TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Contents

INTRODUCTION

This dataset named ‘wine.csv’ details wine review data, including 'country', 'description', 'designation', 'points', 'price', 'province', 'region_1', 'region_2', 'variety', 'winery' were summarised in the dataset. A mixed data types are used, including int64, object, and float64.

Main purposes of this capstone project IV include:

- Import and inspect the dataset
- Data cleaning
- Missing data management
- Data analysis and visualisation

DATASET INSPECTION

There are 1103 wine records and 11 features/columns in this dataset. A total of 22 countries who produce wine are included in this dataset, such as US,Spain,France,Italy etc..There are 124 types of wine were listed, such as Cabernet Sauvignon,Tinta de Toro,Sauvignon Blanc,Pinot Noir, etc..

Here are some explanation of variables appeared in this dataset

- country: The country that the wine is from
- description: A few sentences from a sommelier describing the wine’s taste, smell, look, feel, etc.
- designation: The vineyard within the winery where the grapes that made the wine are from
- points: The number of points WineEnthusiast rated the wine on a scale of 1–100 (though they say they only post reviews for wines that score >=80)
- price: The cost for a bottle of the wine
- province: The province or state that the wine is from
- region_1: The wine growing area in a province or state (ie Napa)
- region_2: Sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank
- variety: The type of grapes used to make the wine (ie Pinot Noir)
- winery: The winery that made the wine

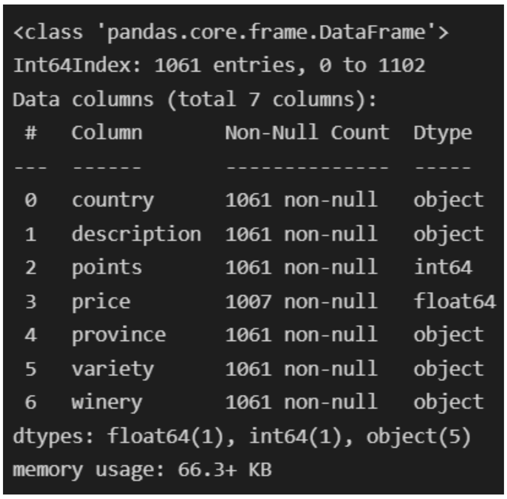
DATA PREPARATION

DATA CLEANING

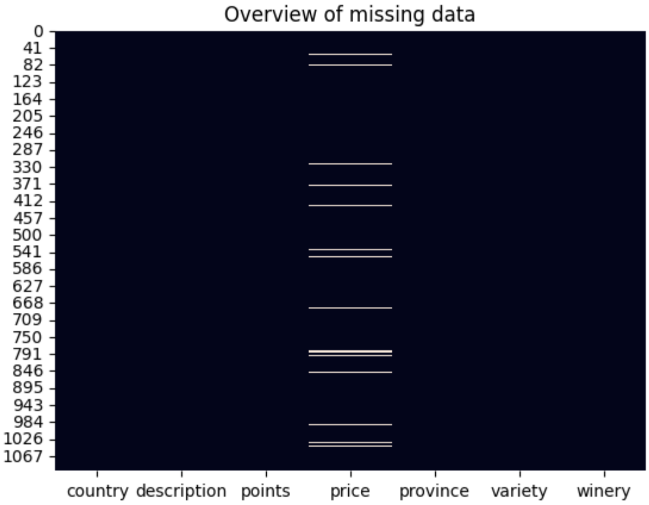
The following steps were carried out for ‘data cleaning’:

1. Identify columns that are redundant or unnecessary
   Based on the data set, it seems the following columns are not needed for the analysis and are removed from the data set: ['Unnamed: 0', 'designation', 'region_1', and 'region_2'].
2. Remove duplicate rows
   Duplicate rows were removed using .drop_duplicates(), if any.

MISSING DATA

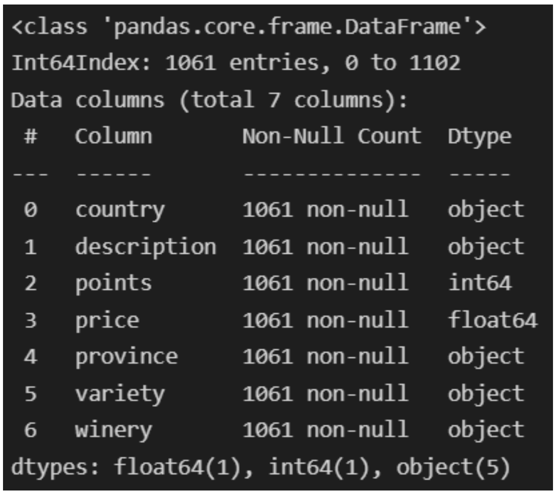


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1061 entries, 0 to 1102
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   country      1061 non-null   object
 1   description  1061 non-null   object
 2   points       1061 non-null   int64
 3   price        1007 non-null   float64
 4   province     1061 non-null   object
 5   variety      1061 non-null   object
 6   winery       1061 non-null   object
dtypes: float64(1), int64(1), object(5)
memory usage: 66.3+ KB
```



From above results, it seems that only small proportion of ‘price’ values are missed. These were filled in with average value of the 'price' column.

After all data preparation procedures, there are a total of 7 columns/features left and all Null values were filled. Detailed information is presented below:



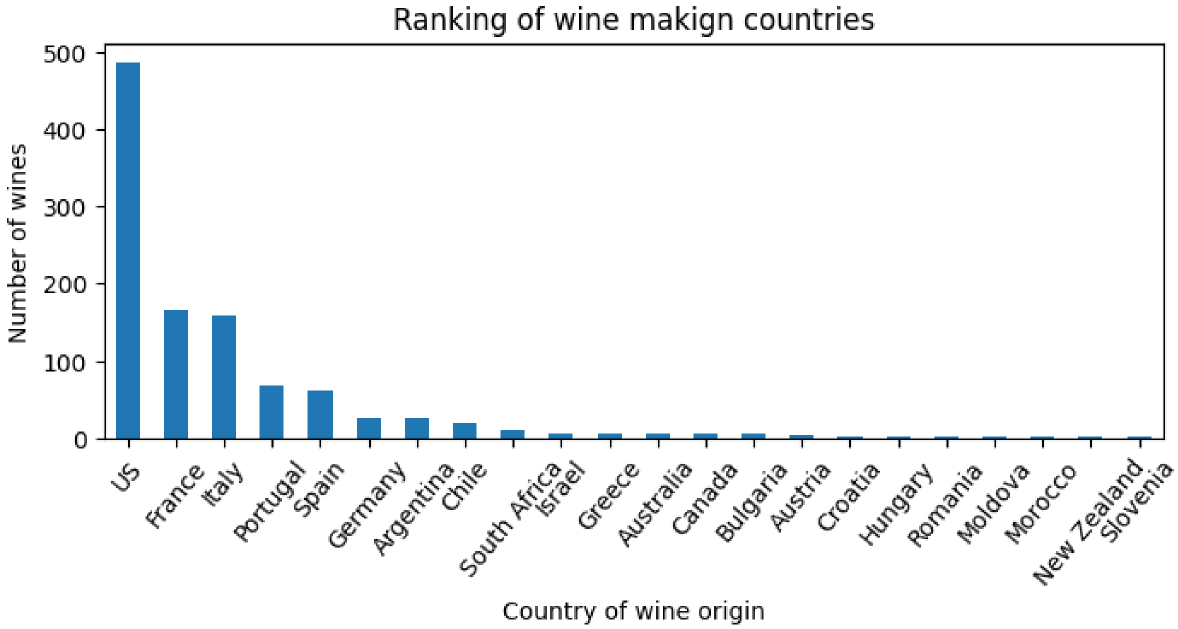
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1061 entries, 0 to 1102
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   country      1061 non-null   object
 1   description  1061 non-null   object
 2   points       1061 non-null   int64
 3   price        1061 non-null   float64
 4   province     1061 non-null   object
 5   variety      1061 non-null   object
 6   winery       1061 non-null   object
dtypes: float64(1), int64(1), object(5)
```

DATA STORIES AND VISUALISATIONS

Data visualisation and story-telling were carried out by identify of relationships between variables/features.

1. Find out which nations are top wine making nations in the world.

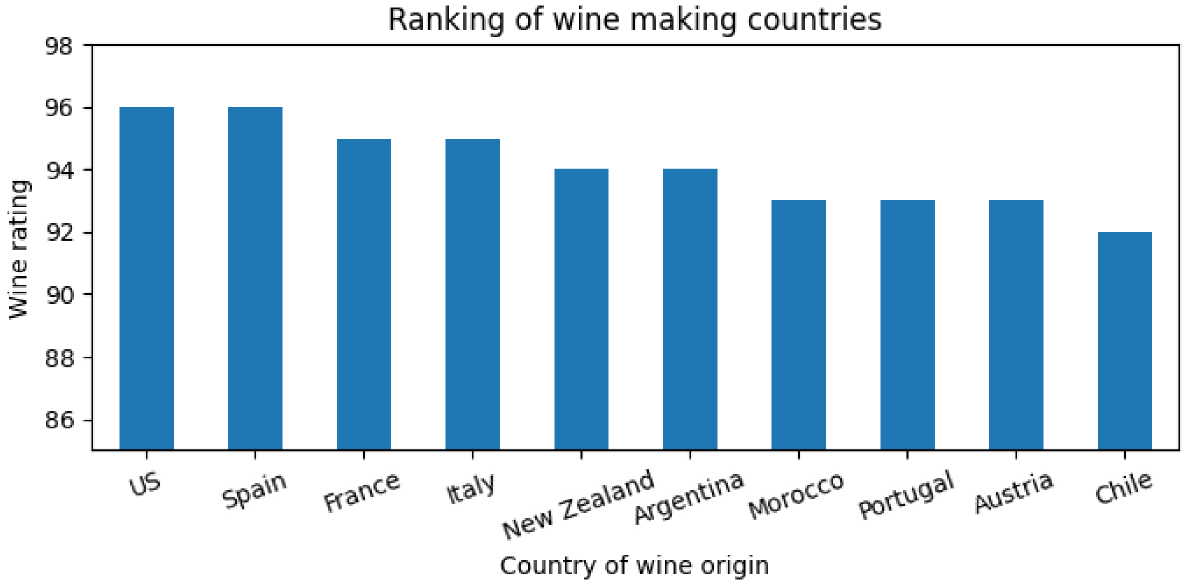
   This was analysed using number of wines produced by each country and a bar graph was plotted.



From above results, it is apparent that US the top one and has nearly 500 types of wine in the wine review dataset. This is over 2 times higher than the next one in the rank: France, thought France has always been known as famous for its wine. Itally has slightly less number of wines than France and is being ranked the 3rd.

2. Which country has top quality wines from the dataset?

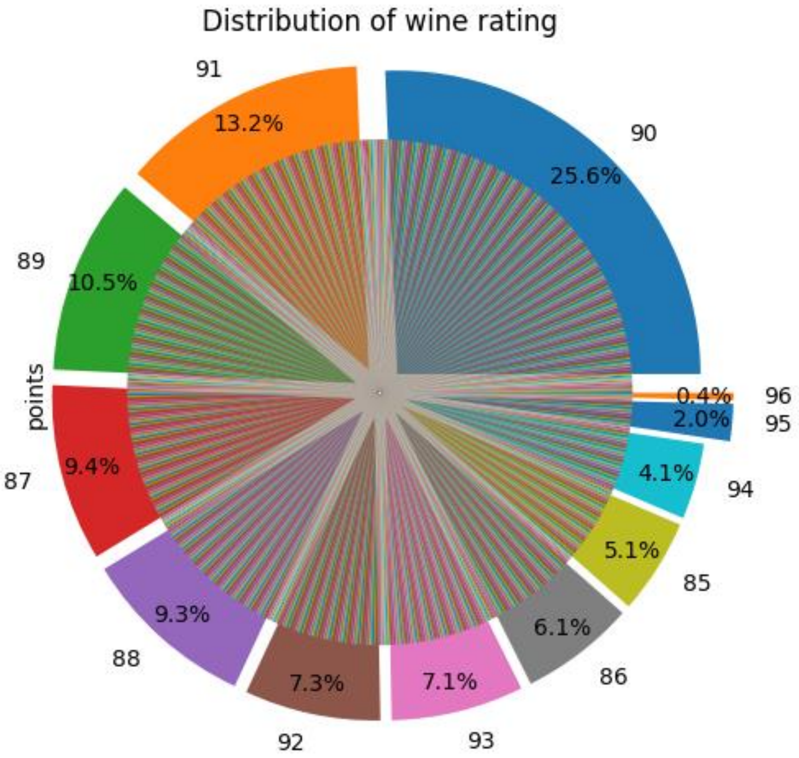
   This was analysed using highest point/rating feature to rank quality of wines produced by each country and a bar graph was plotted for the top 10 countries:



   From results above, all top 10 countries have rating('points') ≥ 92. Among those, both US and Spain have equal rating of 96 and are rated the 1st. France and Italy has wine rating of 95 and are being ranked the 2nd.

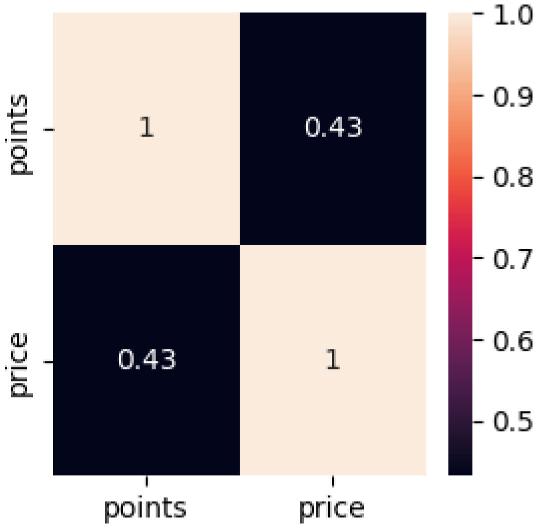
3. What is the distribution of wine rating look like for entire dataset? This will also help interpret how good quality the wine is for the top 10 rated countries.

   This was analysed using point/rating feature to rank quality of wines that are listed in the dataset and a pie graph was plotted:



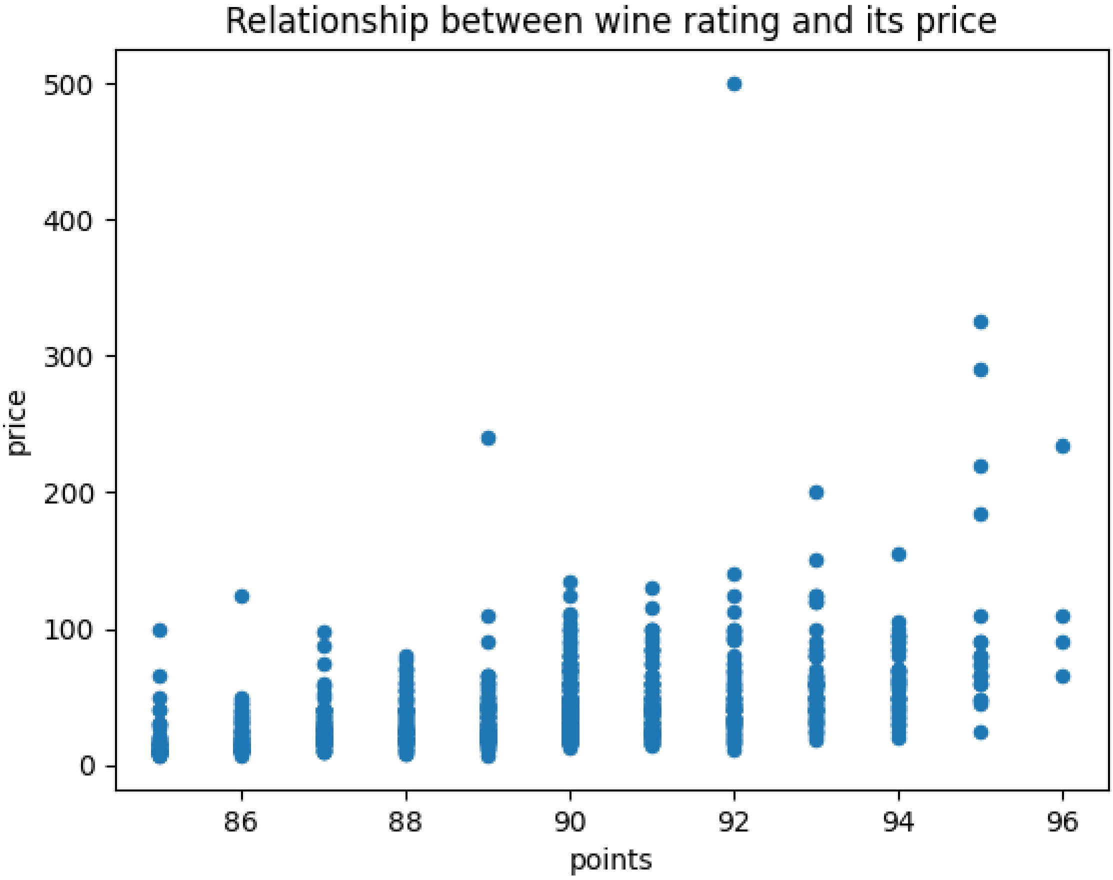
From results above, about 25.6% of wines in the dataset are ranked at 90 and this is the largest proportion for all the rating. Over 50% of wines are ranked under 90. Only 0.4% are ranked 96 and 2.0% ranked 95. Therefore, the top 10 ranking wine countries in the dateset can produce very good quality wines.

4. Please check whether wine ranking is related with price and determine whether you should spend lots of money to buy high quality wines.
   1) A heatmap was created to compare correlation between wine raking('points') and price.



It can be seen that price and points are positively correlated with each other, however, the correlation does not seem to be strong and only have a value of 0.43 with 1.0 being the maximum.

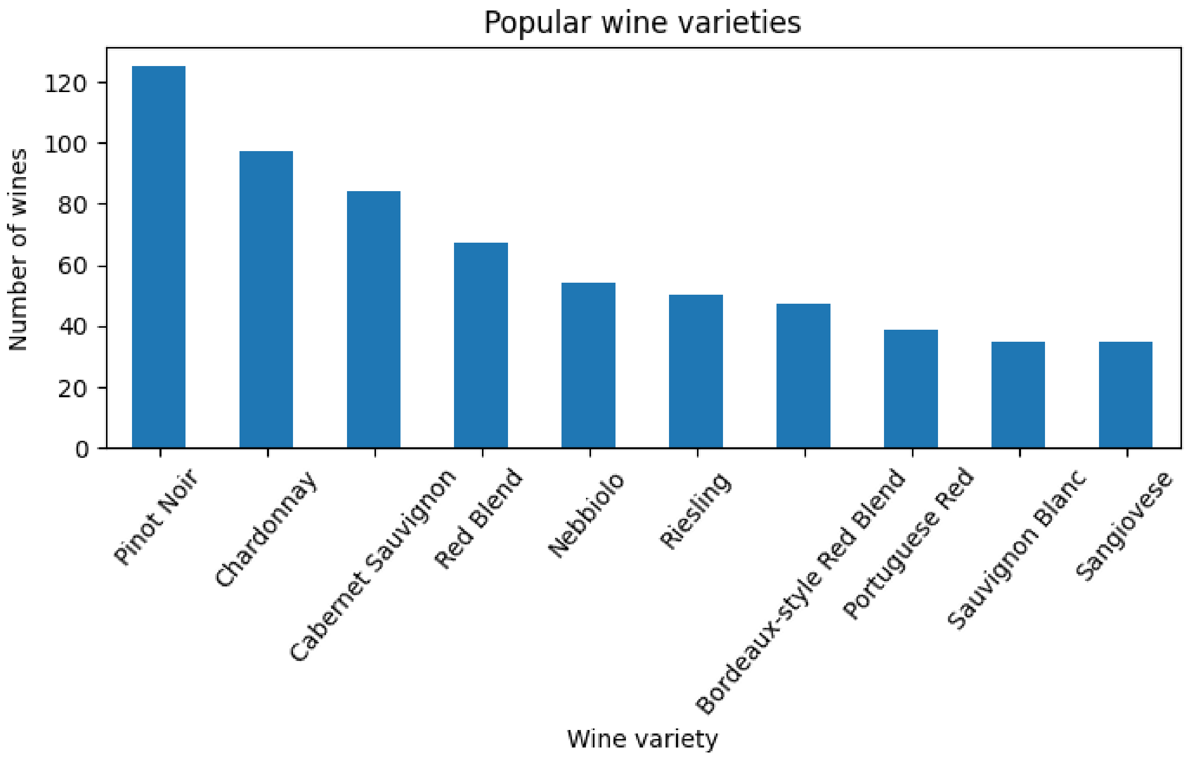
2) To further verify above point, a scattered graph presenting relationship between wine ranking('points') and price was plotted.



From above results, it does not seem there is strong correlation between price and ranking. However, for wines with rating at 96, the price is slightly higher than rest of the rating. A good range of wines are under 100. You could also get very good quality wines with rating over 90 with little spend. Therefore, you do not need to spend lots of money to buy good quality wines.

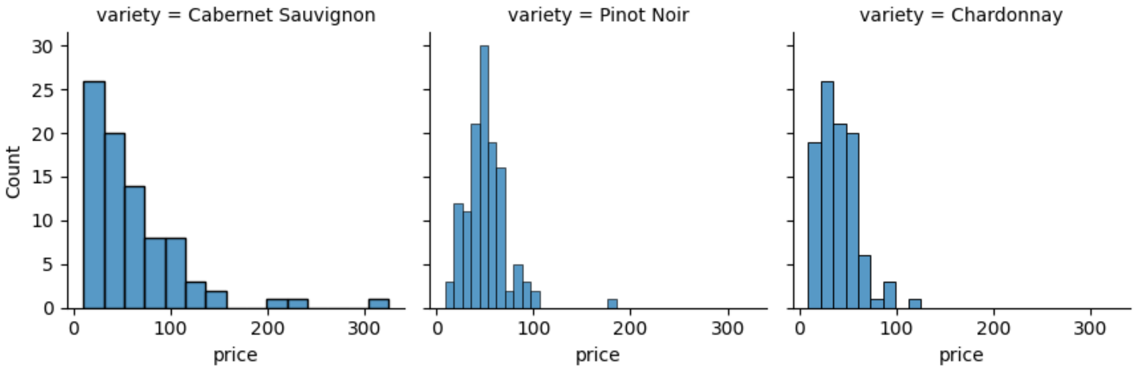
5. Find out which are the top wine varieties in the dataset. Do they cost similar?

A bar graph was plotted using 'variety' feature to select the top 10 wine varieties, as shown below:



From the result, the top 3 most popular wine varieties are named ‘Pinot Noir’, ‘Chardonnay’, and ‘Cabernet Sauvignon’, respectively.

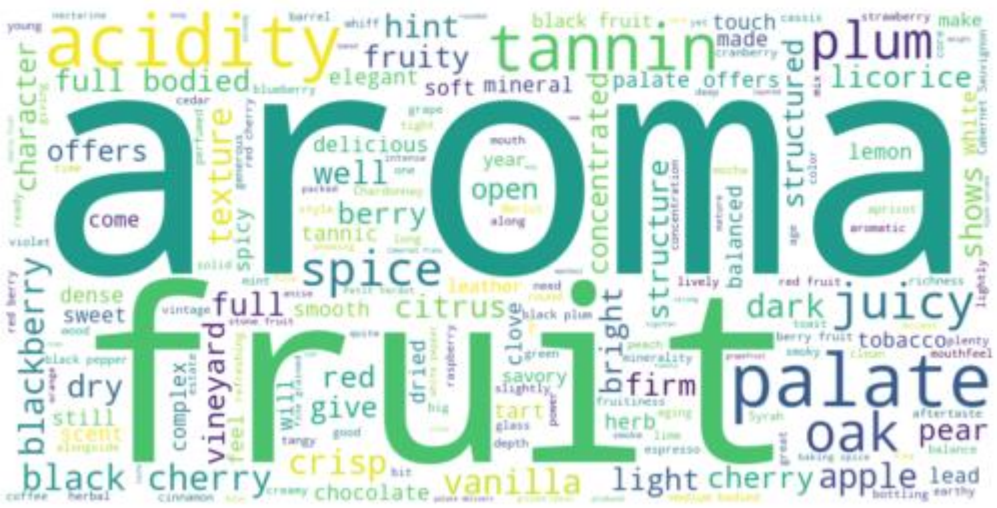
A multi-plot grid was generated to present price for each variety using histogram, as shown below:



From the results, there are few wines for ‘Cabernet Sauvignon’ variety cost higher than 200. Prices for ‘Pinot Noir’ and ‘Chardonnay’ types are generally under 100, except very few being slightly over but still under 200 for all of the wines.

6. Create word clouds of the flavours and name the most popular flavour.

A wordcloud image was generated using WordCloud with information in ‘description’ column. A range of words that are not flavour related were being removed using stopwords. The image obtained is shown as below:



It can be seen that the mostly popular flavours from descriptions of wines in the dataset are aroma, fruit, tannin, acidity, plum, spice, etc..

THIS REPORT WAS WRITTEN BY : Feifei Zhang