Capstone Project 2 Milestone Report – Wine Reviews

Client:

The client here is a Wine Enthusiast that is looking to predict an unknown wine variety given the description about the wine made by a wine expert, just like how a wine expert is able to tell a wine variety using his or her five senses.

Problem:

Given several input variables such as price, description, region, etc, is it possible to create a model to predict which wine variety does the unknown wine belong to? What machine learning technique would be best suited to achieve highest accuracy for the testing data?

Dataset:

The data contains 12 fields where there are nine independent variables, and one dependent variable. The data itself consists of around 130,000 rows of observations.

The 11 independent variables include:

- Points: The number of points a person rates the wine on a scale of 1-100
- Title: The title of the wine review
- Description: A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.
- Country: The country that the wine is from
- Province: The province or state that the wine is from
- Region1: The area where the wine is grown in a province or state
- Region2: The more specific area about where the wine is grown, can have missing values
- Winery: The winery that made the wine
- Designation: Vineyard within the winery where the grapes that made the wine are from
- Price: Cost of a bottle of wine
- Taster Name
- Taster Tweeter Handle: Taster's twitter account name

Dependent Variable: Wine Variety, which is the type of grapes used to make the wine for example: Pinot Noir.

Data Wrangling:

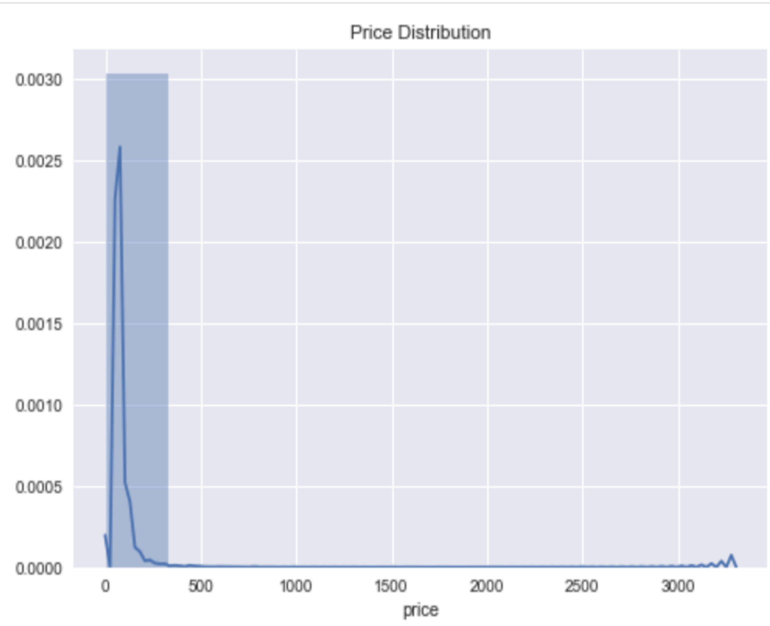
There are several problems in this dataset, which include: Duplicate Values as well as missing values in multiple rows. Thus, there are several steps that need to be taken to clean the data before being able to analyze it.

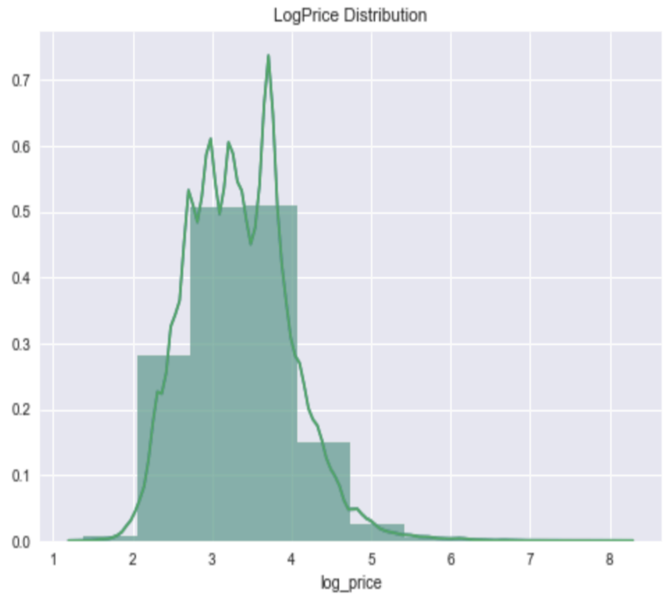
- Duplicate Values: Since there are around 20,000 duplicates, it shows that our data is actually not as many as what we though it should be. Thus, we have to remove all of the duplicated data since it will distort our analysis by double counting.

- Missing Values: I will talk about how I handle missing values in different columns.
  - Country: For countries that are missing, it turns out that there are only 63 out of them, which is only 0.05% of the entire dataset, so I simply drop all those rows.
  - Price: Since there are quite a significant number of rows that have missing values for price, I figured out the average price a wine costs in each country and fill in the missing values for the wine based on which country it is from
  - Region 1: For region 1 that has missing values, I figured out that it would skew the data and possibly over emphasize a certain area if I replace missing values with the area, I decided to change them to unknown
  - Region 2: More than 50% of the data has missing values for this column. It would not be possible for me to simply drop all of the rows. Thus, I decided to simply exclude this variable from the analysis as the main problem aims to predict wine variety based on its description and this is only an additional feature.
  - Taster Name: There are a lot of missing values as well for this variable, thus I decided to exclude this variable as well.
  - Taster Tweeter Handle: I decided to exclude this variable as well since there are a lot of missing values and this variables is not the main variable in question.

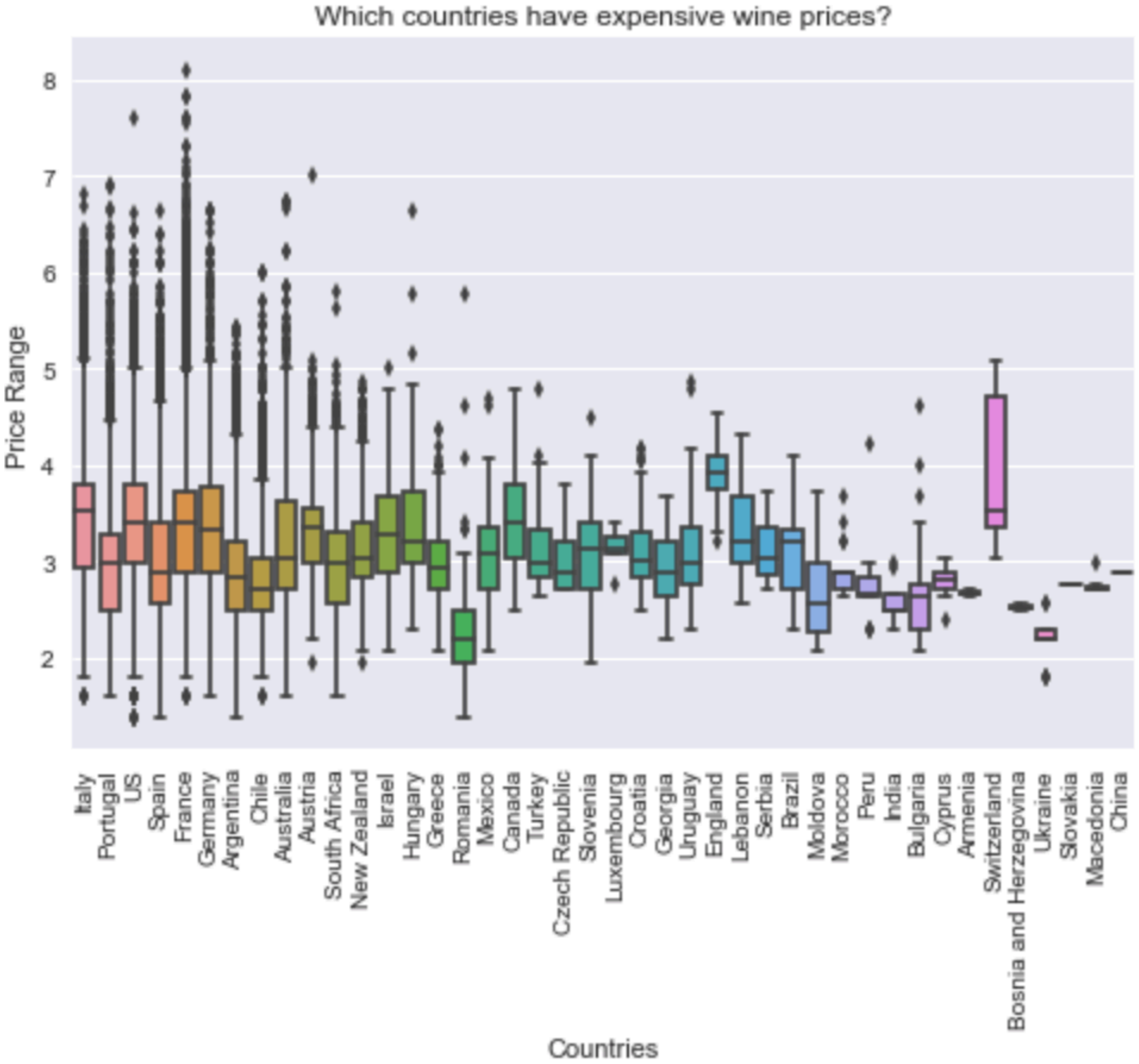
Exploratory Data Analysis:

Price: As expected, we can see that the price of wine is very right-skewed in this case and most prices are centered around the $0-$300 range. For this, I decided to take a log out of it to make the distribution more normal, as seen on the graph on the right.

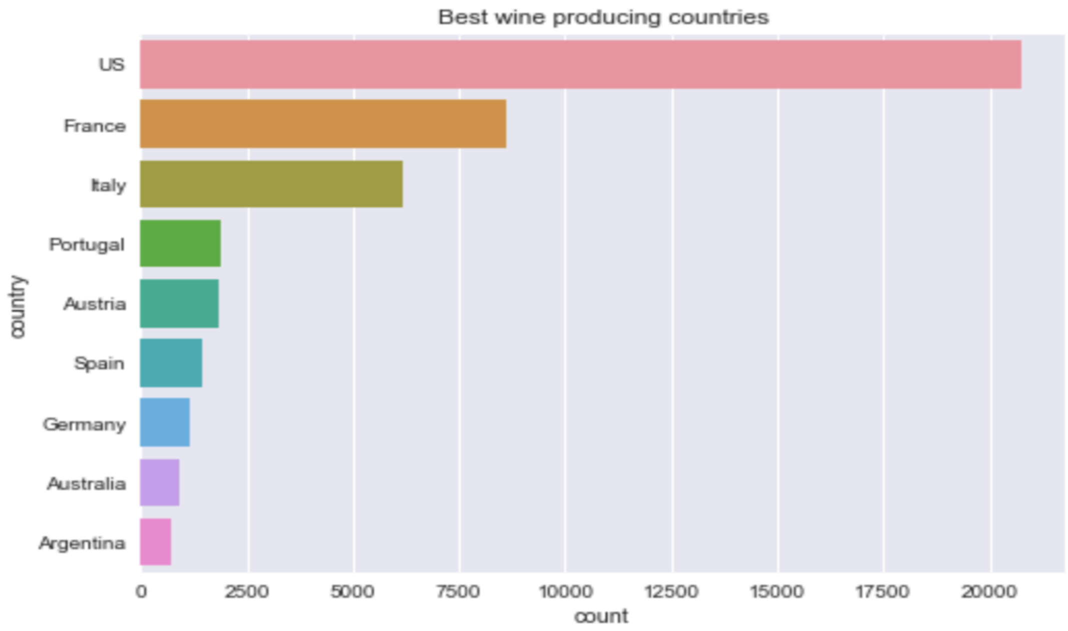




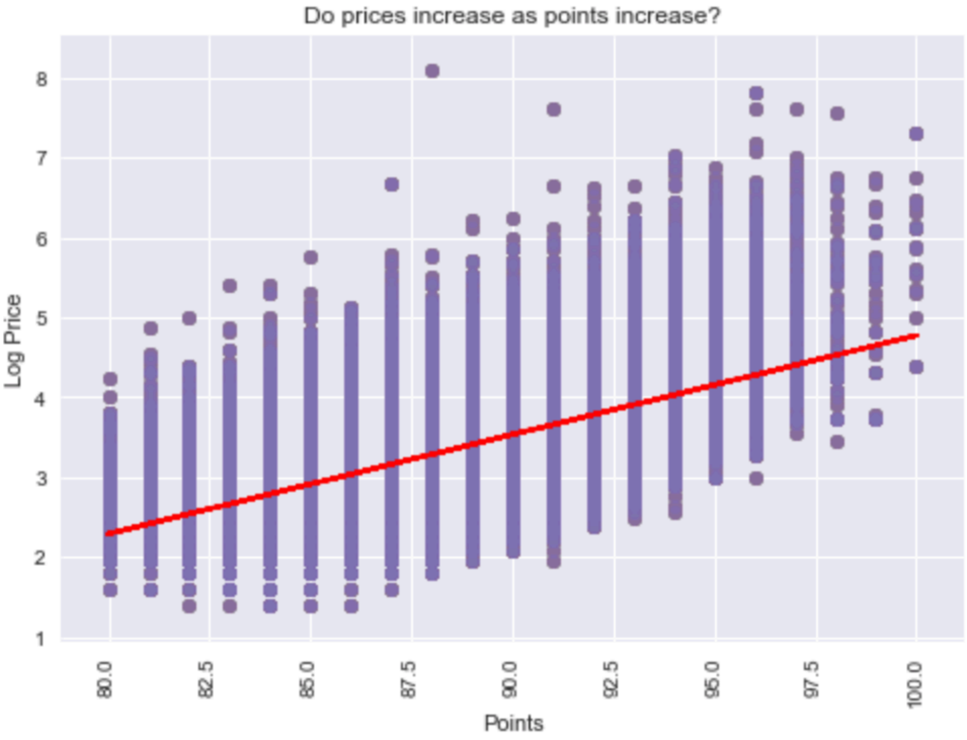
Wine Prices across Countries: From the graph of boxplot below, we can see that China, Macedonia, Slovakia, Bosnia, and Herzgovina, and Switzerland a large range of wine prices. The country with the highest average price of wine is England, followed by Italy, and Switzerland. Countries with lowest average prices of wine include Ukraine, Romania, India, and Moldova.



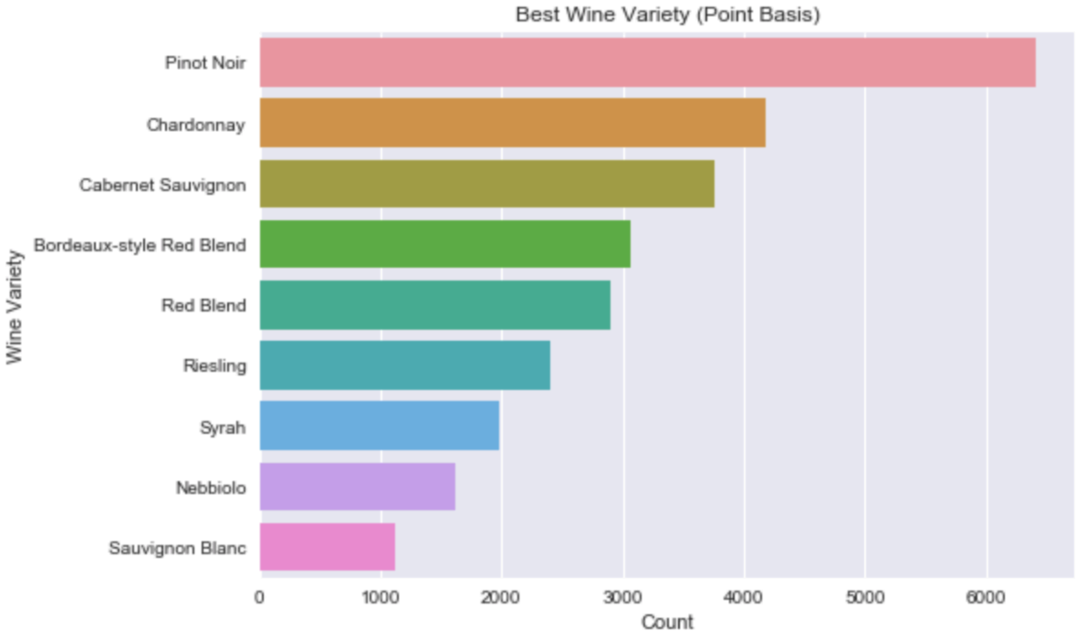
Top 10 countries with high count of high-points wine: The bar graph below shows that US tops the best wine producing countries list, followed by France, Italy, and Portugal. Let's see if there is any relationship between price and points.



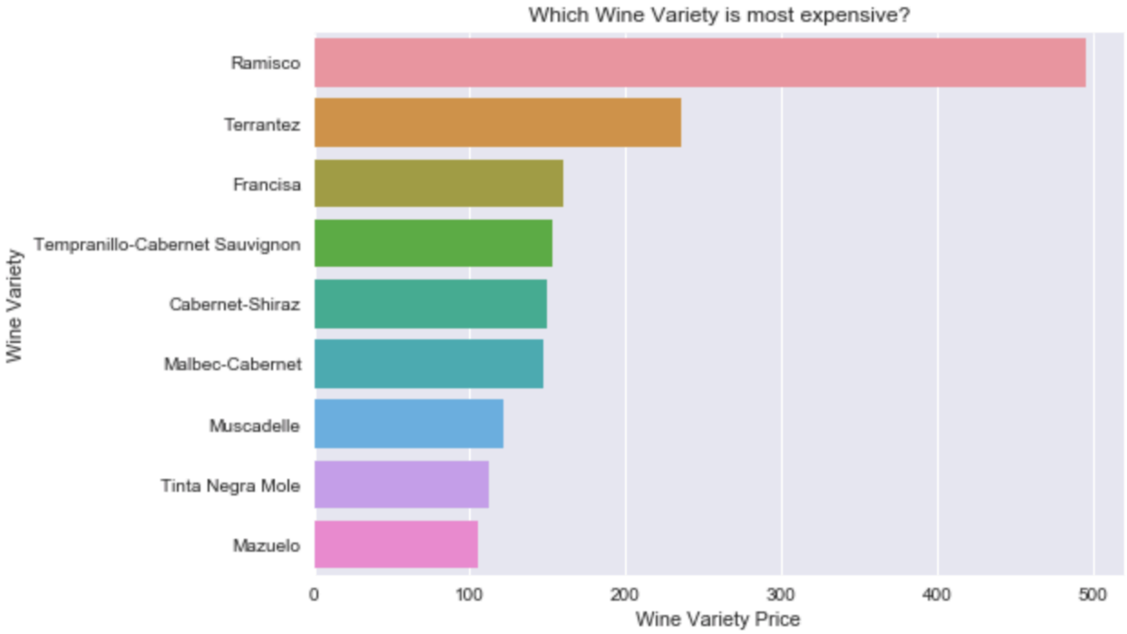
Price Vs Points: The graph shows that there is a linear relationship between price and points. As point increases, the price of a wine increases as well.



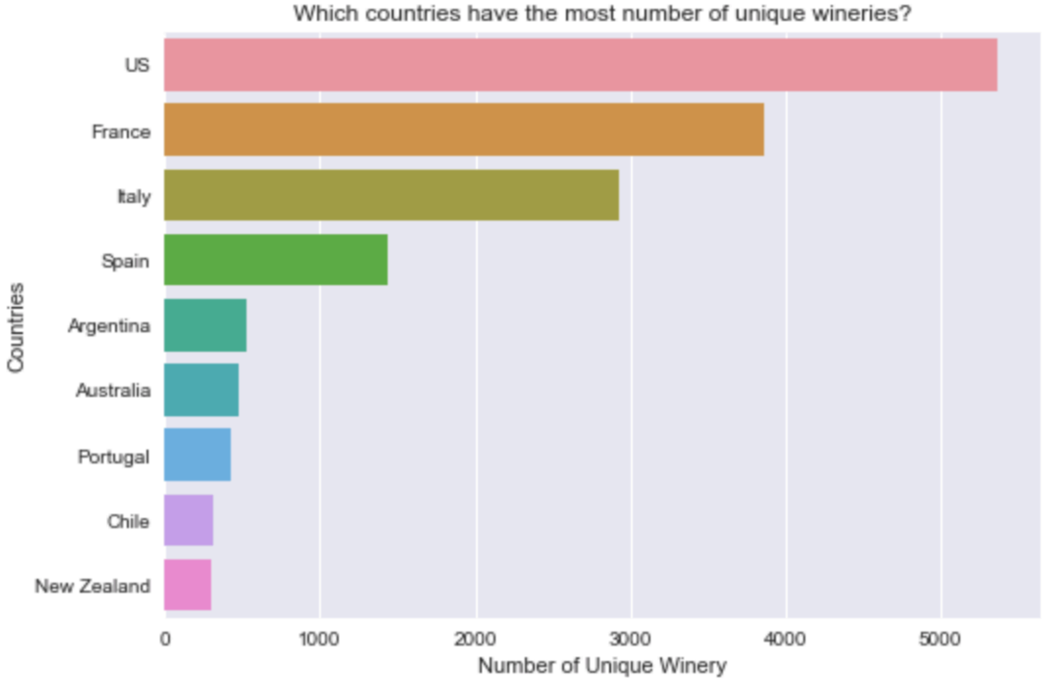
Best Wine Variety based on Points: The wine variety that is deemed to be high quality based on points given by wine experts is Pinot Noir. Given the list of best wine variety, let's see if these wine also costs the most.



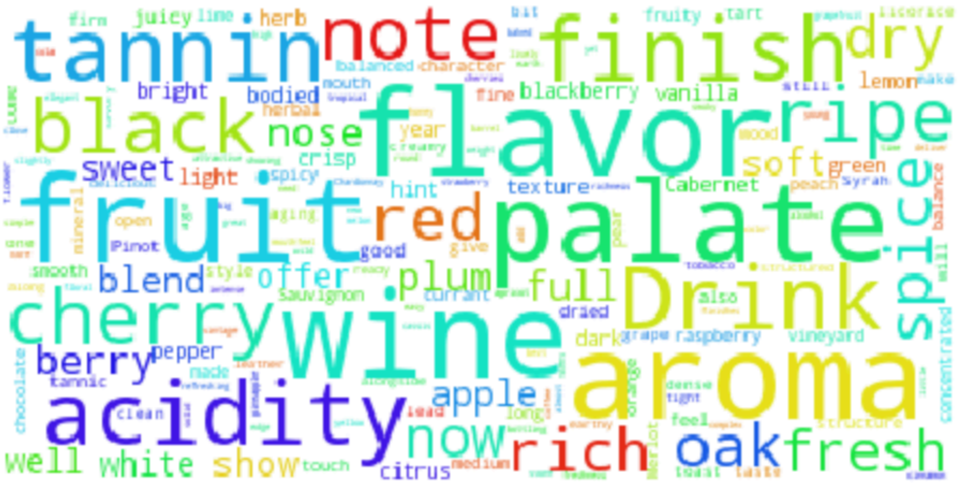
Which wine variety is most expensive? The graph below shows that Ramisco is the most expensive wine variety, and wine variety that made into the top 10 most expensive wines are not consider of high quality, which is weird because we saw that points seem to have a linear relationship with points. However, that does not seem to apply to the top 10 list but more towards the average.



Which Countries have most number of Unique Wineries? US topped the list of countries that have most number of unique wineries, followed by France, Italy, and Spain. The list here seems to be similar to the countries that made into the list of 'Best Wine Producing Countries'. Perhaps, there is a correlation between the quality of wine and number of wineries there are in a country.



Popular Words Used by Wine Experts: In describing the wine experts taste, turns out that most wine experts use the words flavor, fruit, palate, aroma, and acidity. The wordcloud below shows what kind of words are most commonly used by wine experts.


tannin
note
finish
dry
firm
juicy
lime
herb
fruity
tart
licorice
black
bright
bodied
herbal
mouth
balanced
character
fine
blackberry
vanilla
lemon
flavor
ripe
nose
crisp
year
creamy
soft
green
Syrah
sweet
light
spicy
hint
texture
Cabernet
peach
fruit
red
palate
spice
balance
open
Pinot
aging
good
give
pear
mineral
smooth
blend
style
offer
plum
full
Drink
Sauvignon
currant
dried
cherry
wine
dark
grape
raspberry
vineyard
berry
pepper
made
tannic
aroma
orange
chocolate
concentrated
acidity
clean
apple
lead
long
dense
tight
feel
structure
well
white
show
touch
citrus
now
medium
rich
Merlot
oak
toast
taste
fresh