1. My wife likes Sauvignon Blanc from South Africa. My mother-in-law likes Chardonnay from Chile. Both agree that €15 is the right amount to spend on a bottle of wine.

a. i. Which type of wine is better rated? How much better?

ii. Suppose I buy a South African Sauvignon Blanc and a Chilean Chardonnay, both priced €15. What is the probability that the Sauvignon Blanc will be better?

Dataset

The dataset consists of detailed information about wines available all over the world. The dataset is taken from Kaggle and was collected from Wine Enthusiast. The dataset has the reviews and description for all the varieties of wines from different wineries. The regions to which these wines belong to are also part of the dataset. The price of the wines is also given along with their ratings given by the tasters. The data is imported into the R using read.csv() function.

i. Which wines are better, Sauvignon Blanc or Chilean Chardonnay?

Data Preparation

To answer this question, the data has to be filtered first based on the conditions. The first condition would be to subtract the wines that are priced are more than €15. Then a subset of data having country, variety, price and points are taken as this analysis is only based on these columns. The next filter that is applied to the data is ((variety == "Sauvignon Blanc" AND country == "South Africa") OR (variety == " Chardonnay " AND country == "Chile")). There are no empty rows in this data, so the data is now ready for the analysis.

Data Analysis

The mean and variance of the points of wines were calculated as part of the basic analysis to check which of these two is better. The below table shows the mean and variance of wines:

Wine	Mean	Variance
Chardonnay	85.08108	4.854354
Sauvignon Blanc	87.21429	2.950549

Table 1 Mean & Variance

It can be clearly seen from the table that the mean and variance of the Sauvignon Blanc is greater than the Chardonnay wines. Also, the mean of the Sauvignon Blanc has less variance compared to the Chardonnay wines. So, the basic analysis suggests that the Sauvignon Blanc are better than the Chardonnay wines. This can also be proved visually using a box plot.

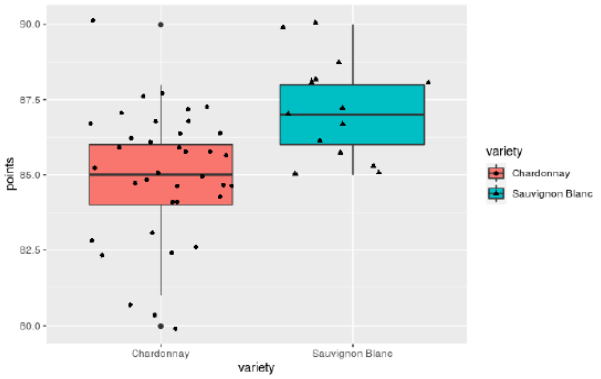

Fig 1 Box Plot

To prove this statistically, the next step in the analysis would be Student's T-test. The basic requirements for the T-test are independent samples having the same variance and they must be normally distributed. To check if the sample is normally distributed, we plot them using the QQ plot using the qqnorm() function in R.

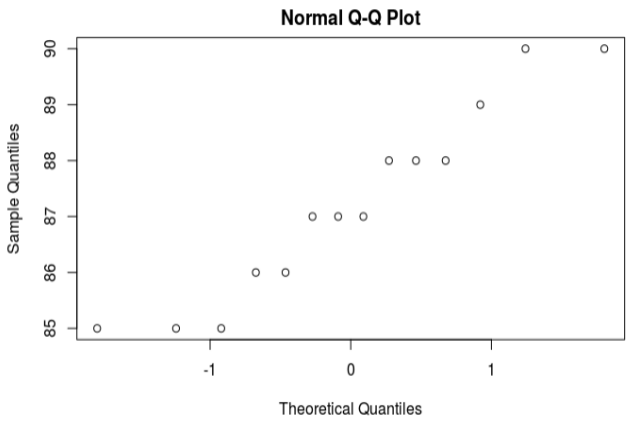

Fig 2. Sauvignon Blanc

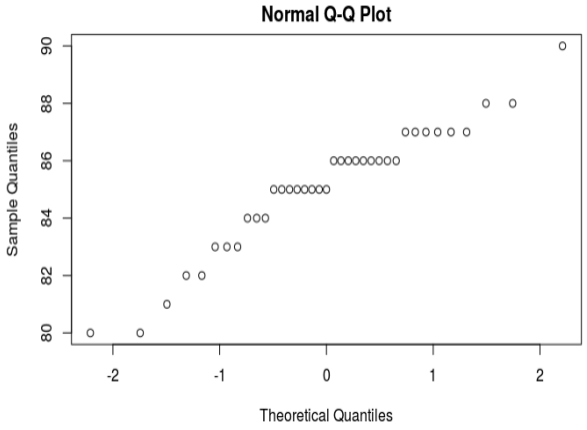

Fig 3. Chardonnay

The above QQ plots show that the data follows a linear trend and hence we can say that it is normally distributed. In addition to this, the Shapiro-Wilk test can be used to check if the data satisfies the condition of normality.

```
> shapiro.test(Chardonnay$points)

        Shapiro-Wilk normality test

data:  Chardonnay$points
W = 0.94394, p-value = 0.06162

> shapiro.test(sauvignon_blanc$points)

        Shapiro-Wilk normality test

data:  sauvignon_blanc$points
W = 0.92273, p-value = 0.2407
```

The null hypothesis for the Shapiro-Wilk test is that the data sample is normally distributed. From figure 4 we can see that the value of p is greater than the significance level(0.05) for both the samples. The null hypothesis holds and we can say that the data is normally distributed.

Fig 4. Shapiro-Wilk Test

To verify that the population must have the same variance, the F-test is used. The null hypothesis for F-test states that there is no difference in the variance of the population.

```
        F test to compare two variances

data:  points by variety
F = 1.6452, num df = 36, denom df = 13, p-value = 0.3372
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5877808 3.7659436
sample estimates:
ratio of variances
          1.645237
```

Figure 5 shows the F-test result for the population for both the wines from the dataset. The results show that the value of p is more than the significance level and therefore we can say that both the wines have the same variance. The null hypothesis holds, and all the conditions for the T-test are satisfied now.

Fig 5. F-test

To check if there is any difference in the mean of points for Sauvignon Blanc and Chardonnay wines, Two-sample T-test will be used for the analysis. The null hypothesis for this test is that there is no difference between the mean of the points for both the wines.

```
>   t.test(points ~ variety, data = new_df, var.equal=TRUE)

        Two Sample t-test

data:  points by variety
t = -3.2599, df = 49, p-value = 0.00203
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.4482245 -0.8181847
sample estimates:
     mean in group Chardonnay mean in group Sauvignon Blanc
                     85.08108                      87.21429
```

Fig 6 Two-Tailed T-test

From the above figure, it can be seen that the value of p is less than the significance level (0.05) and hence the alternate hypothesis holds which states that there is a difference between the mean. But to examine which of the wines has the value of mean greater than the other, One-tailed T-test will be used for this inequality. The null hypothesis for this test says that the mean for the points for both wine types has no difference.

```
>   t.test(points ~ variety, data = new_df, alternative = "less", var.equal=TRUE)

        Two Sample t-test

data:  points by variety
t = -3.2599, df = 49, p-value = 0.001015
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.036108
sample estimates:
     mean in group Chardonnay mean in group Sauvignon Blanc
                     85.08108                      87.21429
```

Fig 7 One-Tailed T-test

As per the One-tailed T-test, the value of p is equal to 0.001015 which is less than the significance level of 0.05 and therefore the alternative hypothesis holds that states that the first mean is less than the second one. Hence, we can say that the mean of Chardonnay is less than the Sauvignon wine that implies that Sauvignon Blanc is better than the Chardonnay wine.

As the alternative hypothesis holds, the confidence interval based on T-test from figure 7 says that there is a 95% chance that the Sauvignon Blanc will be better than Chardonnay wine.

The next analysis is based on the Gibbs Sampling that finds the probability of Sauvignon Blanc is better than the Chardonnay wine. In this method, data is generated from the distribution of both the wines and then calculating the difference of the means for each simulation. The same filtered data used for the previous analysis will be used for this analysis.

Gibbs sampling is a repetitive sampling process that compares the mean based on five parameters. The parameters used for the Gibbs sampling are mean of the overall data, the precision of data, the difference between the sample mean and the α and β parameters considered as 1 and 50 respectively.

Mean of the overall data – $\mu = 85.6667$
Precision τ = 1/Variance = 0.19280
Difference between the sample mean $\mu_0 = 2.133$

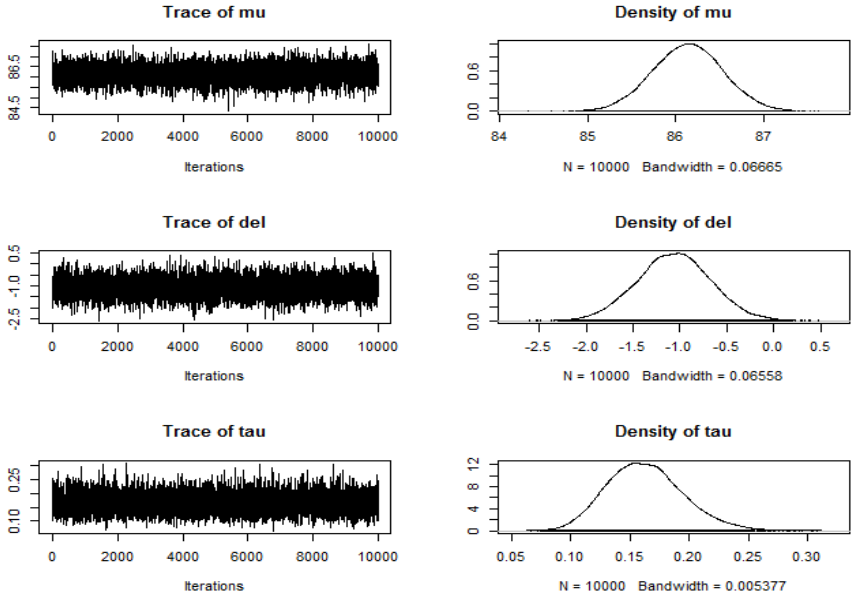

Fig 8 Gibbs Sampling

Figure 8 shows the traces and density of the parameters, μ, τ and delta δ. The number of iterations ran for Gibbs sampling with these parameters was 10000. The density curve is normal and the trace is also not so wide that means all, μ, τ and δ follow normal posterior distributions.

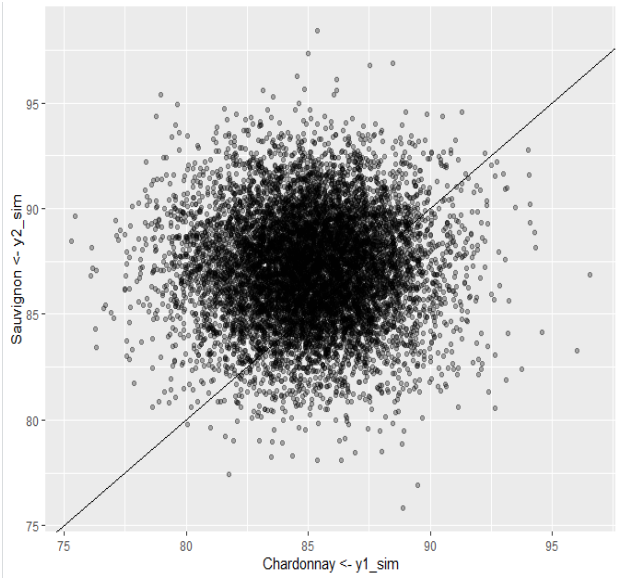

Fig 9 Simulated Samples

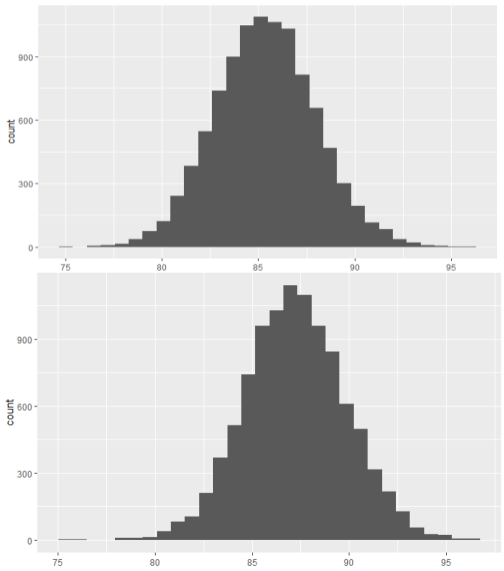

Fig 10 Point Distributions

The scatter plot shown above is a representation of simulated values of point for both wines and the divides the points of both the points. It can be seen that the cluster is slanted towards the Sauvignon Blanc wine. Also, from the point distributions charts, we can see that the Chardonnay wine has high variance than the Sauvignon Blanc wine.

The probability for the ratings is calculated based on each iteration from simulation data. For one iteration, if the value Sauvignon Blanc was better than Chardonnay wine it will result as 1 and 0 otherwise. The mean of these calculated values resulting in 1 came to be 0.723922. Therefore, we can say that the probability of Sauvignon Blanc wine will be better than the Chardonnay wine is approximately **72.39%** at the price of €15.

Conclusion

With the help of T-test, we can say that Sauvignon wine is better than the Chardonnay wine with the confidence interval is 95%. For the considered sample of the dataset, the mean of Sauvignon wine is better than Chardonnay wine by 2.10. By Gibbs sampling, we proved that the probability of Sauvignon wine being better than Chardonnay wine is 72.39%.

1b. Consider the Italian wines in the dataset. Which regions produce better than average wine? Limit your analysis to wines costing less than €20 and to regions which have at least four such reviews.

Data Analysis

As the analysis is based on Italian wines, the first step would be to take a subset of data that only consists of wines from Italy. The next filter would be to subtract the wines whose cost is more than €20. The number of reviews is grouped based on the regions and then the observations are discarded that do not have 4 reviews.

After filtering the dataset as required, the final subset of data has 152 regions. To compare the points for these wines, boxplot is used and to check the distributions of points for this data, a histogram is plotted. As we can see from figure 11 & 12, most of the wines from Italy have ratings ranged from 85 to 88.

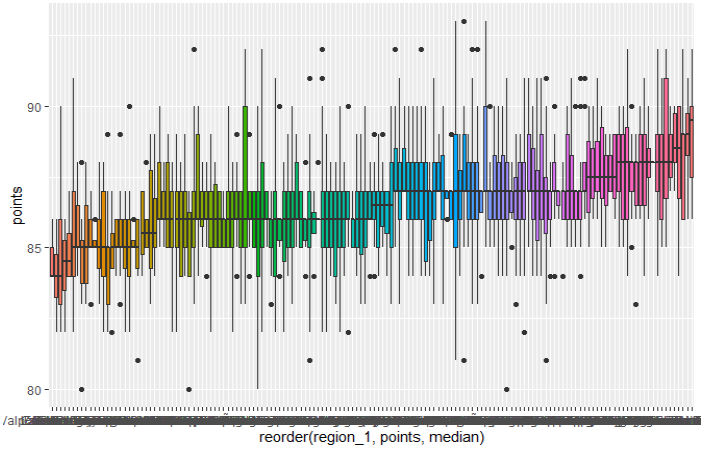

Fig 11. Comparison of Wine Ratings

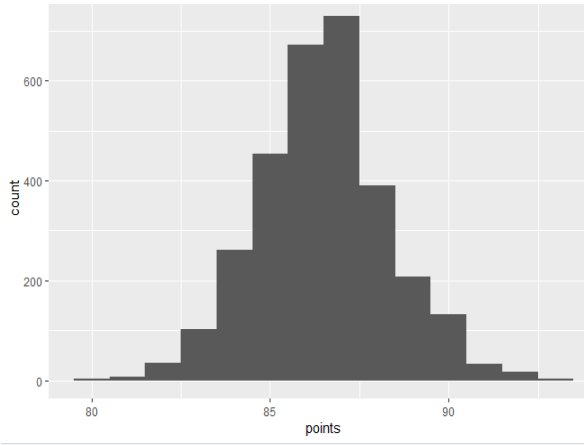

Fig 12. Points Distributions

Figure 13 gives insights into the points for the sample size of the selected regions. The points are taken as the mean of points for the sample. We can deduce from the plot the regions with small sample size tend to have a higher mean score of the points. The variation in the size of the sample leads to different mean points. To estimate this difference, the sample size for every population should be considered and we will have use Gibbs sampling for modelling posterior.

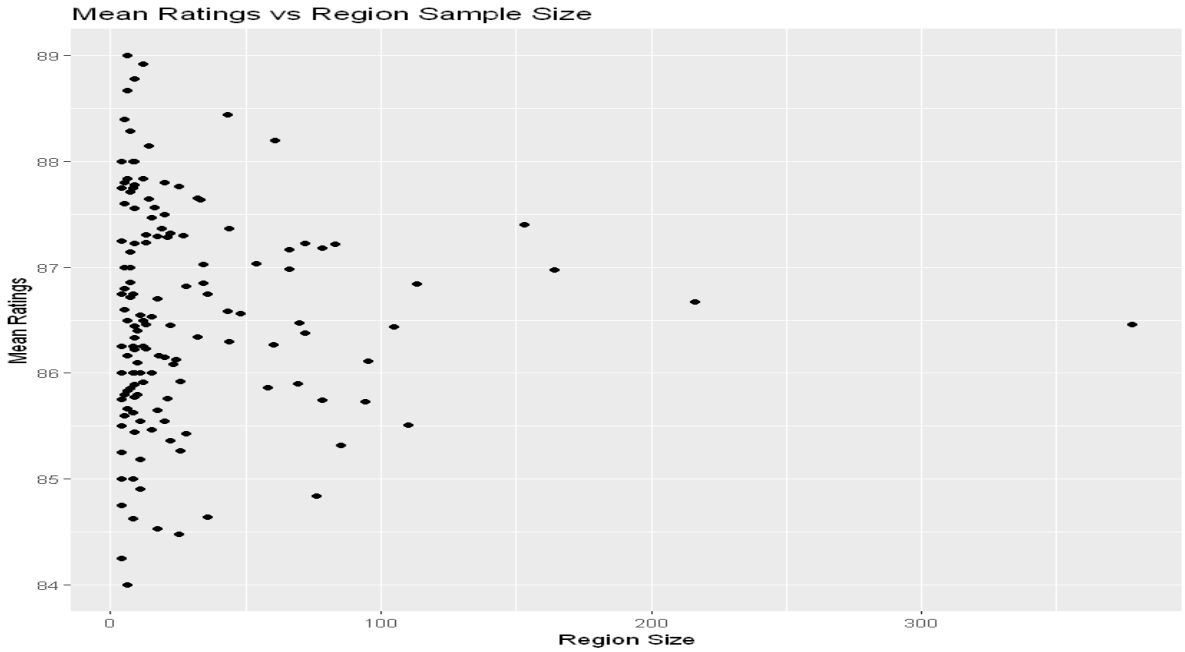

Fig 13. Mean Scores For Each Region (Sample Size)

Using Gibbs sampling to generate the posterior for all 152 regions needs input parameters as the mean across regions (μ), the precision of variance between regions (τb), the precision within the region (τw), and mean point for a specific region k (θk). The number of iteration for the Gibbs sampling is 10000 and the mean & standard deviation is set as 50 & 20 respectively. The Gibbs model returns two objects as params and theta. Params object is for holding the value of μ, τb, and τw and theta hold the value for sample sets of 152 regions.

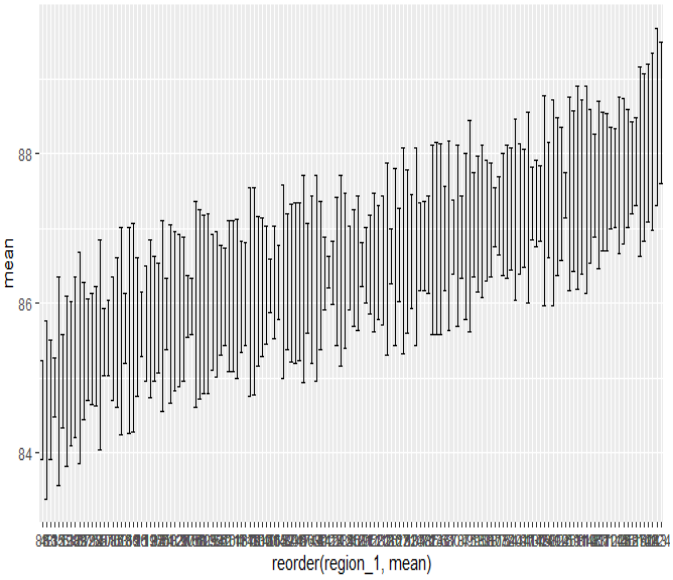

Fig 14 Mean bounds for all regions

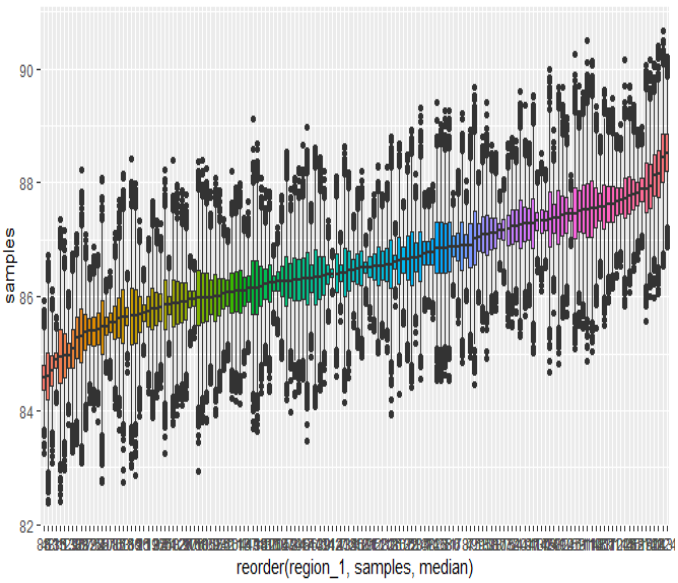

Fig 15 Boxplot for each region

Conclusion

The total average points for all Italian regions is 86.479 and the regions having wine ratings better than this can be found out using :

```
ans <- sorted > avgpoints
view(names(sorted[ans]))
```

'Aglianico del Vulture' 'Alto Adige' 'Alto Adige Valle Isarco' 'Asolo Prosecco Superiore' 'Barbera d\'Alba' 'Barbera d\'Asti' 'Barbera d\'Asti Superiore'
'Bardolino' 'Bardolino Chiaretto' 'Bardolino Classico' 'Bolgheri' 'Campi Flegrei' 'Cannonau di Sardegna' 'Carignano del Sulcis' 'Carmignano'
'Castel del Monte' 'Cerasuolo d\'Abruzzo' 'Cerasuolo di Vittoria' 'Cerasuolo di Vittoria Classico' 'Chianti Classico' 'Chianti Rufina' 'CirÃ²'
'Colline Novaresi' 'Collio' 'Conegliano Valdobbiadene Prosecco Superiore' 'Dogliani' 'Etna' 'Falanghina del Sannio' 'Fiano di Avellino'
'Friuli Colli Orientali' 'Greco di Tufo' 'Irpinia' 'Isola dei Nuraghi' 'Lambrusco di Sorbara' 'Lugana' 'Maremma' 'Maremma Toscana' 'Molise'
'Monica di Sardegna' 'Montefalco Rosso' 'Montepulciano d\'Abruzzo Colline Teramane' 'Morellino di Scansano' 'Nebbiolo d\'Alba' 'Offida Pecorino'
'Orvieto Classico Superiore' 'Primitivo di Manduria' 'Prosecco di Valdobbiadene' 'Roero' 'Romagna' 'Rosso del Veronese' 'Rosso di Montalcino'
'Rosso di Montepulciano' 'Salice Salentino' 'Sant\'Antimo' 'Sardinia' 'Soave Classico' 'Soave Classico Superiore' 'Toscana' 'Trento' 'Umbria'
'Valdobbiadene Prosecco Superiore' 'Valpolicella Classico Superiore Ripasso' 'Valpolicella Ripasso' 'Valpolicella Superiore Ripasso'
'Verdicchio dei Castelli di Jesi Classico Superiore' 'Verdicchio di Matelica' 'Vermentino di Gallura' 'Vermentino di Sardegna'
'Vernaccia di San Gimignano' 'Veronese' 'Vigneti delle Dolomiti' 'Vino Nobile di Montepulciano' 'Vittoria'

Fig 16. Final Result

The regions mentioned above (73) have better ratings than the average Italian wine.

2. Use model-based clustering methods to categorise the wines from the USA based on price and points rating. Can you identify any clusters that are good value for money?

Answer

The aim of this cluster analysis is to summarise data that allows representing the data based on the points rating and price for the wines from the USA in order to identify the clusters that have good ratings and less pricing. The cluster analysis is based on the method of the model-based clustering.

Model-based clustering

Traditional heuristic-based methods such as K-means or hierarchical clustering develop clusters based on the data given as input. They do not consider any measure of probability or uncertainty while deriving the clusters. To overcome this, the model-based clustering algorithm takes the probability into the account while assigning the clusters. It also identifies the optimal number of clusters based on the data.

The analysis of wines data is done using Mclust package in R which is based on the finite Gaussian mixture model. The finite Gaussian models are estimated with the use of a statistical method named Expectation–maximization that finds the maximum likelihood or posterior estimates for parameters in the model.

After loading the data into the R workbook, to get insights into the data, the points and price for the wines from the USA are plotted. The below blot shows the distribution of price and points.

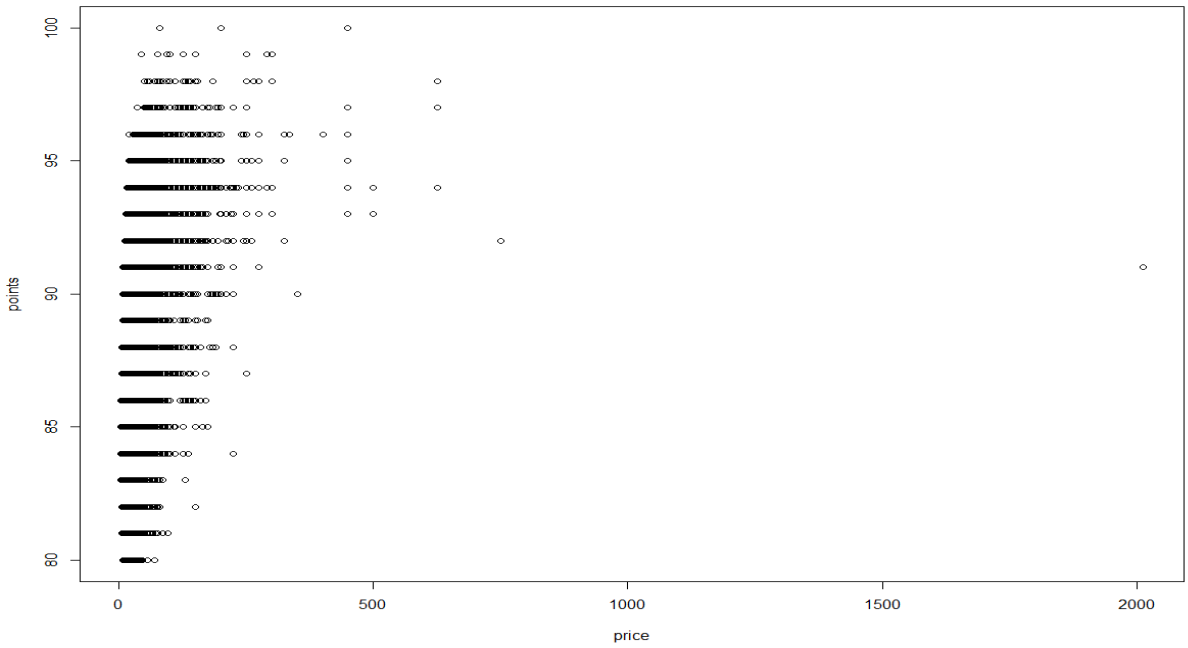

Fig. 17 Price vs Points

The above plot shows that there is no correlation between the points and price columns of the dataset. Also, we can deduce from the visualization that there are a few outliers that may affect the cluster formation and have to be removed before applying the Mclust() function.

After removing the outliers, the plot looks like :

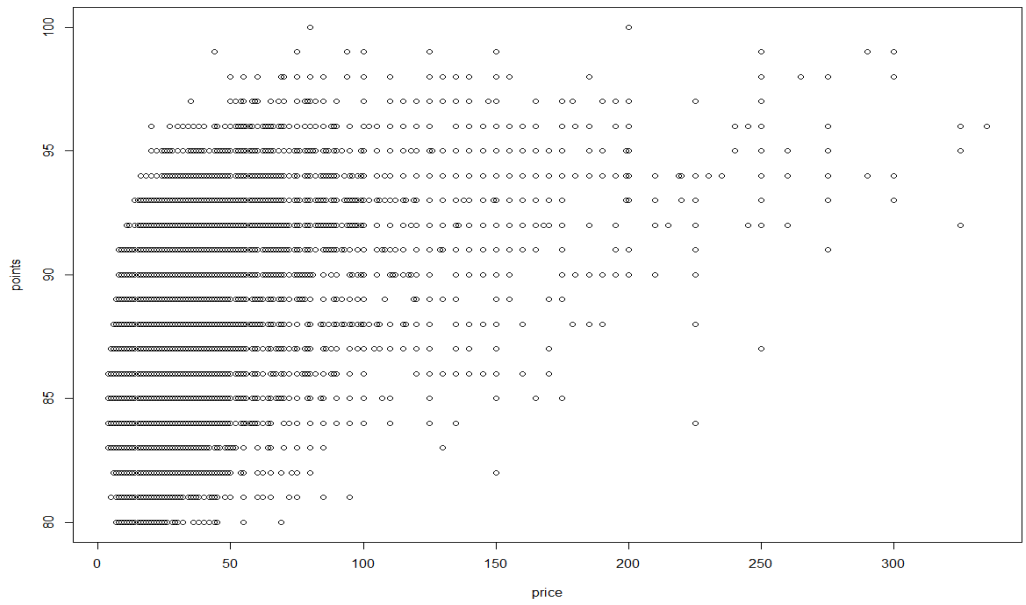

Fig. 18 After Removing Outliers - Points vs price

Now the data looks more clear and distinct for the cluster analysis. The data can now be passed to the Mclust function to fit a model on this data. It uses all the default values to fit a mixture model based on normal distributions using the EM algorithm. The number of mixture components, i.e, clusters, denoted by G, is set to 1:9 as default while fitting the model. To determine the value of G and the covariance structure suitable for the data, the Bayesian Information Criterion (BIC) is used. The BIC function is based on the likelihood function and the model having the highest BIC value is considered to be the best for the data. First, BIC allows us to visualize the models based on the data.

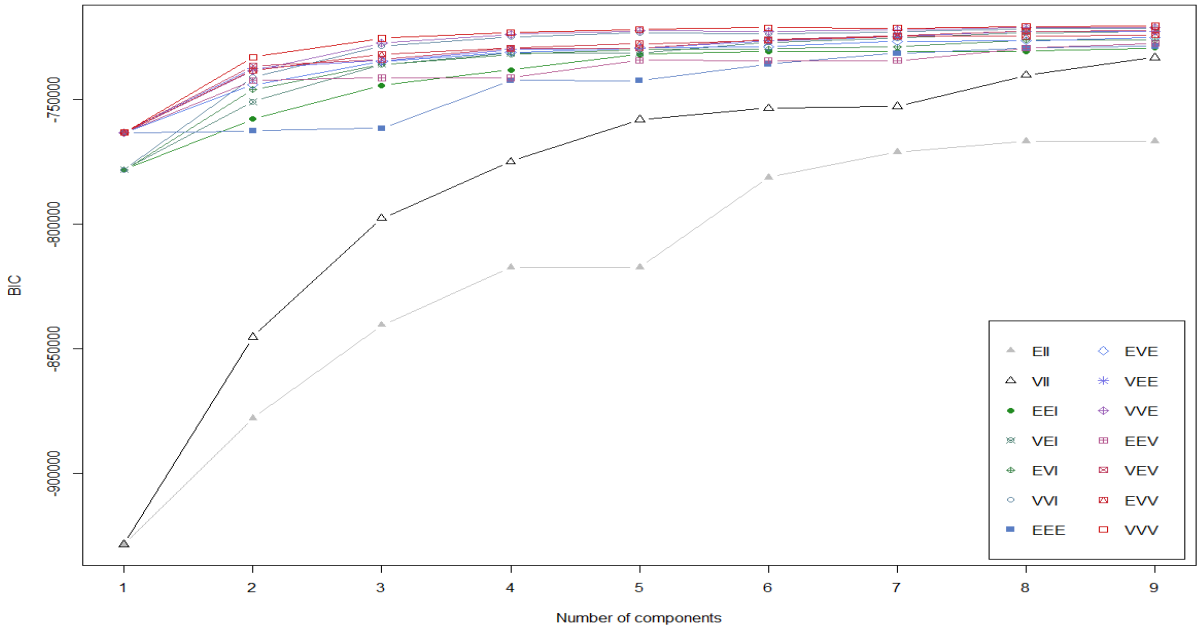

Fig. 19 BIC for the data

The above plot depicts the different covariance structures, the number of clusters and their BIC value. The line charts are the covariance structures highlighting the BIC value for each group size. To describe the BIC in detail we summarise the data with the BIC values.

```
Bayesian Information Criterion (BIC):
         EII       VII       EEI       VEI       EVI       VVI       EEE       EVE       VEE       VVE       EEV
1 -928388.9 -928388.9 -778186.5 -778186.5 -778186.5 -778186.5 -763326.1 -763326.1 -763326.1 -763326.1 -763326.1
2 -877818.3 -845269.5 -757854.6 -750971.8 -746025.9 -741024.4 -762556.3 -744135.2 -737535.0 -738498.9 -762576.4
3 -840485.0 -797697.9 -744352.9 -736048.1 -735954.6 -728516.2 -743090.2 -734823.7 -734484.8 -726954.7 -741447.6
4 -817318.7 -774941.5 -738152.3 -731479.6 -731990.9 -725191.8 -742531.1 -731471.4 -730025.5 -723924.6 -741325.1
5 -797007.5 -758068.9 -731979.7 -728681.3 -729912.1 -723357.2 -735974.1 -729276.1 -726399.6 -725460.2 -734413.5
6 -781343.4 -753500.1 -730767.8 -727107.8 -728461.6 -723741.7 -735991.4 -728442.9 -726482.5 -722889.1 -734456.5
7 -771128.9 -740902.6 -730949.1 -724319.5 -726957.0 -722125.2 -729305.7 -726529.6 -724033.0 -721619.8 -734505.0
8 -765675.8 -739896.2 -730898.4 -723353.9 -727008.6 -721378.5 -729336.3 -726651.6 -722723.7 -721329.7 -734541.4
9 -766766.5 -739868.9 -729481.3 -722920.6 -726872.3 -720972.7 -728518.0 -726808.3 -722384.7 -720818.4 -729528.3
         VEV       EVV       VVV
1 -763326.1 -763326.1 -763326.1
2 -736456.7 -738328.1 -733053.2
3 -733941.2 -731849.9 -725563.5
4 -729711.5 -729367.3 -723069.2
5 -726606.7 -727928.3 -721613.3
6 -726027.8 -726414.8 -722391.4
7 -723665.7 -726346.0 -721069.2
8 -722591.9 -726390.7 -720927.9
9 -722407.5 -724491.5 -720625.1

Top 3 models based on the BIC criterion:
    VVV,9     VVE,9     VVV,8
-720625.1 -720818.4 -720927.9
```

Fig. 20 BIC Description

Based on the above data, it can be seen that the models with variable volume, shape and orientation that assume general clusters show the highest BIC with group size 9. Models that assume spherical and diagonal clusters does not fit the wines data properly. The top three models that suit this data based on BIC analysis are VVV with group size 9 or 8 and VVE with size 9.

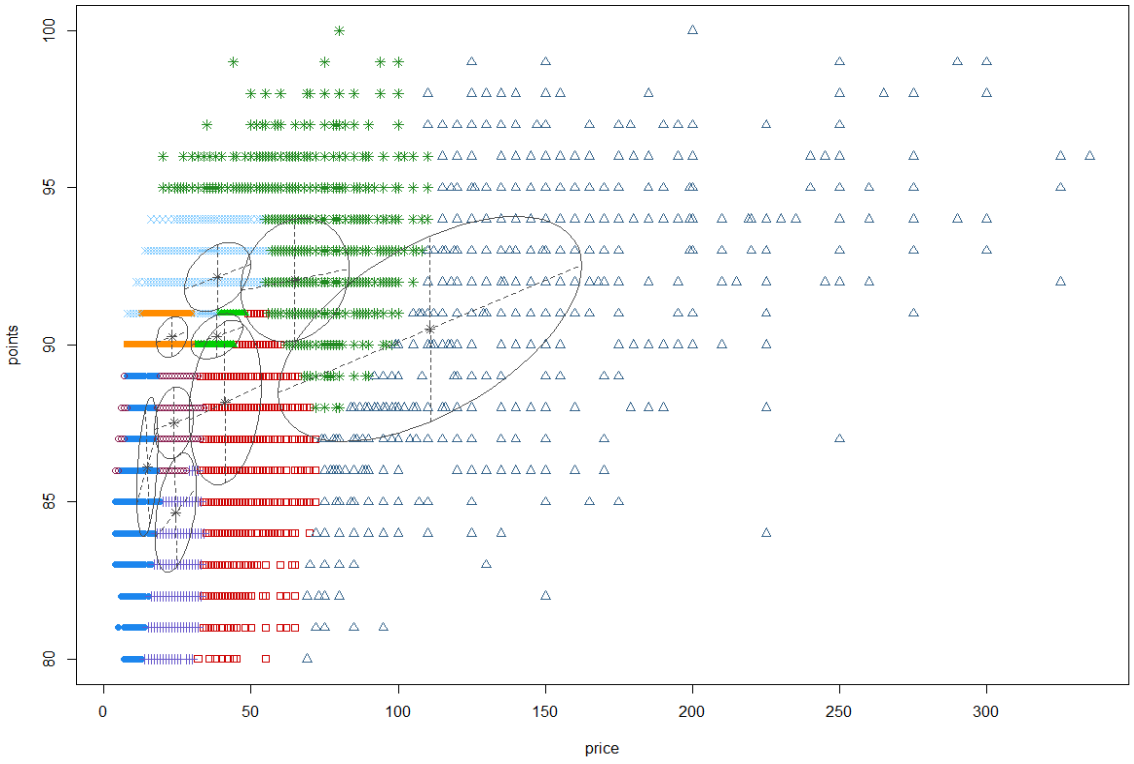

Fig. 21 Classification

The above plot is a classification of the observations based on the new model based on group size 9 and model name as VVV. The chart below shows the assignment uncertainty level for each of the cluster and observations.

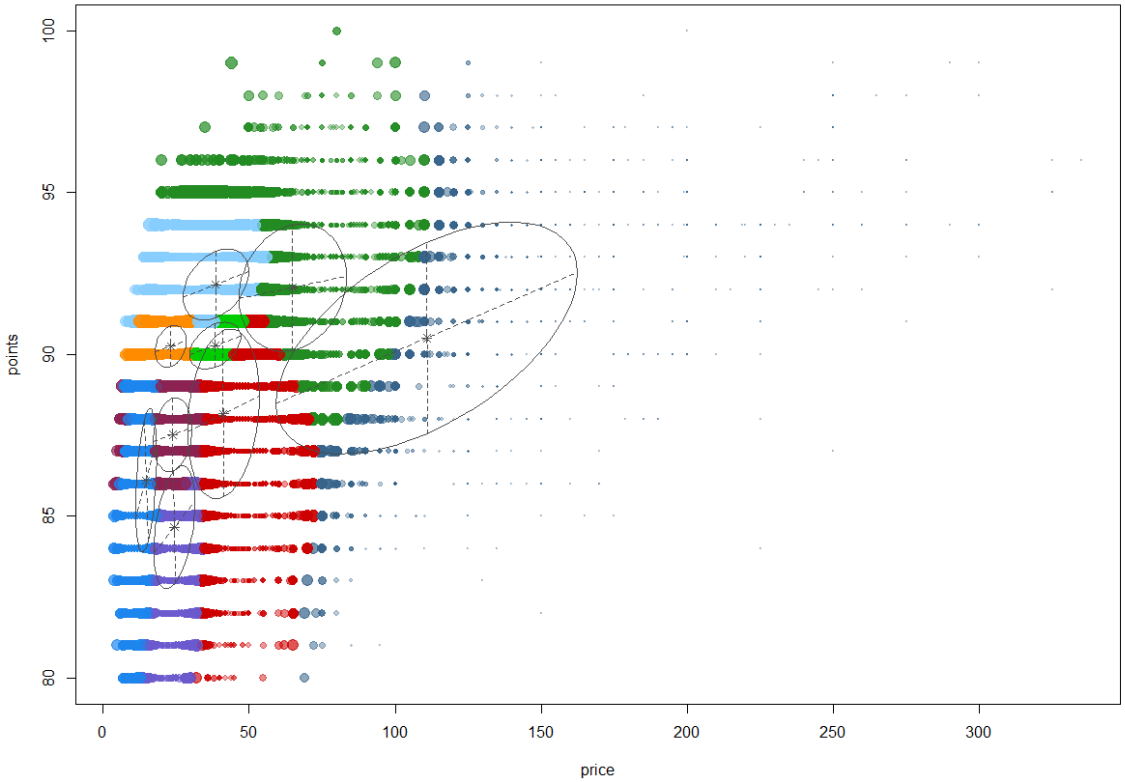

Fig. 22 Uncertainty

The summary of the final model is :

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 9 components:

 log-likelihood     n df       BIC       ICL
      -360162.8 54249 53 -720903.4 -772419.4

Clustering table:
    1     2     3     4     5     6     7     8     9
 8643 10147  2747  4464  4706  6637 10016  5535  1354
```

Fig 23 Summary of the model

Conclusion

Based on the above data and the classification visualization, we can easily identify the clusters that are good value for money. Wines having a high rating in terms of points and cost less than others can be considered as good value for money. The cluster highlighted by the cyan colour can be called as the best value for money cluster amongst all as it has observations having a higher rating with less price. Second to that, cluster denoted by the orange colour also has the wines worth their price. Adding to the list, the clusters having a good combination of price and points are underlined by the lime green and the dark green colour.