Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Assessment Submission Form

Student Name	Tanvi Bagla
Student ID Number	19300699
Course Title	MSc. Computer Science- Data Science
Module Title	Applied Statistical Modelling
Lecturer(s)	Dr. Arthur White
Assessment Title	Main Assignment
Date Submitted	15-05-2020

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
I declare that the assignment being submitted represents my own work and has not been taken from the work of others save where appropriately referenced in the body of the assignment.

Signed: Tanvi

Date: 15-05-2020

Introduction

The wine review dataset (winemag-data-130k-v2.csv) taken from (https://www.kaggle.com/zynicide/wine-reviews) is analysed. Dataset contains wine reviews, the rating of wine (measured in points) and other relevant information obtained from wine enthusiasts from winemag.com. The data is available in two formats – json and csv.

The objective here is to analyse this data to transform it into some useful information that can be used by non-technical people like wine sellers who would like to use the analysis in qualitative way or by technical managers/supervisors who check the correctness of the analysis done. Statistical methods and models like Gibb's sampling and Bayesian model is used to compare the means of different wines corresponding to different countries in order to find out the best rated wines and their regions. Use of Linear Regression model to estimate the rating (points) of the wines depending on other factors.

The report is divided into two parts Question 1 and Question 2, each having sections like Data Handling, Analysis (Analysis of Q1, Analysis for Q2), Conclusions (Summarize results, overall evaluation, and further recommendations).

CS7DS3 Applied Statistical Modelling
Main Assignment

To be submitted on Blackboard by **5pm Wednesday 29th April**

I would like you to analyse the wine reviews dataset. This dataset is available to download from the class page and from the Kaggle website: https://www.kaggle.com/zynicide/wine-reviews

Please put your analysis in a report (page limit: 10 pages). I would like your report to use the statistical methods covered in CS7DS3 to analyse the following questions:

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?
   b. 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.
2. **EITHER:**
   a. Build a linear regression model to estimate the points value for wines from the USA. Using simple language, identify which factors are most important in obtaining a good rating.
   OR
   b. 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?

Q1.a.i

Data Handling - The wine dataset presents information for each wine on the ratings provided by the reviewers. The nation, size, points, region and variety are some of the main columns to be considered for first question. Country and area determine where the vineyard is situated. Price is the quality of a wine that is sold. Points apply to the ratings each customer has given. Column of variety displays the name of the wines. The data is filtered with rows having Chardonnay wine from Chile and Sauvignon Blanc wine from South Africa with price taken is exactly Euro 15. Rows having few missing values are omitted from the filtered data. Also, to treat variety variable as an index value and not as a measurement as.factor() function is used.

Data Exploration

Summary of the data is presented as below:

```
summary(data_1)
```

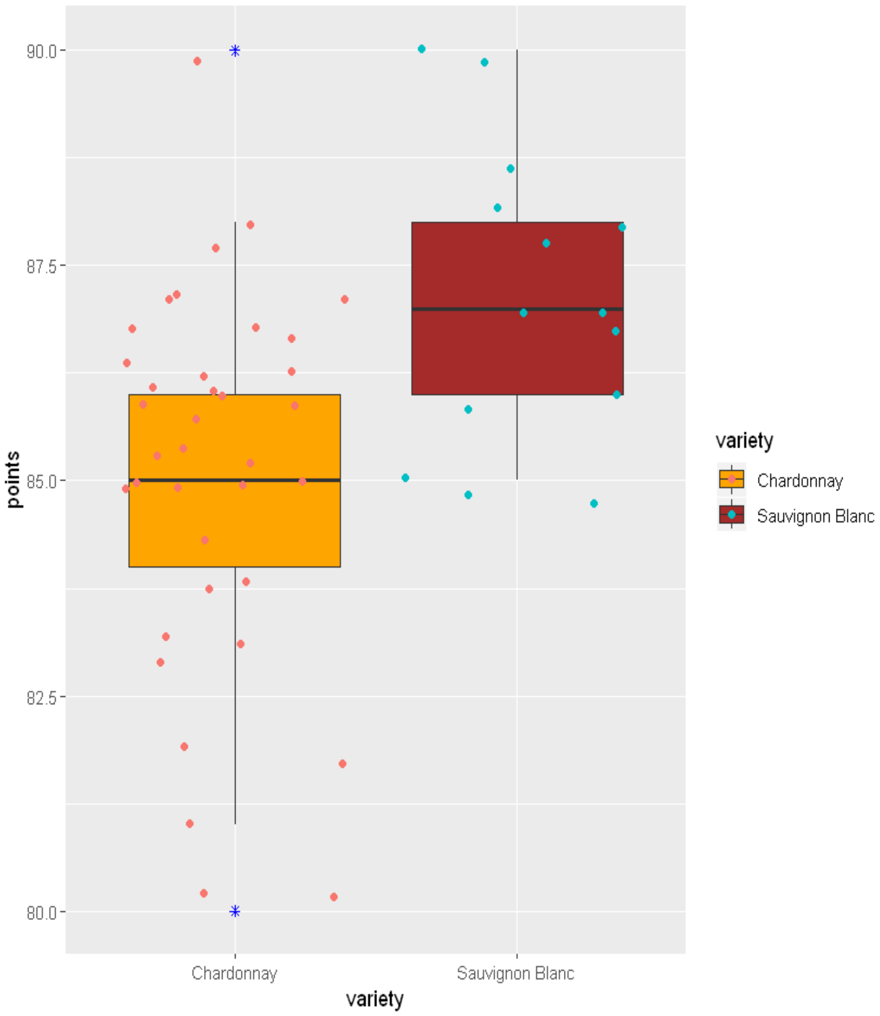
```
          country        points                   variety  
 Chile        :37   Min.   :80.00   Chardonnay      :37  
 South Africa:14   1st Qu.:85.00   Sauvignon Blanc:14  
             : 0   Median :86.00                        
 Argentina    : 0   Mean   :85.67                        
 Armenia      : 0   3rd Qu.:87.00                        
 Australia    : 0   Max.   :90.00                        
 (Other)      : 0                                        
```

```
	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 
```

Table Analysis:

We can note from the above table that the minimum rating(points) given between the two wines is 80 although the maximum shown is 90. The filtered information also shows that the count of reviews given for Chardonnay wine is 37 which is higher than that number of reviews given to other wine, i.e. 14. To understand data distribution, box plot along with 'jittered data' is displayed as below. In addition, t test is applied to compare two sample means.

The box plot shows that both samples are normally distributed as the median of each plot is closer to each of its mean. As per box plot, we can examine that Chardonnay wine's average rating is around near 86 and Sauvignon Blanc's average rating is around 87. Chardonnay wine has a median rating of 85 while Sauvignon has a median rating of 87.



The lowest 25 percent of Chardonnay wine ratings (1st quartile) are less than 84 while the lowest 25 percent of Sauvignon Blanc wine ratings (1st quartile) are less than 86. We can see that there are lots of outliers which are nothing but jittered noise to avoid overlapping of data.

Finally, as the respective median lies outside the box of the comparison box plot, we can assume that there is a discrepancy between two samples of wine. We can prove this difference in sample wines using T test statistic that follows the distribution of student t.

The results of T test indicate that the two samples (Chardona and savoru blanc) have different mean. This result is demonstrated by rejecting the null hypothesis which states that there is zero difference between the mean of two samples. P value - $p < 0.05$ and $t >$ Critical value with a confidence level of 95% means that we can reject null hypothesis implying that the true difference in means is not equal to zero.

Answer - 1.a.i

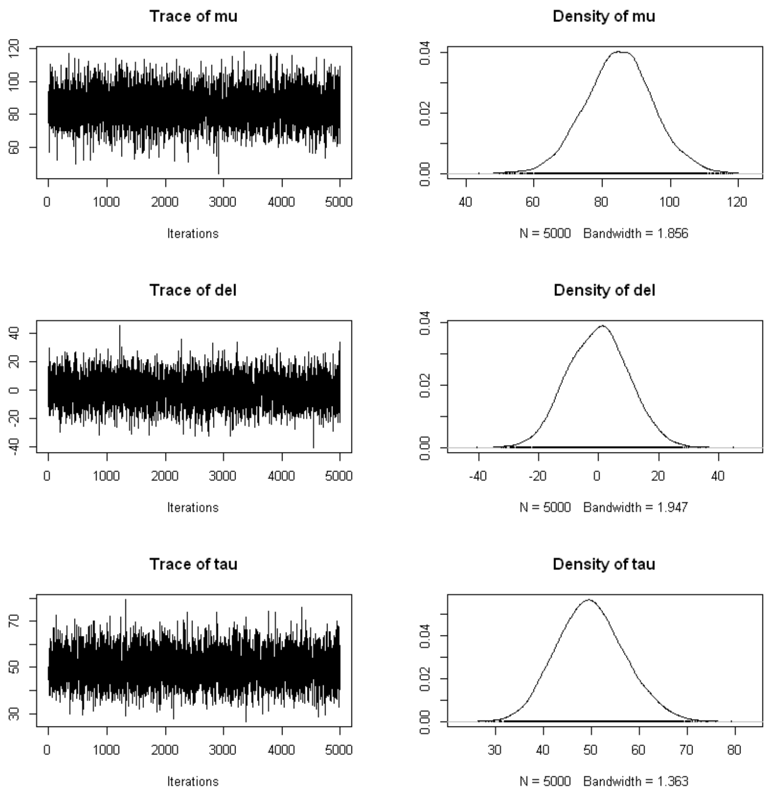
Following the study, we can conclude that Sauvignon Blanc from South Africa is better than Chilean Chardonnay wine. The mean rating difference between Sauvignon Blanc wine and Chardonnay wine is 2.133. This also means the Sauvignon Blanc wine is 2.50 per cent more valuable than Chile's Chardonnay wine price.

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?

In order to quantify the likelihood of Sauvignon Blanc being better than Chardonnay wine, we need to directly model the difference between two wine samples by means of the score. Given that the sample size of two wines is different and low, we cannot strongly predict the probability of better wine. Further samples from each distribution are hard to simulate directly. So that's why we use Gibbs sampling using Markov Chain Monte Carlo (MCMC) approach to compare each wine's marginal probability by first simulating posterior parameters from the joint probability distribution.

Prior parameters are taken as (mu0 = 80, tau0 = 1/100, del0=0, gamma0=1/100, a0 = 50, b0 = 1, maxiter = 5000). There is no fix rule for calculating priors a0 and b0, which can be taken as ambiguous if one is high and another is extremely small.

Below is the plot to understand the properties of the posterior distribution:



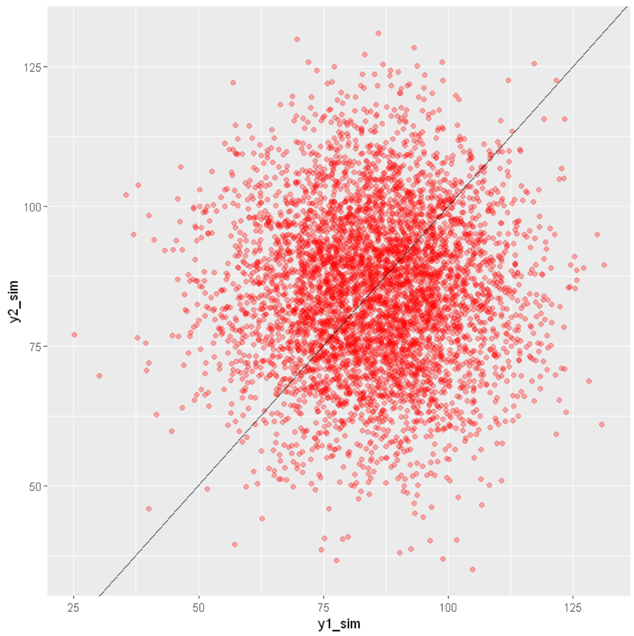
From this analysis we can say that the normal distribution of simulated posterior mean with the highest probability density of customer rating(points) occurs at around 85. Likewise, the parameter of precision (tau) is derived from the gamma distribution (which is biased by little). Now we have sampled subsequent normally distributed parameters from observed normally distributed data, from which we can now produce different samples for each wine, thereby calculating marginal probability.

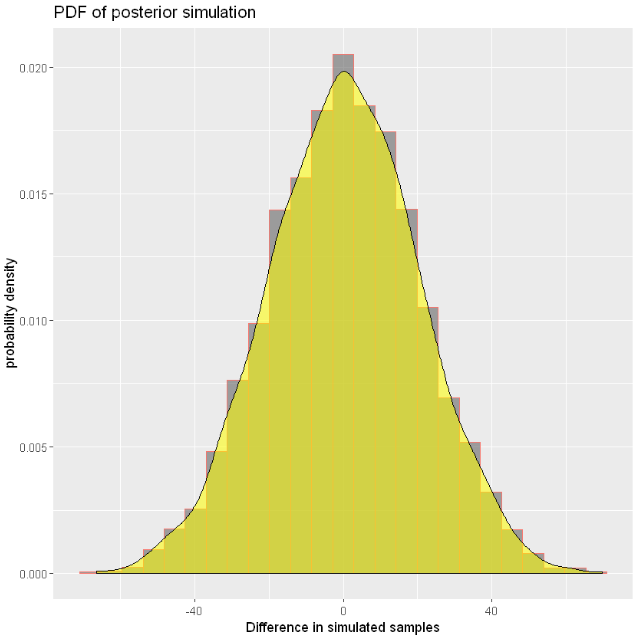
Gibbs sampler Performance

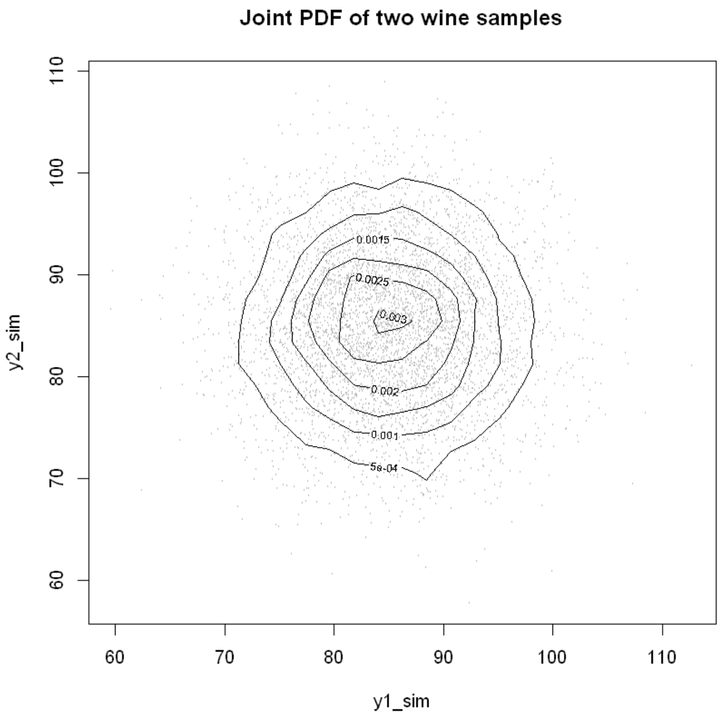
```
Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

     Burn-in  Total  Lower bound  Dependence
     (M)      (N)    (Nmin)       factor (I)
 mu   2       3803   3746         1.020
 del  2       3680   3746         0.982
 tau  2       3620   3746         0.966
```

The **sampler efficiency** can be estimated using the Dependency factor (I). Smaller the dependency factor (closer to 0 and 1), better sampler efficiency. Side fig shows the dependency factor is very small which explains the sampler 's satisfactory efficiency.

Now we're simulating samples for each wine using the normal distribution along with input posterior parameters.







Answer 1.a.ii

Various plots are plotted to summarize the ratings corresponding to each wine sample simulated by the gibbs sampling resulting from the posterior parameters. From the plots we see that there is a little skewness of distribution towards Sauvignon Blanc wine sample.

The probability that the Sauvignon Blanc is better than Chardonnay wine can be calculated as:

mean (y1_sim > y2_sim) → 0.73

b. 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.

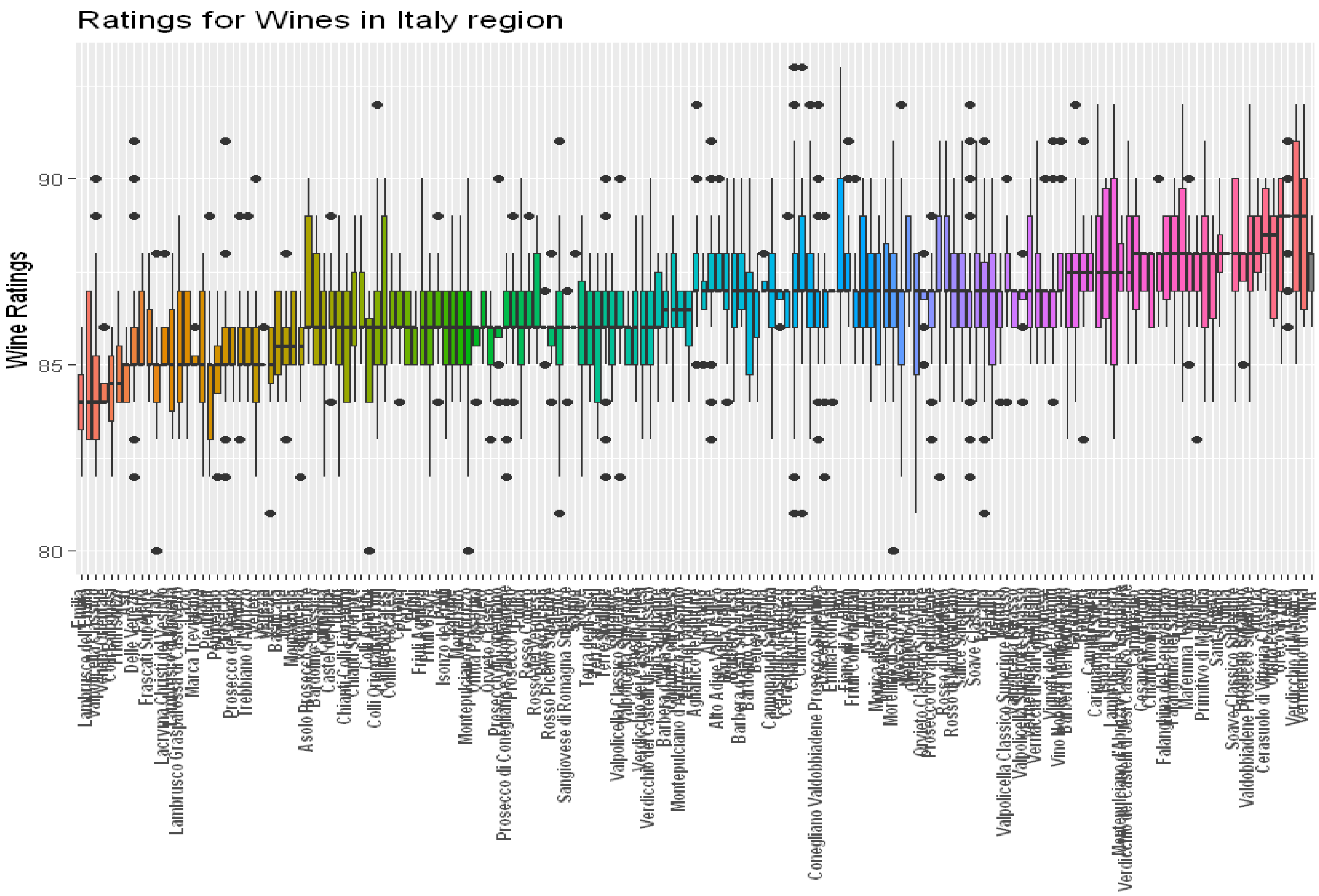
Data Handling

Again, whole 130k size wine dataset is filtered with region as Italy and price as lower than Euro 20. The data set being filtered includes 4702 row counts. In the data set region 1 column there are 8 missing values. Rows with missing region values are omitted because they are not imputable. You can summarize the data using table below:

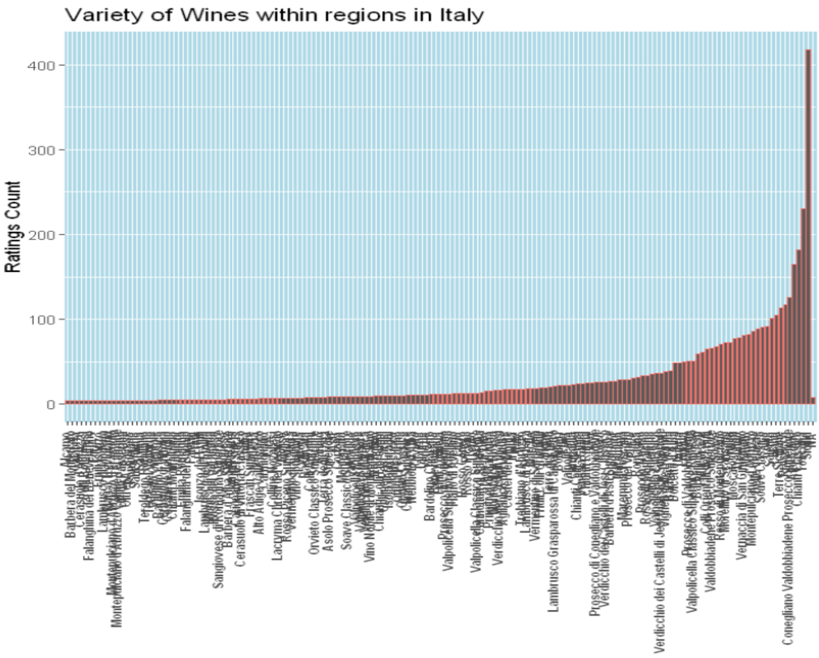
```
summary(df_test_4)
```

```
      country         points          price     
 Italy    :4702   Min.   :80.00   Min.   : 5.00  
          :   0   1st Qu.:86.00   1st Qu.:13.00  
 Argentina:   0   Median :87.00   Median :15.00  
 Armenia  :   0   Mean   :86.59   Mean   :15.02  
 Australia:   0   3rd Qu.:88.00   3rd Qu.:17.00  
 Austria  :   0   Max.   :93.00   Max.   :19.00  
 (Other)  :   0                                  
                                        region__1            variety    
 Sicilia                                     : 418   Red Blend   : 821  
 Toscana                                     : 230   Glera       : 351  
 Chianti Classico                            : 182   Pinot Grigio: 346  
 Alto Adige                                  : 165   Sangiovese  : 310  
 Conegliano Valdobbiadene Prosecco Superiore: 126   White Blend : 250  
 (Other)                                     :3573   Nero d'Avola: 180  
 NA's                                        :   8   (Other)     :2444  
```

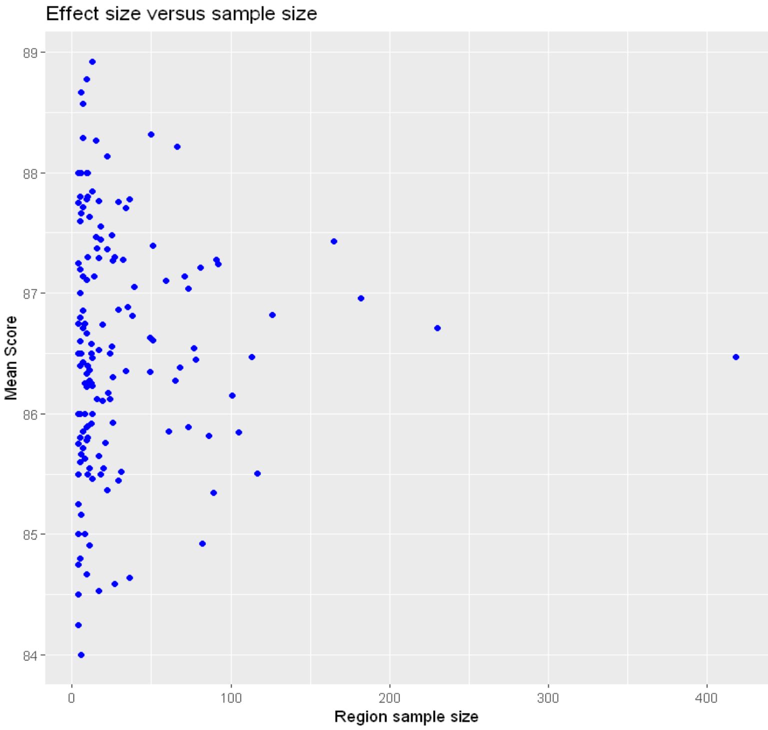
Now we can see that wine varieties belonging to multiple regions are numerous. Using boxplot below we seek to imagine the different regions and their distribution of scores.



Count of ratings given for wines in different regions within Italy can understood using below plot:



We can see in side plot that there are not many regions with significant (large) review counts. There are less than half regions with scores numbered over 40. We can visualize the count of reviews with respect to the ratings.



The count of reviews is found to significantly influence ratings. The average sample-sized ratings are higher than the average sample-sized ratings.

We are now specifically modelling the difference in mean ratings for each region again. Bearing in mind identical prior criteria as before. This model takes longer to run as the dataset is larger, and more parameters need to be sampled. The sampler's two outputs are: parameters that represent the subsequent mean, del and tau, while θ is the simulated group of mean parameters θˆ1, . , θˆm for each region.

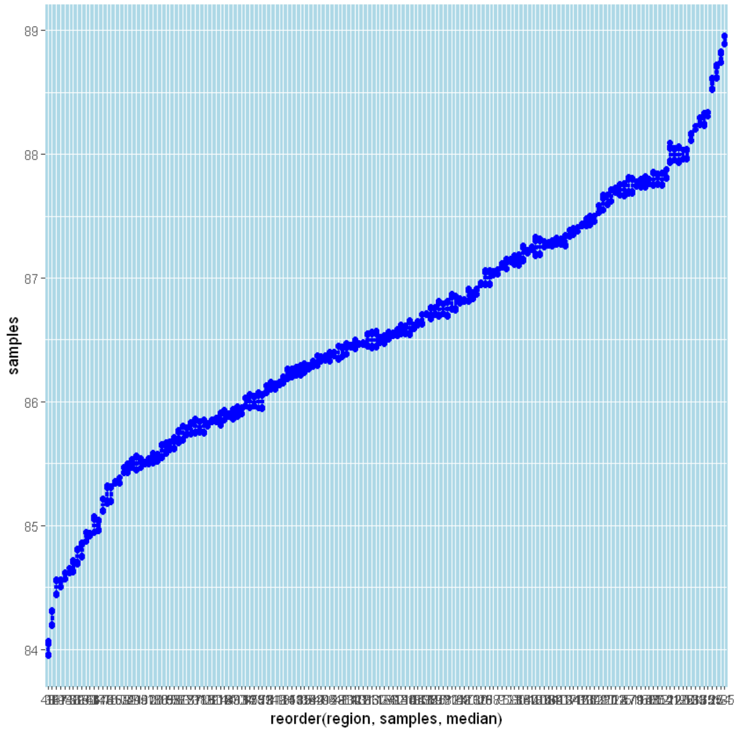
$params

mean	precision(w)	precision(b)
86.58239	770.7800	4.336458
86.52218	781.1695	4.387846
86.60860	741.5375	4.450199

$theta

Aglianico del Vulture	Alcamo	Alto Adige
87.76340	86.75087	87.43314
87.76020	86.74166	87.42482

We can also represent each region's sorted ratings w.r.t (linear relation) as below:



Sorting the average of the ratings for each region, it is observed that Trento region has received the highest rating.

Answer 1.b

To measure the regions which produce better than average wines, the mean of the simulated ratings (theta) is determined for each region and compared with the average of the distributed mean of the posterior joints. The result shows that the regions below yield better than average wines.

Aglianico del Vulture   Alcamo   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   Calabria   Campi Flegrei   Cannonau di Sardegna   Carignano del Sulcis   Carmignano
Cerasuolo d'Abruzzo   Cerasuolo di Vittoria   Cerasuolo di Vittoria Classico   Cesanese del Piglio   Chianti Classico   Chianti Montalbano   Chianti Rufina
CirÃ²   Colline Novaresi   Collio   Conegliano Valdobbiadene Prosecco Superiore   Dogliani   Etna   Falanghina del Beneventano   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   Teroldego Rotaliano   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

Question 2.

2. Build a linear regression model to estimate the points value for wines from the USA. Using simple language, identify which factors are most important in obtaining a good rating.

Data handling

Here the csv format of data (winemag-data-130k-v2.csv) is taken that is filtered to fetch the data for US, which is then checked for 'NA' entries. Out of '54504' entries, 239 rows contained null data in the column 'price' that are omitted for further analysis. Data contains 14 variables (columns) out of which 3 columns ('X', 'points' and 'price') are numerical variables. And variables like 'country', 'description', 'designation', 'province', 'region_1', 'region_2', 'taster_name', 'taster_twitter_handle', 'title', 'variety' and 'winery') are categorical columns.

```
> glimpse(Data)
Observations: 54,265
Variables: 15
$ X                     <int> 2, 3, 4, 10, 12, 14, 19, 20, 21, 23, 25, 29, 33, 34, 35, 41, 43, 45, ...
$ country               <fct> US, US, US, US, US, US, US, US, US, US, US, US, US, US, US, US, US, U...
$ description           <fct> "Tart and snappy, the flavors of lime flesh and rind dominate. Some g...
$ designation           <fct> , Reserve Late Harvest, Vintner's Reserve Wild Child Block, Mountain ...
$ points                <int> 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 86, 86, 86, 86, 86, 86, 8...
$ price                 <int> 14, 13, 65, 19, 34, 12, 32, 23, 20, 22, 69, 16, 50, 20, 50, 22, 14, 4...
$ province              <fct> Oregon, Michigan, Oregon, California, California, California, Virgini...
$ region_1              <fct> Willamette Valley, Lake Michigan Shore, Willamette Valley, Napa Valle...
$ region_2              <fct> Willamette Valley, , Willamette Valley, Napa, Sonoma, Central Coast, ...
$ taster_name           <fct> Paul Gregutt, Alexander Peartree, Paul Gregutt, Virginie Boone, Virgi...
$ taster_twitter_handle <fct> @paulgwineÂ , , @paulgwineÂ , @vboone, @vboone, @mattkettmann, , , @p...
$ title                 <fct> Rainstorm 2013 Pinot Gris (Willamette Valley), St. Julian 2013 Reserv...
$ variety               <fct> Pinot Gris, Riesling, Pinot Noir, Cabernet Sauvignon, Cabernet Sauvig...
$ winery                <fct> Rainstorm, St. Julian, Sweet Cheeks, Kirkland Signature, Louis M. Mar...
```

Columns Derived: Columns ('wordcount', 'year' and 'reviewcount') are derived from the respective columns ('description' and 'title'). **Columns Omitted:** There are certain columns that display redundant data originally or after deriving the new columns. These columns are not used further in the analysis. ('X', 'Country', 'taster_twitter_handle', 'designation', 'description', 'title'). **Categorical columns encoded:** Ordinal encoding of categorical variables ('province', 'region_1', 'region_2', 'taster_name', 'variety' and 'winery') is done to convert it to numeric variables. Here first the factorization of the variables is done to store them as levels and then the ordinal encoding is done.

```
> sapply(wine_dataset_US, class)
     points       price    province    region_1    region_2 taster_name        year     variety 
  "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric" 
     winery reviewcount   wordcount 
  "numeric"   "numeric"   "numeric" 
```

Analysis

Before diving into correlation of ratings (points) with other features, analyze the disttribution of 'points' frequency in the entire data. The histogram below shows that the number of reviews given to the wines with rating between 80 to 90 is greater than the number of reviews given to the wines with ratings between 90 to 100. The frequency distribution of the points seems to be normal.

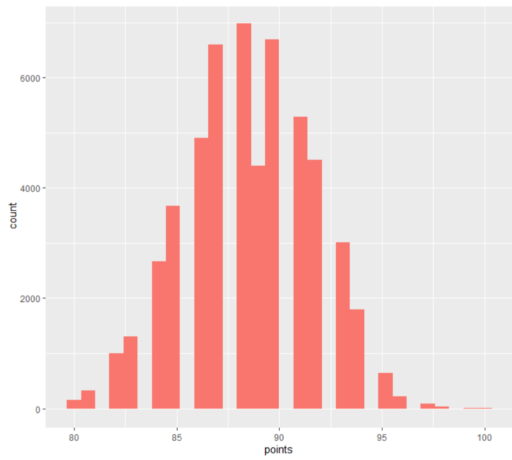


Fig 1. Frequency distribution of 'points' (target variable)

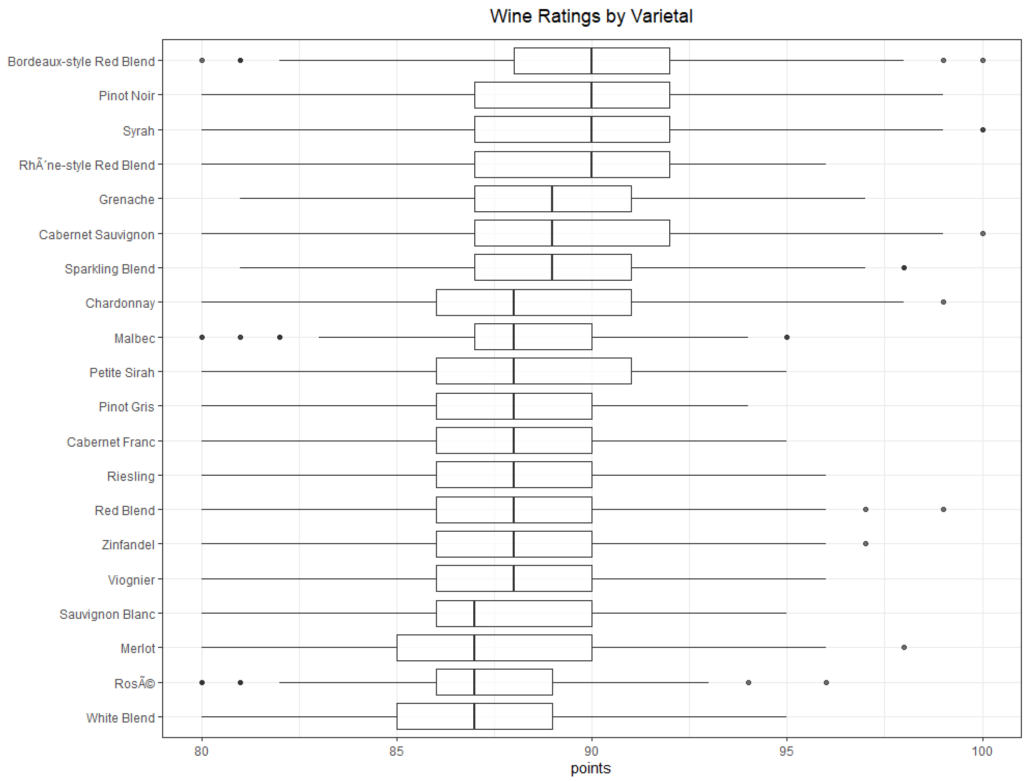


Fig 2. Boxplot showing dependency of Ratings on Variety of wine

As shown in Figure 2, there is not much effect of variety of wine on its rating but still it shows the top 4 variety of wines whose average rating go above 90. Although this can be further analyzed to find out how it is influencing but another factor that affects on the Variety<>Rating dependency is the number of wines analyzed of each variety.

Determining the correlation between 'points' and other variables

Since the question is to estimate the value of 'points' and identify the most influencing factors, 'points' here is the target variable and the other variables are predictors. Correlation matrix is formed between target variable and predictors (not including the derived variables).

	points
points	1.00000000
price	0.45307886
province	-0.11097407
region_1	0.02581571
region_2	-0.13232239
taster_name	-0.15535752
variety	-0.07209281
winery	-0.12633553
reviewcount	0.05021992
year	-0.13097095
wordcount	0.59803721

As seen above there is only one existing feature ('price') that seems to be most correlated to 'points'. Other than that, 'province','region_2','taster_name','winery' and 'variety' are weakly correlated. Notice the derived variables ('wordcount' , 'year') that shows correlation with the 'points' to some extent.

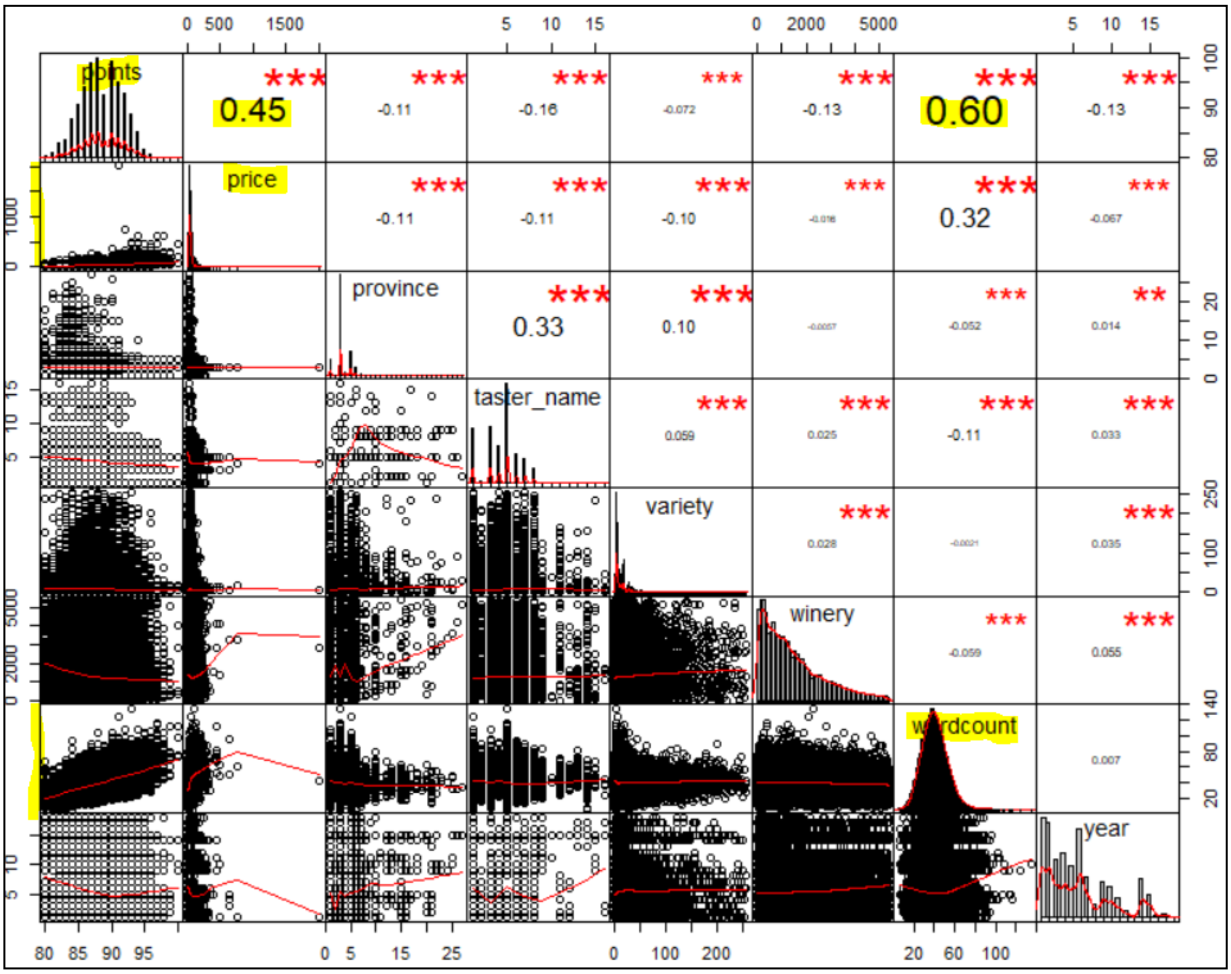


Fig. 3 Graphical representation of correlation matrix

Figure 3 shows the graphical view of correlation matrix. It is clearly shown the most correlated features with 'points' are 'wordcount' and 'price'. Refer to the corresponding coefficient values and the fitted graphs. The other features don't show that much correlation as shown in the graphs.

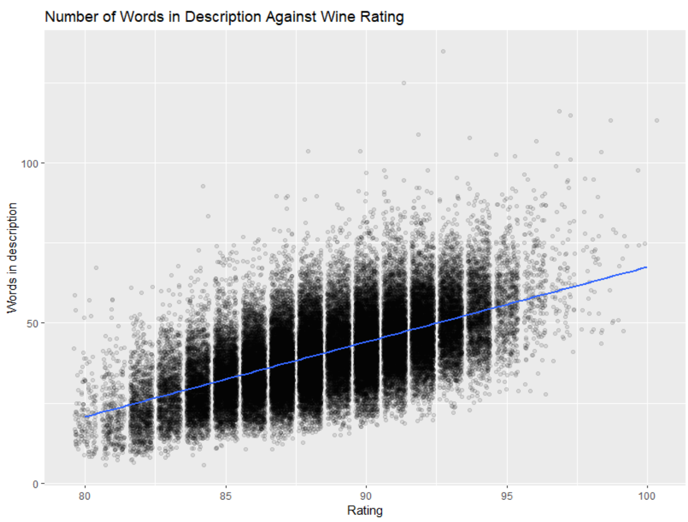


Fig 4. Word count is strongly correlated to points

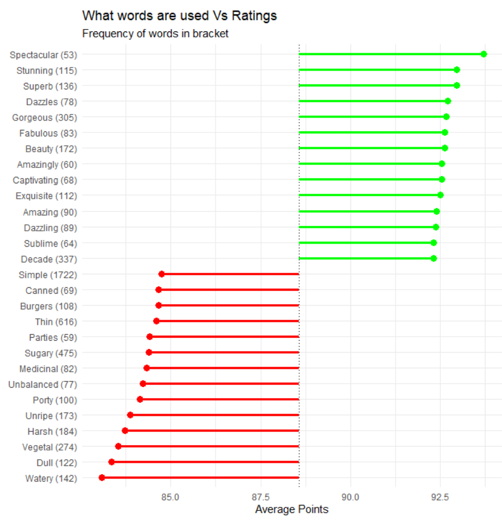


Fig 5. Word quality Vs Wine Ratings

Now as we get the highest correlated factors, lets estimate the 'points' w.r.t those factors and see how it influence the rating. Linear regression model: attempts to establish how X causes Y to change and the results of the analysis will change if X and Y are swapped. Following terms to be referred to interpret the LM summary report.

Formula call	formula R used to fit the data
Residuals	Difference between the actual observed response values and the response values that the model predicted. Ideally when plotted the distribution of the residuals should be symmetrical. The difference values of five parameters (Min, 1Q, Median, 3Q, Max) should be as low as possible for a good fit.
Coefficient Estimate	Contains multiple rows. First one is the intercept (when all the features are at 0, the expected response is the intercept). The other rows represent slope (the effect other variables have on the target variable).
Coefficient Standard Error	Average amount that the coefficient estimates vary from the actual average value of our response variable. This error for each variable should be as low as possible.
Coefficient - t value	A measure of how many standard deviations our coefficient estimate is far away from 0. Ideally it should be far away from zero as this would indicate we could reject the null hypothesis
Coefficient - Pr(>t)	Individual p value for each parameter to accept or reject null hypothesis. Lower the p value allows us to reject null hypothesis.
Residual Standard Error	Measure of the quality of a linear regression fit. Average amount that the response will deviate from the true regression line.
Multiple R-squared:	Measure how well the model fits the actual data. Measure of the linear relationship between predictor variable and response / target variable. High value is better Percentage of variation in the response variable that is explained by variation in the explanatory variable.
Adjusted R-squared	works well for multiple variables
F-Statistic	good indicator of whether there is a relationship between our predictor and the response variables

Model 1 (Base Model)

Estimating 'points' w.r.t price, province, region_1,region_2,taster_name,year,variety,winery,wordcount.

```
> summary(lm_model1)

Call:
lm(formula = points ~ price + province + region_1 + region_2 +
    taster_name + year + variety + winery + wordcount, data = wine_dataset_US)

Residuals:
    Min      1Q  Median      3Q     Max
-59.257  -1.545   0.029   1.605   8.510

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  8.365e+01  5.029e-02 1663.222  < 2e-16 ***
price        3.098e-02  3.936e-04   78.728  < 2e-16 ***
province    -6.652e-02  7.062e-03   -9.419  < 2e-16 ***
region_1     1.892e-03  2.804e-04    6.749  1.5e-11 ***
region_2    -1.317e-02  2.810e-03   -4.686  2.8e-06 ***
taster_name -7.792e-02  5.047e-03  -15.437  < 2e-16 ***
year        -8.052e-02  2.369e-03  -33.986  < 2e-16 ***
variety     -4.573e-03  4.685e-04   -9.761  < 2e-16 ***
winery      -2.231e-04  8.199e-06  -27.215  < 2e-16 ***
wordcount    1.273e-01  8.559e-04  148.687  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.287 on 54255 degrees of freedom
Multiple R-squared:  0.4615,    Adjusted R-squared:  0.4614
F-statistic:  5167 on 9 and 54255 DF,  p-value: < 2.2e-16
```

Here if we check the estimate coefficient of every variable, it is noticed that 'price', 'region_1' and 'wordcount' show the highest influence on the target variable 'points'. Also the't value' is high for price, region_1 and word count which shows some relation between factors and 'points'.

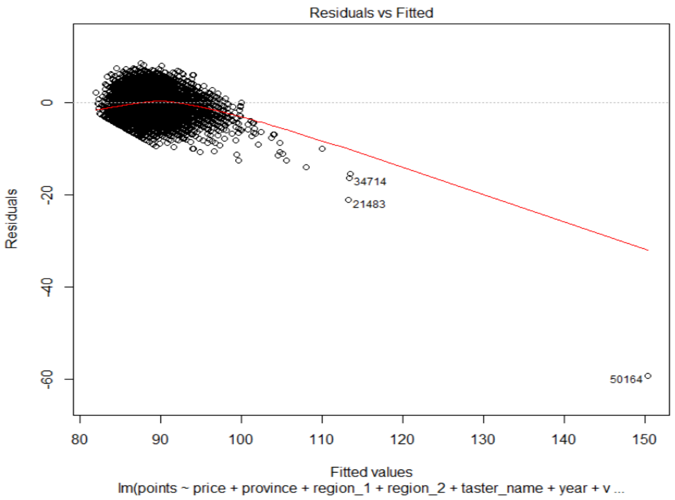


Fig 6. Residuals Vs Fitted plot for Model 1

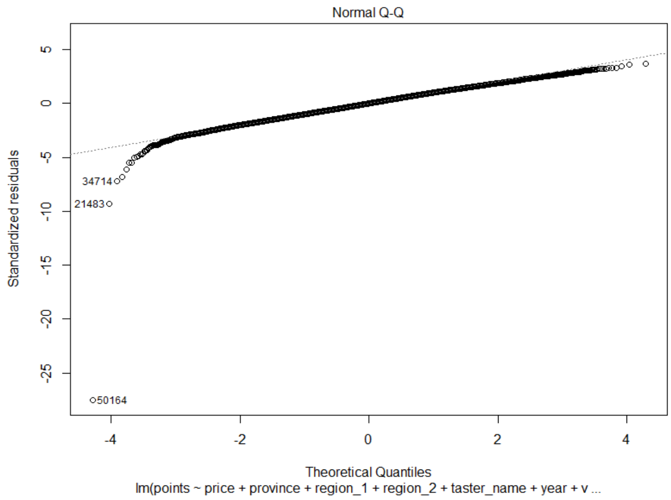


Fig 7. Normal Q-Q plot for Model 1

Figure 6 shows Residual Plot depicting a comparison of the residuals of model against the fitted values produced by model, and is the most important plot because it can tell us about trends in our residuals. It clearly shows that the residuals calculated does not fit and have non-linear patterns.
Figure 7 shows Normal Q-Q plot depicting that the residuals are roughly normally distributed. But there is a slight deviation at the lower end showing the difference in residuals.

Selecting features for improvement in model

Stepwise regression can be done with the measurement criteria as AIC and BIC- penalized-likelihood criteria. They are used for choosing best predictor subsets in regression. On applying AIC backward on Model1 we come to know what are the least important factors that can be excluded to improve the linear model. This is decided on the basis of the AIC value. **It is should be less for the model to be accepted.**

```
> AIC(lm_model1)
[1] 243805.4
> step_AIC_backward <- step(lm_model1)
Start:  AIC=89806.04
points ~ price + province + region_1 + region_2 + taster_name +
    year + variety + winery + wordcount

              Df Sum of Sq    RSS    AIC
<none>                     283855  89806
- region_2     1       115 283970  89826
- region_1     1       238 284093  89850
- province     1       464 284319  89893
- variety      1       498 284354  89899
- taster_name  1      1247 285102  90042
- winery       1      3875 287730  90540
- year         1      6043 289898  90947
- price        1     32428 316283  95674
- wordcount    1    115665 399520 108352
```

Current AIC value is 243805.4.
Since this is AIC_Backward, determining what factors have the least AIC value have to be removed. In other words, on removing what factors, give the best (least) AIC value, In this case 'region_2' and 'region_1'. Note that the AIC is increasing if removing 'price' and 'wordcount' i.e. it has to be included in the model.

Model 2:

Applying log to the columns like 'price' and 'wordcount' to re-scaling so that it matches its neighbors.
Removing columns from model: region_1 and region_2, since it shows least estimate (1.892e-03, (-)1.317e-02) and truth of estimation is with minimum standard error (i.e. the values in the Estimate) are close to the actual values.

```
> summary(lm_model2)

Call:
lm(formula = points ~ log(price) + province + taster_name + year +
    variety + winery + log(wordcount) + region_1 + region_2 +
    reviewcount, data = wine_dataset_US)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4229 -1.4775  0.0474  1.5318  8.1915

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.612e+01  1.222e-01 541.224  < 2e-16 ***
log(price)      1.930e+00  1.867e-02 103.364  < 2e-16 ***
province       -5.450e-02  6.753e-03  -8.072 7.08e-16 ***
taster_name    -6.320e-02  4.828e-03 -13.090  < 2e-16 ***
year           -6.100e-02  2.279e-03 -26.769  < 2e-16 ***
variety        -3.981e-03  4.479e-04  -8.890  < 2e-16 ***
winery         -1.786e-04  7.856e-06 -22.732  < 2e-16 ***
log(wordcount)  4.549e+00  3.233e-02 140.717  < 2e-16 ***
region_1        5.362e-04  2.687e-04   1.995    0.046 *
region_2        1.493e-02  2.717e-03   5.493 3.96e-08 ***
reviewcount     3.741e-05  5.225e-06   7.159 8.20e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.185 on 54254 degrees of freedom
Multiple R-squared:  0.5084,	Adjusted R-squared:  0.5083
F-statistic:  5611 on 10 and 54254 DF,  p-value: < 2.2e-16
```

Highlighted values show the improvement i.e. reduced residual values, increased estimate coefficients of log(price) and log(wordcount); reduced residual standard error and increased multiple r-squared which shows better fit. Note that the derived column 'reviewcount' doesn't show that much improvement.

Model 3:

In order to fit the model to some more extent, Adding and squaring the features (log(wordcount)+log(price))^2. Multiplying features taster_name*province.

```
Call:
lm(formula = points ~ (log(wordcount) + log(price))^2 + variety +
    taster_name * province + winery + year, data = wine_dataset_US)
```

```
Residual standard error: 2.177 on 54255 degrees of freedom
Multiple R-squared:  0.5123,    Adjusted R-squared:  0.5122
F-statistic:  6331 on 9 and 54255 DF,  p-value: < 2.2e-16
```

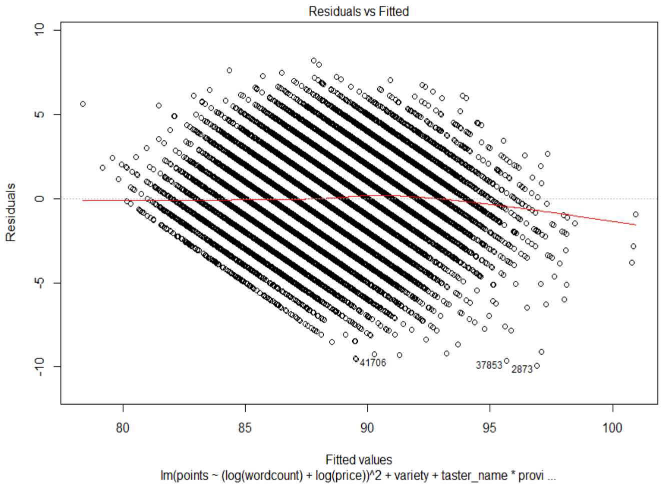


Fig. 8 Residual Vs Fitted plot for Model 3

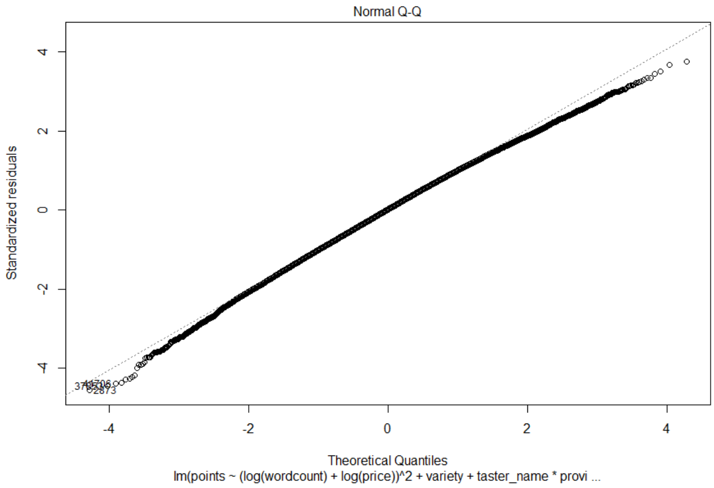


Fig 9. Normal Q-Q plot for Model 3

.

Figure 8 shows Residual Plot depicting a comparison of the residuals of model against the fitted values produced by model, and is the most important plot because it can tell us about trends in our residuals. We see that the red line is almost flat tells us that there is no discernible non-linear trend to the residuals.

Figure 9 shows Normal Q-Q plot which displays the residuals are towards the normal distribution and are somehow linearly related which means a model is good fit.

Conclusion

Using R's LM model, Model 1 is designed mainly with non-normalized existing features, it showed that less value is obtained for "Multiple and Modified R-Squared" which explicitly reveals that it is a badly fitted model. From the correlation coefficients within the correlation matrix and the estimates given within the model, it was evident that features such as 'region 1' and 'region 2' were least important in determining wine ratings. On the other end 'price' and the derived attributes 'wordcount' and 'year' contributed to the achievement of high wine ratings. AIC backward helped in selecting the features and resulting in the creation of Model 3. Such findings can also be interpreted in the sense of potential violations that may arise on assumptions. The ratings are also normally distributed. Various similarities can be seen in multiple distributions. Overall analysis is that it is important to find out the best correlated features as done above best fit the model. In future I would recommend doing more work on the textual characteristics of the data since this seems to be more a problem of categorical feature analysis.

References

[1] https://www.kaggle.com/zynicide/wine-reviews
[2] https://en.wikipedia.org/wiki/Gibbs_sampling#Relation_of_conditional_distribution_and_joint_distribution
[3] http://www.mit.edu/~ilkery/papers/GibbsSampling.pdf
[4] http://www2.stat.duke.edu/~rcs46/modern_bayes17/lecturesModernBayes17/lecture-7/07-gibbs.pdf
[5] https://stephens999.github.io/fiveMinuteStats/gibbs1.html
[6] https://www.statisticshowto.com/gibbs-sampling/
[7] Marginal Posterior Distribution, Harry F. Martz, Ray A. Waller, in Methods in Experimental Physics, 1994
[8] Methods for Computing Posterior Distributions- http://www.fao.org/3/Y1958E/y1958e04.html
[9] https://www.rdocumentation.org/packages/LaplacesDemon/versions/16.1.4/topics/joint.density.plot
[10] http://r-statistics.co/Linear-Regression.html
[11] https://stackoverflow.com/questions/19435773/significant-quadratic-terms-linear-regression-r
[12] https://data.library.virginia.edu/diagnostic-plots/
[13] https://www.kaggle.com/chrisbow/scalable-model-building-with-nested-regression
[14] https://rstudio-pubs-static.s3.amazonaws.com/431281_5df7c95c18984c43be6429f70c339611.html
[15] https://www.tutorialspoint.com/r/r_linear_regression.htm
[16] http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch10.pdf