A Taste of French Wine

<Yun-Chung LIU>

Abstract

In this study, I briefly described the number of products and price level for French wine in the dataset extracted from wine enthusiast website. Then, I used the Aroma Wheel paradigm to dig out the taste of French wine from different provinces and presented a aroma-based recommendation system. Finally, I compared two classification methods, Naïve Bayes and K-Nearest Neighbor. The result suggests that Naïve Bayesian classifier can can predict the origin of French wine more accurately.

Motivation

Have you ever wondered how the wines from the old world taste like? Bordeaux, Chardonnay, Pinot Noir and other words that make you merry? In this study, I used the data from Wine Enthusiast to help you have a glimpse on the amazing world of French Wine. You will ”see” how these wine “taste” and “find” the wine for you! Come check it out!

Dataset

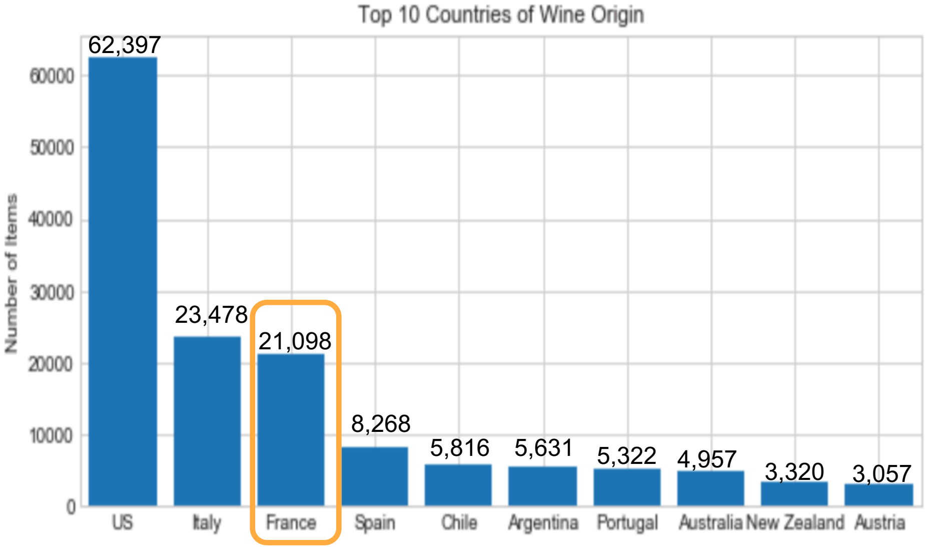
In this project, I used the wine review dataset downloaded from Kaggle (https://www.kaggle.com/zynicide/wine-reviews). The dataset was built by Zackthoutt using wine review data from Wine Enthusiast website (https://www.winemag.com/). The dataset contains more than 150,000 wine items with country of origin, region, province, designation, description, price, points given by tasters.

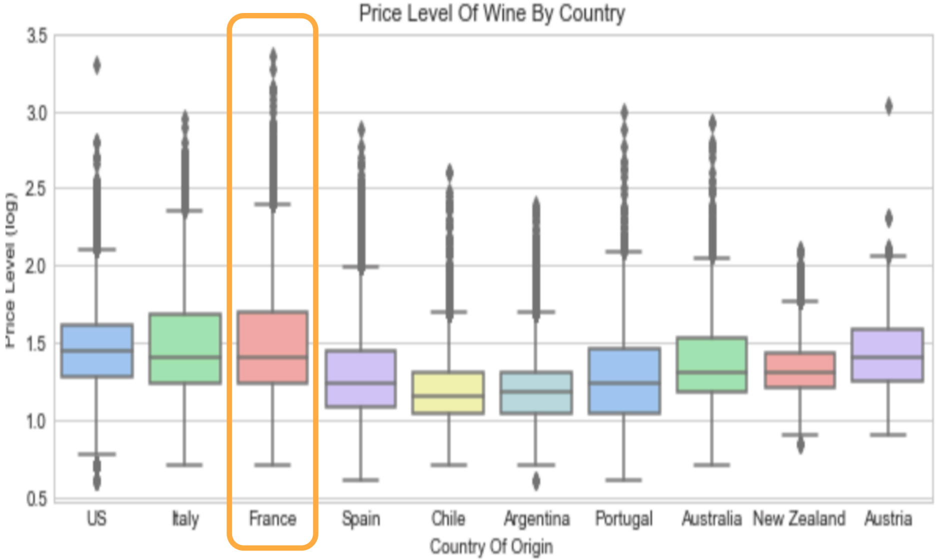
Research Questions

1. What are some characteristics in terms of number of items and price of French wine on the market?
2. What are some common words used to describe French wine? Can we use descriptions to distinguish province of wine origin with similar aroma?
3. Can we use descriptions of wine to predict the province of wine origin? Which classification method, Naïve Bayes or K-Nearest Neighbors, performs better?

French Wine 101

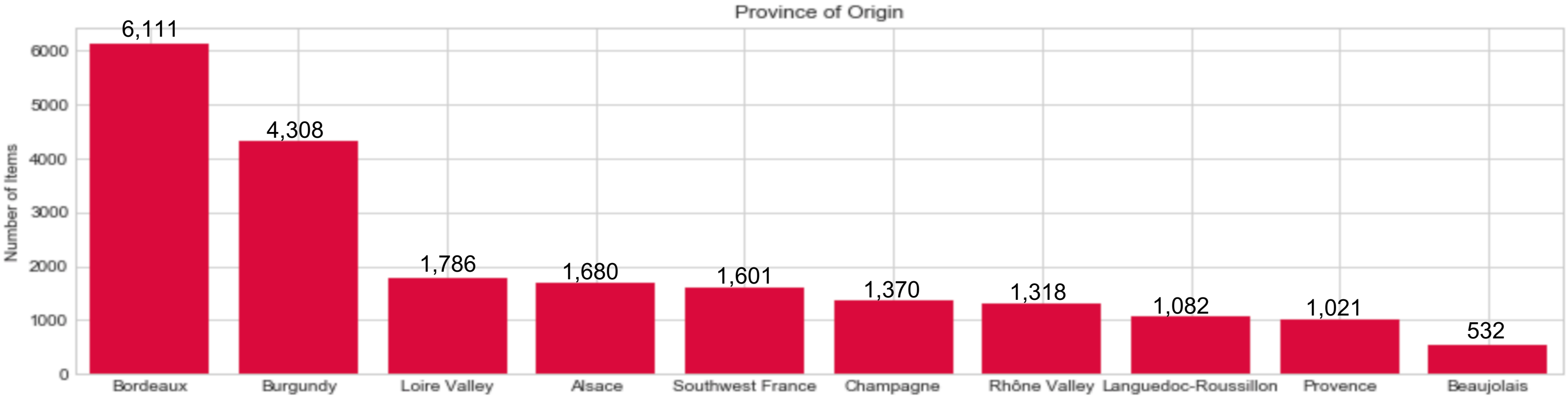
French wine represents 14.0% of items in the Wine Enthusiast platform, ranked the 3rd after the US and Italy. The average price of French wine is $45.6 (3rd among other countries) and the standard deviation is $69.7(1st among other countries), which indicates that price of French wine can vary.



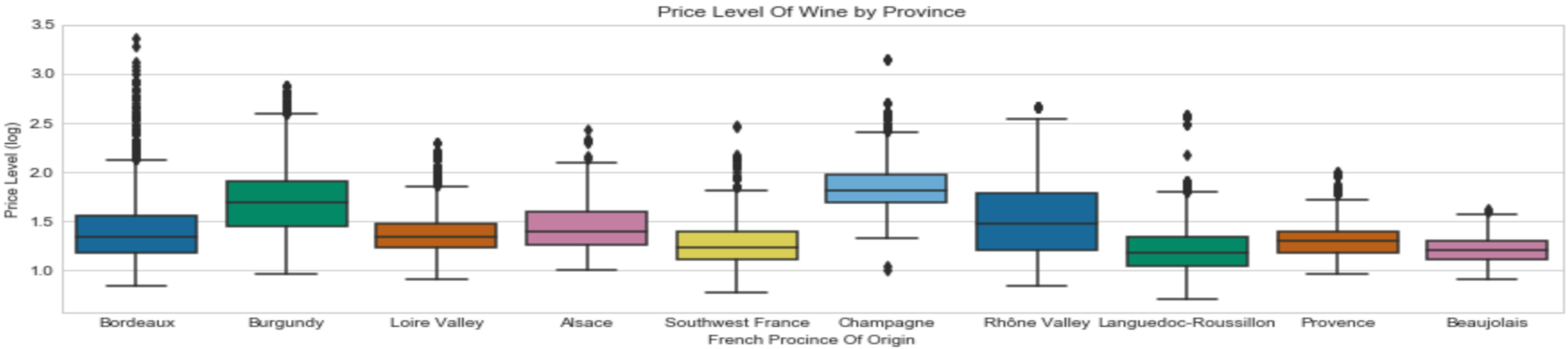


French Wine by Province

Most French wine comes from Bordeaux, representing 29.0% of French wine products. Burgundy wine ranked the second(20.4%). In terms of price, Champagne(avg. $93.4) and Burgundy(avg. $70.6) stand out. The average price of Bordeaux is not too high($42.6), but it has relative large dispersion.



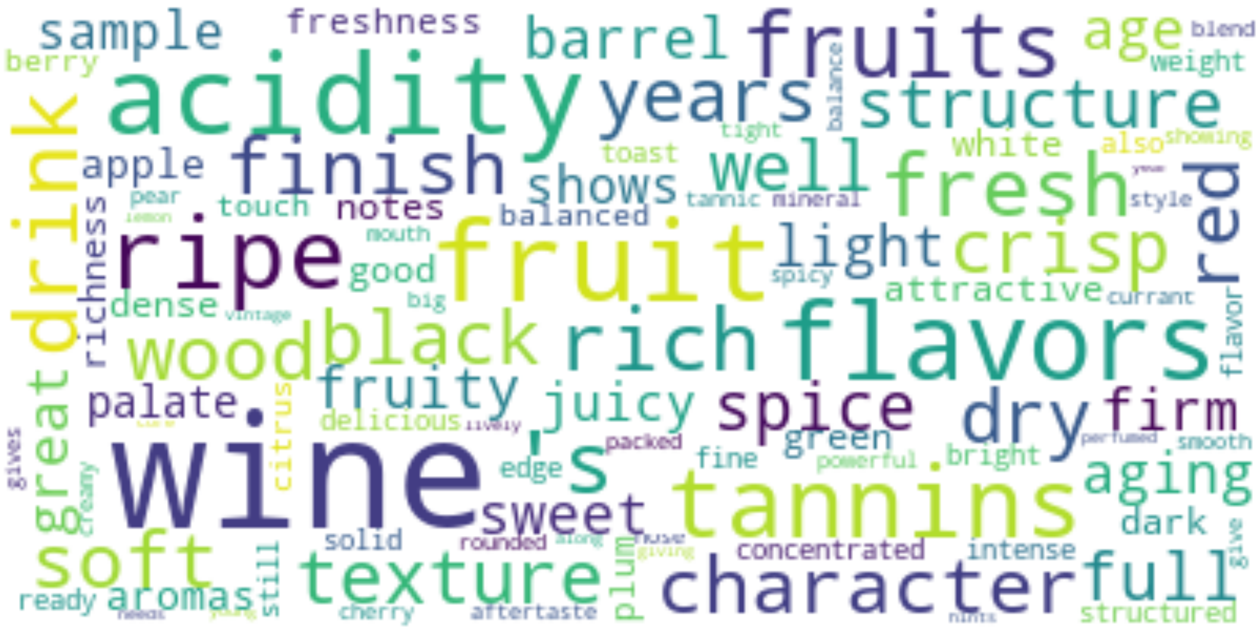
Province of Origin



Price Level Of Wine by Province

A taste of French Wine

How does French wine taste like? Using all the description for French wine, several terms appear repetitively (word size represents the frequency of appearance in the corpus), such as fruit, fresh, tannins, dry, firm, wood, soft..etc. However, some of them are irrelevant to the taste (e.g. wine, structure).



The Aroma Wheel

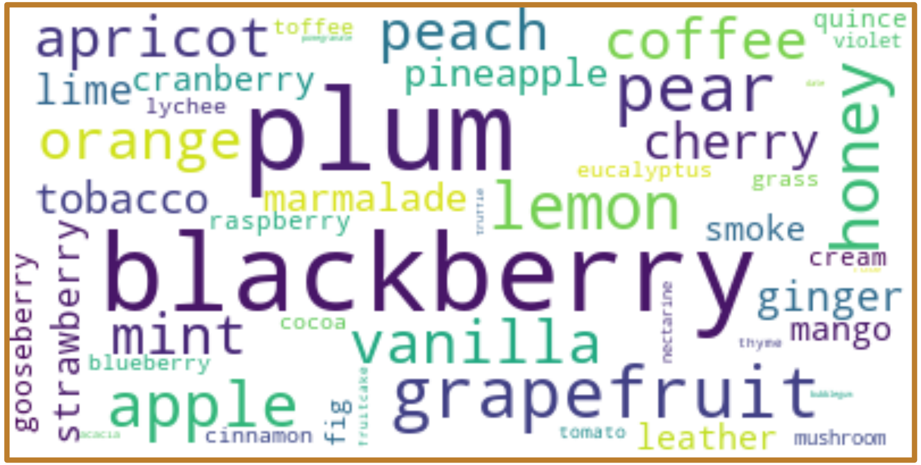
To filter out words that are related to the taste of wine, I used a modified version of the wine aroma wheel paradigm(https://winefolly.com/). In this paradigm, Noble et al. listed a wheel of words related to the wine aroma (related to the underlying chemical substances) to represent the taste of wine. The words are categorized in 3 tiers(see example below). In the project, I use the 3rd tier terms(the most detailed ones).

Table 1. The Aroma Wheel

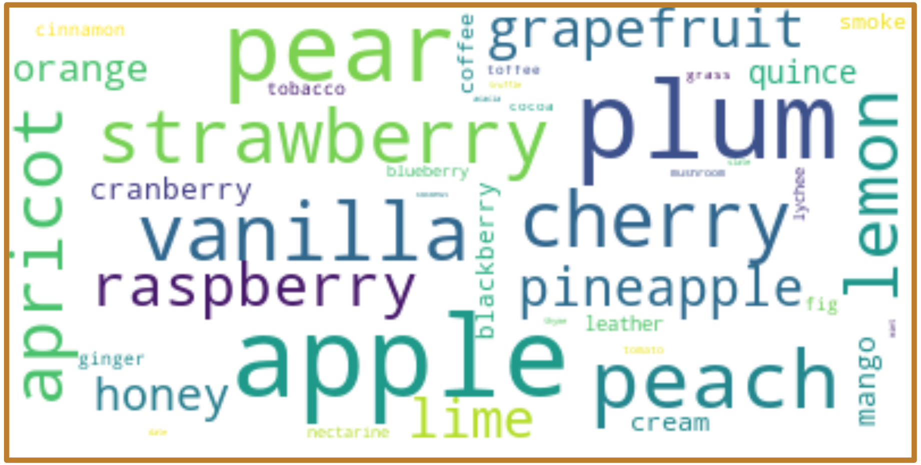
Principal term	2nd tier term	3rd tier term
Caramelized	Carmel	Honey
Wood	Phenolic	Vanilla
	Burned	Smoke
Fruity	Citrus	Grapefruit
		Lemon
	Tropical Fruit	Pineapple
		Mango

The Aromas of Bordeaux and Burgundy Wine

Using the Aroma Wheel Paradigm, we can see the similarity and difference between Bordeaux and Burgundy wine. In Bordeaux wine description, plum (word count: 429), blackberry(word count: 419) and grapefruit(word count: 127) appear the most often. For Burgundy, people consider it to have the aroma of apple(word count: 397), plum(word count: 367) and pear(word count: 295) more often.



The Bordeaux Aroma



The Burgundy Aroma

The Aroma Distance

To quantify the similarity between different origin of wine, I first calculated the prrportion of every aroma word used to describe the wine (e.g. x_1 = wordcount of Aroma 1 / total Aroma word count of Province A).The Aroma Distance is the sum of the square root of squared difference between all Aroma words in the Aroma Wheel between Province A and B (see table 2).

Table 2. The Aroma Distance Methodology

Aroma Word	Province A	Province B	Distance
Aroma 1	x_1	y_1	$\sqrt{(x_1 - y_1)^2}$
Aroma 2	x_2	y_2	$\sqrt{(x_2 - y_2)^2}$
Aroma 3	x_3	y_3	$\sqrt{(x_3 - y_3)^2}$
… Aroma n	… x_n	… y_n	… $\sqrt{(x_n - yn)^2}$
The Aroma Distance between Province A and Province B $(0 \leq \text{distance} \ \leq 2)$			$\sum_{i=1}^{n} \sqrt{(x_n - yn)^2}$

The Aroma Distance Based Wine Recommendation

Using the aroma distance, we can recommend people the wine with the shortest Aroma Distance. For example, a person who likes Bordeaux a lot may be tempted to try a wine from Southeast France. A Burgundy fan can try a wine from Loire Valley. He or she may like it a lot because of the similar aroma.

Aroma Distance from Bordeaux wine

Province of Origin	Distance
Southwest France	0.43
Burgundy	0.79
Languedoc-Roussillon	0.92
Loire Valley	0.91
Rhône Valley	0.99
Provence	1.05
Alsace	1.05
Champagne	1.07
Beaujolais	1.37

Aroma Distance from Burgundy wine

Province of Origin	Distance
Loire Valley	0.59
Champagne	0.63
Provence	0.70
Southwest France	0.71
Alsace	0.74
Bordeaux	0.79
Languedoc-Roussillon	0.84
Rhône Valley	0.92
Beaujolais	1.27

The Aroma Distance Matrix

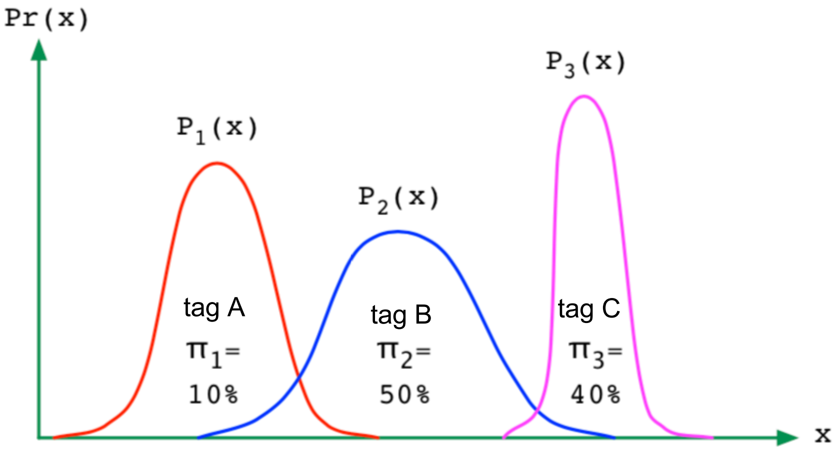
This matrix below summarizes the Aroma Distance between province of wine origins. Pink indicates short distance (0.44 – 0.91); purple represents medium distance(0.92-1.07) and blue implies long distance(1.08-1.63).

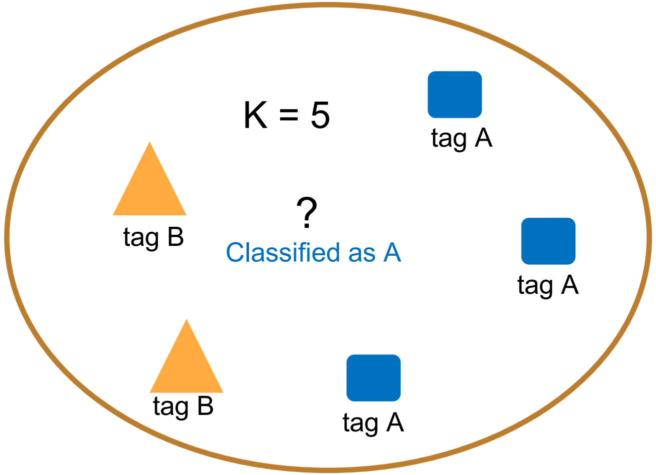
	Champagne	**Bordeaux**	**Burgundy**	**Rhône Valley**	**Languedoc-Roussillon**	**Alsace**	**Southwest France**	**Loire Valley**	**Provence**	**Beaujolais**
Champagne	-	1.07	0.63	1.24	1.20	0.64	1.00	0.44	0.96	1.63
Bordeaux	1.07	-	0.80	0.99	0.92	1.05	0.43	0.91	1.05	1.37
Burgundy	0.63	0.80	-	0.92	0.84	0.74	0.71	0.59	0.70	1.27
Rhône Valley	1.24	0.99	0.92	-	0.69	1.25	0.99	1.15	1.06	1.19
Languedoc-Roussillon	1.20	0.92	0.84	0.69	-	1.30	0.95	1.14	0.85	1.07
Alsace	0.64	1.05	0.74	1.25	1.30	-	0.97	0.54	1.11	1.78
Southwest France	1.00	0.43	0.71	0.99	0.95	0.97	-	0.83	0.95	1.36
Loire Valley	0.44	0.91	0.59	1.15	1.14	0.54	0.83	-	0.99	1.61
Provence	0.96	1.05	0.70	1.06	0.85	1.11	0.95	0.99	-	1.29
Beaujolais	1.63	1.37	1.27	1.19	1.07	1.78	1.36	1.61	1.29	-

Classifying Province of Wine Origin

Lastly, I compared the performance of two classification methods. Since words not included in the aroma wheel could affect the classification result (e.g. “sunshine”, “texture”), I included all the words in the description for the classification tasks.

Naïve Bayes Classification	K-Nearest Neighbors
Classify wine origin based on Bayes theory considering 1) probability of wine origin in the dataset 2) probability of word(features) describing wine from different province.	Calculate the distance between descriptions (the more words appeared in both description A and B, the less the distance is). Classify a new description based on the wine origin of the majority of k-nearest neighbors.





Classifying wine origin based on description

The result suggests that Naïve Bayes classifier performs better (accuracy rate: 70.1%) than K-Nearest Neighbors (accuracy rate: 21.3%).

	Naïve Bayes Classification	**K-Nearest Neighbors**
Preprocessing	Use the Natural Language Toolkit(NLTK) tokenizer to generate word list of wine descriptions, excluding meaningless and repetitive words, and punctuation.	
Train dataset	14,768 randomly selected wine description.	-
Train Accuracy	78.0%	-
Test dataset	6,329 randomly selected wine description.	500 randomly selected wine description * 10 provinces = 5,000 samples.
Test Accuracy	70.1%	21.3%

Discussion

Comparing the 2 classification methods, I found that Naïve Bayes classifier demonstrated far higher accuracy rate than K-nearest neighbors(KNN). This can be due to the fact that wine descriptions has higher sparsity. There are not many words on a single wine description. Thus, it works better when comparing a piece of wine description to the whole set of wine tag and description data(Naïve Bayes) rather than a single piece of wine description (KNN). Also, sample size matters, due to the nature of KNN, the sample size is limited to the wine origin with the lease amount of wine product.

Conclusions

In this study, I first revealed the status of French wine in terms of price and number of products on the market for consumer’s reference. Then, I presented the taste of wine from different origins using word cloud. Based on the proportions of aroma words used to describe wine from different province, I developped an Aroma Distance based recommendation system for people to try to wine with aroma closer to their favorite wine origins. Finally, I compared 2 classification methods, Naive Bayes and K-Nearest neighbors. The result suggests that Naive Bayes classifier performs better on the task(accuracy rate higher than 70%).

This study demonstrates a possible recommendation paradigm based on natural language processing. It can not only be applied to wine recommendation system, but also it also demostrate a more effective methods for classification for few tags and large amount of features.

Limitations

The wine dataset used in this study came from a single US Based website. Some characteristics and products can be under-represented. Also, due to the limitation of time and resources, this study remains an exploratory analysis and demonstration of methodology. For example, I only considered one word aroma in the aroma wheel. Further effort is needed to provide information with business impact or insights on machine learning methodologies.

Future Work

With record of consumer buying behavior and website log, more sophisticated recommendation system can be formed. Thus, the effectiveness of the recommendation system can be measured and fine-tuned continuously.

Acknowledgements

First of all, I would like to dedicate this work to Christophe PALLIER, who inspired me to the world of Python. In this series of endeavor, I would like to first of all give many thanks to Fang-Yu, CHENG, gave me advice along the way of my data science-related endeavors. Also, Pei-Chun CHIANG gave me the idea of using the Aroma Wheel paradigm. Finaly, many thanks to Li-Tsu YANG for reviewing the materials.

References

Noble, A.C., Arnold, R.A., Buechsenstein, J., Leach, E.J., Schmidt, J. O. and Stern, P.M.(1987). Modification of a Standardized System of Wine Aroma Terminology. *American Journal of Enology and Viticulture*, 32(2), p.143-146.