Main Assignment

CS7DS3 Applied Statistical Modelling

Patrick Lee 15325692

Introduction

This report addresses the questions posed in the CS7DS3 Applied Statistical Modelling Main Assignment. Each question involves analysis of a Kaggle dataset containing thousands of wine reviews, the data can be found at: *https://www.kaggle.com/zynicide/wine-reviews*. The dataset contains multiple files with a varying number of reviews and different file formats. The file used for each section of this report is entitled: *winemag-data_first150k.csv*. The data contains lots of information for each wine, including the text review and the reviewer name, but in this report the point score and price is the focus for comparison. Each wine is given a rating of 0-100, however, the majority of the wines lie in the 80-100 range.

The report is split into three sections: (1) that answers question 1 parts a(i) and a(ii) and compares South African Sauvignon Blanc with Chilean Chardonnay, (2) that answers question 1 part b and investigates which regions in Italy produce above-average wines, and (3) which addresses question 2 part b and uses clustering methods to categorise the wines from the USA and compare them based on their price and point scores. Each section will follow the same layout, denoted by the following headings: Data Handling, Analysis and Conclusion. Each section will begin with an explanation of the question that will be addressed and will layout the steps taken in the analysis. Data Handling will discuss how the original 150k dataset was reduced for the particular analysis and the dimensions of the subset to be used. The Analysis section will detail each step, discuss the thought process behind some design decisions, and present the results with explanations. Each section will finish with a Conclusion, that will give a summary of the results of each step and contain the answers to the assignment questions.

Section 1

Stand-off in the Southern Hemisphere

In this section, Sauvignon Blanc produced in South Africa will be compared to Chardonnay produced in Chile. Two questions will be answered: Which wine variety is rated better and by how much? What is the probability a Sauvignon Blanc will be better than a Chardonnay? €15 is asserted to be the right amount to spend on a bottle of wine for this question, therefore the majority of the investigation will be performed on the wines of each variety that are exactly €15. The analysis contains three elements: a visualisation, a t-test, a Gibbs sampler.

Data Handling

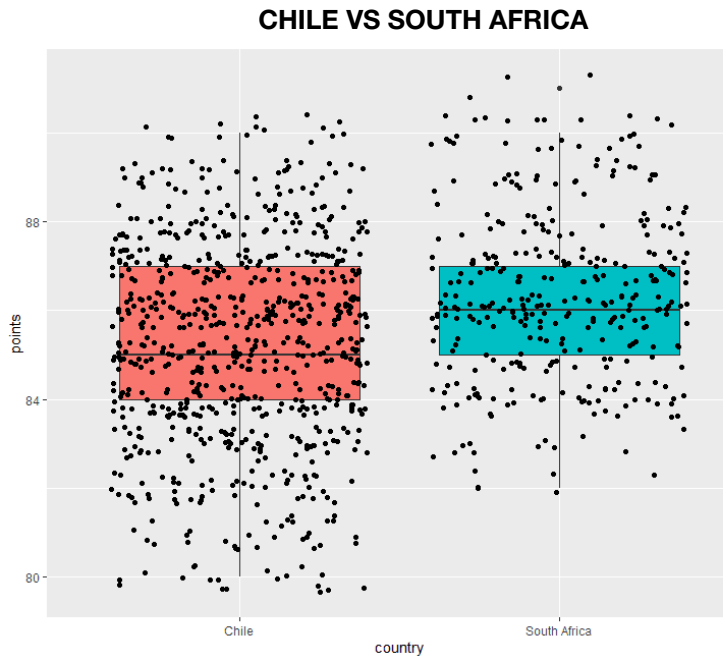
For this analysis, the 150k dataset was reduced to a subset of data items with Variety = Sauvignon Blanc and Country = South Africa, or Variety = Chardonnay and Country = Chile.
The subset was reduced further by removal of unnecessary columns. The remaining columns were price, points, variety and country. Any data items with missing values were removed from the subset. A separate subset was then created to extract the wines equalling €15 in price and this will be used for the majority of

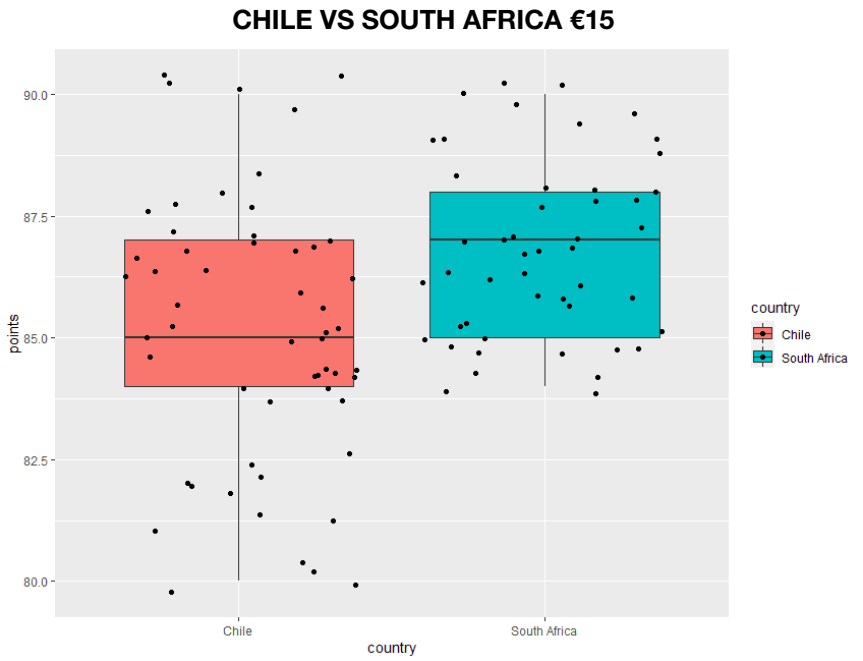
the analysis. The resulting subsets contained 1082 wines for all prices and 103 wines at exactly €15. It is assumed that the data follows a normal distribution.

Analysis

Visualisation

The first step of this analysis was to visualise the data and compare the scores of both wine subsets. This was conducted by graphing the data as a box plot first before including a jitter plot over the top.





The plots suggest that South Africa and the Sauvignon Blanc generally have higher scores than Chile and the Chardonnay, with almost all scores equal or better than the median of Chile for the wines costing €15. South Africa also visibly has much less variation in its scores, suggesting that its wines are of a similar high quality. There is also a very similar number of wines in this dataset (55 Chile, 48 South Africa) for each country, which gives a fairer comparison.

	Mean	Median	Standard Deviation	Range
Sauvignon Blanc	86.81	87	1.81	84 - 90
Chardonnay	85.07	85	2.77	80 - 90

The difference becomes clearer when comparing the values of the mean, standard deviation, and median directly. South Africa's mean is ~1.74 points greater and the median shows a 2 point increase. The standard deviation and range is also telling, suggesting that any Sauvignon Blanc chosen at 15 euro will provide a similar quality taste whereas the Chardonnay might be more hit or miss.

T-test

A t-test can be performed to account for the variability in the data. As each review is conducted by a different reviewer with their own taste, agenda, and opinions on the quality of wine there will be an

arbitrary element to the scoring. A way to account for variability is to perform a t-test. Above is the results of the t-test for this data, it determines whether there is a significant difference in the data. The indicator is the p-value which if $p > 0.05$ then there is no significant difference. In this case, $p = 0.00035$ is much less than the threshold and therefore we can conclude that the scores are different enough to suggest that the Sauvignon Blanc is indeed better.

```
        Two Sample t-test

data:  points by country
t = -3.7043, df = 101, p-value = 0.0003459
alternative hypothesis: true difference in means is not
 equal to 0
95 percent confidence interval:
 -2.6714489 -0.8080965
sample estimates:
       mean in group Chile mean in group South Africa
                  85.07273                   86.81250
```

Gibbs Sampler

The final step is a more explicit comparison between the mean of each wine type. For this analysis the means (θ_x) are defined as such:

- $\theta_1 = \mu + d$
- $\theta_2 = \mu - d$

Where θ_1 represents the mean of the Sauvignon Blanc, θ_2 represents the mean of the Chardonnay, μ is the mean of the whole subset, and d is the difference in the subset means. Determining the exact posterior for μ and d with respect to the data is not possible but can be estimated using a Gibbs Sampler, given the assumption that the data is normally distributed.

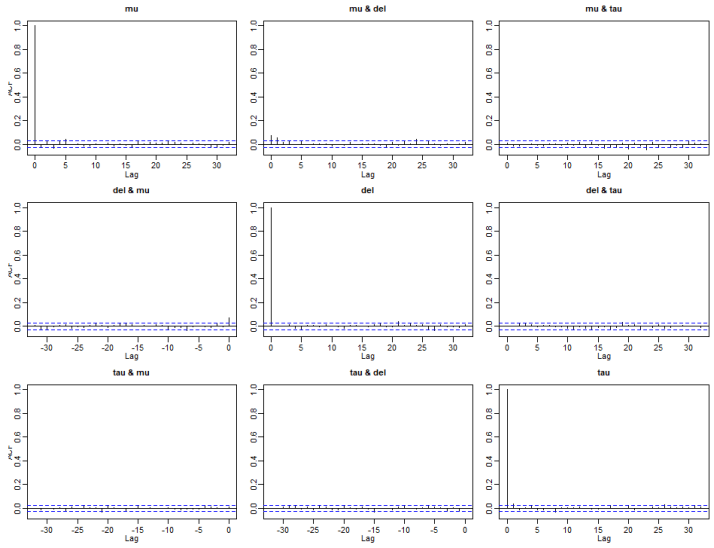
Gibbs Sampler Hyperparameters

Name	Value
μ_0	85.88
a_0	1
b_0	10
γ_0	1/25
τ_0	1/25

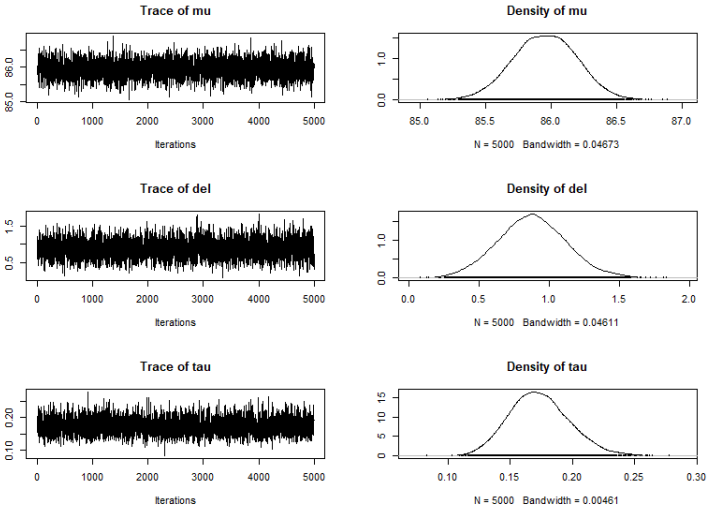
The table below shows some of the hyperparameters for the Gibbs Sampler and the values they were set to before running the sampler. The mean was selected as a calculation from the overall mean of the dataset, where a_0 and b_0 were selected for confidence in the prior knowledge of the data and its deviation respectively. This is a more narrow prior, which makes sense given the range of the data.

Visualising some output of the sampler and collecting some information can help to measure its performance. The Burn-in represents improbable initialised values for the hyperparameters that should not be included, so the lower the burn-in value the better. The dependence factor looks at the thinning required for the model, which is the process of removing samples that are too highly correlated and could bias the model. This value should be close to 1, which is the case for this sampler and the burn-in is low,

AUTOCORRELATION PLOTS



TRACE PLOTS



	Mu	D	Tau
Burn-in	2	2	2
Dependence Factor (l)	1.020	0.999	0.966

suggesting that the sampler performed well. This is also suggested in the autocorrelation and trace plots, as the trace graphs do not show any obvious patterns and the autocorrelation value is low for each term. The output of the Gibbs sampler can then be used to determine the difference between the two means, following the calculation of the differences of θ_1 and θ_2.

Assuming that the data follows a normal distribution then both can be simulated as such using the values of μ, d and σ^2 from the Gibbs sampler. By then taking the proportion of simulated Sauvignon Blanc score samples that are greater than the corresponding simulated Chardonnay sample the probability that the Sauvignon Blanc will be scored higher can be calculated. The result is a 69.9% probability that the Sauvignon Blanc will be better.

Conclusion

It is clear from the analysis that the Sauvignon Blanc from South Africa is, on average, better than the Chardonnay from Chile. The trend was apparent from the first step of visual analysis and comparison of simple values like the the mean, median, and standard deviation. For a more conclusive result a t-test was performed to determine if the difference between the means of each wine was negligible or significant, and the result continued the trend by indicating the latter. Through both steps the Sauvignon Blanc was approximately 1.74 to 2 points better than the Chardonnay. The final step was to answer part a(ii) of the assignment and determine the probability that the Sauvignon Blanc was better, which, by using a Gibbs Sampler to estimate μ and d, was calculated to be 69.9% probability.

Section 2

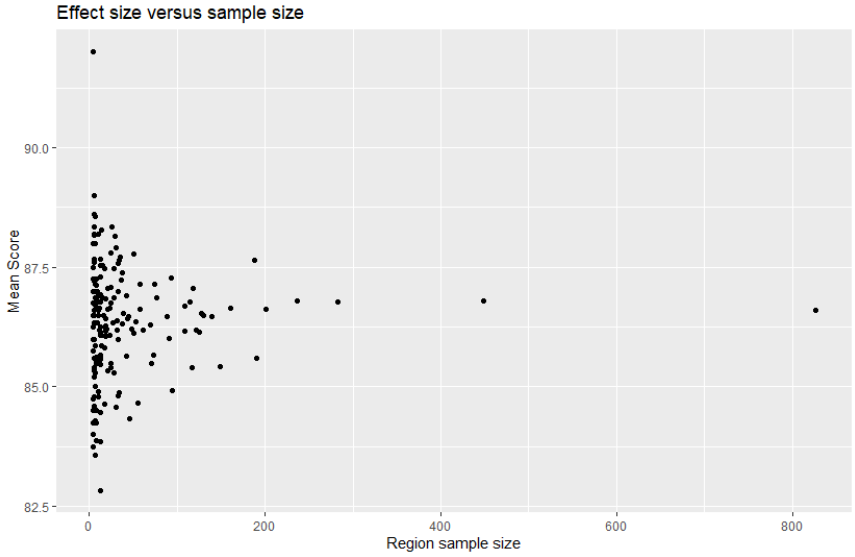
Italian Turf War

This section looks at the wines originating from Italy, specifically investigating the individual regions. The aim is to determine which regions produce above average wine, which this report will assume is in comparison to the full dataset. The analysis undertaken involves an initial visual investigation and then the use of a Gibbs Sampler to estimate the mean of each region.

Data handling

The data items of wines from Italy were extracted from the full dataset for the analysis. As per the instructions, this was reduced further by extracting wines that cost less than 20 euro and then the wines of regions that have at least 4 reviews. The columns isolated were the Region, Price, and Points and any items with missing values were removed. The resulting subset contained 7174 wine reviews. The sample sizes of each region can be seen in the graph below with most regions containing less than 200 reviews. The outliers include *Toscana* with 448 reviews and *Sicilia* with 827 reviews.

SAMPLE SIZE OF REGIONS

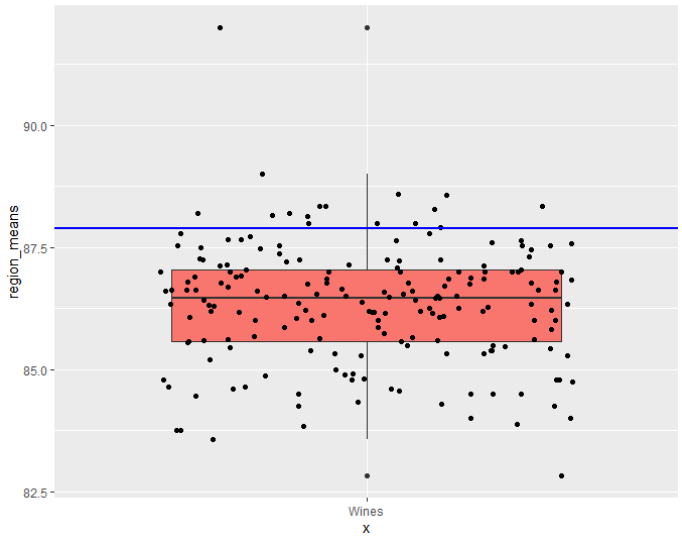


Analysis

Visualisation

The first step is to produce a visual representation of the mean score of each region compared to the mean score in the original dataset. This will give a rough estimate of the number of regions and wines that are above-average compared to the overall average. The mean points of each region is represented by the boxplot and jitter points above with the blue horizontal line representing the mean of the original dataset. This graph shows 16 regions having a higher mean score than the overall mean. This is subject to variation so a more thorough estimation is required.

REGION MEANS VS OVERALL MEAN

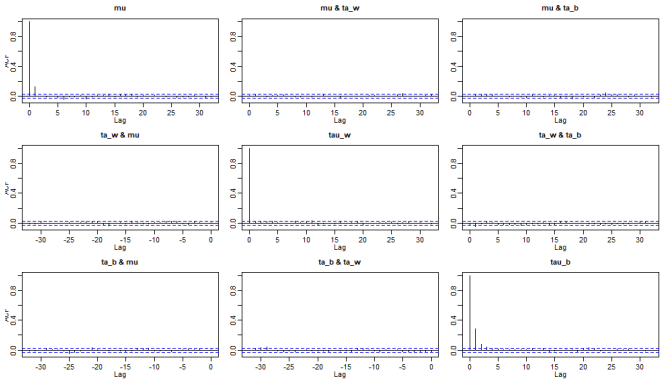


Gibbs Sampler

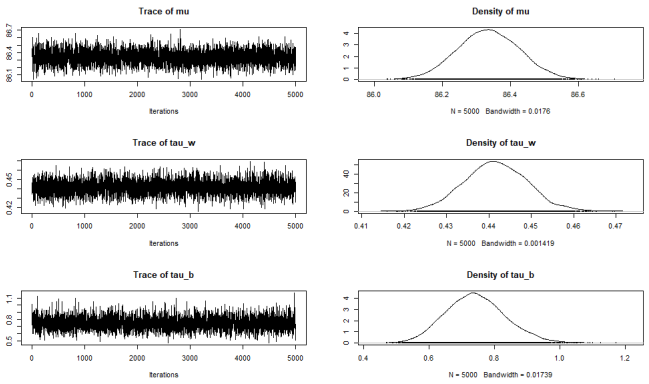
The next step in the analysis was to explicitly model the mean of each region using a Gibbs Sampler, like in the previous analysis of Sauvignon Blanc and Chardonnay. This will allow some variation to be considered and ultimately give a better assessment of the differences between the regions. The following values are estimated from the sampler:

- μ = overall mean across regions
- τ_b= precision (inverse variability) between regions
- τ_w = precision within regions
- θ_m = mean mark of region m

AUTOCORRELATION PLOTS



TRACE PLOTS



	Mu	Tau_w	Tau_b
Burn-in	2	2	3
Dependence Factor (I)	1.05	1.02	1.12

The Gibbs Sampler performance will be assessed in the same way as the previous section and the autocorrelation looks to be low again and the trace graphs are not stuck in a pattern loop. The burn-in and dependence factor values are also favourable. The estimated mean of the entire Italian subset comes to 86.34 compared to 87.88 of the original dataset.

The calculated θ for each region was estimated using the sampler and then displayed in order of highest to lowest mean score. By using the mean of the overall dataset (87.88) as the reference, that includes wines from all over the world, there are only 10 regions out of the 188 Italian regions analysed that can attest to having a higher mean score:

Position	Region	Mean Points
1	Barbaresco	90.30
2	Coline Novaresi	88.50
3	Isola dei Nuraghi	88.20
4	Barbera d'Alba	88.16
5	Soave Classico	88.08
6	Sannio	88.03
7	Primitivo di Manduria	88.03
8	Carignano del Sulcis	87.95
9	Emilia-Romagna	87.90
10	Conegliano Valdobbiadene Prosecco Superiore	87.88

Conclusion

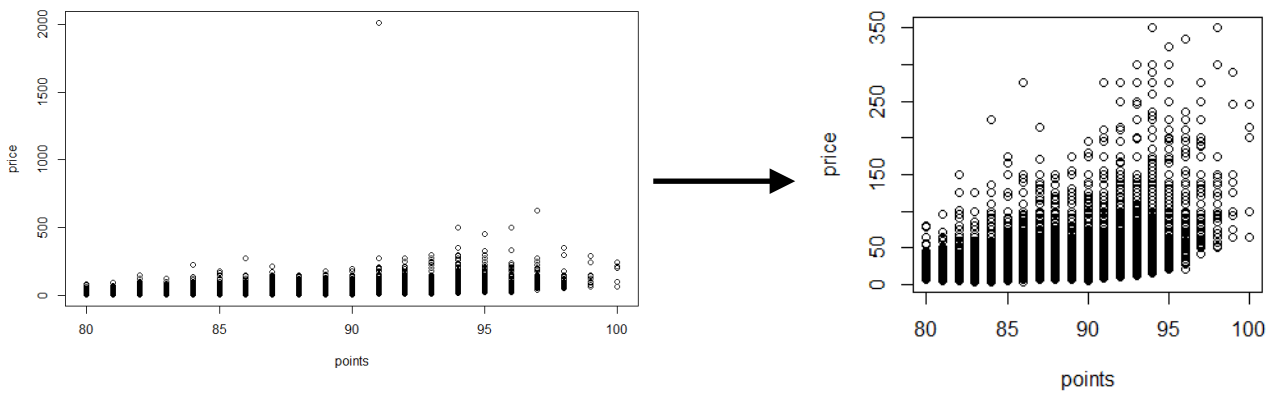
An initial look at the data through a visualisation showed that 16 regions produced above average-wine. Upon more thorough analysis the list was shortened to 10 with a large number sitting just above the overall mean. The representative score of each region was an estimated mean from all the wines they produce and due to the closeness in value to the overall mean it is hard to dismiss other regions or declare the list of 10 significantly better.

Section 3

Bang for your Buck

In this section, a model-based clustering method is used to categorise wines from the USA with respect to their price and points rating. The goal of the modelling is to find clusters that represent good value for money. The model used is from the $Mclust$ package from R, which, using an EM algorithm, fits a mixture of normal distributions to the data. The analysis is split into 3 sections that investigate different numbers of clusters: models with 1 to 9 clusters, 10 to 19 clusters, and then 20 to 29 clusters.

TRIMMING THE DATASET

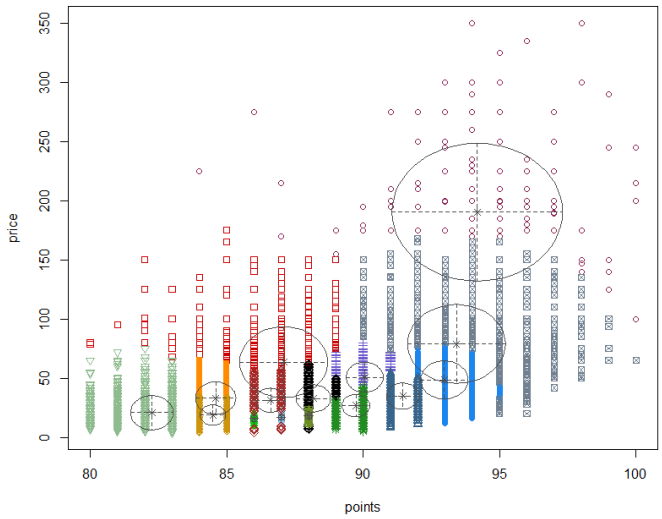


Data handling

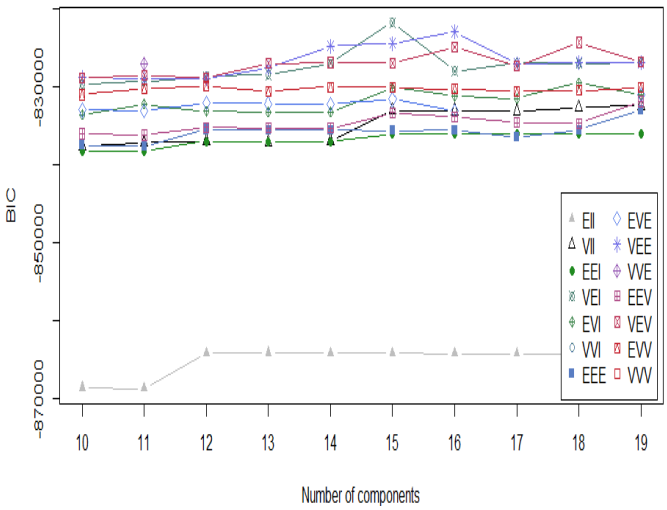
The original dataset was processed to extract the relevant items for this analysis. The final subset includes price and point values for the wines originating from the USA with items containing missing values removed. As the aim of this investigation is to determine which wines deliver quality for an equatable price some assumptions were made to trim the subset further. Wines above 400 euro were rejected and removed from the dataset out of hand. This removed outliers that were significantly more expensive than alternatives but often worse in their points score.

Clusters 10 - 19

CLUSTERS 15



BIC RATING OF MODELS, CLUSTERS 15



The first step is to fit the data using the *Mclust* function in R to generate and identify different clusters. The default of this function is to use 1-9 clusters, but this data requires larger clusters. For this section the limits are set to 10-19. The best fitted model was found to be of 15 clusters with a large number of smaller sized clusters towards the lower priced section of the graph. At a glance, the three groups representing the highest price (Maroon, Red, and Grey) contain the most variation and Red can be dismissed due to its similar cost and worse scoring to Grey and also its similar scoring but higher cost to the cluster beneath it.

The EII structure for the 10-19 cluster groups performs poorly. The best performing models are now VEI with 15 clusters, VEE with 16 clusters, and VEV with 18 clusters.

VEI, 15 Clusters	VVE, 16 Clusters	VVE, 18 Clusters
-821775	-822954	-824333

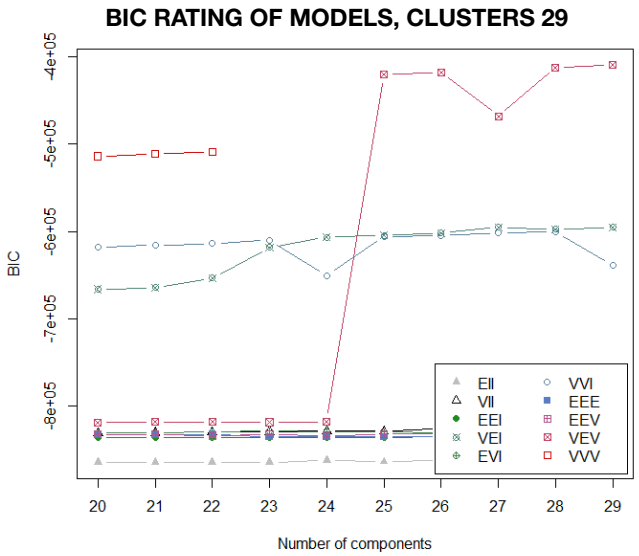
The metric is known as Bayesian Information Criterion, BIC, which will penalise models for having too many clusters. (i.e. being too complex). The model variations that were tested are listed in the legend in the graph below. These represent cluster structures that differ in *volume*, *shape*, and *distribution*. Taking the best performing structure,VEI, as an example: VEI denotes variable *volume*, diagonal *distribution*, and equal *shape*. EII, which is the worst performing structure as shown in the graph below, is circular and fixed orientation but also uses a fixed volume, which would fit the data poorly.

As the better performing models are not close to the limits of 10 or 19 clusters it does not seem likely that more clusters will yield a better score. However, the purpose of this analysis is not to develop a model for accurate prediction and little overfitting but to identify specific groups where little variation is favourable. Therefore, a larger amount of groups could identify more specific and therefore more valuable cluster of similarly priced and scoring wines.

As a references, the large clusters Maroon, Grey, and Red have proportions 0.148, 0.116 and 0.110 respectively of the size of the dataset. This represents around 7000 wine reviews each. The smaller clusters (A-M) have a size range of [~0.02-0.84] with a mean of 0.054, representing around 3355 wine reviews per cluster. This collection of clusters also has a mean uncertainty value of 0.217 across each wine review, suggesting the uncertainty of each review fitting their assigned cluster is quite high.

Clusters 20-29

For this section, the *Mclust* function limit was set to 20-29 clusters. The greater number of clusters means that a legible visual representation is difficult to produce. However, the BIC graph can be produced and shows very little structures that fit the data well. The best fit is VEV for different cluster numbers, with the next best variations, VVI and VEI, some way off. The mean uncertainty of this fit is 0.004 which is a large improvement on 0.217 of Clusters 10-19. The mean size is at 0.034 which is comparable to the smaller clusters of Clusters 10-19. This size represents 2112 wines, although some of the larger clusters contain 4000-6000 wines.



By splitting the clusters into groups depending on point score ranges with means (80-85), (85-90), and (90-100) we can identify some interesting clusters that represent value for money.

VEV, 29 Clusters	VEV, 28 Clusters	VEV, 26 Clusters
-409839	-412480	-417828

Standard (80-85)	1	2	3	4	5	6	7	8
Points	81	82	82	82.4	82.7	83	84	85
Price	20.8	22	75	54	45.1	21	22.5	23.6

Premium (85-90)	9	10	11	12	13	14	15	16
Points	86	87	87	87.9	88	88.3	89	89.5
Price	26.1	80.5	27	33.5	30	185	33.2	126

Extra Premium (90-100)	17	18	19	20	21	22	23	24	25	26	27	28	29
Points	90	90	90.5	91	92	92.5	93	94	95	96	97	98	99.4
Price	94.5	34	187	39	45.8	161	53.2	61.7	61.7	83.4	103.3	104.3	134

In the Standard range, the best value for money would be from Cluster 8 with wines €23.60 euro and a mean points rating of 85.
For the Premium range, Cluster 11 presents good value at around €27 with a mean 2 points improved on Cluster 8 at 87 points. This range also has another interesting option in Cluster 7, with an 89 mean points score and price at €33
In the Extra Premium range, Cluster 7 is out-performed by Cluster 20 that offers an average of 91 points for a mean price of €39. Coming towards the higher end of the point scores the price increases dramatically but a few clusters suggest the high cost could merited with a comparable high quality. Clusters 28 and 29 offer the highest scoring wines with a mean of 98 and 99.4 respectively but for a staggering €104.30 and €134 respectively per bottle. Cluster 25 represents an intriguing middle ground with a mean of 95 points for €61.70 on average.

Conclusion

By limiting the cluster number to 10-19 the resulting groups are too large and varied to be identified as good groups to recommend and this is compounded by the relative uncertainty of each wine's assignment being quite high. The use of a larger number of clusters has reduced the uncertainty and variation within the groups significantly and identified a number of intriguing wine groupings. For wines under €40, Clusters 8, 11, 15, and 20 offer wines of 85, 87, 89, and 91 point scores respectively. The pattern here results in a €6 increase in price equates to a 2 point mean score increase. For wines rated in the high 90s, Clusters 28 and 29 offer 'near perfectly' rated wines for €104-134.