Case Study 4 Wine Analysis & Prediction

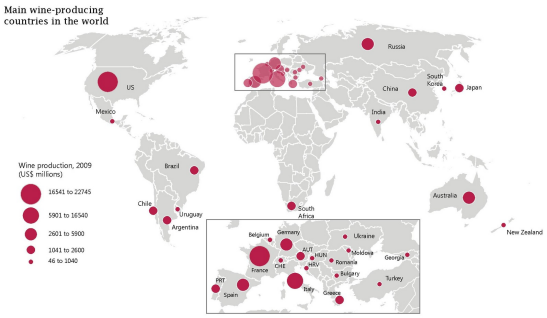
Group 7 Aishwarya Ramakrishnan, Jingyan(Amy) Sun, Xueying(Flora) Zhang

Case Study Topic

Our topic for Case Study 4 is Wine Analysis & Prediction. Wine (Though in limited quantity) is good for human body, and there are increasingly large number of wine drinkers. However, there are too many different wine brands all around the world and there's no standard for what is a good wine, which can bring a huge trouble for wine consumers to choose wine. Is this bottle of wine worth that much? In our case study, we use the data of wine reviews of 130,000 wine reviews, hope that after our analysis consumers could easily find whether a bottle of wine is expensive or cheap. For this study, we used the data of wine reviews with 130,000 wine reviews. We sincerely hope that after our analysis, consumers could easily find whether a bottle of wine is expensive or cheap.

Background and Motivation

Wine manufacturing is a mature industry. The technology advancements have slowed down, and the growth rate is lower than the growth rate of GDP. There are many wine brands all over the world, almost all countries and regions have their local wine brands. There's no monopoly, or major players in this industry. (Information Source: IBIS Global Wine Industry Report http://clients1.ibisworld.com/reports/gl/industry/default.aspx?entid=410 Wikipedia Wine Production https://en.wikipedia.org/wiki/File:WineProduction.pdf) .



Picture 1. Main Wine-producing Countries in the World

The perfectly competitive market enriches the diversity of products, which means there's no absolute standards for wine quality, and this condition brings a lot of trouble to consumers to evaluate a product.

As consumers, we don't have wine-tasting training, and hence it is hard to tell whether a wine is good or not, and whether an expensive bottle of wine is worth that much. The price of wine varies within a very huge range, and there are many variables that could influence the price, like winery, year, and type of grapes. How can we find the real price of a bottle of wine? We want to build a model to help consumers evaluate wines. We are trying to predict the estimated price of the wine when consumers provide some information about the wine(from the label), say variety of grapes, description or year; so that consumers won't feel confused about the real price of wine.

Data Sources

We downloaded the data from Kaggle https://www.kaggle.com/zynicide/wine-reviews . This data set has 13 columns, includes title, country, winery, variety (the type of grapes used to make the wine), points, price and so on. The year is included in title column and we use excel to get them out and append a new column as "Year".

Methodology

After downloading the dataset, we first try to segregate the year out. In the title column, it combines the brand, year, and type of grape together in this one column, and the number of words the brands have is different, from two words to five, and there are many non-English words. So, we use Excel to help us. After viewing the data, we find that the years are all after 2000. So we choose "20" as the beginning words, and set to return the following four words. The function is =MID (K3, FIND (20,K3), 4) (K is the "title" column). Then, we cleaned the rows with missing fields especially the price, point, and year value. The cleaned data file is then used for further processing and analysis in python.

We make some plots first to get a basic statistical insight of data. We get the number of wines per point, top 10 most popular wine producing countries, price distribution, correlation between price and point, country-price distribution, variety of grapes that are used to make the most expensive wines and so on.

In the end, we choose Random Forest to build the models to help us with our prediction. We try to use description and variety to predict the price, and do an experiment on wines made in US.

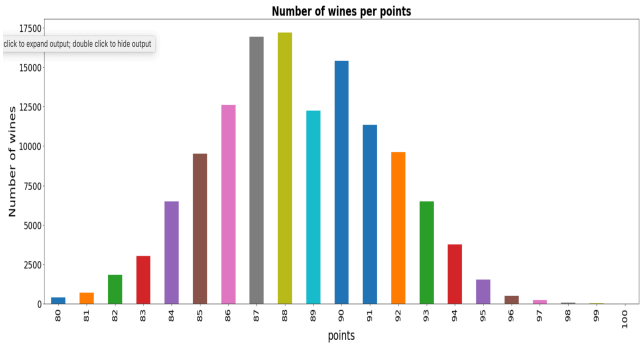
Results

Math Part

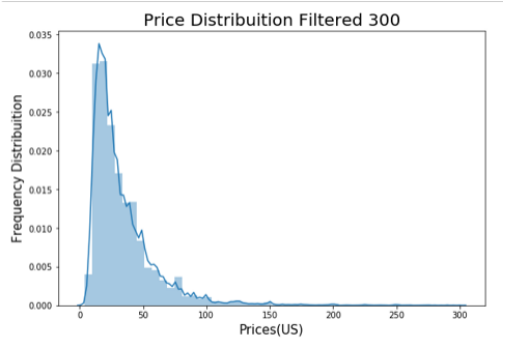
In order to figure out the trend of the wine market, we mainly focus on some numerical variables in term of "points", "price", and "reviews"(in length); as well as categorical variables like "Country" "variety" by applying bar chart, boxplot, pie chart, scatter plot, tree map and so on.

1.About "points", "price" distribution and "point and price" correlation

From the left figure, we can tell the mode of this data is 88, the distribution of ratings is almost a normal distribution, except the slight decrease in 89 points. Most of reviewers or tasters prone to give a general grade (above 85). And from the figure in the right, we filtered in the wine price under 300, because the price higher than 300 USD seem to be outliers in our data, then we found that the mode price is between 25 to 50, which means the approachable wines are popular among costumers.

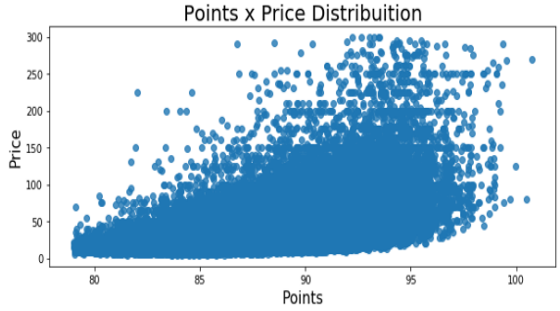


Picture 2. The number of wines per point.



Picture.3 The wine price (<300 USD) distribution

We try to dig out the relationship between price and points. Firstly, we generate the scatter plot to see the trend (left figure below), we found that the higher price of wine tends to have higher points, but not significant, because the large group of wine under 50 USD has been rated varied from 85 to 95. Then verified our assumption by correlation heat map, as seen on the figure in the right, the correlation between these two features are 0.42, which means not strong enough.

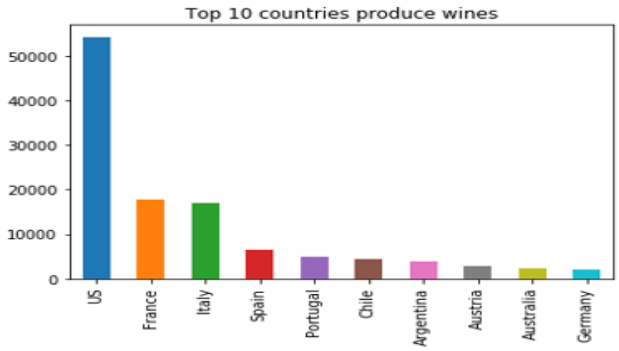


Picture 4. point-price under 300 USD distribution

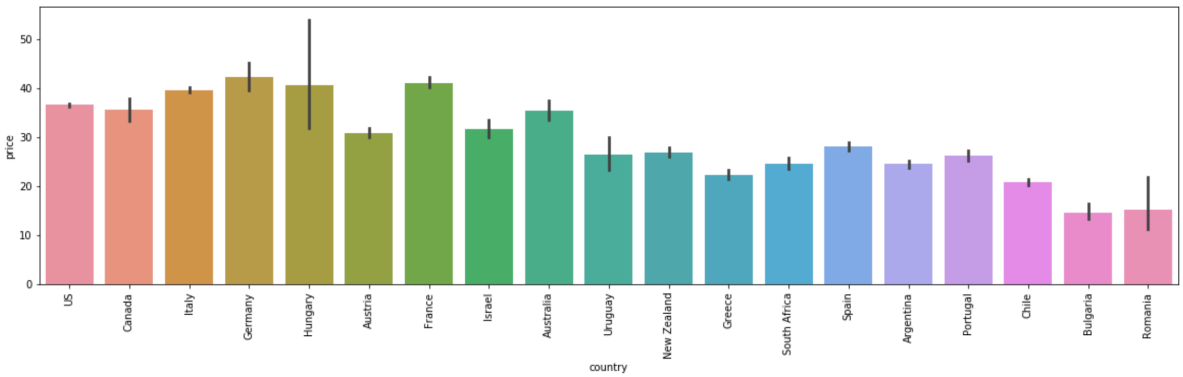
Picture 5. price-point correlation heat map

2.About “Country”, “Country and Price”, “Country and Points” analysis

In our data, USA is the largest wine producing country. There are more than 50,000 types of wine products. This is then followed by the European countries like France, Italy, Spain and Portugal. Then follows the Asian country China and Argentina in South America. Based on the significant producing amount of USA, in following computer part, we will focus the case study on USA wine.



Picture 6. Top 10 Most Popular Wine Producing Countries.



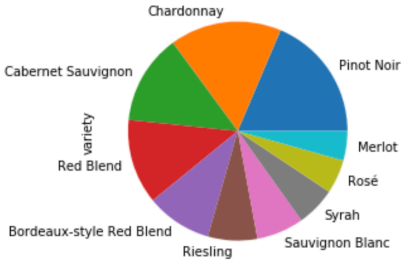
Picture 7. Country-price Distribution

In the price count chart, we see that the price range is huge from $4 to $3,300, with most of them concentrated on $4 to $100 range. Here in price setting, we may assume that their inner competition is unassertive (no cutthroat competition like price competition). While in Hungary, the price is also high though, just a little lower than in Germany, but the price range is very large. This shows that in Hungary, some wineries have to lower down the price to compete with other wineries. The profitability is not steady in this industry here and the quality of wine may change a lot in Hungarian wine brands.

3.About “Variety”, “Variety and Points” analysis

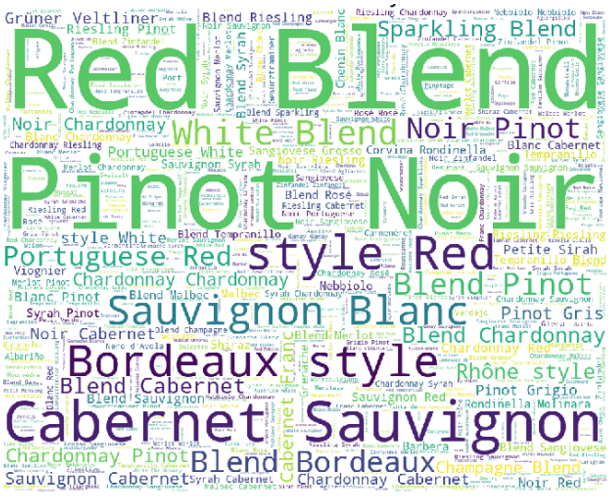
This is the points by varieties. We can see that the grapes we eat daily like *Rose* grape, have lower ratings. The varieties especially grown for wine usually get higher points. Below are the top 10 most popular wine grape varieties. They together take 56.94% of all the wine products. The most popular grape is *Pinot Noir*, which is mainly grown in French, US, and New Zealand. Then follows the *Chardonnay*, *Cabernet Sauvignon*, *Red Blend* and *Bordeaux-style Red Blend*. The wines made from *Bordeaux-style Red Blend* grape only takes 4.58% of all wine products, but it's well welcomed in high-end wineries to make expensive wines.

variety	count	percent
Pinot Noir	12479	10.84%
Chardonnay	10658	9.26%
Cabernet Sauvignon	9082	7.89%
Red Blend	8190	7.12%
Bordeaux-style Red Blend	5273	4.58%
Riesling	4927	4.28%
Sauvignon Blanc	4740	4.12%
Syrah	4035	3.51%
RosV©	3222	2.80%
Merlot	2925	2.54%

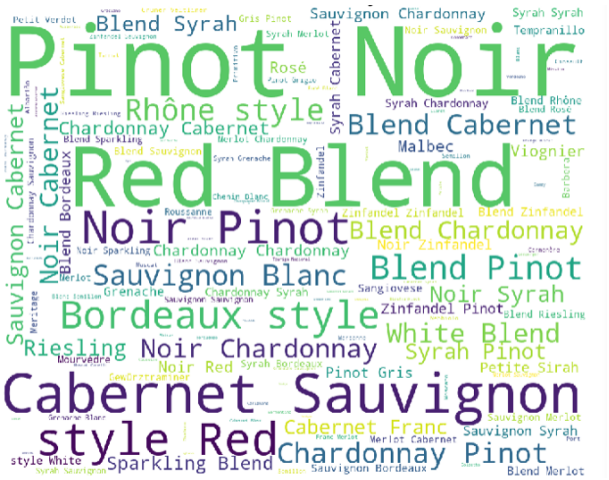


Picture 8 & 9. Top 10 Grape Varieties

We also tried to visualize the data by word cloud to see the worldwide (data-based) variety distribution and the largest production wine country (USA) variety distribution despite without the numerical analysis, as picture 9 and 10 shown. Based on the most popular types of grapes shown in both the word cloud's, we can easily tell that US wine variety has narrowed down those less popular varieties but kept the most popular varieties. So, in further research, we could set US wine as a model to make prediction.



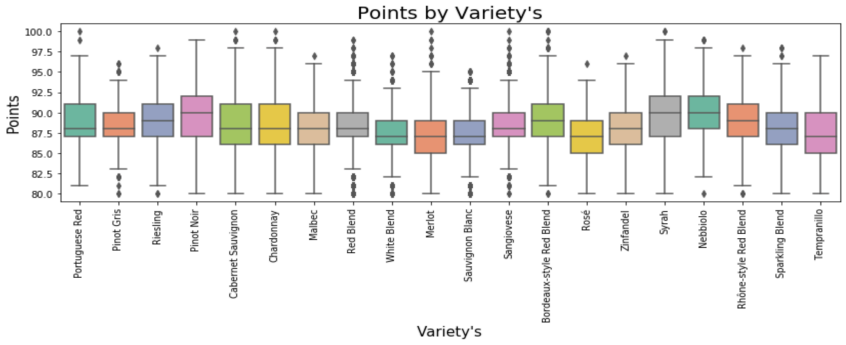
Picture 10. Word Cloud—Worldwide trend



Picture 11. Word Cloud--- US wine trend

Following boxplot is the Points by Variety. We can see that the average point of *Pinot Noir* is relatively high, with highest points at the same time and without outliers; though the *Red Blend*

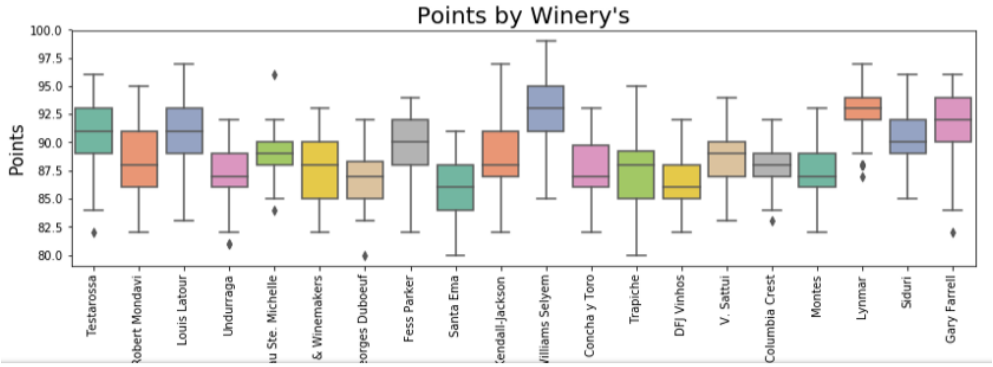
is most popular variety according to the word cloud output, the points for this certain type of grapes seems not the best, with salient high points but with many outliers.



Picture 12. Points by Varieties

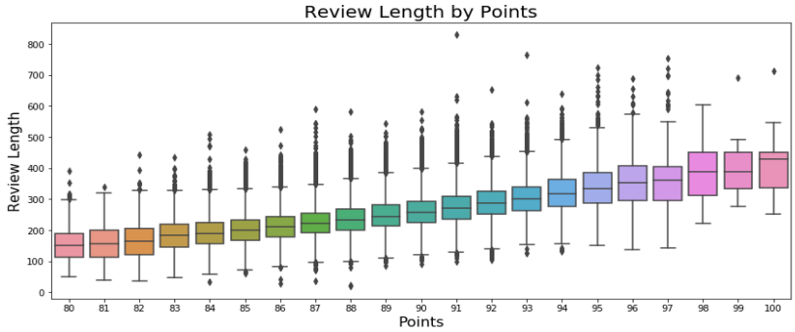
4.“Point and Winery” and “Point and Description length” analysis

In the following points by winery boxplot, we can tell that the Williams selyem winery has the highest points and also highest average points.



Picture 13. Points by Wineries

Remarkably, we found that the review length will affect the points given by tasters in some degree, the larger tasters writing for description, the higher possibility the wine will have a higher point. The data trend is almost fit the positive linear regression.



Picture 14. Review Length by Points

Computer Part

Cleaned data is taken and price range is calculated in order to use it as the dependent variable in the Random forest classification model.

Case1: Description to price - Random Forest

Here the input is the description of the wine, where the composition and type of wine are described. This is the kind of information an end user would have and would like in the one they would like to taste. Tfidf vectorizer was used to transform the text to matrix. We are trying to predict the price of the wine based on the description. The price range classification is done on a a very superficial level, splitting the data into equal bins. As shown in the classification report and confusion matrix, we can see that the accuracy rate is really high when using the Random Forest classifier; which is around 93%. This means that, we would be able to accurately predict the price (expected) for a particular wine based on its description.

```
def transform_price_simplified(price):
    if price < 50:
        return 1
    elif price >= 50 and price < 200:
        return 2
    elif price >= 200 and price < 1000:
        return 3
    elif price >= 1000 and price < 2500:
        return 4
    else:
        return 5
```

	precision	recall	f1-score	support
1	0.92	1.00	0.96	3135
2	0.98	0.69	0.81	626
3	1.00	1.00	1.00	8
5	1.00	0.66	0.79	237
micro avg	0.93	0.93	0.93	4006
macro avg	0.98	0.84	0.89	4006
weighted avg	0.93	0.93	0.92	4006

Actual \ Predicted	1	2	3	5
1	3131	4	0	0
2	197	429	0	0
3	0	0	8	0
5	77	4	0	156

```
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 2 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 5 and Predicted outcome :: 5
Actual outcome :: 2 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 2 and Predicted outcome :: 1
```

Picture 15 & 16 & 17. Input and Output of Case 1 (93% Accuracy)

Case2: Variety to price - Random Forest

Here the input is the variety of the wine, where the type of grapes used in the wine is mentioned. This is the kind of information an end user would have (present in the wine bottle label) and would like in the one they would like to taste. Tfidf vectorizer is used to transform the text to matrix. Here the price categorization is more detailed targetting the data distribution based on

price. We are trying to predict the price of the wine based on what grapes are used. As shown in the classification report and confusion matrix, we can see that the accuracy rate is decent enough when using the Random Forest classifier; which is around 60%. This means that, we would be able to predict the price (expected) with 60% accuracy for a particular wine based on the variety of wine.

```
def transform_price_simplified(price):
    if price < 25:
        return 1
    elif price >= 25 and price < 100:
        return 2
    elif price >= 100 and price < 1000:
        return 3
    elif price >= 1000 and price < 2500:
        return 4
    else:
        return 5
```

	precision	recall	f1-score	support
1	0.62	0.56	0.58	24481
2	0.60	0.77	0.67	29133
3	0.33	0.00	0.00	1937
4	0.00	0.00	0.00	8
5	0.25	0.00	0.00	4114
micro avg	0.60	0.60	0.60	59673
macro avg	0.36	0.27	0.25	59673
weighted avg	0.57	0.60	0.57	59673

Actual \ Predicted	1	2	3	5
1	13596	10878	1	6
2	6661	22460	3	9
3	302	1632	3	0
4	0	8	0	0
5	1448	2659	2	5

```
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 2
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 5 and Predicted outcome :: 1
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 1 and Predicted outcome :: 1
```

Picture 18 & 19 & 20. Input and Output of Case 2 (60% Accuracy)

Case3: Variety to price (USA Only) - Random Forest

This is a similar case as above, the difference being that the data taken is only for USA brands(as we see that USA has most production of wines) of wine. Here the input is the variety of the wine, where the type of grapes used in the wine is mentioned. Tfidf vectorizer was used to transform the text to matrix. We do further, in-detail split for the price factor in order to find more accurate contry based pricing. We are trying to predict the price of the wine based on what grapes are used. As shown in the classification report and confusion matrix, we can see that the accuracy rate is average, when using the Random Forest classifier; which is around 55%. This means that, when a subset of variety is taken (say USA) we could predict the price (expected) correctly, 55% accurately for a particular wine.

```
def transform_price_simplified(price):
    if price < 25:
        return 1
    elif price >= 25 and price < 50:
        return 2
    elif price >= 50 and price < 150:
        return 3
    elif price >= 150 and price < 250:
        return 4
    elif price >= 250 and price < 400:
        return 5
    elif price >= 400 and price < 600:
        return 6
    elif price >= 600 and price < 800:
        return 7
    elif price >= 800 and price < 1000:
        return 8
    else:
        return 9
```

	precision	recall	f1-score	support
1	0.67	0.36	0.47	4256
2	0.57	0.64	0.60	8623
3	0.49	0.59	0.53	4842
4	0.00	0.00	0.00	135
5	0.00	0.00	0.00	17
6	0.00	0.00	0.00	1
7	0.00	0.00	0.00	2
micro avg	0.55	0.55	0.55	17876
macro avg	0.25	0.23	0.23	17876
weighted avg	0.56	0.55	0.55	17876

Actual \ Predicted	1	2	3
1	1533	2208	515
2	728	5538	2357
3	39	1961	2842
4	2	33	100
5	0	4	13
6	0	0	1
7	0	1	1

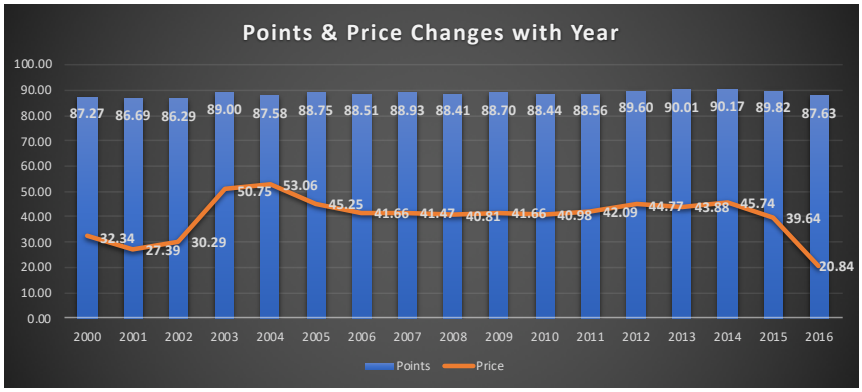
```
Actual outcome :: 2 and Predicted outcome :: 3
Actual outcome :: 3 and Predicted outcome :: 3
Actual outcome :: 3 and Predicted outcome :: 3
Actual outcome :: 1 and Predicted outcome :: 2
Actual outcome :: 3 and Predicted outcome :: 2
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 3 and Predicted outcome :: 3
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 3 and Predicted outcome :: 3
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 2 and Predicted outcome :: 1
Actual outcome :: 2 and Predicted outcome :: 1
Actual outcome :: 1 and Predicted outcome :: 1
Actual outcome :: 2 and Predicted outcome :: 3
Actual outcome :: 3 and Predicted outcome :: 3
Actual outcome :: 2 and Predicted outcome :: 2
Actual outcome :: 3 and Predicted outcome :: 3
Actual outcome :: 2 and Predicted outcome :: 2
```

Picture 21 & 22 & 23. Input and Output of Case 3 (55% Accuracy)

Conclusions

Low-ending Wines

For normal wines price under $100, wines that were made in 2004 are overpriced, with a very high price, but a modest taste. The wine made in 2014 is of good quality, and the wine made in 2016 is highly cost effective, only about half price and the point is 2.54 lower than that of the wine made in 2016.



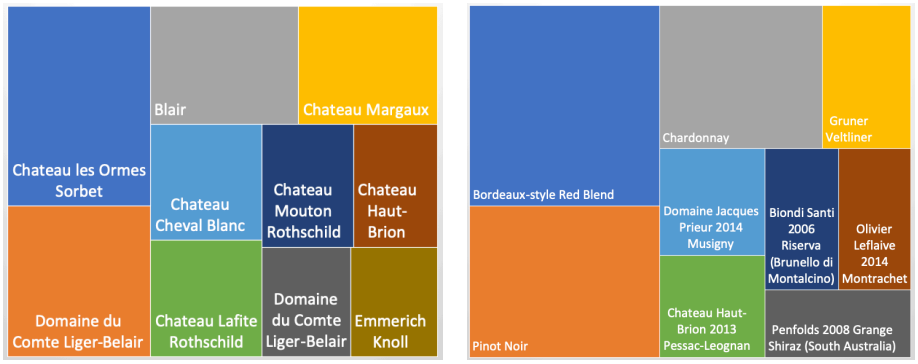
Picture 24. Point & Price by Year

Based on this set of data from all the experiments, we can see we are able to predict the price of the wine (be it a new model or existing one) based on the model. Wine price is a comprehensive reflection of all the variables, including place of production, year, varieties, and winery.

Expensive wines don't necessarily taste that good, consumers must pay attention when choose a wine. Don't be cheated by price!

High-ending Wines

For the high-ending wines, we study the most expensive wines. The first chart is the most expensive wineries. We can see that 8 of them (name begins with "*Chateau*" or "*Domaine*") are in Bordeaux Province, French, and one, *Blair Winery* is in California, US, and *Emmerich Knoll* from Australia. The second chart is the top 10 types of grapes to make expensive wines. In French, especially in Bordeaux Province, most wineries use *Bordeaux-style Red Blend* grape to make wine, and these wines takes 7 places in the top 10 most expensive wines, and *Pinot Noir* grape (also in French) takes 2 places, Chardonnay grape, from US, take one place. The Australian wine made from grape *Gruner Veltliner* grape only takes the 14th place in the most expensive wines list.



Picture 10 & 11. Top 10 Most Expensive Wineries & Top 10 Varieties of Grapes

So, for high-ending consumers, we recommend that they could choose French wineries, and wines made from *Bordeaux-style Red Blend* grape. Also, make sure about the winery. Maybe when visiting France, visitors could go to these wineries who make wines of whopping price for a visit!

Future Scope:

- Find a better model to predict the price of wines more accurately.
- Create an App that would help the consumers to evaluate the price.
- Based on the consumer base, we can gain profit by advertising.