# Sentiment Analysis on Wine Reviews

Name: Qingyang Yu

Professor: Robert McDougal

# Contents

# Figures



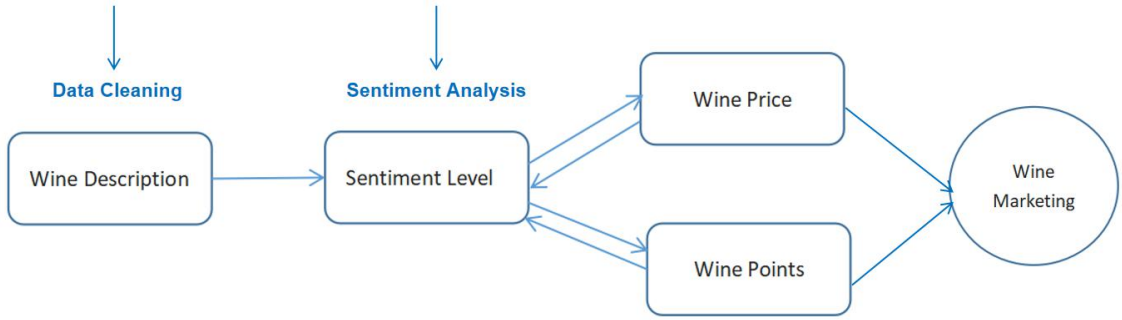
# References

# 1 Introduction

## 1.1 Background and Motivation

Sentiment analysis[1], uses natural language processing, text analysis, computational linguistics and biometrics to systematically identify, extract, quantify and study subjective emotions, which are usually classified into three categories: neutral, positive and negative.

In this study, I perform data analysis on different types of wine features. More importantly, I use sentiment analysis on wine description, to investigate whether description could reflect sentiment level of subjective emotions, and whether wine price, wine points are associated with sentiment level. In this case, although I have little background about wine tasting before, I could use predictive model to predict sentiment level based on wine text messages, which could be used in wine marketing.




## 1.2 Description on Dataset

The dataset was originated from WineEnthusiast website[2] during the week of Jun.15th, 2017, which contains 130,000 wine reviews with 14 columns:

1. Unnamed:0: Wine identifier.
2. country: The country that the wine is from.
3. description: Detailed description about wine.
4. designation: The vineyard within the winery where the grapes that made the wine are from.
5. points: The number of points. WineEnthusiast rated the wine on a scale of 1-100.
6. price: The cost for a bottle of the wine.
7. province: The province or state that the wine is from.
8. region_1: The wine growing area in a province or state.
9. region_2: Sometimes there are more specific regions specified within a wine growing area.
10. taster_name: Name of the taster.
11. taster_twitter_handle: Twitter handle of the wine taster.
12. title: The title of the wine review, often contains the vintage.

13. variety: The type of grapes used to make the wine.
14. winery: The winery that made the wine.

| | Variable | Data type | Description |
|---|---|---|---|
| 1 | Unnamed:0: | Numerical | Wine identifier. |
| 2 | country | Categorical | The country that the wine is from. |
| 3 | description | Descriptional | Detailed description about wine. |
| 4 | designation | Categorical | The vineyard within the winery where the grapes that made the wine are from. |
| 5 | points | Numerical | The number of points. WineEnthusiast rated the wine on a scale of 1-100. |
| 6 | price | Numerical | The cost for a bottle of the wine. |
| 7 | province | Categorical | The province or state that the wine is from. |
| 8 | region_1 | Descriptional | The wine growing area in a province or state. |
| 9 | region_2 | Descriptional | Sometimes there are more specific regions specified within a wine growing ar... |
| 10 | taster_name | Categorical | Name of the taster. |
| 11 | taster_twitter_handle | Categorical | Twitter handle of the wine taster. |
| 12 | title | Descriptional | The title of the wine review, which often contains the vintage. |
| 13 | variety | Categorical | The type of grapes used to make the wine. |
| 14 | winery | Categorical | The winery that made the wine. |

Fig.1 Data description

## 1.3 Data Acquisition(method, license)

The data was downloaded as **a csv file at kaggle public data repositories**[3] on Dec. 05th. The data itself does not contain metadata. However, the well-annotated metadata is accessible on kaggle website.

The license is **Attribution-NonCommercial-ShareAlike 4.0 International(CC BY-NC-SA 4.0)**, stating that I am free to:

- share(copy and redistribute the material in any medium or format);
- adapt(remix, transform, and build upon the material).

# 2 Data cleaning

## 2.1 Missing values

First, I extract None and NaN values in all columns, and plot the null values rate per column using seaborn heatmap and barplot.

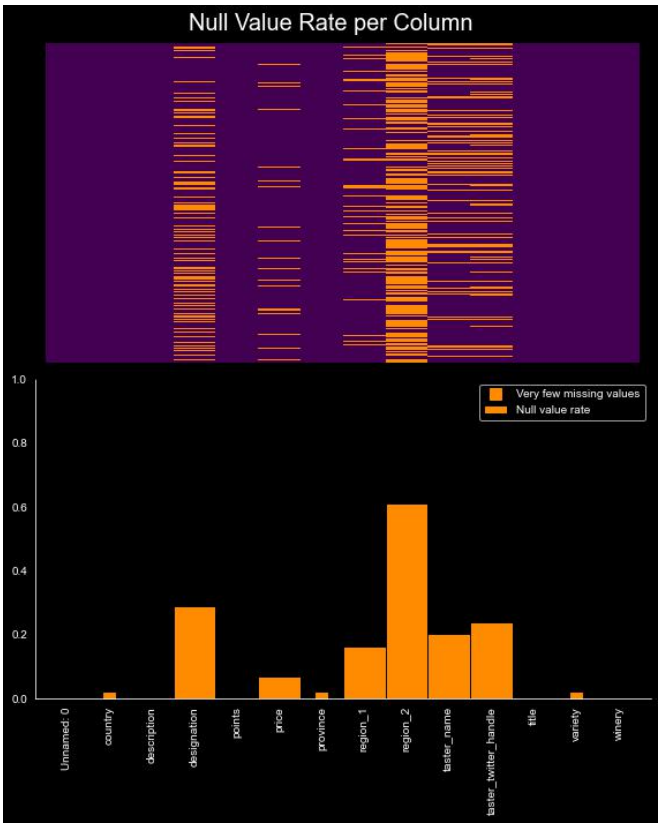



Fig.2 Null value rate per column

I use five strategies to deal with null values:

- For those columns that contain very few missing values(i.e., country, province and variety), drop the rows that include null values.
- For columns that contain many null values and not contain too much important information(i.e., Unnamed:0 and region_2), simply drop the columns.
- For columns that contain many null values but the most common value in these columns are not representative(i.e., designation, region_1, only occur in 1%), instead of replacing missing values with most common value of the column, I replace null values with 'Unknown'.
- For columns that have similar meanings(i.e., taster_name and taster_twitter_handle), first count to see how often a taster name is given but a twitter handle is not, and vice versa. According to Figure 3, taster name is more comprehensive, taster and taster twitter handle are one-to-one relationship(there is no taster that use multiple twitter handles), thus simply drop the column taster_twitter_handle. And replace missing values in column taster_name with Unknown.
- For numeric column(i.e., price), replace null values with median of the price.

```
Rows that contains a taster name but no twitter handle: 4969
Rows that contains a twitter handle but no taster name: 0
```

Fig.3 Rows that only contain taster name or taster twitter handle

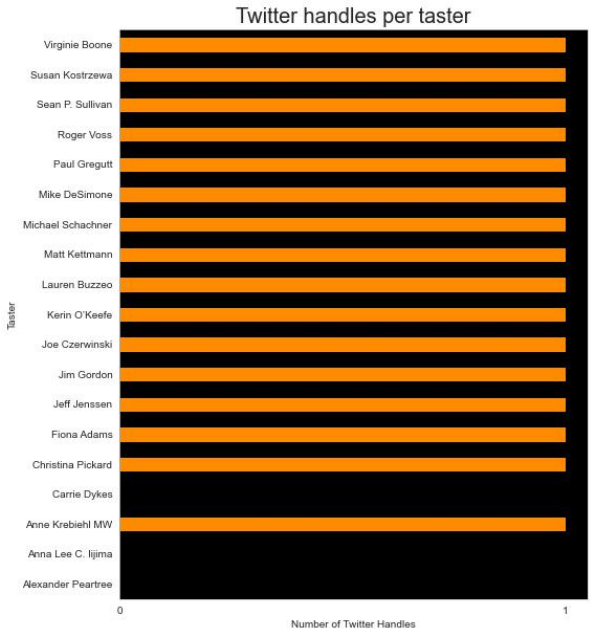



Fig.4 One-to-one relationship between taster and taster twitter handle

## 2.2 Duplicates

Check if the data contains any duplicates. Find duplicates that are same in column description and title, and remove all the duplicates.

## 2.3 Text mining

For text data, I use natural language processing to clean the wine description.
- Turn words into lowercase, remove punctuation and irrelevant terms.
- Use regular expression to tokenize sentences into list of words.
- Lemmatize each token(e.g. walk, walked, walking -> walk).
- To visualize the difference brought by text mining, I summarize the 100 most common words before and after text mining process.
- Analyze on N-gram, when n = 2[4].

```
4937774 words total, with a vocabulary size of 30452

[('and', 321309),
 ('the', 204668),
 ('a', 166325),
 ('of', 159396),
 ('with', 111439),
 ('this', 104965),
 ('is', 89535),
 ('it', 79887),
 ('wine', 74246),
 ('flavors', 58016),
 ('in', 57888),
 ('to', 51898),
 ('s', 49361),
 ('fruit', 46032),
 ('on', 41934),
 ('that', 37566),
 ('aromas', 36229),
 ('palate', 35107),
 ('acidity', 32127),
```

```
3010589 words total, with a vocabulary size of 21848

[('wine', 76823),
 ('flavor', 65592),
 ('fruit', 58926),
 ('aroma', 37541),
 ('finish', 37167),
 ('acid', 36618),
 ('palat', 35220),
 ('drink', 31234),
 ('cherri', 30909),
 ('tannin', 30397),
 ('ripe', 26752),
 ('black', 26666),
 ('dri', 24411),
 ('note', 23161),
 ('spice', 21709),
 ('rich', 20454),
 ('red', 20196),
 ('fresh', 20132),
 ('berri', 17826),
 ('show', 17205),
```

Fig.5 Changes of 100 most common words due to text mining

# 3 Data Analysis

## 3.1 Summary statistics

**Research Question**

**- Is there any ways summary statistics might be misleading?**
**- Any solutions?**

**Results**

Based on characteristic of each column, I use different ways to visualize the result.

- Use bar charts to visualize categorical data/numerical data(i.e., province, country, variety, winery).
- Use histogram to visualize wine price.
- Since **histogram of price is skewed by outliers:** Some of the wine price is over 300, and the chart would grow to include these outliers, thus ruin the rest of data being shown. I **exclude wine price that over 300.**

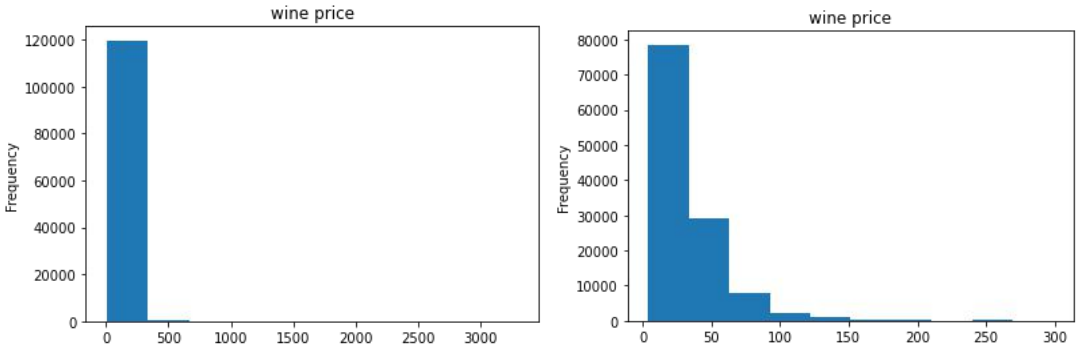



Fig.6 Wine counts distribution according to price before and after adjustment

## 3.2 Univariate analysis

**Research Question**

**- What is the wine province distribution(relative proportions)?**
**- What is the wine country distribution(relative proportions)?**
**- What is the most reviewed type of grape?**
**- Is there any winery the major supplier of wines?**
**- What is the pattern in terms of wine points or price?**
**- Who reviewed most number of wines?**

**Results**

Plot results are from above summary statistics.

We could see that from province level, **California produces almost a third of wines** reviewed in WineEnthusiast.

From country level, **US produces over 40% of wines** reviewed in WineEnthusiast,

followed by France, Italy and Spain.

In terms of variety(type of grape), **Pinot Noir is the most reviewed wine variety, and produces over 10% of wines**, followed by Chardonnay, Cabernet Sauvignon, Red Blend, and Bordeaux-style Red Blend.

Since the most number of wine produced by a single winery is only 200, I suppose wines come from different wineries, and **there is no winery acts as the major supplier of wine.**

In terms of wine points, wine points lie between 80 and 100, and **most wines are scored 87, 88 or 90.** Wine points seem very similar with a normal distribution.

**Most wines are not too expensive**, most of the wine price **lie between 5 and 100**.

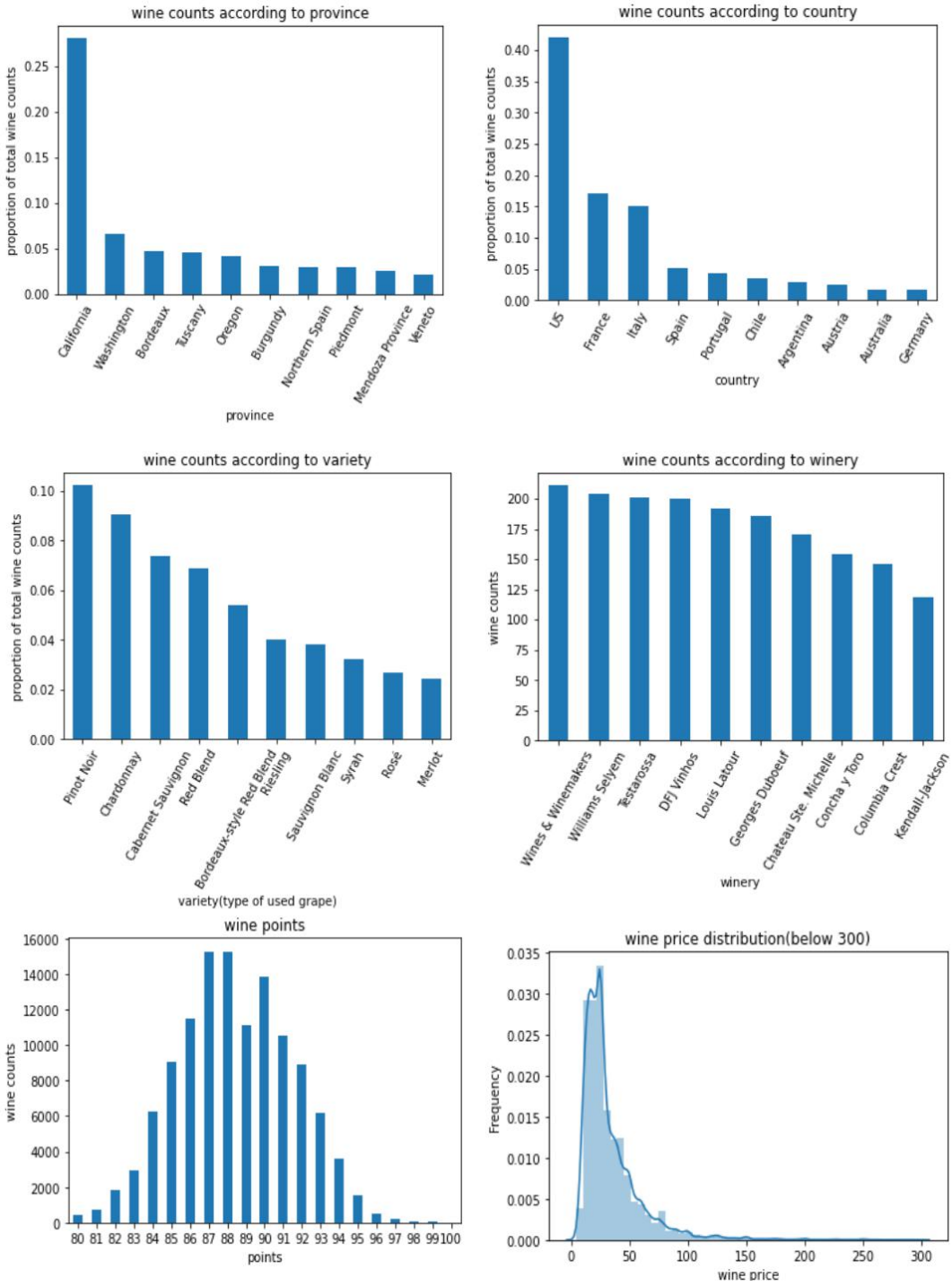


Fig.7 Summary statistics of wine features

I also plot wine points into six categories: category1: point(80,83); category2: point(83,87); category3: point(87,90); category4: point(90,94); category5: point(94,98); category6: point(98,100).
Find that most wines lie in categories 3 and 4, which score between 87 and 94.

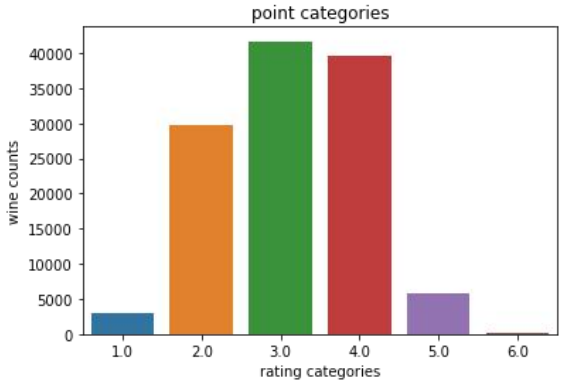



Fig.8 Wine points into categories

Apart from unknown tasters, **Roger Voss has reviewed the most number of wines**, it is 10000 more than the second person Michael Schachner.

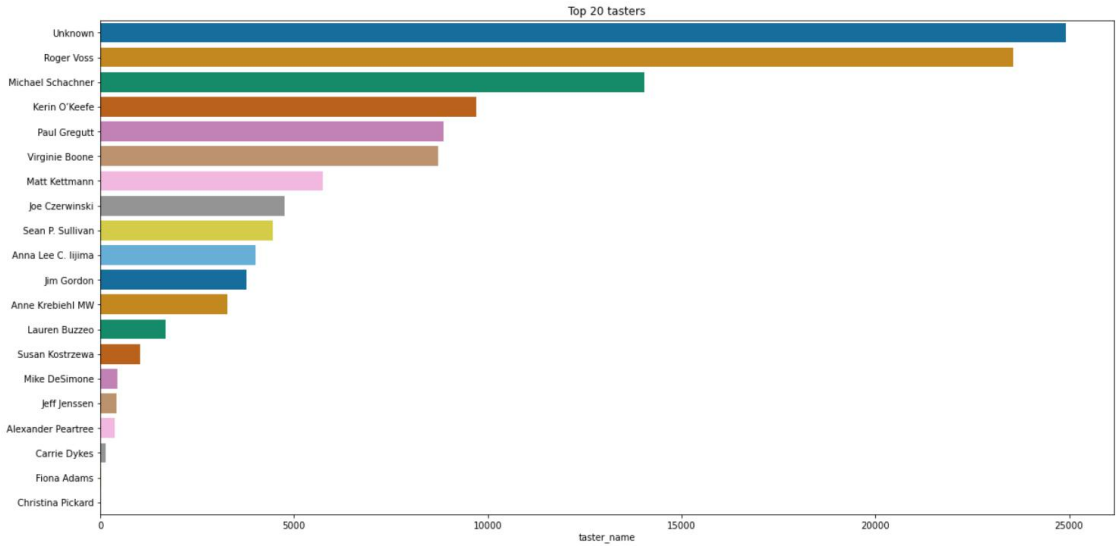



Fig.9 Most frequent tasters

## 3.3 Bivariate analysis

### Research Question

**- Which variety of grape would be best to make wine?**
**- Is there any relationship between wine points and price?**

### Results

I find the wine varieties that are highest rated or cheapest, and merge two groups on

wine variety.

Results indicate that **Merlot, Cabernet Sauvignon, Syrah, Portuguese Red, Chardonnay are best to make wine. They all earn 100 points with price for 4.**

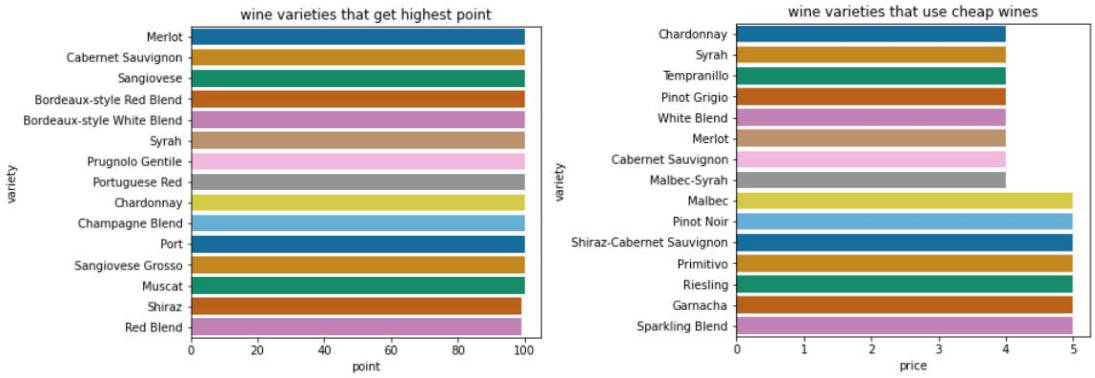



Fig.10 Wine varieties that are highest rated or cheapest

To see the relationship between wine points and price, first I use hex plot, and exclude wine price that is over 200.

For wine price that is less than 200(majority), most wines cluster around 87.5 points and price around 15.

For further investigation, I use scatterplot.

From the scatterplot, wine with highest price only got a moderate points(~87). However, the highest rated wine is not the most expensive ones. Overall, **there is a weak linear relationship between wine points and price, but there are exceptions.**

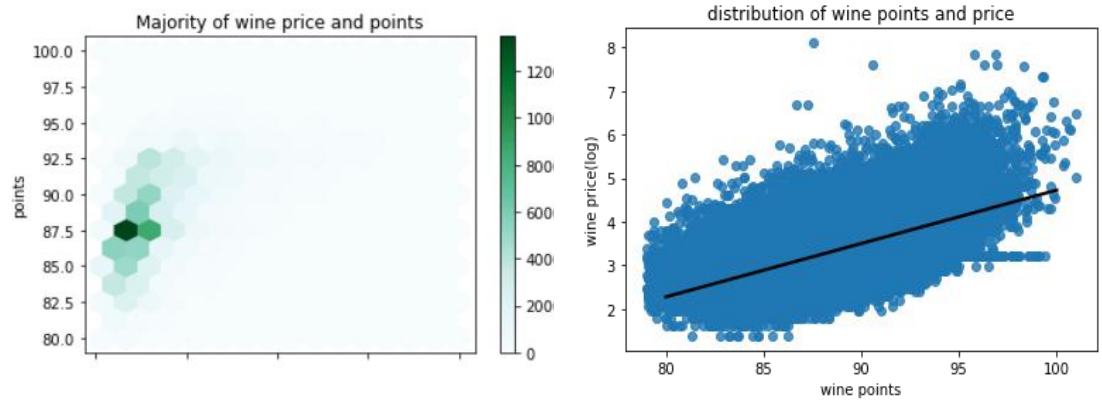



Fig.11 Hexplot and scatterplot of wine points and price

## 3.4 Multivariate analysis

### Research Question

**- Is there any relationship among all wine features? Is the relationship weak/moderate/strong?**

### Results

I use heatmap to visualize the correlation among all wine features.
Although there is a relationship between wine price and points, which is a positive correlation, the correlation is rather weak(~0.4). There is also **a stronger correlation between length of description and points(~0.51).**

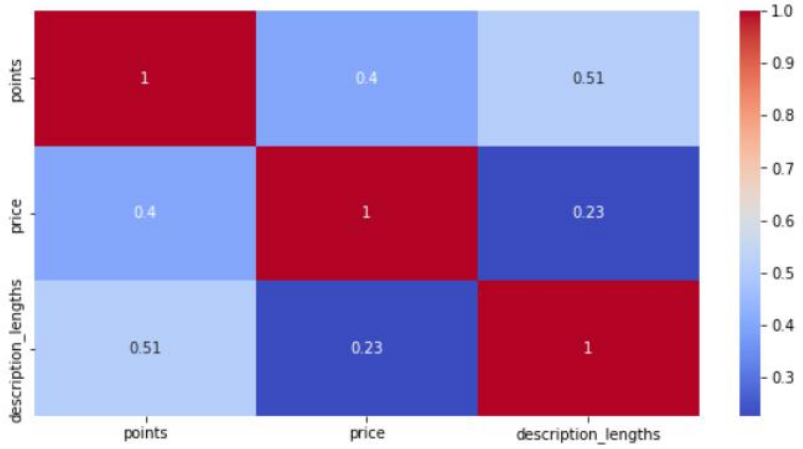



Fig.12 Heatmap of all wine features

## 3.5 Word frequency analysis

**Research Question**

**- What topics taster mention most when they taste wine?**

**- Is topics mentioned by taster differed by points? Is topics mentioned by taster differed by price?**

**- What are some common terms appeared in the lowest rated and highest rated wine?**

**- What should a wine company keep in mind to get good reviews?**

**Results**

I performed word frequency analysis on wine description, title, lowest rated wine, highest rated wine, expensive wine and cheap wine. Stopwords are common wine-related words, such as 'fruit', 'Drink', 'wine' and 'drink'.

The most frequently used words in wine description are: black cherri, full bodi, rich, cabernet sauvignon, sweet, fresh, medium bodi, dri and black currant. **These are either main ingredients of wine or standards on wine appraisal.**

The most frequently used words in wine title are: Pinot Noir, Cabernet Sauvignon, Valley WA, Napa Valley, Columbia Valley, Sauvignon Blanc. **These are either type of grapes or origins of high quality wines.**

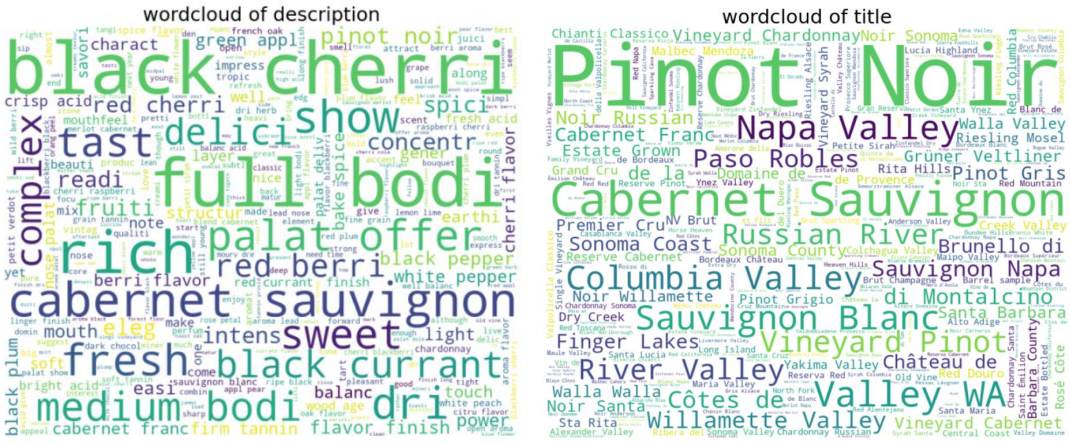



Fig 13. Wordcloud of wine description and title

The most frequently used words in lowest rated wine are: flavor, finish, palat, tast, aroma, sweet, bitter, smell, acid. These are mostly about negative palate or taste.
The most frequently used words in highest rated wine are: flavor, structur, rich, tanninacid, complex, power, ripe, great, age. These are mostly about positive taste.
**Common terms appeared in the both lowest and highest rated wines are flavor, and adjectives used to describe palate or taste.**
**For a wine to get good scores, aging, concentration and aroma play an important role.**


Fig 14. Wordcloud of lowest and highest rated wine

The most frequently used words in cheap wine are: flavor, aroma, finish, palat, fresh, sweet, note, dri, soft, fruiti. These are mostly about palate or taste.
The most frequently used words in expensive wine are: flavor, tannin, rich, acid, ripe, age, structur, concentr, finish. These are mostly about positive palate or taste.
**Expensive wines are more associated with aging. This might be a important factor for customers that prefer expensive wines.**


Fig 15. Wordcloud of expensive and cheap wine

## 3.6 N-gram Analysis

### Research Question

**- What are some country-wise characteristics of wine?**

### Results

N-gram is a contiguous sequence of n items from a given sample of text or speech[4]. N-grams are typically collected from text.

First I select 10 countries that have most wine reviews, split their wine description into 2-grams, refers to 2 words element, and summarize the frequency of each 2 words.

According to Fig 16., **each country has its own characteristic and description of wine, indicates different wine drinking style, different wine tastes and wine ingredients.** For instance, US wine tend to have fruit flavors, main ingredients include black cherry or black pepper. France wine are ready drink, also have fruit flavors, main ingredients include black currant and red fruits. Italy wine have palate offers, main ingredients include black cherry, stone fruit and white pepper.

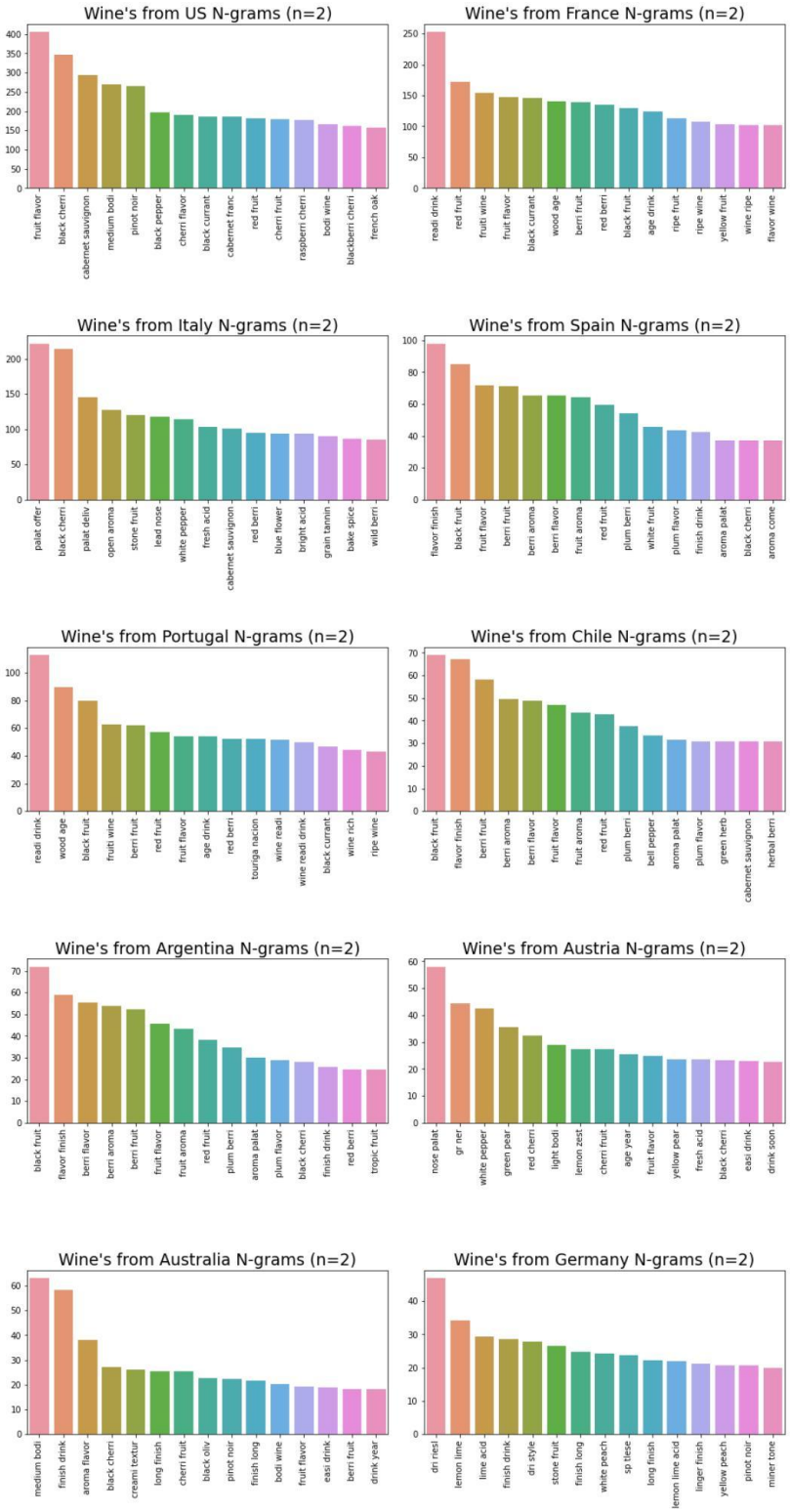



Fig 16. 2-gram analysis according to countries

## 3.7 Sentiment Analysis

### Research Question

**- Is sentiment level related to wine points or price?**

### Results

Finally I perform sentiment analysis.

First I create the sentiment intensity analyze model. Based on description after text mining, I give polarity score(total score), neutral score, negative score, positive score and overall sentiment level to each wine variable.

I use box plot to visualize the relationship between sentiment level and wine points or wine price.

From the Figure 17.,, we could see that **sentiment level does not have much effect on the wine price, while sentiment level really affects the wine points: If wine description are predicted as positive sentiment level, it is more likely to have a higher wine point.**

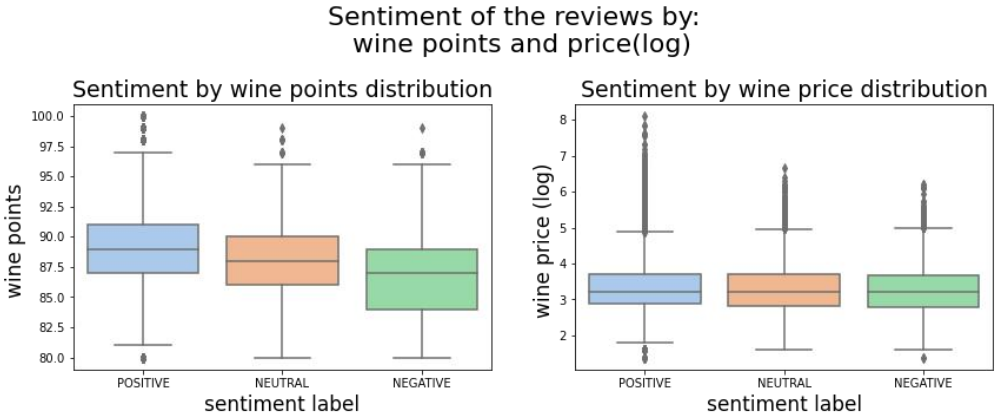



Fig 17. Distribution of sentiment level by wine points or price

# 4 Web Interface

## 4.1 API Server

The framework of API server is showed in the Fig 18. You could access the web page through http://localhost:10010/.

First I create a flack application. Then I define ten routes which could render each corresponding page from html file. Noticed that route number is designed to show fixed number of rows within wine data, and route Analysis is designed to do sentiment analysis, by which user could enter wine description displayed in the table, and receive sentiment score results.

In the main function, it run the application(app.py).

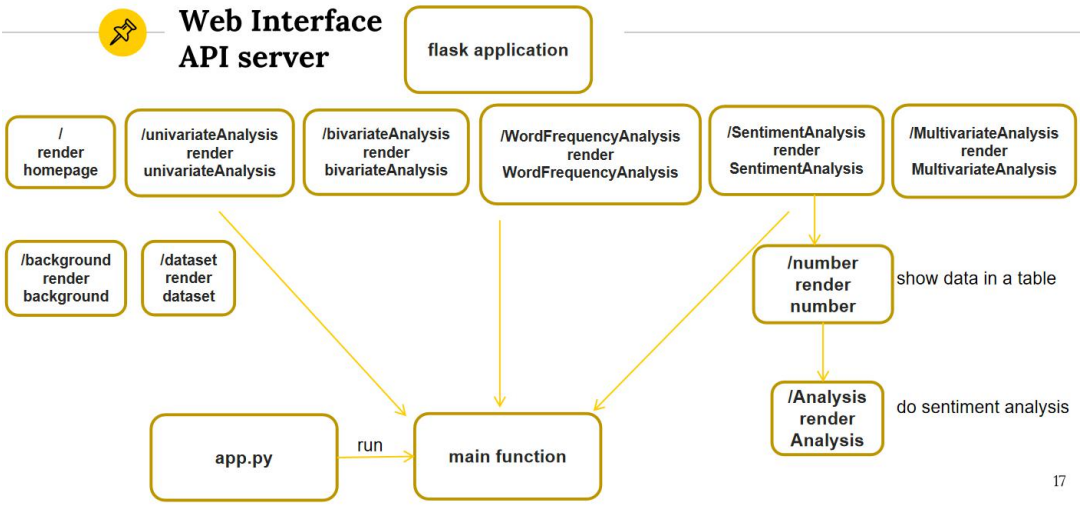



Fig 18. Framework of API server

## 4.2 Web Front-end

The web front-end has eight web pages in total. User could access each page by clicking the above menu.

- The Homepage: Display the project title.
- The Background page: Display the background and motivation: What is sentiment analysis? Why do I do this project using sentiment analysis? User could get access to the wikipedia page of sentiment analysis by clicking on the link.
- The Dataset page: Display the data source, detailed information about dataset and data license. User could get access to the WineEnthusiast(where the dataset is originated) or kaggle website(where I downloaded the dataset) by clicking on the links.
- The Univariate Analysis page: Display the figures and findings of univariate analysis. User could press selection button, and choose the variable that is interested to see the result.
- The Bivariate Analysis page: Display the figures and findings of bivariate analysis. User could press selection button, and choose the relationship between two variables that is interested to see the result.
- The Word Frequency Analysis page: Display the figures and findings of word frequency analysis.
- The Sentiment Analysis page contains a prediction model. User could click the selection button to choose number of rows displayed in this page(n=10,20, 50, 100), and copy the wine description from displayed data, enter text into the below box to perform sentiment analysis. Results include negative score, positive score and neutral score of the text.
- The Multivariate Analysis page: Display the figures and findings of multivariate analysis.

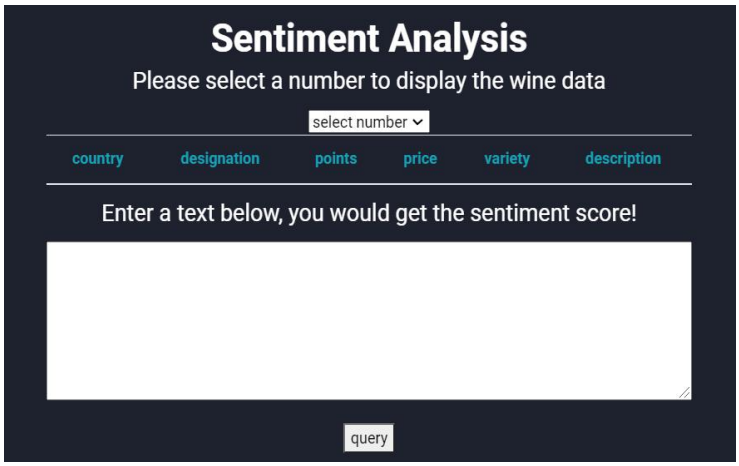


Fig 19. Web page of sentiment analysis

# 5 Discussion

## 5.1 Surprising results

- For a wine to get good scores, aging, concentration and aroma play an important role.
- Expensive wines are more associated with aging. This might be a important factor for customers that prefer expensive wines.
- Compared with the correlation between wine price and points, length of description and wine points has a stronger correlation: If wine description is longer, wine points are higher.
- Each country has its own characteristic and description of wine, indicates different wine drinking style, different wine tastes and wine ingredients, which is very useful for wine marketing. Wine companies need to adjust their product's flavor in different countries.
- Sentiment level does not have much effect on the wine price. Sentiment level really affects the wine points: If wine description are predicted as positive sentiment level, it is more likely to have a higher wine point, vice versa.

## 5.2 Difficulties

- Cleaning the wine description is a lot of work.
- API implementation: it really take times to perform sentiment analysis, I have to minimize the sampled wine reviews used to built the sentiment model.

# References

[1] https://en.wikipedia.org/wiki/Sentiment_analysis
[2] https://www.winemag.com/?s=&drink_type=wine
[3] https://www.kaggle.com/zynicide/wine-reviews
[4] https://en.wikipedia.org/wiki/N-gram