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.

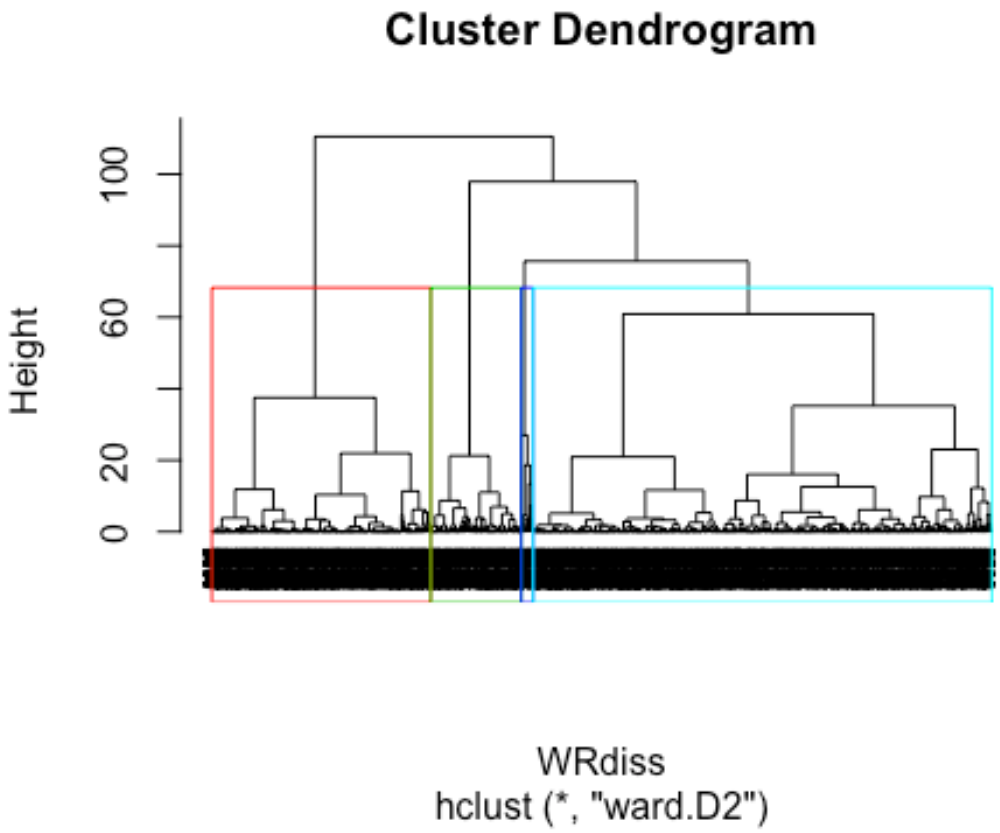
The height of the vertical line or vertical axis, indicates the (dis)similarity between two observations. The higher the height of the vertical line/fusion, the less similar the observations are.

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 dendogram 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.

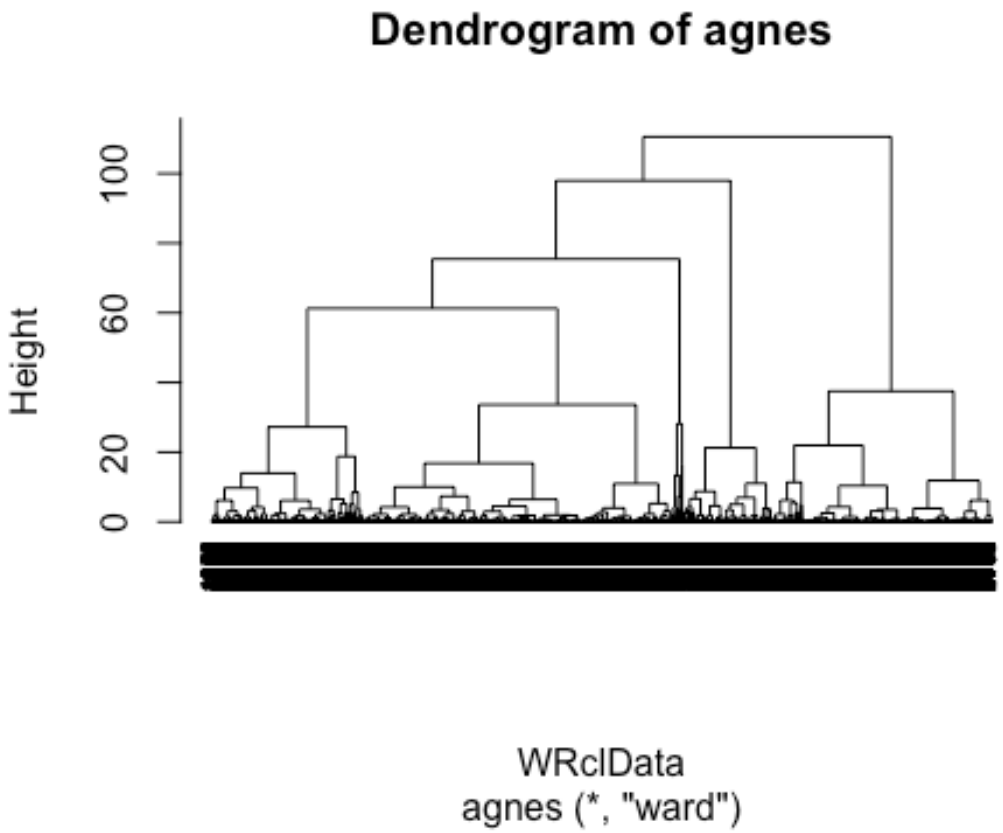
Methods to assess the coefficient

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:

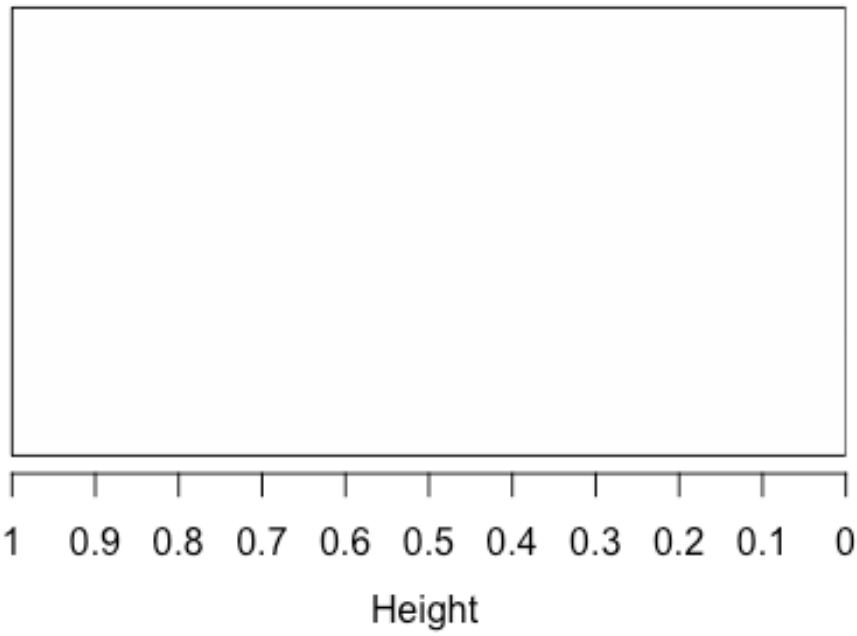


Divisive Hierarchical Clustering

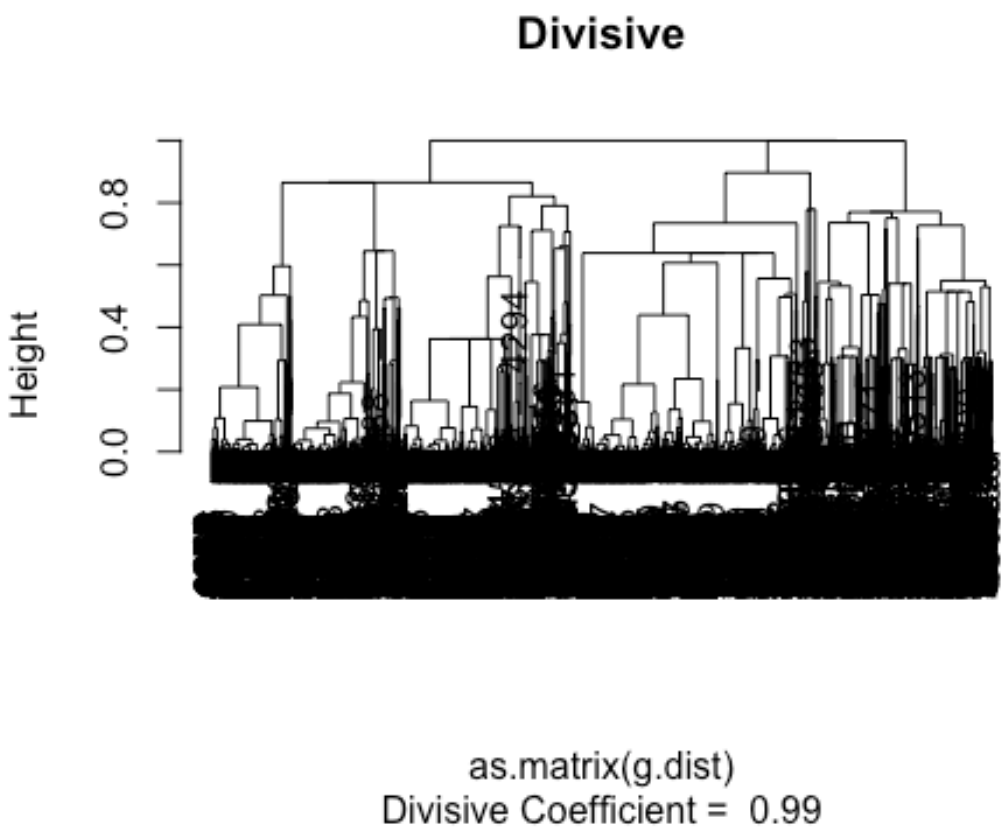
This variant of hierarchical clustering is called top-down clustering or divisive clustering . We start at the top with all documents in one cluster. The cluster is split using a flat clustering algorithm. This procedure is applied recursively until each document is in its own singleton cluster.

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

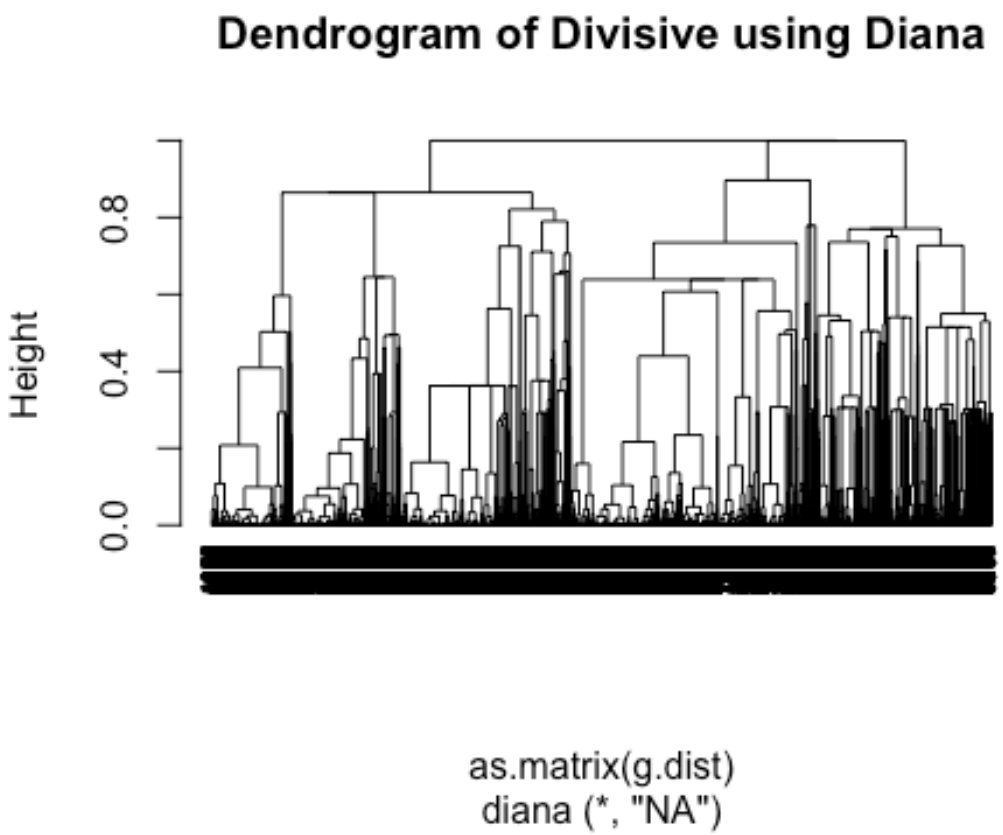
Divisive



Divisive Coefficient = 0.99


Divisive
Height
0.0
0.4
0.8
as.matrix(g.dist)
Divisive Coefficient = 0.99

```
## [1] 0.9909801
```

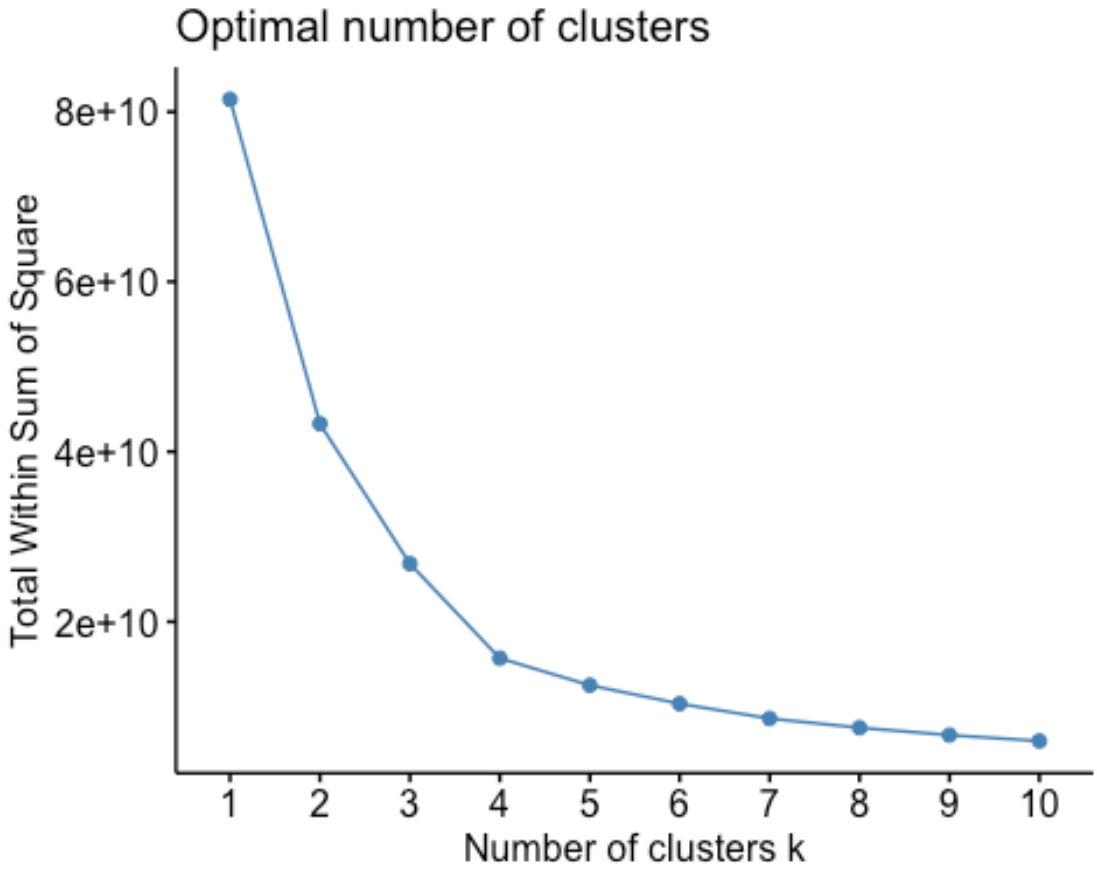


The height of the cut to the dendrogram controls the number of clusters obtained; similar to K-means, used to identified sub-groups.

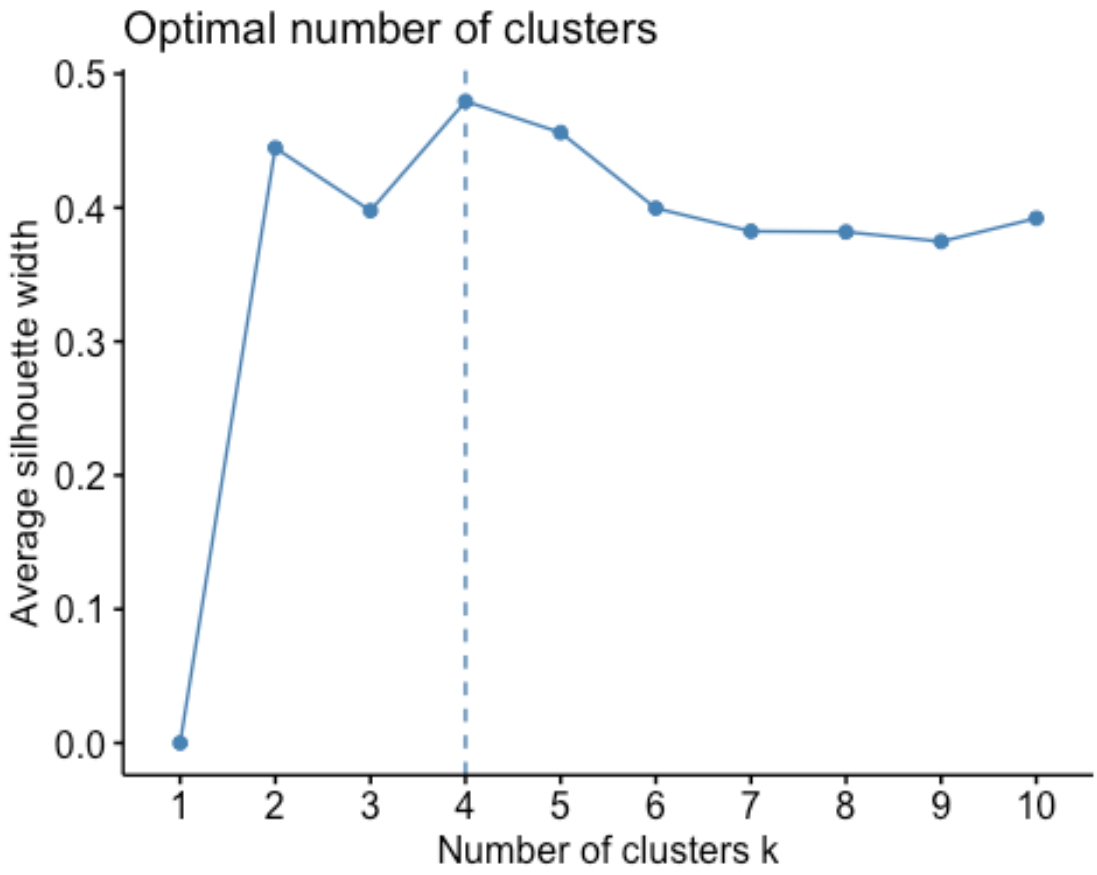

Cluster plot
Dim2 (16.6%)
40
0
-40
-50
-25
0
25
50
Dim1 (26.7%)
cluster
1
2
3
4

Determining Optimal Clusters

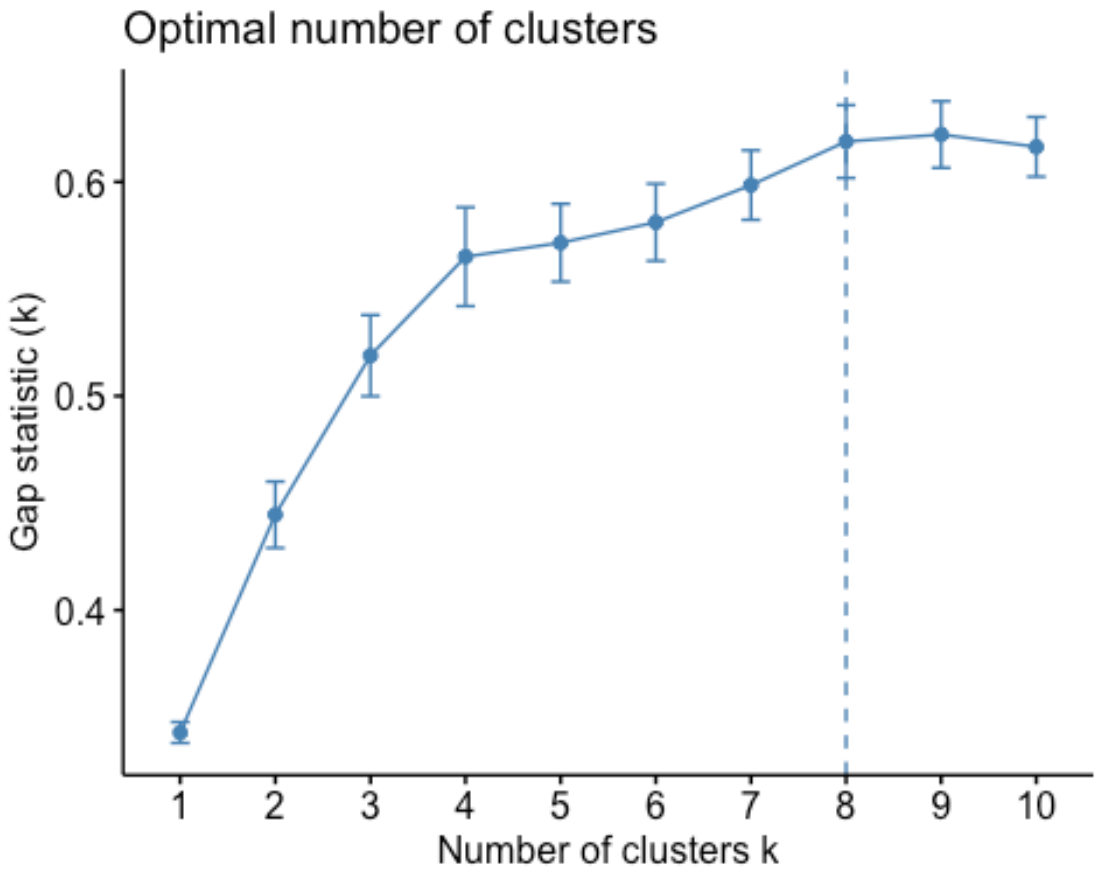
Elbow Method



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:

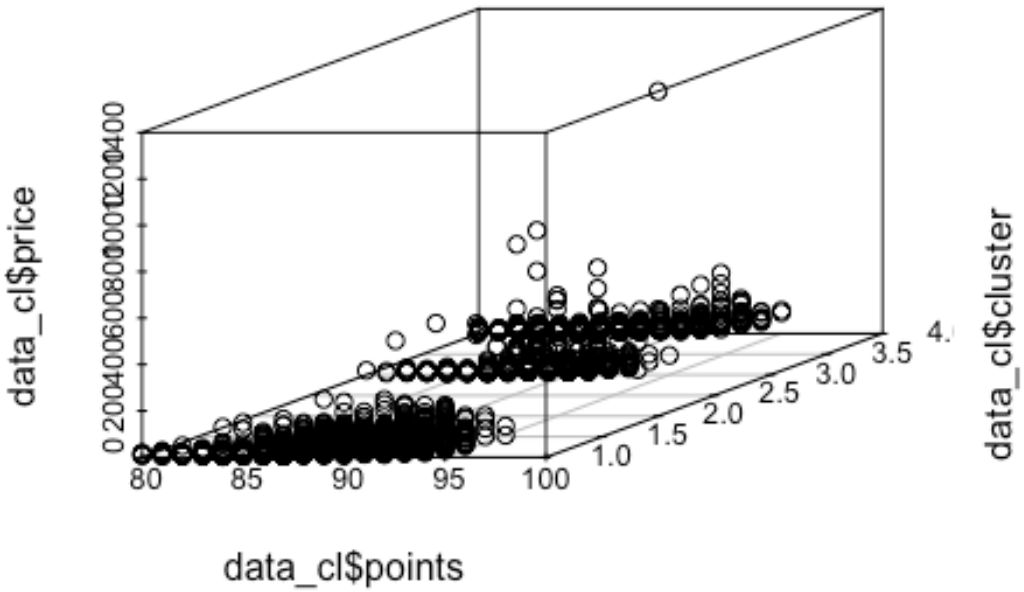
Let's see the distribution of the Clusters.

```
## # A tibble: 4 x 2
##   cluster     n
##     <int> <int>
## 1       1  2940
## 2       2    76
## 3       3   580
## 4       4  1404
```

As we see the biggest cluster is the number 1. Cluster 1: 34.2% Cluster 2: 33.8% Cluster 3: 18% Cluster 4: 14%

Plotting in 3D

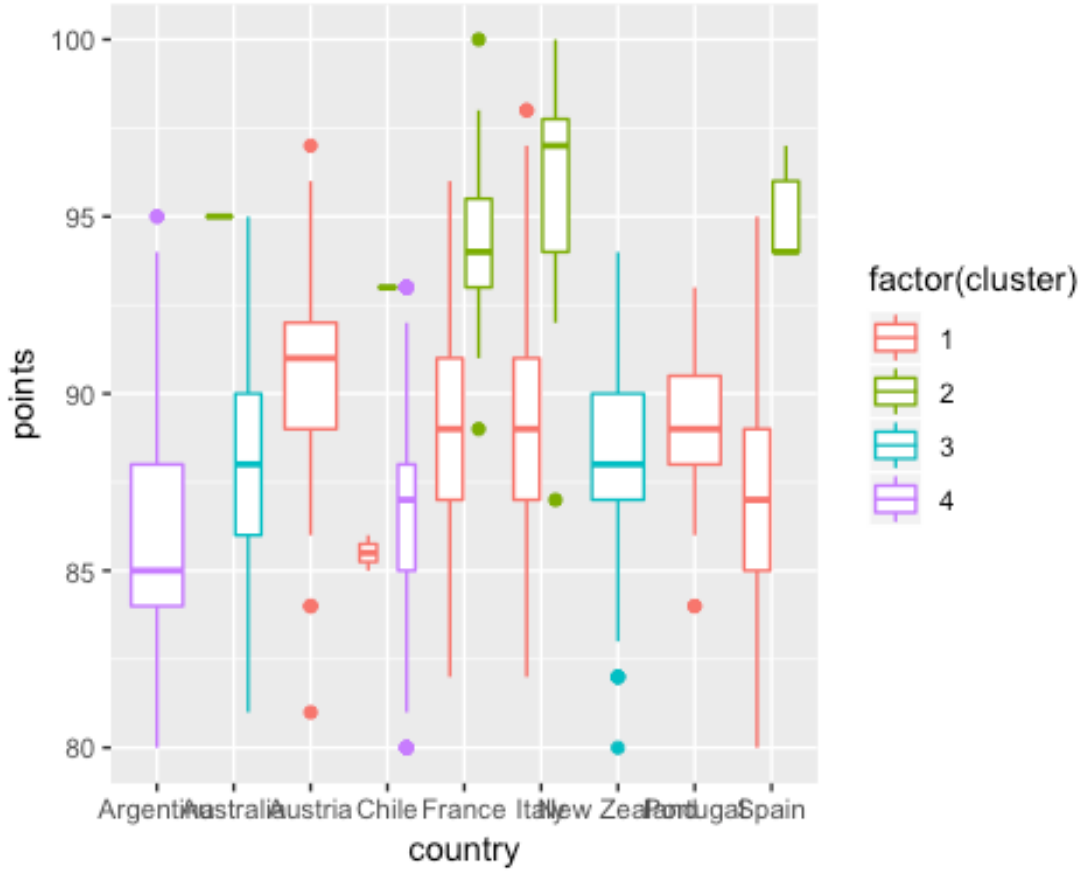
How this looks like in 3D



From this chart, we can gather the cluster 1 and 2 has less distance between each observation, while the fourth cluster is a lot more spreadout. Very interesting behavior of the third cluster.

Plot Country aganst Points

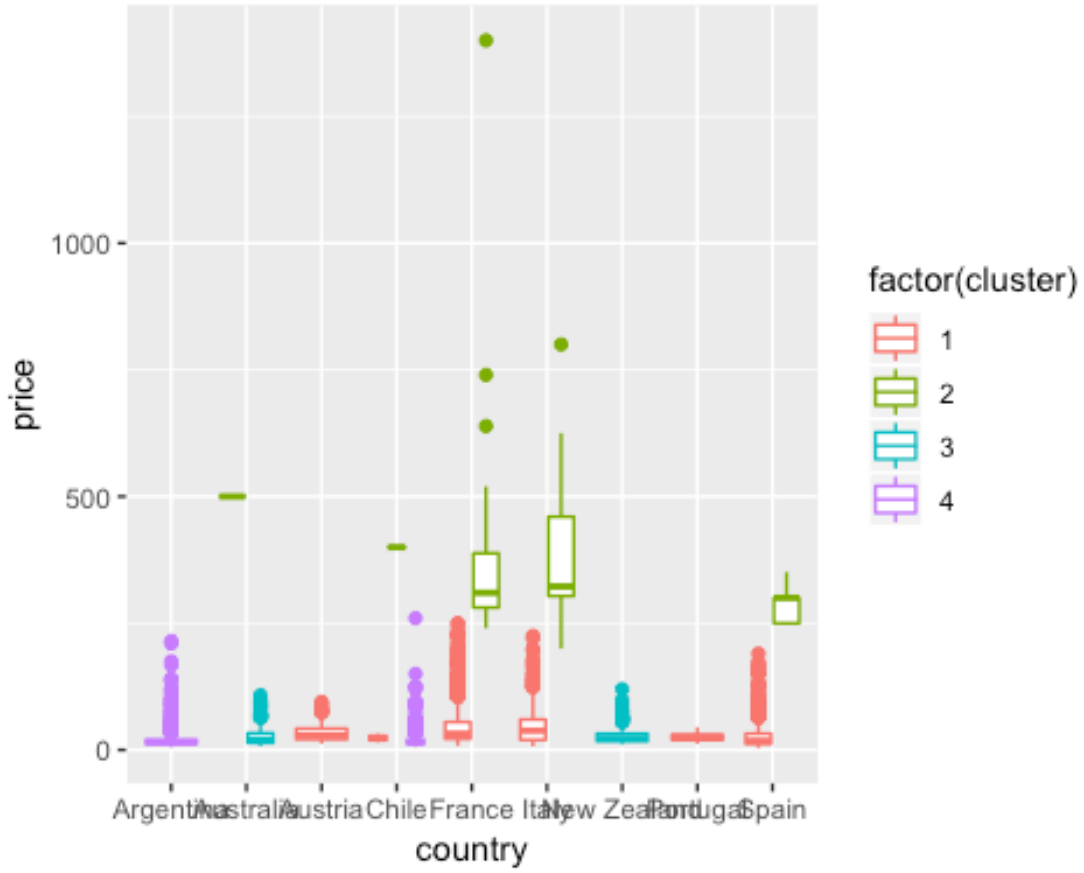
Let's determine which Country has the biggest acceptance across the clusters.



Wines from France has the biggest representation in the two biggest clusters, surprisely, Italy made it between the 3rd and 4th clusters. What it calls the attention is that New Zealand seems to have a growing acceptance.

Plot Country by Price

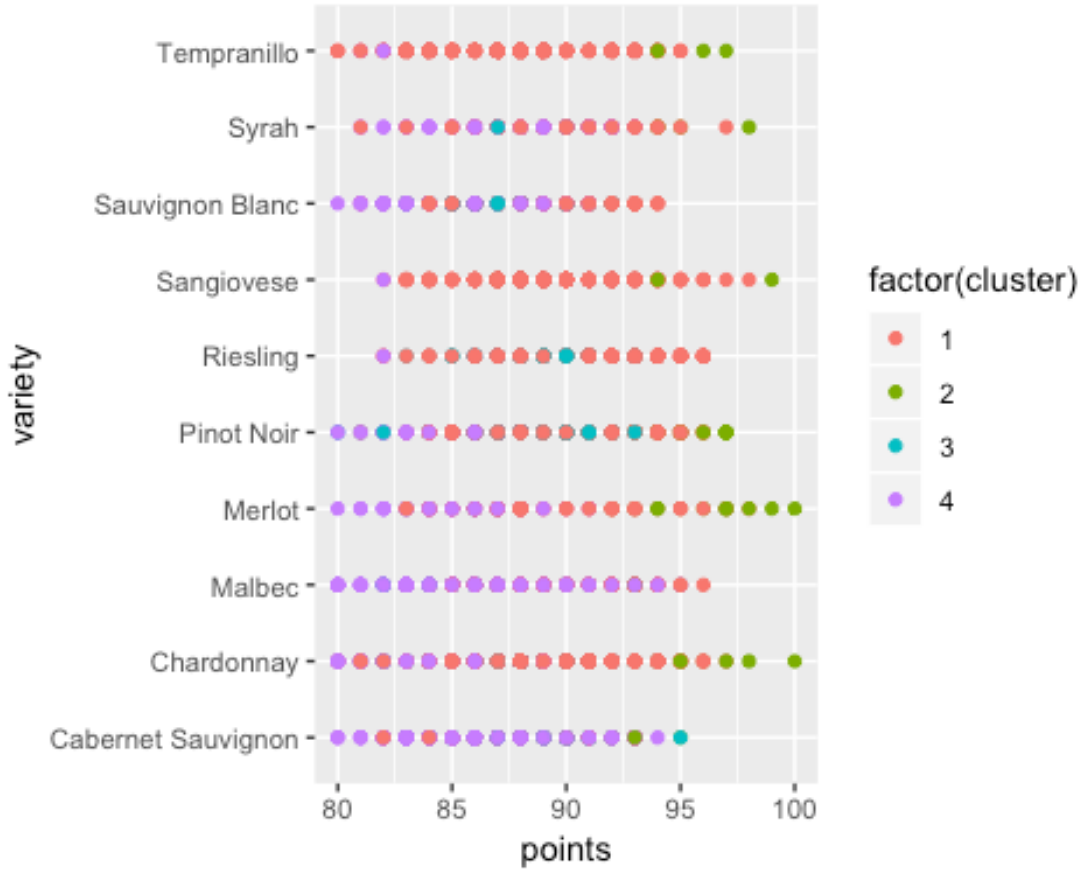
From our analysis, France has a great acceptance, at the same time, New Zealand has a potential for a growing market. Let's see what price tell us.



As expected as well; the most expensive wines are from European wines; however, New Zeland is not that remarkably expensive, which reinforce our previous assumption.

Variety by Points

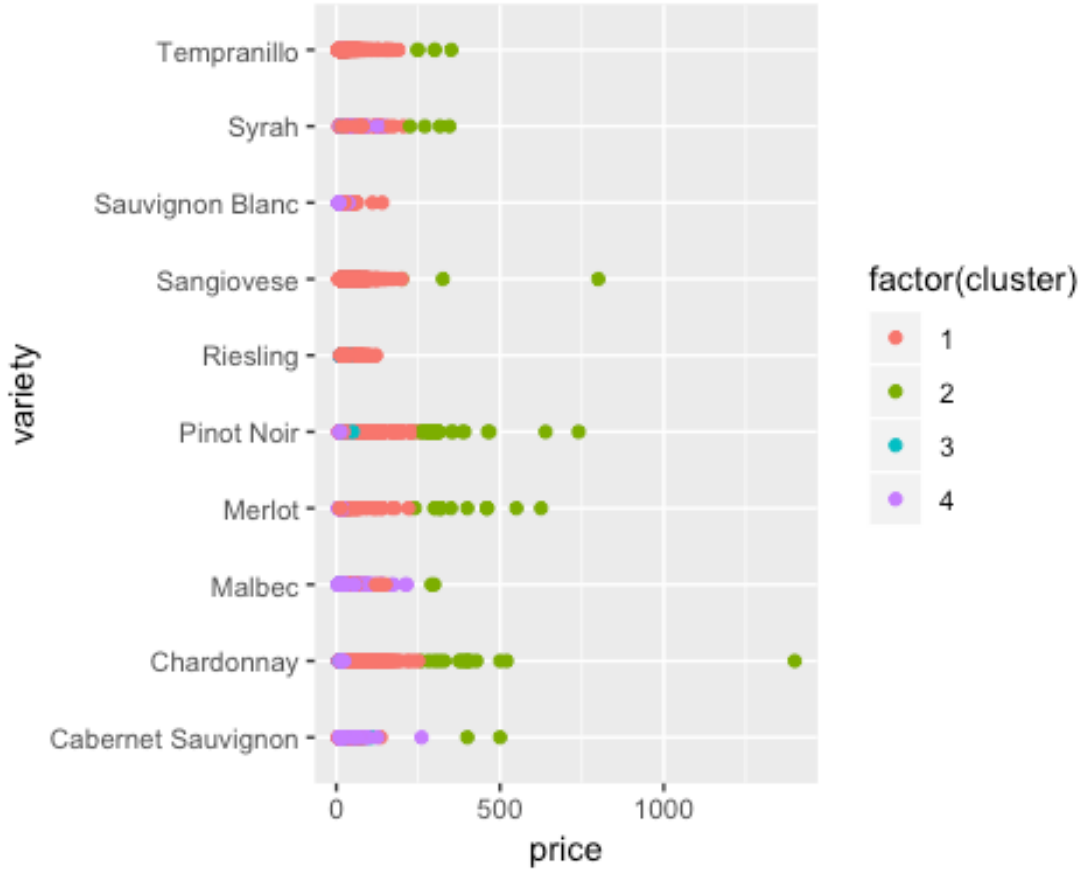
The Variety or type of wine plays a significant role in the type of wine produce by each region. Each category has a predominant acceptance within the tasters.



Interesting, Merlot has made a great come back since 1995; however, it has been mainly placed in the 4th cluste, similar to Sangiovese; at first thought, I might think to look into these two wines for future market, but we need to understand why they are under the two smallest clusters. In the other hand Tempranillos are making a great impresion. Spain is one of the best producers, and from our first chart, is quite well ranked within the wines and price of wines from spain are very raseable. What calls my attention is the Cabernet Sauvignon and Malbec; which could be a great potential to open a market from South America.

Varience by Price

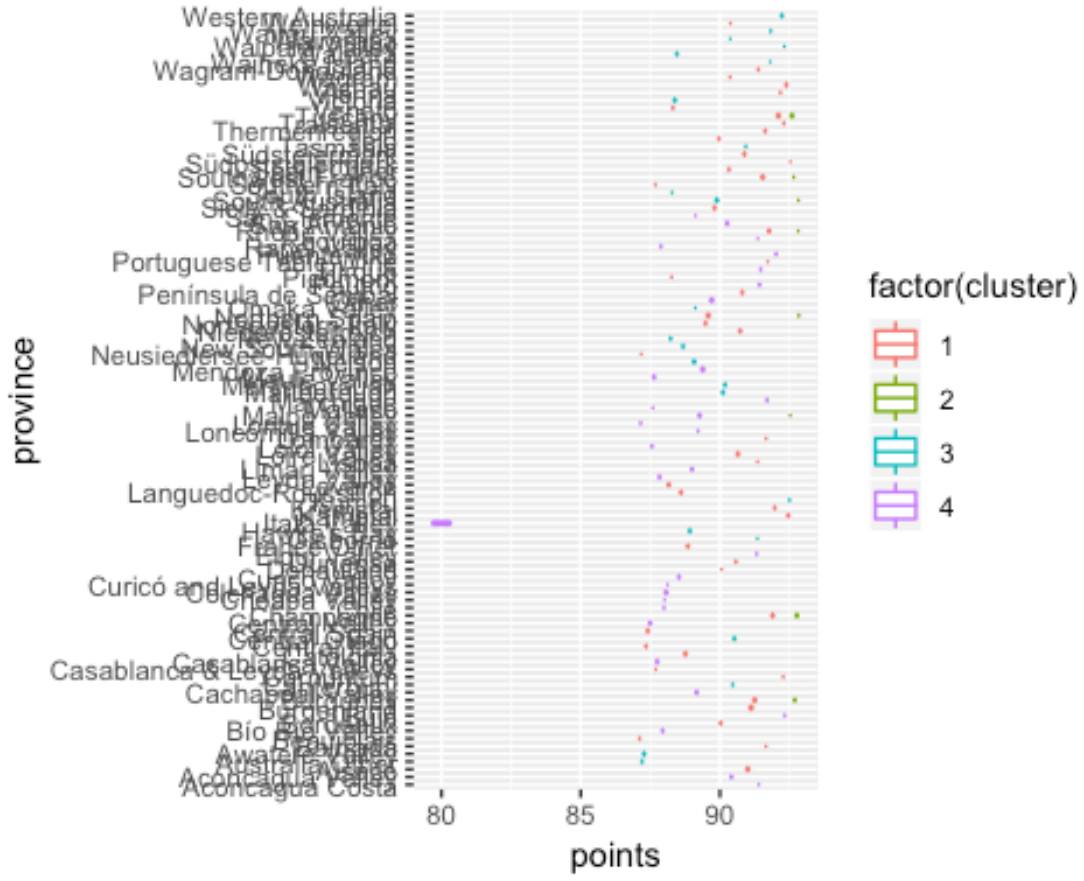
Let's analize the price per type of wine.



Chardoney ranked as one of the highest it's also an expensive wine. Depending of the region. However, it's placed in the 3rd cluster. Pinot is not a surprise to be at the top. Again, temnpranillo making a great impresion on price. We should make a film about wines in spain; that will boots people's interest even more. But I'm definetly thinking on Tempranillos as a great candidate for next promotion.

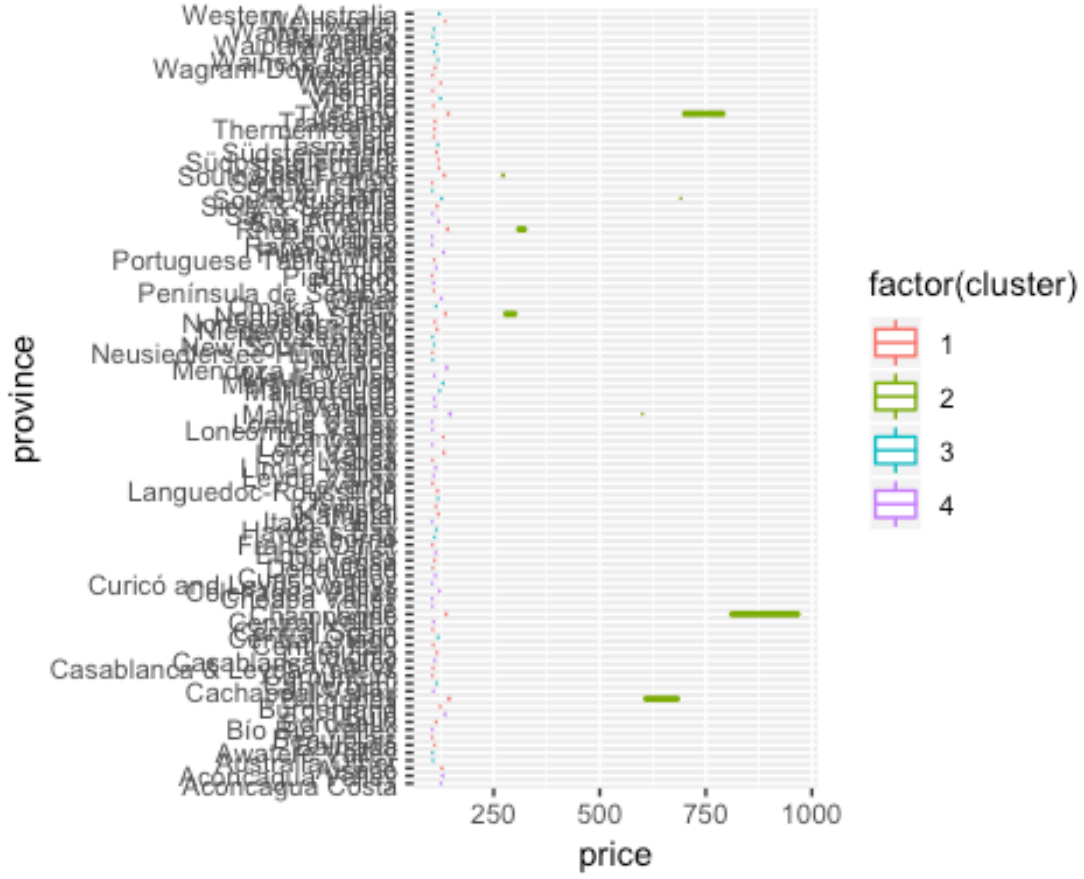
Province most ranked

Let's review people's opinion's about the Province; not all type of wines taste the same across the provinces; neither all the wines from the same province are the same.



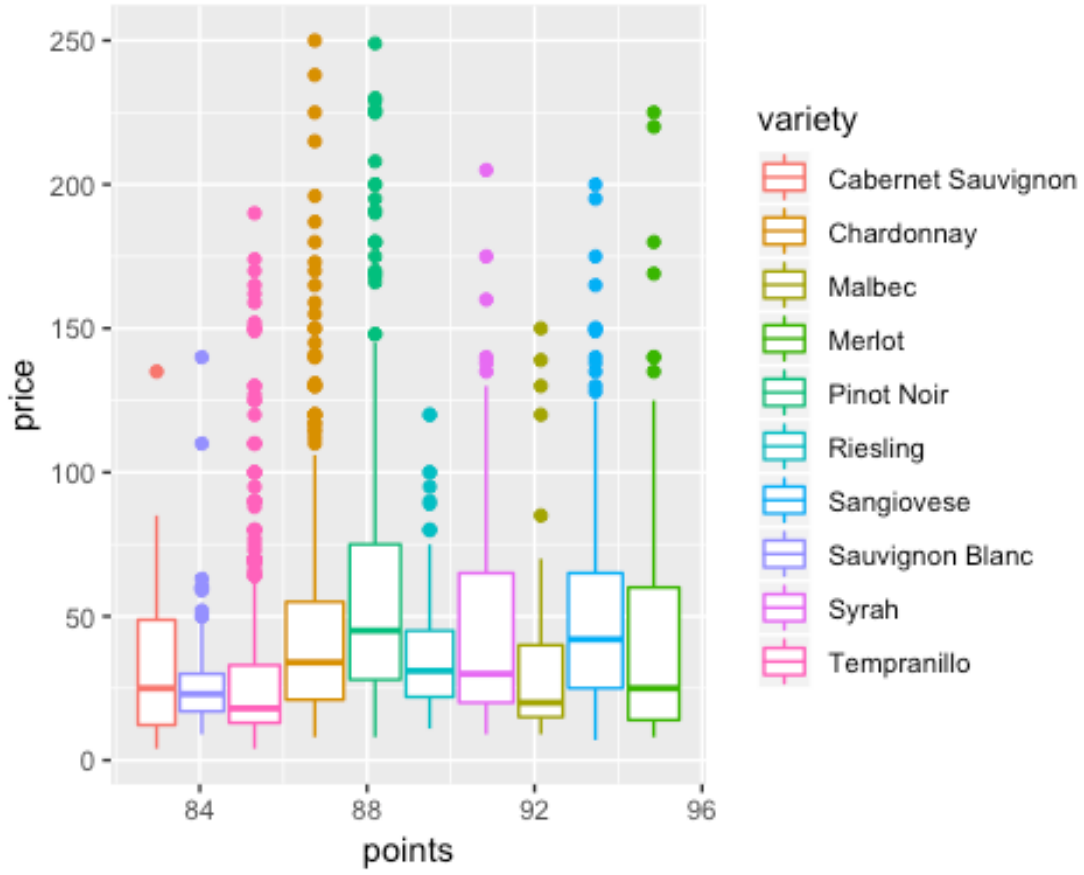
Pretty much all the provinces are spreadout across the X axes. Exept for one which is located in the 3rd cluster; which suspect it's one producing Cardoney from France. Cardoney, potentially from France has a very high acceptance, but also has a range of price. We'll need to drill down by Cluster 3 to appreciate better the name of the Province.

Price by Province



There is not much Province can tell us. In this case, re-inforce what we discussed previously, cluster number 3 might be associated to Chardonnay and is a expensive wine, very likely coming from France.

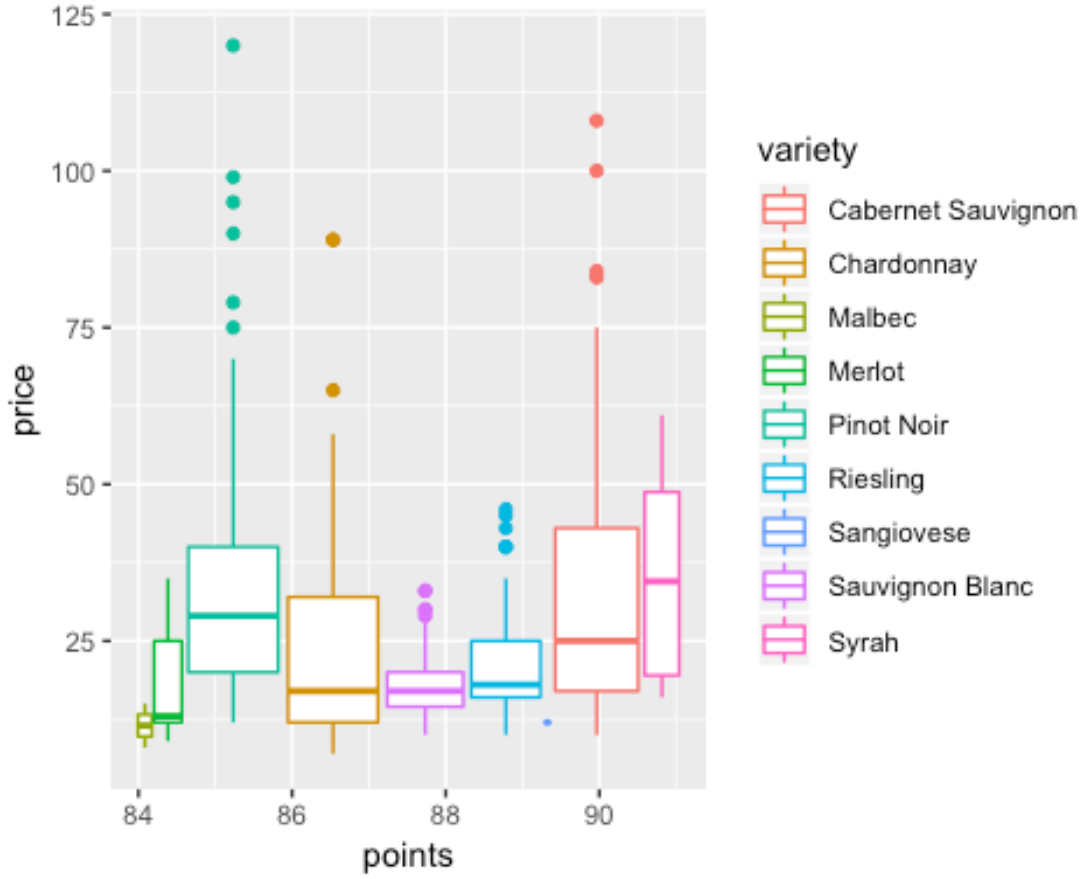
What Cluster 1 can tell us:



Tempranillos reinforced the decision; but one we didn't detect before is Malbec. There is great potential for Malbec to get into the top wines.

Investigate what's in the Cluster 3

As Chardonnay has not only great acceptance and also great price. Great candidate to get it in the tables.



Conclusion

Wine investment

From the Hierarchical Clustering Analysis we can detect that definetly Tempranillos from Spain, and Chardonnay and Pinot Noir from France are the best potential for import. One wine not being mentioned, but consistently showing results in the scores is Malbec from Chile. Not only a great wine, but the range of price.

Analysis perspective

Clustering is a very useful tool for data analysis in the unsupervised setting. This experience made us consider the below questions more carefully during our 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?

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.