Retail Group - Unisupervised Learning - Wine Review Assignment2

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

2019-02-24

## Installing packagaes:

Please make sure the libraries are included:

# Importing the data

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

## [1] 21044

## [1] 21044

## [1] 21044

## [1] 8

## 'data.frame': 10522 obs. of 8 variables:  
## $ country : chr "Spain" "New Zealand" "Italy" "France" ...  
## $ description: chr "Nicely oaked blackberry, licorice, vanilla and charred aromas are smooth and sultry. This is an outstanding win"| \_\_truncated\_\_ "Yields were down in 2015, but intensity is up, giving this medium-bodied, silky wine the potential to drink wel"| \_\_truncated\_\_ "Aromas of forest floor, violet, red berry and a whiff of dark baking spice unfold in the glass while wild cherr"| \_\_truncated\_\_ "Dark in color and in flavor profile, this medium-bodied Cornas boasts aromas and flavors reminiscent of chocola"| \_\_truncated\_\_ ...  
## $ points : int 95 94 90 90 90 90 90 91 91 91 ...  
## $ price : int 80 57 29 69 50 68 28 36 45 85 ...  
## $ province : chr "Northern Spain" "Kumeu" "Tuscany" "Rhône Valley" ...  
## $ variety : chr "Tempranillo" "Chardonnay" "Sangiovese" "Syrah" ...  
## $ latitude : num 43.3 -36.8 43.8 44.1 43.8 ...  
## $ longitude : num -4.25 174.56 11.25 4.83 11.25 ...

## 'data.frame': 10522 obs. of 4 variables:  
## $ points : int 95 94 90 90 90 90 90 91 91 91 ...  
## $ price : int 80 57 29 69 50 68 28 36 45 85 ...  
## $ latitude : num 43.3 -36.8 43.8 44.1 43.8 ...  
## $ longitude: num -4.25 174.56 11.25 4.83 11.25 ...

## [1] 10522

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 [1:10522, 1:4] 2.201 1.893 0.659 0.659 0.659 ...  
## - attr(\*, "dimnames")=List of 2  
## ..$ : chr [1:10522] "1" "3" "5" "7" ...  
## ..$ : chr [1:4] "points" "price" "latitude" "longitude"  
## - attr(\*, "scaled:center")= Named num [1:4] 87.9 35.3 14 -22.2  
## ..- attr(\*, "names")= chr [1:4] "points" "price" "latitude" "longitude"  
## - attr(\*, "scaled:scale")= Named num [1:4] 3.24 44 37.8 85.56  
## ..- attr(\*, "names")= chr [1:4] "points" "price" "latitude" "longitude"

# 

# For the Hierarchical Cluster Analysis we select:

# 1. Points Number

# 2. Price Number

# 3. Latitude Number

# 4. Longitude Number

# 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:

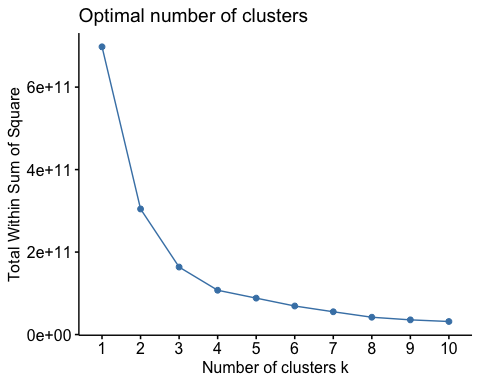
The agglomeration method that can be used is (an unambiguous abbreviation of) one of “ward.D”, “ward.D2”, “single”, “complete”, “average” (= UPGMA), “mcquitty” (= WPGMA), “median” (= WPGMC) or “centroid” (= UPGMC).

For our analysis, we will use Ward.D2:

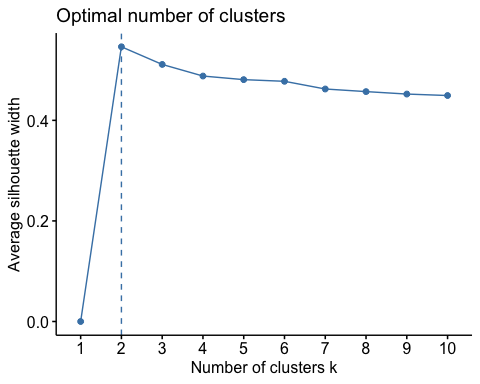
**WRhc\_hclust <- hclust(WRdiss, method = "ward.D2" )**

# Determining Optimal Number of Clusters

## Elbow Method



## Average Silhouette Method



## 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.

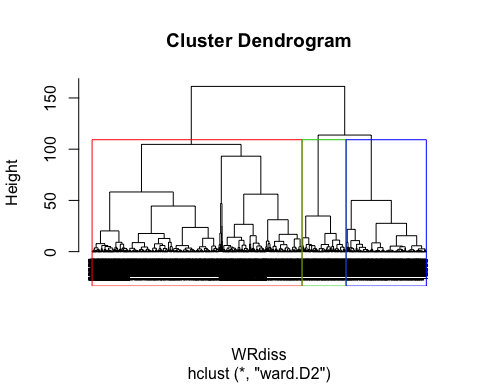
The height of the vertical line or vertical axis, indicates the (dis)similarity between two observations. The higher the height of the vertical line/fusion, the less similar the observations are.

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 3 groups, although Silhouette method suggested 2, we’ll have a better analysis when using 3.

## sub\_grp  
## 1 2 3   
## 6610 1386 2526

Drawing the dendrogram with a border around the 3 clusters

 From the dendogram we are able to identify the first Cluster has the majority of the observations.

|  |  |
| --- | --- |
| 6610 | 62% |
| 1386 | 13% |
| 2526 | 24% |

We can conclude HCluster might not be the best to interpret the data and it can be the dataset has lots observations. However, we will continue with our analysis and evaluate the Agglomerative method.

## 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)

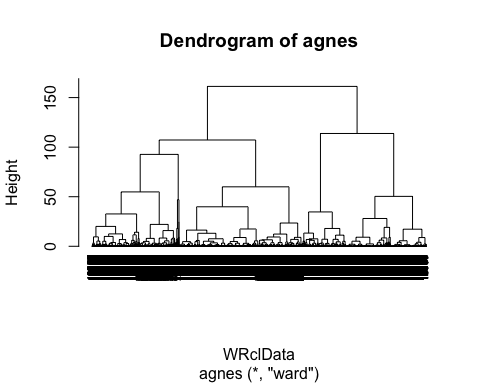
The Agglomerative coefficient allows us to find certain hierarchical clustering methods that can identify stronger clustering structures.

# 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:

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:



## # A tibble: 3 x 2  
## cluster n  
## <int> <int>  
## 1 1 6610  
## 2 2 1386  
## 3 3 2526

## Divisive Hierarchical Clustering

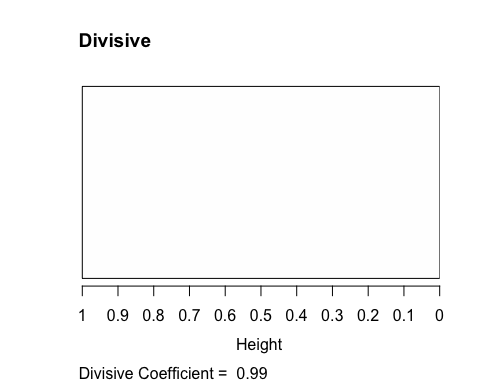
This variant of hierarchical clustering is called top-down clustering or divisive clustering . We start at the top with all documents in one cluster. The cluster is split using a flat clustering algorithm. This procedure is applied recursively until each document is in its own singleton cluster.

For this analysis we will use:

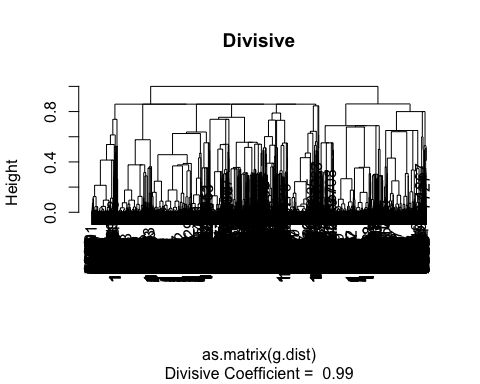
1. Variety Nominal
2. Country Nominal
3. Province Nominal
4. Points Price

We will find the distance through the gower function.

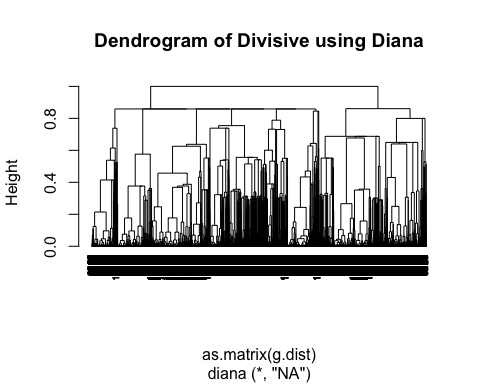
Let’s run the Diana method and plot the results:



With a Coefficient of 0.99 is giving us great confidence to continue with this analysis.

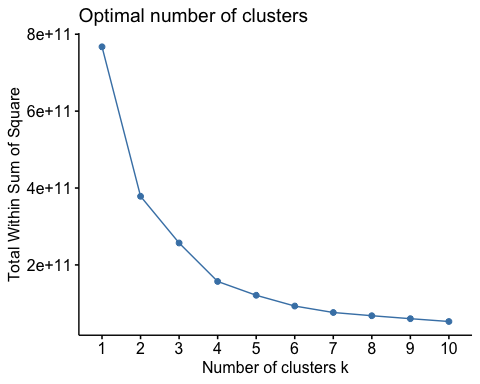


## [1] 0.9944146

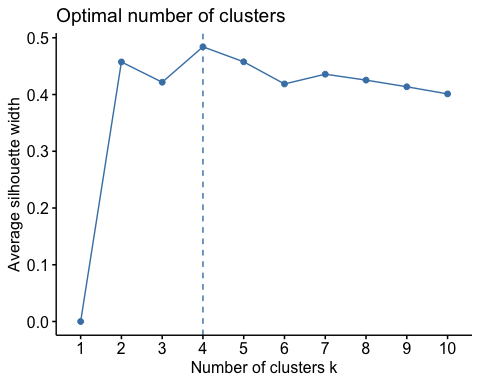


# Determining Optimal Number of Clusters for Divisive Hierarchical Clustering analysis

## Elbow Method

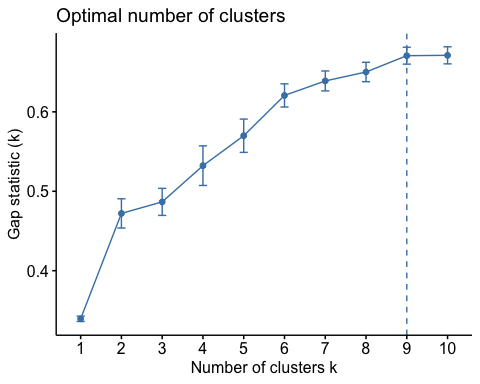


## Average Silhouette Method



Silhouette giving us 4 clusters.

## Gap Statistic Method



## 

## Interesting how Gap Analysis is providing 9 clusters.

## Cut tree into 4 groups

As per the Silhouette method, we use K=4

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



# Further Analysis

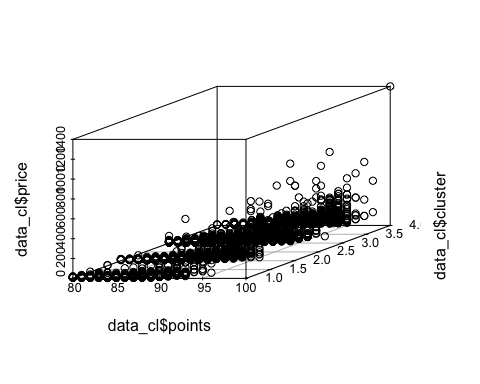
Let’s also cutree output to add the the cluster each observation to the original data to drice further analysis against the nominal variables:

Let’s see the distribution of the Clusters.

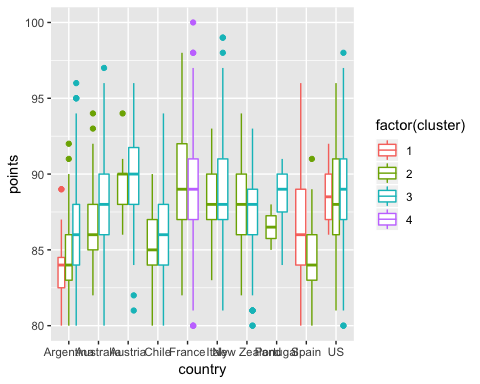
## # A tibble: 4 x 2  
## cluster n  
## <int> <int>  
## 1 1 806  
## 2 2 2013  
## 3 3 6331  
## 4 4 1372

As we see the biggest cluster is the number 1 and 4

## Plotting in 3D

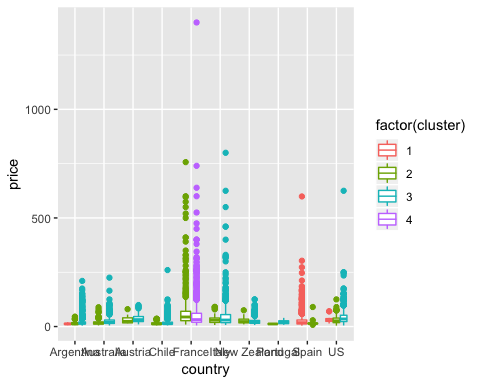
How this looks like in 3D  From this chart, we can gather the cluster 3 and 4 has less distance between each observation, while the fourth cluster is a lot more spreadout. Very interesting behavior of the third cluster.

## Plot Country aganst Points

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

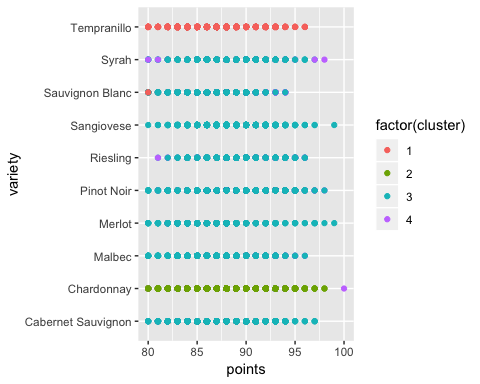
Wines from France, France, Italy and Spain have a great acceptance across the clusters. Wines from France on its own is in one single cluster.

## Plot Country by Price

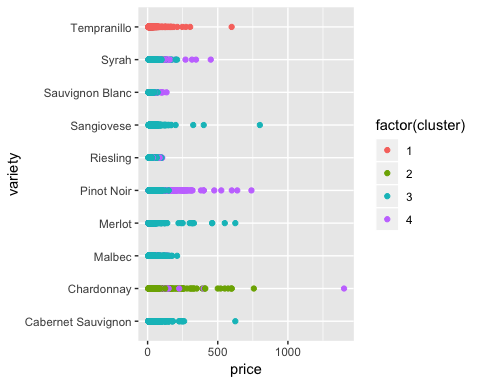
From our analysis, France has a great acceptance, at the same time, New Zealand has a potential for a growing market. Let’s see what price tell us. 

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

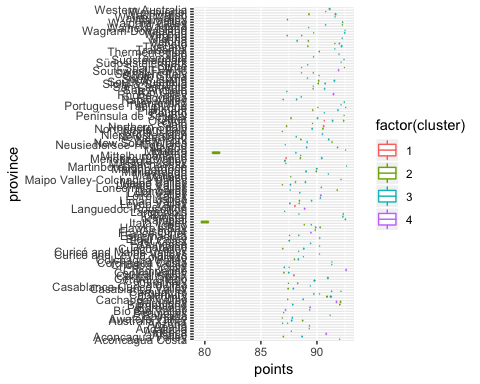
## Variety by Points

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 3rd cluster with other similar wines such as Sangiovese; at first thought, I might think to look into these two wines for future market. 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 our attention is the Cabernet Sauvignon and Malbec; which could be a great potential to open a market from South America.

# Varience by Price

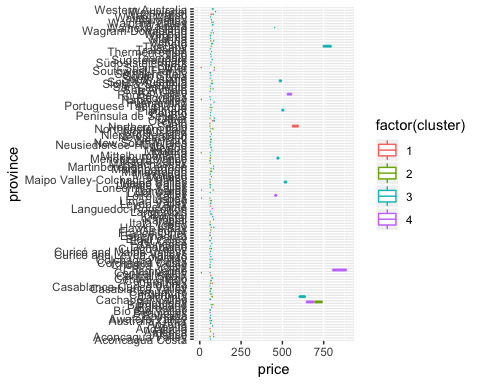
Let’s analize the price per type of wine.  Chardoney ranked as one of the highest it’s also an expensive wine. Depending on the region. However, it’s placed in the 2nd cluster. Pinot is not a surprise to be at the top as its quite high price for those coming from France. Again, Tempranillo making a great range on price as well as its acceptance. We should make a film involving wines from Spain to boots people’s interest in that market.

## 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. 

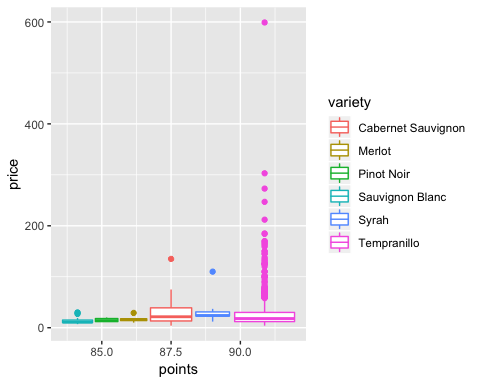
Pretty much all the provinces are spreadout across the X axes. Exept for one which is located in the 2nd cluster; which suspect it’s one producing Cardoney from France. Cardoney, potentially from France has a very high acceptance, but also has a range of price. We’ll need to drill down by Cluster 3 to appreciate better the name of the Province.

## Price by Province

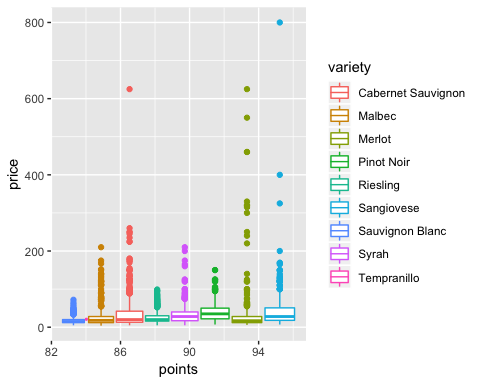


There is not much that Province can tell us for our analysis. In this case, re-inforce what we discussed previously, cluster number 3 might be associated to Chardonnay and Pinot Noir which are expensive wine, usually coming from France.

## What the smallest cluster, number 1 can tell us:

 Tempranillos reinforced the decision to look for Tempranillo to be introduced in the Market.

## Investigate what’s in the biggest Cluster number 3 tell us

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

# Conclusions

## From the Wine investment

From the Hierarchical Clustering Analysis, we can detect that definitely Tempranillo wine from Spain, and Sangiovese 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. Highly rated wine and with a great range of price.

We should consider to drop Province for a new iteration of the analysis.

## From the Analysis perspective

Agglomerative does not perform well with extensive amount of data. Divisive gave us a greater number of clusters. Is this case, we could find more ways the data can be segmented.

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.