Wine Price Prediction

Hua-lai, Tseng(Justin)
Springboard Capstone Project 1
Oct -10th, 2019

Purpose of Wine price prediction

- ***To discover if there are correlated factors, between wine price and other features of a bottle***
  - *Is the point of wines correlated to their price?*
  - *Are year, country and grape variety significant factors that might affect the price?*

- ***To know the true value of a bottle***
  - *If a certain aroma, flavor, and certain grape variety of wine could easily achieve a higher price, it could be a good reference for winemakers interested in targeting higher profit or market.*
  - *To pick up some best value bottle when entering a Wine & Liquor.*

Methodologies

- ***Data Preprocessing:***
  - *Extract year, red_or_white columns information from existing columns.*
  - *Drop null and meaningless values, identified outliers, and organize variables in each columns.*
- ***Statistics Inference & Story telling:***
  - *Select columns to see if there exist any correlations between columns and price.*
  - *Interpret the correlation and in finding into stories.*
- ***Machine Learning:***
  - *Use linear regression model to predict the price of wine.*
  - *Use Advance regression model to predict the price of wine.*
- ***Data Source:***
  - *The data file in this project are acquired from Kaggle at: https://www.kaggle.com/zynicide/wine-reviews#winemag-data-130k-v2.csv, which was scraped from WineEnthusiast during the week of June 15th, 2017.*

Data Wrangling

- ***Deal with missing values:***
  - *Drop region_2 column since it contains more than 40 % null value*
  - *Drop any rows in price column(target variable) with nan values.*
  - *Fill or drop rows with missing values in other columns.*
- ***Introduce new variables:***
  - *Year: From designation column extract information of vintage of the wine.*
  - *Red_or_white: By initializing a dictionary of grape varieties for red or white wines are made out of, create a new columns to identified if the bottle is red or white.*
- ***Outliers and organize each column:***
  - *For columns like country and variety, only keep the variables with top occurrence.*
  - *Analyze and drop values in each columns if there are abnormal or meaningless values.*

Cleaned Data for Analysis purpose

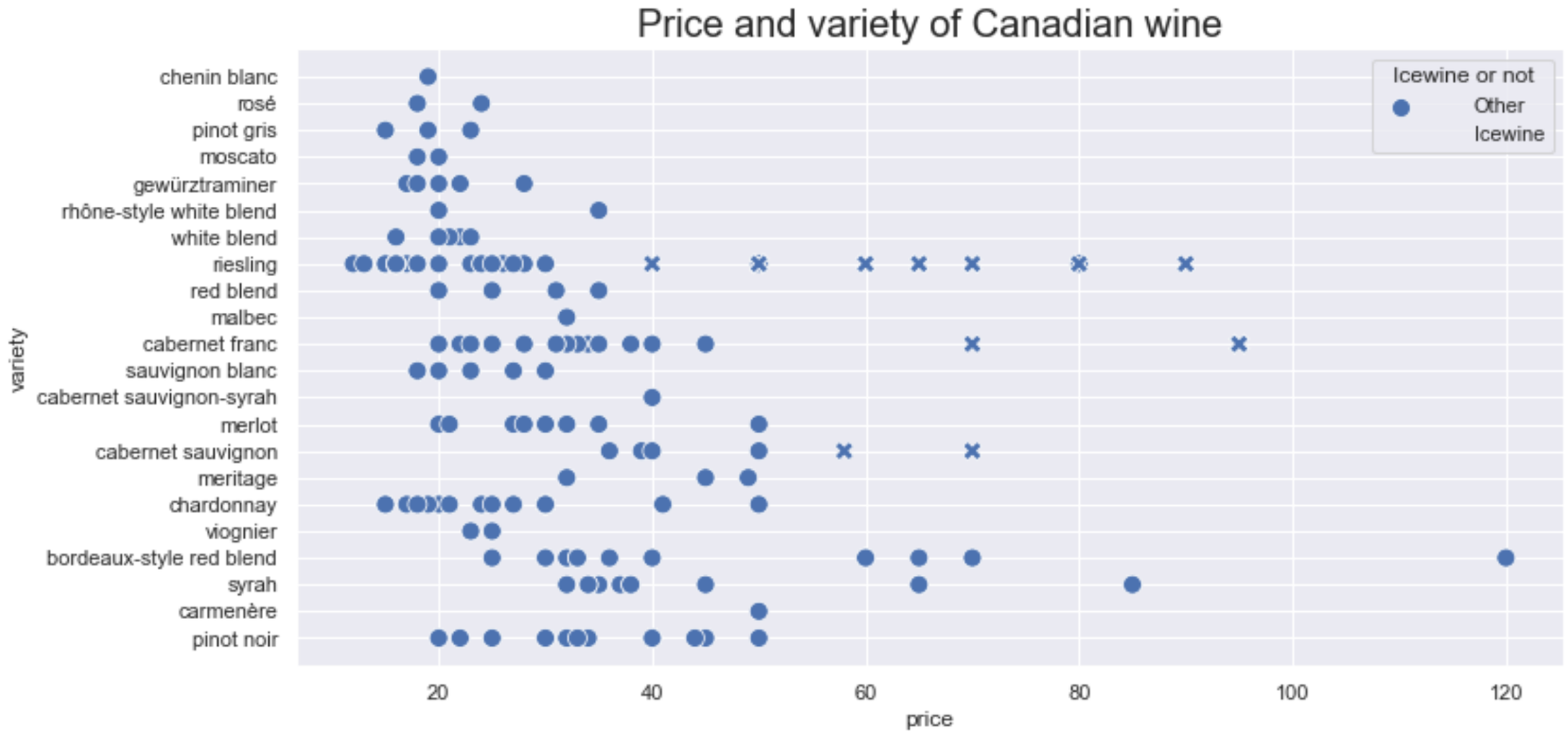
	price	country	points	variety	year	red_or_white
0	18.0	Spain	82	cabernet sauvignon	2008	r
1	35.0	US	81	grenache	2009	r
2	21.0	US	85	chardonnay	2006	w
3	21.0	US	83	rhône-style white blend	2006	w
4	22.0	US	84	viognier	2010	w
5	18.0	US	86	moscato	2015	w
6	18.0	US	89	red blend	2012	r
7	18.0	US	84	sauvignon blanc	2014	w
8	19.0	US	91	zinfandel	2013	r
9	19.0	US	90	zinfandel	2014	r

EDA & Storytelling - Country and wine price

➤ *Country and wine price*

- *After Mann-whitney U test, since non of the above variables are normally distributed, the more expensive bottles are produce from US, Italy and Canada.*
- *Speculation: Is expensiveness of Ice wine the main reason to push the price of Canadian wine up?*

EDA & Storytelling - Canadian wine

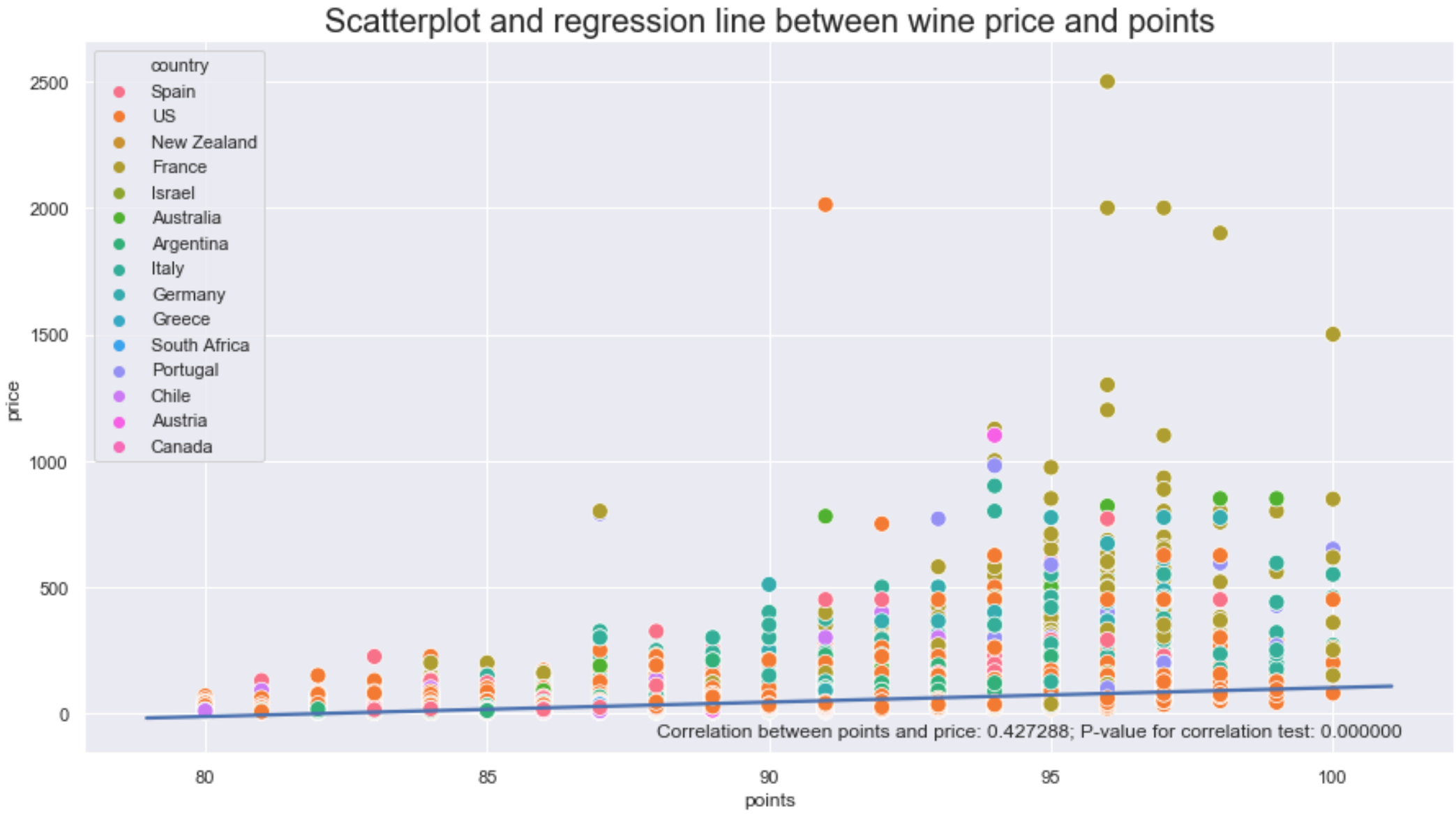



```
The amount of Canadian icewine: 21
Amount of Canadian icewine with price >= 40: 21
percentage of icewine among Canadian wine with price >= 40: 0.3684210526315789
Percentage of icewine among all Canadian wine: 0.0933333333333334
```

➤ *Canadian wine*

- *By looking at the result after analyzing Canadian wine, we can see that wine from Canada is, in general , very pricy, since Ice wine only weigh 9.3% from all the wine from Canada.*

EDA & Storytelling - Points and wine price I



➤ *Points and wine price I*

- *The result of the scatterplot shows that points and price have a positive correlation, which is not unexpected, though the correlation (0.427288) does not seem to be strong.*

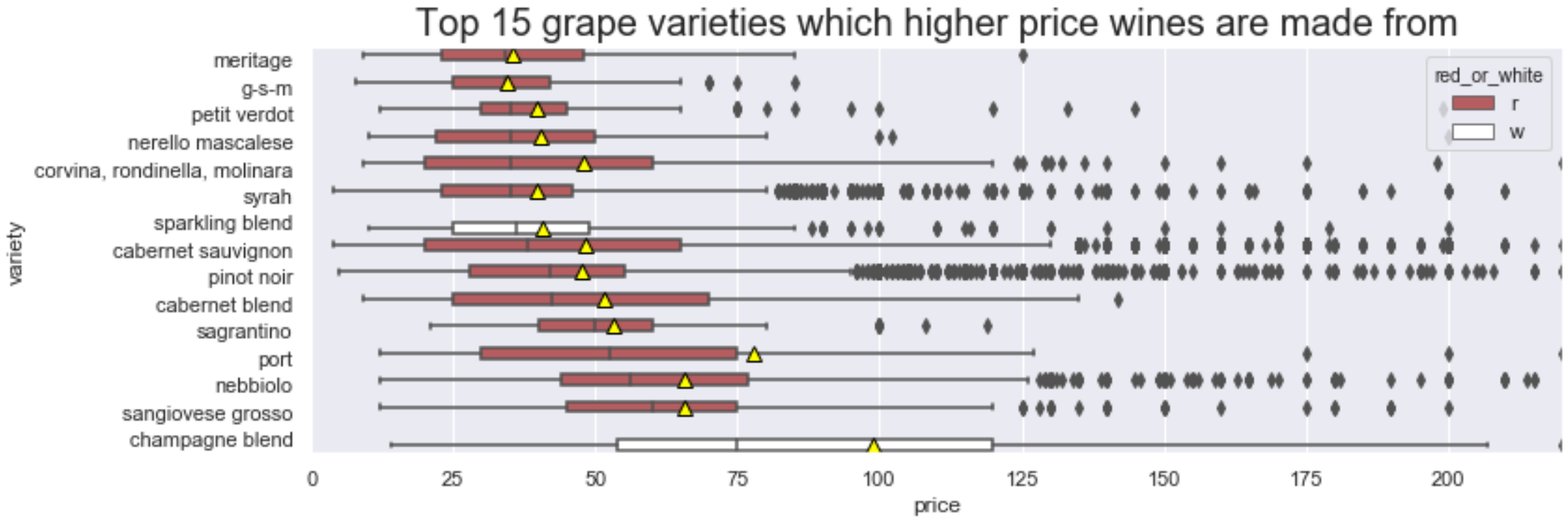
EDA & Storytelling - Points and wine price II

```
Average price for best wine (points > 98): 374.27906976744185 USD
Median price for best wine (points > 98): 268.0 USD
Minimum price to get a bottle which scored 100: 80.0 USD


Country counts for best wine (points > 98) :
US            16
France        11
Italy         11
Portugal       4
Australia      1
Name: country, dtype: int64

Percentage of US produced wine that have scored at least 99 in points: 0.372093023255813
```

➤ *Points and wine price II*

- *Winemag ranks wines from points 80 to 100, what is interesting is that only a few bottles make it to the top 10% in scoring (points 99 & 100). By looking into the top tier, I found some interesting facts. First, the median price for the best wine (points 99 & 100) is approximately 268 USD. Second of all, you can get a perfectly ranked wine (points 100) with only 80 USD. Finally, the country that produces the most of these best bottles (points 99 & 100) is the US, which weighed 37.2% among the best wines.*

EDA & Storytelling - Variety and wine price



➤ *Variety and wine price*

- *Extracting out 15 most expensive grape varieties, we can see that red wines dominant this tier, the only two white wine grape variety that makes it to the top band are both sparkling wine. After conducting several statistical tests, we can conclude that red wine is more expensive than white wine.*
- *From the print out of the most expensive grape varities and there dominant production country, we can see that Italian endemic grape varieties(nerello mascalese, corvina & rondinella & molinara, sagrantino, nebbiolo, sangiovese grosso) produce high price.*
- *By looking at the boxplot of grape variety, the champagne blend has the highest unit price, this is foreseeable due to the rigorous and complexity of the production procedure. It is also clear that red wine's grape variety is more expensive than white wine's grape variety.*

EDA & Storytelling - Year of production and wine price I (before year 1980)


```
Correlation between year & price for Port produced before 1980: -0.7875162835677858
P-value for correlation: 0.00017484716784353833
```

➤ *Year of production and wine price I (before year 1980)*

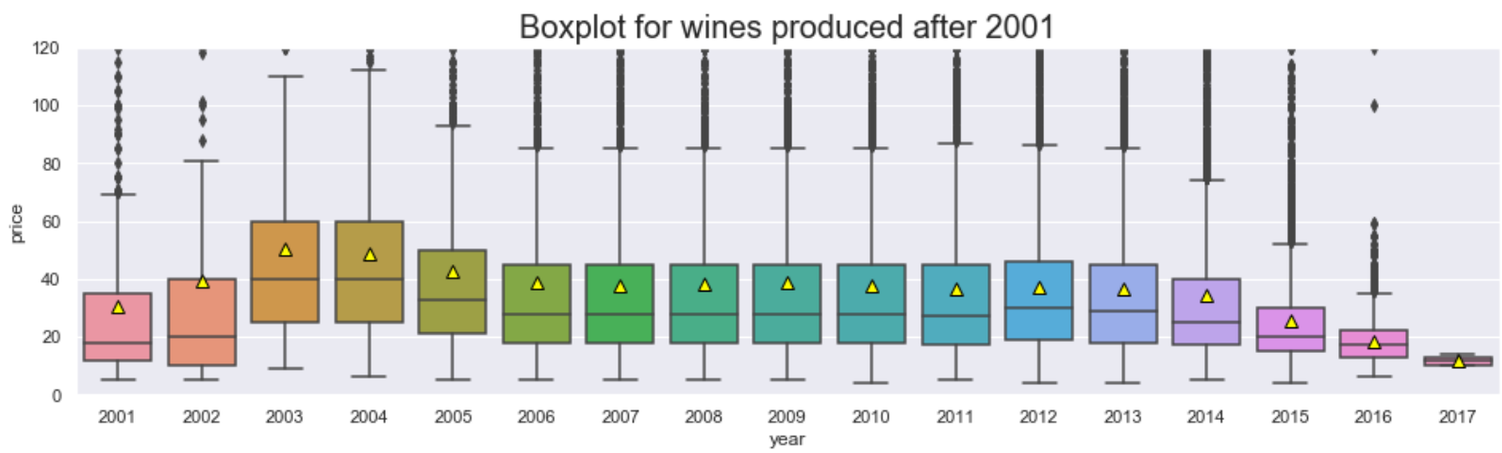
- *By printing out info for wine produce before 1980, I realized that most of them are port, after plotting out the regression line for port, we can see that there exist a strong correlation between amount of years port aged and price of port produced before 1980.*

EDA & Storytelling - Year of production and wine price II (between year 1981 and 2000)

➤ *Year of production and wine price II (between year 1981 and 2000)*

- *By plotting out the scatterplot for red wine and white wine produced between 1981 and 2000, we can see that there exist weak correlations between year and price for both kinds of wine. We can also see that price of wine production within this time frame do not go higher than 420USD per bottle, and white wine contributes 72% to the higher price wine (>= 200USD). By printing out the info of higher price white wines, we can see that champagne is one of the major variety for high price wine produced in this time period.*

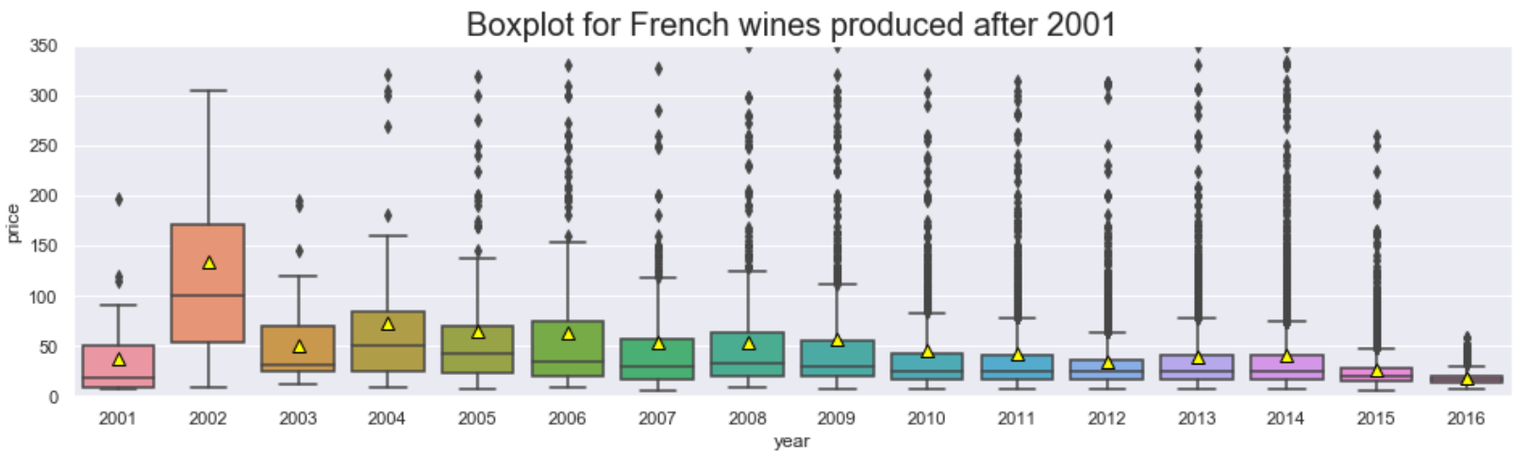
EDA & Storytelling - Year of production and wine price III (after year 2001)

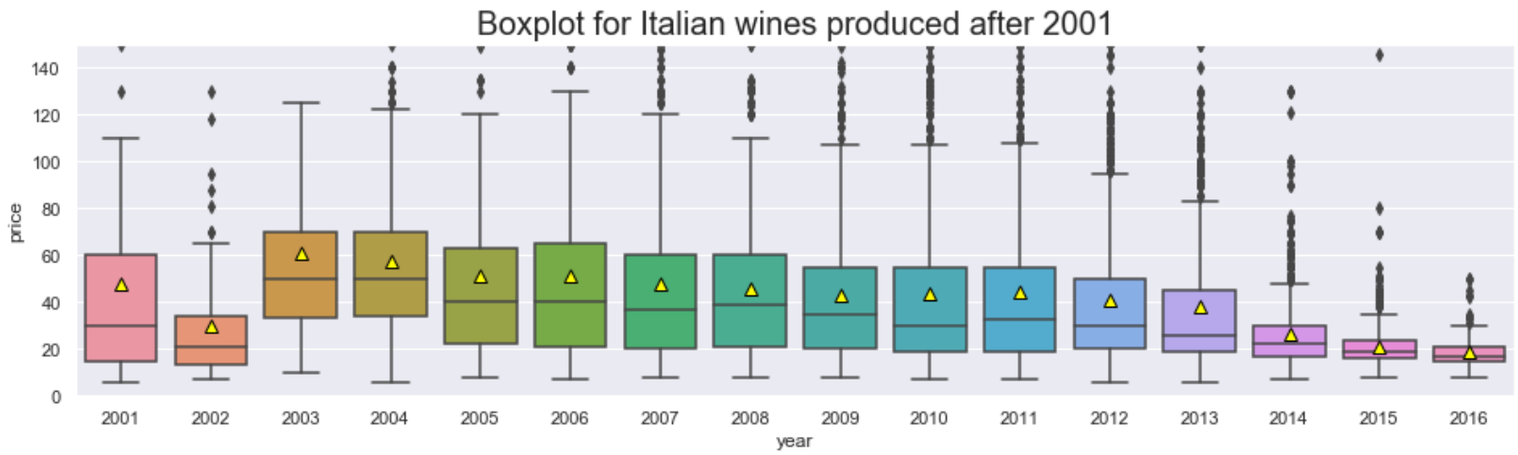


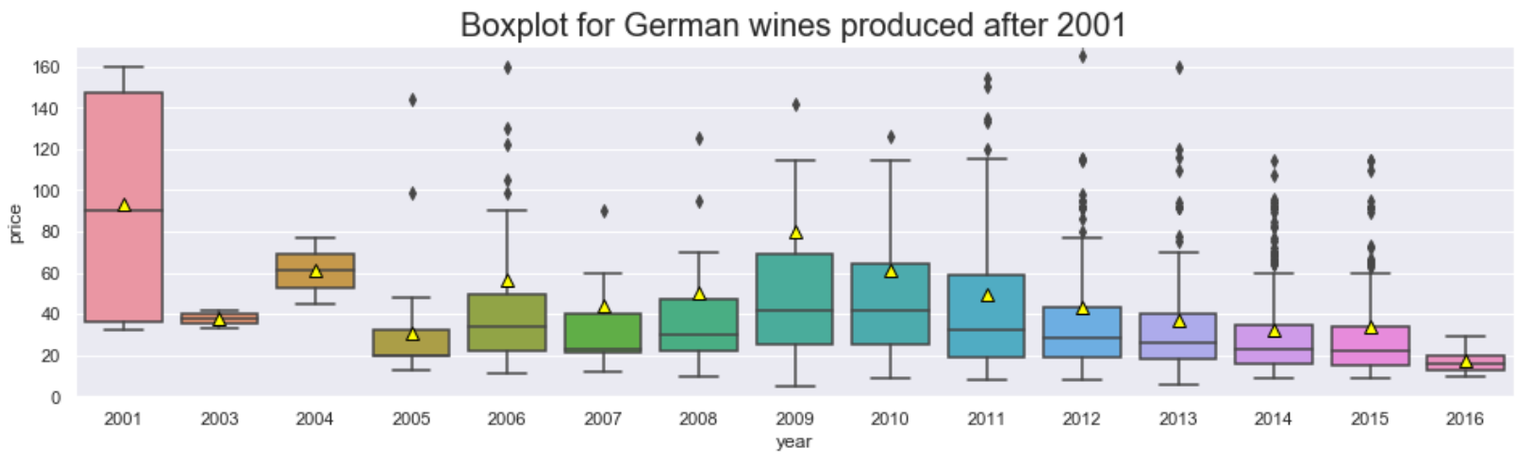
➤ Year of production and wine price III (after year 2001)

- *From the first boxplot, we cannot really see a correlation between year and price, but certain year seems to have a higher average price such as 2003 and 2004. From the markers of the scatterplot below, I realized some countries have higher wine price in certain year, so I decide to plot out the price and year boxplot for each countries.*

EDA & Storytelling - Year of production and wine price III (after year 2001)







➤ *Year of production and wine price III (after year 2001)*

- *From the boxplot of French wines produced after 2001, we can see that year 2002 seems to have the highest price in this timeframe. By printing out the info of each year, we can see that the abnormal high average price of 2002 is due to lack of amount of data. And this is also the reason why other countries have extreme high or low price in certain year. If we only look at the year with sufficient amount of data for each country, we can conclude that year is not a strong factor that will affect the price in this time period.*

Machine Learning - Linear Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.210
Model:                            OLS   Adj. R-squared:                  0.210
Method:                 Least Squares   F-statistic:                     5829.
Date:                Sat, 28 Sep 2019   Prob (F-statistic):               0.00
Time:                        13:19:18   Log-Likelihood:            -5.5041e+05
No. Observations:              109875   AIC:                         1.101e+06
Df Residuals:                  109869   BIC:                         1.101e+06
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -408.1479      3.371   -121.066      0.000    -414.756    -401.540
points         5.7079      0.036    158.748      0.000       5.637       5.778
year          -1.2214      0.031    -39.873      0.000      -1.281      -1.161
variety       -0.1135      0.005    -25.040      0.000      -0.122      -0.105
country       -0.0028      0.025     -0.111      0.911      -0.051       0.046
red_or_white  -7.7103      0.236    -32.699      0.000      -8.172      -7.248
==============================================================================
Omnibus:                   234432.810   Durbin-Watson:                   1.321
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       2873703699.986
Skew:                          18.475   Prob(JB):                         0.00
Kurtosis:                     794.415   Cond. No.                     3.36e+03
==============================================================================
```

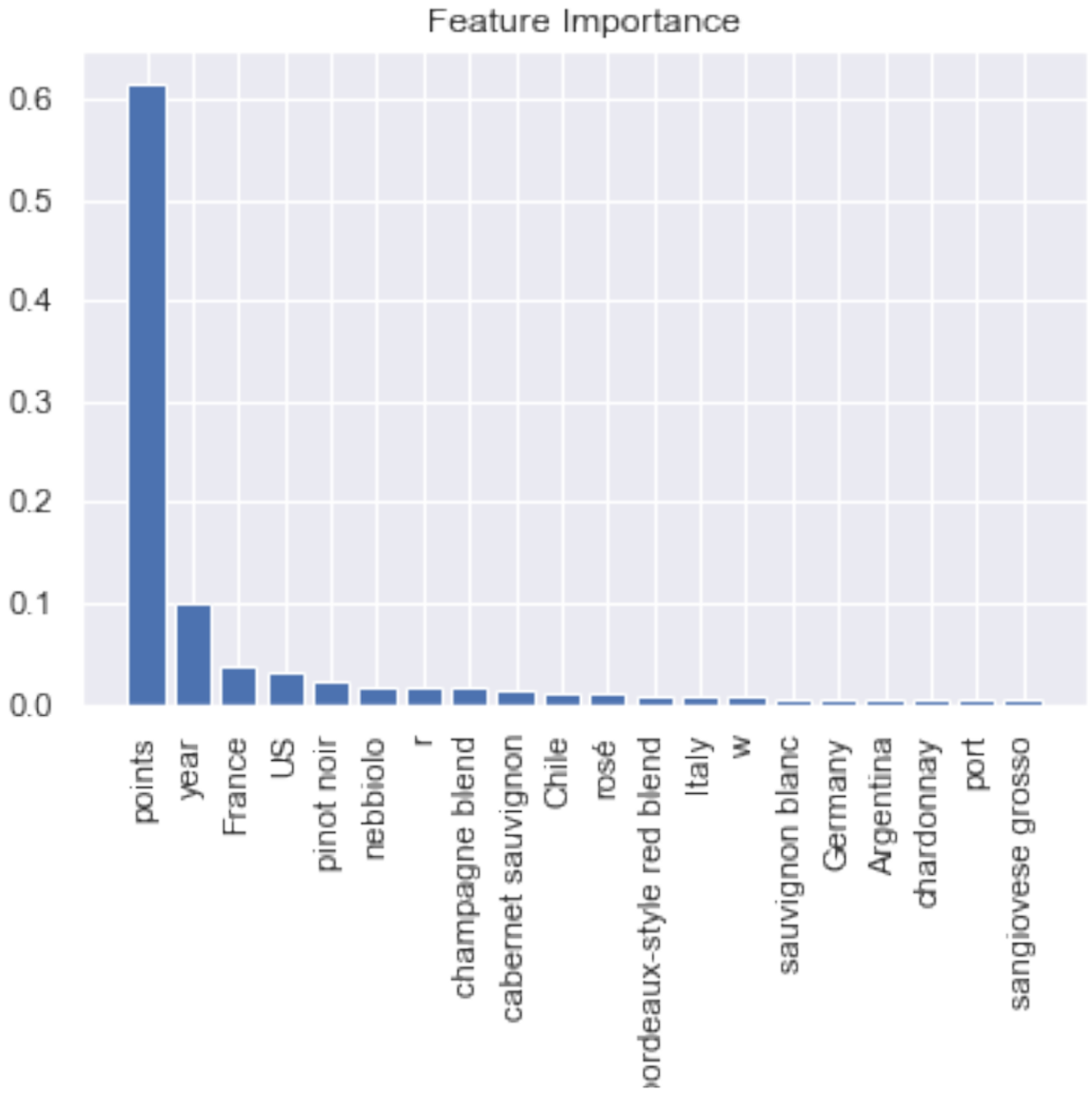
```
                                      MSE_train R^2_train MSE_test  R^2_test
Linear Regression(label encode) Ridge   1398.86  0.208191   977.55  0.253191
Linear Regression(one hot encode) Ridge 1348.18  0.237826  930.599   0.28906
```

➤ *Machine Learning - Linear Regression*

- *Compare with label encoded data, one hot encoded data have better linear regression performance in general. One of the interesting finding is that the model will perform really bad if we one hot encoded the year column. Another interesting finding is that the score of one hot encoded data dropped after we scale the data using pipeline.*

- *For one hot encoded data, ridge regression perform slightly better compare to other methods here, with MSE: 930.6, R^2: 0.28906*

Machine Learning - Advance models



- ***Machine Learning - Advance Models***

  - *All the more advanced regressors generate better result compare to linear regression. Among all of them, Gradient boosting method with a MSE: 705.62 and R^2: 0.46 perform the best. However, Catboosting also conduct an impressive result within a relatively short amount of time. For the advanced methods, the top 5 important features for them to make decisions are: points, year, France, US, pinot noir.*

	MSE_train	R^2_train	MSE_test	R^2_test
Linear Regression(label encode) Ridge	1398.86	0.208191	977.55	0.253191
Linear Regression(one hot encode) Ridge	1348.18	0.237826	930.599	0.28906
RandomForestRegressor	1119.28	0.356199	758.734	0.420358
GradientBoostingRegressor	1047.03	0.410215	705.62	0.460935
CatBoostRegressor	1051	0.404586	736.256	0.437531
XboostRegressor	1054.88	0.402816	766.288	0.414587

Future Work

- ***Extract more hidden variable from original data columns, for example, turn year columns into three columns before machine learning.***

- ***Combine the original data with other wine or weather related data to improve the score.***

- ***Try out other models or tune the parameters in Gradient Boosting to see if the score can be improved.***

Summary

- ***Wine price pattern:***
  - *Among all five columns, points of wine seems to have the strongest correlation with price of wine.*
  - *Countries that produce more expensive bottle of wines are: US, Italy, Canada.*
  - *Red wine have higher price than white wine in general, sparkling wines are the most expensive white wine type.*
  - *Year do not really affect price of wine, except for bottles produced before 1980(Porto are majority), which have a strong correlation between years aged & price.*
  - *Italian endemic grape varieties(nerello mascalese, corvina & rondinella & molinara, sagrantino, nebbiolo, sangiovese grosso) produce high price.*

- ***Machine Learning:***
  - *Points, year, France, US, pinot noir are the top important features to predict price of wine.*
  - *Winning model: Gradient Boosting with MSE: 705.62 and R ^ 2: 0.46.*

Acknowledgement

- ***Mentor(Sorted_value): Dipanjan (DJ) Sarkar, Harsh Singh , Kenneth Gil-Pasquel, Max Sop***

- ***Springboard***

- ***Kaggle***