Unsupervised Clustering Analysis Wine Review Dataset

Retail Group: Jairo Melo, Vikram Khade, Ignacio Palma, Mahboob Jamil

2019-02-24

Installing packagaes:

tidyverse: data manipulation cluster: clustering algorithms stats: clustering algorithms factoextra: clustering visualization

Importing the data

```
## [1] "/Users/jairomelo/Desktop/ML/YORK/Assigment2"

## [1] 21044

## [1] 21044

## [1] 21044

## 'data.frame':    5000 obs. of  4 variables:
##  $ points   : int  95 95 94 90 90 90 90 90 90 90 ...
##  $ price    : int  80 290 57 135 29 23 69 90 50 100 ...
##  $ latitude : num  43.3 46.2 -36.8 43.8 43.8 ...
##  $ longitude: num  -4.25 2.21 174.56 11.25 11.25 ...
```

As we don't want the clustering algorithm to depend to an arbitrary variable unit, we start by scaling/standardizing the data using the R function scale:

```
##  num [1:5000, 1:4] 2.094 2.094 1.791 0.578 0.578 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:5000] "1" "2" "3" "4" ...
##   ..$ : chr [1:4] "points" "price" "latitude" "longitude"
##  - attr(*, "scaled:center")= Named num [1:4] 88.09 39.01 13.58 1.34
##   ..- attr(*, "names")= chr [1:4] "points" "price" "latitude" "longitude"
##  - attr(*, "scaled:scale")= Named num [1:4] 3.3 54.7 39.1 64.9
##   ..- attr(*, "names")= chr [1:4] "points" "price" "latitude" "longitude"
```

Agglomerative Hierarchical clustering

This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

First, we will find the dissimilatiry values and then use the distance matrix to run the Hierarchical clustering to plot the dendogram:

The agglomeration method that can be used is (an unambiguous abbreviation of) one of “ward.D”, “ward.D2”, “single”, “complete”, “average” (= UPGMA), “mcquitty” (= WPGMA), “median” (= WPGMC) or “centroid” (= UPGMC).

For our analysis, we will use Ward.D2:

Cutting the Tree

Each leaf of the Dendrogram corresponds to one observation. As we move up the tree, observations that are similar to each other are combined into branches, which are themselves fused at a higher height.

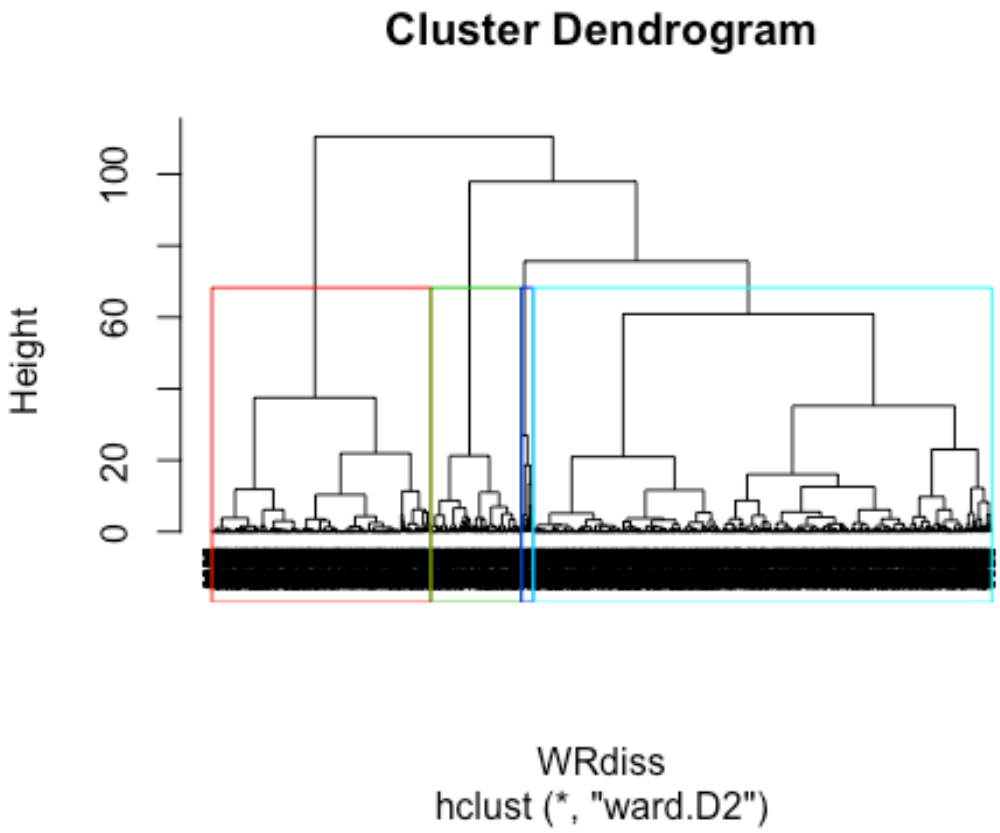
The height of the vertical line or vertical axis, indicates the (dis)similarity between two observations. The higher the height of the vertical line/fusion, the less similar the observations are.

Note: conclusions about the proximity of two observations can only be based on the height where branches containing those two observations first are fused. We cannot use the proximity of two observations along the horizontal axis as a criteria of their similarity.

Let’s cut the tree in 4 groups

```
## sub_grp
##    1    2    3    4
## 2940   76  580 1404
```

Drawing the dendrogram with a border around the 4 clusters



From the dendrogram we are able to identify ……

Comparing between different Agglomerative methods:

Using agnes we can calculate the Agglomerative coefficient. The agglomerative coefficient measures the amount of clustering structure found (values closer to 1 suggest strong clustering structure)

The Agglomerative coefficient allows us to find certain hierarchical clustering methods that can identify stronger clustering structures.

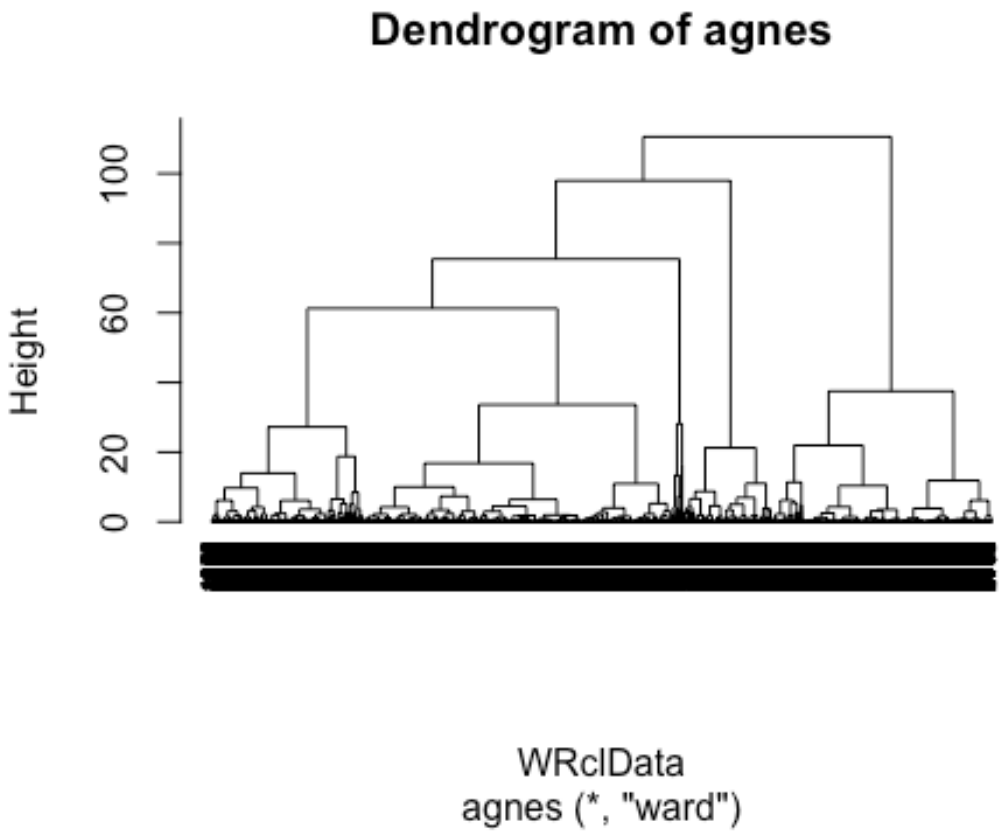
Methods to assess the coefficient

We will compare the use coefficient from average, single, complete and ward to understand the differences between each method and select one to continue with our analysis:

```
##   average    single  complete      ward
## 0.9984433 0.9971121 0.9984540 0.9996464
```

From the table, we conclude that Ward is giving the highest Coefficient; and cut the tree at 4 groups.
ward= 0.9996464

Let's run the method and plot its results:

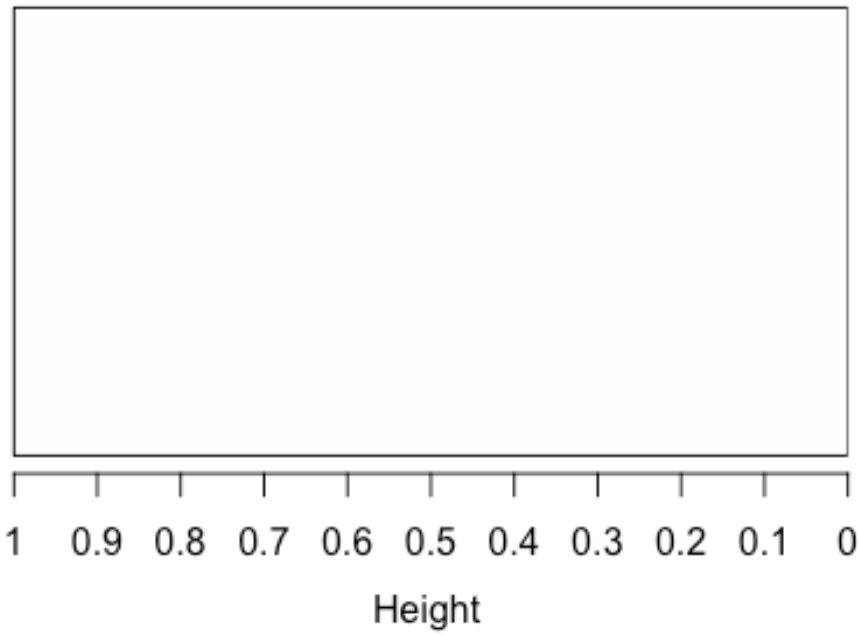


Divisive Hierarchical Clustering

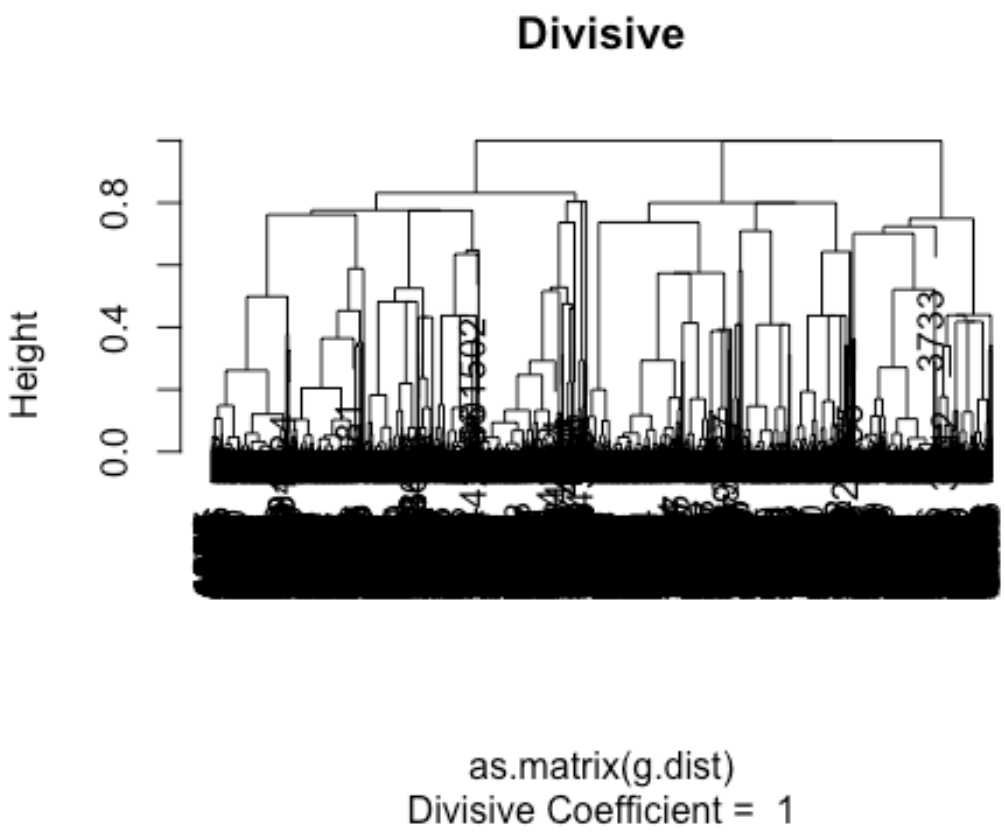
This variant of hierarchical clustering is called top-down clustering or divisive clustering . We start at the top with all documents in one cluster. The cluster is split using a flat clustering algorithm. This procedure is applied recursively until each document is in its own singleton cluster.

Let's run the Diana method and plot the results:

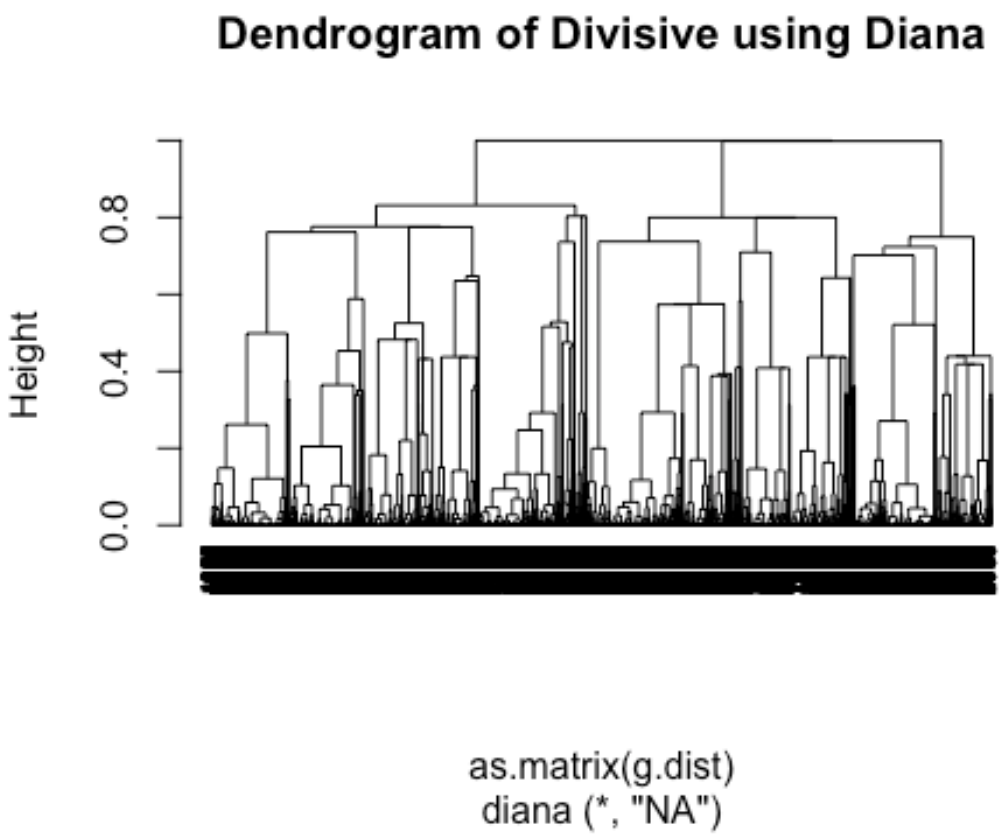
Divisive



Divisive Coefficient = 1


Divisive
Height
0.0
0.4
0.8
as.matrix(g.dist)
Divisive Coefficient = 1

```
## [1] 0.9971214
```



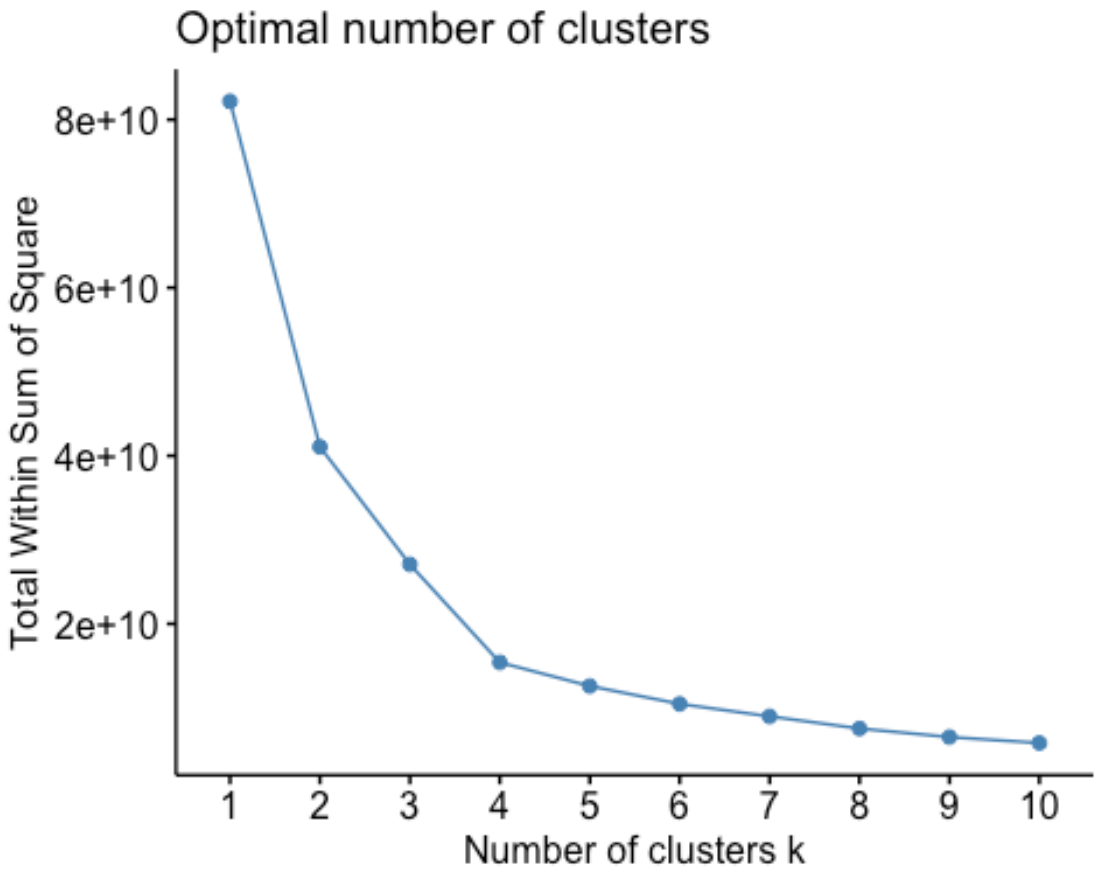
The height of the cut to the dendrogram controls the number of clusters obtained; similar to K-means, used to identified sub-groups.

Visualize

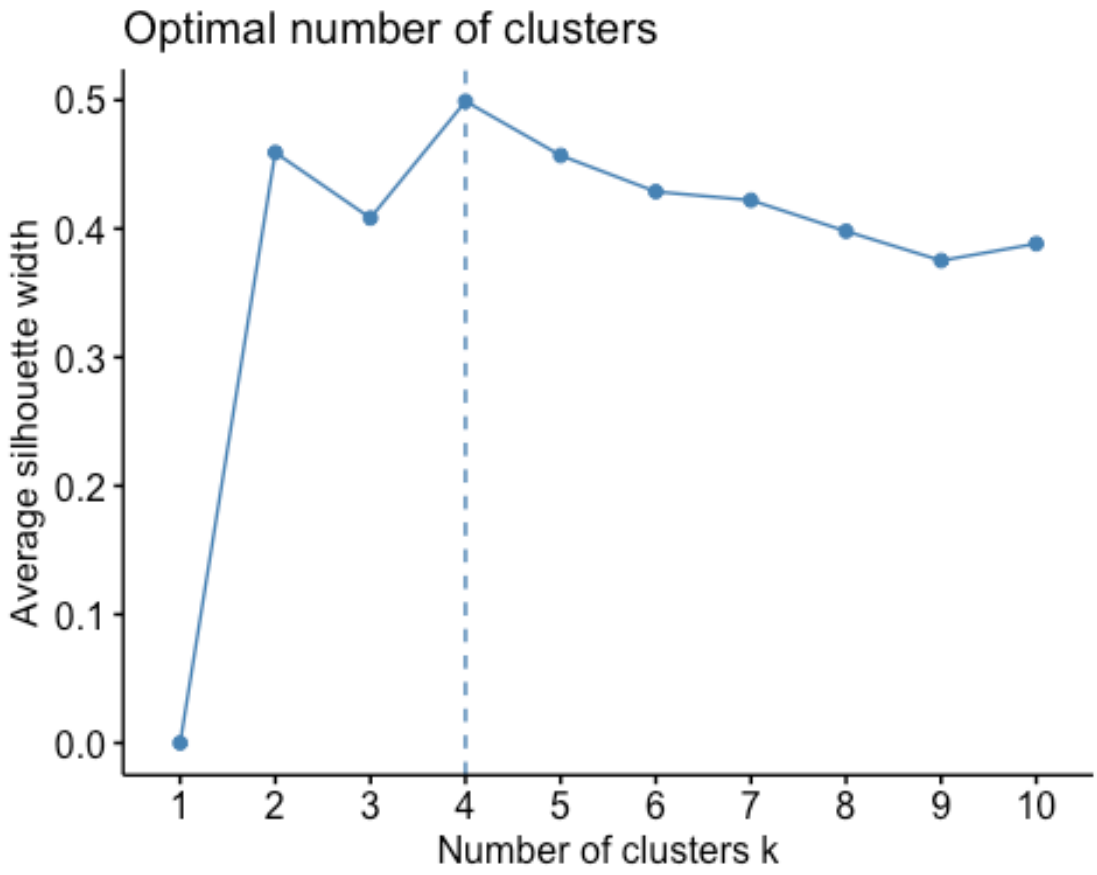
Similar to how K-means represent the cluster, we can visualize the result in a scatter plot.

Determining Optimal Clusters

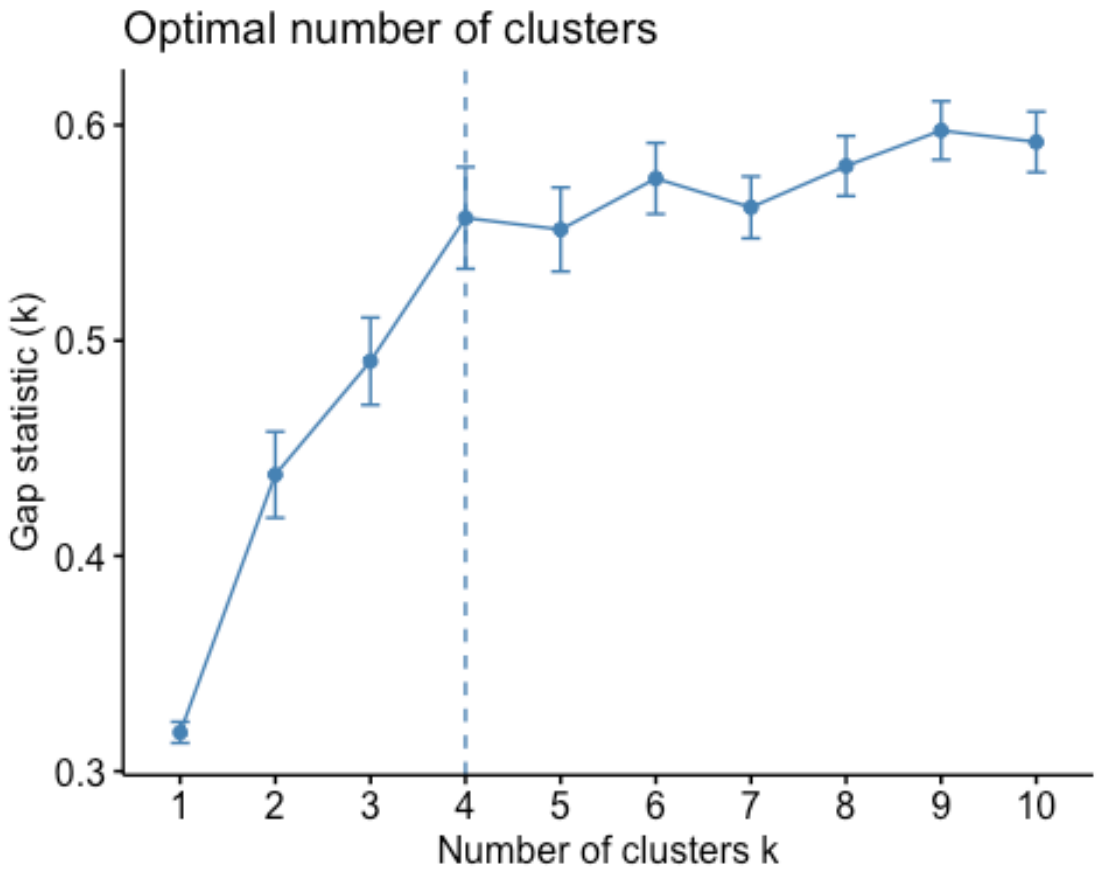
Elbow Method



Average Silhouette Method



Gap Statistic Method



Further Analysis

Let's also cutree output to add the the cluster each observation to the original data to drice further analysis against the nominal variables:

```
##       country
## 1       Spain
## 2      France
## 3 New Zealand
## 4       Italy
## 5       Italy
## 6       Italy
##
description
## 1                                    Nicely oaked blackberry, licorice,
vanilla and charred aromas are smooth and sultry. This is an outstanding wine
from an excellent year. Forward barrel-spice and mocha flavors adorn core
blackberry and raspberry fruit, while this runs long and tastes vaguely
chocolaty on the velvety finish. Enjoy this top-notch Tempranillo through
2030.
```

```
## 2 Coming from a seven-acre vineyard named after the dovecote on the
property, this is a magnificent wine. Powered by both fruit tannins and the
28 months of new wood aging, it is darkly rich and with great concentration.
As a sign of its pedigree, there is also elegance here, a restraint which is
new to this wine. That makes it a wine for long-term aging. Drink from 2022.
## 3
Yields were down in 2015, but intensity is up, giving this medium-bodied,
silky wine the potential to drink well through at least 2025. Hickory smoke
outlines white peach before ending in a long flurry of lime zest.
## 4
Forest floor, tilled soil, mature berry and a whiff of new leather combine on
this. The ripe palate offers fleshy black cherry, dried aromatic herb and
tobacco, while fine-grained tannins give the finish some grip. Drink 2018–
2023.
## 5
Aromas of forest floor, violet, red berry and a whiff of dark baking spice
unfold in the glass while wild cherry, black raspberry, ground pepper and
star anise drive the palate. Framed in firm, fine-grained tannins, this is a
classic Tuscan red. Drink through 2020.
## 6
This has a charming nose that boasts rose, violet and red berry while the
juicy, easy-drinking palate offers ripe wild cherry, chopped mint, white
pepper and a hint of star anise. There isn't much complexity but it is savory
and balanced, with fresh acidity and supple tannins.
##   points price         province     variety  latitude  longitude cluster
## 1     95    80   Northern Spain Tempranillo  43.26912  -4.250079       1
## 2     95   290 Southwest France      Malbec  46.22764   2.213749       2
## 3     94    57            Kumeu  Chardonnay -36.77726 174.558802       3
## 4     90   135          Tuscany  Sangiovese  43.77105  11.248621       4
## 5     90    29          Tuscany  Sangiovese  43.77105  11.248621       4
## 6     90    23          Tuscany  Sangiovese  43.77105  11.248621       4
```

Let's plot the categorical variables by cluster

Cluster sizes

```
## integer(0)

## character(0)

## character(0)

## character(0)

## character(0)
```

Conclusion

Clustering is a very useful tool for data analysis in the unsupervised setting. Bellow are some items we must consider while performing hierarchical clustering analysis.

1. What dissimilarity measure should be used?
2. What type of linkage should be used?
3. Where should we cut the dendrogram in order to obtain clusters?

Each of these decisions has a strong impact on the results obtained.

We should try several different choices, and look for the one with the most useful or interpretable solution. With these methods, there is no single right answer - any solution that exposes some interesting aspects of the data should be considered.