PROFUSION

Dilpa Rao

Predicting profitable wine ingredients

Date of submission (March 2020)

Contents

Abstract

Wine classification is a difficult task since human taste varies. The main objective here was to do data visualization and data modelling and using the insights to analyse the different attributes in the dataset and how it is impacting the prediction of our most profitable ingredients. Feature extraction of the description attribute was done by using the NLTK library of python which is basically used for text analytics to predict the most profitable wine ingredient based on good quality. The different types of wines using the profitable ingredients can be obtained by KMeans and to predict the qulaity or points we can train a classifier using Random Forest or Decision Trees.

Keywords: nltk, pos tagging, profit percentage, random forest, stopwords.

1 Introduction

Drink Today, who is leader in manufacturing and selling alcoholic beverages in UK is an important client of Profusion. They are about to start a new business 'Wine Today', which will specialize in creating and selling wines in France. The new business line will have three branches of wine:Low price, Medium price And VIP. They have approached Profusion to advise them, the 10 most profitable ingredients for each category. Based on the requirements of the client, the data science team in Profusion has already gathered an initial data-set with some information for 10,000 different wines across the world.

2 Business Objectives

The data science team aims to do a review on the different wine ingredients and find the profitable ingredients. Firstly, to do an exploratory data analysis based on the information gathered on the initial data-set. Secondly, they intend to present a market analysis of France's share of wine with the rest of the world. Thirdly, to present an analysis of of wine tasters all across the globe. Finally, to obtain the main ingredients from each wine description, used in its development.

3 Method

3.1 Dataset

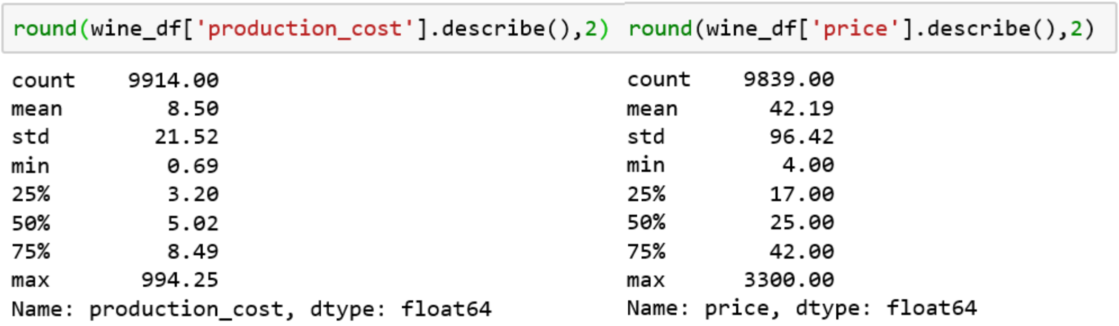
The dataset consisted of 10587 observations of wines from all over the world with 8 feature columns consisting of ('country', 'province', 'description','taster name', 'winery') were object variables containing text and the remaining ('production cost','price') were float variables while ('points') was an integer type.

3.2 Preprocessing of the data

On calculating the null values using the isna(), it was found that there were 4 missing values in both country and province, 673 missing values in price and 2076 missing values in taster name.Since price is a deciding factor for points and profitability of wine, the missing rows were dropped as we had enough data to work with. The country and province missing values were found using the winery name and then it was imputed in the respected rows.The missing values with respect to the tasters name was left as it is, as it comprised of nearly 1/5 of the data. The description column also had duplicates so we dropped those rows having duplicate description values, after cleaning the data set we had rows 9839 and the only missing values were in the taster column (1947) which we left it the way it was.

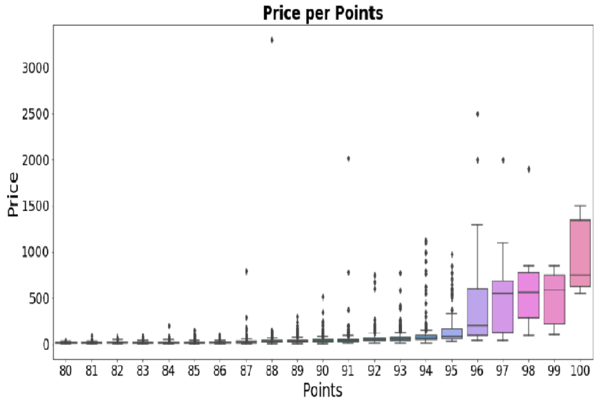
3.3 Exploratory Data Analysis

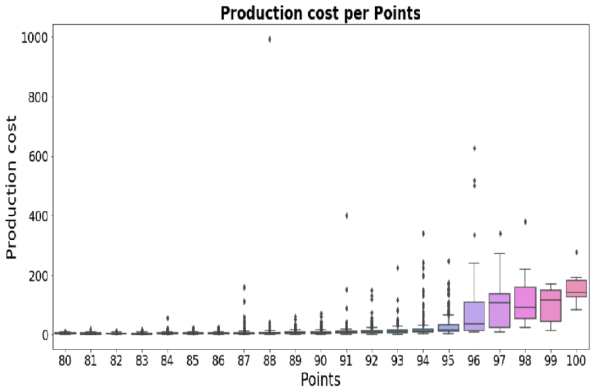
On examining the price, points and production cost column it was seen that they all have outliers and they have a positive correlation with each other.



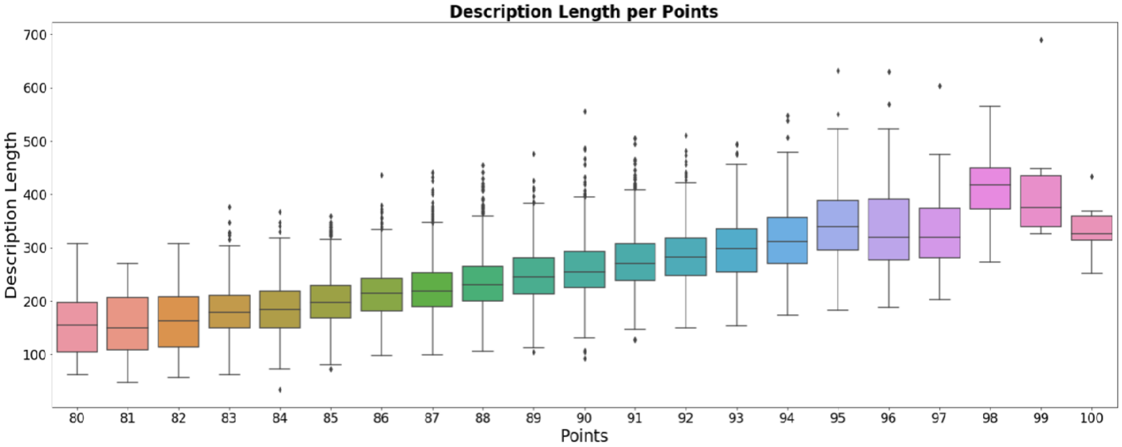
```
round(wine_df['production_cost'].describe(),2) round(wine_df['price'].describe(),2)

count    9914.00                                 count    9839.00
mean        8.50                                 mean       42.19
std        21.52                                 std        96.42
min         0.69                                 min         4.00
25%         3.20                                 25%        17.00
50%         5.02                                 50%        25.00
75%         8.49                                 75%        42.00
max       994.25                                 max      3300.00
Name: production_cost, dtype: float64            Name: price, dtype: float64
```

It is seen that the 75th percentile of the production cost is at 8.49 whereas the maximum value is at 994.25, which is 110 times more than the 75th percentile.Similarly for the price, the 75th percentile is at 42 whereas the maximum value is at 3300 which is roughly 80 times more than the 75th percentile. The most expensive wine is for 3300 and its production cost is 994.25 at 88 points belonging to Bordeaux province of France. The relationship can be further demonstrated with the box plots given below, which clearly shows a positive correlation among them.



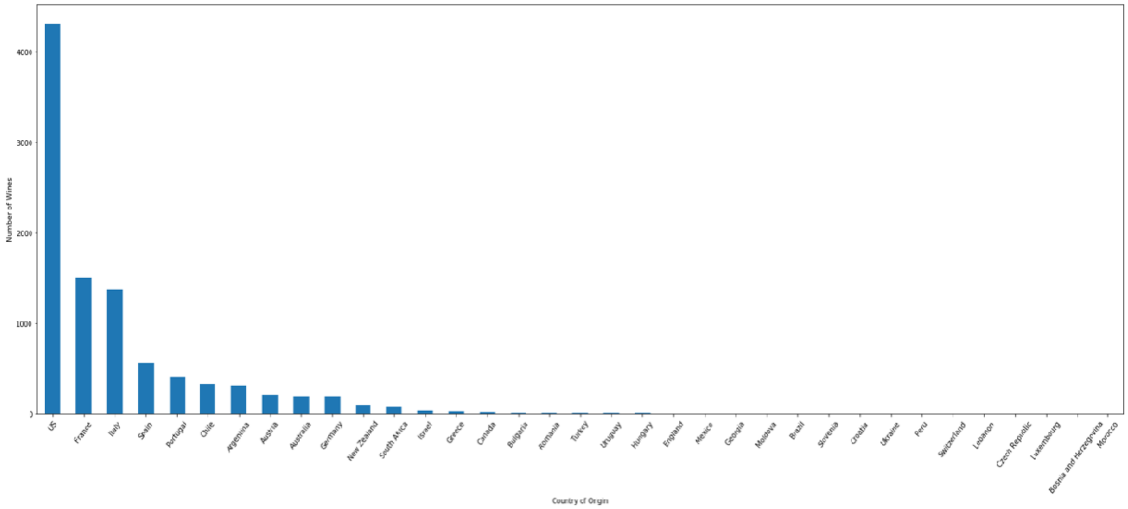


The Description column also has a positive correlation with price, points and production cost. The box plot below shows the positive correlation as the points are more so is the length of the description column. We were able to analyse this by finding the length of the description column for each wine and then comparing it with its respective points. It can be easily inferred that people would be eager to give a longer comment about the wine which they like.



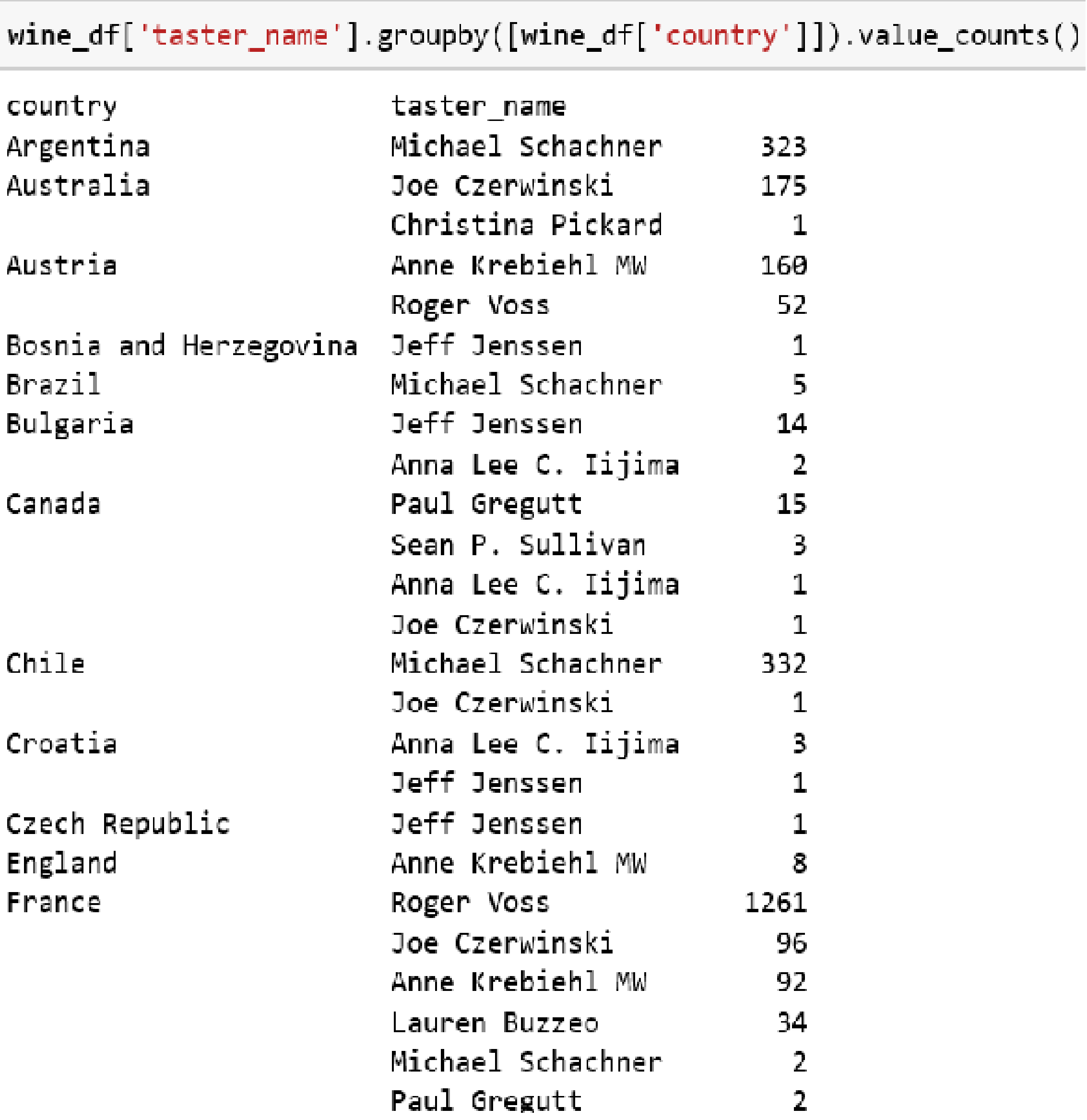
3.4 Business Objective Analysis

On doing a market analysis of France on the number of wines, it is observed that US accounts for 40% of the wines in the world having a count of 4312 wines followed by France having 1501 wines which accounts for 16% of the world share. The graph shows the number of wines from each country.With US leading, followed by France,Italy,Spain and Portugal.



There are 18 tasters world wide. These tasters ('Anna Lee C. Iijima', 'Michael Schachner', 'Alexander Peartree', 'Paul Gregutt', 'Roger Voss','Sean P. Sullivan', 'Kerin O'Keefe', 'Virginie

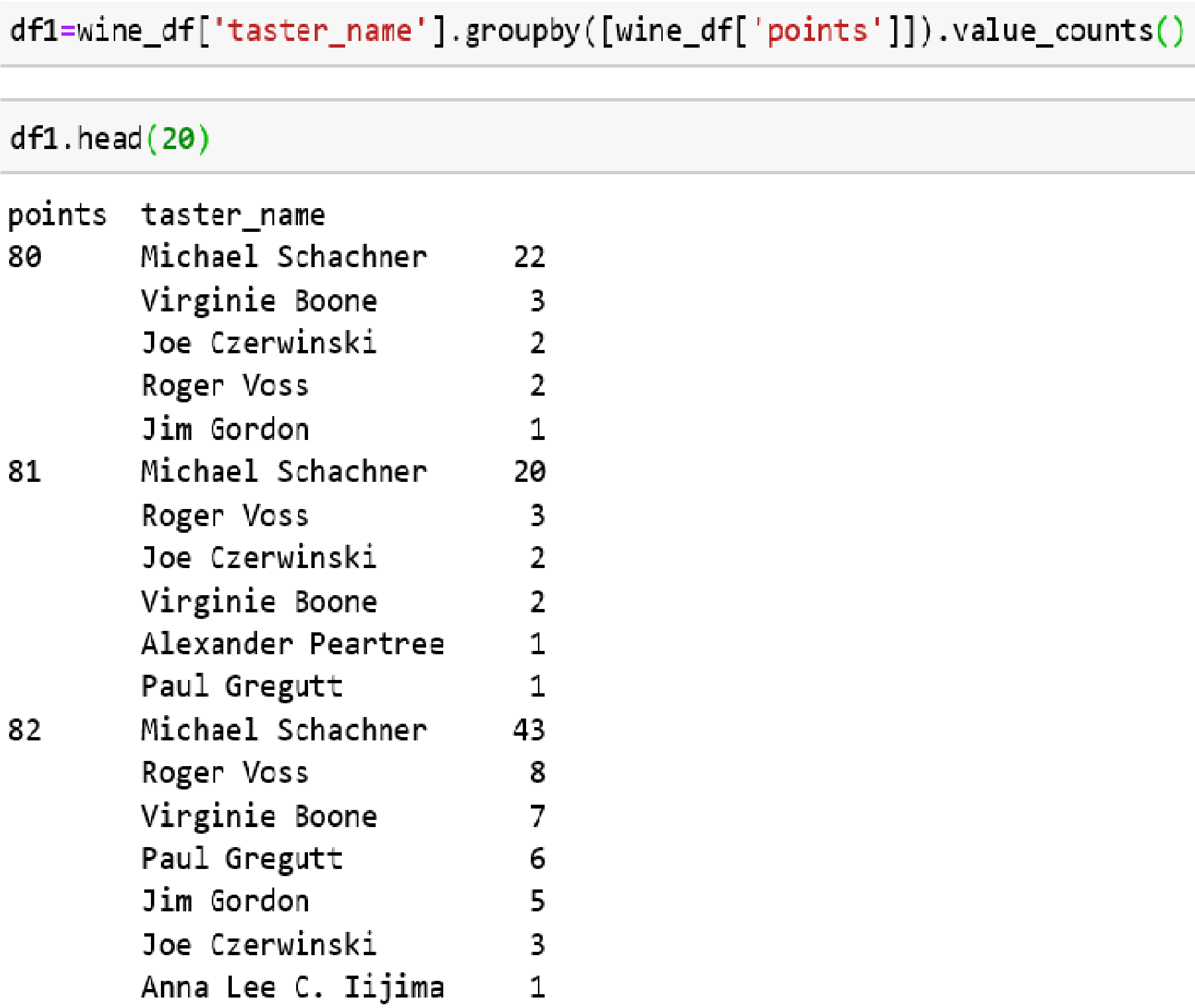
Boone', 'Matt Kettmann', 'Joe Czerwinski', 'Lauren Buzzeo', 'Anne Krebiehlxa0MW', 'Jim Gordon', 'Susan Kostrzewa', 'Jeff Jenssen', 'Mike DeSimone', 'Christina Pickard', 'Carrie Dykes') have tasted wines world wide.The chart below gives us the names of taster who tasted the wines country-wise.



```
wine_df['taster_name'].groupby([wine_df['country']]).value_counts()
```

```
country                 taster_name
Argentina               Michael Schachner      323
Australia               Joe Czerwinski         175
                        Christina Pickard        1
Austria                 Anne Krebiehl MW       160
                        Roger Voss              52
Bosnia and Herzegovina  Jeff Jenssen             1
Brazil                  Michael Schachner        5
Bulgaria                Jeff Jenssen            14
                        Anna Lee C. Iijima       2
Canada                  Paul Gregutt            15
                        Sean P. Sullivan         3
                        Anna Lee C. Iijima       1
                        Joe Czerwinski           1
Chile                   Michael Schachner      332
                        Joe Czerwinski           1
Croatia                 Anna Lee C. Iijima       3
                        Jeff Jenssen             1
Czech Republic          Jeff Jenssen             1
England                 Anne Krebiehl MW         8
France                  Roger Voss            1261
                        Joe Czerwinski          96
                        Anne Krebiehl MW        92
                        Lauren Buzzeo           34
                        Michael Schachner        2
                        Paul Gregutt             2
```

On the other hand the table below shows which taster had alloted the points from 80-100 to the different wines.

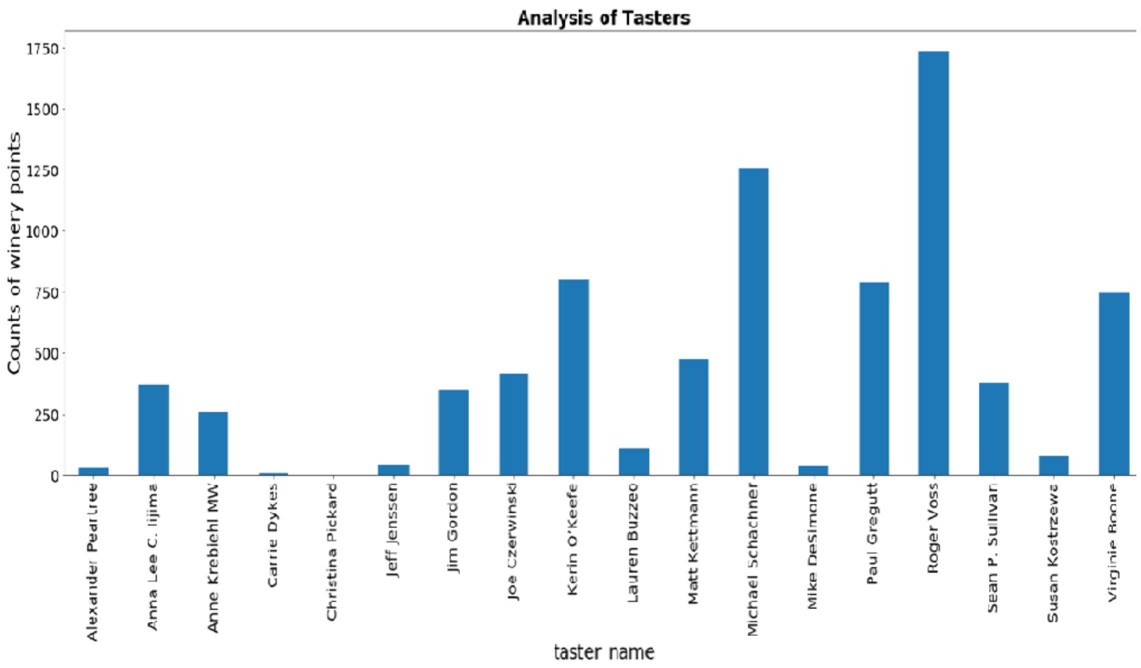


```
df1=wine_df['taster_name'].groupby([wine_df['points']]).value_counts()
```

```
df1.head(20)
```

```
points  taster_name
80      Michael Schachner     22
        Virginie Boone         3
        Joe Czerwinski         2
        Roger Voss             2
        Jim Gordon             1
81      Michael Schachner     20
        Roger Voss             3
        Joe Czerwinski         2
        Virginie Boone         2
        Alexander Peartree     1
        Paul Gregutt           1
82      Michael Schachner     43
        Roger Voss             8
        Virginie Boone         7
        Paul Gregutt           6
        Jim Gordon             5
        Joe Czerwinski         3
        Anna Lee C. Iijima     1
```

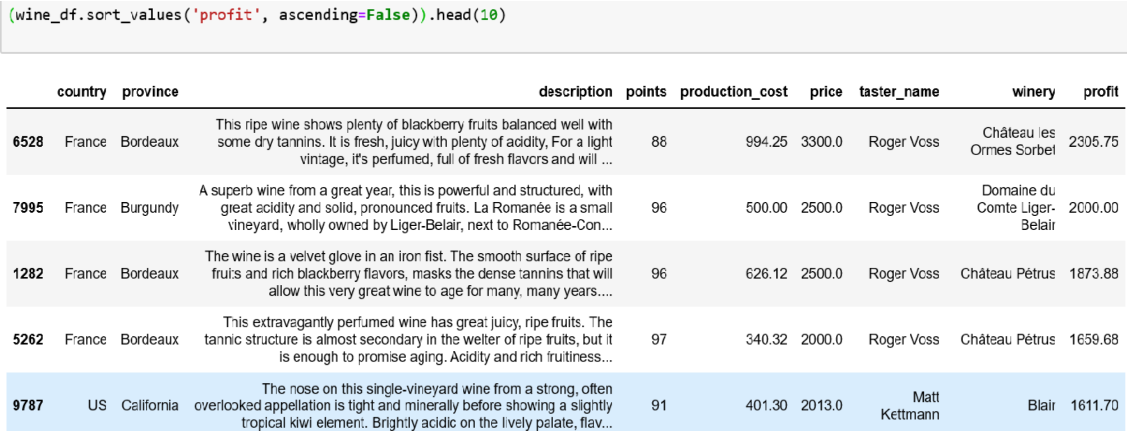
The graph below further enumerates the number of wines each taster tasted. From the table below we can analyse that Roger Voss is very famous and well known in wine industry having tasted 1737 wines all over the world followed by Michael Schanner who has scored 1258 wines.



4 Results

To calculate the most profitable wines world wide two additional columns for calculating profit and profit percent were added to the dataset.Since the new business line of Wine Today has three branches namely,Low price, Medium price and VIP, a new column priced grouped was added. The Low price category comprised of prices within the 25th percentile of the price column, the Medium price category comprised of prices from the 25th percentile to the 75th percentile and finally the VIP was from the 75th percentile to the 100th. It was further analysed that Low price and VIP both accounts for roughly 25% of wines, whereas Medium price accounts for about 50% of the total wines.

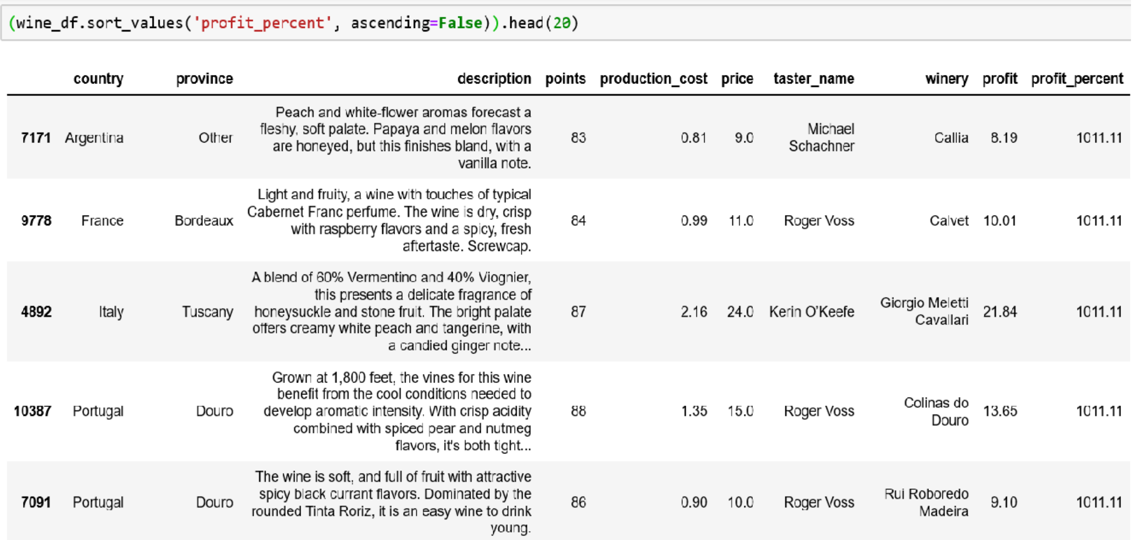
On calculating the profit it was observed that out of the top 5 profit making wines, four of them belonged to France.



```
(wine_df.sort_values('profit', ascending=False)).head(10)
```

	country	province	description	points	production_cost	price	taster_name	winery	profit
6528	France	Bordeaux	This ripe wine shows plenty of blackberry fruits balanced well with some dry tannins. It is fresh, juicy with plenty of acidity, For a light vintage, it's perfumed, full of fresh flavors and will ...	88	994.25	3300.0	Roger Voss	Château les Ormes Sorbet	2305.75
7995	France	Burgundy	A superb wine from a great year, this is powerful and structured, with great acidity and solid, pronounced fruits. La Romanée is a small vineyard, wholly owned by Liger-Belair, next to Romanée-Con...	96	500.00	2500.0	Roger Voss	Domaine du Comte Liger-Belair	2000.00
1282	France	Bordeaux	The wine is a velvet glove in an iron fist. The smooth surface of ripe fruits and rich blackberry flavors, masks the dense tannins that will allow this very great wine to age for many, many years....	96	626.12	2500.0	Roger Voss	Château Pétrus	1873.88
5262	France	Bordeaux	This extravagantly perfumed wine has great juicy, ripe fruits. The tannic structure is almost secondary in the welter of ripe fruits, but it is enough to promise aging. Acidity and rich fruitiness...	97	340.32	2000.0	Roger Voss	Château Pétrus	1659.68
9787	US	California	The nose on this single-vineyard wine from a strong, often overlooked appellation is tight and minerally before showing a slightly tropical kiwi element. Brightly acidic on the lively palate, flav...	91	401.30	2013.0	Matt Kettmann	Blair	1611.70

On comparing the profit of the wines with the basis of profit percent, it was observed that the top 5 profit making wines weren't the same one with the highest profit percent. The most expensive wine of France priced at 3300 shows a profit percent of only 231.91 whereas the Low price wines have a profit percent of 1011 as shown in the table below.



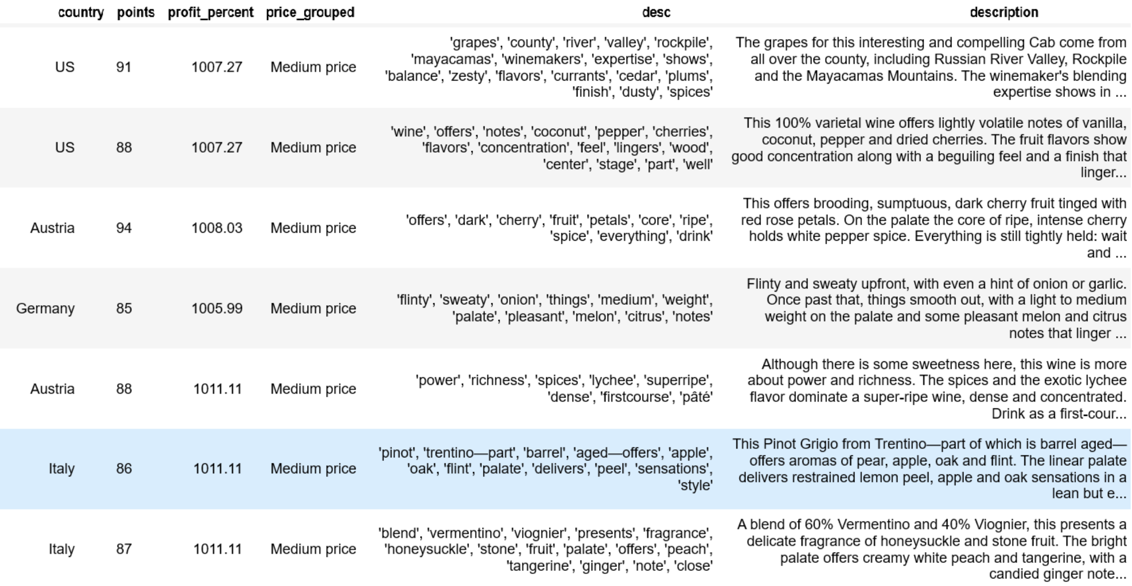
```
(wine_df.sort_values('profit_percent', ascending=False)).head(20)
```

	country	province	description	points	production_cost	price	taster_name	winery	profit	profit_percent
7171	Argentina	Other	Peach and white-flower aromas forecast a fleshy, soft palate. Papaya and melon flavors are honeyed, but this finishes bland, with a vanilla note.	83	0.81	9.0	Michael Schachner	Callia	8.19	1011.11
9778	France	Bordeaux	Light and fruity, a wine with touches of typical Cabernet Franc perfume. The wine is dry, crisp with raspberry flavors and a spicy, fresh aftertaste. Screwcap.	84	0.99	11.0	Roger Voss	Calvet	10.01	1011.11
4892	Italy	Tuscany	A blend of 60% Vermentino and 40% Viognier, this presents a delicate fragrance of honeysuckle and stone fruit. The bright palate offers creamy white peach and tangerine, with a candied ginger note...	87	2.16	24.0	Kerin O'Keefe	Giorgio Meletti Cavallari	21.84	1011.11
10387	Portugal	Douro	Grown at 1,800 feet, the vines for this wine benefit from the cool conditions needed to develop aromatic intensity. With crisp acidity combined with spiced pear and nutmeg flavors, it's both tight...	88	1.35	15.0	Roger Voss	Colinas do Douro	13.65	1011.11
7091	Portugal	Douro	The wine is soft, and full of fruit with attractive spicy black currant flavors. Dominated by the rounded Tinta Roriz, it is an easy wine to drink young.	86	0.90	10.0	Roger Voss	Rui Roboredo Madeira	9.10	1011.11

The NLTK library was used to find the most profitable ingredients from the description column of each row of the dataset. In order to get only the ingredients we did a preprocessing using the different NLTK library functions. From the description column, we removed the numbers, punctuation, stopwords('it', 'they', 'and', 'the','of', 'in','to','from' and many more) after that we used the pos tag function and assigned each word to its figure of speech and finally, we selected the nouns(NN, NNS, NNP) from the pos tagged words and added those selected keywords to each row in a new column called desc.

After that we sorted the dataframe according to maximum profit percentage and took the first 20 rows. It was observed that out of 20 rows, 6 of them were categorised as Low price and VIP price each and the remaining 8 rows were categorised as Medium price. Based on the desc column we

took 10 ingredients from each of these categories.The table below gives us the maximum profit percentage of the wines with their key ingredients.



country	points	profit_percent	price_grouped	desc	description
US	91	1007.27	Medium price	'grapes', 'county', 'river', 'valley', 'rockpile', 'mayacamas', 'winemakers', 'expertise', 'shows', 'balance', 'zesty', 'flavors', 'currants', 'cedar', 'plums', 'finish', 'dusty', 'spices'	The grapes for this interesting and compelling Cab come from all over the county, including Russian River Valley, Rockpile and the Mayacamas Mountains. The winemaker's blending expertise shows in ...
US	88	1007.27	Medium price	'wine', 'offers', 'notes', 'coconut', 'pepper', 'cherries', 'flavors', 'concentration', 'feel', 'lingers', 'wood', 'center', 'stage', 'part', 'well'	This 100% varietal wine offers lightly volatile notes of vanilla, coconut, pepper and dried cherries. The fruit flavors show good concentration along with a beguiling feel and a finish that linger...
Austria	94	1008.03	Medium price	'offers', 'dark', 'cherry', 'fruit', 'petals', 'core', 'ripe', 'spice', 'everything', 'drink'	This offers brooding, sumptuous, dark cherry fruit tinged with red rose petals. On the palate the core of ripe, intense cherry holds white pepper spice. Everything is still tightly held: wait and ...
Germany	85	1005.99	Medium price	'flinty', 'sweaty', 'onion', 'things', 'medium', 'weight', 'palate', 'pleasant', 'melon', 'citrus', 'notes'	Flinty and sweaty upfront, with even a hint of onion or garlic. Once past that, things smooth out, with a light to medium weight on the palate and some pleasant melon and citrus notes that linger ...
Austria	88	1011.11	Medium price	'power', 'richness', 'spices', 'lychee', 'superripe', 'dense', 'firstcourse', 'pâté'	Although there is some sweetness here, this wine is more about power and richness. The spices and the exotic lychee flavor dominate a super-ripe wine, dense and concentrated. Drink as a first-cour...
Italy	86	1011.11	Medium price	'pinot', 'trentino—part', 'barrel', 'aged—offers', 'apple', 'oak', 'flint', 'palate', 'delivers', 'peel', 'sensations', 'style'	This Pinot Grigio from Trentino—part of which is barrel aged—offers aromas of pear, apple, oak and flint. The linear palate delivers restrained lemon peel, apple and oak sensations in a lean but e...
Italy	87	1011.11	Medium price	'blend', 'vermentino', 'viognier', 'presents', 'fragrance', 'honeysuckle', 'stone', 'fruit', 'palate', 'offers', 'peach', 'tangerine', 'ginger', 'note', 'close'	A blend of 60% Vermentino and 40% Viognier, this presents a delicate fragrance of honeysuckle and stone fruit. The bright palate offers creamy white peach and tangerine, with a candied ginger note...

In the Low price category, based on maximum profit percentage of wines, the most popular ingredients are raspberry,black currant, pear, nutmeg, red currant, vanilla, green melon, tangerine, peach and papaya. The first five ingredients were mostly found in all European wines but the last five ingredients were mostly found in wines from North and South America.

For the Medium price wines, the most popular ingredients were cherry, melon, citrus, lychee, apple, peach, tangerine, honeysuckle ,oak and plums.

For the VIP price wines, the most popular ingredients were pomegranate, cherry, coffee, strawberry, grapefruit, plum, eucalyptus, oak, butterscotch and vanilla. The last two ingredients were found in wines from North America. the rest of the ingredients were found in European wines.

5 Discussions

We could have used the Kmeans clustering algorithm with k = 5 to find 5 different types of wines with key ingredients and based on these key ingredients we could create different wine types with atleast 3 of these ingredients.We could split the dataset into 80-20 train test ratio and train a classifier like Random Forest with the target variable as points to create different wine types with atleast a score of 87, but whether this is the correct approach is still open to debate[1].

Moreover, while doing text analytics on the description table and using the stop word function we could have added 'flavor', 'wine' as it mostly appears in all the rows and it doesn't provide us with any useful information.

6 Limitations

There are a lot of complexity involved when calculating whether the ingredients can be available locally or need to be imported. Moreover, due to the complex nature of the problem to create a combination of 3 ingredients out of 30 ingredients selected from the top 20 rows of maximum profit percentage and to create a wine with a score of atleast 87 points needs more research and time to obtain accurate results.

7 Conclusion

In recent years, the interest in wine has increased, leading to growth of the wine industry. As a consequence, companies are investing in new technologies to improve wine production and selling. Quality certification is a crucial step for both processes and is currently largely dependent on wine tasting by human experts[2]. The data science team at Profusion aims at the prediction of key wine ingredients from objective analytical tests to make the business more profitable and a world leader.

We just found out the 10 most profitable ingredients from each category of wine on the basis of maximum profit percentage. Wine Today can use these ingredients to make different varieties of wine to increase their profit capacity and become the world leader in wine creating and selling.

References

[1] Yogesh Gupta. Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science*, 125:305–312, 2018.

[2] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.