Analysis of Wine Popularity and Suggestion to Customers for Wine Selection

Abstract

Why wine is so popular among drinks? One of the main reasons is that a favorite wine can be associated with a particular time, place, or memory. That's the thing with great food and great wine — it is woven into our life experiences. When it comes to wine, the taste, the bouquet, or simply the wine-drinking experience itself can elicit a memory of a time or place. In addition, wine makes food taste better, wine tells a story about where it comes from, wine is tradition and history, it enhances special occasions, wine improves health and so on. Based on these appealing points, I decided to implement the analysis of wine popularity and try to figure out suggestions to wine buyers when they select wine. In this project, I will do it in that pattern, which is commonly used for solving this type of problems: data wrangling, data visualization, feature engineer, model training and feature exaction.

Introduction

My project goal is to find out why wine is so popular in the world through data analytics perspective. In other words, I want to figure out which feature or features influence(s) wine rating points most. I collected dataset from Kaggle, which is the world's largest community of data scientists and machine learners. The dataset is made up of 13 features (2 numerical features and 11 categorical features) and 129,970 records. After loading this dataset as csv file, I did data wrangling by using Pandas, Numpy and missingno library to deal with duplicated recording removal, missing value removal and null object column removal. The next step is to experiment data visualization by importing library such as pyplot, seaborn, squarify and so on.

These library helps me to figure out how these variables affect each other and how their trends look like. In the third step, I did feature engineer by analyzing every single feature and had a general idea on which several features have influenced wine review. In the fourth step, from SKlearn importing Decision Tree Classifier, Random Forest Classifier, KNeighbors Classifier, Gradient Boosting Classifier and Cat Boost Classifier, I trained these models and calculated the accuracy. By comparing these accuracy, we went to the fifth step, which is to extract features affect wine popularity most. After finish coding project, I summarized suggestion how customers should pay more attention while selecting wine.

Related Work

Below links is the works which helped me prepare my project:

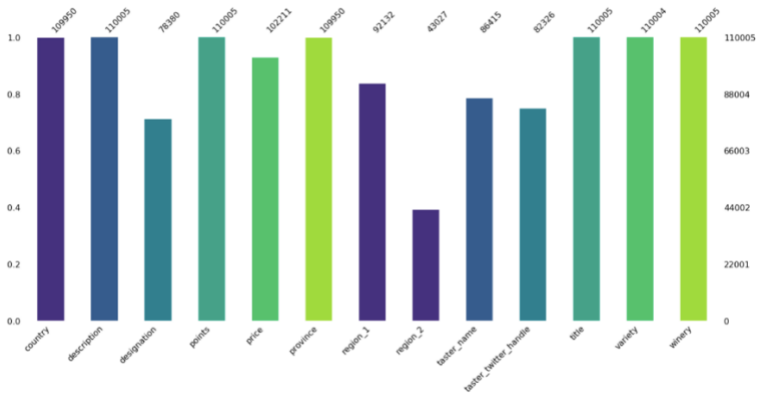
- "Plotting with Seaborn | Kaggle." *Countries of the World | Kaggle*, www.kaggle.com/bharatsingh213/plotting-with-seaborn/notebook.
- Towards Data Science. (2018). *How to Handle Missing Data – Towards Data Science.* [online] Available at: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4.
- Missing Values | Stata Learning Modules. [online] Stats.idre.ucla.edu. Available at: https://stats.idre.ucla.edu/stata/modules/missing-values.

Experiments

- **Data wrangling**

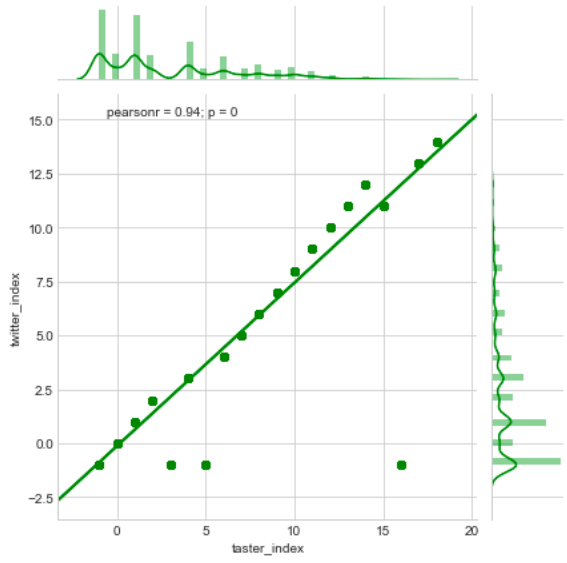
By loading Wine Review file as csv file and removing duplicated records, we found 110,005 unique rows. There is almost 20,000 duplicates records(129971-110005=19966). As below image shows, we found 9 columns missing values. As for columns "province", "country", and

"variety", their missing values percentages are less than 1%, so we can just drop these records. As for other 6 columns, we can use corresponding methods to process numerical column and categorical columns.



	MissingValue	Percent
region_2	66978	60.886323
designation	31625	28.748693
taster_twitter_handle	27679	25.161584
taster_name	23590	21.444480
region_1	17873	16.247443
price	7794	7.085132
province	55	0.049998
country	55	0.049998
variety	1	0.000909
winery	0	0.000000
title	0	0.000000
points	0	0.000000
description	0	0.000000

From the below plot we can find these two features are highly related. What's more, under common sense we know feature 'taster_twitter_handle' is feature 'taster_name' twitter accounts, so I'm going to drop one feature of these two. As column 'taster_twitter_handle' has more missing value, I'm going to drop it.



I use 'unknown' to replace missing value in these four columns which stand for another category among variables. This is the simplest way to impute categorical variables.

designation: The vineyard within the winery where the grapes that made the wine are 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.

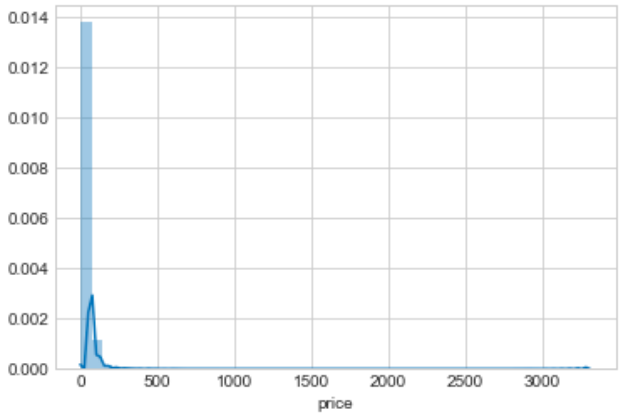
taster_name: Name of the person who tasted and reviewed the wine.

	country	description	designation	points	price	province	region_1	region_2	taster_name	title	variety	winery
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	25.0	Sicily & Sardinia	Etna	unknown	Kerin O'Keefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	unknown	unknown	Roger Voss	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos
2	US	Tart and snappy, the flavors of lime flesh and...	unknown	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm
3	US	Pineapple rind, lemon pith and orange	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	unknown	Alexander Peartree	St. Julian 2013 Reserve Late Harvest Riesling	Riesling	St. Julian

So far, we have impute all missing value and it is time to visualization our data.

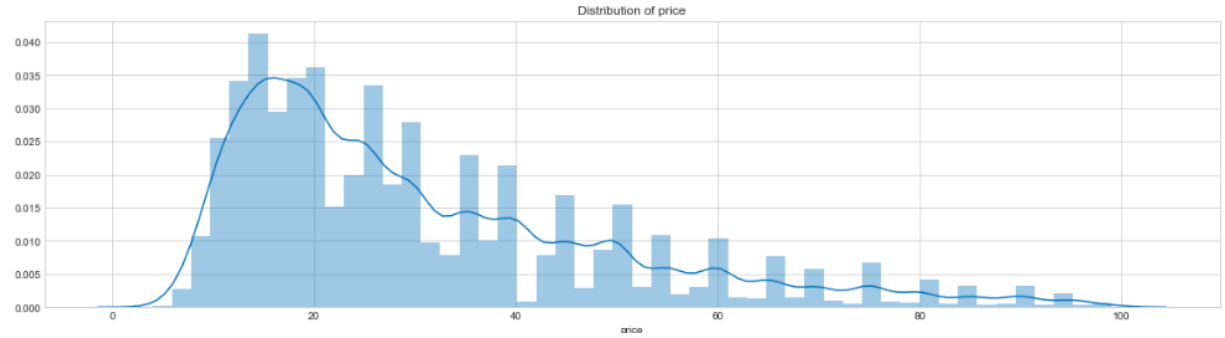
- **Data visualization**

Price

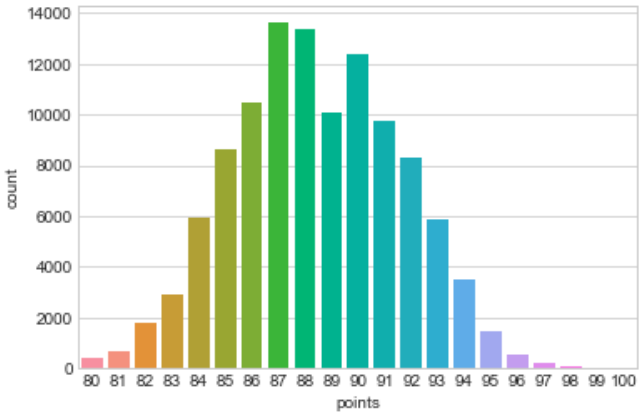


If we want to see better price distribution we have to scale our price or drop the tail. By dropping almost three percent of wines, we got the normal distribution whose prices are less than 100 usd. By visualize "price" in distribution plot and compare mean, median and mode value (mean is 35 and median is 25, which suggests a heavily skewed dataset with some outlier observations influencing the mean of the distribution), we fill the unavailable values with the median of the distribution.

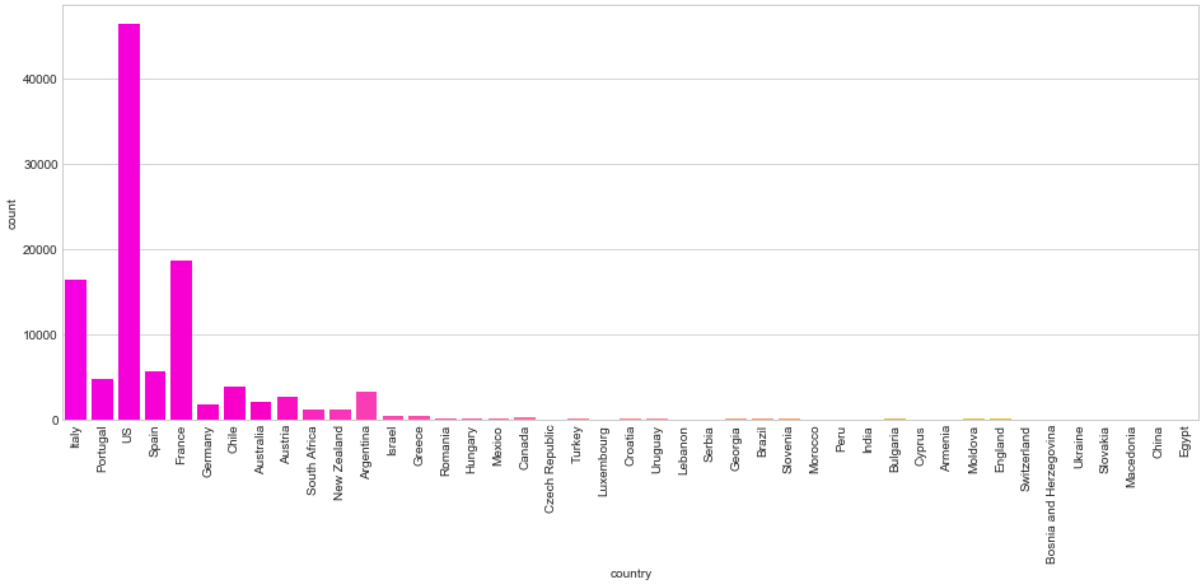
```
There are : 2.8122129350880862 % wines more expensive then 100 USD
```



Points



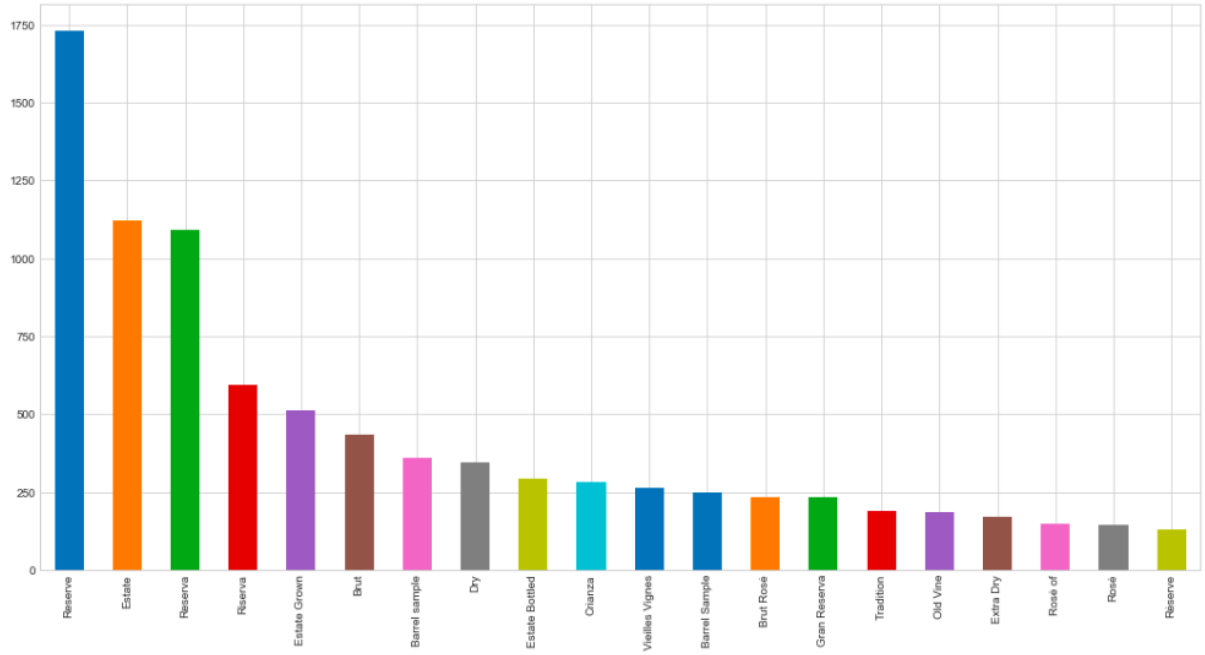
Country



Description

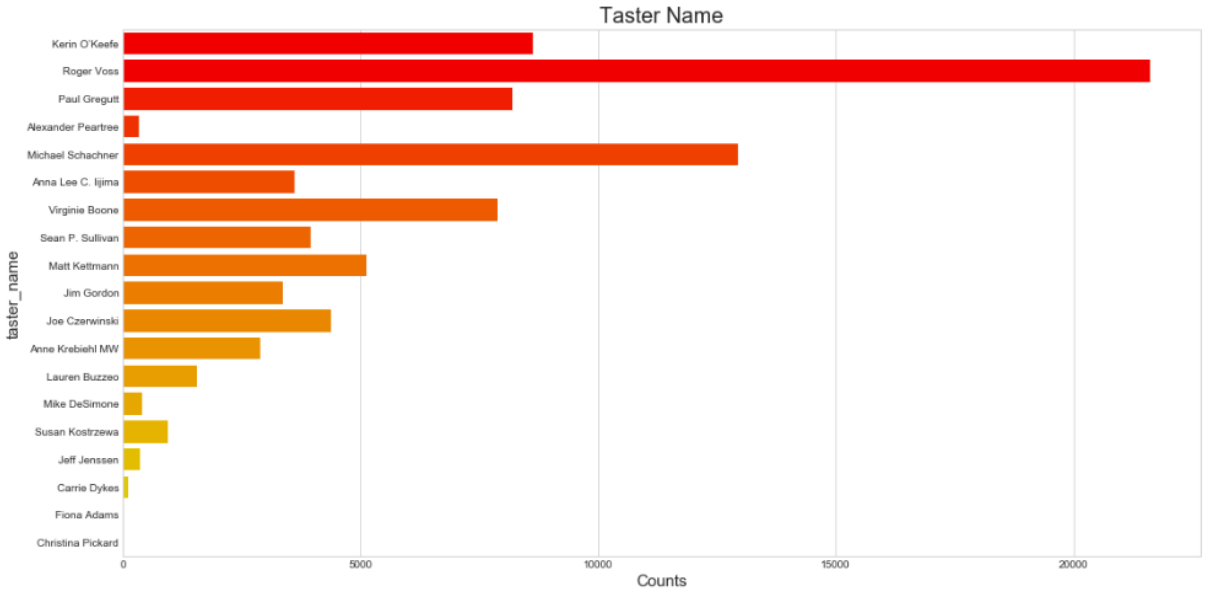
Dataset "Subdata1" is the subset of 'data' whose points are over average points 88. By visualizing the description, we can find the top 10 words which entered into our eyes are: "black", "cherry", "flavor", "drink", "now", "black", "note", "finish", "bodied" and "full" .

Designation



Above table shows top 20 most frequent designation (where grapes are planted). The first one is Reserve and count almost 1750.

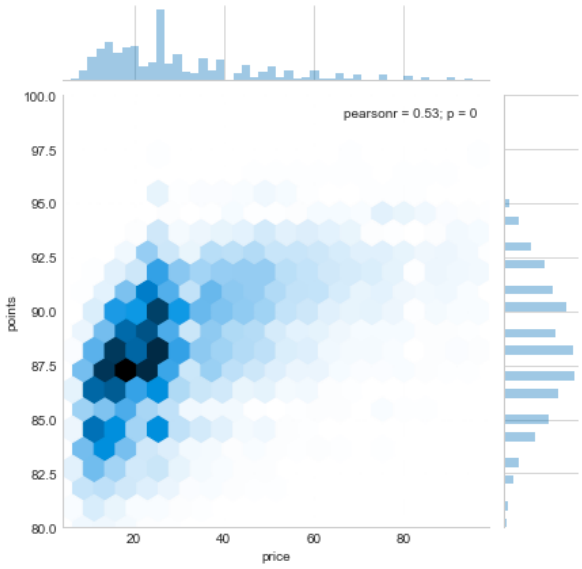
Taster Name



Tasters Roger Voss, Michael Schachner and Kerin O'Keefe rated most wines.

- **Feature engineer**

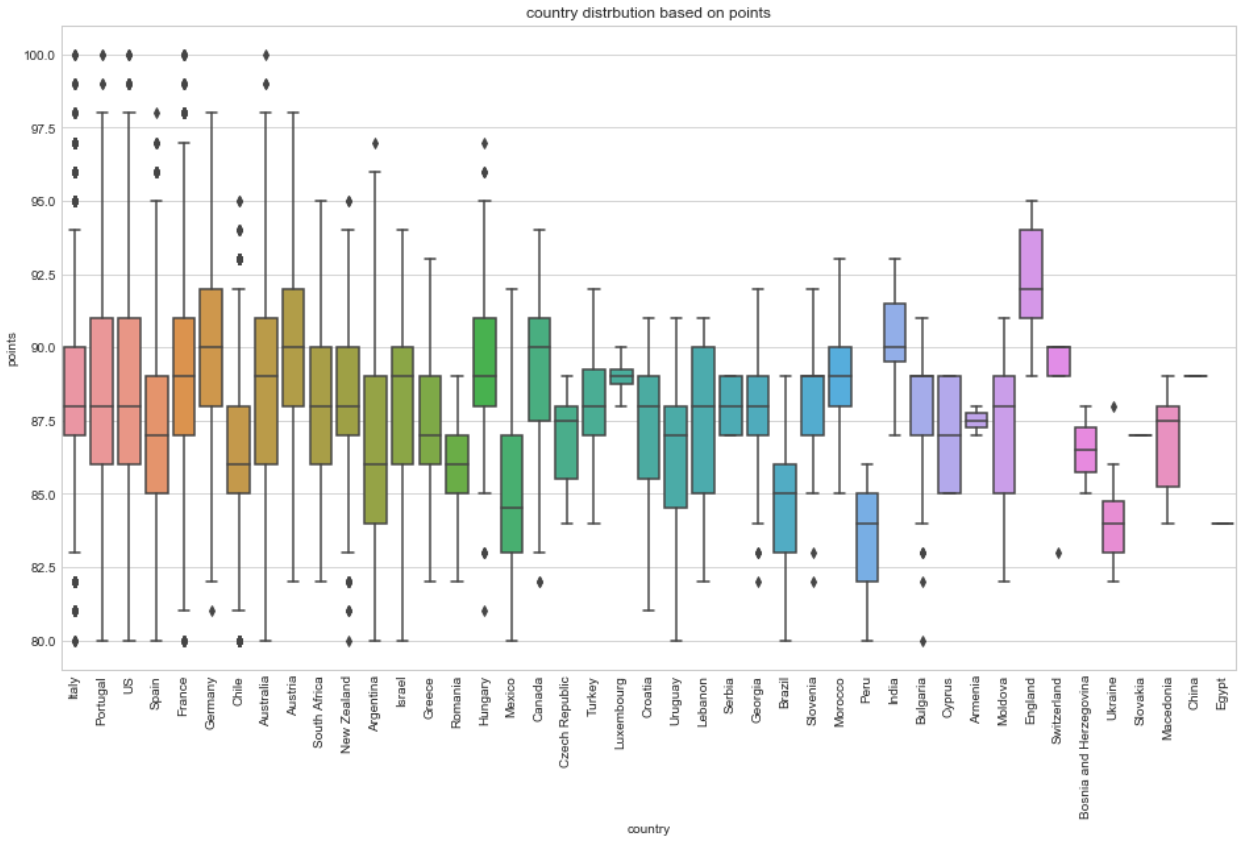
Points VS Prices

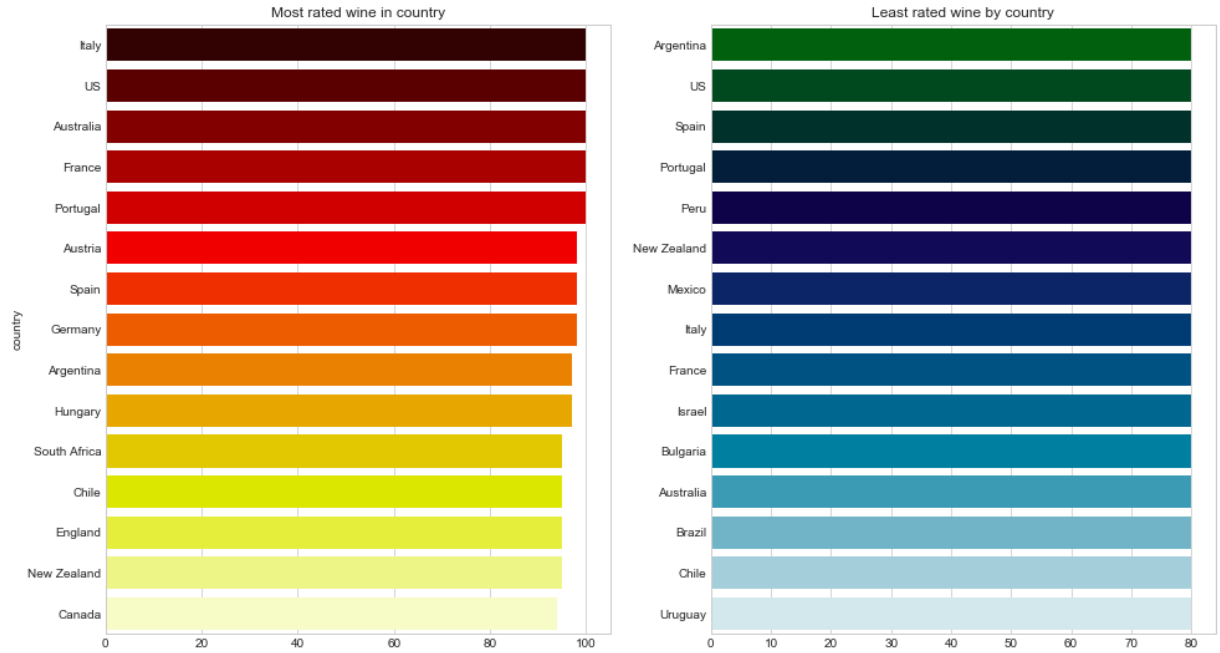


Country VS Price

The top country, Switzerland, which have the highest average wine price 55.4 USD. Both France and Italy have the so-called wine countries, but their average prices are listed as fifth and sixth.

Country VS Points



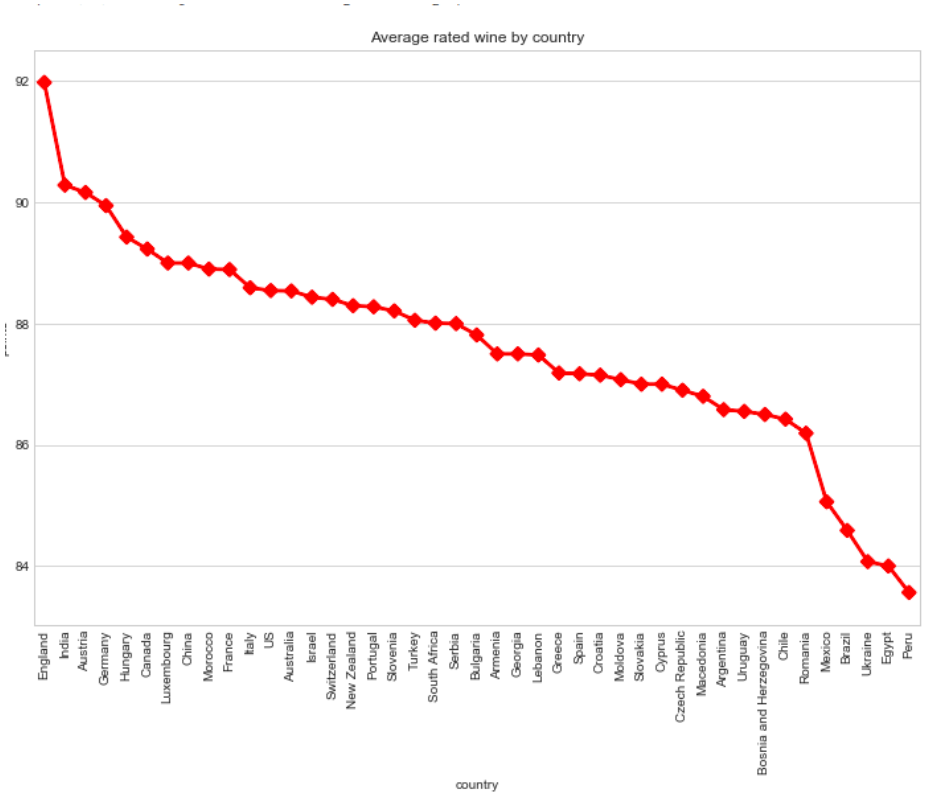


The rating of wine is varies from 0 to 100, but least rating is 80.

The wine produced in Italy, US, Australia, France and Portugal receives top rating.

The wine produced in Argentina, US, Spain, Portugal and Peru reveices least rating.

Both US and Spain received most rated wines and least rated wines.



The average rating of wine for England is top in the table, which is 92.

Coutries like India, Austria and Germany are top quality wine producing countries, which are over 90.

Countries like Ukraine, Egypt and Peru produced least quality wine.

Province VS Points

Based on average points, the top three provinces are Sudburgenland, Madeira and Mitterlrhein.

They are from Austria, Portugal and Germany.

	province	country	mean
0	Südburgenland	Austria	94
1	Madeira	Portugal	93.2308
2	Mittelrhein	Germany	92
3	England	England	91.9808
4	Wachau	Austria	91.8885

Province VS Price

Based on average wine price, the top three provinces are Colares, Vanju Mare and Switzerland.

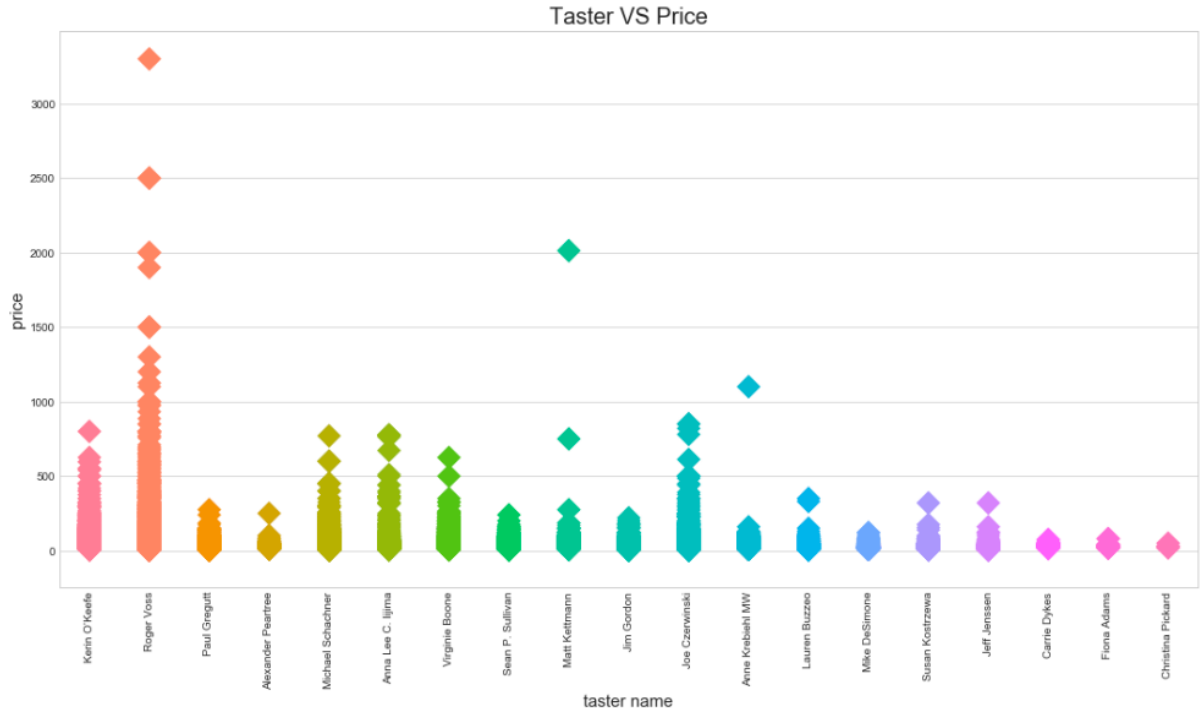
They arefrom Portugal, Romania and Switzerland.

	province	country	mean
0	Colares	Portugal	262.5
1	Vânju Mare	Romania	166
2	Switzerland	Switzerland	160
3	Madeira	Portugal	101.308
4	Rheingau	Germany	85.8524

Taster Name VS Price

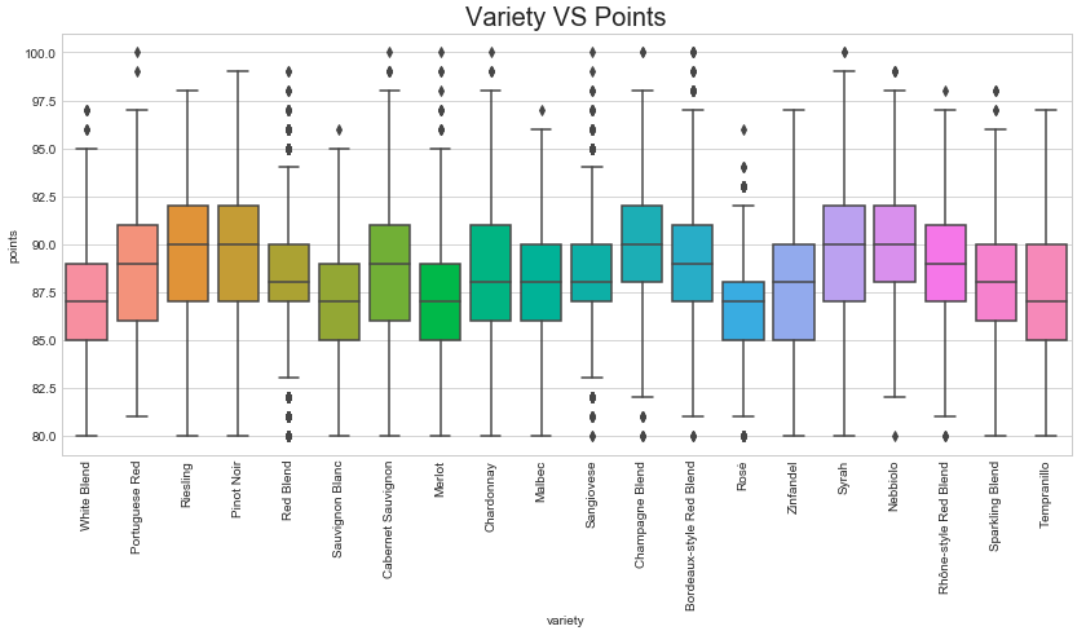
As we can see, both Roger Voss and Matt Kettmann rated wines whose prices are over 2,000

USD.



Variety VS Points

The average point of Pinot Noir is 90. The lowest point for this variety is 80, and the highest point is 98.



- **Model training**

When I was training models, I set target as 1 or 0, in which 1 means rated points are over 88 (average point) and 0 as points lower than 88. I set training data and test data ratio as 7:3 and calculate average accuracy, average precision, average recall and average F1 score by working on decision tree classifier, random forest classifier, cat boost classifier and so on.

Model	Accuracy
Decision Tree Classifier	92.23%
Random Forest Classifier	87.99%
KNeighbors Classifier	86.07%
Gradient Boosting Classifie	76.85%
Cat Boost Classifier	95.65%

Above table show cat boost classifier generated the highest accuracy. Therefore, I'm going to use cat boost classifier to extract which feature have the most important influence.

- **Feature exaction**

Feature selection technique is much more useful with larger dataset, where a lot of columns are useless. As we can see the most important feature is price. Tester has also big impact for the points score. By combining cat boost classifier and panda feature score, we get the below score ranking.

Features	Score
Price	41.64
Taster	15.27
winery	9.58
country	8.2
variety	7.55

Results

Based on the above graph, we can see the top five features that influence wine popularity in this data set. Of course, we can evaluated price always influenced points. What surprised me is taster feature. Before I did this project, I never think taster would have so much influence. Winery, country and variety followed up.

Based on my project result, I would like to give suggestions to wine customers. When you select wine, you do not only think about price, taster (do not care if no taster), winery, country and variety. But think deeper. As for price, there is no 100 percent that higher price lead to higher quality wine, the average great wine is clustered at 20 USD. As for taster, Roger Voss taste most and best wine in this data set. As for winery and country, Italy, US, Australia, France and Portugal always produced higher rated wines. As for variety, The top variety is Pinot Noir.

Conclusion

Dealing with missing value takes me up to half time of this project. It always lead me to think the importance of data preprocessing.

When we were handling with one feature or model and have no idea how to go along, combining other features or models can be a better way to analyze feature.

I always think training data model and result could be the most important part during project, this exercise lead me to think data preprocessing also played a key role in this project.

I would go further by implement more data analytics project and learning new library knowledge.

Appreciate all help from Dr. Yanjun Li during this summer.

Thanks!