Mouhammad Bazzi
A20522180

Assignment 5

CSP 571—Data Preparation and Analysis

I. Recitation Exercises

1. Chapter 12

Exercise 1:

a) We want to prove that:

$$\frac{1}{|C_k|}\sum_{i,i' \in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i \in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

Let's decompose the left part:

$$\sum_{i,i' \in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = \sum_{i,i' \in C_k}\sum_{j=1}^{p}(x_{ij}^2 - 2x_{ij}x_{i'j} + x_{i'j}^2) = \sum_{i \in C_k}\sum_{i' \in C_k}\sum_{j=1}^{p}(x_{ij}^2 - 2x_{ij}x_{i'j} + x_{i'j}^2)$$

$$\sum_{i \in C_k}\sum_{i' \in C_k}\sum_{j=1}^{p}(x_{ij}^2 - 2x_{ij}x_{i'j} + x_{i'j}^2) = |C_k|\sum_{i \in C_k}\sum_{j=1}^{p}x_{ij}^2 - 2\sum_{i \in C_k}\sum_{i' \in C_k}\sum_{j=1}^{p}x_{ij}x_{i'j} + |C_k|\sum_{i \in C_k}\sum_{j=1}^{p}x_{ij}^2$$

Let's now focus on the middle term:

$$-2\sum_{i \in C_k}\sum_{i' \in C_k}\sum_{j=1}^{p}x_{ij}x_{i'j} = -2\sum_{j=1}^{p}\sum_{i \in C_k}x_{ij}(\sum_{i' \in C_k}x_{i'j})$$

But we also know by definition that:

$$C_k : \bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i \in C_k}x_{ij}$$

So therefore, this part in the middle term can be rewritten like this:

$$\sum_{i \in C_k}x_{ij}(\sum_{i' \in C_k}x_{i'j}) = \sum_{i \in C_k}x_{ij}(|C_k|\bar{x}_{kj})$$

So finally, the left part could be rewritten as this:

$$2|C_k|\sum_{i \in C_k}\sum_{j=1}^{p}x_{ij}^2 - 2|C_k|^2\sum_{j=1}^{p}\bar{x}_{kj}^2$$

We don't forget to divide by |C_k| and we got this:

$$2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - 2|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2$$

Now let's rewrite the left term:

$$-2|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2 = -4|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2 + 2|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2 = -4\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}\bar{x}_{kj} + 2\sum_{i\in C_k}\sum_{j=1}^{p}\bar{x}_{kj}^2$$

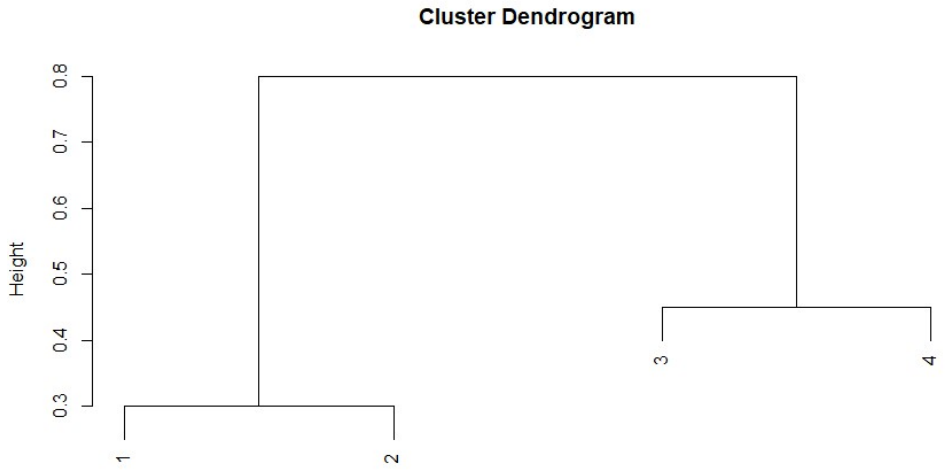
So, by replacing with the rewritten left term and simplifying we got what we were looking for:

$$2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - 4\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}\bar{x}_{kj} + 2\sum_{i\in C_k}\sum_{j=1}^{p}\bar{x}_{kj}^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij} - \bar{x}_{kj})^2$$

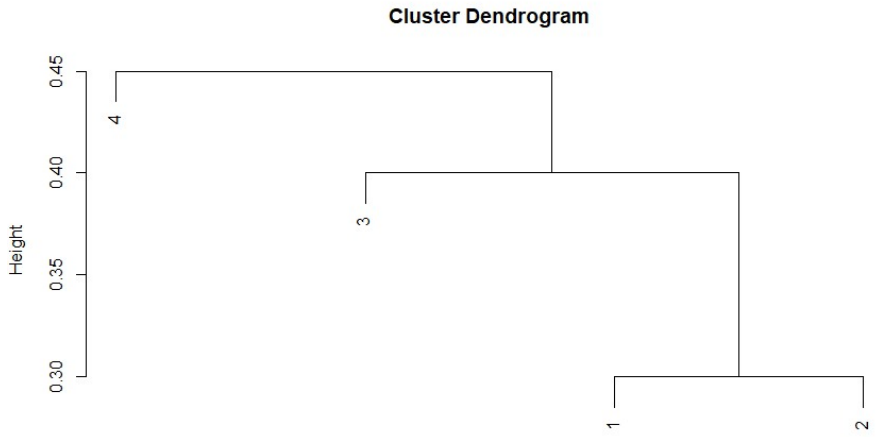
b) Algorithm 12.2 decrease the objective 12.17 at each iteration because of the previous identity. In k-means we minimize the sum of the squared Euclidian distance for each cluster, and our objective is the minimize the within-cluster variance for each cluster and both are the same according to the previous identity.

Exercise 2:

a)



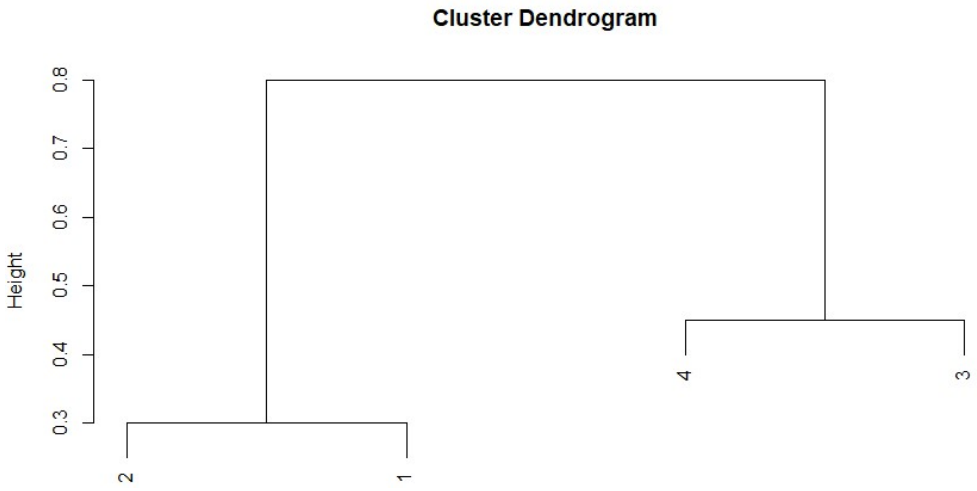
b)



c) Observations **1 and 2** will be in the same cluster and observation **3 and 4** will be in another cluster together.

d) Observations **1, 2 and 2** will be in the same cluster and observation **4** will be in another cluster alone.

e) (2 and 1 are swapped & 4 and 3 are swapped too)

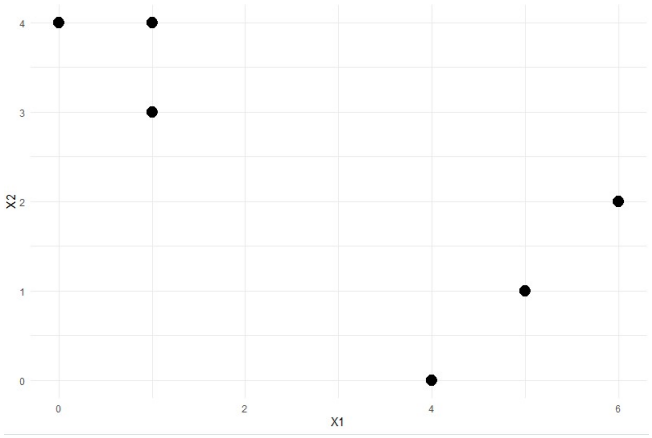


Mouhammad Bazzi
A20522180

Assignment 5

Exercise 3:

a)



b) Here the random clusters assigned to each point:

```
> data$Cluster
[1] 1 1 1 1 2 2
```

c) Here the centroids of each cluster:

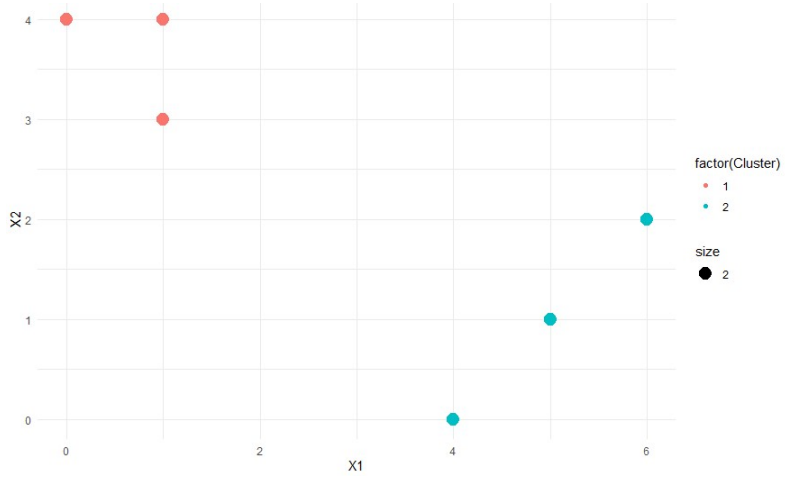
```
> centroids
  Cluster   X1 X2
1       1 1.75  3
2       2 5.00  1
```

d) After this iteration the 3 points near the left corner are assigned same cluster and the one near the right corner are assigned to another cluster.

```
> data$Cluster
[1] 1 1 1 2 2 2
```

e) At iteration 2 I obtained the same cluster, so I stopped the algorithm and moved to question **f.**

f)



Mouhammad Bazzi
A20522180

Exercise 4:

a) In single linkage clustering, the distance between two clusters is defined as the minimum distance between any two observations belonging to different clusters. In complete linkage clustering, the distance between two clusters is defined as the maximum distance between any two observations belonging to different clusters. Therefore, if the maximum and minimum values are equal, they will fuse at the same height, if not, the single linkage will fuse at a lower height. **But without knowing this information, we can't conclude so we can only conclude that we need more information**.

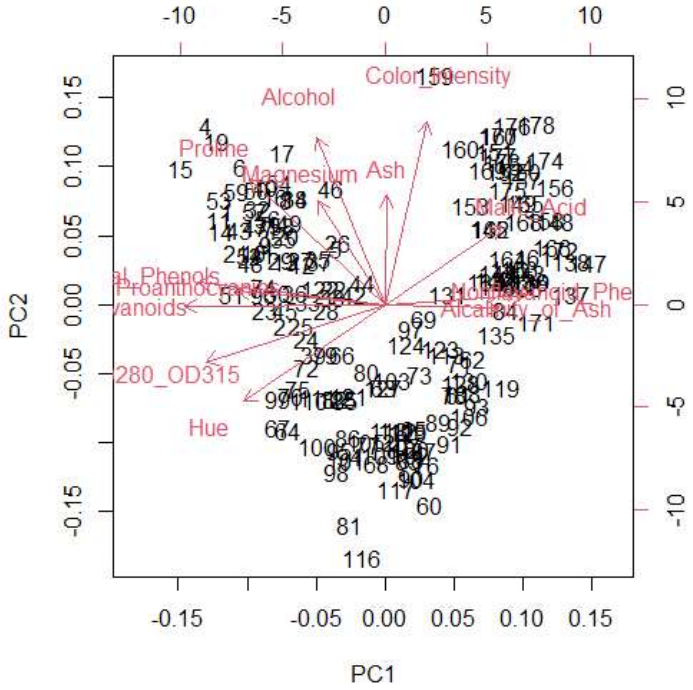
b) Since we are fusing singleton clusters (in that example her {5} and {6}) in both single linkage and complete linkage, the minimum distance between the observations (single linkage) is equal to the maximum distance (complete linkage). As a result, the fusion of the 2 clusters (here {5} and {6}) will necessarily occur at the **same height in both dendrograms**.

Mouhammad Bazzi **Assignment 5**
A20522180

II. Practicum Problems

Problem 1:

In PCA scaling should be used on the inputs because it allows us to give equal importance to each feature.

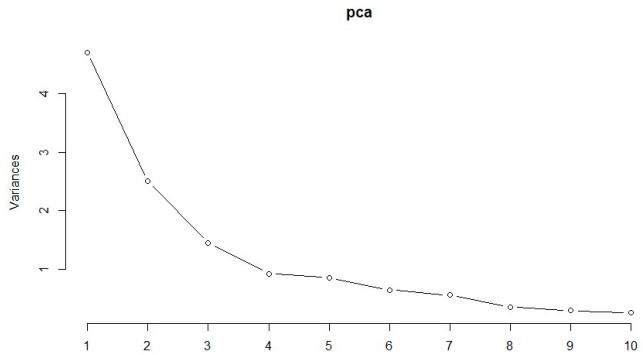


From the biplot, it appears that **'Malic_Acid'** is pointed in the opposite direction of **'Hue'**. And that means that the two features are **negatively corelated**, when one tends to increase the other one tends to decrease.

This can be confirmed by computing the correlation between both features, it should be negative and significantly far from 0:

```
> cat("Correlation between Malic_Acid and Hue:", correlation, "\n")
Correlation between Malic_Acid and Hue: -0.5612957
```

That's the scree plot:



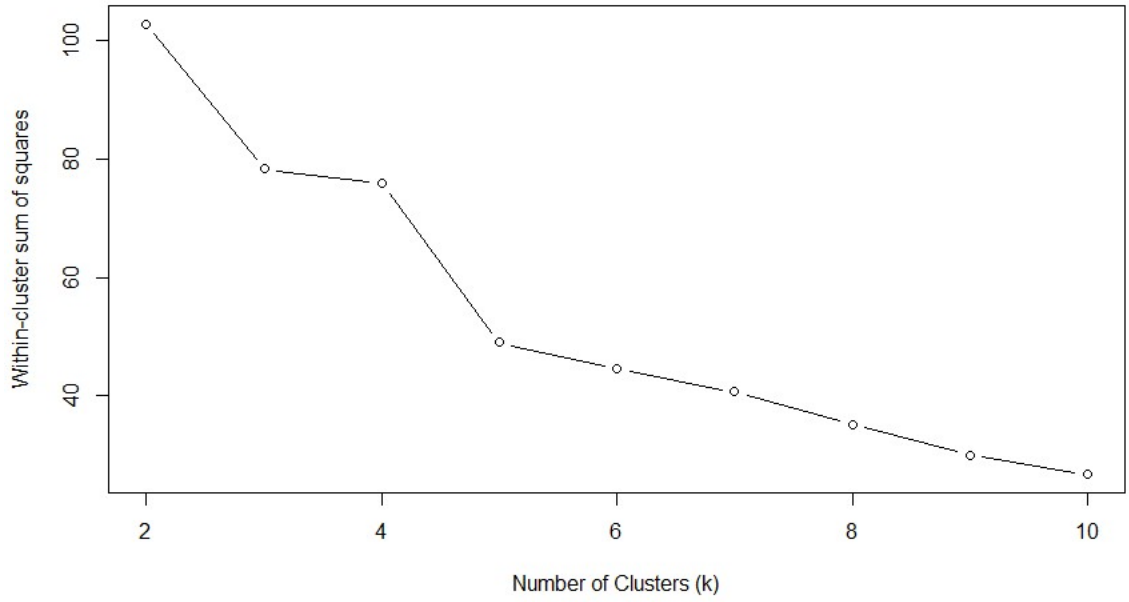
And here the percentage of total variance explained by PC1 and PC2

```
> cat("Percentage of total variance explained by PC1:", var
Percentage of total variance explained by PC1: 36.19885 %
> cat("Percentage of total variance explained by PC2:", var
Percentage of total variance explained by PC2: 19.20749 %
```

Problem 2:

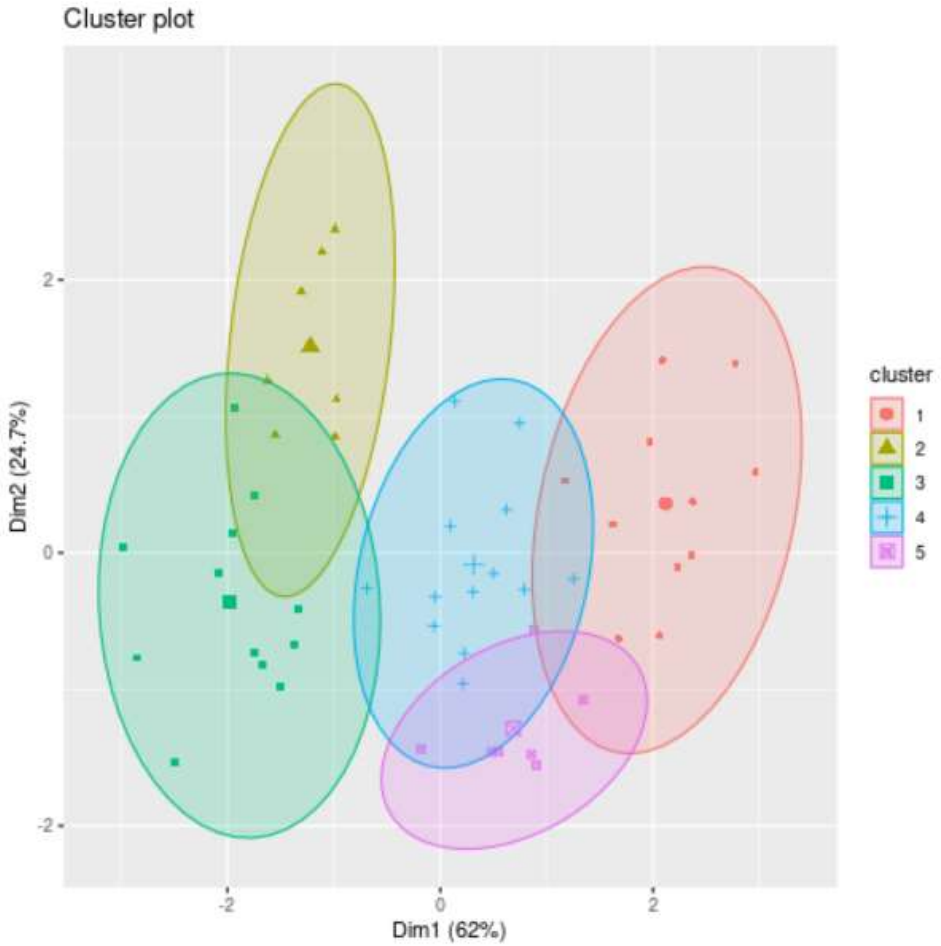
I decided to center/scale the observations because the variables in the dataset have different units and we should scale them to make them comparable to perform our k-means algorithm well. (all the variables should have the same importance regardless of their units).

Here is the plot of the within-cluster sum of squares for each value of k:



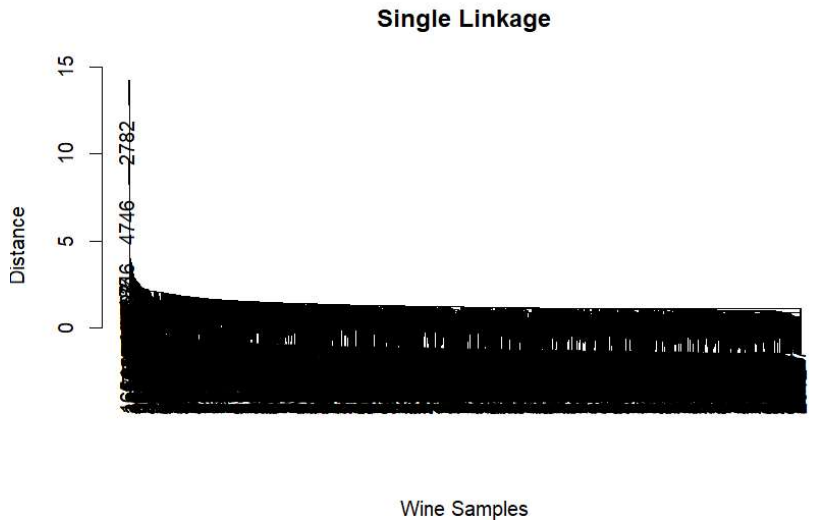
And we can see that the elbow is at a number of clusters equal to 5. So, it seems that the **optimal number of clusters is equal to 5**.

And here is the **optimal clustering**:

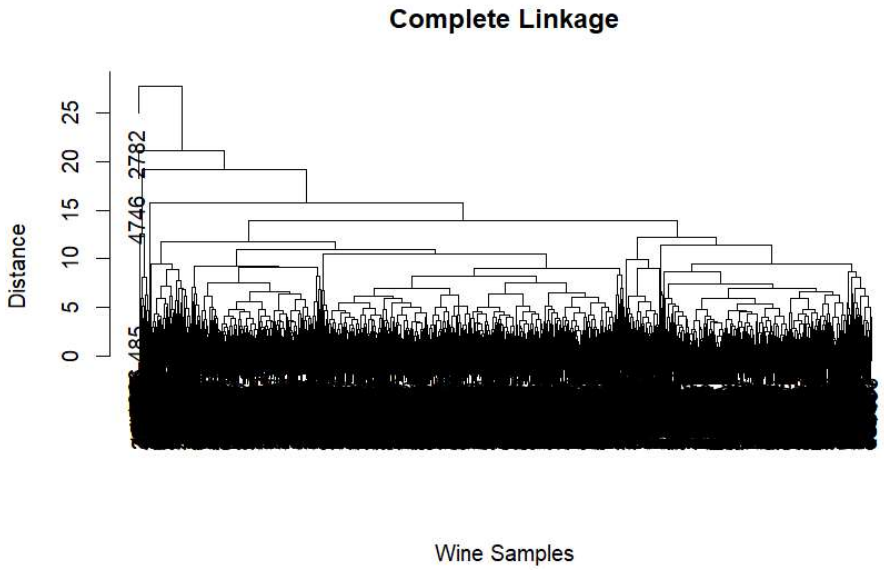


Problem 3:

Here is the single linkage plot of the hierarchical clustering:



And here is the complete linkage plot of the hierarchical clustering:



I decided to center/scale the observations because the variables in the dataset have different units and scales and we should scale. (all the variables should have the same importance regardless of their units).

Here is the distance value of the two penultimate clusters merged for both hierarchical clustering:

```
> cat("Single linkage penultimate distance:", single_penultimate, "\n")
Single linkage penultimate distance: 9.706043
> cat("Complete linkage penultimate distance:", complete_penultimate, "\n")
Complete linkage penultimate distance: 21.08566
```

Here are the summary statistics of both clusters and we can see that they are the same:

```
Summary statistics for single linkage clustering:
> print(wine_data_single)
# A tibble: 2 × 25
  cluster `fixed acidity_fn1` `volatile acidity_fn1` `citric acid_fn1` `residual sugar_fn1`
    <int>               <dbl>                  <dbl>             <dbl>                 <dbl>
1       1                6.85                  0.278             0.334                  6.38
2       2                7.8                   0.965             0.6                   65.8
Summary statistics for complete linkage clustering:
> print(wine_data_complete)
# A tibble: 2 × 25
  cluster `fixed acidity_fn1` `volatile acidity_fn1` `citric acid_fn1` `residual sugar_fn1`
    <int>               <dbl>                  <dbl>             <dbl>                 <dbl>
1       1                6.85                  0.278             0.334                  6.38
2       2                7.8                   0.965             0.6                   65.8
```

Feature means with the largest differences for both hierarchical clustering is residual **sugar_fn1 with an absolute difference equal to 59.42072**.

In that case**, complete linkage and single linkage are produced as balanced clustering**.