How Good Is Your Wine?

Pinmanee Eowpittayakul, James Gilson, Frederic Li, Leah Reinhard, Xinyuan Zhu

4/24/2019

Contents

Have you ever thought about the quality of the wine you just bought? Is it really that good? Is it really worth the price? Or, to be honest, what is a good wine? These are the motivations behind our projects to predict wine quality. The report is organized as follows. Section 1 introduces the project summary and problem statment. Section 2 talks about some early explorations of the dataset we chose. Section 3 details our data cleaning process. Section 4 demonstrates our text mining, sentiment analysis and clustering results. Section 5 highlights a special note we made during the feature selection process. The final section concludes our project with some recommendations and caveats.

1. Introduction

Our group chose a dataset originating from Wine Enthusiast - a wine focused publication, featuring reviews for 129,972 wines with a total of 13 variables such as variety, location, winery, price and description. The variables *country*, *province*, *designation*, *region 1*, *region 2*, and *winery* all relate to wine location. There are 20 tasters, 708 wine varieties, and wine points that rank wine rating on a scale of 80 - 100 (ratings below 80 are possible, but not included in this dataset). We originally planned to determine which variables in the wine ratings dataset most accurately predict wine **price**, however, after a group discussion, we ultimately decided to build a model that predicts **rating** instead of price. Our reasoning was largely based on utility - a purchaser will always know the price of the wine they are buying, but quality generally cannot be determined until the bottle has been opened.

Our main question is determining what variables are the most useful in predicting wine ratings for this dataset. Our follow-up supplement questions are: What descriptive words from the reviews help determine wine quality? Which words tend to come up most frequently? Do specific wine tasters have a tendency to rate wine points a certain way? Which countries receive the highest points ratings for wine?

2. Early Exploration

We began our exploratory analysis looking at the relationships between price and other wine variables to see if they were related. An analysis of the mean price for each country revealed that Switzerland had the highest average price at $72.83 (with mean wine points of 88.5) while Ukraine had the lowest average price at $9.21 (with mean wine points of 83.8). The analysis of mean wine points for each country revealed that

England had the highest average wine points at 91.76 (with mean wine price of $53.83) and Peru had the lowest average wine points at 83.56 (with mean wine price of $18.05). Our goal was to determine if countries that had the highest average price also had the highest level of wine points. The analysis revealed that the countries ranking for highest average wine points was different than the ranking for highest average wine price. However, it was noticeable from the data that the wines with the highest wine prices and highest wine points were all from west Europe.

We further explored the relationship between wine tasters and the wine points they assigned in their reviews. The data revealed that the average wine points were 88, only 8 points above the lowest possible score in our dataset. Alexander Peartree consistently gave the lowest wine point scores with an average of 85.83934. Anne Krebiehla gave the highest wine point scores with an average of 90.75175. This exploratory analysis answered our question concerning wine taster's tendencies to rank wine points a certain way. The highest average wine point score was more than 9 points below the highest possible score for wine points, indicating that, in this dataset, it is uncommon for wines to be rated above 90 points. An analysis of wine price versus the variable *has_twitter* revealed that the existence of a twitter handle had no effect on wine pricing.

3. Cleaning the Data

Price and origin are arguably the most important factors to a customer when considering purchasing a wine. Therefore, we cleaned our wine quality dataset by removing rows with duplicate values and missing values for price and location variables. Intuitively, price is a significant predictor for wine quality since vinters will charge more for better quality wine. Also, geography is important for wine quality since soil fertility, moisture levels, and weather conditions aka terroir vary with location. These factors affect the grapes' taste which in turn affect the wine quality and price. There were a significant amount of blank entries for taster names, so we renamed them as *other* to retain the important information in their rows without accidentally attributing their info to our known tasters.

Certain wine tasters also had listed twitter handles while others had no social media listed. Previously, we believed that tasters who had twitter handles might rate wine points differently than those who were not active on social media. A quick analysis between our created *has_twitter* column and wine points revealed no correlation between the two factors. Therefore, we removed *has_twitter* and *taster_twitter_handle* from our wine quality analysis.

In our first round of data cleaning, we extracted the wine year from the *title* column and created a new variable since the time a wine is bottled affects its taste and price. We further edited this variable to subtract the wine year from our current date (2019) to give the wines' current age and relabeled the column *age*. Since a wine's taste changes with age, we reasoned it would be a useful variable in predicting wine quality. We also removed a small number of records with no wine age/or and country listed, as we reasoned these would be important in our modeling later on. We added a unique identifier column *id* to the data set to help with data wrangling later on. To prevent our data from being skewed, we eliminated outlier data points by only including wine prices below $200 (which is 97% of the dataset). Our final data cleaning steps were to change the variables *description* and *designation* data types to character values.

4. Fundamental Analysis

4.1 Description Sentiment Analysis

Once we felt our dataset was sufficiently clean for analysis, we began text mining both the description and title variables. We created variables analyzing the character count, word count, and sentence count of wine reviews (named *description*). The correlation between character count and wine points was relatively high at 0.579, as was the correlation between word count and wine points (0.535). This revealed that a longer review typically indicated a higher wine points scoring. Our analysis of the correlation between wine price and character count, word count, and sentence count revealed that there was a low correlation between these variables. These low correlations reinforced our decision to predict wine points instead of wine price, since

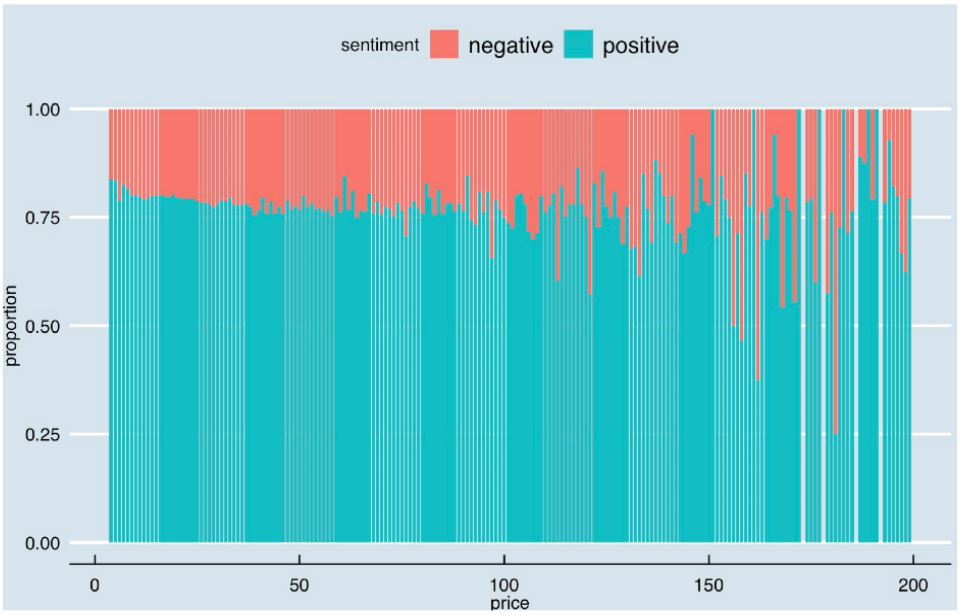
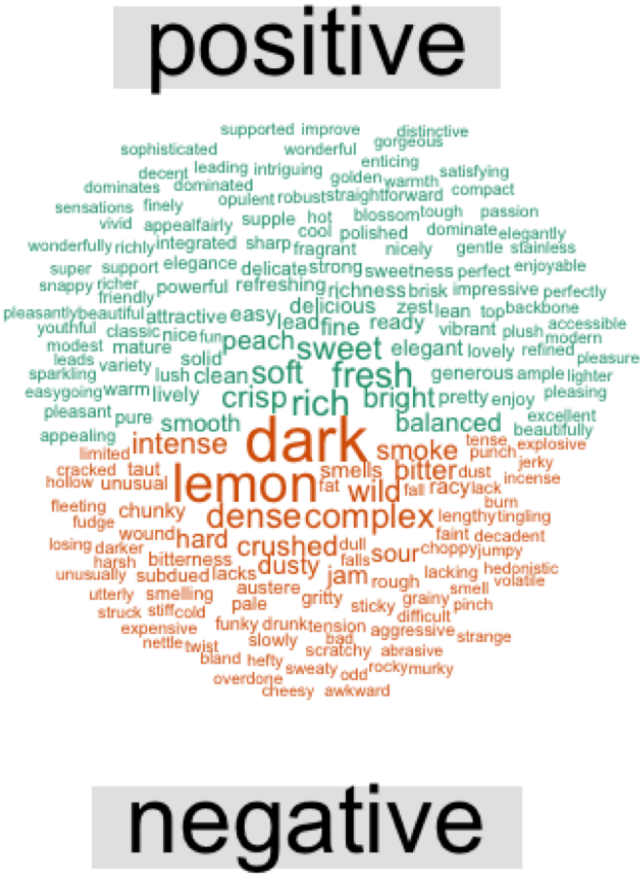


Figure 1: Eh or wow?

the relationship between variables was less obvious with price. We then measured how much of the review variable (*description*) contained positive sentiment using the **bing** lexicon, and created a new variable using the positive sentiment proportion (*positivity*). The majority of reviews were largely positive according to our graph which plotted the positive and negative sentiments for reviews of all prices.

When we ran a correlation between positivity and price, the result was a small negative correlation between the two variables. Our hypothesis was that the wine tasters had higher expectations due to the higher wine price, and therefore they were more harsh in their critiques leading to greater negative sentiment levels in the wine descriptions. Next, we analyzed the descriptions to determine what terms were used most frequently to describe the wines. The top ten descriptive words (excluding **wine** and **it's** from the summaries) were flavors, fruit, aromas, palate, finish, acidity, tannins, drink, cherry, and black. We also further used the **bing** lexicon to determine that 21.8 percent of the wine description words were negative while 78.2 were positive. While interesting, this analysis revealed the need for a more relevant lexicon. The wordcloud below nicely illustrates the issues with the lexicon applied to this data set – words like **tough** and **dominates** are flagged as positive, while words like **complex**, **dark** and **intense** are flagged negatively. In the context of wine descriptions we know this is an incorrect assignment since we conducted outside research on wine terminology. **Dark** refers to the wine coloring which is a neutral descriptor, while **complex** is actually a positive descriptor that means the wine flavor changes from the time you taste it to the time you swallow it. (Wine Folly) We also used the **afinn** lexicon to score the sentiment level of words for wine description. The analysis revealed that wine descriptions contained mostly positive sentiments, but we determined it has the same issue as the bing lexicon regarding properly identifying wine terminology. Sentiment analysis techniques applied to the title variable did not yield any noteworthy results.



4.2 Predictive Text Analysis

After the description sentiment analysis, we continued our text analysis with more predictive text mining techniques. We created a corpus from wine description and cleaned up the inputs by transforming the text to lowercase and removing punctuation, stopwords, and white space. We changed the sparsity threshold to .985 resulting in an overall matrix sparsity of 95% with a total of 322 terms. Given the large size of our dataset, this relatively high threshold helped ensure the retention of descriptive words that occur most frequently and ultimately improved model performance. Using document term metrics (xdtm), we created a corpus to identify the most frequent terms. This analysis was made to answer our question concerning what descriptive words from the reviews help determine wine quality. After attempting various different models (linear regression, decision tree, random forest), we determined that linear regression had the best rmse at 2.11. Words like **rich**, **black**, and **fruit** were highly predictive for wine points according to this model.

We also used term frequency - inverse document frequency weighting (tfidf) to determine if model performance was improved by using a frequency evaluator that ranked importance as well as a frequency count. This

model returned different terms as the most important predictor for wine points, however, test rmse was comparable to the document term metrics model. There is little difference between the two models with both decision tree and linear regression models returning the same rmse. The tfidf model weighted a few descriptive words differently in its output. It inverted the order of **ripe** and **tannin**, and also substituted **dribble** with **cherried**. These small differences in descriptive word frequency/importance changed the random forest rmse by a slight degree, yet identified the xdtm model as the better predictive model for wine points.

```
corpus <- VCorpus(VectorSource(wine$description))
corpus <- tm_map(corpus, FUN = content_transformer(tolower))
corpus <- tm_map(corpus, FUN = removePunctuation)
corpus <- tm_map(corpus, FUN = removeWords, c(stopwords('english'),'wine'))
corpus <- tm_map(corpus, FUN = stripWhitespace)
dict <- findFreqTerms(DocumentTermMatrix(Corpus(VectorSource(wine$description))),
                      lowfreq = 0)
dict_corpus <- VCorpus(VectorSource(dict))
corpus <- tm_map(corpus, FUN = stemDocument)
dtm = DocumentTermMatrix(corpus); dtm
xdtm = removeSparseTerms(dtm,sparse = 0.985)

xdtm = as.data.frame(as.matrix(xdtm))
colnames(xdtm) = stemCompletion(x = colnames(xdtm),
                                dictionary = dict_corpus,type='prevalent')
colnames(xdtm) = make.names(colnames(xdtm))
sort(colSums(xdtm),decreasing = T)

dtm_tfidf = DocumentTermMatrix(x=corpus,control = list(weighting=function(x)
  weightTfIdf(x,normalize=F)))
xdtm_tfidf = removeSparseTerms(dtm_tfidf,sparse = 0.985)
xdtm_tfidf = as.data.frame(as.matrix(xdtm_tfidf))
colnames(xdtm_tfidf) = stemCompletion(x = colnames(xdtm_tfidf),
                                      dictionary = dict_corpus,type='prevalent')
colnames(xdtm_tfidf) = make.names(colnames(xdtm_tfidf))
sort(colSums(xdtm_tfidf),decreasing = T)

wine_corp <- cbind(points = wine$points, xdtm)
wine_corp_tfidf <- cbind(points = wine$points, xdtm_tfidf)
```

We attempted to run a recommender system for the data set, transforming taster name, winery, and points to matrix format. We explored the data and attempted User-based collaborative filtering and Non-Personalized Recommendation. After reviewing different models, we concluded that recommender systems do not produce any insights within our data set. For example, according to the User-based collaborative filtering and the Non-Personalized Recommendation, Williams Selyem was the most popular winery among tasters. However, when we looked at the wine points, the winery's average was only slight higher than wine points average overall. We scrapped our findings but gained a better understanding of our data and the utility of recommender systems through this attempted analysis.

4.3 Clustering

Our next step was to use K means clustering to cluster the data by their nearest means. We subset the data to numeric variables: price, character count, wine age, and review sentiment. These variables were then scaled separately for train and test, and evaluated for potential clustering solutions. We tried both a 2 cluster solution and a 3 cluster solution since there were some discrepancies between individual group members' results on the ratio plot. Based on our final combined ratio plot, the k means clustering was 3 because the elbow in the graph indicated a clustering with the lowest sse. We wanted the point with minimum errors

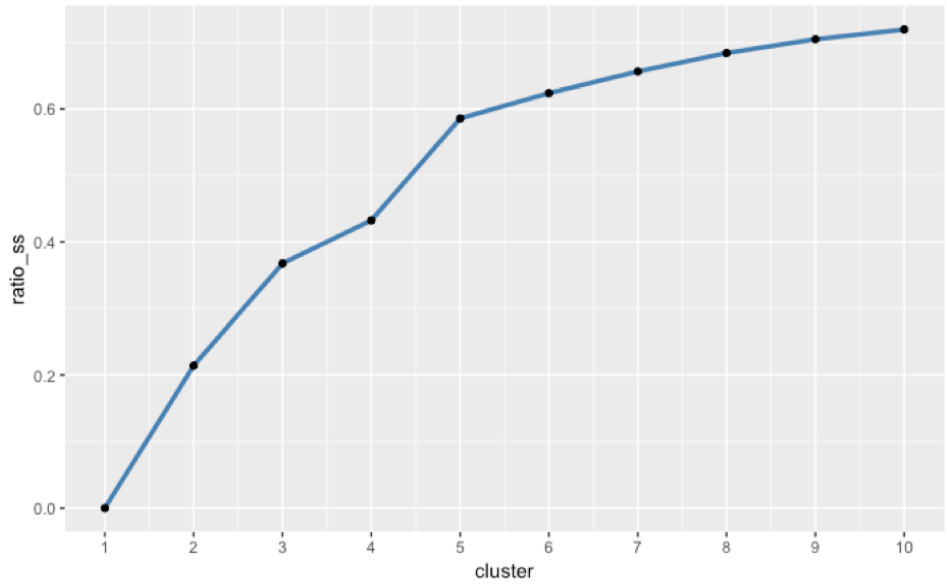


Figure 2: Where's the elbow?

because the larger ks gave diminishing returns and more errors.

We applied the cluster solution to the test set, added the clusters back to the original data set, and then evaluated the train and test set clusters individually. Cluster evaluation revealed interesting segments in the data. Cluster 1 corresponds with "Expensive but good wines"' – slightly older, higher priced wines, with high sentiment scores and the highest ratings. Cluster 2 corresponds with "Overpriced wines"' – moderately priced wines with negative review sentiment and average ratings. Cluster 3 corresponds with "Value wines" – lowest price, high review sentiment with average ratings. While we eventually determined that creating separate models for each cluster did not improve overall predictive accuracy, using the clusters as a feature in our final model did provide some benefit.

We then continued our exploration on transformation by creating a feature to score wineries based on the average rating of their wines (named *wineryScore*). The large number of levels associated with this factor made it infeasible to include in a model as is, making this transformation necessary. We also created a new feature capturing the number of wines rated per winery (named *numWinesByWinery*) with the intention of compensating for overfitting introduced by wineries with a smaller number of ratings (this ultimately proved to be outside the scope of our analysis). These variables were then incorporated back into both the train and test set. Given that some wineries in the test set (with only a small number of rated wines) did not initially receive ratings, we imputed missing values from the mean of wineries in the set that did receive ratings. There was a strong correlation between average winery score and points on the train set (0.733), though this correlation was inflated by wineries that had a small number of wines. However, the correlation remained strong on the test set (0.58459) despite no longer being inflated by one off wineries.

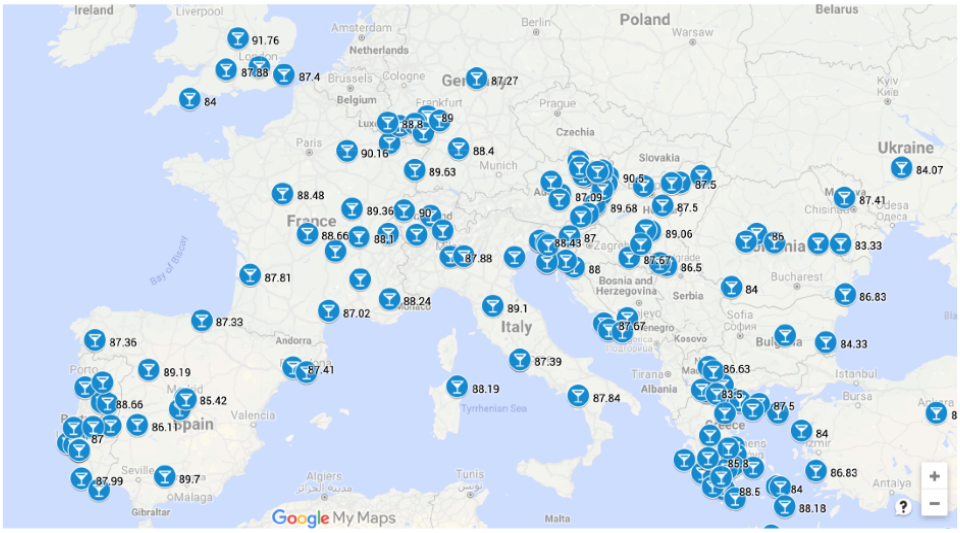


Figure 3: Come what may, my wine

5. Feature Selection

We imported a data set containing a large number of coordinates associated with countries, and then performed some minor cleaning to make sure this new information could be joined properly with our existing data. We attempted to retrieve more granular geographic information based on winery location, but since we were unable to use the API provided by Google without payment, we were only able to conduct a quasi-spatial analysis to visualize the wine quality of provinces identified as locations that had the best quality wine i.e. west Europe.

That said, we decide to remove the *province* factor from our final model because most of the provinces are not statistically significant in our 10-fold cross-validated Lasso model, and it is unreasonable to substitute one province for another as each province has a unique territory.

6. Conclusion

Based on all our findings and feature selection techniques, we decided to use linear regression for our final model because it was able to provide us with highest accuracy and relatively good understanding of the model with a RMSE of 1.738 on our test set.

```
final.model <- lm(points~reviewSentiment+positivity+price+taster_name+age
                  +word_count+longitude+latitude+pred_lm+avgScore+numWines+cluster,train)
```

See Figure 4 for the model's summary below.

Looking at the statistical analysis, all variables except *latitude* are statistically significant. Variables in the final model include review sentiment, positivity, price, tasters, age, length of tasters' reviews, longitude, average score, and how many wines the tasters have reviewed per winery.

```
##
## Call:
## lm(formula = points ~ reviewSentiment + positivity + price +
##     taster_name + age + word_count + longitude + latitude + pred_lm +
##     avgScore + numWines + cluster, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3107 -0.9986  0.0221  1.0179  9.5403
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -6.0567166  0.3764357 -16.090  < 2e-16 ***
## reviewSentiment               0.1456693  0.0065024  22.402  < 2e-16 ***
## positivity                    0.5123225  0.0240444  21.307  < 2e-16 ***
## price                         0.0187321  0.0003221  58.151  < 2e-16 ***
## taster_nameAnna Lee C. Iijima  0.3210271  0.1013718   3.167 0.001542 **
## taster_nameAnne KrebiehlÂ MW   0.7475500  0.1045914   7.147 8.93e-13 ***
## taster_nameCarrie Dykes       -0.1773739  0.1992104  -0.890 0.373262
## taster_nameChristina Pickard  -1.5524217  1.1149683  -1.392 0.163822
## taster_nameFiona Adams        -1.0427575  0.4832357  -2.158 0.030941 *
## taster_nameJeff Jenssen        1.0748762  0.1362673   7.888 3.11e-15 ***
## taster_nameJim Gordon          1.0114579  0.1012900   9.986  < 2e-16 ***
## taster_nameJoe Czerwinski      0.1478337  0.1067621   1.385 0.166148
## taster_nameKerin  231Keefe     0.0629149  0.1002360   0.628 0.530224
## taster_nameLauren Buzzeo      -0.3157476  0.1098112  -2.875 0.004037 **
## taster_nameMatt Kettmann       0.9763198  0.0997583   9.787  < 2e-16 ***
## taster_nameMichael Schachner   0.5914692  0.0990705   5.970 2.38e-09 ***
## taster_nameMike DeSimone       0.4879131  0.1333704   3.658 0.000254 ***
## taster_nameOther               0.2675131  0.0978461   2.734 0.006258 **
## taster_namePaul Gregutt        0.8774667  0.0987431   8.886  < 2e-16 ***
## taster_nameRoger Voss          0.2489397  0.0989171   2.517 0.011850 *
## taster_nameSean P. Sullivan    0.7866310  0.1006308   7.817 5.48e-15 ***
## taster_nameSusan Kostrzewa    -0.5084756  0.1151354  -4.416 1.01e-05 ***
## taster_nameVirginie Boone      0.5749624  0.0985984   5.831 5.52e-09 ***
## age                           -0.0510302  0.0019392 -26.315  < 2e-16 ***
## word_count                     0.0279471  0.0007571  36.914  < 2e-16 ***
## longitude                      0.0032576  0.0001729  18.843  < 2e-16 ***
## latitude                       0.0002217  0.0003370   0.658 0.510710
## pred_lm                        0.5190604  0.0036699 141.437  < 2e-16 ***
## avgScore                       0.5219281  0.0034803 149.966  < 2e-16 ***
## numWines                       0.0011667  0.0002982   3.912 9.16e-05 ***
## cluster                        0.0903949  0.0098337   9.192  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.571 on 74841 degrees of freedom
## Multiple R-squared:  0.7372, Adjusted R-squared:  0.7371
## F-statistic:  6997 on 30 and 74841 DF,  p-value: < 2.2e-16
```

Figure 4: Model summary

6.1 Recommendation

Given the implications of our final model, we recommend wine enthusiasts seriously consider the age and origin of the wine when looking for great wines, as well as whether any oenophilia has ever recommended the wine. Experts' opinions do matter.

Additionally, if you want to sound like an expert, do not say the wine tastes good. Say the aroma is fruity and sweet, the palate thick and chewy with light tannins, the finish nutty and ripe and in sum, a very idiosyncratic and balanced variant with a distinctive cellar character. Time for a wine dictionary!

6.2 Limitation

The final model does not have a statistically significant latitude, contradicting the conventional idea that many of the best vineyards under the sun are located on specific degrees of latitude (e.g. 45th parallel north). We believe this finding is largely a result of the limitations of our location data. Our country level coordinate data does not consider the large variety of potential geographic variations within the area of a specific country. For example, even though the U.S. is a large country and has vineyards on both coasts, we only have one pair of coordinates for the U.S. More granular location data would allow for a more thorough exploration of this conventional wisdom in the context of this data set.

Another limitation of the model is not having an appropriate lexicon for wine. As aforementioned, there are many words that are miscategorized in our sentiment analysis. A more appropriate lexicon would go a long way towards extracting more useful and persuasive information from this sentiment analysis.

7. Reference

40 Wine Descriptions and What They Really Mean. Wine Folly, 13 Mar. 2019, https://winefolly.com/tutorial/40-wine-descriptions/.