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

School of Computer Science and Statistics

Assessment Submission Form

Student Name	Vishal Kumar
Student ID Number	19304373
Course Title	MSc. Computer Science- Intelligent Systems
Module Title	Applied Statistical Modelling
Lecturer(s)	Dr. Arthur White
Assessment Title	Assignment- Main
Date Submitted	14-05-2020
Word Count	

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: Vishal Kumar

Date: 14-05-2020

GitHub link for code: https://github.com/vishalkumarmishra7/TCD_Applied_Statistical_Modelling_CS7DS3.git

Introduction

Wine dataset was extracted from WineEnthusiast[1]. This dataset can be used with deep learning or statistical model can be fitted to check the prediction and interpret some useful information. Wine dataset (winemag-data-130k-v2.csv) has serval information about wine like country, description, designation, points, price, province, region_1, region_2, taster_name, title, variety, taster_twitter_handle, winery. Here, In this report there are analyse using statistical methods for few questions raised as main assignment. In this report, different statistical methods and models as Gibbs sampling or Bayesian model are evaluated and used to compare rating points of wines available.

Question 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.

i.e. From two wines "Sauvignon Blanc - South Africa" and "Chardonnay - Chile" are given, with price limit €15.

Part a.1

Which type of wine is better rated? How much better?

Pre-processing of Data

- Only few columns can be considered here i.e. **country, price, points, region and variety**. We need to select variety from region as 'Sauvignon Blanc from South Africa' and 'Chardonnay from South Africa', points columns represent rating for each wine, variety is type of wine and dataset has around 130k data points.
- Performed Data cleaning after the complete dataset is loaded into R.
- A subset 'winedata' is created from columns- country, price, points, region and variety and, check is performed for missing values, but none found.
- Wines are filtered for price as €15 and converted variety data to be a measurement data so that can be treated as index i.e. changed class of variety object to become factor.

Analysis:

- To understand selected data, generated Summary as:

Table 1: Summary of selected Dataset

```
> summary(test_winedata)
         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
```

- Here in table 1, we can see mean value for points data is 85.67 and is uniformly distributed with minimum and maximum ratings as 90 and 80 respectively. For variety Chardonnay and Sauvignon Blanc type wine count of ratings are 37 and 14 respectively i.e. Chardonnay wine is rated more in count.
- To make further understand of data distribution generated box plot with extra 'jittered data' for variety based on points data
- In Fig 1, wine Chardonnay is represented by red coloured area and Sauvignon Blanc wine is represented by green coloured area. From boxplot it is clear that data for both variety of wine is uniformly distributed and will be fat tailed due to some higher number of outliers present in both wine type.

GitHub link for code: https://github.com/vishalkumarmishra7/TCD_Applied_Statistical_Modelling_CS7DS3.git

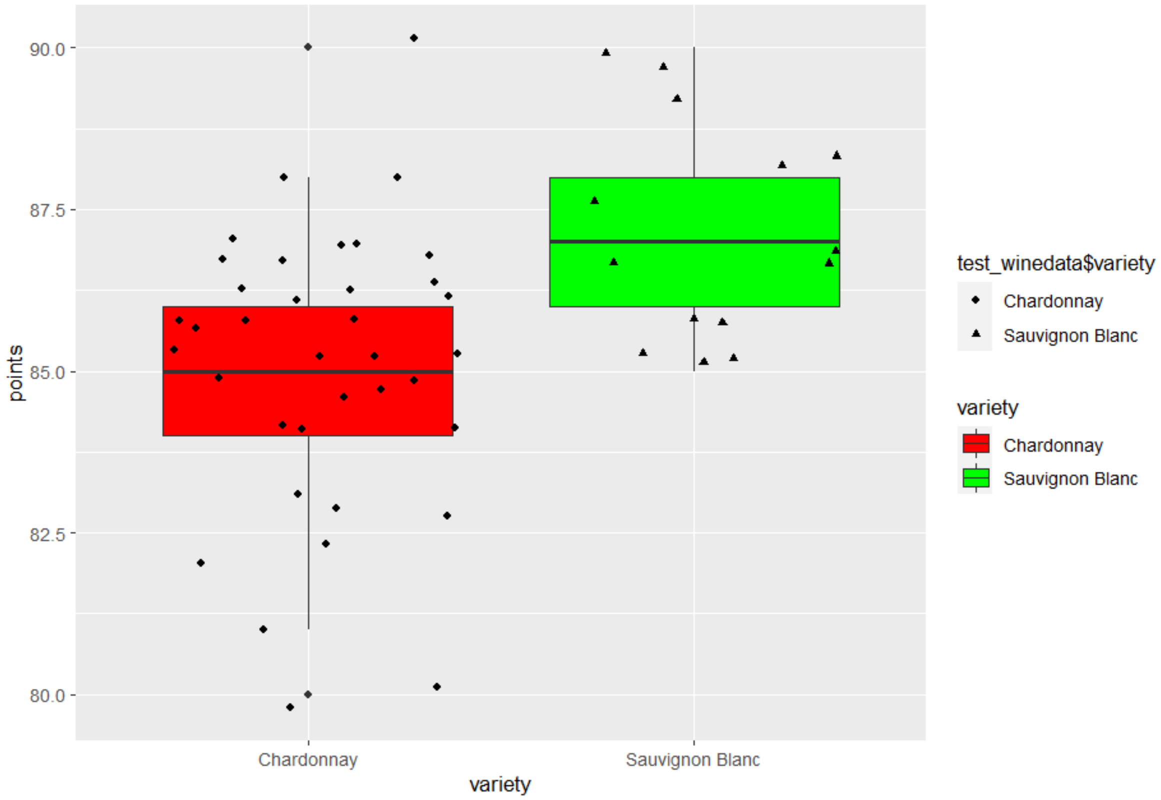


Figure 1 Boxplot for wine variety and rating points data

- Here in table 2, we can analyse average rating and median rating for both variety of wine i.e. average rating for Chardonnay and Sauvignon Blanc type wine are 85.08 and 87.21 respectively, whereas median rating for Chardonnay and Sauvignon Blanc type wine are 85 and 87. The standard deviation for rating points for each Chardonnay and Sauvignon Blanc wine are 2.2 and 1.71.
- Also, lowest 25% of rating i.e. 1st quartile for Chardonnay wine and Sauvignon Blanc wines are given below ~84 and ~86 respectively.
- In addition, median rating for both wine types lies outside the box of comparison boxplot, which implies – There must be some significant difference between two wine types. From this analysis we can assume the null hypothesis: H_0 as there is no difference in means of both wine types i.e. 0 and conduct a t-test.
- From table 3, we can analyse result for t-test. P-value is 0.00203 < 0.05 which is very small value and less than 5% significance level. So, we can reject null hypothesis H_0 with 95% confidence interval i.e. we are 95% confident to say mean of both wine types are different and interval is 0.81 to 3.44

```
	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 3: t-test result for difference in wine types

```
Chardonnay Sauvignon Blanc
  85.08108        87.21429
Chardonnay Sauvignon Blanc
        85              87
Chardonnay Sauvignon Blanc
  2.203260        1.717716
```

Mean

Median

Standard Deviation

Table 2: Mean, median and standard deviation for Wine types

GitHub link for code: https://github.com/vishalkumarmishra7/TCD_Applied_Statistical_Modelling_CS7DS3.git

Conclusion

From all of the above analysis, we can conclusively interpret that **Sauvignon Blanc wine from South Africa is better than Chardonnay from Chile**. And, we came to fact that difference in average rating points for both wines is **2.133 and** quality of Sauvignon Blanc wine is ~2.44% is higher than Chardonnay wine.

Part a.2

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?

Pre-processing of Data

- Here Dataset is already pre-processed, so using same dataset 'test_winedata' with data points as **country, price, points, region and variety.**

Analysis:

- In order to find probability of Sauvignon Blanc is better than Chardonnay, we must look for the difference in means of rating points in both sample (i.e. wine type) which can be done by explicitly modelling it. Here sample of both wine types are small in size, so predicting probability of getting better wine is not easy, also, simulating extra samples from both sample distribution is difficult to perform.
- Keeping these points in focus, we must use Gibbs sampling technique in association with Markov Chain Monte Carlo (MCMC) method. We will find the marginal distribution for both wine types by initially simulating posterior-parameters derived from the generated joint probability distribution.
- Taking Prior parameters as: $mu_0 = 85$, $tau_0 = 1/100$, $del_0=0$, $gamma_0=1/100$, $a_0 = 50$, $b_0 = 1$, maxiter = 5000. Since we do not have any specific method to set a_0 and b_0 so assuming some vague data for both, high and relatively small respectively.
- From analysis of Fig 2, we can interpret normal distribution of simulated posterior mean exists. Maximum probability density for rating points exists at ~86. In addition to this, we have a little right skewed simulation of precision parameter (tau) from gamma distribution.
- Since we have obtained normally distributed observed data and sampled normally distributed posterior-parameter, so we can use these to generate different sample data for both wine type and marginal probability.
- In table 4, we can analyse Gibbs sampler's performance. Dependence factor(I) has value closer to 0 and 1 which implies sampler's performance is better and satisfactory.
- Now we can simulate samples for both wine types using normally distributed input posterior parameters. And, to clearly visualise if auto-correlation of simulated samples do not exist, we can evaluate Auto Correction Functions (ACF) for both samples.
- From Fig 4, we can find a highly correlation of initial lag value only. And, rest other lags have significant level which implies samples generated do not show a high correlation with previous lags at any specific point given.
- Other plots are generated to demonstrate rating points correspondence to both wine types sample, these sample are simulated using posterior parameters obtained by performing Gibbs Sampling. Like Figure 5(a), shows PDF for difference in two simulated data samples and 5(b), represents Joint PDF of 2 wine samples with trace range of 1 to 10.

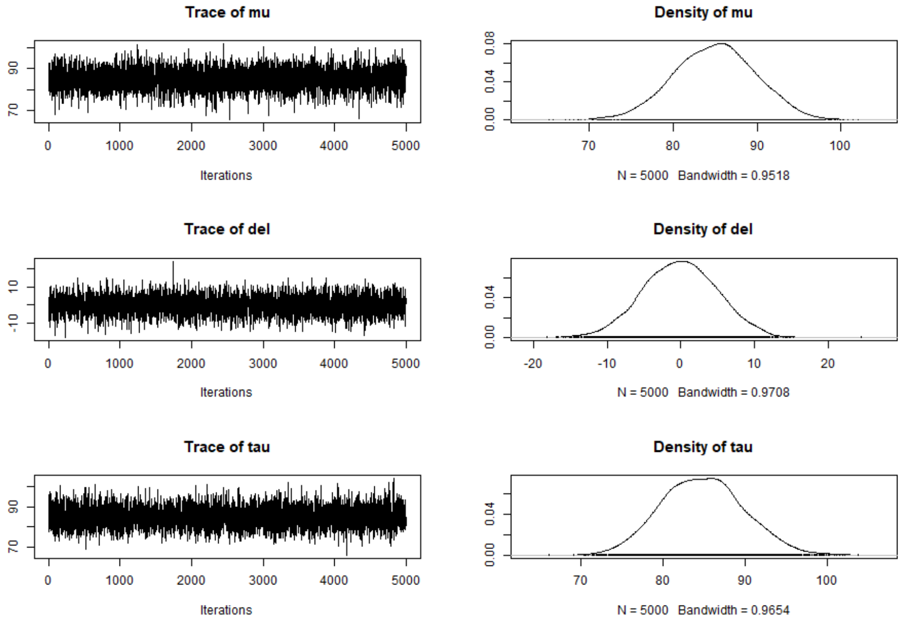


Figure 2: Displaying basic properties of posterior distribution

Table 2: Displaying Gibbs Sampler performance

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

	Burn-in (M)	Total (N)	Lower bound (Nmin)	Dependence factor (I)
mu	2	3620	3746	0.966
del	2	3803	3746	1.020
tau	2	3741	3746	0.999

Figure 3: Summary of posterior distribution parameters

```
> apply(fit, 2, mean)
         mu         del         tau
85.06597230  0.02318141 84.98665496
>
> apply(fit, 2, sd)
      mu      del      tau
5.025802 4.924854 5.000691
> mean(1/sqrt(fit[, 3]))
[1] 0.1086151
> sd(1/sqrt(fit[, 3]))
[1] 0.00320712
```

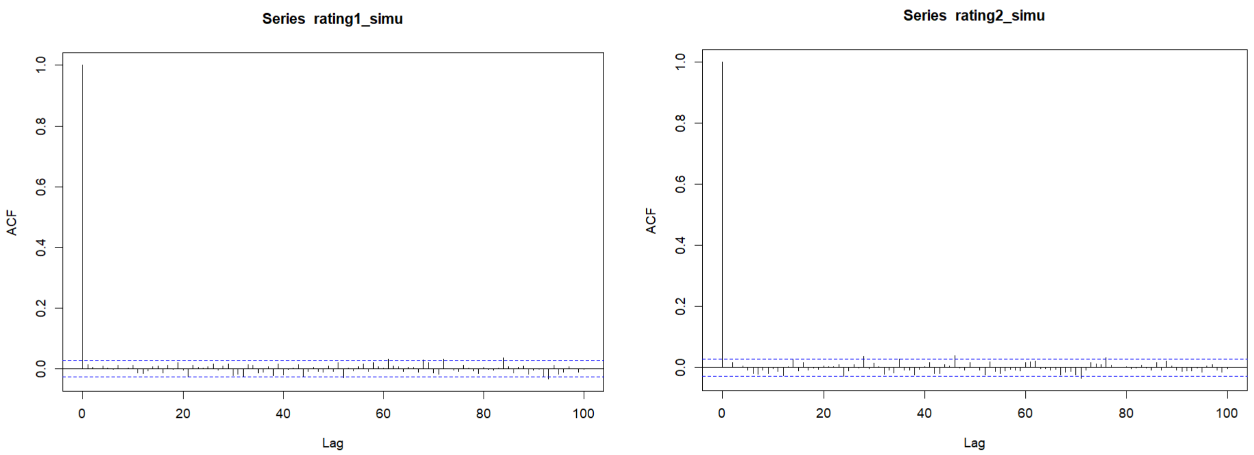


Figure 4: Auto Correlation Function for both wine type samples

GitHub link for code: https://github.com/vishalkumarmishra7/TCD_Applied_Statistical_Modelling_CS7DS3.git

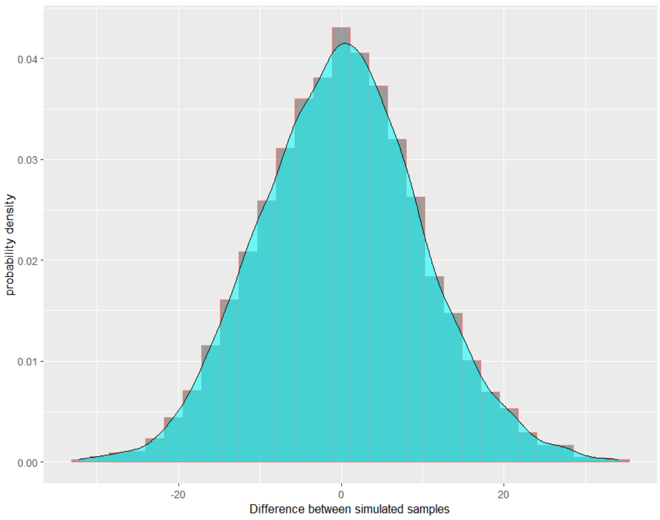


Figure 5(a): PDF for variation in two simulated data sample

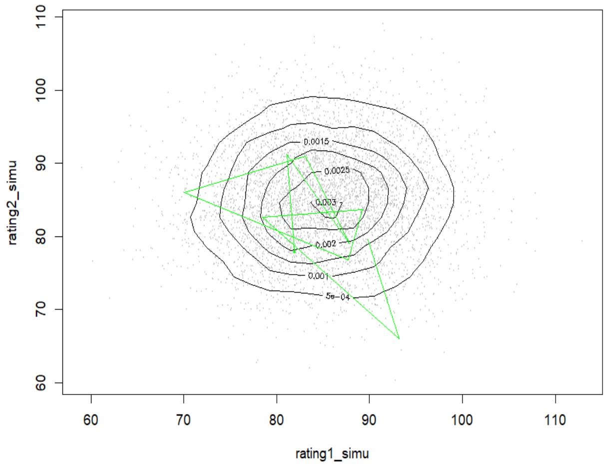
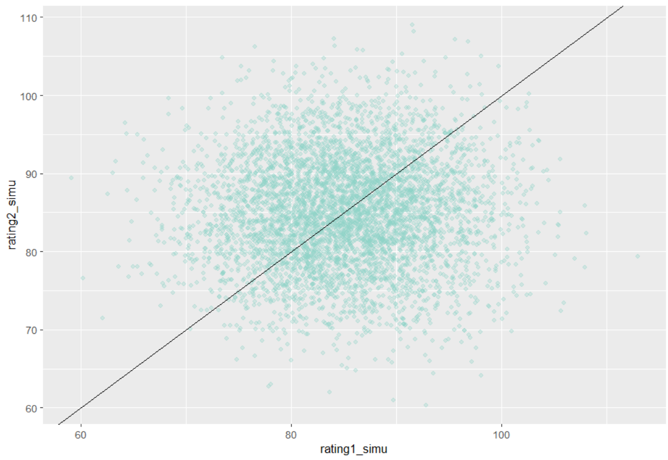


Figure 5(b): Joint PDF of 2 wine samples




```
> mean(rating1_simu > rating2_simu)
[1] " 0.7102"
```

Figure 6: Compare probability of both wine type

Conclusion

From Fig 6, We can conclude that there is **0.71 probability of Sauvignon Blanc wine is better than Chardonnay**. Here, rating1_simu and rating2_simu are simulated samples representing Sauvignon Blanc wine and Chardonnay wine respectively.

Part 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.

Pre-processing of Data

- Here we must consider complete dataset and filter by region as 'Italy', wines with price only below €20 and minimum count of reviews as 4. After filtering we observe ~4.7k data points.
- Some missing are present for region_1 so removed them. Summary of data is presented in figure 7.

```
> summary(test_winedata_4)
      country          points          price                                                 region_1              variety
 Italy     :4702   Min.   :80.00   Min.   : 5.00   Sicilia                                    : 418   Red Blend   : 821
           :   0   1st Qu.:86.00   1st Qu.:13.00   Toscana                                    : 230   Glera       : 351
 Argentina:   0   Median :87.00   Median :15.00   Chianti Classico                           : 182   Pinot Grigio: 346
 Armenia  :   0   Mean   :86.59   Mean   :15.02   Alto Adige                                 : 165   Sangiovese  : 310
 Australia:   0   3rd Qu.:88.00   3rd Qu.:17.00   Conegliano Valdobbiadene Prosecco Superiore: 126   White Blend : 250
 Austria  :   0   Max.   :93.00   Max.   :19.00   (Other)                                    :3573   Nero d'Avola: 180
 (Other)  :   0                                    NA's                                       :   8   (Other)     :2444
```

Figure 7: Summary of filtered dataset

Analysis:

- To analyse various wines available in multiple region_1, from **Fig. 8** we can visualise distribution of rating points for wines among several region_1.
- In **Fig. 9**, we can also analyse count of rating points for wines belonging to different region_1. Here we say that there are very few regions with significant wine review count more than 50. We can also analyse review count w.r.t wine rating points (**Fig. 10**), histogram shown represents most of the reviews given has rating in range of 85 and 90. From **Fig. 11**, in scatter plot it is visible that wine review rating point shifts towards mean with increasing sample size.
- We can also infer; wine rating points is greatly influenced by user review counts available i.e. higher wine ratings are found when sample size is small.
- Again, we model difference in mean wine rating points among available regions. Here, taking same prior parameters as before, since this time dataset is larger in size and there are much more parameters to be sampled, model takes extra time for execution. We can get 2 outputs as- params representing posterior mean, del and tau and theta(θ) is the simulated mean parameters for every region.
- In **Fig. 12**, we can analyse a linear relationship between- sorted mean of wine rating points and their available regions.
- In addition, generating **Table 4**, upon sorting average of wine rating points among region_1, we can find ‘Trento’ region is on top i.e. has received highest rating points for its wine.

Figure 8: Ratings for Wines with regions as Italy

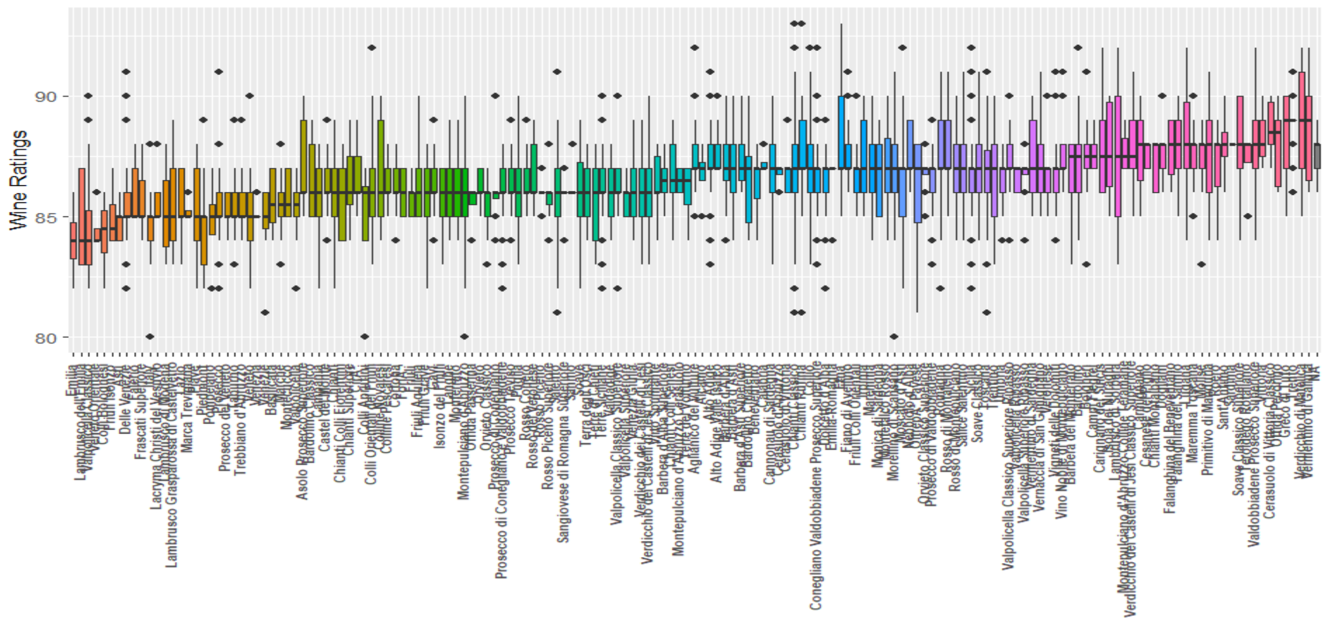
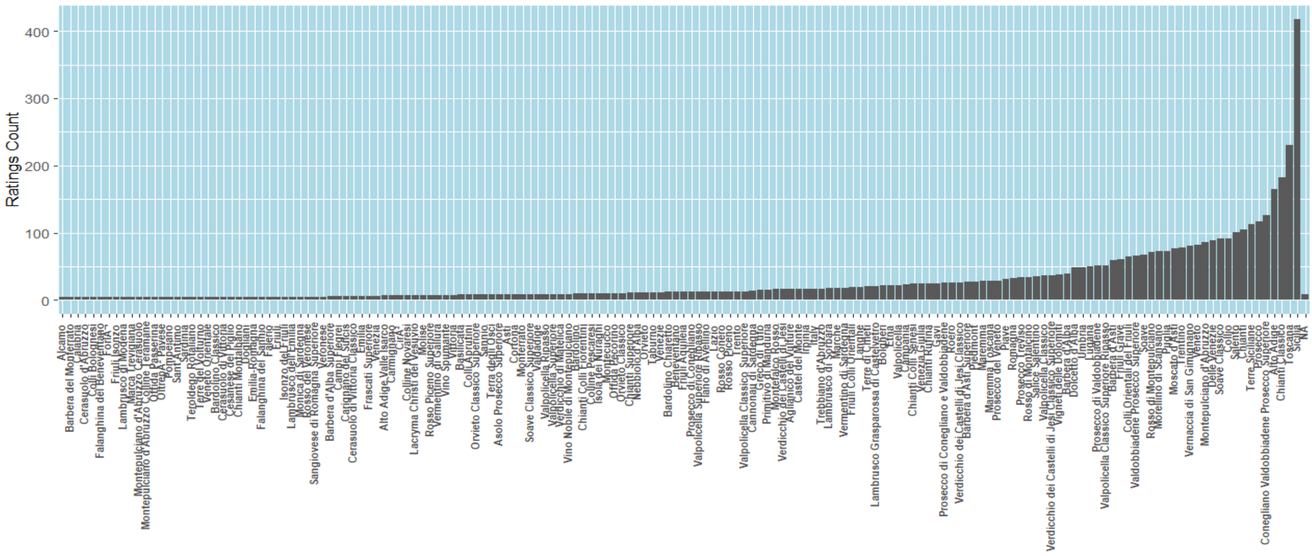
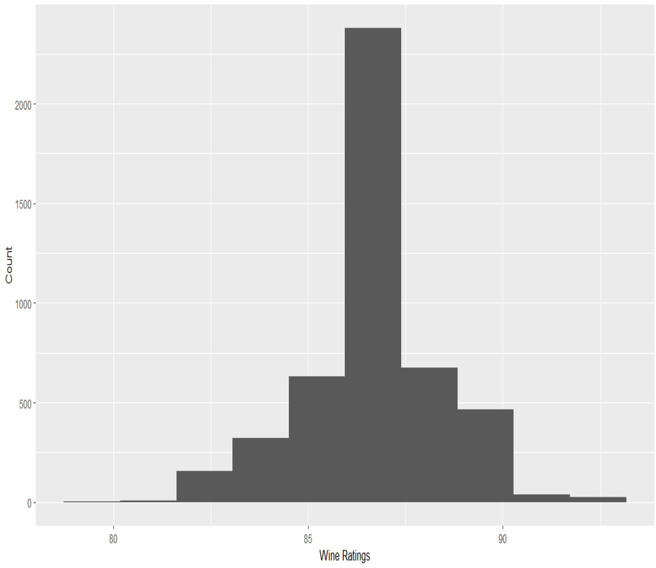


Figure 9: Variety of Wines with regions as Italy





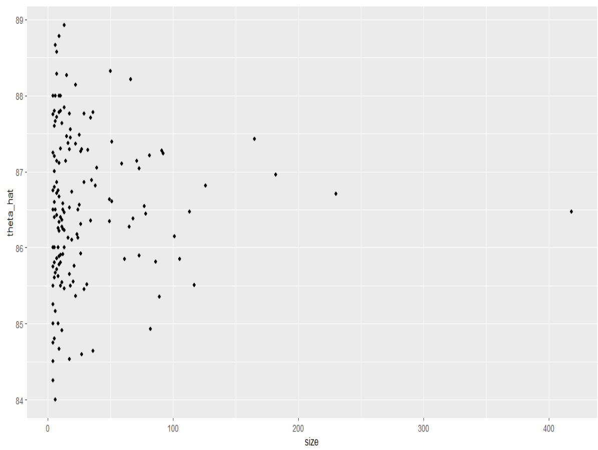


Figure 10: Range and Frequency of Wine Ratings in Italy

Figure 11: Effect size vs Sample size

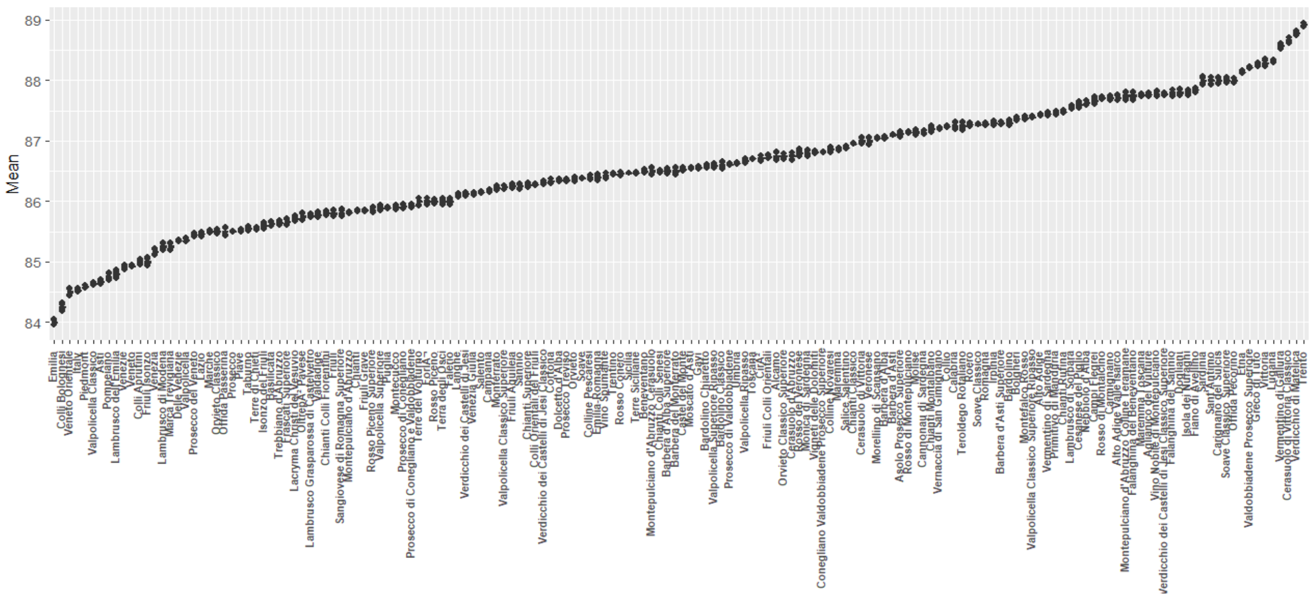


Figure 12:Sorted mean ratings for wines for Italy region from generated samples

GitHub link for code: https://github.com/vishalkumarmishra7/TCD_Applied_Statistical_Modelling_CS7DS3.git

Table 3: Sorted value of rating for each region_1

	sort.theta_hat..decreasing...TRUE.
Trento	88.92301
Verdicchio di Matelica	88.77746
Cerasuolo di Vittoria Classico	88.66602
Vermentino di Gallura	88.57069
Lugana	88.32076
Vittoria	88.28458
Greco di Tufo	88.26756

Conclusion

- After analysis from all tables and figures, we can find regions which produce wines whose better than available average wines i.e. we compared average of posterior joint distribution mean with found theta (simulated rating point) for each region present.
- From theta results we can find regions whose wine are better than average (few are shown below)

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 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

Part 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.

Pre-processing of Data

- Wine dataset (winemag-data-130k-v2.csv) is selected and filtered for country as 'US'. After checking missing values are removed rows where price is not available. Only 239 rows are removed and new dataset contains ~54K datapoints.
- After taking a glimpse on columns type, only 2 columns(point and price) have in Integer type data and fields are in categorical form.
- Extracted new features from description data as word count and sentence count.
- Removed some fields we are not interested in for our model i.e. country, taster_twitter_handle, designation and region_2.
- Added new field as log of price and length of description.
- Here, For categorical fields like ('province', 'region_1', 'taster_name', 'variety' and 'winery') ordinal encoding method is used to encode categorical fields into numerical type. Encoding is done after factorization method to get levels.
- Created variety subset with only top 30 data counts and will use in one specific model to check its influence.

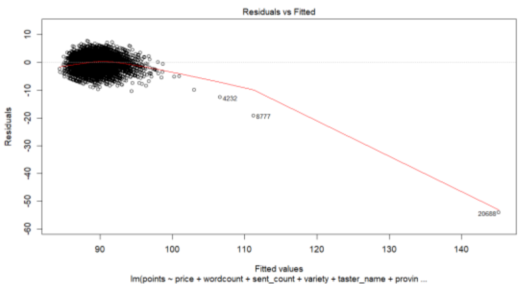
Analysis:

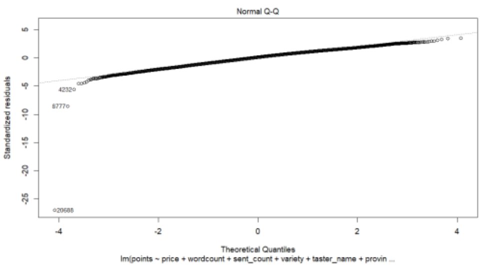
- To start analysis initially checked frequency for rating points available. In **Fig. 13**, from histogram it is clear that rating points are normally distributed, and number of ratings given in range 85 to 95 are much than rating points available more than 95.
- Since our objective is to find most influential factors for rating points. Here our response variable is points and other variables are predictors. In **Table 4,** we must analyse correlation of all selected predictors and response altogether. It is evident that only price and wordcount are significantly correlated to points.
- Also, we can check correlation of points with log of price and description length using scatter plots in **Fig. 15** and in **Fig.14** we can find which wines are rated higher on average.
- In Fig. 16, it is interesting to know what kind of words are usually seen in good of bad reviews and how much points are affected.
- Here, in **Table 5** we have found maximum correlation between variables and its is clear wordcount and price are influencing points significantly. Now we will check different models with variables selected using AIC methods and compare the linear regression models
- With summary of linear regression model, for each model we will generate 2 diagnostic plots namely- residuals vs fitted value and Norma Q-Q plot.
- **Residual vs fitted plot** is most significant as it informs about patterns found in residuals. It is mainly used to find linear relationship among assumptions, it also tells about the comparison of residuals under experiment with values fitter in model. A straight line is a good indicator of strong relationship.
- **Normal Q-Q plot** helps in describing normal distribution of residuals. When residuals follow the straight line, model is considered good.
- Before selecting predictors for model, we must know AIC values for all, which is illustrated in **Fig. 17.** Using this List as reference we will create model and calulate AIC values and compare Multiple R^2 values
- **Table 7,** describes terms used in illustrate summary of models.

Model 1:

- Estimating points from selecting predictor as log of wordcount, price, sent_count, variety, taster_name, province and winery.

```
Call:
lm(formula = points ~ price + wordcount + sent_count + variety +
    taster_name + province + winery, data = us_dataset_lm)

Residuals:
    Min      1Q  Median      3Q     Max
-54.134  -1.484   0.135   1.619   7.751

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.446e+01  8.833e-02 956.202  < 2e-16 ***
price        2.846e-02  5.380e-04  52.892  < 2e-16 ***
wordcount    1.044e-01  1.670e-03  62.547  < 2e-16 ***
sent_count  -5.661e-03  2.584e-02  -0.219  0.82662
variety     -4.595e-03  8.426e-04  -5.453    5e-08 ***
taster_name -5.748e-03  1.185e-02  -0.485  0.62760
province    -7.492e-02  2.480e-02  -3.020  0.00253 **
winery      -4.210e-04  2.342e-05 -17.977  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.27 on 22379 degrees of freedom
Multiple R-squared:  0.3471,    Adjusted R-squared:  0.3469
F-statistic:  1700 on 7 and 22379 DF,  p-value: < 2.2e-16
```



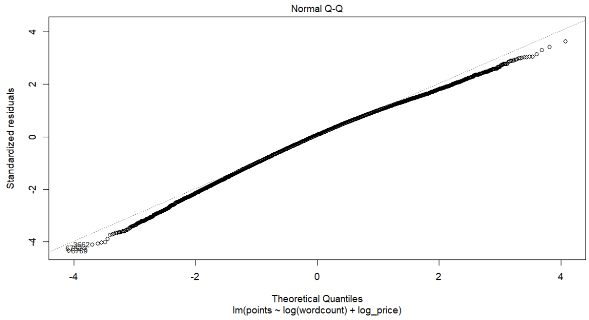


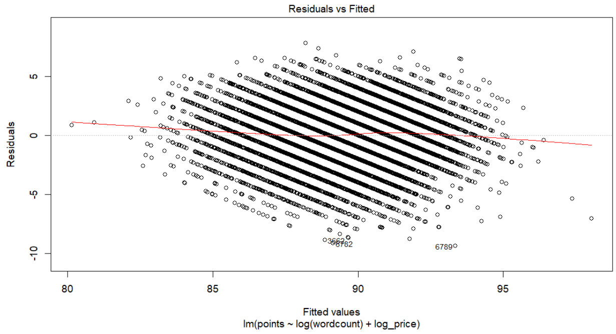
- Here we can find in summary that price and wordcount has higher influence and R-squared value is not quite good implying model fit is not very well. Also Regression line is not well fitted in Residual vs fitted plot and Normal Q-Q plot, representing large error values.

Model 2:

- Estimating points from selecting predictor as log of wordcount and log of price.

```
# MODEL 1:
lm_1 <- lm(points ~ log(wordcount)+log_price, data = us_dataset_lm)
summary(lm_1)

Call:
lm(formula = points ~ log(wordcount) + log_price, data = us_dataset_05)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7091  -1.4788   0.0955   1.5790   8.3969

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    64.78823    0.11162   580.4   <2e-16 ***
log(wordcount)  4.60814    0.03243   142.1   <2e-16 ***
log_price       2.01704    0.01791   112.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.223 on 54262 degrees of freedom
Multiple R-squared:  0.4911,	Adjusted R-squared:  0.4911
F-statistic: 2.618e+04 on 2 and 54262 DF,  p-value: < 2.2e-16
```



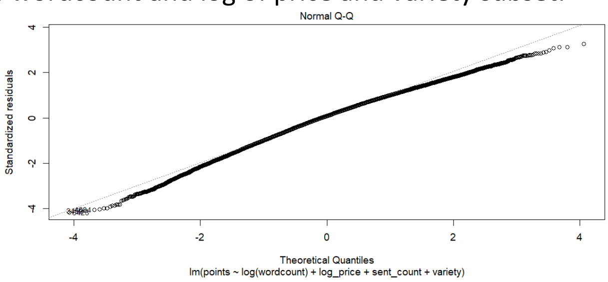


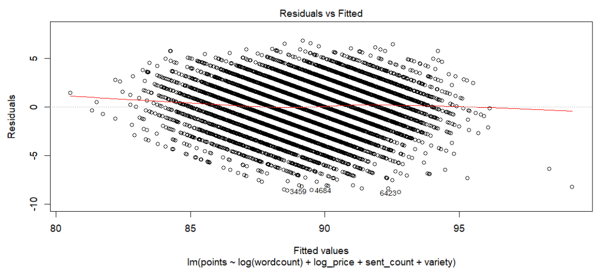
- Here, we can find effective coefficients for log of wordcount and price values. And. Significant change in R-squared value and well fitted regression line in Residual vs fitted plot and Normal Q-Q plots.

Model 3:

- Estimating points from selecting predictor as log of wordcount and log of price and variety subset.

```
Call:
lm(formula = points ~ log(wordcount) + log_price + sent_count +
    variety, data = us_dataset_lm_new)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3495  -1.4521   0.0955   1.5450   7.5788

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 63.68053    0.18797 338.775  < 2e-16 ***
log(wordcount)               4.72547    0.04002 118.077  < 2e-16 ***
log_price                    2.18096    0.02061 105.835  < 2e-16 ***
sent_count                  -0.07816    0.01463  -5.342 9.22e-08 ***
varietyBordeaux-style Red Blend -0.23802    0.15070  -1.579  0.11425
varietyCabernet Franc       -0.38928    0.15720  -2.476  0.01328 *
varietyCabernet Sauvignon   -0.03668    0.14389  -0.255  0.79878
---
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.179 on 50996 degrees of freedom
Multiple R-squared:  0.5154,	Adjusted R-squared:  0.5151
F-statistic:  1695 on 32 and 50996 DF,  p-value: < 2.2e-16
```





GitHub link for code: https://github.com/vishalkumarmishra7/TCD_Applied_Statistical_Modelling_CS7DS3.git

Here, we can find effective coefficients for log of wordcount and price values. And, much better change in R-squared value and well flat regression line in Residual vs fitted plot and Normal Q-Q plots, indicating best fit as per model generated.

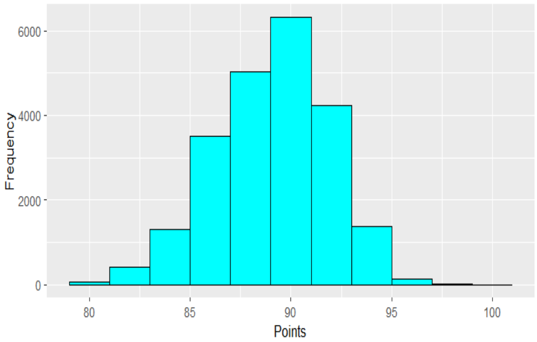
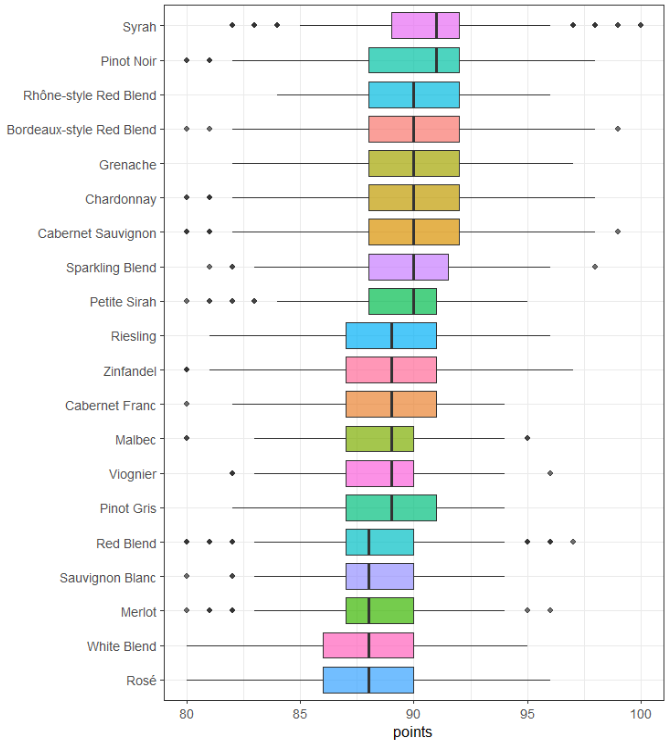


Figure 13: Frequency distribution of rating points

```
> cor(subset(us_dataset_04, select=c(points,wordcount,price,province,region_1,taster_name,winery
                points   wordcount       price    province    region_1 taster_name      winery
points      1.00000000  0.49743670  0.39462635 -0.07449474 -0.02578292 -0.08542033 -0.11881395
wordcount   0.49743670  1.00000000  0.20681615 -0.06729789  0.01192498 -0.07879875 -0.04236022
price       0.39462635  0.20681615  1.00000000 -0.10137637 -0.01532389 -0.11270864 -0.01098285
province   -0.07449474 -0.06729789 -0.10137637  1.00000000  0.02082468  0.33760314 -0.05933227
region_1   -0.02578292  0.01192498 -0.01532389  0.02082468  1.00000000  0.20385914  0.06237331
taster_name -0.08542033 -0.07879875 -0.11270864  0.33760314  0.20385914  1.00000000  0.06882017
winery     -0.11881395 -0.04236022 -0.01098285 -0.05933227  0.06237331  0.06882017  1.00000000
```

Table 4: Correlation matrix for all selected columns



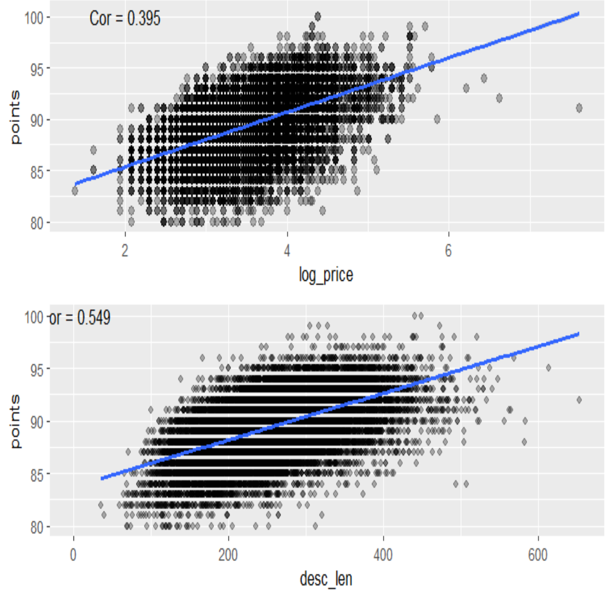


Figure 14: Wine Ratings by Varietal

Figure 15: Correction of Price with log of price and description length

Conclusion

With all tables and figures that we have used much better fitted model is generated to estimate wine rating point. And, we can conclude that **wine price and wordcount derived from review description can influence mostly the rating points,** must be considered to obtain good rating. Apart from these, other features can be derived from review description like good or bad words can also be utilized to make a better fit model

GitHub link for code: https://github.com/vishalkumarmishra7/TCD_Applied_Statistical_Modelling_CS7DS3.git

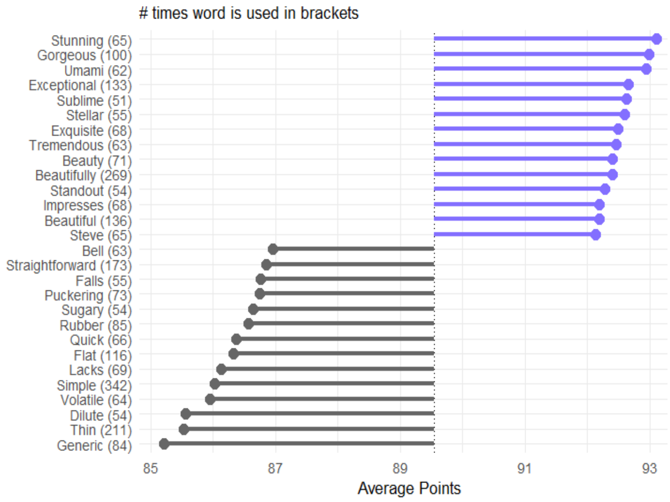


Figure 15: What Words Show up In Good/Poor Wine Reviews

```
> AIC_step_forward <- step(lm_base)
Start:  AIC=34502.04
points ~ log_price + wordcount + sent_count + variety + taster_name +
    province + winery

              Df Sum of Sq    RSS   AIC
- province     1       0.0 104473 34500
- variety      1       2.2 104475 34501
<none>                     104473 34502
- sent_count   1      14.5 104488 34503
- taster_name  1      47.0 104520 34510
- winery       1    1231.4 105704 34762
- wordcount    1   14782.9 119256 37463
- log_price    1   25295.3 129768 39354
```

Tabe 6 AIC summary

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

Table 7: Terms used in summary

GitHub link for code: https://github.com/vishalkumarmishra7/TCD_Applied_Statistical_Modelling_CS7DS3.git

Reference

1. https://www.kaggle.com/zynicide/wine-reviews
2. http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/
3. https://www.rdocumentation.org/
4. https://online.stat.psu.edu/stat504/node/168/
5. https://www.methodology.psu.edu/resources/AIC-vs-BIC/
6. https://www.scss.tcd.ie/~arwhite/Teaching/CS7DS3/Regression_Case_Study.pdf

GitHub link for code: https://github.com/vishalkumarmishra7/TCD_Applied_Statistical_Modelling_CS7DS3.git