It’s Wine’o Clock Somewhere !

CMPT- 353 Project Report

Amazing Fact: Nearly all red wines are made from one species of grape!

Wine is one of the most famous drinks in the world; with its comforting nature and health benefits, it's a drink for almost any occasion. However, selecting the perfect bottle is no easy task . With wine becoming more popular each year, there are more options than ever before. This has put pressure on winery owners and MNC’s to provide high quality wine to stay competitive in the market.

The First Question ? But how can we determine the quality of wine?

Is it better for the importers look at the chemical composition of the wine? Or should they consider looking at the geographical locations?

Overview:

Almost every wine is a mixture of water, ethanol, and a combination of acids, sugars, volatile flavours, etc. This unique mixture makes all the difference. The different proportions of these ingredients make one's mouth sweet, sour, salty, and bitter. Thus, the chemical composition of wine becomes an interesting topic to research.

But wine is made from grapes, and grapes need the right conditions to grow. The humidity, temperature, and precipitation of the region will have a large affect the quality of grapes. There is a big difference between the climate of British Columbia and the south of France, and the soil conditions in Australia versus that in Germany. Thus, where the wine is grown will also have a huge effect on the quality of the bottle.

Let's dive straight in..

Exploring the datasets

In order to answer two different questions, two different datasets were exported from kaggle. The first data-set(*link)that we explored was scraped from the WineEnthusiast during the week of June 15th, 2017 with 150,931 rows.

This data-set 1 has the following features:**Country**: The country of wine. **Description**: Description of wine.**Designation**: The vineyard location. **Points**: Points given on scale 1-100 **Price**: Cost of wine **Province**: The province or state name .**Region 1**: Wine growing area in a province or state.**Region 2**: More specific wine growing area .**Variety**: The type of grapes used to make the wine.

The second data-set(*link) deals solely with the chemical compositions of the wine with 7500 rows. The features included in second data-sets are : **Acidity**: Fixed or non-Volatile acids.**Volatile Acidity:** Amount of acetic acid in wine.**Citric Acid:** Amount of Citric Acid . **Residual sugar:** The amount of sugar after the fermentation stops.**Chlorides:** The amount of salt.**Free Sulphur Dioxide:** The dissolved gas in wine.**Density:** The density of the wine. **pH:** Acidic or Basic scale. **Quality:** The rating of wine from 10.

Data Cleaning:

The first step involves removing missing values. As analysis involves geographical locations, we removed unrelated columns. For instance, the Price and Wine Description columns. After a preliminary cleaning, it was found that Region 2 was missing approx. *50,000* values. For this analysis, the decision was taken to perform analysis with and without *Region* 2 column. Next, the outliers were removed. Since the rating is out of 100, any values greater than 100 and less than 0 will be an outlier.

Data-set 2 comes in 2 separate CSV files, as the chemical composition of white and red wine is different. Both data-sets required cleaning of columns names. For Example: "Unknown" and having "." in column names which can be confusing for pandas data-frame. After analyzing the scatter plots, the density of one wine was found close to density of milk (1.4ml/l). Sulphur Dioxide and free sulphur dioxide of one data-point were around 300 while others were less than 150.

Data Transformation:

For **Dataset 1**, most of the values were composed of categorical data. The data sets were locations, for instance, countries: the US, France, Italy, etc. Similarly for Province and Regions. Since we can't apply Machine Learning tools to categorical data, the data as converted to numerical categories.

For instance, US = Category 1, France = Category 2. After changing each column our data is ready for performing Machine Learning analysis and making predictions.

For **Dataset 2**: As our data is composed of numbers, looking at the bar plot for each feature has shown that the data is normal. Since, some ML techniques require the data to be normal, doing this analysis made us confident that it is ready for the Machine Learning algorithms.

The Figure 1on the left shows the density of white wine before cleaning and Figure 2 shows the density after removing outlier.

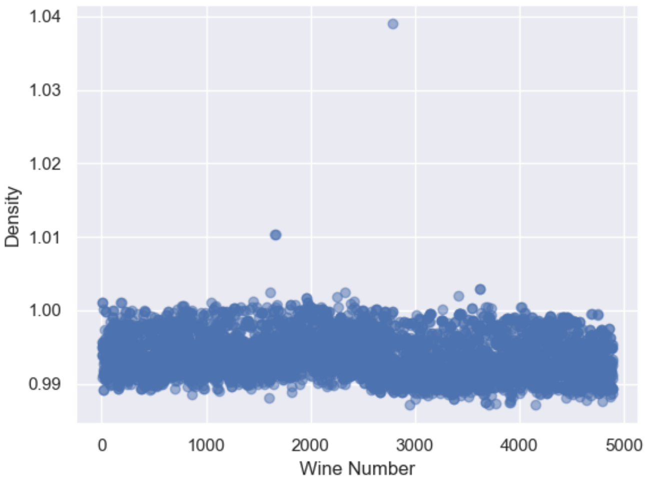


Figure 1

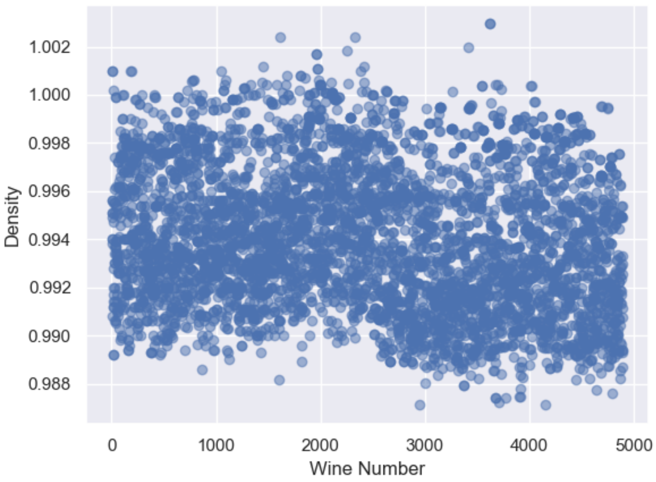


Figure 2

Data Analysis:

Let's analyze the ***Location Dataset*** first. The first Machine Learning algorithm used was **Decision Trees**. Decision trees excel at making complex decisions, so was a natural choice when dealing with many features. The next technique we explored for analyzing the results was **K-Nearest Neighbours**. Given a new wine, K-Nearest Neighbours looks at ratings of wines that are most similar to the new wine. This technique is well suited for this data. For instance, the scores in US will be related, scores in common provinces will be related, and so on for regions and Provinces. Although, other algorithms were used for predictions, these two algorithms seemed the best fit for the analysis.

Let's analyze the ***Physiochemical Properties Dataset***: This dataset needed to have all the physicochemical properties scaled to the same range, so as not to over-value certain properties. This was achieved using the Standard Scalar.

Scaling values increased the scores of the ML models considerably.
Since, our data was fairly normal, The first ML technique used was **Gaussian Naive Bayes.**
Which Machine Learning technique should we try next? An accomplished data scientist provides the answer:
"Boosted Trees are often a good first choice for a new problem."

- ML notes, Greg Baker

Who's the Winner ?

Models	Scores(Red Wine/ White Wine)
Gaussian Naive Bayes	56 / 44.8 %
Gradient Boosting	68 / 70 %
K-Nearest Neighbours	60 / 57 %
Decision Trees	59 / 50 %

Models	Scores(With/ Without Region 2)
Gaussian Naive Bayes	14 / 13 %
Gradient Boosting	16 / 16 %
K-Nearest Neighbours	22 / 25 %
Decision Trees	21 / 23 %

Conclusion: From the above results, the validity score of Chemical Composition is more than location reviews, Thus, we can conclude that winery owners and MNC's can use chemical composition of wines to predict the quality of their new wines. This may assist their decision in pricing the wine.

The Second Question ?

Wine is a huge global market. With so many bottles changing hands each year, it is important to stay educated if you are looking to purchase. This section begins by giving a brief statistical overview of wine ratings, and then performs comparison tests between ratings given by professionals and amateurs. The report then looks at the most commonly rated wine regions to determine which regions produce, on average, better wines.

Overview:

Each year, billions of bottles of wine are produced and consumed[4]. In Canada alone we consume over 600,000,000 bottles of wine each year, that's right, six-hundred million bottles[4]. With so many bottles changing hands, it's important for the consumer to be able to make an educated decision about what wines to buy. Fortunately, there are many easily available sources for wine ratings. In fact, there are professional wine reviewers, called sommeliers, that rate wines*. Amateurs also enjoy reviewing bottles of wine; there is a website called Vivino[5], where anyone can give their own opinions and ratings for a bottle of wine. This paper focusses on bringing to light some statistics about the world of wine and then looks at a comparison between the ratings of sommeliers and amateurs. We finish up by looking at the most popular regions for wine and whether there are some regions that are better than others for purchasing highly rated wines.

* If you're interested in seeing what it takes to become a sommelier, I strongly recommend watching the SOMM documentaries.

Methods:

Data from Vivino was scraped [see work done by Kevin/Raman], to produce data for nearly eight thousand wines (n = 7731). The data was composed of red and white wines from 26 countries, with ratings between 3.0 to 5.0 (ratings out of 5.0). When the csv file was read in, all numeric data types were defined as float due to a lack of NA support by numpy[6]. After cleaning the data, the Vivino data frame had the following schema:

[**wine** : string, **vintage** : float, **winery** : string, **country** : string, **region** : string, **ratings** : float, **number of ratings** : float].

Data rated by sommeliers was also obtained as a dataset from Kaggle. This dataset was composed of red and white wines from the United States with ratings above 80/100.

After some data processing, the schema was:

[**country** : string, **ratings** : float, **price** : float, **region** : string, **wine** : string, **variety** : string, **winery** : string, **vintage** : float]

The points were converted to ratings out of 5 to match with the Vivino Data. For both datasets, any rows with missing or null values were dropped. Figures were created as either scatter, histogram, or boxplot using matplotlib.

Summary statistics containing the mean, standard deviation, min, and max for the ratings of each dataset was obtained using the scipy stats module. The covariance between ratings and the number of ratings of the vivino set was evaluated using the linregress function of the scipy stats module, as was the covariance between price and ratings for the professional data set.

To compare the two datasets, their ratings were scaled to a range of 0-1. A normality test was performed on the ratings of the datasets using the normaltest function for the scipy stats module and the variance was calculated from the levene function.

A Student's T test was performed on the scaled data using the ttest_ind function from the scipy stats module, followed by a Mann Whitney U test using the mannwhitneyu function.

For the regional comparisons, regions from the Vivino data set with at least 40 ratings total were selected, and the means of each region were calculated. The data was then split by region so that an ANOVA test could be performed. A pairwise Tukey HSD test was conducted post hoc.

Results: To get a good idea of what the data looked like, some preliminary graphs were generated for both the Vivino and sommelier data set. Supplementary Figures 1-3 provide examples of some of these graphs (more graphs can be found in the figures folder provided).

Box plots and histograms (Figures 1 & 2) show the distribution of the ratings overall from both datasets. The preliminary stats for the Vivino and sommelier data are summarized in Table 1.

From the Vivino dataset, I thought it may be that the popularity of a wine could influence its rating. To determine if this was the case, a covariance test between number of ratings and ratings was performed. The result obtained was a correlation coefficient equal to 0.084. Since the value is close to zero, this indicates that there is no linear relationship between the number of ratings and the ratings.

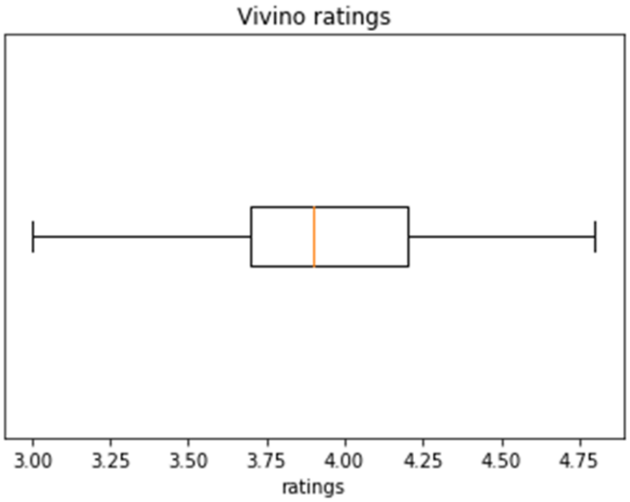


Figure 1A: Boxplot of Vivino Ratings

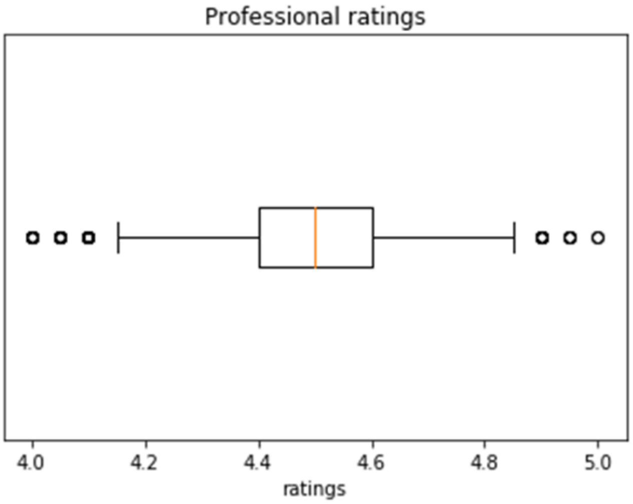


Figure 2A: Boxplot of Professional Ratings

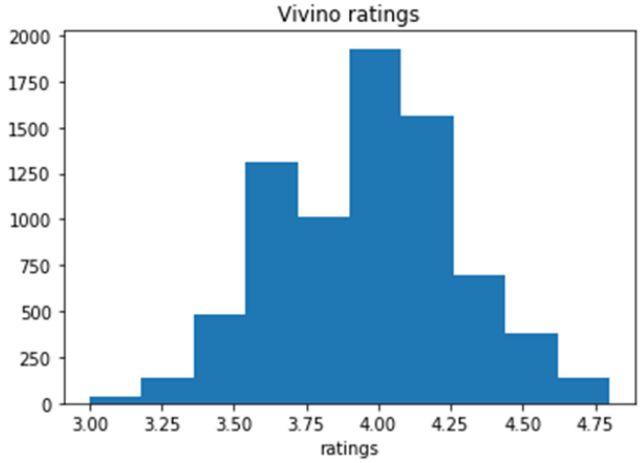


Figure 2A: Histogram of Vivino Ratings

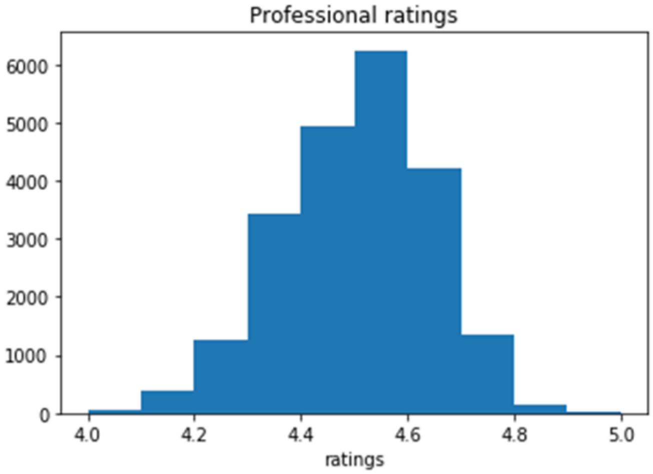


Figure 2B: Histogram of Professional Ratings

I also suspected that the ratings of the professionals may increase linearly with price. The correlation coefficient for the covariance test between price and ratings was 0.39. There was therefore a linear relationship between price and rating but a weaker one than expected. Whether this relationship is due to sommeliers rating wines higher based on their price tag or whether higher quality wines are more expensive is an area for future research.

The ratings from both data sets were scaled to the range 0-1 so that their means could be compared. A normality test was performed on each data set, and both failed (Table 2). However, since we have a large number of samples and the histograms of the data looks roughly normal (Figure 2), it was determined that the data was normal enough.

Data Set	Count	Mean	Standard Deviation	Minimum Value	Maximum Value
Vivino	7693	3.95	0.31	3.0	4.8
Professional	22037	4.48	0.14	4.0	5.0

Table 1: Summary Statistics for Vivino and Professional Ratings

Data Set	Mean	Normal test p-value
Vivino	0.53	9.9e-09
Professional	0.47	3.4e-50

Table 2: Means and Normal Test Results for Vivino and Professional Scaled Ratings

A levene test was then performed on the two data sets to determine if they had equal variance. The result was a p-value of 4.3*10-76, indicating that the two samples do not have equal variance.

Since the two samples did not have equal variance, we must use the version of a T-test that does not assume equal variance. The result of the T-test was a p-value of 4.3*10-117, meaning that we conclude that the means of the scaled ratings from the Vivino and professional data sets are different.

A Mann WhitneyU test was also performed since it makes no assumptions of the probability distributions. The result was a p-value of 7.4*10-97, also meaning that we conclude that the samples have different means.

I then looked at the regions of the Vivino data set that had more than 40 ratings total, which returned 40 different regions. These regions and their means are summarized in Table 3.

An ANOVA test was then performed on the ratings of each of the 40 most popularly rated regions. The resulting p-value was 4.6*10-278, indicating that there is a difference between the means of the groups. Since a significant result was obtained, a post hoc analysis was performed using Tukey's HSD test. The result of the post hoc analysis is summarized in Figure 3.

Region	Rating
Alentejano	4.0
Alsace	3.8
Barbera d'Alba	3.9
Barolo	4.1
Barossa Valley	3.9
Bolgheri	4.0
Brunello di Montalcino	4.2
California	3.8
Chianti Classico	3.9
Châteauneuf-du-Pape	4.2
Côtes-du-Rhône	3.7
Douro	4.0
Dão	3.8
Langhe	3.9
Marlborough	3.8
McLaren Vale	3.9
Mendoza	3.9
Montefalco	3.8
Napa Valley	4.3
Niagara Peninsula	3.7
Okanagan Valley	4.0
Paso Robles	4.0
Pauillac	4.4
Pessac-Léognan	4.3
Pomerol	4.4
Rioja	3.9
Russian River Valley	4.1
Saint-Julien	4.2
Saint-Émilion Grand Cru	4.2
Sancerre	4.0
Sonoma Coast	3.9
Sonoma County	3.8
South Australia	4.0
Stellenbosch	3.9
Südtirol - Alto Adige	4.0
Terre Siciliane	3.8
Toro	3.9
Toscana	4.0
Veneto	4.1
Willamette Valley	4.0

Table 3: Means of 40 Most Popularly Rated Regions

Conclusion: This paper brought to light some insight surrounding the distributions of wines rated by sommeliers and those rated by amateurs on Vivino. It was found that the average rating given by the amateurs was higher than that of the professionals.

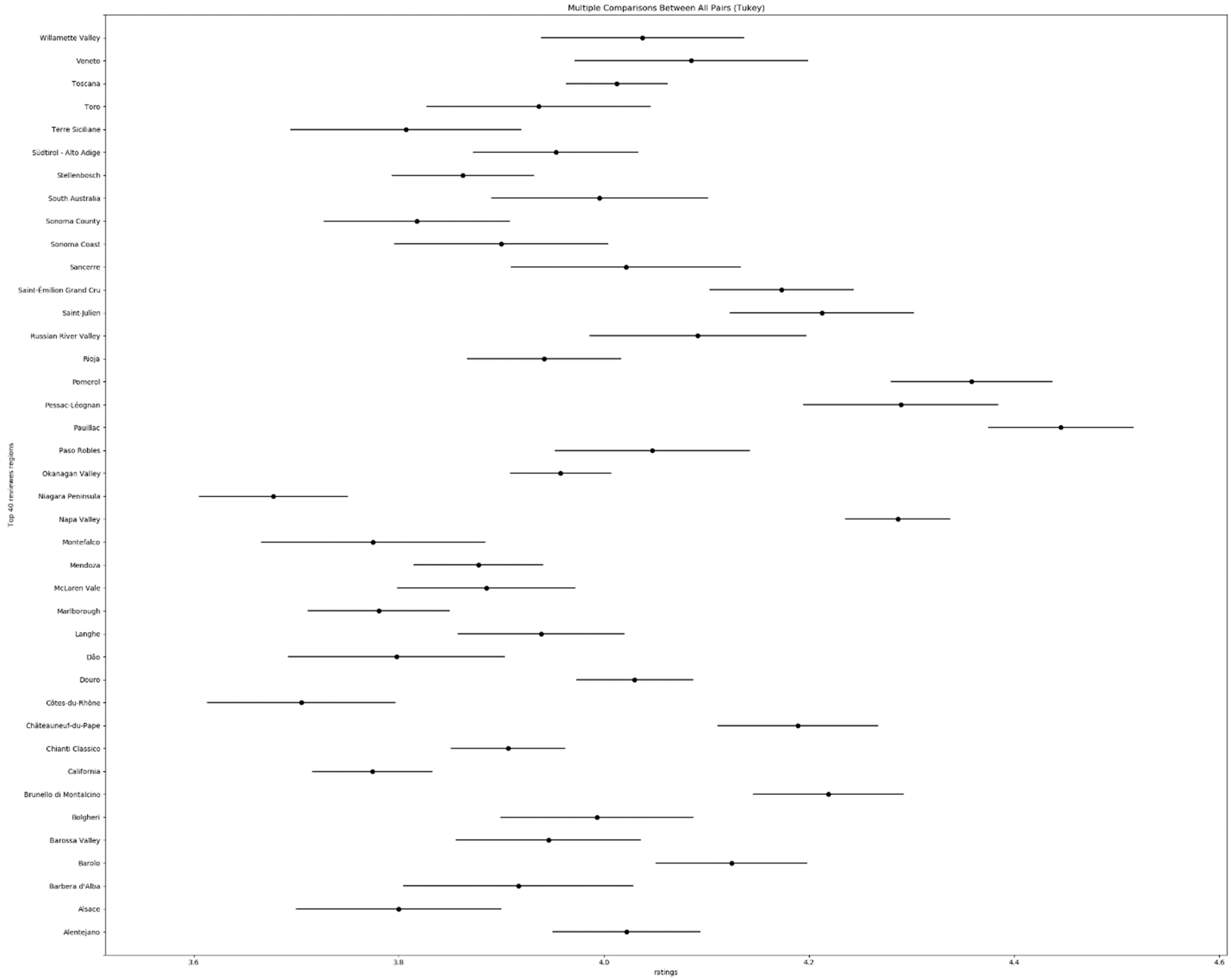


Figure 3: Tukey's HSD Test Post Hoc Analysis of Ratings for Each Region

Additionally, it was found that if you are looking for a bottle of wine, the best region to buy is from Pauillac. This region had a statistically higher average rating amongst Vivino users than nearly all other popularly rated regions. This makes sense since the Pauillac region is in Bordeaux, which is commonly believed to contain some of the best wines in the world. The region that was found to be worse than many other regions is the Niagra peninsula. This personally does not surprise me since Canada is only getting started with wine production and that region gets very cold in the winter, making it not ideal for many grape varietals. Some of the ratings that surprised me was California and Cotes-du-Rhones. Both California and Cotes-du-Rhones are big wine producers and are generally thought to be quite good. Also, of note, the Okanagan Valley performed well with an average of 4.0. This is in the same league as regions such as Sonoma and Rioja, both prestigious regions.

The Third Question ?

Can you tell which wine type it is if someone described the taste of wine? Or, if you want wine that has a specific taste, what type of wine should you get? In other words, if you want wine that has flavours of tannins, or apple, what type of wine will be satisfying? As it turns out, if you want wine that tastes of tannins, you want Bordeaux-style Red Blend or Cabernet Sauvignon. If you want the apple flavoured wine, Chardonnay is the one for you. A variety of wine refers to the type of grape used in making the wine and is referred to as a type of wine. There are over 10,070 types of grapes according to James the Wine Guy.{[1]} In this study, we'll use classification to predict the variety of wine from text reviews.

Data:

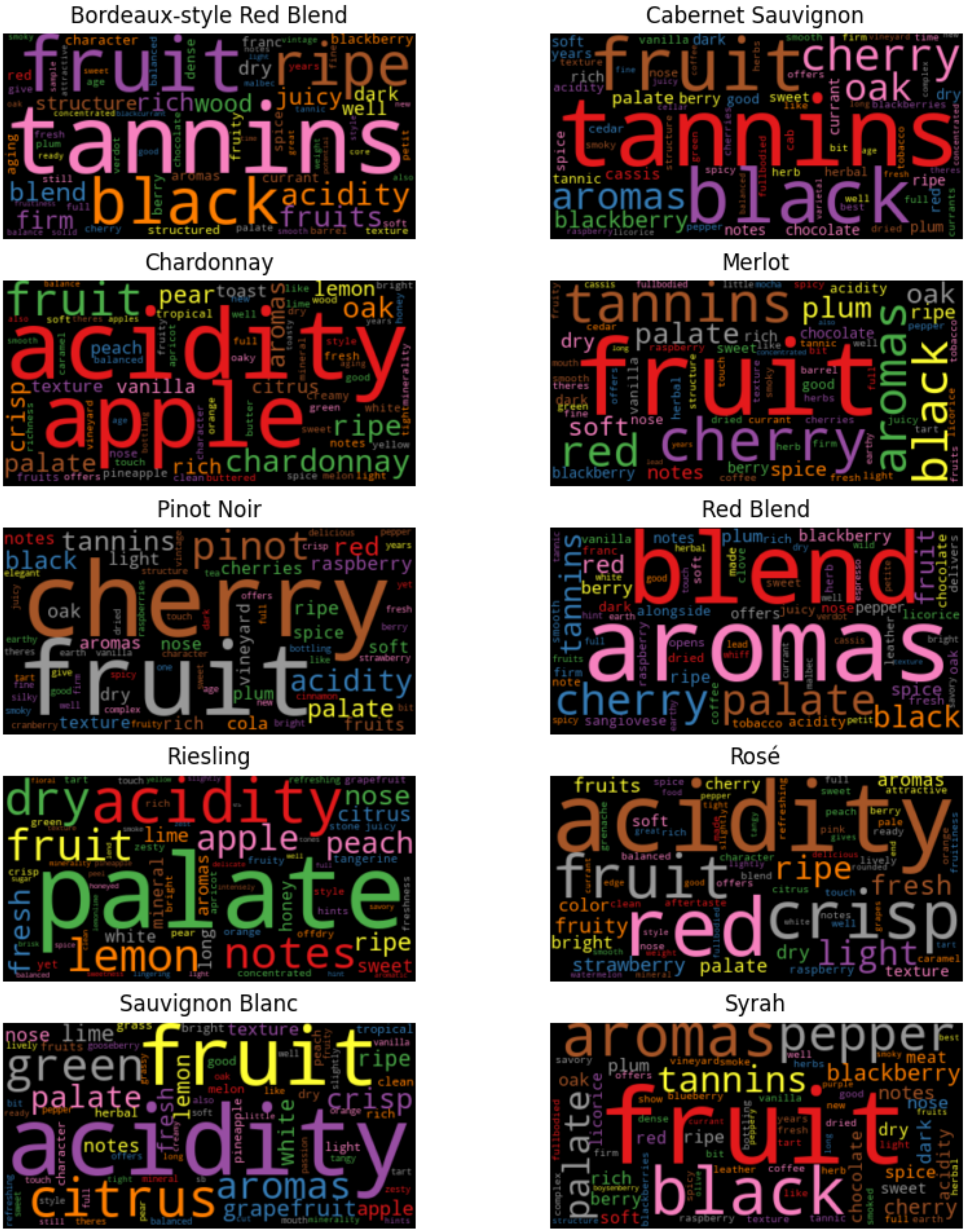
The dataset used in this study was obtained from www.kaggle.com.{[2]} The dataset contains 130,000 reviews on 707 different types of variety. The features included in the dataset are country, description, designation, points, price, province, region 1, region 2, taster name, taster twitter handle, title, variety, and winery. The description is the feature column for text reviews. For this study, however, variety and description are the main focus.

Cleaning the data:

Since the features interested in are variety and description, the rest of the features are removed. Punctuations and numbers are removed as well as some words that do not have much meanings for classification. Such words removed are wine, flavour, drink, and etc. Also, the reviewers tend to mention the variety of wine in their reviews. This might help with higher training score but could potentially have adverse effect on validation dataset. Therefore, the names of variety are removed from the description, e.g. Bordeaux-style Red Blend and Pinot Noir. As already mentioned, there are 707 varieties in the dataset. From those, the most popularly or frequently reviewed 10 varieties are chosen. The 10 most reviewed varieties are Bordeaux-style Red Blend, Cabernet Sauvignon, Chardonnay, Merlot, Pinot Noir, Red Blend, Riesling, Rosé, Sauvignon Blanc, and Syrah. All words in description are lower cased for ease of analysis.

Exploring the data: Before starting the classification, it would be useful to get to know the data in hand. Two visualization methods, word cloud and bar plots, were used to capture the most frequently appearing words for each variety in the description. Figure 1 shows the word cloud, in which the size of the words are related to the number of occurrences; the largest word appears the most frequently in the wine reviews in the dataset. Word cloud makes it easy to understand which keywords have the strongest interpretation of the variety. At a glance, it is easy to see that Bordeaux-style Red Blend and Cabernet Sauvignon are tannin tasting, whereas Chardonnay has strong acidity and tastes like apples. Even a non wine expert can easily see that Pinot Noir has a strong cherry taste. Word cloud provides essential information quickly. However, there are shortcomings with Word Cloud. It is impossible to know the actual number of occurrences for each keyword, or how frequent the largest keyword is from the second largest. Figure 2 is bar charts that show the number of occurrences for each keyword. The bar chart shows how frequent the most frequent keywords are in comparison to less frequent keywords. For the variety, Rosé, acidity is far ahead of the next frequent keyword, red, while for Merlot, fruit, cherry, and tannins have almost the same occurrence. It can be said that Rosé has a strong acidic taste and Merlot has mixture of fruity cherry, tannin taste.

Classification : The dataset was split into training and validation. To pre-process the text for classification, Bag of Words technique was used: a table or matrix with the count of occurrence of each word, where each word is converted to a numeric value. The word with the greater number of count has greater significance. This can potentially mislead the classification due to the stop words. Stop words are words that do not have meaningful information in the context. An example of stop words are i, me, my, myself, we, our, and etc.[3] Therefore, stop words are removed when vectorizing words. A number of classification methods were tried to make the best prediction, namely, multinomial Naive Bayes, Logistic Regression, support vector machine, random forest, neural network, and k-nearest neighbours. The logistic regression classifier resulted in the highest accuracy score on the validation dataset with 0.753. The accuracy scores of other classification methods are as follows: support vector machine = 0.739, neural network = 0.658, multinomial Naive Bayes = 0.699, k-nearest neighbors = 0.458, and random forest = 0.673. Although the validation accuracy score of support vector machine classifier comes close to that of the logistic regression classifier, the running time for the support vector machine classifier was considerably longer.



Bordeaux-style Red Blend

Cabernet Sauvignon

Chardonnay

Merlot

Pinot Noir

Red Blend

Riesling

Rosé

Sauvignon Blanc

Syrah

Figure 1. Word cloud for wine reviews by frequency

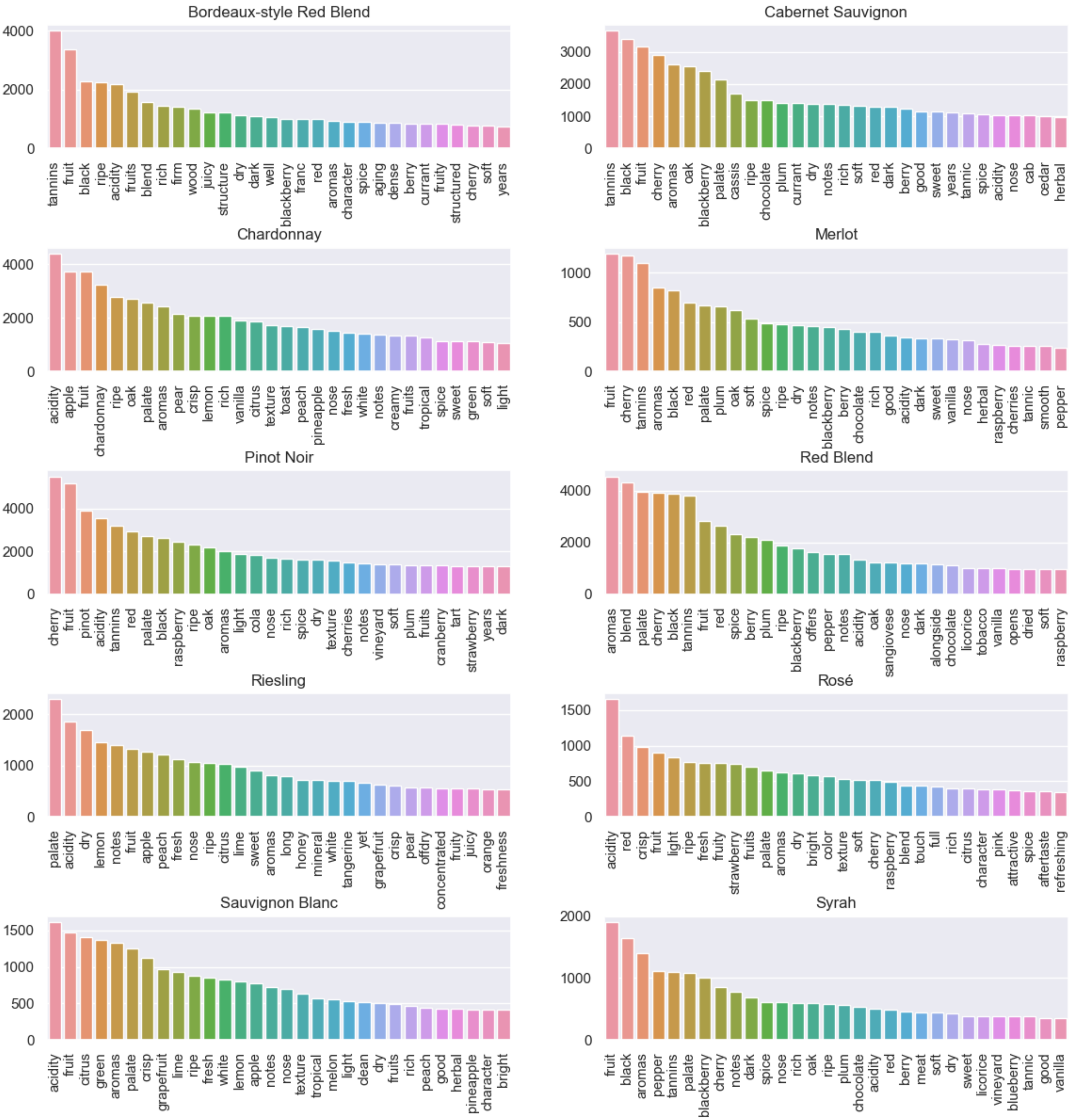


Figure 2. Bar plots for wine reviews by frequency

Conclusion:

In this study, wine reviews were classified to predict the variety of wine being reviewed. The dataset was explored to bring out the best keywords to describe a wine type. Word cloud provided a quick way to capture the relevant information whereas the bar chart was more suitable for detailed look at the data. With the validation accuracy score of 0.753, the logistic regression classifier performed the best of the classifiers tested in the study. It's worth noting that a binary classification method logistic regression outperformed multi-class classification methods such as k-nearest neighbors and random forest. A further study can include a look into whether keywords in wine reviews vary depending on the ratings given, or whether keywords in wine reviews affect the rating of the wine.

References:

[1] https://jamesthewineguy.wordpress.com/2018/11/12/how-many-wine-grape-varieties-exist-james-melendez/
[2] https://www.kaggle.com/zynicide/wine-reviews
[3] https://gist.github.com/sebleier/554280
[4] https://www.alcohol.org/guides/beer-wine-production-consumption-worldwide/
[5] https://www.vivino.com/
{6} https://pandas.pydata.org/pandas-docs/stable/user_guide/gotchas.html#support-for-integer-na

Project Experience

- Analyzed text wine reviews and performed classification of varieties of wine from the wine reviews.
- Scraped vivino.com to obtain wine data.
- Using Python and its libraries such as Numpy, Pandas, sklearn, nltk, wordcloud, matplotlib, and seaborn, cleaned the data first, which helped with exploring the data as well as the classification. Performed data exploration by visualization of the most frequently appearing keywords in the reviews using word cloud and bar chart. These visualizations highlight which keywords associate strongly with which varieties of wine. Performed classification to make prediction on varieties of wine from text reviews. This included utilizing a natural language processing library and machine learning library such as nltk and scikit-learn.
- Using a piece of code borrowed from online and added more functionality to it to scrape wine data. Used the combination of Beautiful Soup and Selenium in order to extract the necessary information.
- Created visualizations that represent keywords describing wine types. Obtained a satisfactory result from the classification. Gained knowledge on natural language processing and text classification.
- Gained insights and skills on web scraping. Learned the basics to use Beautiful Soup and Selenium.

- Kevin Park(301322108)

- Successfully cleaned two data sets to obtain data with the necessary parameters and no missing values.
- Produced scatter, box, and histogram plots using matplotlib to easily describe the data
- Analyzed the two datasets for normality and variance to inform which statistical tests could be conducted.
- Conducted appropriate statistical tests, such as the Student's T test, Mann Whitney U, and ANOVA, resulting in significant results.
- Performed post hoc analysis using Tukey's HSD test and graphically modelled the data obtained.

-Megan Fowler (301378374)

- Analyzed two different large data-sets and applied Machine Learning techniques to get to conclusions.
- Experienced challenges involved in cleaning and transforming the data, Missing values, outliers, etc. Converted categorical data into numerical data.
- Experienced how removing data can add bias to the results.
- Experienced Machine Learning tools can not perform well if we don't have the knowledge of data.
- Applied Statistics and Machine Learning techniques like Gaussian Naive Bayes, K-Nearest Neighbours, Gradient Boosting.
- Scraped vivono.com to collect data for my team-mates.
- Worked together to come up with project topic and questions. Interacted with each other on Zoom for meetings.

-Raman Sehmbi(301326893)