**CS7DS3 Main Assignment**

**Student Name: Rohit Rawat**

**Student Number: 19301289**

**Declaration: "I have read, and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.**

**I have also completed the Online Tutorial on avoiding plagiarism ‘Ready Steady Write’, located at** [**http://tcd-ie.libguides.com/plagiarism/ready-steady-write**](http://tcd-ie.libguides.com/plagiarism/ready-steady-write)**."**

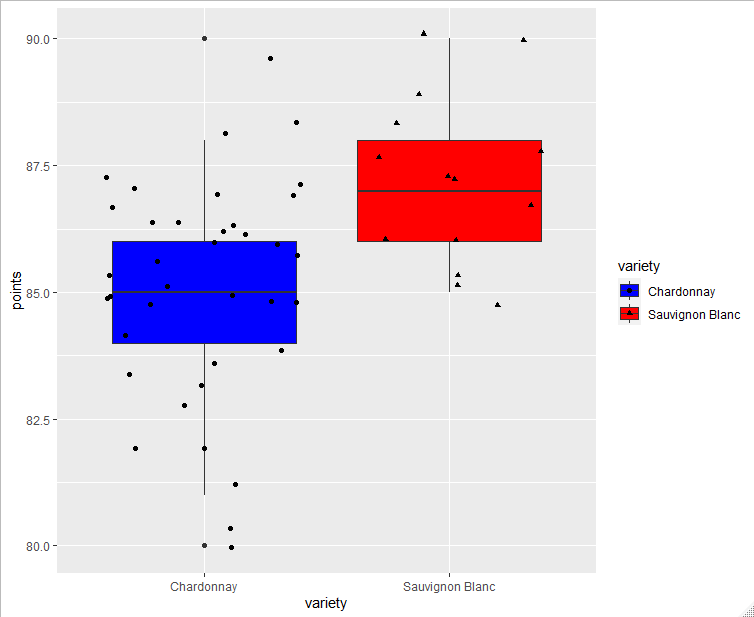
**Introduction:**

Wine review dataset is being used for the purpose of analysing which variety of wine is better based on the description, points, price and origin. It is a very huge dataset and consists of wine data almost from all the countries. For the analysis purpose, only subset of the data is being used and that too for the specific region, wine variety or country. This analysis helped to understand which wine is better among Sauvignon Blanc from South Africa and Chardonnay from Chile. Further it helped to decide how many regions in Italy provides the best wines above the average rating. Clustering of the wine variety based on the price and points using model clustering and deciding on the cluster which gives better value for the price which means high rating and low price.

**Question 1.**

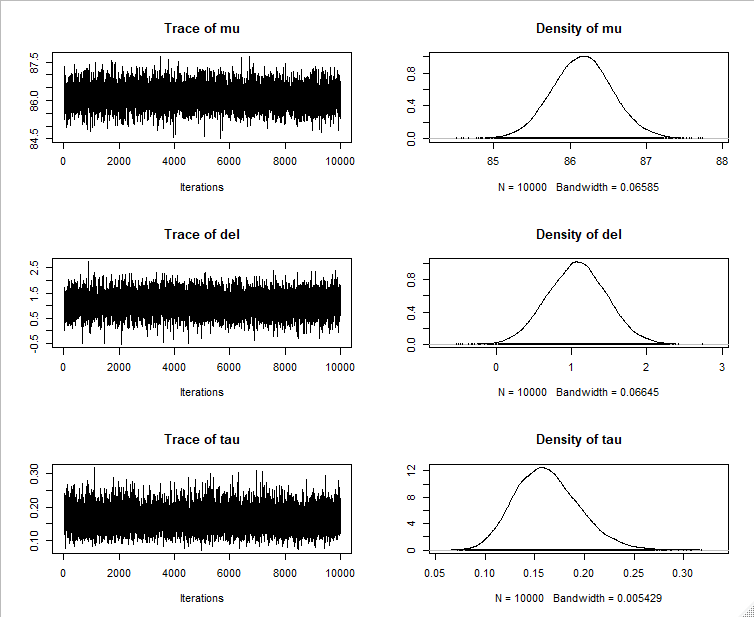
**Part a**

For this part, Sauvignon Blanc from South Africa is being compared with Chardonnay from Chile. Both are priced at 15 Euro. To calculate this, first wine data is loaded from csv file and then individually data has been filtered out for each wine respectively and only two columns have been selected that is winery and points. As based on these points, which wine is better will be decided. This filter and selection are done using dpylr package in R. After this, both the dataset has been combined into one dataset which is then used for the further analysis. To get proper information from data, boxplot has been plotted for both the wines along with the jitter for all the datapoints. This plot gives clear information about which wine is better and number of observations for each wine.

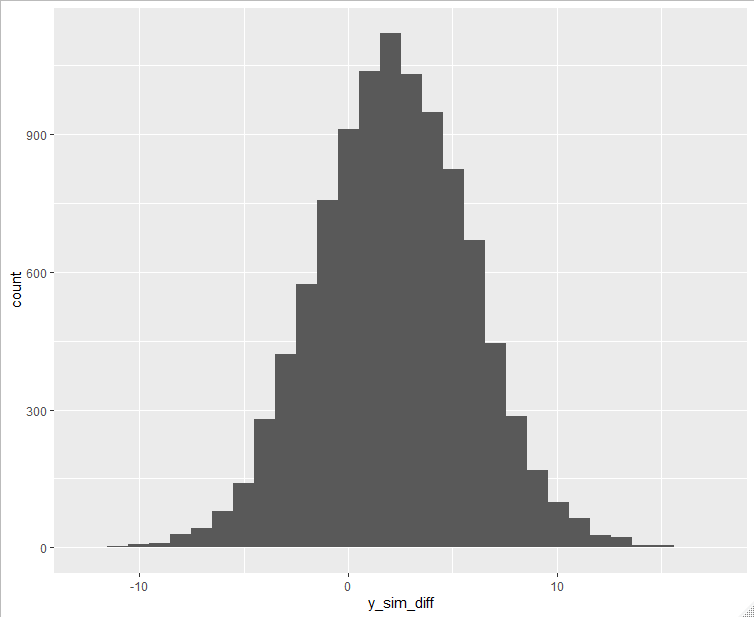


Blue colour boxplot indicates Chardonnay and red boxplot is Sauvignon Blanc. Observations for Chardonnay is marked by circle and for Sauvignon Blanc by triangle. Sauvignon Blanc has higher points compared to the Chardonnay. Median of Sauvignon Blanc is higher than Chardonnay. Chardonnay has many outliers compared to the Sauvignon Blanc. Calculating the mean of each wine we get 85.08 for Chardonnay and 87.21 for Sauvignon Blanc. Even from this we can say Sauvignon Blanc is better than Chardonnay. As we have very less number data, to strengthen this result t-test is conducted to check whether two datasets are following the same normal distribution. Basically, we consider null hypothesis which is that both the means are same. This test will give a p- value which we must compare with the significant level. The p-value tells us whether to accept or reject null hypothesis. In this case, 95% confident level is being considered for which significant level is 0.05 and p-value should be less than this value. After running t-test, p-value is 0.002 which is less than the significant level and null hypothesis can be rejected. This means there is a difference in the mean of Sauvignon Blanc and Chardonnay. As mean of Sauvignon Blanc is greater than the Chardonnay, **Sauvignon Blanc is better wine than Chardonnay**.

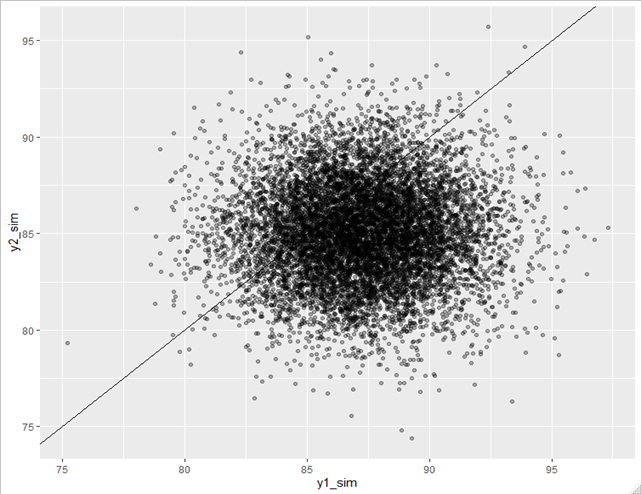
To calculate the difference between the two means and to tell by how much, one wine is better than the other, Gibbs sampling is used. Gibbs sampling take in initial mean value which is set to 90 as points range from 80 to 100 and standard deviation to 10, so tau will be 1/100 and default value start a and b i.e. 1 and 50. It will be run for 10000 iterations. In this y1 has been set to Sauvignon Blanc and y2 to Chardonnay. After running Gibbs sampling for our dataset, del, mean and tau value is generated. Positive value of del (which is the difference between means) indicates mean 1 is greater than mean 2. We get del value as 1.07 and the difference between mean is 2\*del, which is 2.14. **So, Sauvignon Blanc is better than Chardonnay by 2.14.** These three parameters can be plotted using mcmc, which tells that they follow the normal distribution.



Next, probability needs to be calculated for each variety of wine. From this, we will get the probability for the Sauvignon Blanc to be better. Same sample will be used as above generated using Gibbs sampling. This will be used to generate two set of samples for Sauvignon Blanc and Chardonnay each, that will give the probability of each wine. For both variety, 10000 samples have been created and based on that we get probability of 0.7231 for Sauvignon Blanc. **Hence, Probability for the Sauvignon Blanc to be better is 0.7231.** Difference between these simulations have been plotted which is normal distribution, which means posterior value follows normal distribution.

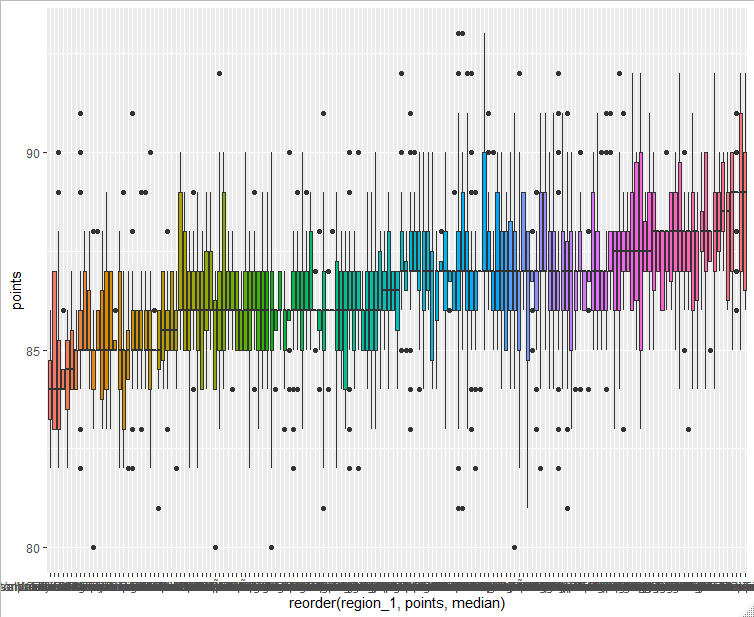


Following plot show the simullations that was generated above to calculate the probability.



**Part b**

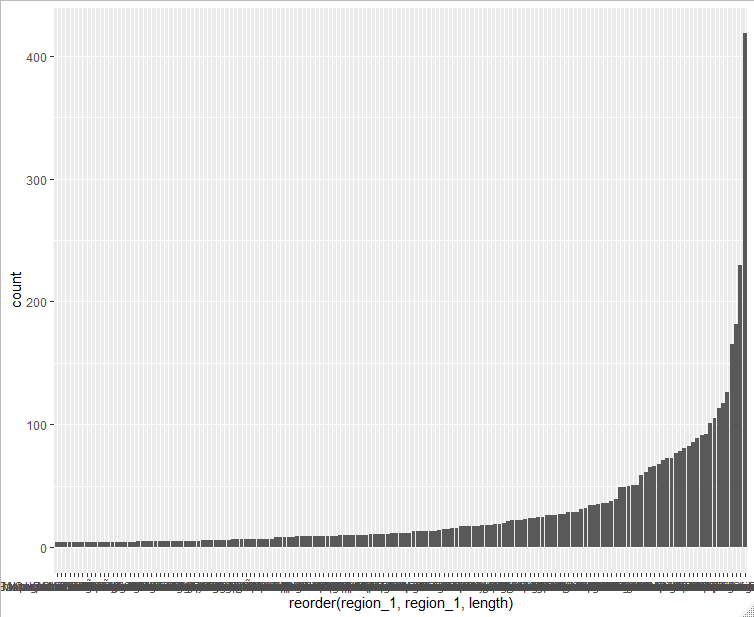
In this, how many regions in Italy produce the best wine above average needs to be calculated. This analysis has been restricted to the wines costing less than 20 Euro and the region which has four our more review. After sub-setting and filtering a data we will be having two columns region and points. These regions have 161 value after filtering. Out of which only **75 regions generate wines above average rating**. This is just looking at the dataset. Sardinia has been excluded from this as it has same value for points for all the observation. Following plot shows, all 161 region and their boxplot to display the median and the spread of data.



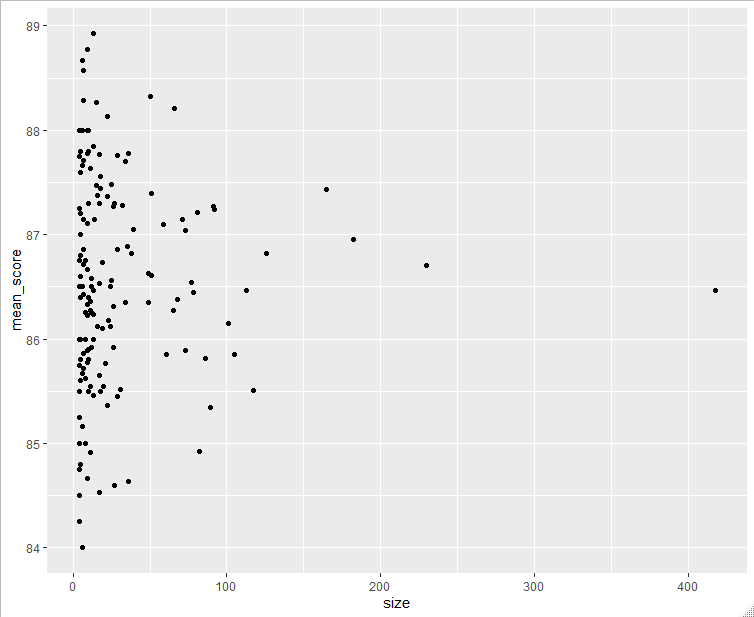
Wordcloud chart shows region names having points above average.



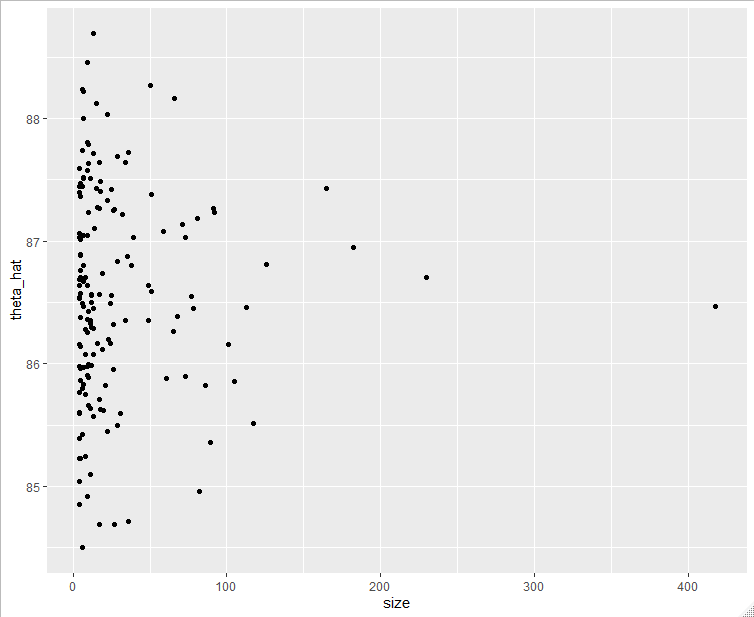
Next plot is used to display number of samples in each region, and it can be inferred that some regions have extremely large number of data which can create bias in calculating mean.



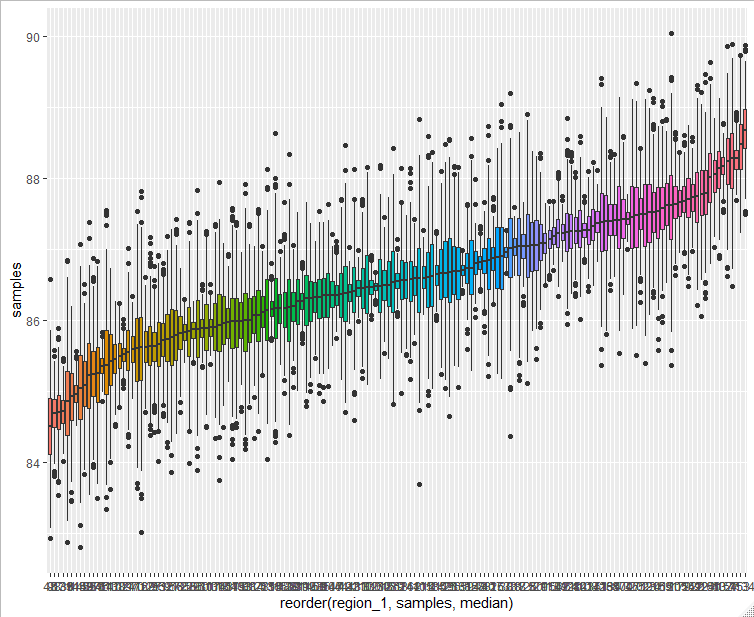
Next plot displays that most of the points are for the sample less than 100. This suggests there can be a huge difference in the mean, hence using Gibbs sampling to compare the means of all the region and using this variation to generate reliable results.



After running Gibbs sampling for 500 iterations, **79 regions are above average mean producing best wines.** Which is almost like the number we got from the dataset. Even, variability of posterior value is reduced for the smallest sample. Plot is shown below.



Even from below figure, variability can be seen between the regions as compared to same plot above plotted for the normal population.



**Thus, Sauvignon Blanc is better than Chardonnay by 2.14. Probability of Sauvignon Blanc to be better is 0.7231. 79 regions are above average in Italy producing the best wines after calculating it using Gibbs sampling.**

**Question 2.**

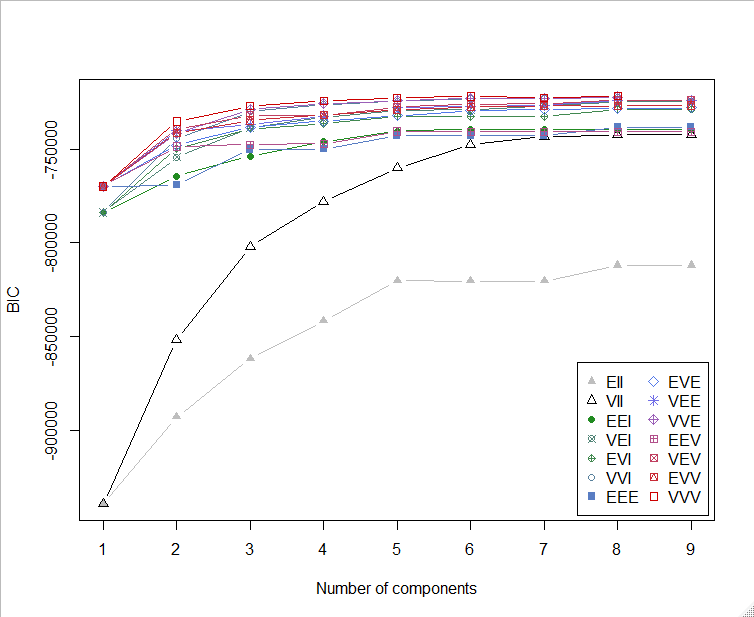
**Part b**

In this part, clustering of wines needs to be done based on the price and points for the USA. A new dataset is made, consisting of two columns points and price for the USA. After filtering, nan values are removed from it. For further clarity, I have considered observations which has price less than 800. Clustering model is run on this dataset to find the cluster which will be better in terms of the price value i.e. the cluster which will have less price for maximum points.

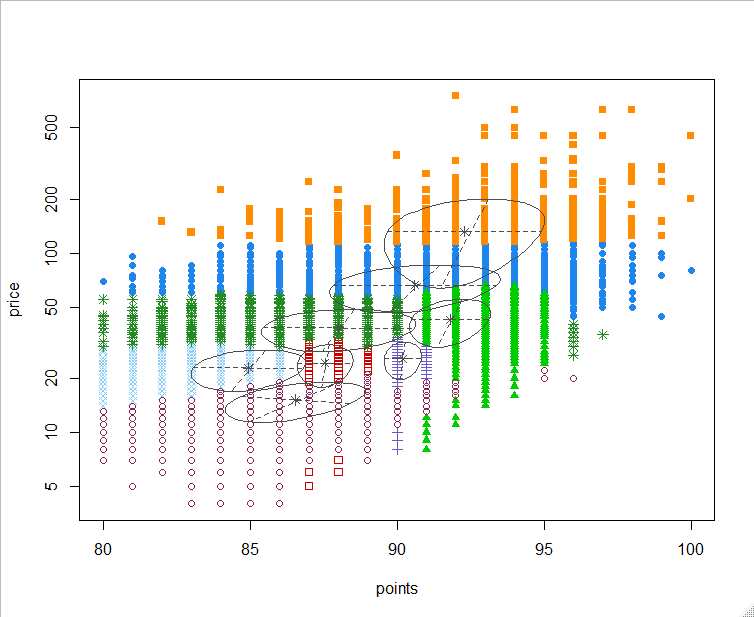
For this purpose, Model based clustering is used, which considers the probability as well as uncertainty to the cluster unlike k-means and hierarchical clustering. Model based clustering itself selects the optimal number of clusters. For this purpose, R package mclust is used to do the clustering. Gaussian mixture model approach is used and EM algorithm to select the optimal model. This helps in differentiating a high and low uncertainty clusters which is useful in providing unique solutions. There 14 covariance parameters used in model-based clustering and how which the best one is used for clustering. This is decided based on the BIC value.

First, data is using Mclust package and it is clustered as per the best covariance structure and no. of clusters which is decided based on the EM algorithm. This model can have highest BIC value. By default, G’s is 1:9, even you can check for 10:20 by explicitly mentioning in the function. By plotting BIC, you will get BIC for all the covariance structure and you can select the one which has highest BIC or similar BIC with a smaller number of clusters.

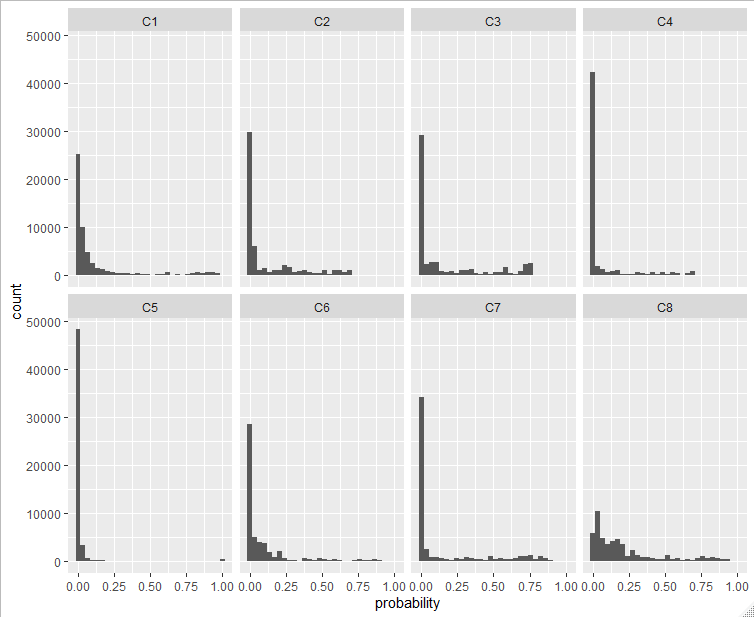
This is a BIC plot for G 1:9



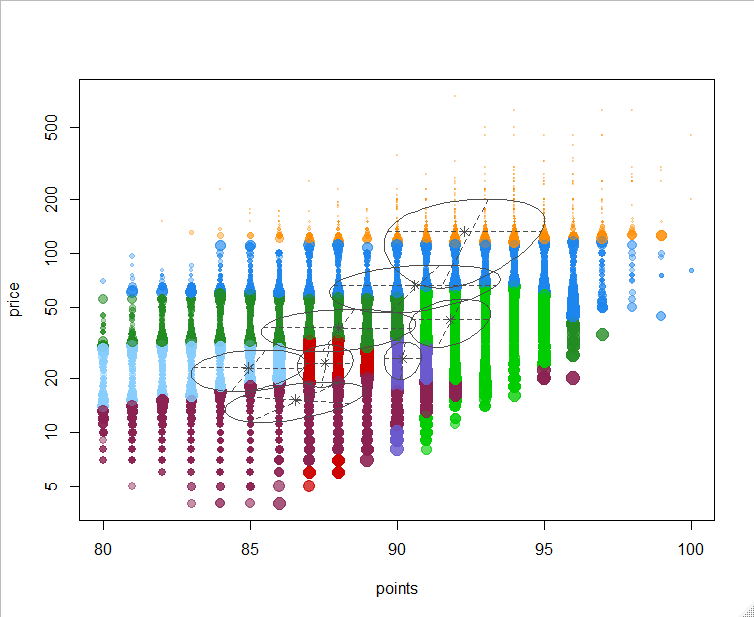
From this, we can say VVV is better and it has 8 clusters. We will be using this to plot the classification and the uncertainty plot. Classification plot is to visualise how data has been assigned to the clusters. Below is the plot of classification.



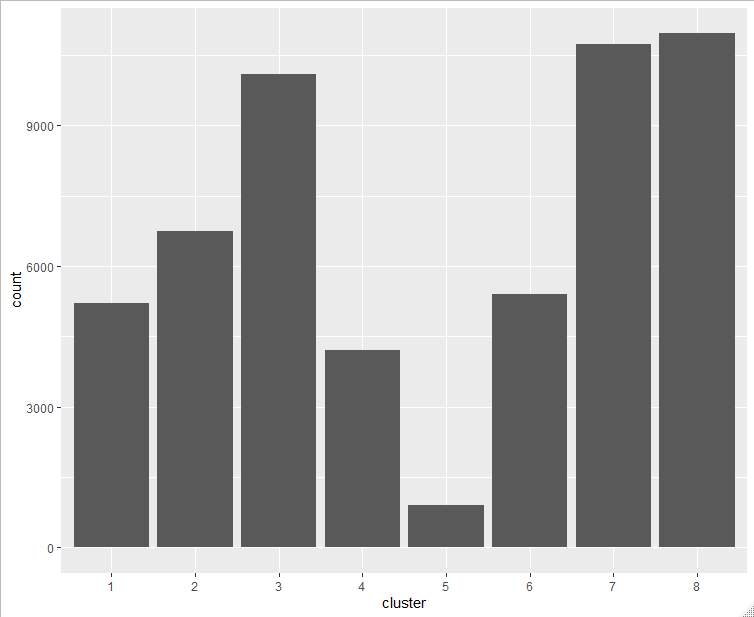
Now we will see in each cluster how the data is assigned to them. Cluster 5 is the smallest cluster as it does not have any observations in the middle probability, and it is also a compact cluster. Observations with probability greater than 0.5 is added to that cluster.



Following plot shows the uncertainty which means whether data is identified with higher or lower uncertainty. If observations are more, then uncertainty is higher.



Next plot displays the count of observations in each cluster. This proves that the cluster 5 is compact and cluster 8 has the greatest number of observations.



**Cluster 8 is good value for money, because it has maximum number of observations and observations in that cluster does not have high price for low points and it has spread over all the points.**

One problem with Model based clustering is it needs underlying model like GMM for clustering data and it takes heavy toll on the computation resources. If we have high dimensions, then this will be taking lots of time to form clusters. To avoid this, we can take heuristic approach of clustering like k means and hierarchical which is not heavy to run on resources.

**Conclusion:**

* Sauvignon Blanc is better than Chardonnay wine, it has been computed using t test and Gibbs sampling. Difference between mean is 2.14
* 0.7231 is the probability of the Sauvignon Blanc to be better.
* 75 regions are better than average mean calculated from the dataset and 79 regions based on the Gibbs sampling. Wordcloud has been used to display the regions name.
* VVV with cluster 8 is better than other covariates as it has the highest BIC. Cluster 8 is the value for money as it has observations spread over the points for reasonable price.