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

School of Computer Science and Statistics

Assessment Submission Form

Student Name	Kul Gaurav
Student ID Number	19307204
Course Title	MSc. Computer Science (Data Science)
Module Title	APPLIED STATISTICAL MODELLING
Assessment Title	Main Assignment
Date Submitted	15 May 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.

Introduction

The given wine review dataset consists of 129,971 observations, which is gathered from WineEnthusiast portal. There are fourteen variables in total covering the following columns:

- X: the numerical value of the row
- Title: contains the name of the review and even vintage for a few observations
- Variety: type of grapes the wine is made of
- Points: rating points distributed between 1 to 100
- Description: review text for the observation
- Country: origin country of the wine in consideration
- Province: origin state in the given country
- Region_1: area in the province
- Region_2: closer proximity to the wine-growing farm
- Winery: distillery for the wine manufacturing
- Designation: name of the vineyard
- Price: cost for a bottle of wine
- Taster_name: name of the person who gave the review
- Taster_twitter_handle: Twitter id for the person giving the review

The report covers various statistical models and data analysis techniques to answer the evaluation questions. R and Python languages are in use for the process. This document provides the analysis in simple language which can help non-technical readers and also contains in-depth analysis for technical readers. Complete code and the mechanism are available on the GitHub repository with the link provided at the end of the report.

Data Handling

We are reading the data using version two of the available CSV file. We perform the data sanity check using str() in R and find that all the columns are matched with the right data types, and we do not need to change the data type for any column as of now. We are considering the “Description” column to verify if there exists any duplicate observation as this column contains free text, and the exact two rows containing the same value will imply duplicate values. If we do not remove the duplicates, they may bias our models. Duplicates are taken out of the dataset, and the file is saved as a new version. The new file will serve as the main file for the rest of the assignment. Further data handling and data preparation are done individually for each part of the assignment.

Question Wise Analysis

This section provides data handling, analysis, and conclusion for each question and sub-question independently. The following approach is considered as there is little overlapping subproblem in various questions.

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

We are considering two kinds of wines for this question:

- Sauvignon Blanc from South Africa
- Chardonnay from Chile

Data Handling

We subset the data for the two types of wines with a price of 15 units. There are a total of 51 observations into consideration. As we need to identify the better-rated wine, we will focus on two variables, namely:

- Points
- Variety

Analysis

As we have our "Variety" limited to two unique values, we factor the data object to categorize and store it as two distinct levels. We plot the distribution of the "variety" and "points" on a box plot:

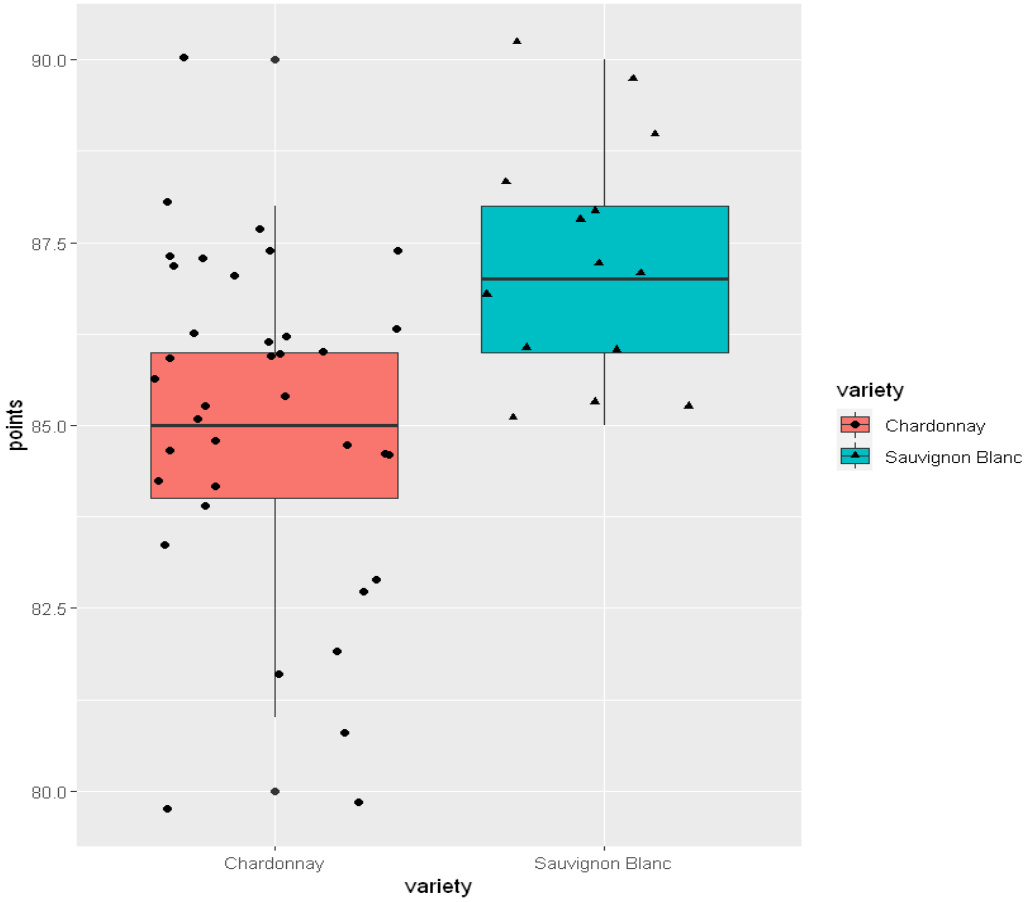

Figure 1: Points vs. Variety Box plot

The median of "points," representing the standard measure of the center of the observations, is higher for Sauvignon Blanc. Also, the points for this variety are skewed to a more upper side. With the

consideration that both the wines have a normal distribution with nearly equal variances, we perform the t-test to determine if the means of the two wines are identical.

```
1  t.test(points ~ variety, data=wine, var.equal = TRUE)
2
```

```
        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
```

Figure 2: t-test output

Following results are inferred from the t-test:

- t-test statistic value is at -3.259
- degrees of freedom are 49
- p-value or the significance level of the test is 0.00203, which is less than significance level alpha, and hence the two wines' average rating is significantly different.
- The confidence of 95 % validates our hypothesis of Sauvignon Blanc to be better.

Conclusion

The Sauvignon Blanc is rated better than Chardonnay with the difference in mean of 2.13 points. Though the t-test won't give the exact value of by how much the Sauvignon Blanc is better, we can obtain the percentage of difference as 2.50%. Further, running Gibbs sampler will identify the quantitative difference in terms of a probability for the two wines.

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 determine the probability of Sauvignon Blanc being better, we perform Gibbs sampling, which is a Markov chain Monte Carlo (MCMC) algorithm.

Analysis

Considering the function for compare_2_gibbs in the case study, we update the hyperparameters mu0, tau0, del0, gamma0 per the value of a0 and b0. These hyperparameters specify the prior distribution for the model. For the case, a0 is 85, and b0 is 3. We obtain the placement of remaining variables, which is a posterior distribution with the condition on observed data. The model generated is fit to the data.

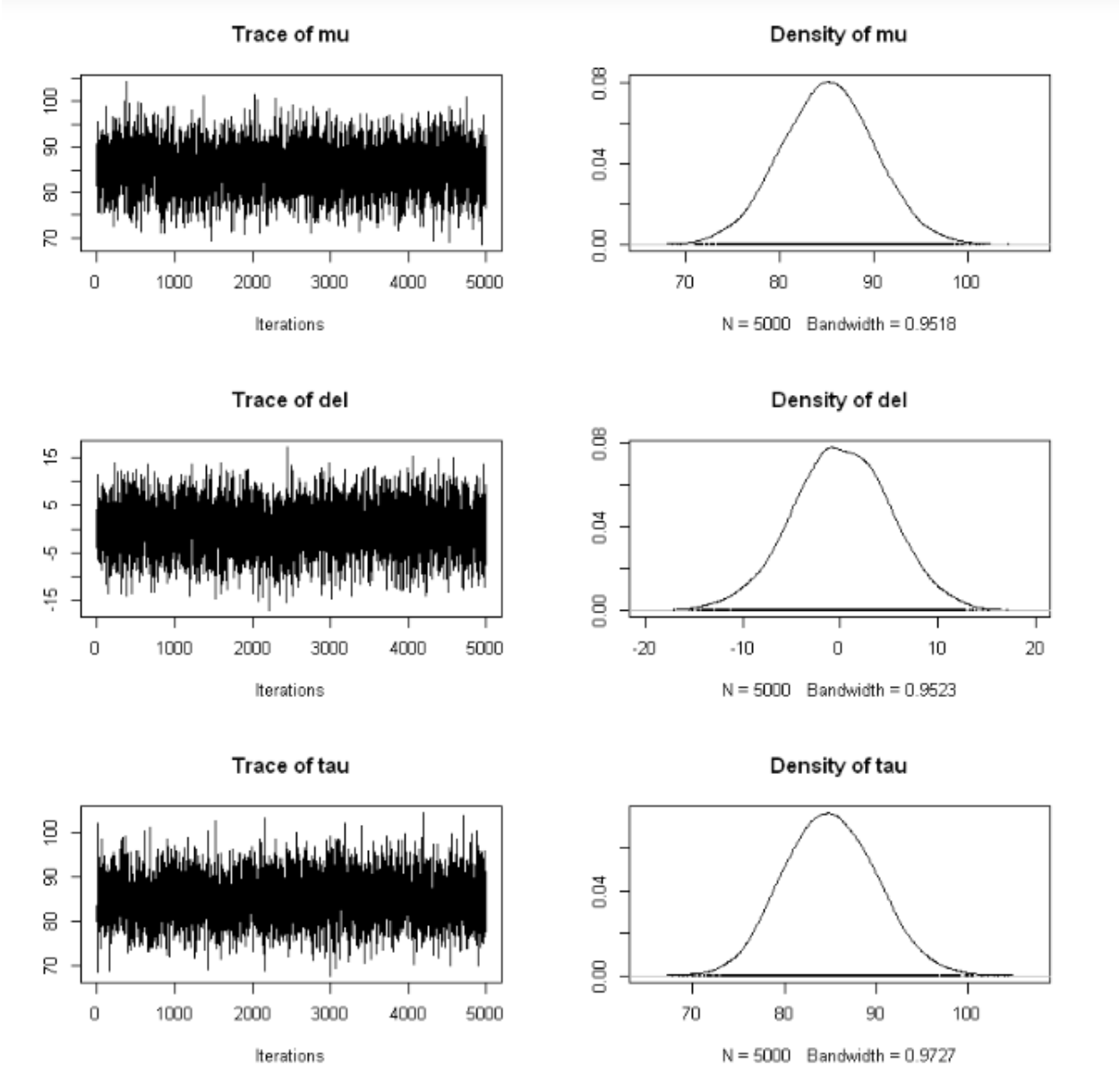

Figure 3: Trace and density plots after running Gibbs sampler

The trace plots represented the time series division of the MCMC algorithm and distributed over 5000 iterations. The density plots for mu, del, and tau are appropriately typical.

The model using Gibbs sampler can help in predictions for the new data set. We generate two multivariate normal random variates of size 5000. The distribution of the difference between the two generated sets is shown in figure 4.

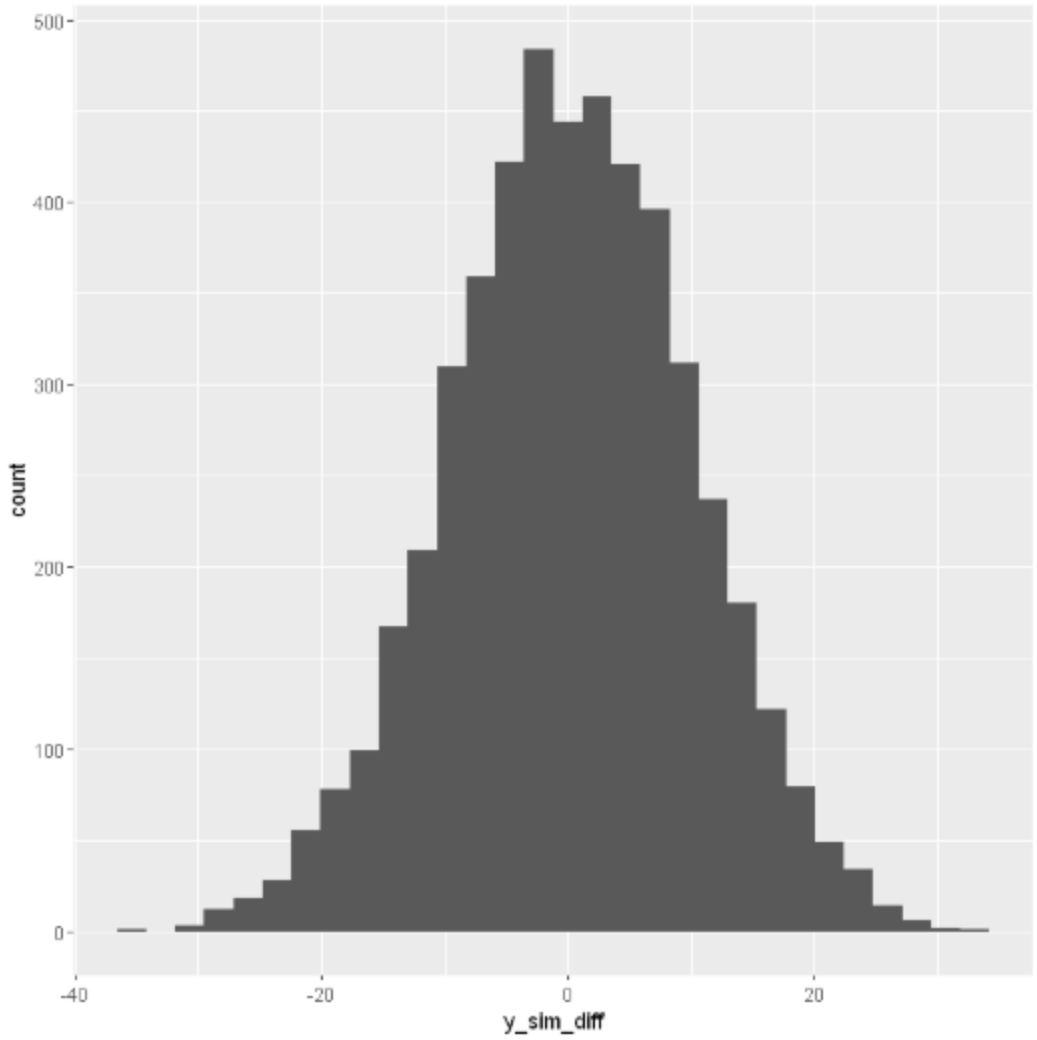

Figure 4: Distribution of simulated data points difference

Conclusion

From the mean probability for Sauvignon Blanc to be better is around 0.712 or 71.2%. Though much difference was not obtained on changing the hyperparameters; a0, and b0, which is gamma prior to overall precision tau, if chosen abruptly gives distorted skewed density plot.

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.

We aim at identifying the areas which produce better than average wine in Italy.

Data Handling

For the case, we need to filter the dataset with a price of less than 20, country to be Italy, and regions with a minimum of four reviews.

```
wine2 <- wineFullData[ which((wineFullData$price < 20) & (wineFullData$country == "Italy"))  , ]
```

```
wine2 <- wine2 %>%
  group_by(region_1) %>%
  filter(n() >= 4)
```

Figure 5: Filter for Italy wines in consideration

Analysis

We factor the observations on “region_1” field and obtain 163 levels. The box plot for the levels after ordering based on the median of points is in the figure below.

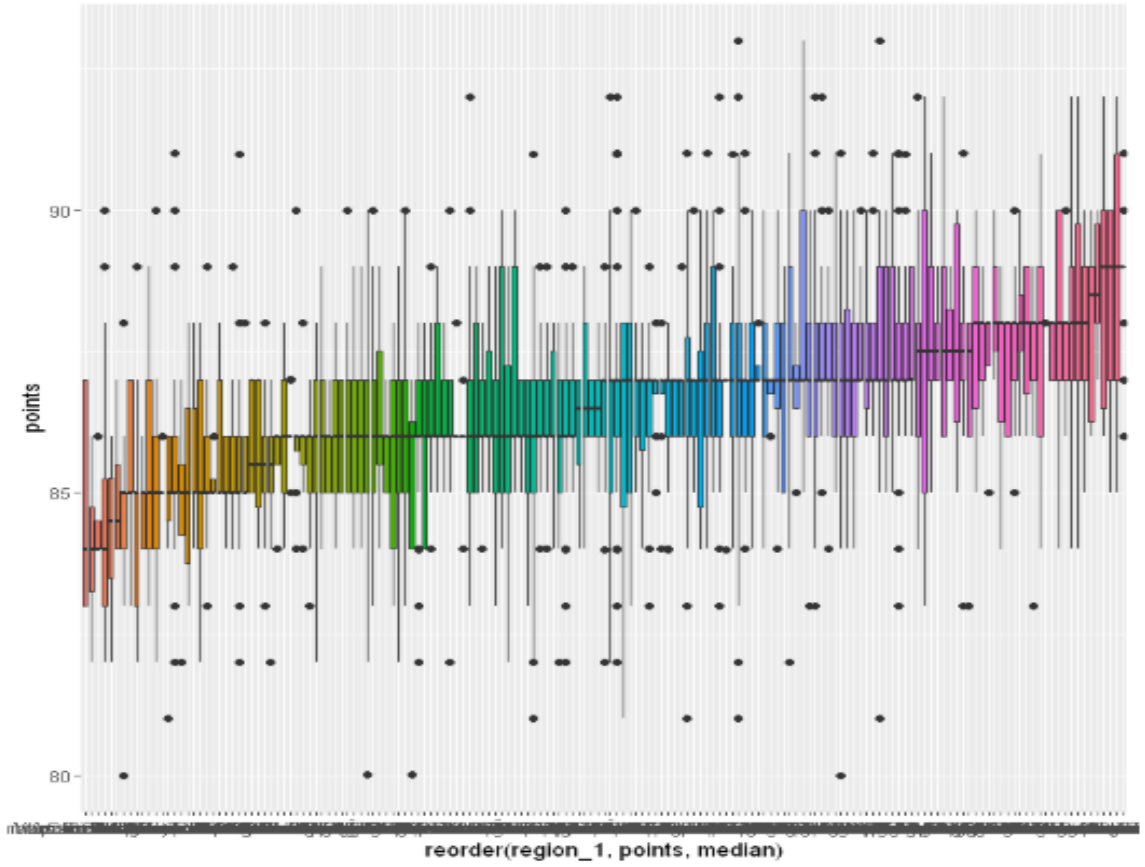

Figure 6: Box plot distribution of points in different regions

The points in “region_1” are distributed evenly, and we do not see much skewness in the pattern.

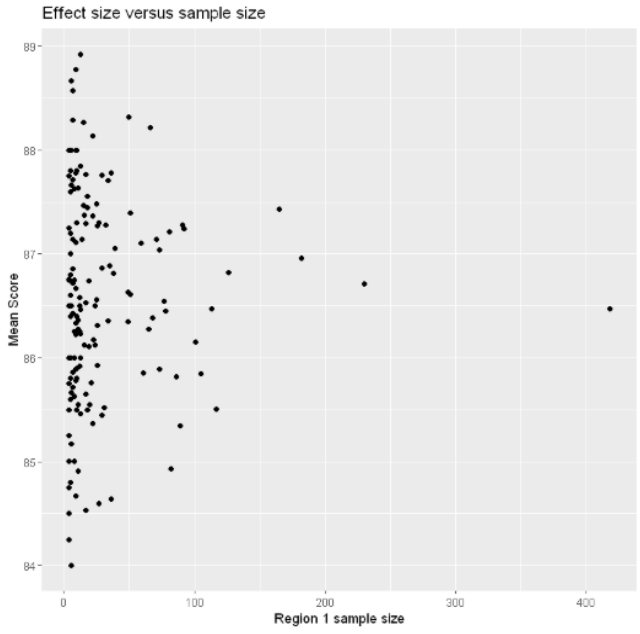

Figure 7: Mean score with increasing sample size

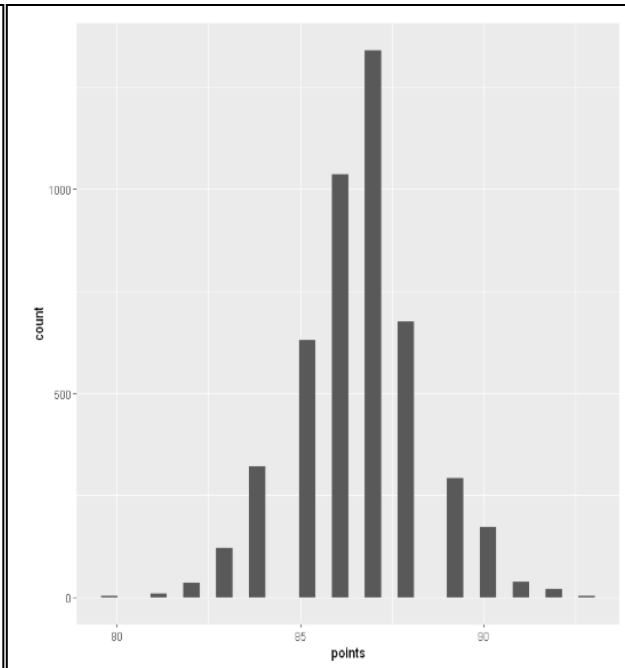

Figure 8: Distribution of points

From figure 7, we see that that most of the observations are concentrated for a smaller sample size of around 50. Though with the increase in sample size, the mean tends to normalize. Also, the points are distributed normally and with the weak around 87 points. The trace and density plot for mean, delta, and precision after performing Gibbs sampling gives normal distribution and evenly distribution of time series for iterations.

Conclusion

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

Figure 7: List of regions in Italy with better than the average wine

Around 71 regions are present in Italy that produce better than average wine. Figure 9 lists all those regions. It will be interesting to find the changes if we iterate the *compare_m_gibbs* method for a larger value.

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.

For this part of the assignment, we perform analysis on the dataset observations for the wines from the US.

Data Handling

Most of the columns in our dataset are object type, so we need to identify which can contribute to the model if we do not perform text analytics. We determine the number of unique instances for each such column in the data.

Column	Unique Instances
country	1
description	50449
designation	14184
province	27
region_1	265
region_2	18
taster_name	16
taster_twitter_handle	13
title	50229
variety	257
winery	5375

Table 1: Number of unique values for columns

Following significant points are considered for data handling:

- Since description, designation, the title have a lot of unique values, we do not consider them for building models. But we save the word count for "description" from verifying later if it is related to "point." Also, we find that the year of wine release is available in "title," we extract and save it as another column.
- Twitter handle and taste name represent the same person, so we drop taster's twitter handle too.
- After removing the unnecessary columns, we remove the duplicate values, which will increase the bias of the model if left unattended.
- We fill the NA values for "region_1", "region_2," and "taster_name" with "unknownR1", "unknowR2," and "unknownT" respectively.
- There are observations with empty price value; we fill it with the mode of price grouped by variety.
- Outliers for "price" and "points" are removed, calculating lower and uppercut limits. We calculate the lower cut limit as (mean – (standard deviation * 3)) and uppercut limit as (mean + (standard deviation * 3)).

Analysis

We start with plotting word count for description and points to identify if there is a trend. Figure 7 shows the obtained result.

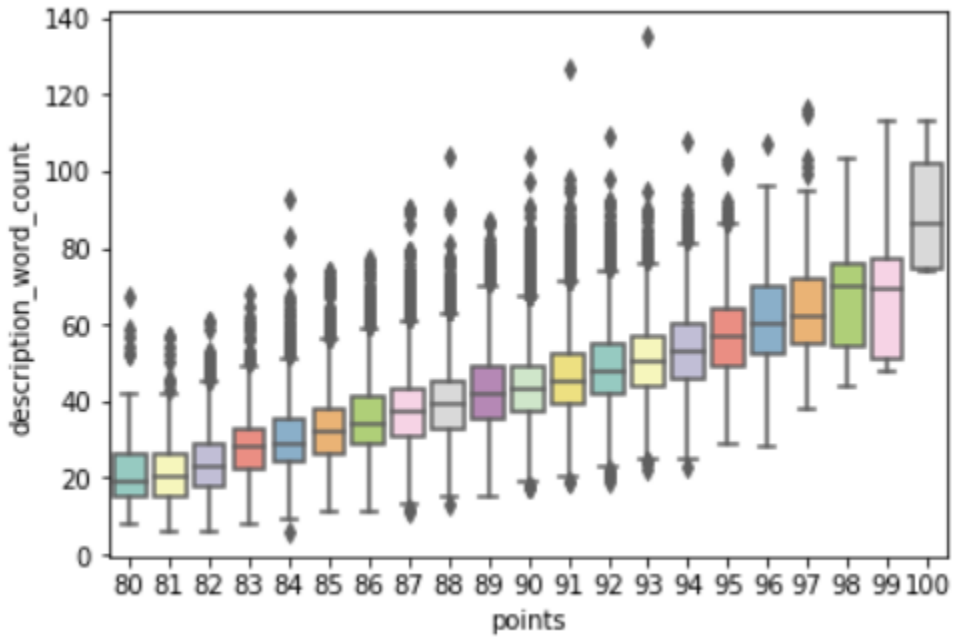

Figure 8: Relation between word count of description and points

The word count seems one of the factors for increasing "points" in an observation. Further, plotting the "price," we find that the data is skewed, as shown in figure 8. For linear regression, we assume that data has a normal distribution. Also, there is multivariate normality. We perform the log transformation on "price" and obtain figure 9.

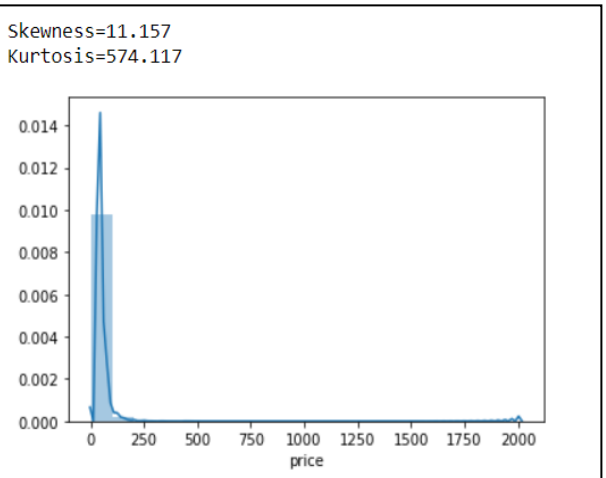

Figure 11: Price distribution

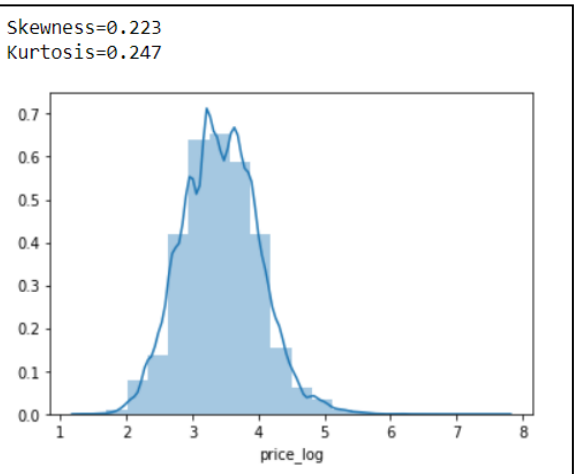

Figure 12: Log (Price) distribution

We apply linear regression on the cleaned dataset and analyze the coefficients assigned for various variables. “Price,” “Region_1,” and “description_word_count” are the most useful variables for the model.

Following the above model, we run stepwise regression and obtain a reduction in the AIC approximation score. After excluding the “Region_2” and “Winery,” the result is enhanced.

Conclusion

We received a good initial model as a result of better data handling. The three variables affecting most further increases their coefficient with step regression and reducing AIC. Analyzing the normal Q-Q and residual plot, we may conclude that rescaling and normalizing the variables may further enhance the model. Also, text features and sentiment analysis of reviews will add dimensions to data, which may help the model in better prediction of points.

Code repository: https://github.com/kulgaurav/Wine-Review-Analysis