# A Statistical Analysis

OF

# Wine Reviews Dataset


By

Nishant Mohan

# 1. Introduction

The Wine Reviews Dataset provides a wonderful opportunity to explore and analyse the quality of wines from across the world. The dataset has close to 130k reviews of wines. The presence of a textual description provides an opportunity for text analysis as well. Here I study the dataset with a statistical approach.

Specifically, I present my findings while trying to answer following questions:

- Given two classes of wine (Sauvignon Blanc from South Africa and Chardonnay from Chile, prices at $15), which wine is better rated and by how much? I answer this question in section 3.1.2. For these classes, what is the probability that a bottle of Sauvignon Blanc will be better? Section 3.1 analyses this. I answer this question in section 3.1.3.
- For the Italian wines costing less than $20 and regions having at least 4 reviews, which regions produce better than average wines? Section 3.2 studies this, I answer this in section 3.2.2.
- Which factors are most important while trying to predict wine rating for wines from USA, using a linear model? I dedicate section 3.3 to study this wherein 4 different models are proposed. I answer the question in section 3.3.4.

Finally, I present a conclusion in section 4. Appendices 1-4 contain well-documented R code for an exploratory data analysis as well as each of above questions.

# 2. Getting to know the Data: Exploratory Data Analysis

We start with a basic exploration of data. There are several fields of interest, I picked country as the first one. In the following chart (figure 1), I sort countries by the median of their price and plot boxplots of their points.

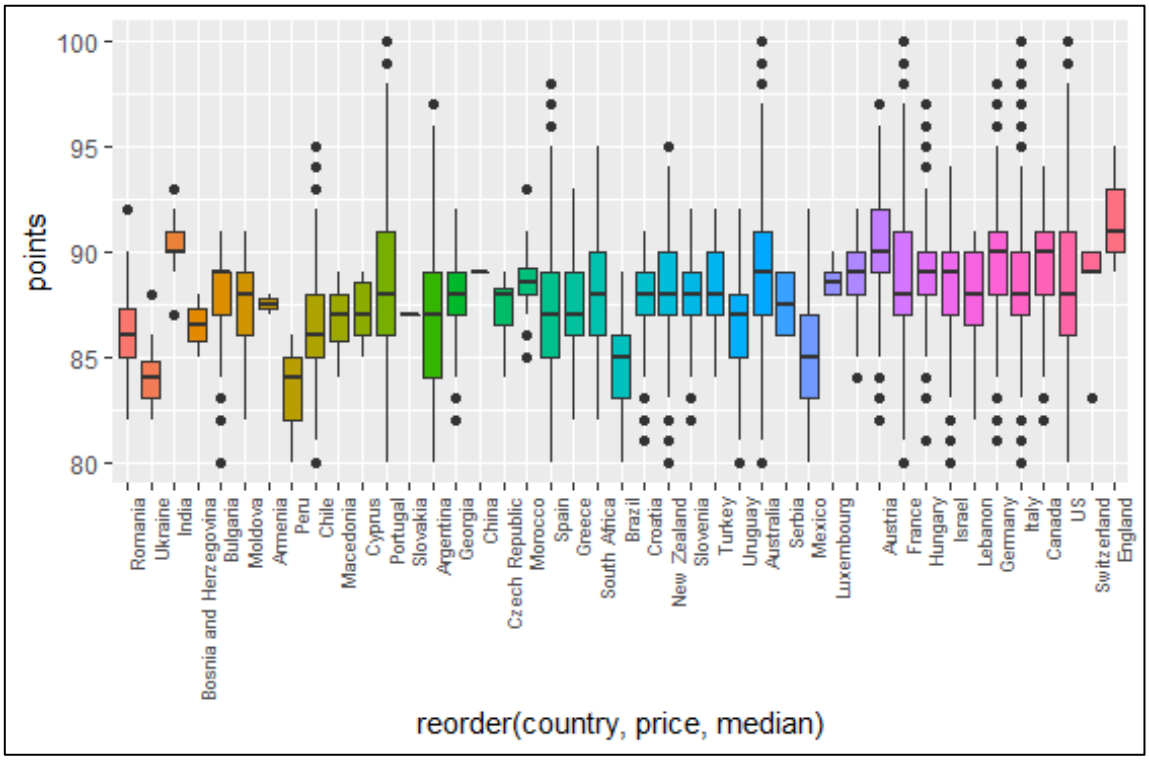


*Figure 1: Comparison of quartiles of points across countries. Countries are sorted by median prices on the horizontal axis.*

Looking at the x-axis in the above chart, a country towards left side of the axis has low median price and towards right has high median price. Therefore, England has some of the most expensive wines, while Romania has cheapest ones. Y-axis shows the boxplot of points for these countries. We see a general upwards trend of median line of each boxplot from left to right, which signifies that higher points mean more expensive wines.

Table 1 shows the top and bottom 10 countries by the count of reviews and the count of reviews. Table 2 gives the top variety of wines by count in the dataset. We can see that Pinot Noir and Chardonnay are the most reviewed varieties of wines.

*Table 1: Top and Bottom Countries by Review Count*

| Top 10 Countries | Count | Bottom 10 Countries | Count |
|---|---|---|---|
| US | 54265 | Macedonia | 12 |
| France | 17776 | Serbia | 12 |
| Italy | 16914 | Cyprus | 11 |
| Spain | 6573 | India | 9 |
| Portugal | 4875 | Switzerland | 7 |
| Chile | 4416 | Luxembourg | 6 |
| Argentina | 3756 | Armenia | 2 |
| Austria | 2799 | Bosnia and Herzegovina | 2 |
| Australia | 2294 | China | 1 |
| Germany | 2120 | Slovakia | 1 |

*Table 2: Top Wine Varieties*

| Variety Name | Count |
|---|---|
| Pinot Noir | 12787 |
| Chardonnay | 11080 |
| Cabernet Sauvignon | 9386 |
| Red Blend | 8476 |
| Bordeaux-style Red Blend | 5340 |
| Riesling | 4972 |
| Sauvignon Blanc | 4783 |
| Syrah | 4086 |
| RosÃ© | 3262 |
| Merlot | 3062 |

Next, we look at the actual ratings or points awarded to the wines. The ratings follow (figure 2) roughly a normal distribution, centred around 88 and spread between 80 and 100. The price of wines is highly skewed as it has some outliers. Figure 3 shows the histogram of price with maximum limit of the plot placed at 97$^{th}$ percentile of prices.

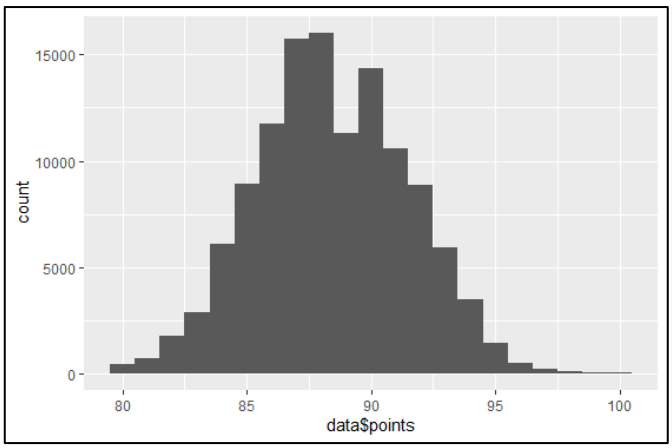


*Figure 2: Histogram of points*

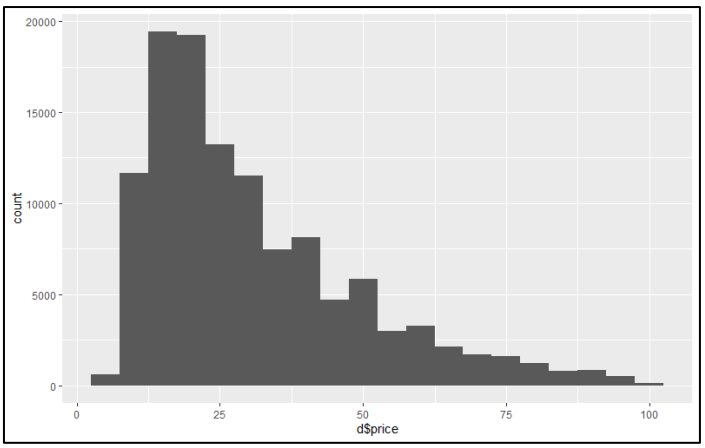


*Figure 3: Histogram of Price (Max at 97$^{th}$ percentile)*

I finally took a closer look at the reviews description. I made a corpus of the text using **tm** library. Then I processed the text as: stripping extra white spaces -> converting to lowercase -> removing numbers -> removing punctuation -> removing stop words -> stemming of words.

| BEFORE CLEANING | AFTER CLEANING |
|---|---|
| *Soft, supple plum envelopes an oaky structure in this Cabernet, supported by 15% Merlot. Coffee and chocolate complete the picture, finishing strong at the end, resulting in a value-priced wine of attractive flavor and immediate accessibility.* | *soft suppl plum envelop oaki structur cabernet support merlot coffe chocol complet pictur finish strong end result valuepr wine attract flavor immedi access* |

*Figure 4: Text Cleaning Results*

After cleaning I make a document-term matrix of the corpus and finally built a word-cloud to visualize the most frequent words in the reviews. Looking at figure 5, we can see that apart from wine and flavor, which are naturally the most frequent words, words such as fruit, aroma, finish, acid cherri are present in high numbers.

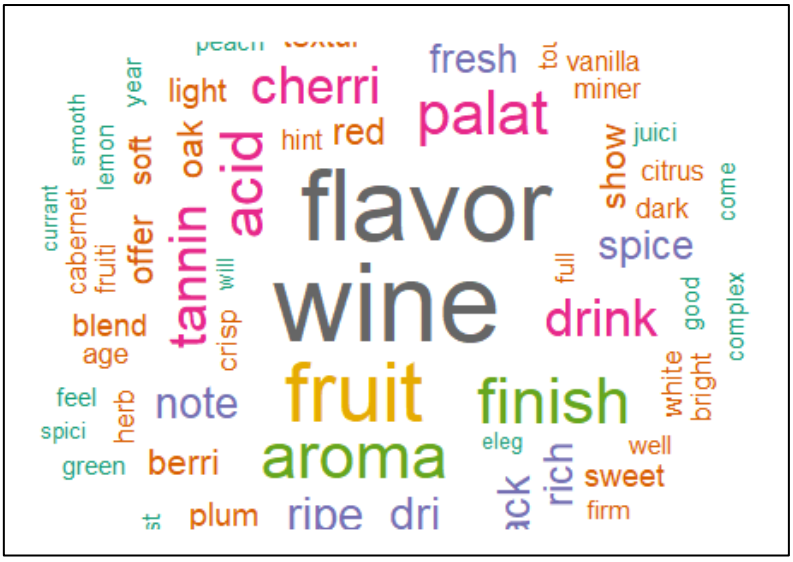


*Figure 5: Most frequent words in Review description*

There could be much more extensive exploratory analysis of the data. But for the purpose of this report, I limit myself here, and continue with the actual analysis in the following sections. The code for this EDA can be found in Appendix 1.

# 3. Analysis

## 3.1. Q1a: Comparison of 2 Means

### 3.1.1. Data Preparation

I filter out the Chilean Chardonnay and South-African Sauvignon Blanc, which are priced at $15.

Some key statistics of the two classes we are interested are provided in table 3. Figure 6 gives a boxplot of the two classes which shows that Sauvignon Blanc are generally placed above the Chardonnay wines. Also, there are more sample observations for Chardonnay than are for Sauvignon Blanc.

*Table 3: Key Statistics of the two Classes*

| Measure | Chardonnay- Chile | Sauvignon Blanc- South Africa |
|---|---|---|
| Count | 37 | 14 |
| Mean | 85.08 | 87.21 |
| Median | 85 | 87 |
| Standard Deviation | 2.2 | 1.7 |

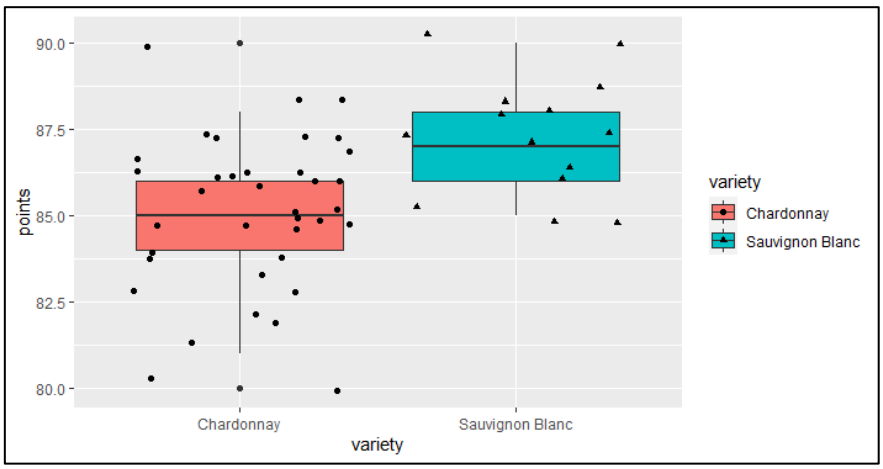


*Figure 6: Distribution of points across varieties of interest*

In next section I try to answer if the difference of these two means is significant enough for a general understanding.

### 3.1.2. Q1a(i)- Two-sample t-test

I performed a t-test to test the hypothesis that the difference in means is significant.

```
        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
```

The t-test confirms that the null hypothesis "the difference in two means is zero" is rejected, the difference in means of Chardonnay and Sauvignon Blanc is indeed not 0.

**This provides us with the answer to the question: "Which type of wine is better rated? How much better?"**

**We see that Sauvignon Blanc of South Africa has mean points score 87.21, while Chardonnay of Chile has 85.08. Therefore, the Sauvignon Blanc of South Africa is better by (87.21-85.08)/85.08=2.5% or 2.13 points.**

### 3.1.3. Q1a(ii)- Comparing Means in Bayesian Model

In order to answer the second question, I need more data samples in order to get the probabilities. I use the comare_2_gibbs() function as provided in the hierarchical case study to model the two classes.

I chose mu0=86, which is close to the overall population mean, precision tau0=1/400, the difference in means del0=5 with gamma0=1/400.

Figure 7 gives the basic properties of the posterior and performance of the sampler. The traces do not follow a specific pattern. We see that the mean is centred around 86, as expected. Mean of del is at 1.06 and that of precision is at 0.16. The code can be seen in Appendix 2.

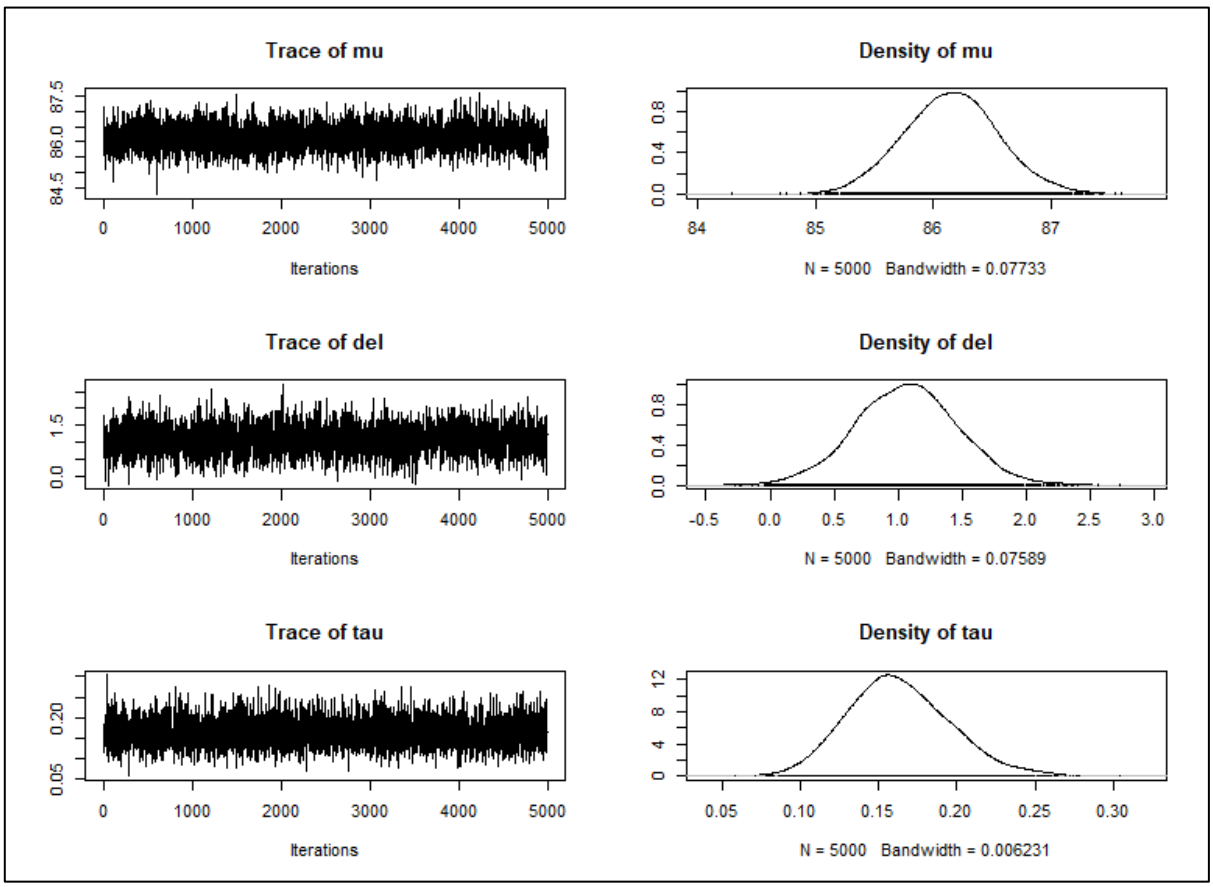


*Figure 7: Performance of Gibbs Sampler*

Using the resulting posterior, I sampled the two distributions. Figure 8 and 9 give an overview of the modeled simulations and difference in points of the two classes respectively. The difference is normally distributed and centred around 1. More points lying below the line than above the line in figure 8 signify that more samples from Suvignon Blanc have higher ratings than Chardonnay.

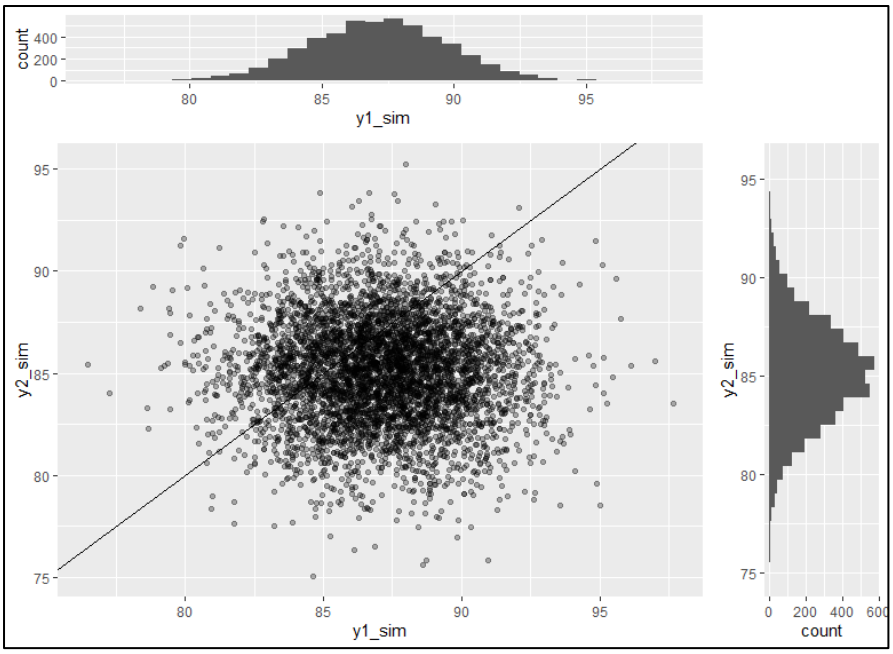


*Figure 8: Simulations of the two classes*

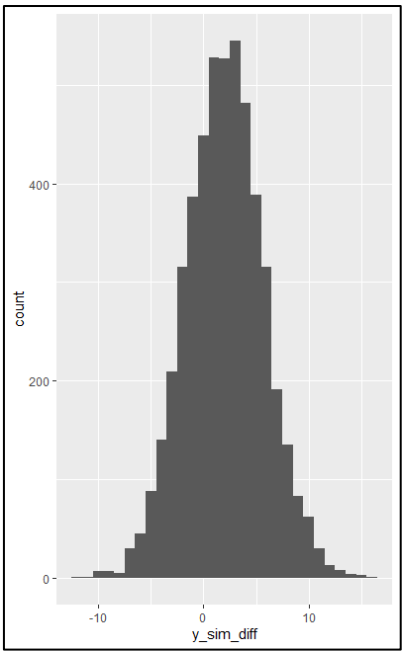


*Figure 9: Difference in the Points*

Finally, in order to answer the question: **"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?"**, we use the simulated data and check mean of the stepwise result of all instances wherein one class was better than the other.

**This gives us a 70.7% probability that Sauvignon Blanc will be better in the given situation.**

## 3.2. Q1b: Comparison of More Than Two Categories

### 3.2.1. Data Preparation

I filter the data to take only wines from Italy with price less than or equal to $20, and having at least 4 reviews. This gives 5647 observations.

```
#preparing the data
data_q1b<-data[data$country=='Italy' & data$price<=20,]
counts<-aggregate(data_q1b$region_1,list(data_q1b$region_1),length)
counts<-droplevels(counts[counts$x>3,'Group.1'])
data_q1b<-droplevels(data_q1b[data_q1b$region_1 %in% counts,c('region_1','points')])
nlevels(data_q1b$region_encoded)
```

Therefore, we have 174 distinct regions. I take a further look at the data using figures 10,11,12 and 13.

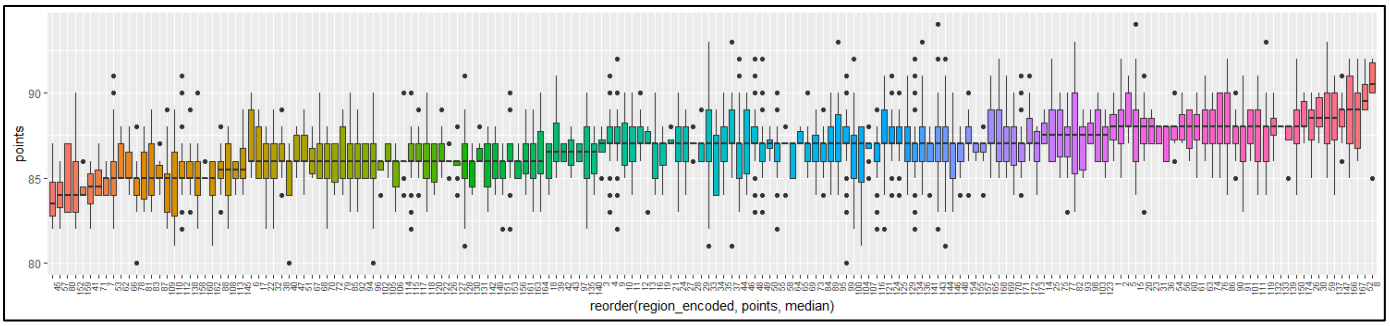


*Figure 10: Boxplots of points of each Region*

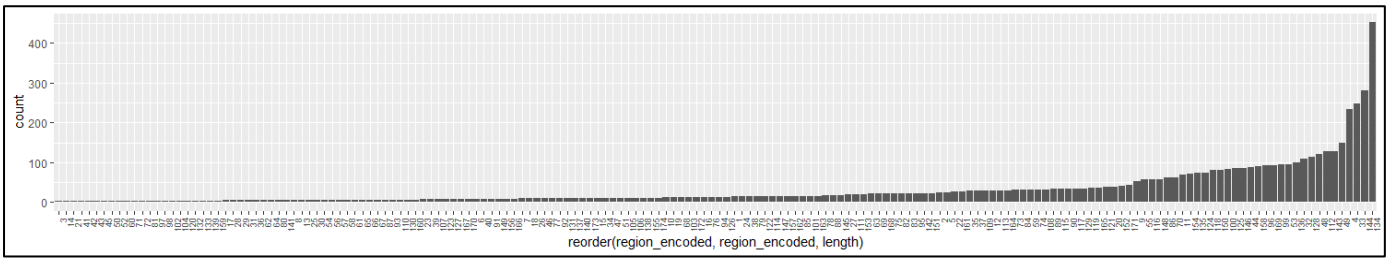


*Figure 11: Count of Observations of each Region*

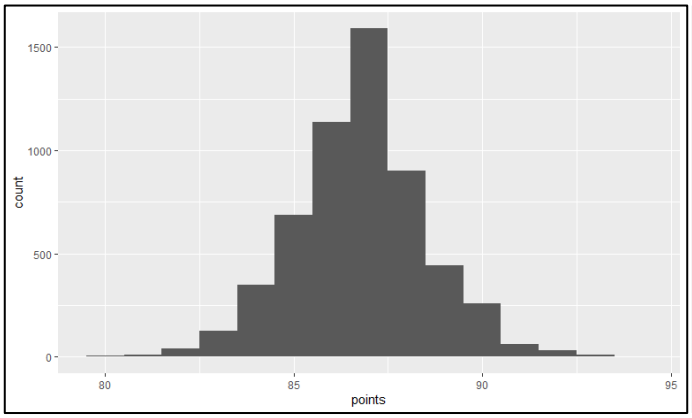


*Figure 12: Histogram of Points*

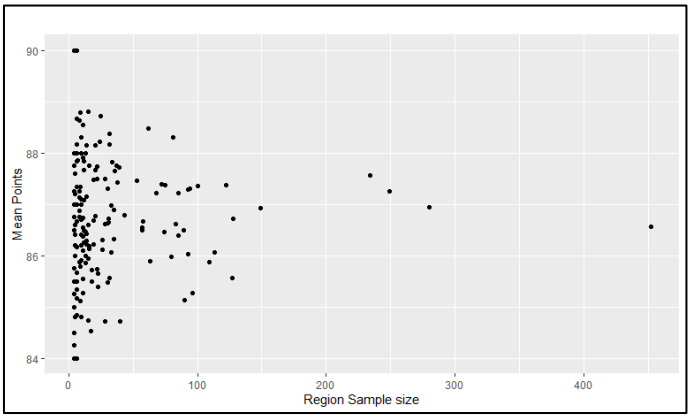


*Figure 13: Mean Points versus Sample Size*

Figure 10 shows boxplots of each region sorted by median of points. Figure 11 shows that there is considerable difference in the sample size of each region, ranging from 1 to close to 400. While points are normally distributed (figure 12), the variance in sample size across regions has an impact on mean of region points (figure 13). It can be clearly seen in figure 13 that higher number of observations in a region suggest that the points mean would be closer to the population mean. This suggests that comparing means across regions is not a good idea since the sample sizes are highly skewed. We should explicitly model the mean scores from different populations and take into account this variation to produce reliable estimates of region-wise points mean.

### 3.2.2. Sampling

I used the compare_m_gibbs() function provided in the hierarchical case study for this task.

I chose mu0=86, which is the population mean and a small value for tau0 to make the prior vague. As a result, I get a mean points value of 86.72 as opposed to 86.76 of the original population.

```
> apply(fit2$params, 2, mean)
       mu     tau_w     tau_b
86.7237224  0.4072722  1.1170166
> apply(fit2$params, 2, sd)
        mu      tau_w      tau_b
0.080418913 0.007767757 0.151270656
> ## within region standard variation
> mean(1/sqrt(fit2$params[, 2]))
[1] 1.567173
> sd(1/sqrt(fit2$params[, 2]))
[1] 0.01494958
> ## between school standard variation
> mean(1/sqrt(fit2$params[, 3]))
[1] 0.9526758
> sd(1/sqrt(fit2$params[, 3]))
[1] 0.06459531
```

Following figures explain the results of sampling in greater detail.

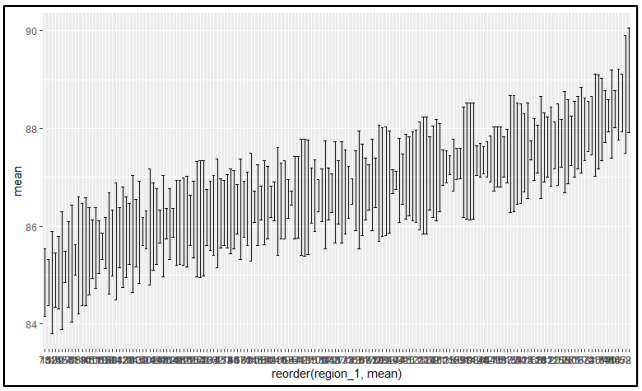


*Figure 14: Error bars for estimates of each region*

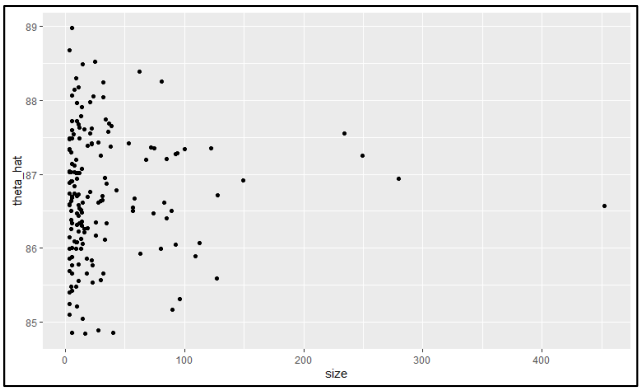


*Figure 15: Mean points versus sample size of estimates*

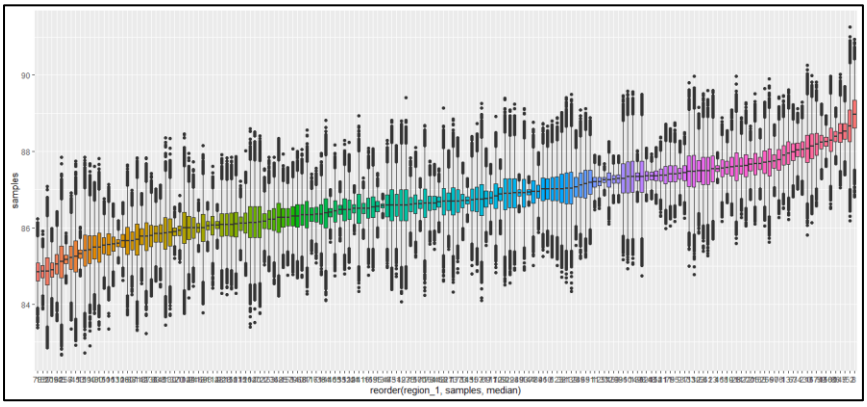


*Figure 16: Boxplots of estimates, ordered by medians for estimates*

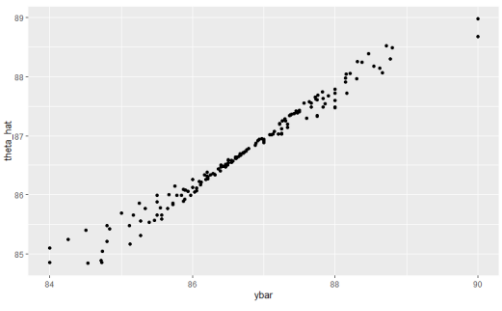


*Figure 17: Comparison of estimates and samples*

Figure 15 shows that sample size of estimates is not as highly related to mean as it was previously. Figure 17 shows that estimates and samples mean now coincide, so convergence is attained.

Now we can answer the question, **"Which regions produce better than average wine?" Table 4 gives list of the 77 regions for which average wine rating is higher than overall average**.

| | | | | |
|---|---|---|---|---|
| 1. Coste della Sesia | 17. Dogliani | 33. Isonzo del Friuli | 49. Orvieto | 65. Sangiovese di Romagna.. |
| 2. Trento | 18. Sardinia | 34. Conegliano … | 50. ForlÃ¬ | 66. Trebbiano d'Abruzzo |
| 3. Verdicchio di Matelica | 19. Montefalco Rosso | 35. Monica di Sardegna | 51. Friuli | 67. Terre di Chieti |
| 4. Vermentino di Gallura | 20. Primitivo di Manduria | 36. Controguerra | 52. Prosecco Treviso | 68. Friuli Isonzo |
| 5. Cerasuolo di Vittoria.. | 21. Romagna | 37. Orvieto Classico.. | 53. Valpolicella Classico .. | 69. Piave |
| 6. Greco di Tufo | 22. Rosso di Montalcino | 38. Alcamo | 54. Chianti Colli Fiorentini | 70. Orvieto Classico |
| 7. Maremma Toscana | 23. Barco Reale di .. | 39. Valpolicella Ripasso | 55. Campania | 71. Valpolicella |
| 8. Verdicchio dei Castelli di.. | 24. Falanghina del .. | 40. Terre Siciliane | 56. Roero Arneis | 72. Lambrusco dell'Emilia |
| 9. Vino Nobile di.. | 25. Morellino di Scansano | 41. Castel del Monte | 57. Rosso Piceno | 73. Pompeiano |
| 10. Nebbiolo d'Alba | 26. Barbera d'Asti Superiore | 42. Emilia-Romagna | 58. Monferrato | 74. Delle Venezie |
| 11. Carignano del Sulcis | 27. Rosso di Montepulciano | 43. Valpolicella Superiore .. | 59. Salento | 75. Colli Bolognesi |
| 12. Roero | 28. Molise | 44. Montepulciano .. | 60. Montepulciano .. | 76. Piedmont |
| 13. Vermentino di Sardegna | 29. Franciacorta | 45. Beneventano | 61. Montecucco | 77. Valpolicella Classico |
| 14. Campi Flegrei | 30. Calabria | 46. Terre del Volturno | 62. Valdadige | |
| 15. Offida Pecorino | 31. Rosso del Veronese | 47. Prosecco di .. | 63. Offida Passerina | |
| 16. Falanghina del Sannio | 32. Barbera d'Alba .. | 48. Sannio | 64. Chianti | |

*Table 4: Regions with higher than average wine ratings*

## 3.3. Q2- Linear Regression

I built four different models for comparison. The first two models use two libraries- **syuzhet** and **SentimentAnalysis** respectively, to extract sentiments from the review description. For the third model, I encode the features- winery, region_1, province, taster_twitter_handle, variety using means of respective category, called as target encoding. In the final model, I incorporate best performing features from each of the previous three models.

### 3.3.1. Data Preparation

I filter the data for the US and keep only relevant columns for the linear regression model. I keep *description,* as I intend to extract sentiment from the text. I keep province and region to extract possibly high rated areas. I keep taster_twitter_handle assuming that one or more reviewer may be more lenient than the others, and therefore give higher ratings. Variety, Winery and Price are natural predictors for the points.

### 3.3.2. Converting text review to emotion (Sentiment Analysis)

In order to extract sentiment from the text review description I use two approaches. First I use **syuzhet** library to get a score of each review across 10 different sentiments. This shows, as in figure 18, that most reviews are positive. This makes sense because we know from our data that the points values range from 80 to 100.

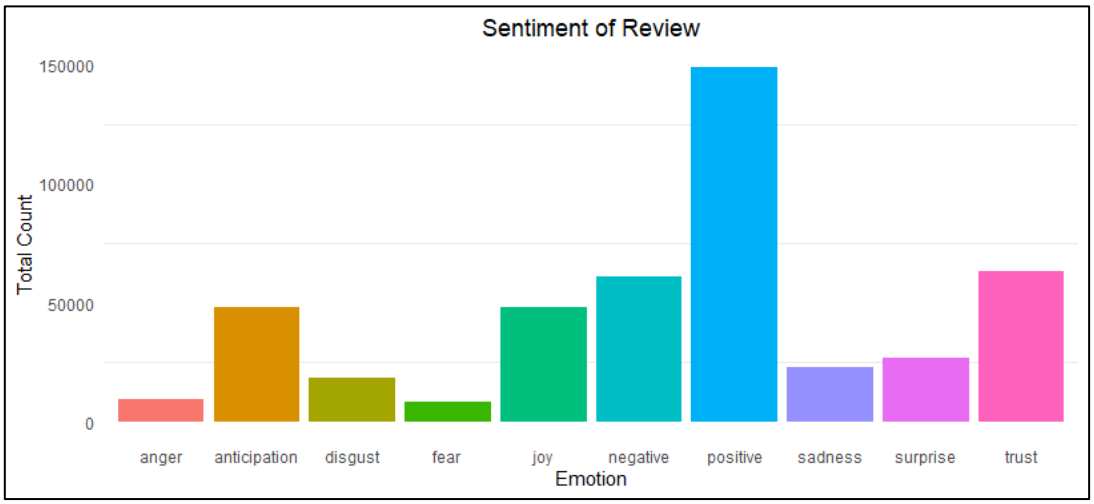


*Figure 18: Sentiment Analysis of Wine Review*

I also use **SentimentAnalysis** library to get a score of emotions as per a variety of dictionaries. The reported features using this package are 14 in total: WordCount, SentimentGI, NegativityGI, PositivityGI, SentimentHE, NegativityHE, PositivityHE, SentimentLM, NegativityLM, PositivityLM, RatioUncertaintyLM, SentimentQDAP, NegativityQDAP, PositivityQDAP. The complete code can be seen in Appendix 4.

### 3.3.3. Encoding Categorical features to Numeric features

Now that the sentiments from the review have been extracted, we are still left with categorical columns: *region_1, province, taster_twitter_handle, winery and variety.* For these, I define a function **target_encode()**, which helps in encoding these features based on the mean of *points* for each category in these. I also use smoothening of the mean of each category, which is based on the fact that we should give more weight to the overall mean if that category has small number of observations. The code can be seen in Appendix 4.

Mathematically,

$$\mu = \frac{n \times \bar{x} + m \times w}{n + m}$$

Where $\mu$ is the mean that replaces our categorical variable, n is the number of observations for this category, $\bar{x}$ is the mean of all observations of this category, m is the weight we assign to the overall mean, and w is the overall mean. Therefore, if the weight m is zero, then $\mu$=$\bar{x}$. When we provide a weight m, more weightage is given to the overall mean in case the category in question has a small number of observations.

### 3.3.4. Analysis of Models

As mentioned earlier, I used four approaches for modelling. For each of these I also applied Backward Elimination to identify best features. While simpler models did help reduce AIC by a small factor, the errors increased by a small factor too. Therefore, I omit those models from the report shown in Table 5. This table gives the important result summaries of each of the four models. I retrieve median of residuals, R-squared and adjusted R-squared and F-statistic p-value from the lm summary function. For errors MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percentage Error), I use a library called **DMwR**. I also record the best predictor from each of the models by comparing the coefficients of all the predictors and choosing ones with the highest coefficients by magnitude, having p-value less than 0.05. Baseline is the model which gives overall mean value of points as the prediction.

| | Models | | | | |
|---|---|---|---|---|---|
| **Metric** | **Baseline** | **Sentiments from *syuzhet*** | **Sentiments using *SentimentAnalysis*** | **Target Encoding** | **Best Features from Other Models** |
| Median of Residuals | - | -0.0298 | -0.0298 | 0.0375 | 0.0051 |
| Multiple R-Squared | - | 0.1937 | 0.4218 | 0.4707 | 0.5346 |
| Adjusted R-Squared | - | 0.1935 | 0.4217 | 0.4706 | 0.5346 |
| AIC | - | 111717.00 | 93674.19 | 88869.19 | 81886.61 |
| MAE | 2.57281 | 2.26756 | 1.90656 | 1.79954 | 1.68594 |
| MSE | 9.71442 | 7.83283 | 5.61718 | 5.14213 | 4.52093 |
| RMSE | 3.11680 | 2.79872 | 2.37006 | 2.26763 | 2.12625 |
| MAPE | 0.02913 | 0.02567 | 0.02155 | 0.02038 | 0.01908 |
| F-statistic p-value | - | 2.20E-16 | 2.20E-16 | 2.20E-16 | 2.20E-16 |
| Best Predictor 1 | - | sadness | RatioUncertaintyLM | winery | RatioUncertaintyLM |
| Best Predictor 2 | - | surprise | NegativityHE | aster_twitter_handle | NegativityHE |
| Best Predictor 3 | - | positive | NegativityQDAP | province | winery |

*Table 5: Comparison of vital statistics from all the Models. Green represents favorable values; Red represents unfavorable values.*

I further explain the main characteristics in observations.

**Residuals**: The residuals are the error between the actual value of points and the predicted value of points. Figure 19 shows that while the residuals normally distributed for all four models, they have least variance in the fourth model. Since the mean and median are close to 0, the model fit seems good. The fourth model gives a Residual standard error

which is the same as RMSE of 2.12. This is average amount (standard deviation) the prediction varies from the true regression line. Considering that the mean value of points is 88.5, this much error is acceptable.

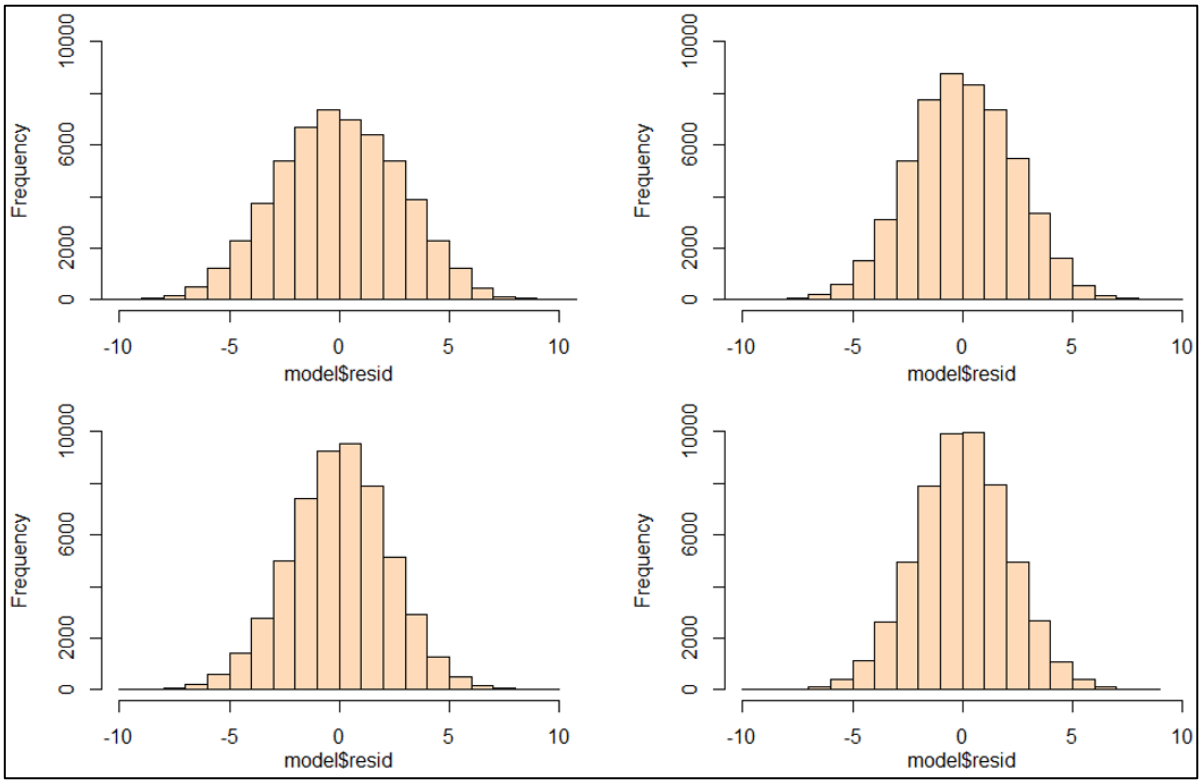


*Figure 19: Histograms of Residuals: Top Left- Model with sentiments from syuzhet, Top Right- Model with sentiments from SentimentAnalysis, Bottom Left- Model with Target Encoding, Bottom Right- Model with Best Features from all other models*

**Important Features**: Looking at the highest coefficient values by magnitude for each of the model (where p-val<0.05), I select best three features (table 5) from each of the first three models for inclusion in the fourth model.

**AIC**: AIC helps to identify the risks of overfitting and underfitting. A lower AIC means the model would be more robust. The fourth model has the lowest AIC.

**Coefficients**: In the models, all the features that are deemed important have standard error comparable to the coefficient values. On the other hand, features which were not important have standard error on order of magnitude less than the coefficient value. This is conformed from the p-values, which are all greater than 0.05 for features related to sentiments, thus accepting the null hypothesis that "beta coefficient associated with the variable is zero".

Looking at the value of coefficients, highest magnitude of coefficients are for winery, taster_twitter_handle, RatioUncertaintyLM, NegativityHE, latter two features being derived from the review description using *SentimentAnalysis* library. For example, the coefficient of winery is 1.23. This means that for one unit increase in winery, the points increase by 1.23. Now, since I encoded the categories using points itself, it is ambiguous understanding what '1 unit increase in winery' means. However, we can say still understand that winery does affect points more than other predictors. We can now answer the question: **"Which factors are the most important in obtaining a good rating?". We see that winery, description and taster_twitter_handle are the most important in obtaining a good review.**

## 4. Conclusions

1a (i). Sauvignon Blanc of South Africa is better than Chilean Chardonnay by 2.13 points or 2.5%.

1a (ii). There is a 70.7% probability that Sauvignon Blanc will be better than Chardonnay.

1b. Out of 174 filtered regions, there are 77 regions for which the average wine rating is higher than population mean. The list is given in table 4.

2. *Winery, description and taster_twitter_handl*e are the most important features in obtaining a good review.

# Appendices: All the R Codes

Appendix 1: Exploratory Data Analysis - github link

Appendix 2: Q1a- Comparison of 2 Means - github link

Appendix 3: Q1b- Comparison of More Than Two Categories - github link

Appendix 4: Q2- Linear Regression Model - github link