## Final Project Report

Aitong Su, Emily Ziyi Xiao, George Li, Han Wang

## Research problem

## The wine market has continued to slump in recent years. In 2018, the global wine market was oversupplied due to unusually high harvests and trade conflicts between the U.S. and Europe (Loose & Nelgen, 2021). Until 2020, there are still stocks left over from 2018 in some European countries. France lost exports to the United States due to a retaliatory 25 percent tariff on wine. COVID-19 pandemic has also had a significant impact on global wine consumption. Demand for wine has declined due to the intermittent closure and removal of dine-in regulations in restaurants and other event venues, coupled with changes in consumer spending habits (Ducman et al., 2021). Falling demand and extensive inventories pose threats and challenges to the wine industry.

According to the previous literature, there are generational differences in consumer behavior in wine consumption. Millennials value social occasions, packaging design, and brand labels to gauge wine needs. However, the older consumers emphasize wine variety, origin, and knowledge (Martinho, 2021). The older generation has a better understanding and engagement with wine than millennials. Conversely, the younger generation has a clear tendency towards sustainable wine consumption (Gazzola et al., 2022).

Other precedent studies have proven that understanding wine quality before buying increases the likelihood of wine purchases. In Tennessee, the Quality Assurance Program (QAP) reduces risk when purchasing products by informing consumers of product quality (Rihn et al., 2022). QAP presents wine information to consumers through intrinsic and extrinsic attributes, such as origin, expert rating, aroma, taste, etc. QAP is not only a key to consumers' wine buying decisions but also a means of increasing consumer interest in Tennessee wines.

## Questions and hypotheses

From the market perspective, the US wine market has been shrinking as Gen-Z are not as engaged with wine consumption as the previous generations. One of the possible reasons is that their first-time wine selection experiences are not personal, interesting and customized enough to introduce them to the world of wine. We would like to propose a recommendation system for consumers so that they only need to input their favored smell, flavor, taste into the questionnaire and we can automatically recommend the wine.

To create this recommendation system, we will need to make accurate predictions on consumers’ preferences based on their descriptions. In order to do that we would like to create a predictive model to predict the wine varieties based on the attributes from wine review notes. We are planning on training the model with the review notes from the Wine Enthusiast Magazine and then test the results.

Our research questions are:

1. How does the sentiments of consumers affect their likelihood of the wine?

* correlation between positive reviews and wine scores

1. What are the overall performance of the classification models（regression model, decision tree and recommenderlab) in predicting wine variety ?

* Accuracy, Precision and Sensitivity
* Confusion Matrix(Type I and Type II error)

Our hypothesis is that the sentiment analysis will show that there is a positive correlation between the amount of positive words in one’s wine review and the score rated by the same person for that wine. For the prediction, it is hypothesized that the recommendation lab will perform best among all the classification models, as it can generate results in a user-based model.

## Data Description

In order to predict wines by inputs such as flavor, smell, and taste, we would like to use a wine review dataset to train our model. This dataset is collected from Wine Enthusiast, which is a magazine and website that specialize in wines, spirits, food, and travel. In total, the dataset has 130 thousands of unique wine samples that are either produced in the US or imported from other counties to the US wine market.

Our dataset includes 10 variables that could be sorted to 2 types. The first type records the basic information and geographic information of each sample, which includes serial number, price, country, province, designation, region\_1, and region\_2. Designation reports the vineyard within the winery where the grapes that made the wine are from. Region\_1 specifies the wine growing area within a province or states such as Napa. Region\_2 records more specific regions within a wine growing area such as Rutherford inside the Napa Valley. All of the sample has the country and province information but only a portion has the specified region. The second type of variable collects the reference information that Wine Enthusiast provides for their customer to choose wines, such as points, description, and taster\_name. Points are the number of points that Wine Enthusiast rated the wine on a scale of 1-100. Description is the unique illustration for each sample that detailedly describes the flavor, smell, and taste of the wine. Taster\_name records the name of the person who tasted each sample and wrote the description.

After some basic analysis of the dataset, we could see that 42% of the samples are produced in the US, 17% are produced in France, and 41% are from other countries. Most of the samples are priced between $4 and $70. The average price is $35.36, with a maximum price of $3300 and minimum price of $4. The points that Wine Enthusiast rated to the sample has an average of 88.45, with a standard deviation of 3.04. Since our model is going to predict by tastes and flavors, the variable we will focus on is the description, which will create all of the new variables in our modeling process.

## Data Preparation

## The ideal data set is easily analyzed and classified, but our raw data is a mixture of simply analyzed pure numbers, formatted worded data and complex context especially like the description of wine. The description of wine all looks like “This tremendous 100% varietal wine hails from Oakville and was…” , this obviously needed to transform to something easily analyzed. The current method is to find clusters or groups that can represent those descriptions. But at this point, the main job for Data preparation is to do data cleaning.

The data cleaning procedure is divided into two main steps, fulfilling the missing values and cleaning the complex description. Creating corpus is utilized to make all words to lowercase, remove punctuation, remove stopwords and strip whitespaces. Then remove the sparse words by stemming the corpus and tokenizing it, then using “removeSparseTerms” to remove the sparse words.

## Working R code for Analysis: Share R code used to run the various analyses attempted. This block of R code should contain all the different analyses conducted, those that were used for the final conclusions and the ones that were not. The R code may be submitted as an R script (.R), R Markdown (.rmd), or a knit R Markdown (.html) file.

## Final R Code: Share R code that contains only the final analysis. In addition, this should also include code for any charts created to corroborate the conclusions. The R code may be submitted as an R script (.R), R Markdown (.rmd), or a knit R Markdown (.html) file.

## Choice of Analytical Techniques

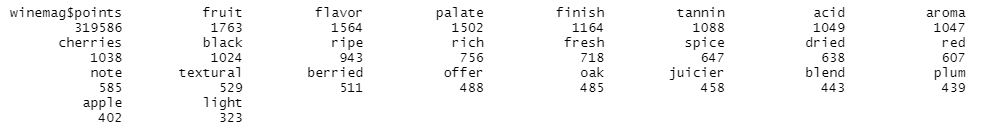
To fully understand our data and ultimately provide professional recommended wine varieties, we select several suitable analysis techniques. First, we use text mining to gain an intuitive and preliminary understanding of the data, such as exploring the relationship between feedback and scores. Binary sentiment analysis and Sentiment Score are commonly used to analyze and discover customer feedback. Then we use clustering to understand the distribution of wine varieties. Clustering helps us group data with similar characteristics together. By understanding more of the different segments, we can target consumer preferences more accurately. We will focus on the Mclust method, which is more efficient and provides the best set of underlying distributions to explain the observed data. Finally, after understanding the varietal distribution, we start making predictions for the wine category.

We use the binomial regression model because of the categorical variables that we selected. Variety is the dependent variable and independent variables are words that we extracted from text mining. The regression model not only has the advantages of simplicity and ease of implementation but also explains the relationship between binary response variables and other explanatory variables (Muschelli et al., 2014). Decision tree models also fit our predictions for variety. Considering the visual analysis and easy extraction of information from this model, we try to consider all possible outcomes of the decision and trace their paths. For example, we use the variables to predict whether consumers will buy and which category of wine they will buy. We consider the limitations of decision trees and try a highly accurate random forest model. In the end, we found that this model is not a good choice for high-dimensional datasets such as text classification based on actual data (Yıldırım, 2020).

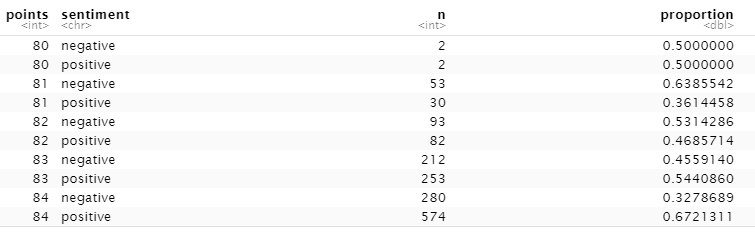
Finally, we decided to use the recommender system model to make predictions, because the recommender system fits our goal to help users find the wine based on their preferences. The system can recommend wines to customers based on user ratings of products. More specifically, there are two main types of collaborative recommender systems (Wu, 2019). We first use the User-Based Collaborative Filtering to perform the nearest neighbor search. After determining the similarity of users' scores, we can predict the top 5 recommended varieties for users. Second, we use an Item-Based Collaborative Filtering to find similarities between products by considering the ratings they have obtained. For example, after a user adds a few new product descriptions, the system can immediately provide recommendations for new wines. Using a recommendation is important because it allows us to weigh the recommendations we give: we will give preference to wines that are closer to a 90 rather than an 80.

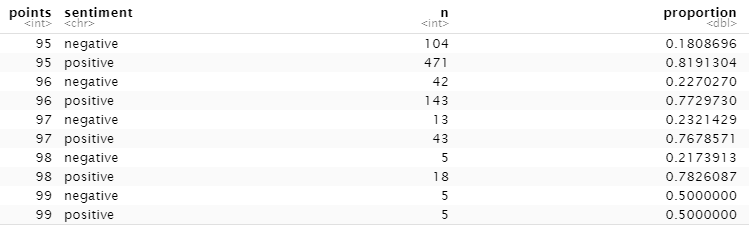
## Results Discussion

Our first technique is to do text mining which is for understanding the reviews for top-rated(score > 90 points) wines. Our results showed that the top 5 most frequently used words among the reviews of top rated wines are ‘fruit’, ‘flavor’, ‘palate’, ‘finish’ and ‘tannin’. ‘Fruit’ occurs 1763 times, ‘flavor’ occurs 1564 times and ‘palate’ occurs 1502 times, which are more frequent than ‘finish’(1164 times) and ‘tannin’(1088 times). ‘Fruit’, ‘flavor’, ‘palate’ and ‘finish’ aren’t meaningful to tell the characteristics of top-rated wines. Other less frequently occurring words could be more worthy to tell something about those top-rated wines, like ‘tannin’, ‘acid’, ‘aroma’ and ‘cherries’. Those words are more specific to describe what kinds of wines will get a higher score. Below is the screenshot of our results:



Binary sentiment analysis could tell us the relationship between score and the proportion of positive words. From our results, the proportion of positive words for those wines with scores under 85 are around 0.5, but those top-rated wines(score > 90) have a large variety of the proportion. There is no clear evidence showing that wines with higher score-point can have higher proportion of positive words in reviewers’ description.

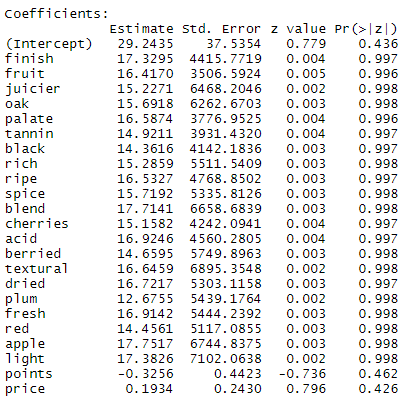




Another thing we did is to get the sentiment score for the score-point, the result shows that the value of review sentiment of score-point under 85 is all smaller than 1, but that of score-point above 90 is even around 2. Thus we can say that a higher score has a larger value of sentiment.



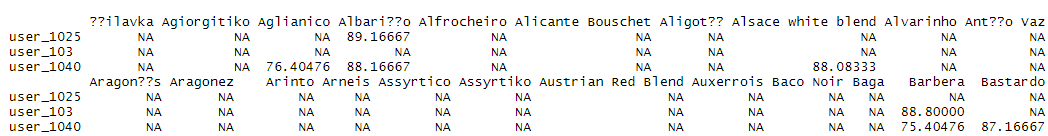
As per the result of the binomial regression model, there are several words, like ‘finish’, ’fruit’, ‘juicier’ and ‘oak’, from the review that have the relationship to the variety. ‘Points’ and ‘price’ have weak relationships to the variety, which only have coefficients of -0.3256 and 0.1934 respectively. In total, there are 21 words that have certain relationships with the variety and the snip is attached after.



The last technique we used is a recommendation lab including user-based collaborative filtering and item-based collaborative filtering. Under UBCF, we get a list of top 5 recommended varieties for users, for example, we get the list of "Ros??", "Chardonnay", "Pinot Noir", "Cabernet Sauvignon" and "Sangiovese" for user 103,



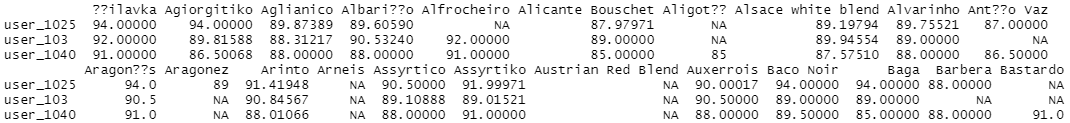
and also get a prediction of all varieties for this reviewer.



Under IBCF, we get a list of top 5 recommended varieties for users as well, for example, we get the list of "Alfrocheiro", "Aragon??s", "Arinto", "Assyrtiko" and "Baco Noir" for user 103,



and also get a prediction of all varieties for this reviewer too.



Conclusion and Recommendation

Based on our sentiment analysis, we see that a higher score that Wine Enthusiast reviewers gave to a sample does not lead to a higher proportion of positive words in the description of that sample. Besides, the proportion of positive words is not the only standard of measuring wine. Some consumers might like the products with flavor described by negative words such as “bitter” or “astringent”.

On the other hand, the result on the sentiment score shows that a higher score of Wine Enthusiast represents a larger value of sentiment. We believe that a sample with more emotional content means that it attracts the reviewer more as he or she tastes and writes the review. However, sometimes it is possible that samples with extreme negative taste also have a high sentiment score.

Therefore, combining two results, we recommend that users should not make their choices by just looking at the scores that Wine Enthusiast has. It would be more helpful to read the description rather than the score. As a reference, users should pay careful attention to products with scores that are higher than 91 and lower than 85, which are the scores 1 unit of standard deviation from the mean score.

As our hypothesis stated, we would like to create a recommendation system that makes accurate predictions on consumers’ preferences. According to the result, our model is able to predict the estimated product for consumers based on their description. Consumers will be asked to give a few words to describe the taste or flavor, then our recommendation system will come up with the top 5 products that match their description. Our result shows that the 2 recommendation methods we used have come up with completely different products. In our case, user based collaborative filtering matches samples with users and recommends products based on patterns of similar reviewers. Item based collaborative filtering matches the testing sample with all other samples and recommends the products with the most similarities. We would like to recommend users to look at both solutions before making decisions. It is possible for either of them to be more helpful for users’ decision making.

## 

## 

## Citation

Ducman, A. A., Oprea, S., Adela, B., Ceaparu, C., & Ene, G. D. (2021, March). *Analyzing the impact of nonformal education in Wine Marketing*. Retrieved July 5, 2022, from https://www.researchgate.net/publication/350421468\_ANALYZING\_THE\_IMPACT\_OF\_NONFORMAL\_EDUCATION\_IN\_WINE\_MARKETING

Gazzola, P., Grechi, D., Pavione, E., & Gilardoni, G. (2022, January 6). *Italian wine sustainability: New trends in consumer behaviors for the millennial generation*. British Food Journal. Retrieved July 4, 2022, from https://www.emerald.com/insight/content/doi/10.1108/BFJ-05-2021-0493/full/html?skipTracking=true

Loose, S., & Nelgen, S. (2021, April). *State of the German and international wine markets*. Retrieved July 5, 2022, from https://www.researchgate.net/profile/Simone-Mueller-Loose-2/publication/350343082\_State\_of\_the\_German\_and\_International\_Wine\_Markets/links/605ae76892851cd8ce61c9e6/State-of-the-German-and-International-Wine-Markets.pdf

Martinho, V. (2021, July). *Contributions from literature for understanding wine marketing*. Retrieved July 5, 2022, from https://www.researchgate.net/publication/352999761\_Contributions\_from\_Literature\_for\_Understanding\_Wine\_Marketing

Muschelli, J., Betz, J., & Varadhan, R. (2014, November 27). *Binomial regression in R*. Handbook of Statistics. Retrieved August 15, 2022, from https://www.sciencedirect.com/science/article/abs/pii/B9780444634313000073

Rihn , A. L., Jensen, K. L., & Hughes, D. W. (2022, February). *Consumer perceptions of Wine Quality Assurance Programs: An opportunity ...* Retrieved July 5, 2022, from https://www.researchgate.net/publication/358760728\_Consumer\_Perceptions\_of\_Wine\_Quality\_Assurance\_Programs\_An\_Opportunity\_for\_Emerging\_Wine\_Markets/fulltext/62195a0c579f1c04171a0822/Consumer-Perceptions-of-Wine-Quality-Assurance-Programs-An-Opportunity-for-Emerging-Wine-Markets.pdf

Wu, J. (2019, May 27). *Types of Recommender Systems*. Medium. Retrieved August 15, 2022, from https://medium.com/@jwu2/types-of-recommender-systems-9cc216294802

Yıldırım, S. (2020, February 11). *Decision tree and Random Forest - explained*. Medium. Retrieved August 15, 2022, from https://towardsdatascience.com/decision-tree-and-random-forest-explained-8d20ddabc9dd