Predicting Wine Quality

Richard Collins

***Abstract*—This report details the steps taken in building and testing two binary classifiers that are able to classify wine into high and low quality categories. The first classifier considered numeric and categoric input variables whereas the second implemented some preliminary sentiment analysis before classification was carried out. Both classifiers were built using logistic regression models. Through careful consideration of input variables and adjustments to thresholds, a classifier accuracy of 72.88% was achieved. Auxiliary analysis was also performed on the data set to better understand its structure and suitability as training data for the classifiers.**

I. Introduction

A. Background and Motivation

THE global wine market is estimated to be worth $304 billion and is expected to rise to $ 380 billion by 2020 [1]. Wine has enjoyed millennia as a cornerstone of European society and culture, and is experiencing new found popularity in Asian markets such as China [2]. With the market saturated, it can be difficult to determine which wine is worth paying luxury prices for, and which wines should be avoided. Entire professions are now dedicated to help solve this problem; courses in wine tasting can be taken, lead by expert wine tasters who, with their superior knowledge and expertise, offer advice on wine selection. Often, this advice is accompanied with vivid descriptions of the wines' colour, aroma, and taste.

The question we seek to answer in this report is whether it is possible to determine the quality of a wine given other attributes of the wine. Special attention will be given to the case of classification through sentiment analysis.This paper aims to build a binary classifier that is able to distinguish between wine of high quality, and wine of low quality. It will be able to do this by considering various attributes of the wine such as country of origin and price. As a further area of consideration, the descriptions of the wines will be analysed for sentiment polarity and word count to investigate whether these features are able to be leveraged in the classifier.

B. Overview of Report

This report is divided into three main parts: *Initial exploration of the data*, *building a general binary classifier*, and *building a text-based binary classifier*. Results and interpretations will be presented throughout and summarised in the *Conclusion* section. Unless otherwise stated, the programming language `R` was used throughout.

II. Exploring the Data

This section describes the data that was investigated. From this initial exploration, decisions on how to proceed were made. This takes into account the size of the data set, its sparsity, and suitability for classification, among other factors.

A. The Data Set

The data was scraped from WineEnthusiast [3], a monthly lifestyle magazine, during the week of June 15th, 2017. The data set was then updated on November 24th, 2017 to eliminate duplicate reviews and include taster names. The data was uploaded to Kaggle [4] where it could be freely downloaded. A summary of the attributes can be found in Fig. 1. Attributes of particular interest at this stage are points, country and price.

Attribute	Description of attribute
country	Origin of the wine
description	A short review of the wine
designation	Vineyard within the winery
points	Review score (80 - 100)
price	Price of one bottle in US$
province	Province/state that the wine is from
region_1	Growing region in the state/province
region_2	Further sub-region
taster_name	Name of taster/reviewer
title	Title of the wine review
variety	Type of grapes used
winery	Winery that made the wine

Fig. 1. Table of attributes

The data set is large; there are $\sim 130,000$ rows where each row represents a single wine. This amounts to 50.48 MB of data. Upon initial inspection, there appears to be many missing values, represented as `NA` in the rendering in RStudio (the interface with which analysis of the data was possible).

B. Summary Statistics

The first attribute to be explored was *country*. Fig. 2 shows a pie chart with every country of origin with over 1000 wines in the data set. There are over 40 countries represented, however, many only have a few dozen wines. Therefore, in order to make the piechart easier to interpret, any country with fewer than 1000 wines were included in the "OTHER" category. US wines are heavily represented in the data set. This is unsurprising considering that WineEnthusiast is an American based magazine. Famous wine-producing countries (such as France, Italy, Spain etc.) also feature heavily.

According to WineEnthusiast, a wine review is only published if it receives 80 points or more from one of its expert tasters. The range of points in this data set is $80 - 100$, discretised into whole points (fractions of points are not given). The distribution of points awarded is shown in Fig. 3.

The points appear to be a near perfect normal distribution with $\mu = 88.4$ and $\text{median} = 88$. Wine lovers from large

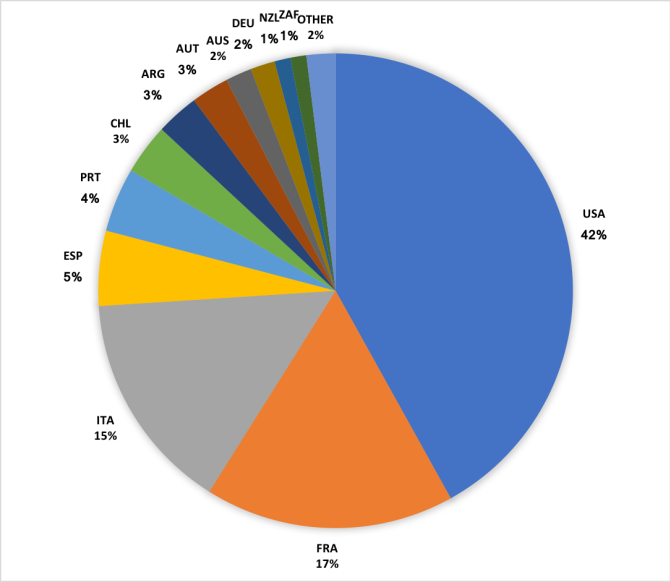

Fig. 2. Proportion of wines reviewed by country

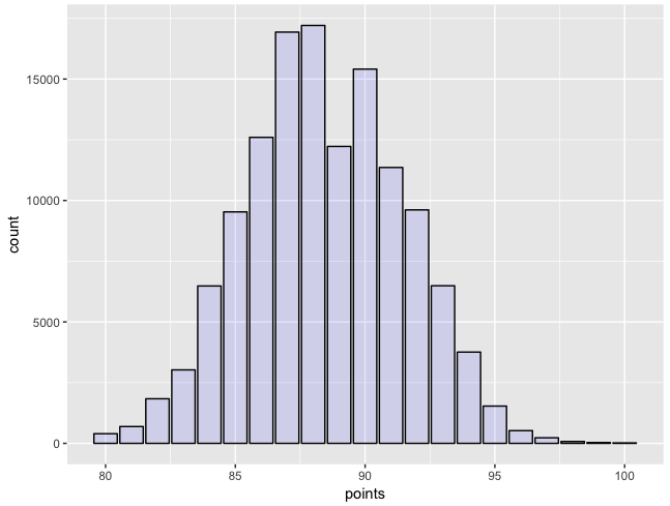

Fig. 3. Distribution of points awarded. $\mu = 88.4$, median $= 88$

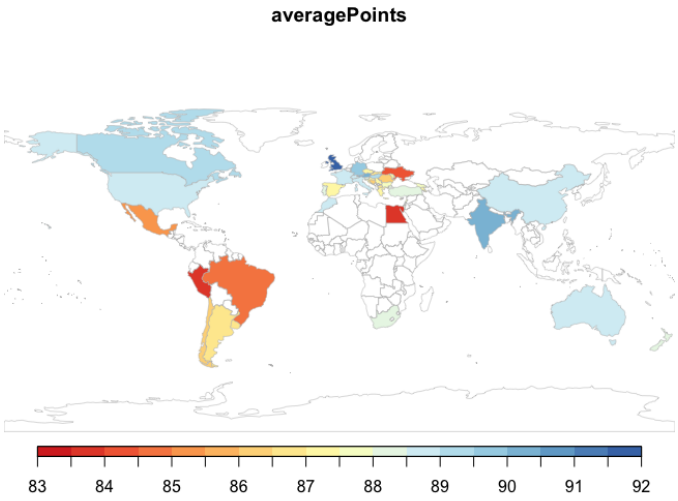

Fig. 4. The average points awarded to a bottle of wine from each country

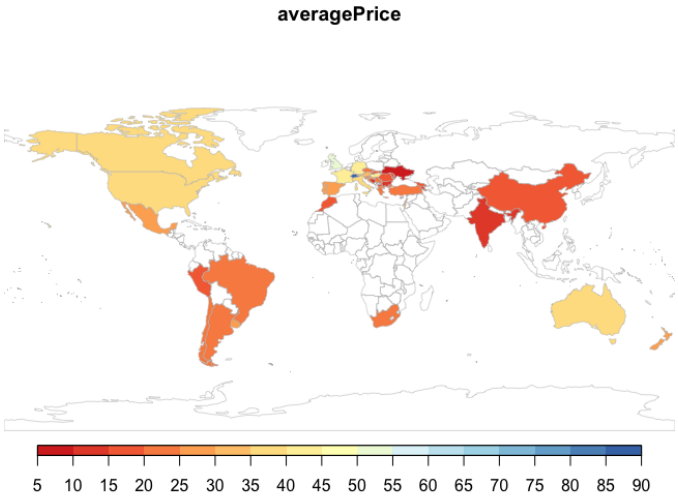

Fig. 5. The average price of a bottle of wine in each country ($)

wine-producing countries will often advocate wine from their own country, claiming it to be superior to others. To explore this further we can plot the average points awarded to a bottle of wine on a world map, as shown in Fig. 4. For this particular plot, the `R` package `rworldmap` was used [5].

The more *red* the country is, the lower its wine's average score. It may be quite surprising that such well-know wine producing countries such as Argentina and Chile score relatively poorly compared to more obscure countries like the U.K. and India. This can be explained by referring again to Fig. 2 and realising that Argentina and Chile contribute a total of 6% of the wines in the data set, where as India and the U.K. contribute a total of 83 bottles combined. There need only be a handful of exceptional wines to increase their average points, whereas the effect of exceptional wines in larger wine-producing countries are diluted by many unexceptional ones.

Using the same `R` package [5] we can breakdown other attributes by country. For example, the average price of a bottle of wine in each country is shown in Fig. 5.

Now, the more red the country is, the lower its wine's average price. We can see that price is more heavily skewed than points, with almost all countries in the red or yellow part of the scale ($\sim \$5 - \$40$). In fact, 80% of the wines in the data set are below $46. The country with the highest average wine price is Switzerland at $85.29 with the U.K. coming second at $51.68. Again, owing to these countries' low representation in the data set, average prices can be skewed easily by extremes.

A natural line of enquiry is the relationship between price and quality. We as consumers often believe that paying a higher price for a product or service begets a higher quality. This belief is especially strong in the wine market were people will spend a great deal more money if they believe they will get a superior product. To investigate this argument a scatter plot[1] of price and points was plotted. This is shown in Fig. 6.

[1] A small amount of jitter was added to the plot to blend out any discreteness

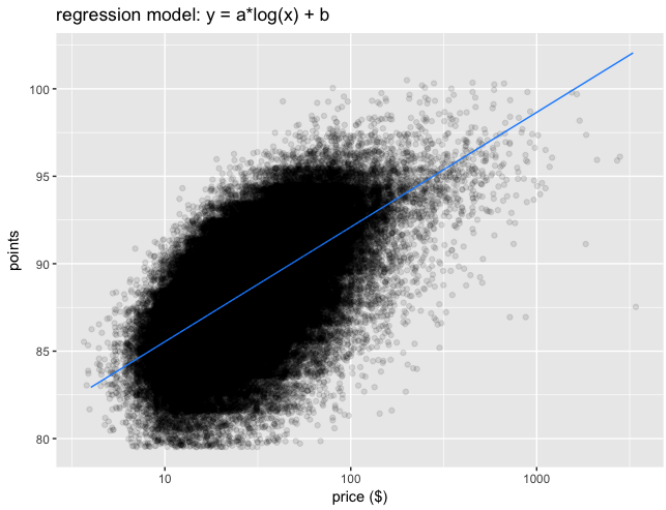

Fig. 6. Scatter plot of the price of a bottle of wine against the points it was awarded

	Estimate	Std. Error
Intercept	78.981	0.036
log(price)	2.848	0.011
R	0.612	

Fig. 7. Regression model summary statistics

In addition to the scatter plot, a regression model was fitted to the data. Because of the distribution of prices in the data set, a linear model was not deemed appropriate for this relationship. Indeed, when the x-axis is converted to a $\log_{10}$ scale the data displays logarithmic behaviour. There is a strong correlation between these two attributes as demonstrated by the high correlation coefficient shown in Fig. 7.

A logarithmic relationship between price and quality is fairly typical for luxury items such as wine. The shape can be interpreted in the following way: At low prices, every incremental increase in price is met with a fairly substantial increase in quality (a $100 bottle of wine tastes substantially better than a $10 bottle of wine). However, at high prices, an equal increase in price does not equate to the same change in quality (A $1000 bottle and a $1090 bottle taste similar).

III. Building the Classifier

This section describes the processes and decisions that were made while building and refining the wine quality classifier.

A. Problem Statement

Is it possible to classify a wine's quality based on other attributes of the wine?

B. Logistic Regression with Price

Logistic regression is used to predict a binary dependent variable. Therefore, the dependent variable in this case, *points*, was converted to a new variable, *highQuality*, that takes the value 0 or 1:

$$\text{highQuality} = \begin{cases} 1 & \text{points} \geq 88 \\ 0 & \text{points} < 88 \end{cases} \tag{1}$$

The threshold 88 was chosen because it is the median of the points. This effectively separates the wines into two equally sized categories: high quality and low quality.

The attribute *price* was used as the sole initial independent variable. In later sections, more attributes will be considered. But in the interest of simplicity, the first regression model only considers price. Although the points column (and by extension the highQuality column) has a full compliment of entries, the price column is missing values. Therefore, before proceeding, all rows where there exists an `NA` value were dropped. This reduced the number of rows to 120,975 (a reduction of $\sim 7\%$), which still gave ample data to work with.

The remaining data were then split into training and testing sets. The training:testing ration split was 67:33, which is typical for machine learning problems such at this and is done primarily to avoid overfitting. The split was performed via random sampling so that any underlying pattern in the original data was not inherited.

The logistic regression was carried out using the `glm()` function in `R` [6]. This function takes in the independent variable(s) and calculates the probability that the dependent variable is 1, given by the following equation:

$$P(y = 1) = F(x) = \frac{1}{1 + e^{(-\beta_0 + \sum_{i=1}^{m} \beta_i x_i)}} \tag{2}$$

where x_i it the ith independent variable and β_i is the ith coefficient as calculated by the `glm()` function. In a very general sense, the higher the coefficient, the greater the effect that that attribute has on the probability of $y = 1$. A plot[2] of the wine data space with the logistic regression curve (equation 2) is shown in Fig. 8.

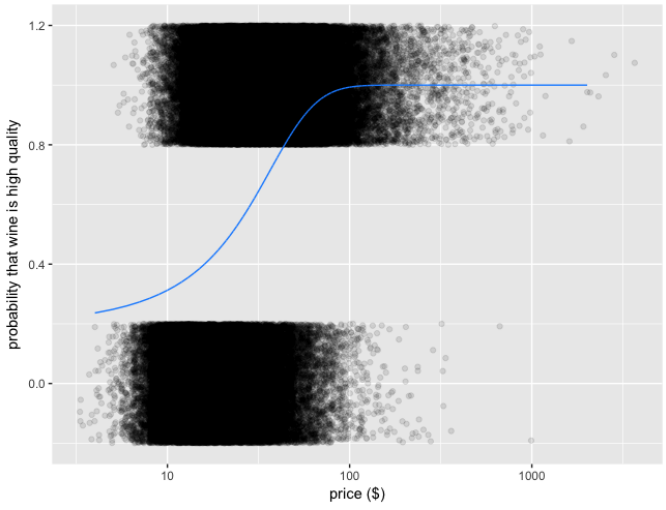

Fig. 8. Data space with logistic regression curve showing the probability that a wine is high quality

From the regression curve generated from the training set, wine quality classification was then predicted using the testing

[2]Again, jitter was added to help show the distribution

set. If $P(y = 1) > 0.5$ for a given wine in the test set, then that wine's $\mathrm{highQuality} \rightarrow 1$. Any probability lower than 0.5 then $\mathrm{highQuality} \rightarrow 0$. Once the testing set was classified, they were then compared to their *true* highQuality classification, generating a confusion matrix shown in Fig. 9.

		Predicted class	
		1	**0**
Actual class	**1**	17,870	6,049
	0	5,307	10,695

Fig. 9. Confusion matrix for regression model in Fig. 8

The confusion matrix can be summarised in the following way: The accuracy of this logistic regression model is given by the percentage of correctly classified instances, which in this case is 71.55%. The false-positive (FP) rate is the percentage of wines that were predicted to be highQuality = 1, but were in fact highQuality = 0. The false-negative (FN) rate is the opposite; it's the percentage of wines that were predicted to be highQuality = 0, but were in fact highQuality = 1. Therefore, we can see that a FN is actually a desirable outcome, despite not correctly classifying the wine, whereas a FP is undesirable.

In a real-world scenario, we would like our classifier to not only be able to correctly classify highQuality (a high true-positive TP and true-negative TN rate), but also to be extra safe when it comes to ambiguous cases and be more likely to classify highQuality = 0. This behaviour may be desirable for someone who wants to avoid lower quality wine, even if it means some high quality wine is overlooked. Although there is no formal definition for this type of *"safe accuracy"*, in this report it is given by the following formula:

$$a_\gamma = \frac{\mathrm{TP} + \mathrm{FN} + \mathrm{TN}}{n} \tag{3}$$

where n is the total number of observations and in the testing set, and γ is the threshold probability at which the binary dependent variable is divided (0.5 in the example above). Of course, γ could be set arbitrarily high to ensure the highest possible value of a_γ. For the above confusion matrix $a_{0.5} = 84.85\%$.

C. Logistic Regression with Multiple Attributes

So far, only price has been considered as the independent variable in the logistic regression. We can improve the accuracy of classification if more attributes are included in the model (please refer to Fig. 1 for a full list of attributes). There are a few key points to keep in mind when choosing appropriate input variables:

- Sparsity of variable; logistic regressions cannot be performed if there exists `NA` values. Rows containing `NA` values would need to be dropped, so in order to maintain as much of the original data as possible, input variables should contain as few `NA`s as possible.
- Balanced variables; variables with many values that appear very few times should be avoided. Take for example the attribute taster_name, there are names that appear very few times compared to others; it seems that they are in fact international correspondents or guest tasters. It might be that they taste one or two very good wines, to which they award a high number of points. If this taster_name appears again in the testing set, recognising that this is a very high information feature, it will be highly weighted by the model.
- Uncorrelated variables; it is difficult to asses the importance of variables if they are correlated with others. The attributes region_2 is a subregion of region_1, so knowing the value for region_2 would automatically give us the value for region_1.

Inspecting the data set, there are no *"perfect"* attributes that avoid the problems listed above. The most suitable is the attribute *country* as this has relatively few `NA`s and is the most balanced in terms of value frequency. Therefore, country and price were used to classify the dependant variable *highQuality*. After running the logistic regression with $\gamma = 0.5$ the accuracy increased to 72.30% and $a_{0.5} = 86.30\%$. This is an appreciable improvement on using only price for the input variable.

D. Refinement of Classifier - Adjusting γ

It is not necessarily the case that a threshold value of $\gamma = 0.5$ gives the best accuracy. The value of γ can be adjusted to maximise accuracy. However, this could be at the expense of the newly proposed value of a_γ. If the optimised threshold is found to be *lower* than 0.5, it would be more likely to classify a high quality wine as low quality, an undesirable effect. It would be at the individual's discretion whether to accept a higher FN rate for an overall better accuracy. A script was written to identify the threshold γ that maximises the accuracy of the classifier. For completeness, the code snippet is included in appendix A.

For the logistic regression model presented so far, the adjusted threshold was found to be $\gamma = 0.45$. The accuracy with this adjusted threshold was 72.88% and $a_\gamma = 82.92\%$. As expected, the accuracy increased but it was at the expense of a higher FN rate, and therefore a higher value for a_γ.

IV. Text-Based Classifier

So far, the descriptions of the wine have been largely ignored. This is principally due to maintaining clarity by separating attributes of very different classes. However, this text is feature rich, and can be used to classify the wine into different quality bands. Analysis of text such as this is a very complex undertaking, and to be treated fully would require a separate study of its own. However, this section will describe the initial steps needed to build a text based classifier.

A. Problem Statement

Is it possible to classify a wine's quality based solely on a description of that wine from an expert wine taster?

B. Exploring the Text

Each wine has associated with it a brief description written by an expert wine taster. Samples of these descriptions have been included in appendix B. Before any analysis of the the text could be carried out, the descriptions had to be parsed and cleaned into an appropriate format. A custom function was implemented to perform such cleaning. The function first lowers every letter to lower case, removes punctuation, and removes *stop words* (these are function words such as *it* or *the*). What is left after this process are content words ready for analysis.

It is important to note here that the process of parsing and cleaning the text is very computationally intensive. When the function was initially applied to every description in the original data set the runtime was over 20 minutes, making it impractical for further analysis. Therefore, the decision to reduce the data set was made. This involved taking a random sample of 13,000 wines ($\sim 10\%$ of the original number of wines). Although this is a large reduction in the number of descriptions, 13,000 samples still provides a large enough data set with which to carry out analysis. After the reduction, the parsing function still took an appreciable length of time, but did not significantly disrupt workflow.

In order to gain an insight into the nature of the descriptions, a word cloud was generated using the `tm` and `wordcloud` libraries in `R` [7][8]. A word cloud displays the most common words closer to the centre and in a larger font, which allows for an effective method for quickly assessing the character of the descriptions. The word cloud is shown in Fig. 10 and a table of the most frequent words shown in Fig. 11.

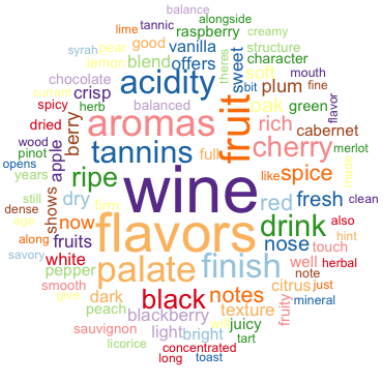

Fig. 10. Word cloud of the most frequent words in the wine descriptions

Unsurprisingly, the most common word used in the wine descriptions is *wine* itself. Other common words are related to the flavour and smell of the wine, which of course is not untypical of descriptions of this nature.

Word	Frequency
wine	7733
flavors	6200
fruit	4498
aromas	4046
palate	3803
finish	3478
acidity	3415
tannins	3113
drink	2995
cherry	2770

Fig. 11. Top ten most frequent words used in wine descriptions

The frequency distribution of words in a corpus follow a law known as *Zipf's Law* [9][10]. The law states that the rth most frequent word is inversely proportional to its frequency rank, r:

$$f(r) \propto \frac{1}{r^{\alpha}} \tag{4}$$

where $\alpha \approx 1$ is a scale factor. To determine whether the descriptions follow Zipf's Law a $\log - \log$ plot was constructed with word frequency on the y-axis and frequency rank on the x-axis. This is shown in Fig. 12.

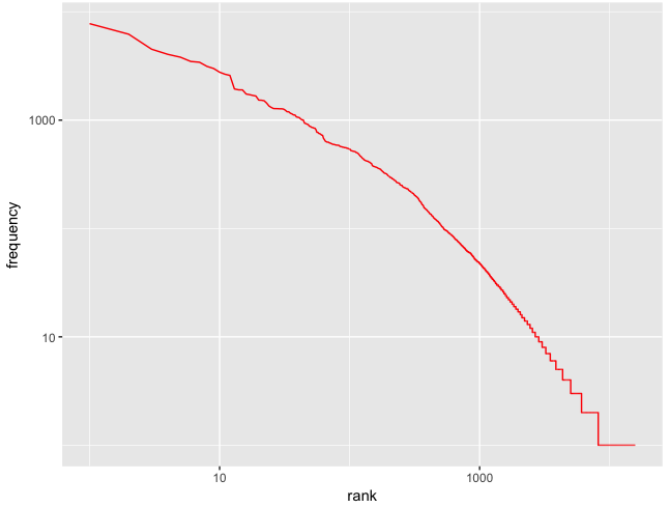

Fig. 12. $\log_{10} - \log_{10}$ plot of word frequency against it's overall rank in the descriptions

The shape of the line graph is approximately linear, meaning a power law is present. We can interpret this observation by saying that the descriptions are composed of a few high-frequency words (shown in Fig. 10 and Fig. 11), and many low-frequency words (such as *"frangelico"* and *"buoying"*).

C. Length and Sentiment of Descriptions

Returning to the problem statement, we want to classify the wine quality by considering only these descriptions. Two lines of enquiry were explored to this end: the length of each description, and the sentiment polarity and magnitude. The word count was determined by applying an adapted version of

the text parser and cleaner used to create Fig. 10 and Fig. 11, however, the stop words were not removed before counting. The word counts were then grouped by description.

In order to measure the sentiment of each description, the `sentimentr` package [11] was used. This package provides a function that calculates sentiment at the sentence level. A sentiment score can take any value, but in practice it is confined to the range of $-10 \leq s \leq 10$. A positive sentiment score indicates positive sentiment (e.g. *this wine is good*), and a negative score a negative sentiment (e.g. *this wine is bad*). The magnitude of the score indicates the degree to which the sentence is positive or negative. Each description is comprised of between 1 - 5 sentences, so the mean average sentiment score was taken as the overall sentiment score for a given description.

Before building the classifier, the word count and sentiment score were plotted against each other to understand their relationship. This is shown in Fig. 13.

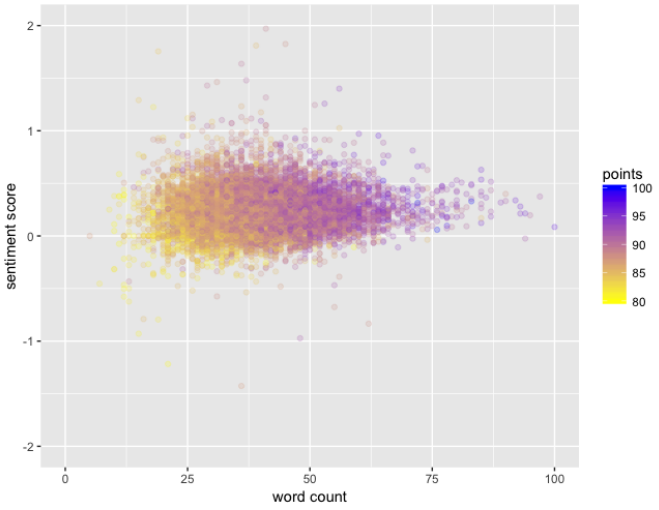

Fig. 13. Plot of description word count and sentiment polarity where colour represents the points awarded to each wine

Firstly, we note that the central density is positioned in the *positive sentiment* region of the y-axis. The sentiment score does not appear to change significantly with word count, remaining approximately constant at slightly above 0. There is however a tendency for shorter descriptions to receive more varying sentiment scores than longer ones. Assessing the correlation coefficient of 0.032 between these two variables confirms that any correlation is negligibly small (see Fig. 14).

Incorporated into this plot are the points awarded to each wine, represented as yellow for low points and blue for high points. One might expect that wines where the description has a high sentiment score would also receive higher points. Examining Fig. 13 reveals that this is not the case; the colour does not vary significantly over the y-axis. This is again confirmed by the low correlation coefficient in Fig. 14.

The final variable relationship is between word count and points. Here we can see a noticeable correlation between the two, indicated by the gradual change in colour over the x-axis from low word count to high word count. A correlation

	sentiment score	word count	points
sentiment score	1	0.032	0.150
word count	0.032	1	0.524
points	0.150	0.524	1

Fig. 14. Correlation coefficients between three attributes

coefficient of 0.524 confirms the initial observation that points are indeed correlated to the length of a description.

D. Building the Text-Based Classifier

The classifier was built in a similar manner to that in the previous section. First, training and testing data were made via a 67:33 split of the original data, ensuring the split was carried out from a random sample. The input variables in this case were *sentiment score* and *word length*. Neither of these attributes contain any `NA` values because they were calculated from the reduced data set of 13,000 rows (as described above), and the correlation between the two was shown to be negligibly small. Furthermore, since both variables are treated as continuous numeric values, unique instances are not considered a problem as they were with categoric data in the previous section. The threshold γ was also optimised, as described previously.

The accuracy from this regression model was 70.29% and $a_{0.49} = 81.18\%$. This is noticeably lower than the accuracy achieved using the attributes *price* and *country*, but still adequate given the preliminary nature of the sentiment analysis involved.

V. Conclusion

This section briefly describes the merits and demerits of the analysis in this report, as well as a review of the main results.

A. The Data Set

The data set had some interesting features that were explored in the report. However, two main issues were noticed:

- The abundant presence of `NA` values meant that much of the data become inadequate for analysis. Once all rows with `NA`s were removed, the data set was reduced from $\sim 130,000 \rightarrow \sim 22,000$ rows. The revealed that only wines with their country value as *"USA"* had complete entries, rendering the country attribute inadequate as an input variable for a logistic regression model (since every wine would come from the USA).
- The size of the data set become a problem during the sentiment analysis section when a computationally intesive function was applied over entire columns. Optimising the function would have gone some way to elevating this issue.

Ultimately, a smaller, more complete data set would have been more desirable.

B. Initial Exploration of the Data

The country that contributed the highest number of wines was the USA with 42%, with other famous wine-producing countries also contributing significantly. It was found that the points were approximately a normal distribution with $\mu = 88.4$ and $\text{median} = 88$. Points and price were broken down at the country level, where it was noticed that price was not balanced. Plotting price and points revealed a logarithmic relationship.

C. First Binary Classifier

A table of the performance is summarised in in Fig. 15

Input Variable(s)	γ	Accuracy	a_γ
price	0.50	71.55%	84.85%
price, country	0.50	72.30%	86.30%
price, country	0.45	72.88%	82.92%

Fig. 15. Performance summary of the first binary classifier

The performance of the classifier improved by carefully selecting appropriate input variables and optimising the threshold γ.

D. Analysis of Descriptions

The frequency of words approximately followed Zipf's law, where the frequency of a word and that word's frequency rank are inversely proportional.

E. Text-Based Classifier

A table of the performance is summarised in in Fig. 16

Input Variable(s)	γ	Accuracy	a_γ
word count, sentiment score	0.49	70.29%	81.18%

Fig. 16. Performance summary of the second binary classifier

The accuracy of this classifier, while lower than the previous classifier, is still admirable given the difficultly of this type of analysis.

VI. Possible Extensions

- A natural extension to a binary classifier is a multinomial classifier, that is, classifying more that two different categories. In this report, points were converted to highQuality which took a value of either 1 or 0. However, this attribute would also be suited to multinomial classification as the points are discretised to whole point values. A method such as a random decision tree would be able to determine a predicted score for the wine.
- Although this report focused on the points as the dependant variable, other attributes would work also. For example, a classifier could be built that predicts the *variety* of a wine based on the other attributes in the data set.
- It would be interesting to further develop the sentiment analysis introduced in this report. In practice, a robust sentiment analysis would take into account the domain from which a target document comes. For example, the word *"formidable"* may have a completely different sentiment value depending on whether it is found in a wine review, or a crime report. The sentiment analyser function described in this report didn't take the domain into account, and therefore sentiment values didn't always align with what was expected.

Appendix A
Code Snippet

```
#Find best threshold
thresholds <- seq(0.01, 1, 0.01)
accuracy <- NULL
for (i in seq(along = thresholds)){
  prediction <- ifelse(
    Test$model_prob > thresholds[i],
    1, 0
    ) #Predicting for threshold

  accuracy <- c(
    accuracy, length(
      which(Test$highQuality == prediction)
      )/length(prediction)
    )
}

best_threshold <- thresholds[
  which.max(accuracy)
]
```

Appendix B
Sample Wine Descriptions

"This is dominated by oak and oak-driven aromas that include roasted coffee bean, espresso, coconut and vanilla that carry over to the palate, together with plum and chocolate. Astringent, drying tannins give it a rather abrupt finish."

"With attractive melon and other tropical aromas, this is a Torrontés that rises above the masses. It smells great and tastes like a pure blend of lychee fruit, tangerine and honeydew melon. The palate feel is smooth and round, and the finish is dry, clean and healthy. It's everything that Torrontés should be. Drink now."

"Disagreeable for its harsh acidity and vegetal streak that mars the cherries and suggests a bitter green herb. Disappointing."

References

[1] Wine Market - Growth, Trends and Forecasts (2017 - 2022) - https://www.mordorintelligence.com/industry-reports/wine-market
[2] Angela Mariani, Eugenio Pomarici, Vasco Boatto, The international wine trade: Recent trends and critical issues, In Wine Economics and Policy, Volume 1, Issue 1, 2012, Pages 24-40, ISSN 2212-9774, https://doi.org/10.1016/j.wep.2012.10.001.
[3] http://www.winemag.com
[4] https://www.kaggle.com
[5] https://cran.r-project.org/web/packages/rworldmap/rworldmap.pdf

[6] https://www.rdocumentation.org/packages/stats/versions/3.4.3/topics/glm
[7] https://cran.r-project.org/web/packages/tm/tm.pdf
[8] https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf
[9] Zipf G. The Psychobiology of Language. London: Routledge; 1936.
[10] Zipf G. Human Behavior and the Principle of Least Effort. New York: Addison-Wesley; 1949.
[11] https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf