linkage



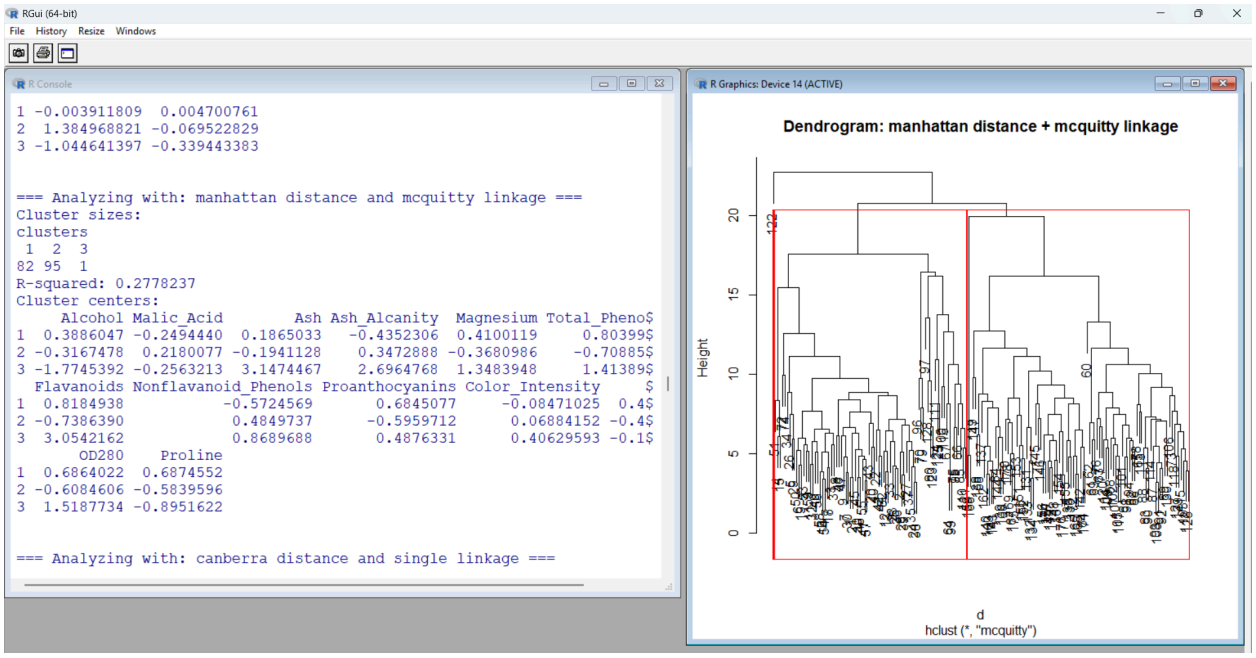
k) Manhattan + Centroid Linkage



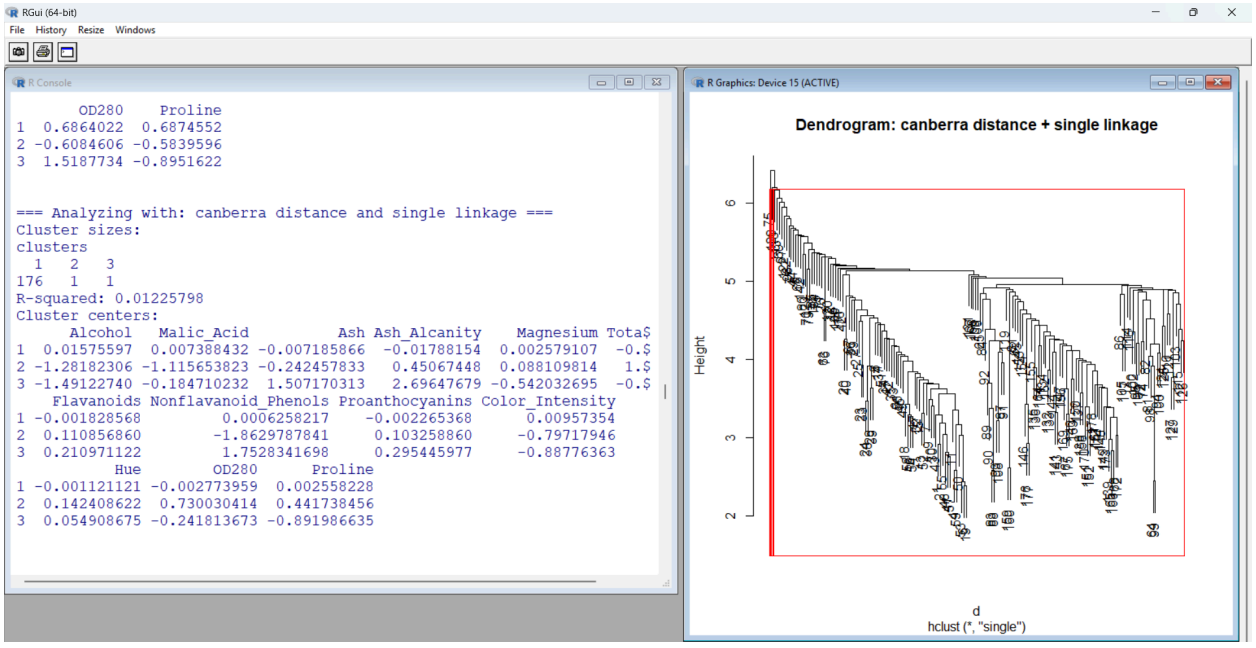
l) Manhattan + Median Linkage



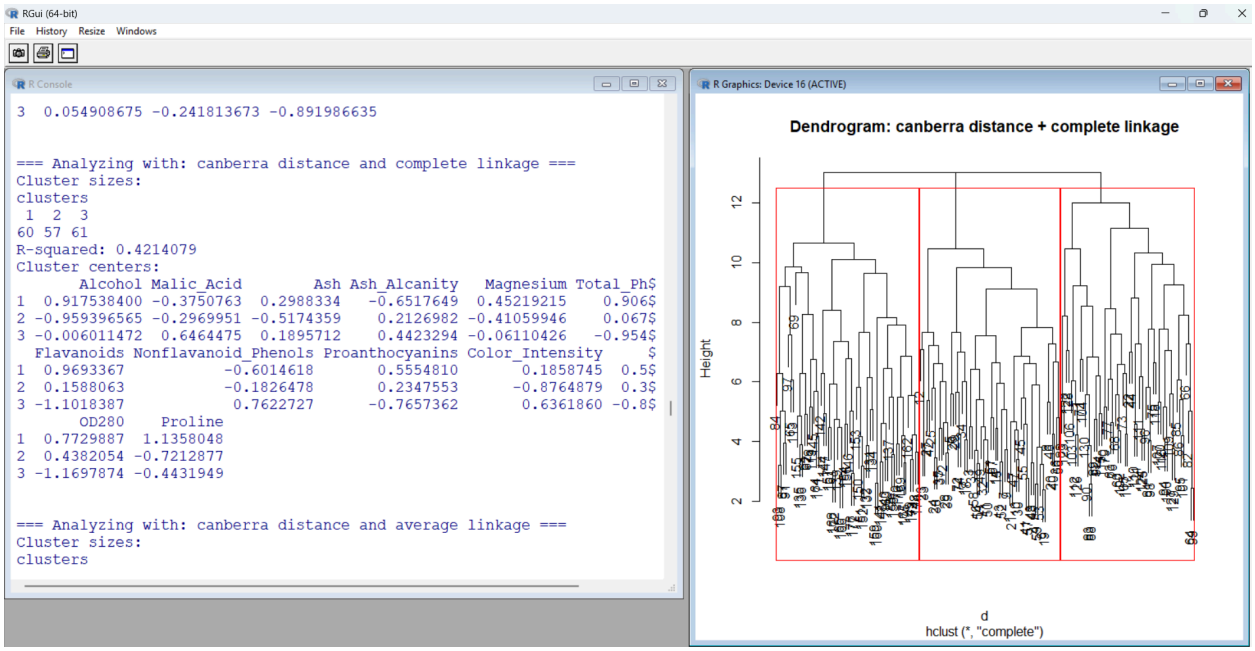
m) Manhattan + Mquitty Linkage



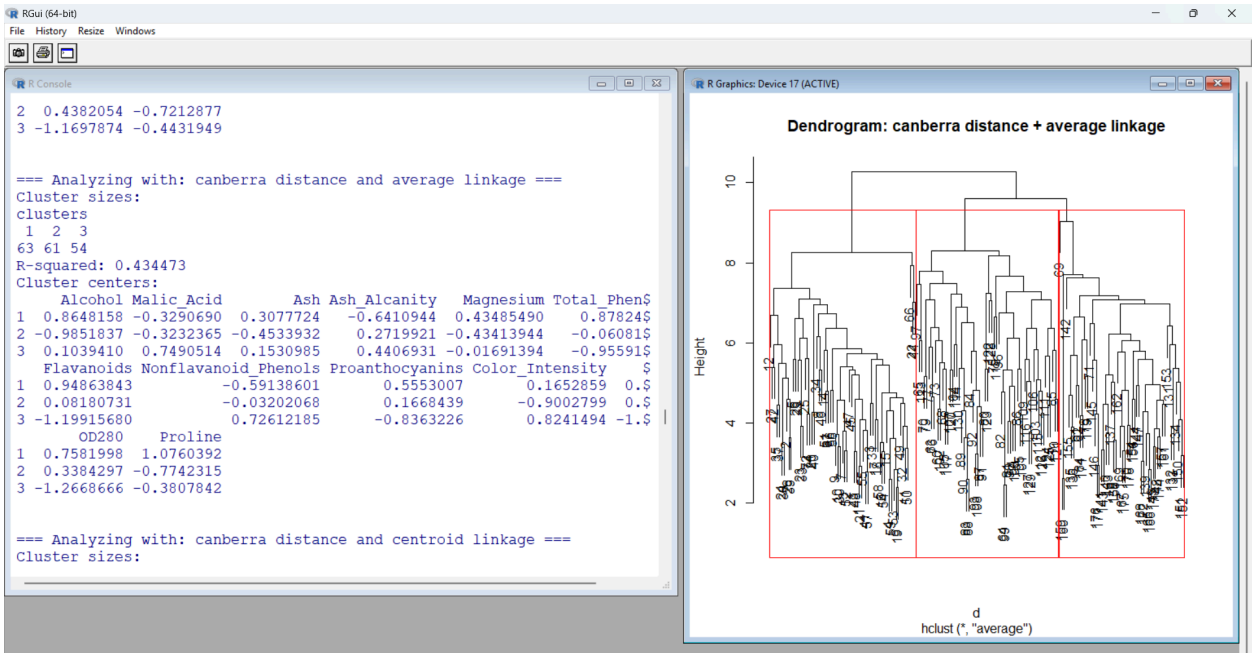
n) Canberra + Single Linkage



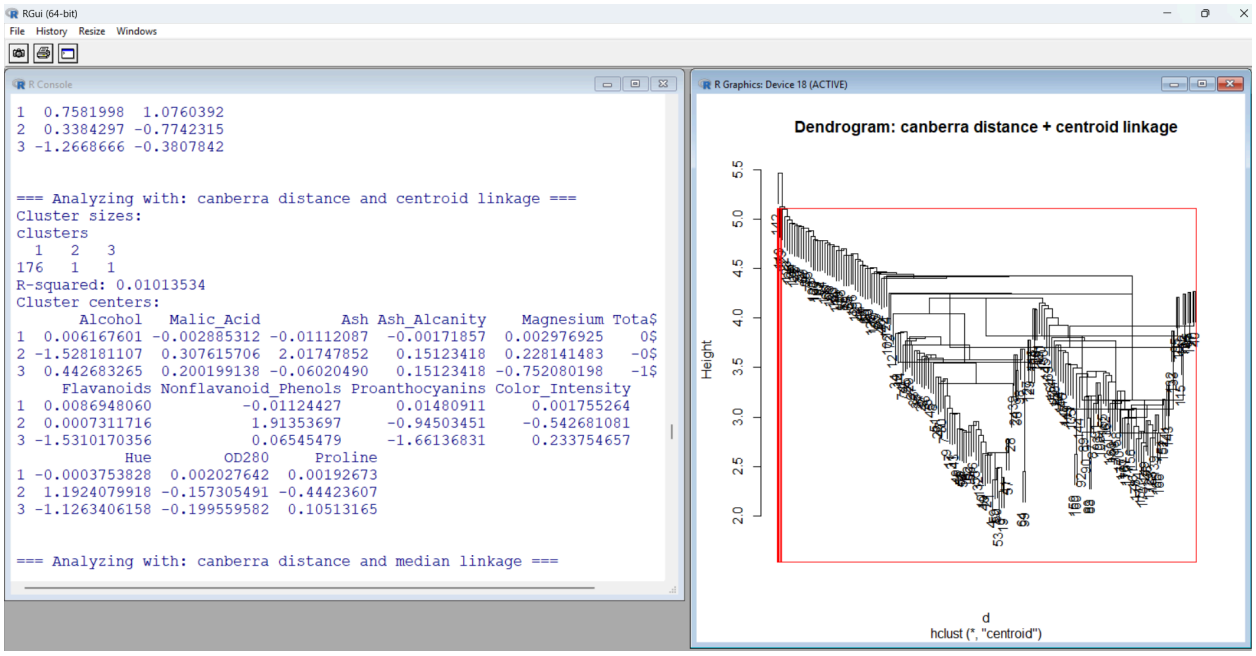
o) Canberra + Complete Linkage



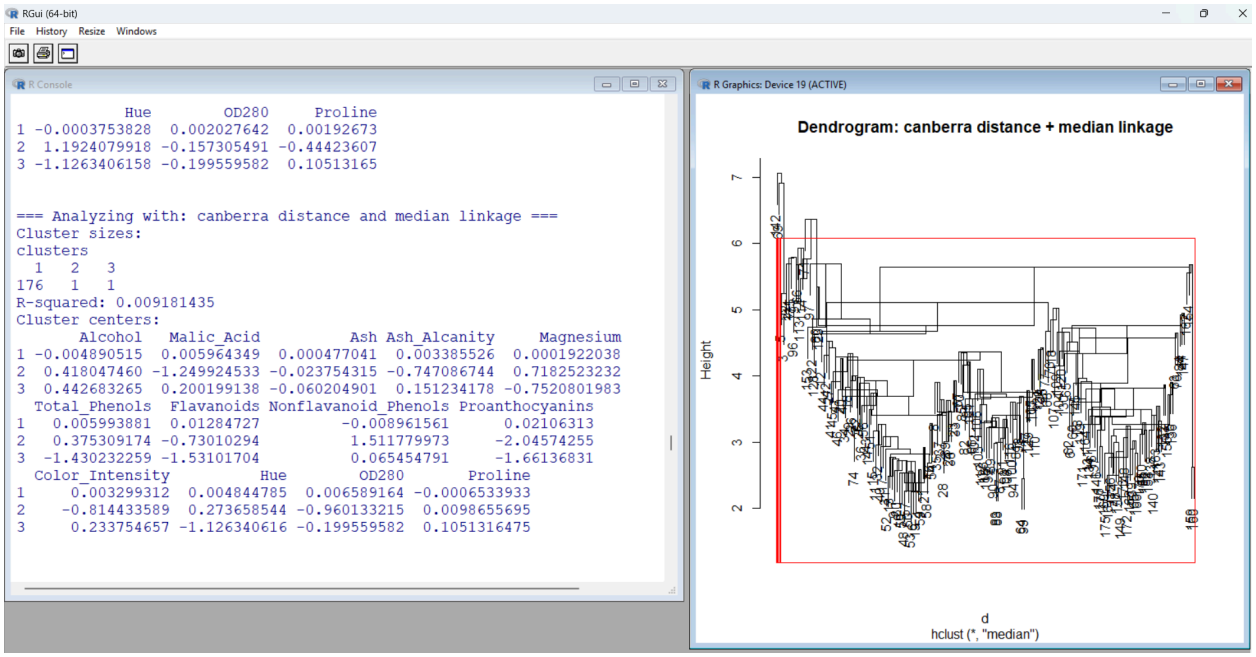
p) Canberra + Average Linkage



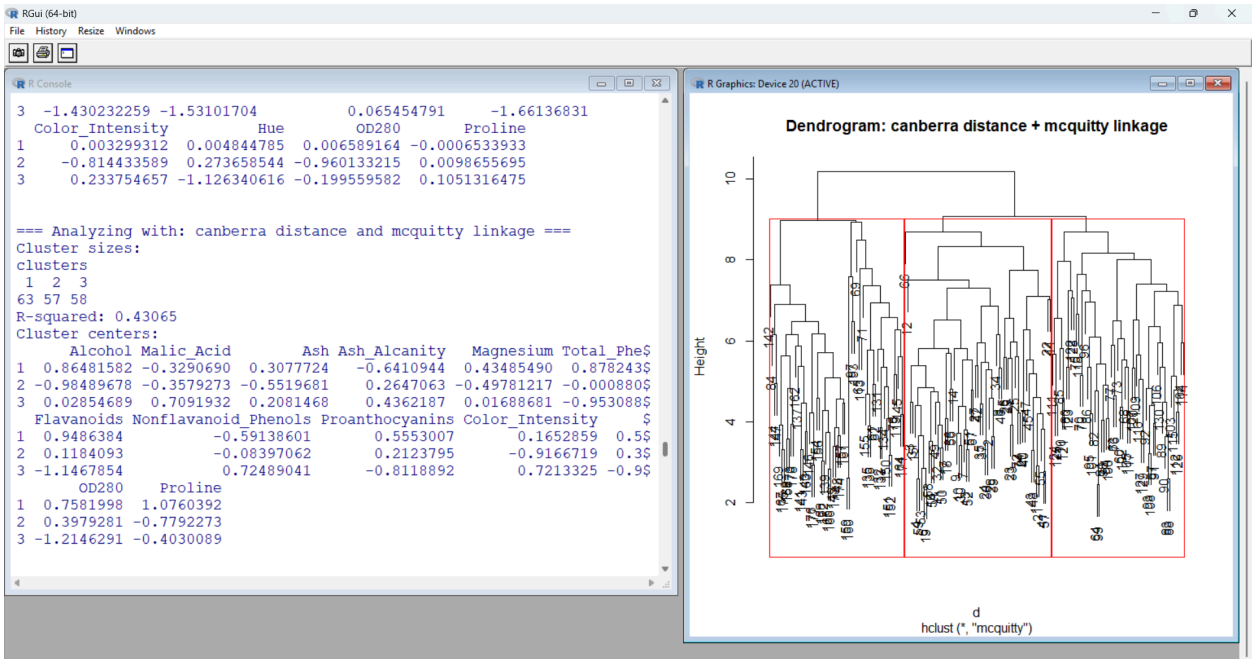
q) Canberra + Centroid Linkage



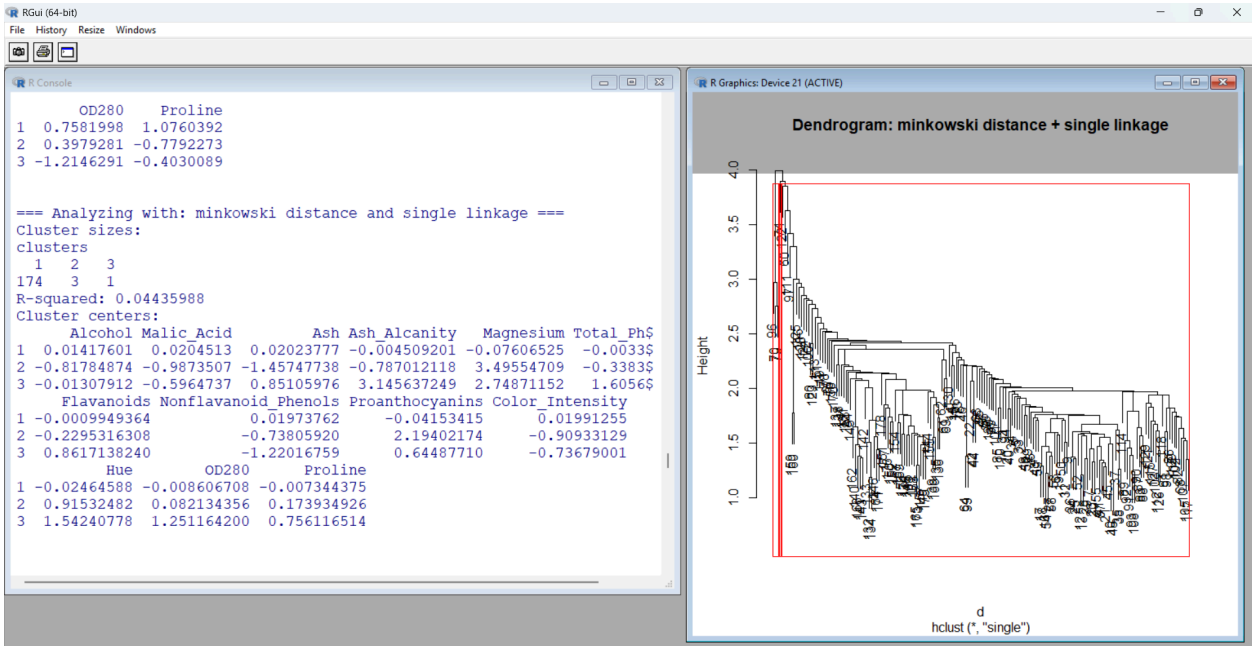
r) Canberra + Median Linkage



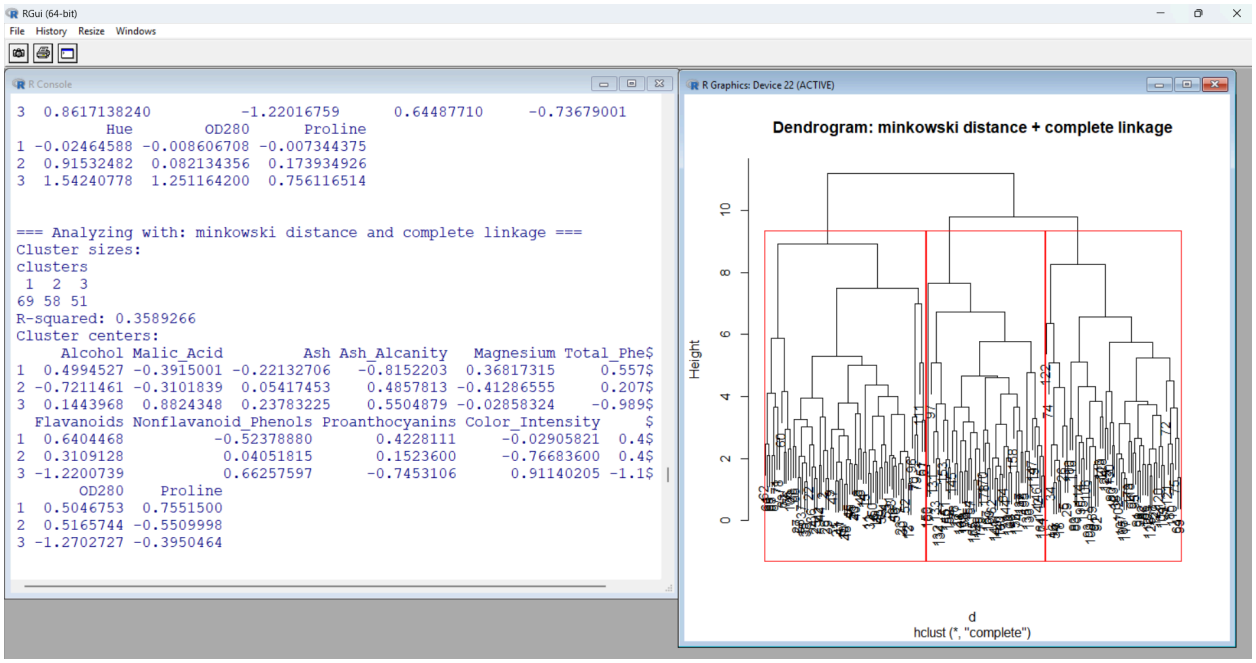
s) Canberra + Mquitty



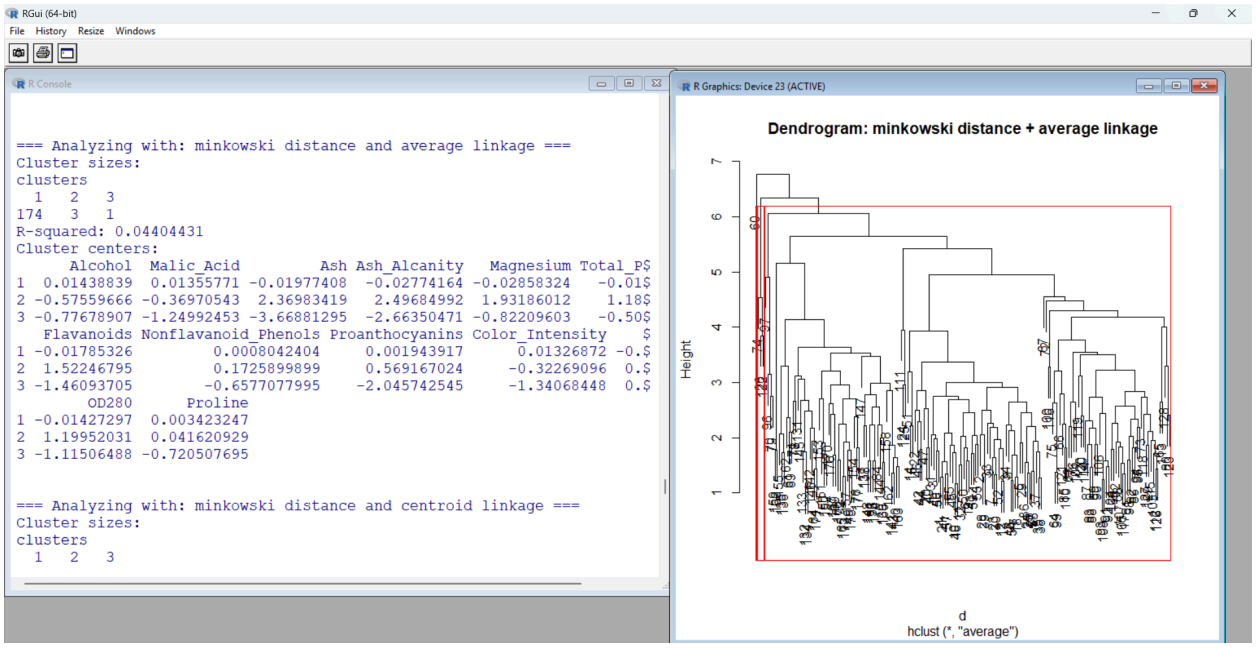
t) Minkowski + Single Linkage



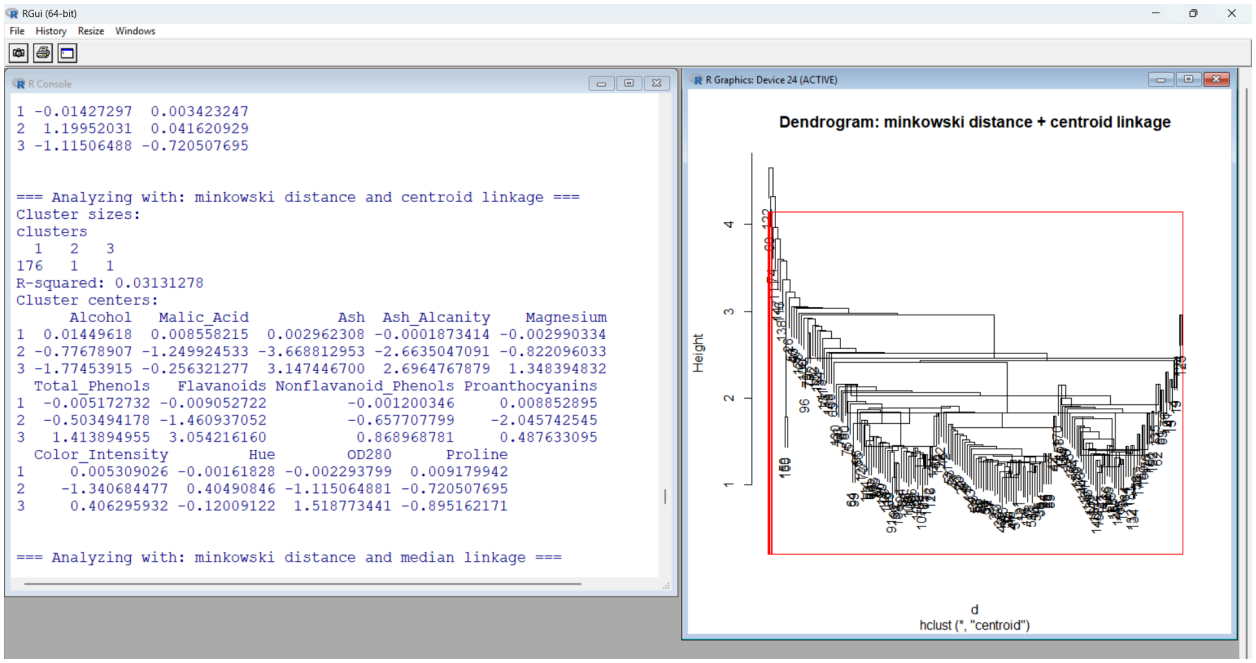
u) Minkowski + Complete Linkage



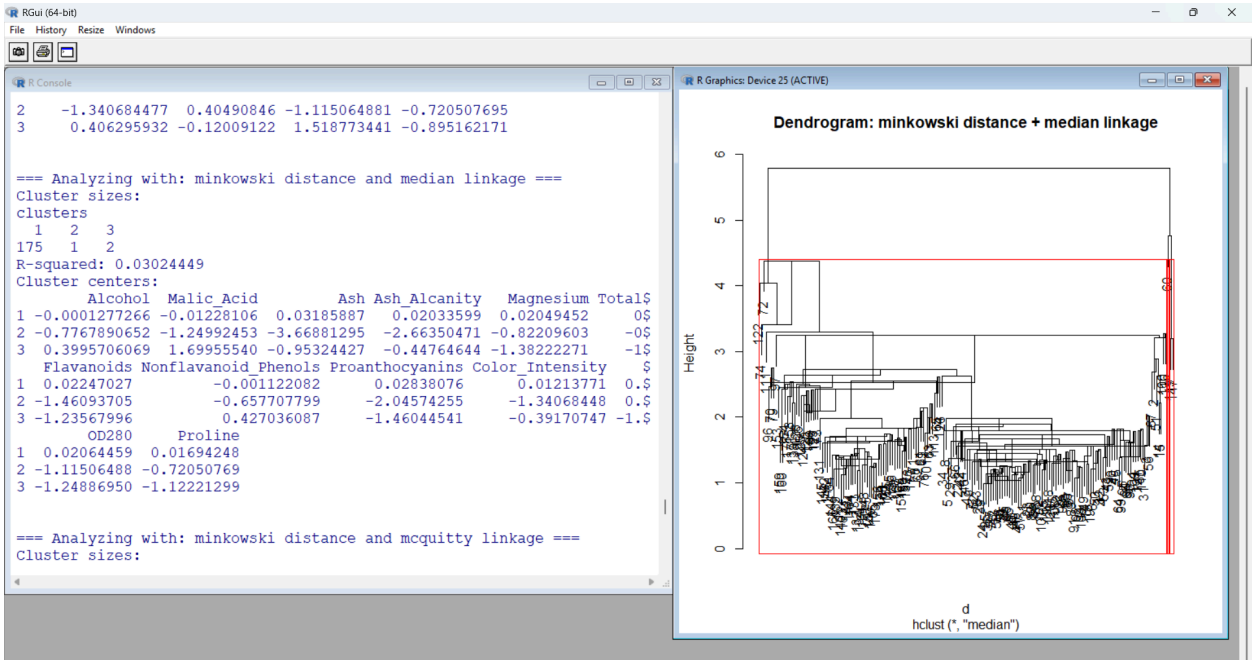
v) Minkowski + Average Linkage



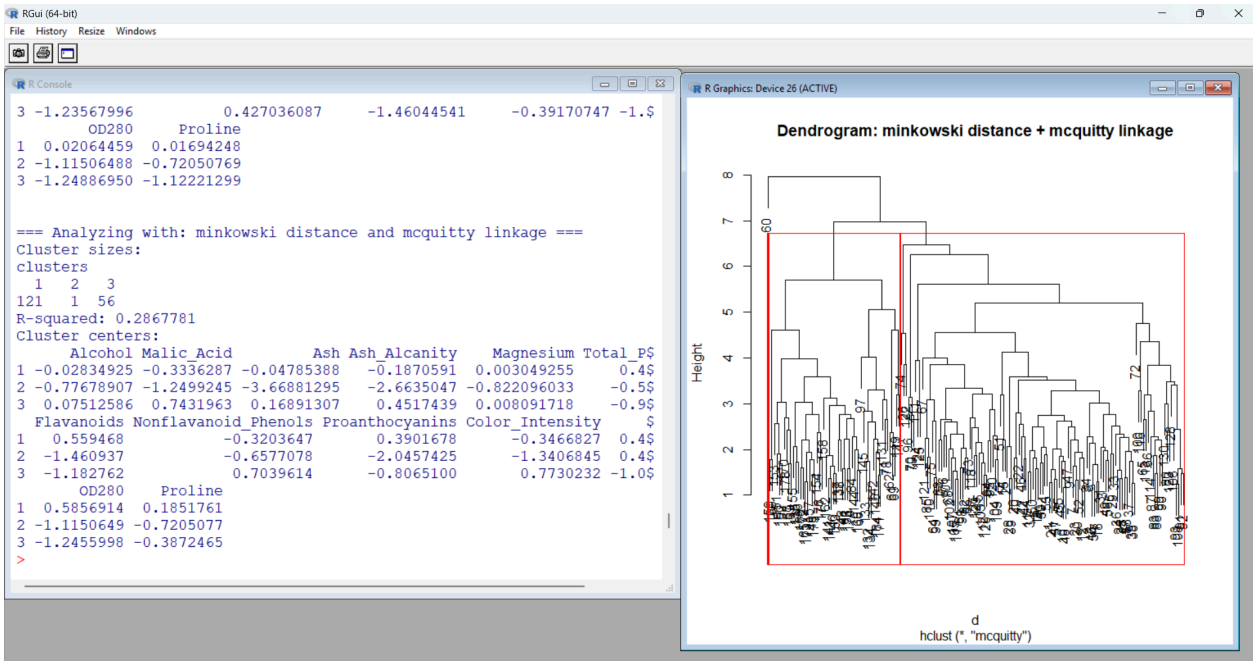
w) Minkowski + Centroid Linkage



x) Minkowski + Median Linkage



y) Minkowski + Mquitty



Best Combination

Now, after examining all these combinations, the best one, based on the calculated R^2 = 0.436, is the first one, which was between Euclidean distance as distance between observations and Ward's method as distance between clusters.

```
=== BEST PERFORMING COMBINATION ===> cat("\nMethod:", best_method)

Method: euclidean_ward.D2> cat("\nR-squared:", results[[best_method]]$$

R-squared: 0.4360204> cat("\nCluster sizes:\n")

Cluster sizes:
> print(results[[best_method]]$cluster_sizes)
clusters
 1  2  3
64 58 56
```

Comparison between K-means & Best Combination

As illustrated in the following code output, K-means and the best combination of the hierarchical method gave quite the same results.

```
=== CLUSTERING PERFORMANCE COMPARISON (k=3) ===
> cat("Best hierarchical R²:", hc_r2, "(method:", best_method, ")\n")
Best hierarchical R²: 0.4360204 (method: euclidean_ward.D2 )
> cat("K-means R²:", km_r2, "\n")
K-means R²: 0.4477405
>
> cat("\n=== CLUSTER SIZE COMPARISON ===\n")

=== CLUSTER SIZE COMPARISON ===
> cat("Hierarchical cluster sizes:\n")
Hierarchical cluster sizes:
> print(hc_best$cluster_sizes)
clusters
 1  2  3
64 58 56
> cat("\nK-means cluster sizes:\n")

K-means cluster sizes:
> print(table(kmc$cluster))

 1  2  3
62 51 65
>
```