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

School of Computer Science and Statistics

Assessment Submission Form

Student Name	Tanmay Bagla
Student ID Number	19300702
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: Tanmay

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 about the ratings given by customers for each wine. Some of the key columns to be considered for first question are country, price, points, region and variety. Country and region define the location of the vineyard. Price is the cost of each wine sold. Points refer to the ratings given by each customer. Variety column displays the name of the wines. The data consist of around 130k data points.
After the csv file is read, the data is filtered out as per the constraints given in question 1 i.e. rows with wine as Chardonnay from Chile and Sauvignon Blanc from South Africa is filtered whose price is exactly Euro 15.After the data is filtered, missing values are checked. As a result, no missing values were found in the filtered data. Also, to treat variety variable as an index value and not as a measurement we have changed the class of this object to be a factor. as.factor() function is used to convert the variable variety into a factor and preserve the variable label attribute.

Analysis

Summary of the data is presented as below:

```
summary(data_1)
```

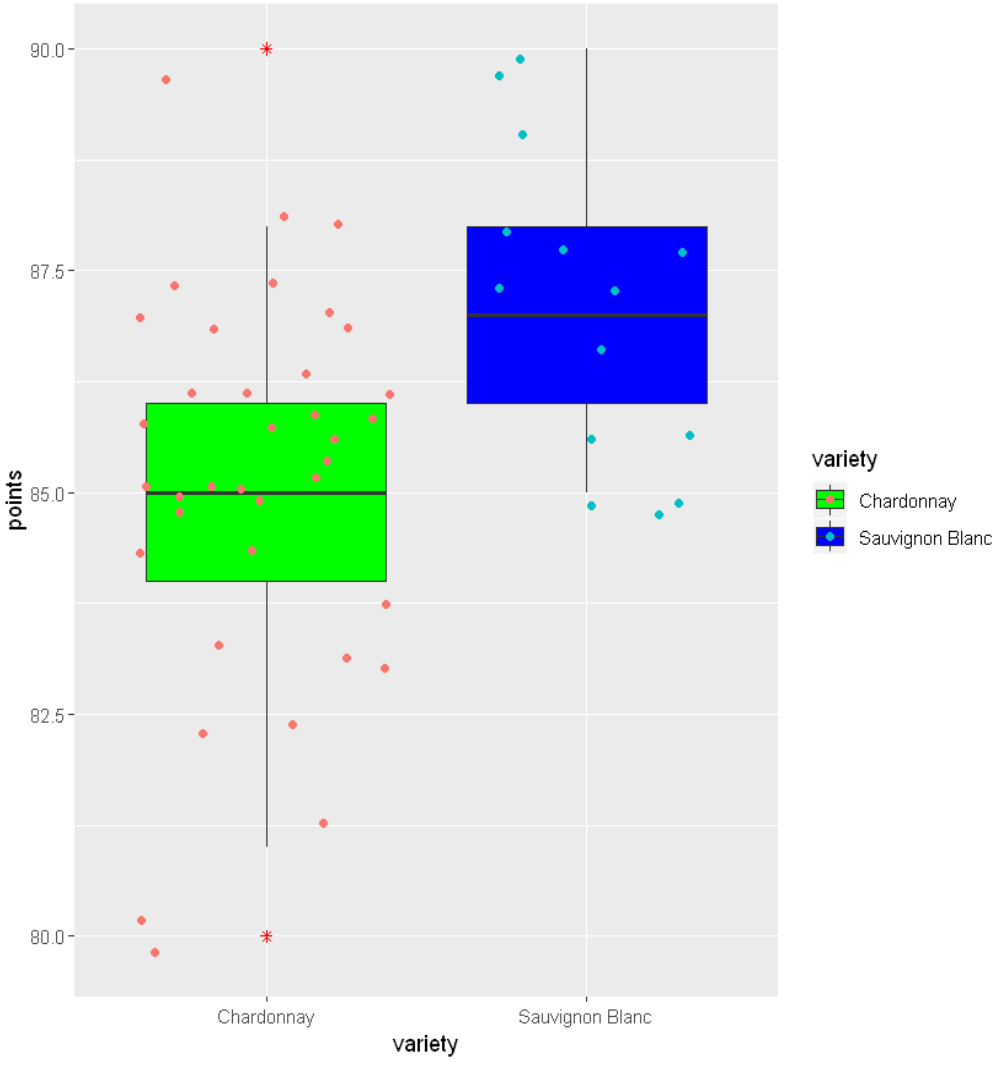
```
         country       points                 variety
Chile       :37   Min.   :80.00   Chardonnay     :37
South Africa:14   1st Qu.:85.00   Sauvignon Blanc:14
            : 0   Median :86.00
Argentina   : 0   Mean   :85.67
Armenia     : 0   3rd Qu.:87.00
Australia   : 0   Max.   :90.00
(Other) 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

T test (Student t distribution)

```
        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:

From the above table we can observed that minimum rating(points) given amongst the two wines is 80 whereas maximum rating given in 90. The filtered data shows that the count of ratings given for Chardonnay wine is 37 which is higher than the count of ratings i.e. 14 given for the other wine. To understand the distribution of data, box plot with addition of 'jittered data' is plotted as below. Also, t test statistic is applied to compare means of two samples.



The **box plot** shows that both the samples are normally distributed (fat tailed due to outliers) as the median of each of the plot is closer to each of its mean. According the box plot we can analyse that the average rating for Chardonnay wine is around 86.5 and average rating for Sauvignon Blanc wine is around 87. The median rating for Chardonnay wine is 85 whereas the median rating for Sauvignon Blanc is 87.
Lowest 25% of the ratings (1st Quartile) given to Chardonnay wine are less than 84 whereas lowest 25% of the ratings (1st Quartile) given to Sauvignon Blanc wine are less than 86. We can see that the there are many outliers which are nothing but the added jittered points in order to avoid overlapping of points.

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

The results of **T test** show that there is a difference in the means of the two samples (Chardona and savoru blanc). This result is proved by rejecting the null hypothesis which states that the difference between the mean of two samples is zero. $p < 0.05$ and $t >$ critical value with 5% significance level, it means we can reject null hypothesis with 95% confidence level which also means that the alternative hypothesis is true i.e. the true difference in means is not equal to zero. The 95% confidence interval is also providing a range that we are 95% confident the true difference in means of Chardonnay wine rating and Sauvignon Blanc wine rating is between 0.8181847 and 3.4482245.

1.a.i

After the analysis we can conclude that Sauvignon Blanc from South Africa is better than Chardonnay wine from Chile. The difference between average rating for Sauvignon Blanc wine and Chardonnay wine from Chile is 2.133. This also means that quality of Sauvignon Blanc wine is 2.50% higher than the quality of Chardonnay wine from Chile.

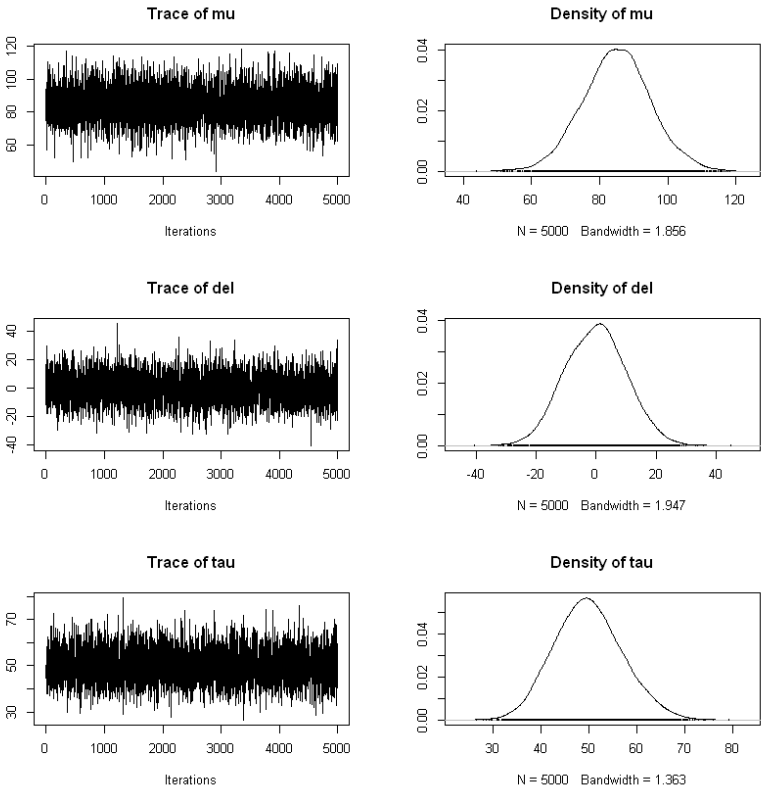
---

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?

To calculate the probability that Sauvignon Blanc is better than Chardonna wine, we need to explicitly model the difference in means of the rating between two wine samples. However, as the sample size of two wines are different and small, we cannot strongly predict the probability of better wine. It is difficult to directly simulate more samples from each distribution. So here we use Gibbs sampling using Markov Chain Monte Carlo(MCMC) method to calculate the marginal distribution of each wine by first simulating posterior parameters from the joint probability distribution.

Prior parameters are taken as (mu0 = 85, tau0 = 1/100, del0=0, gamma0=1/100, a0 = 50, b0 = 1, maxiter = 5000). There is no fix rule to calculate priors a0 and b0, they can be taken as vague with one being high and other being relatively small.

<u>Visualizing the basic properties of posterior distribution as below:</u>



From this visualization we can say that simulated posterior mean is distributed normally with highest probability density of customer rating(points) occurs at around 85. Similarly, precision parameter (tau) is simulated from gamma distribution (which is little rightly skewed). Now from observed normally distributed data, we have sampled posterior normally distributed parameters through which we can now generate different samples for each wine thereby calculating marginal probability.

Performance of Gibbs sampler

```
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
```

Performance of the sampler can be calculated from the Dependence factor (I). Smaller the dependence factor (closer to 0 and 1), better the performance of the sampler. Side fig shows that the dependence factor is quite small which explains the satisfactory performance of the sampler.

Summarising parameters of posterior distribution

```
apply(fit, 2, mean)
```

mu	85.0167282909926
del	0.0473199087241606
tau	84.9403792410249

```
apply(fit, 2, sd)
```

mu	5.02543696636854
del	4.92908030609616
tau	4.98936925045677

```
# to interperate tau we convert it to sd
mean(1/sqrt(fit[, 3]))
sd(1/sqrt(fit[, 3]))
```

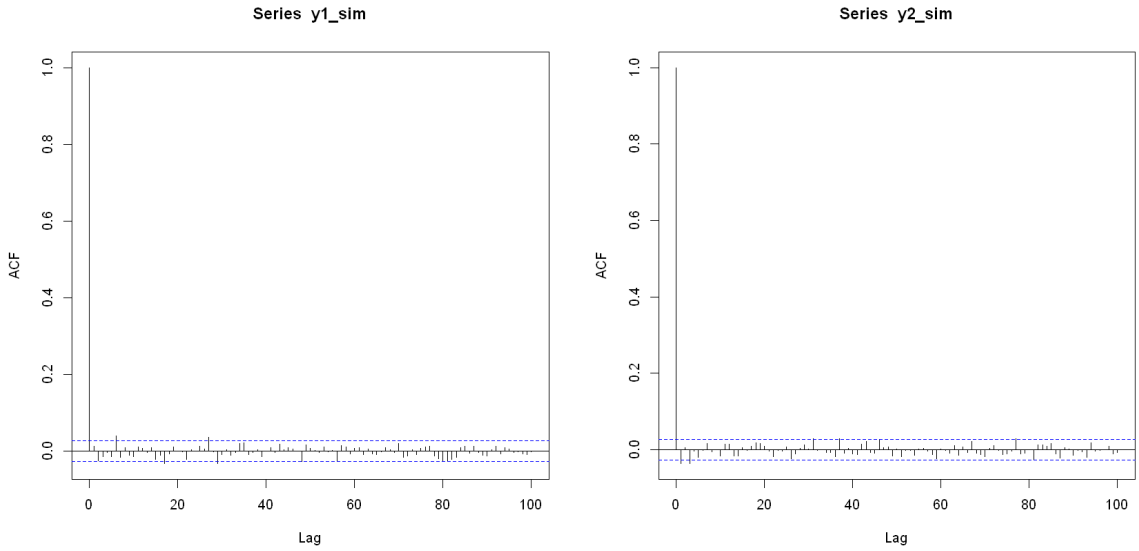
```
0.108644095685789

0.00320064793572873
```

Posterior mean has an average value of 85.01, and precision value of 84.9. Similarly, we can observe that posterior standard deviation has an average value of 5.02 and precision value of 4.98.

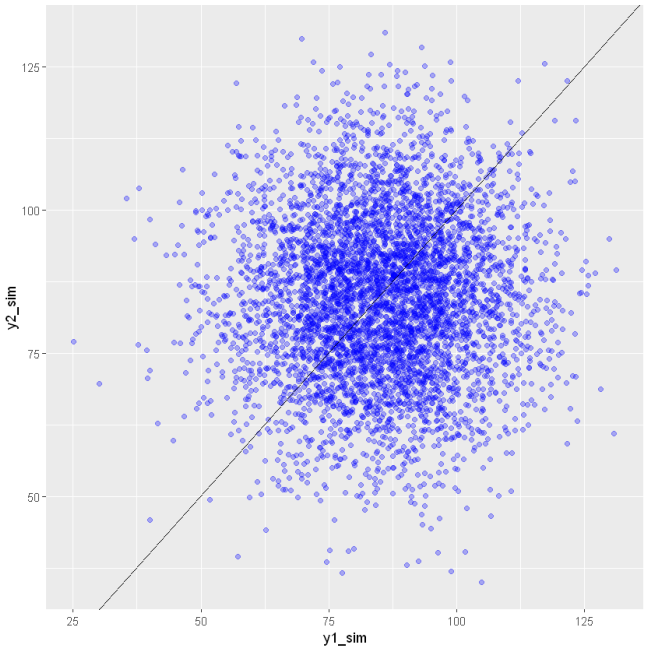
Now using the normal distribution along with input posterior parameters, we **simulate samples** for each wine.

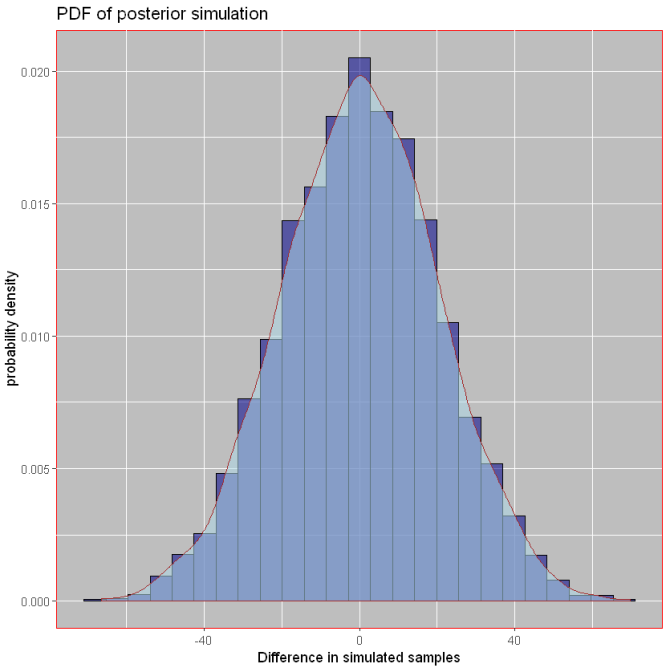
ACF(Auto Correlation Function)

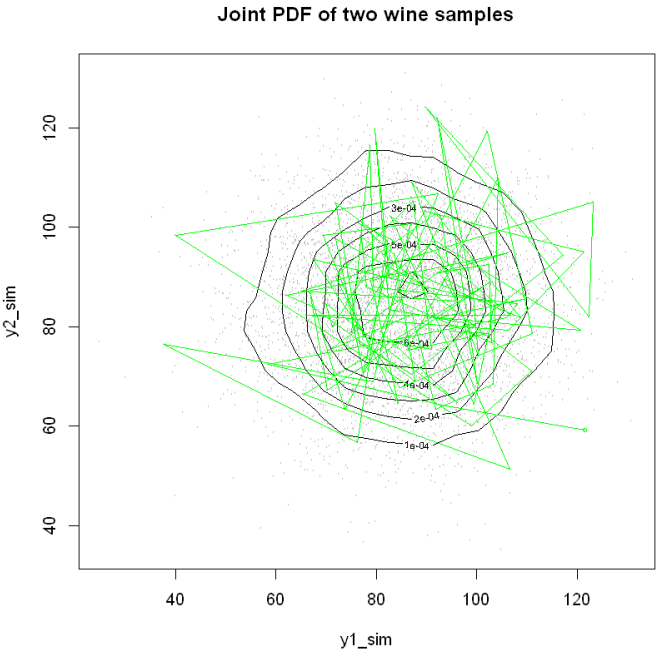
To understand if the samples simulated are not autocorrelated we can calculate the **ACF(Auto Correlation Function)** of each:



Except first value which is highly correlated with itself, we can see that all other lags are in the significance level which means at any given point the value of generated sample does not majorly depend(correlated) on the previous lags.







In Markov Chain Monte Carlo simulation, each sample is generated from previous value which can be show in side plot.

Various graphs have been plotted to summarize the ratings corresponding to each wine sample simulated through the posterior parameters resulted from gibbs sampling.

1.a.ii - The probability that the Sauvignon Blanc is better than Chardonnay wine is:

mean(y1_sim > y2_sim) → 0.74

y1_sim = Simulated samples for Sauvignon Blanc wine

y2_sim = Simulated samples for Chardonnay wine

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.

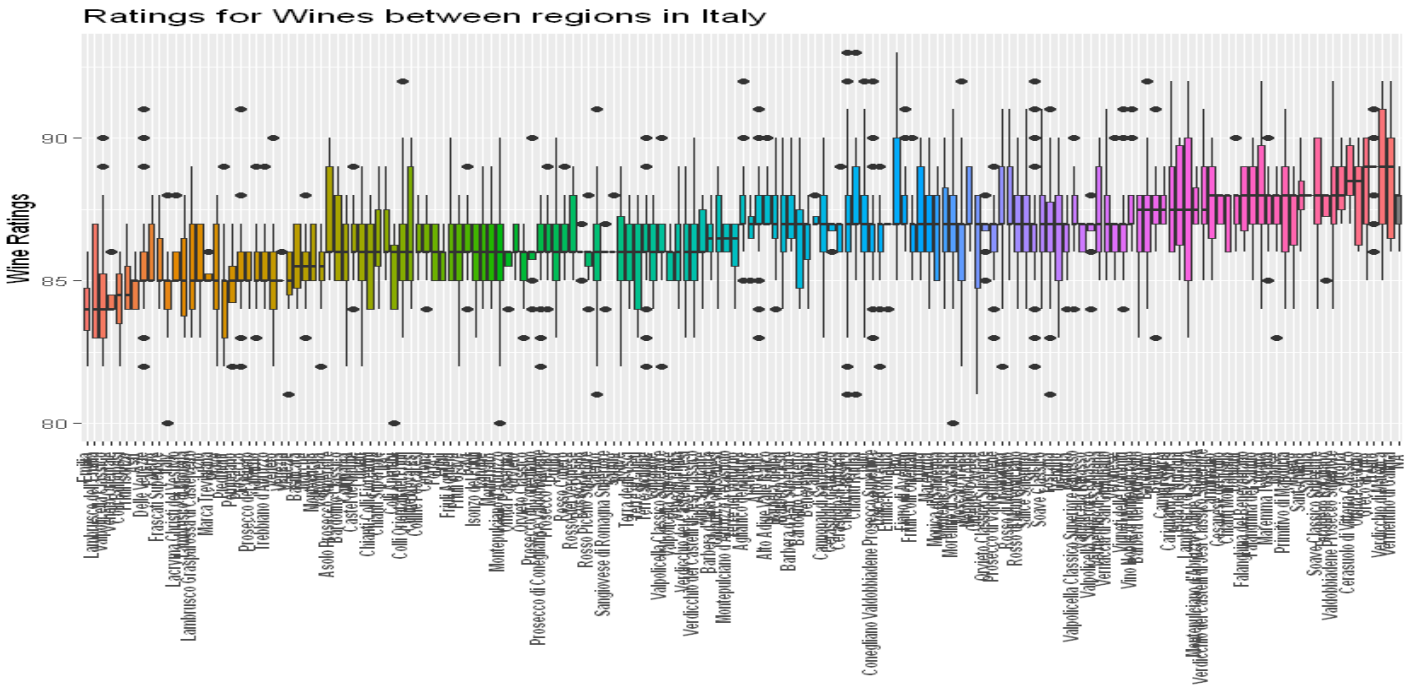
Data Handling

Once again whole wine dataset of 130k size is filtered with region as Italy and price as less than Euro 20. The filtered dataset has 4702 row counts. There are 8 missing values in the region 1 column of the dataset. Rows with the missing region values are omitted as they can't be imputed. Data can be summarized using below table:

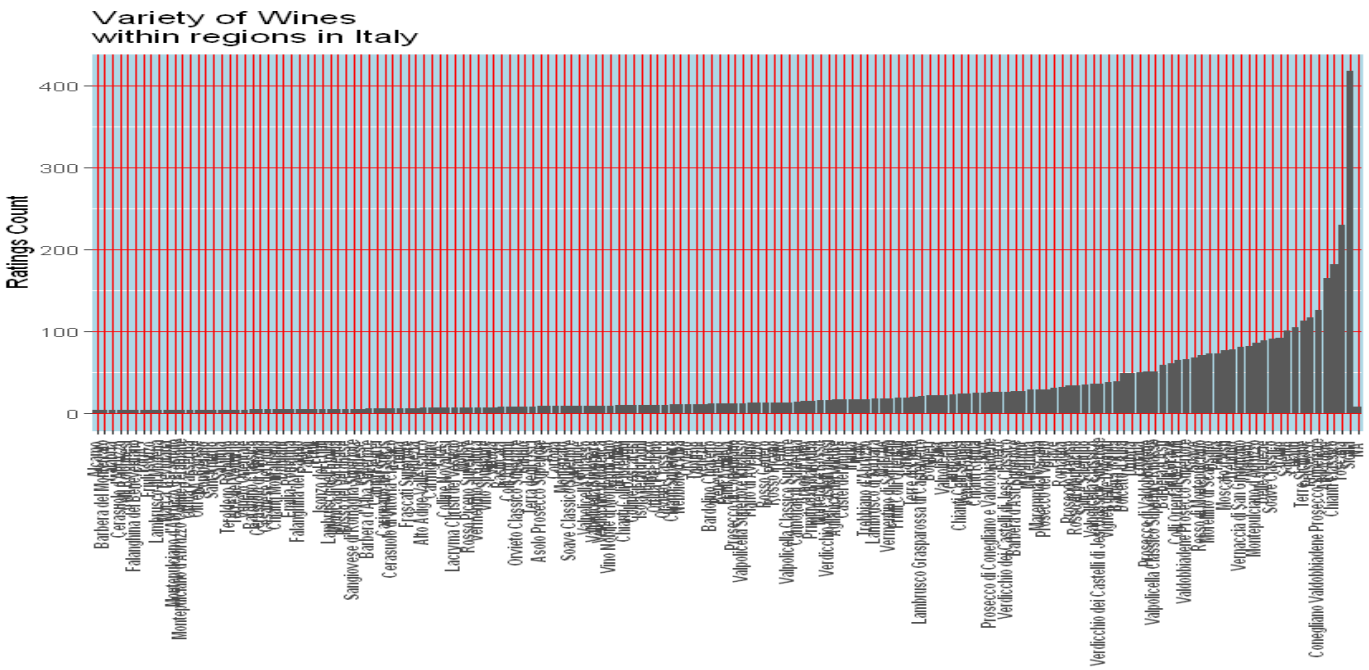
```
summary(df_test_4)
```

```
      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 there are many varieties of wines belonging to multiple regions. We try to visualize the different regions and their rating distribution using below boxplot.

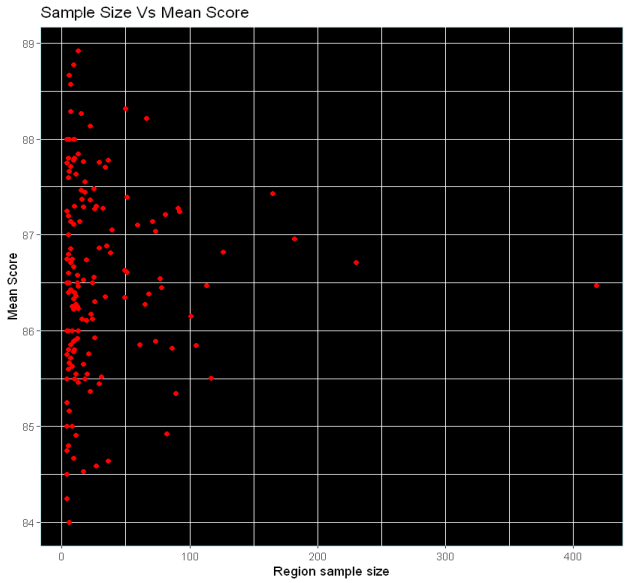


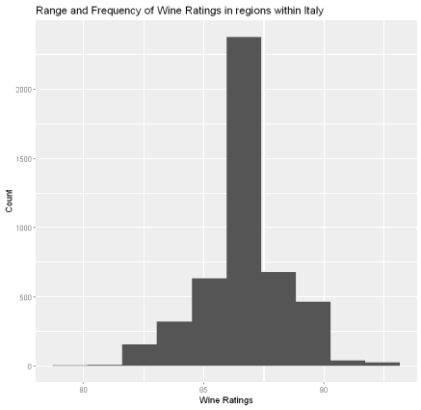
Count of ratings given for wines in different regions within Italy can understood using below plot:



In above plot we can see that there are not many regions with significant(large) review counts. There are very few regions with number of reviews more than 100.

We can visualize the count of reviews with respect to the ratings as below.





From the **histogram** we can observe that the maximum reviews given by the customer are having rating in between 86 to 88. Also from the **scatter plot** we observe that with increase in sample size the ratings plunge towards mean.
It is also observed that the count of reviews greatly influences ratings. The ratings with less sample size are higher than the ratings with higher sample size.

Now again we explicitly model the difference in the mean ratings for each region. Similar prior parameters as before are taken into account. This model takes more time to run as the dataset is larger and it has more parameters to be sampled. The two output of the sampler are: params which displays posterior mean, del and tau whereas θ is the simulated group of mean parameters $\theta\hat{}_1, \ldots, \theta\hat{}_m$ for each region.

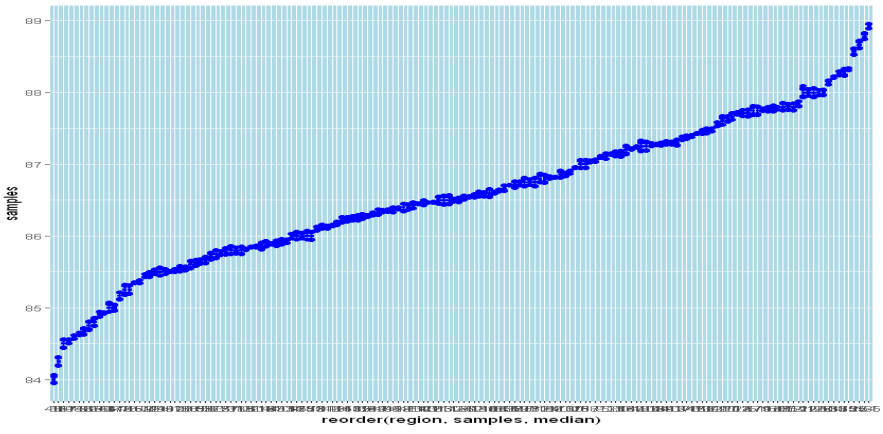
$params

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 visualize the sorted mean ratings of each region (linear relation as below:



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

```
sort(theta_hat, decreasing = TRUE)
```

Trento	88.9229902474753
Verdicchio di Matelica	88.7767370118932
Cerasuolo di Vittoria ...	88.6656463030176
Vermentino di Gallura	88.5704361416656
Lugana	88.3205722990475
Vittoria	88.2852020324851
Greco di Tufo	88.2672269849082

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

1.b - To calculate which regions, produce better than average wines, the mean of the simulated ratings (theta) for each region is calculated and compared with the average of the posterior joint distributed mean. The result indicates that the following regions produce better than average wines.

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

Dataset (winemag-data-130k-v2.csv) is taken and filtered to get the data for US. It is then verified against the number of 'NA' values. 239 rows out of '54504' entries, contained NA data in the column 'price' that is omitted for further analysis. Data contains 14 columns out of which 3 columns ('X', 'points' and 'price') are numerical variables. The remaining 11 variables like 'country', 'description', 'designation', 'province', 'region_1', 'region_2', 'taster_name', 'taster_twitter_handle', 'title', 'variety' and 'winery') are categorical columns.

```
> ExpData(data=Data,type=2)
   S.no          Variable Name Variable Type
1     1                      X         integer
2     2                country          factor
3     3            description          factor
4     4            designation          factor
5     5                 points         integer
6     6                  price         integer
7     7               province          factor
8     8               region_1          factor
9     9               region_2          factor
10   10            taster_name          factor
11   11 taster_twitter_handle          factor
12   12                  title          factor
13   13                variety          factor
14   14                 winery          factor
```

- **Derived** 3 columns ('wordcount', 'year' and 'reviewcount') from the columns ('description' and 'title').
- **Omitted** columns that contained redundant data or by logical reasoning that makes very less sense on the target variable. Columns which are not used for further analysis. ('X', 'Country', 'taster_twitter_handle', 'designation', 'description', 'title').
- **Encoding categorical variables** through Ordinal encoding. Variables('province', 'region_1', 'region_2', 'taster_name', 'variety' and 'winery') are converted to numeric variables. Firstly 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     winery reviewcount
 "numeric"  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"
wordcount
 "numeric"
```

Analysis

Until digging into association of ratings (points) with other factors, examine the distribution of 'points' frequency in the complete dataset. The histogram indicates that the amount of reviews provided to the wines with ranking between 80 and 90 is higher than the amount of reviews given to the wines with scores from 90 and 100. Also this histogram ensures that the data is normally distributed.

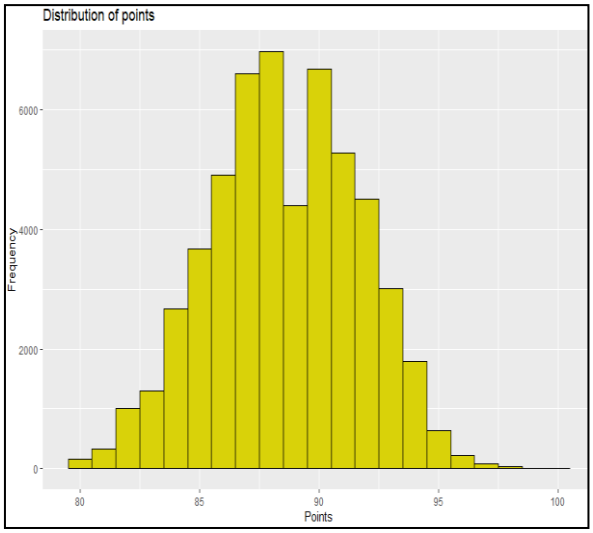


Fig. 1. Frequency distribution of points

Determining the correlation between 'points' and other variables

The objective is to estimate the 'points' and determine the most important factors. Therefore 'points' here is the target variable and the other variables are predictors. In order to get an idea about relation, Correlation matrix is formed between target variable and predictors. As seen above there is only one feature ('price') that seems to be most correlated to 'points'. Other than that, 'province','region_2','taster_name','winery' and 'variety' are weakly correlated.

```
> cor(subset(wine_dataset_US, select=c(points,price,province,region_1,region_2,taster_name,variety,winery)))
                points       price     province   region_1   region_2 taster_name     variety      winery
points      1.00000000  0.45307886 -0.110974072 0.02581571 -0.1323224 -0.15535752 -0.07209281 -0.126335526
price       0.45307886  1.00000000 -0.112733793 0.00506739 -0.2122057 -0.10535167 -0.10100565 -0.015589644
province   -0.11097407 -0.11273379  1.000000000 0.20201136  0.3047703  0.33143015  0.10120025 -0.005723197
region_1    0.02581571  0.00506739  0.202011358 1.00000000  0.2643455  0.06326955  0.07784820  0.022782578
region_2   -0.13232239 -0.21220567  0.304770320 0.26434547  1.0000000  0.20697014  0.11087457 -0.038110196
taster_name -0.15535752 -0.10535167  0.331430151 0.06326955  0.2069701  1.00000000  0.05904396  0.024508992
variety    -0.07209281 -0.10100565  0.101200246 0.07784820  0.1108746  0.05904396  1.00000000  0.027872366
winery     -0.12633553 -0.01558964 -0.005723197 0.02278258 -0.0381102  0.02450899  0.02787237  1.000000000
```

Let's try adding derived variables ('wordcount', 'year', 'reviewcount') and confirm its correlation with the 'points'. Figure 2 shows the graphical view of correlation matrix. It is clearly shown the most correlated features with 'points' are 'wordcount', 'price'. Refer to the corresponding coefficient values and size and color encoding of the circles. The other features don't show that much correlation as shown in the figure.

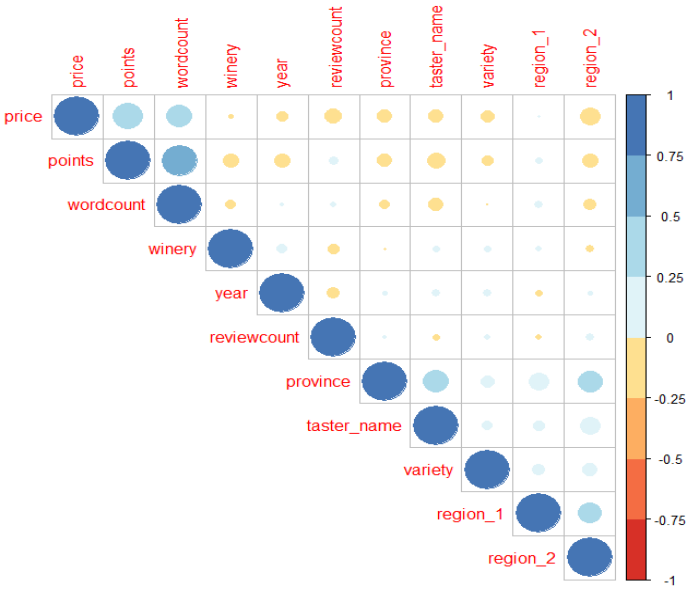


Fig. 2. Graphical representation of correlation matrix

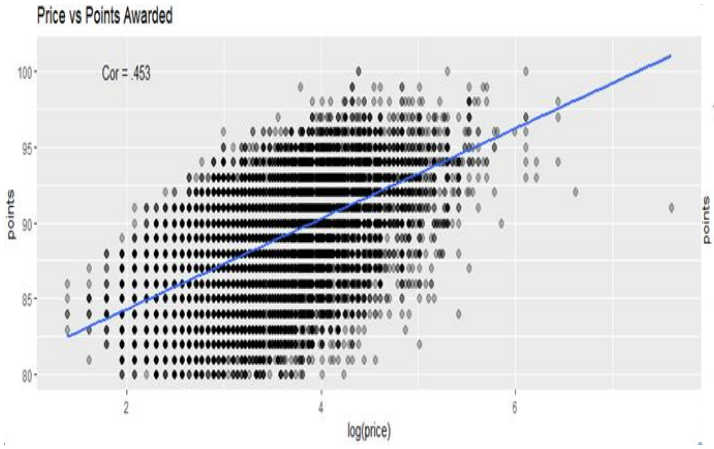


Fig 3. Price Vs Points

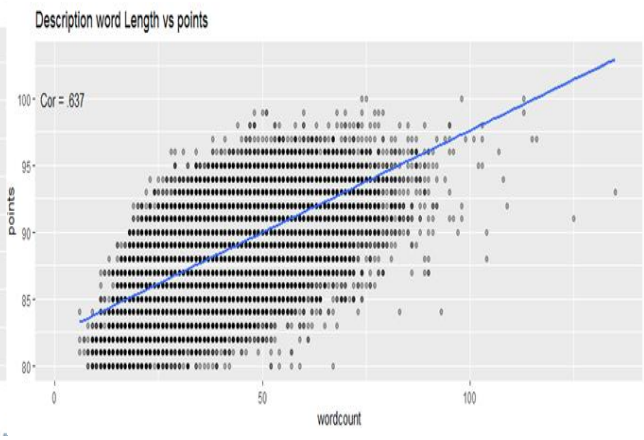


Fig 4. Word count Vs Points

Now that we have the maximum correlation variables, let's approximate the 'points' w.r.t such variables to see if they affect the rating. The word count (quantity) appears to have a significant influence on good or poor scores. Let's dig down to test if the content (quality) of words affects the rating.

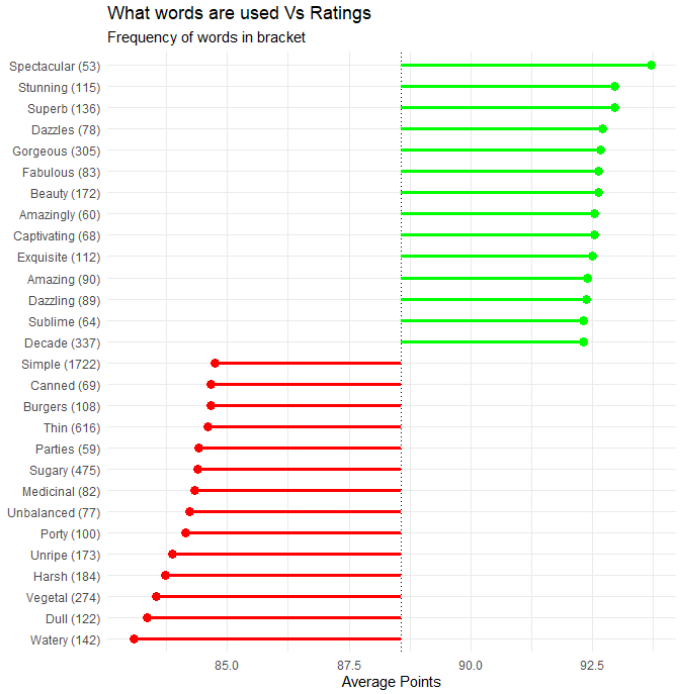


Fig 5. Effect of quality of words on ratings

Figure 5 illustrates how the content of words may be used to determine the ranking. For ex; wines having reviews with the term 'Watery' in them tend to have the least average rating and the term 'Spectacular' is mentioned by wines with the highest rating.

Linear regression model: tries to decide whether X induces Y to shift and the outcomes of the experiment will alter whether X and Y are shared. The terms to be used to interpret the summary report of the LM are as follows.

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, variety, winery (Derived variables wordcount, year and reviewcount is not taken)

```
> summary(lm_model1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-95.047  -1.818   0.078   2.008   9.830

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  8.802e+01  4.350e-02 2023.511  < 2e-16 ***
price        4.948e-02  4.462e-04  110.897  < 2e-16 ***
province    -6.441e-02  8.423e-03   -7.647 2.09e-14 ***
region_1     3.991e-03  3.339e-04   11.954  < 2e-16 ***
region_2    -2.192e-02  3.351e-03   -6.541 6.17e-11 ***
taster_name -1.387e-01  6.001e-03  -23.109  < 2e-16 ***
variety     -2.533e-03  5.583e-04   -4.537 5.71e-06 ***
winery      -3.083e-04  9.748e-06  -31.633  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.728 on 54257 degrees of freedom
Multiple R-squared:  0.234,     Adjusted R-squared:  0.2339
F-statistic:  2367 on 7 and 54257 DF,  p-value: < 2.2e-16
```

Here if we check the estimate coefficient of every variable, it is noticed that 'price' and 'region_1' show the highest influence on the target variable 'points'. Also the 't value' is high for price, region_1 which shows some relation between factors and 'points'. The value of Multiple and Adjusted R-squared is very low which signifies the model is not fitted well.

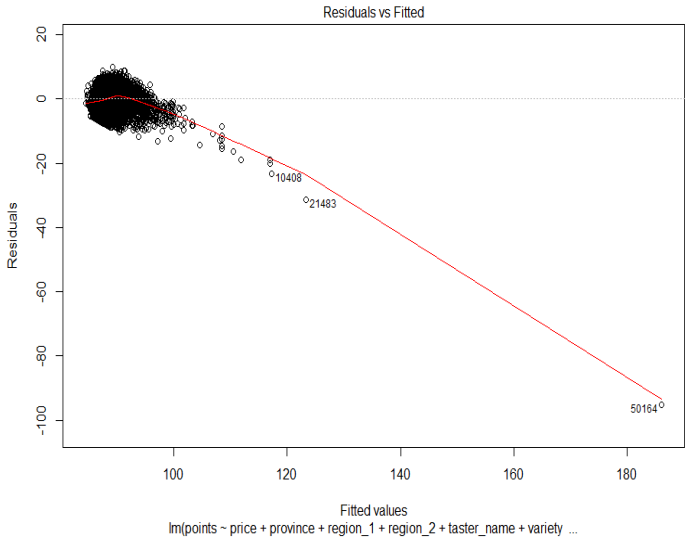


Fig 7. Normal Q-Q plot

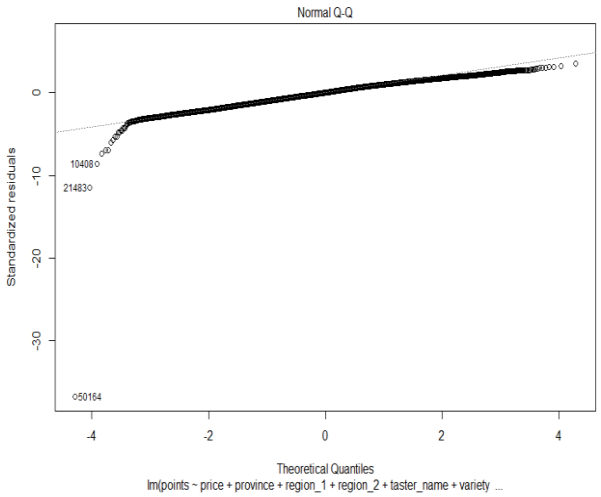


Fig 6. Residuals Vs Fitted plot

Figure 6 shows Residual Plot, which shows a comparison of the residuals of the experiment with the fitted values generated by the model, which is the most significant plot as it can inform us about patterns in our residuals. It clearly shows that the residuals calculated do not fit and have non-linear patterns. This indicates a large variance indicating that the residuals measured in the model have a significant error rate.

Figure 7 shows Normal Q-Q plot depicting that the residuals are roughly normally distributed. But there is a noticeable deviation at the lower end showing the difference in residuals.

Model 2:

Adding derived variables (price, wordcount and reviewcount) in the model. Applying log to the columns like 'price' and 'wordcount' to re-scale so that it matches its neighbors.

```
> 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) and reviewcount. Also it shows reduced residual standard error and a sharp hike in multiple r-squared which shows better fitting model.

Selecting features for improvement in model

Stepwise regression is conducted with the measurement metrics as AIC and BIC penalized-likelihood criteria. They are used to select the best predictor subset of features for regression. When applying AIC backwards to Model1, we come to learn what are the least significant variables that can be omitted to strengthen the linear model. This is calculated on the basis of the value of AIC. It's expected to be less for the model to be accepted.

```
> AIC(lm_model2)
[1] 238863.9
> step_AIC_backward <- step(lm_model2)
Start:  AIC=84864.47
points ~ log(price) + province + taster_name + year + variety +
    winery + log(wordcount) + region_1 + region_2 + reviewcount

               Df Sum of Sq    RSS    AIC
<none>                     259139  84864
- region_1      1        19 259158  84866
- region_2      1       144 259283  84893
- reviewcount   1       245 259384  84914
- province      1       311 259450  84928
- variety       1       377 259516  84941
- taster_name   1       818 259957  85034
- winery        1      2468 261607  85377
- year          1      3423 262561  85575
- log(price)    1     51032 310170  94617
- log(wordcount) 1    94579 353718 101746
```

Current AIC value is 238863.9

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 3:

In order to fit the model to some more extent, Adding and squaring the features (log(wordcount)+log(price)+log(reviewcount)+log(year))^2). Multiplying features taster_name*province. Removing columns from model: region_1 and region_2, since it shows least estimate in the last model (Model 2) and truth of estimation is with minimum standard error (i.e. the values in the Estimate) are close to the actual values. Also the AIC_backward suggests trying removing 'region_1' and 'region_2' from the model.

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

```
Residual standard error: 2.028 on 54249 degrees of freedom
Multiple R-squared:  0.5767,    Adjusted R-squared:  0.5766
F-statistic:  4927 on 15 and 54249 DF,  p-value: < 2.2e-16
```

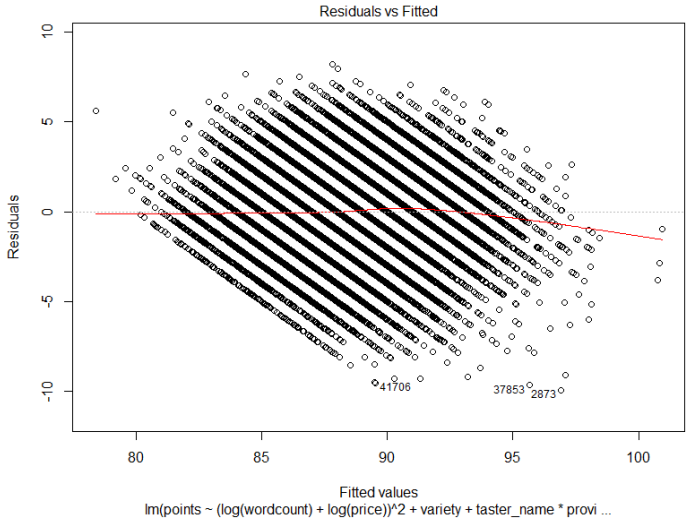


Fig 8. Residual Vs Fitted plot for Model 3

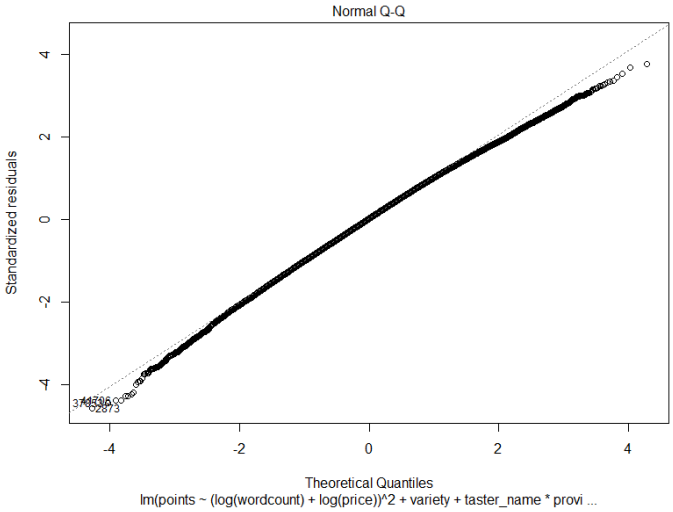


Fig 9. Normal Q-Q plot for Model 3

.

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 depicting that the residuals are roughly normally distributed.

Conclusion

Using LM model of R, Model 1 is built only with the existing features that are not normalized, it showed very less value is achieved for "Multiple and Adjusted R-Squared" which clearly shows that it is a poorly fitted model. It was evident from the correlation coefficients in the correlation matrix and the estimates given in the model that the features like 'region_1' and 'region_2' were least influential in deciding the rating of the wines. **On the other hand 'price' and the derived features 'wordcount' and 'year' contributed in obtaining the good ratings of wines**. AIC backward helped in selecting the features and resulting in the creation of Model 3. These results should also be viewed under possible violations that can happen on assumptions. Like here the ratings have normal distribution. It may be possible to show different correlations in different distributions. Overall evaluation is such that for the model to have good fit it is necessary to know the correlation as done above. I would recommend to work more on the textual features of the data since this more seems to be a textual analysis problem.

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] https://www.rdocumentation.org/packages/LaplacesDemon/versions/16.1.4/topics/joint.density.plot
[9] http://r-statistics.co/Linear-Regression.html
[10]https://stackoverflow.com/questions/19435773/significant-quadratic-terms-linear-regression-r
[11]https://data.library.virginia.edu/diagnostic-plots/
[12]https://www.kaggle.com/chrisbow/scalable-model-building-with-nested-regression
[13]https://rstudio-pubs-static.s3.amazonaws.com/431281_5df7c95c18984c43be6429f70c339611.html
[14]https://www.tutorialspoint.com/r/r_linear_regression.htm
[15]http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch10.pdf