Data Visualization

Module Code : B9DA106

Lecturer : Dr. Basel Magableh

Group Members :

Dhanashree Pisal (10511003)

Jay Parekh (10509524)

Assignment Title :

Interactive Visualization

Exploratory Analysis Example

This visualization report is an example of exploratory data analysis. It concentrates on the visualization of the particular dataset for better understanding of it. This report examines the dataset WINE REVIEWS. We exploring the Wines depending on their different factors.

What is Wine?

Wine is an alcoholic drink made with the juice of grapes, but they are different than what you’ll find at the grocery store. Basically, wine can be made up of any type of fruit (i.e. plum, cranberries, apple, etc) but mostly they are made with wine grapes. Vitis vinifera is the latin name of wine grapes. Wine grapes are small in size, sweet in taste and have thick skins, they contain seeds.



Dataset

Firstly, we collected the data from the Kaggle (<https://www.kaggle.com/zynicide/wine-reviews>).

This dataset contains 130k rows of wine reviews and 14 columns.

Following are the 14 columns of dataset:

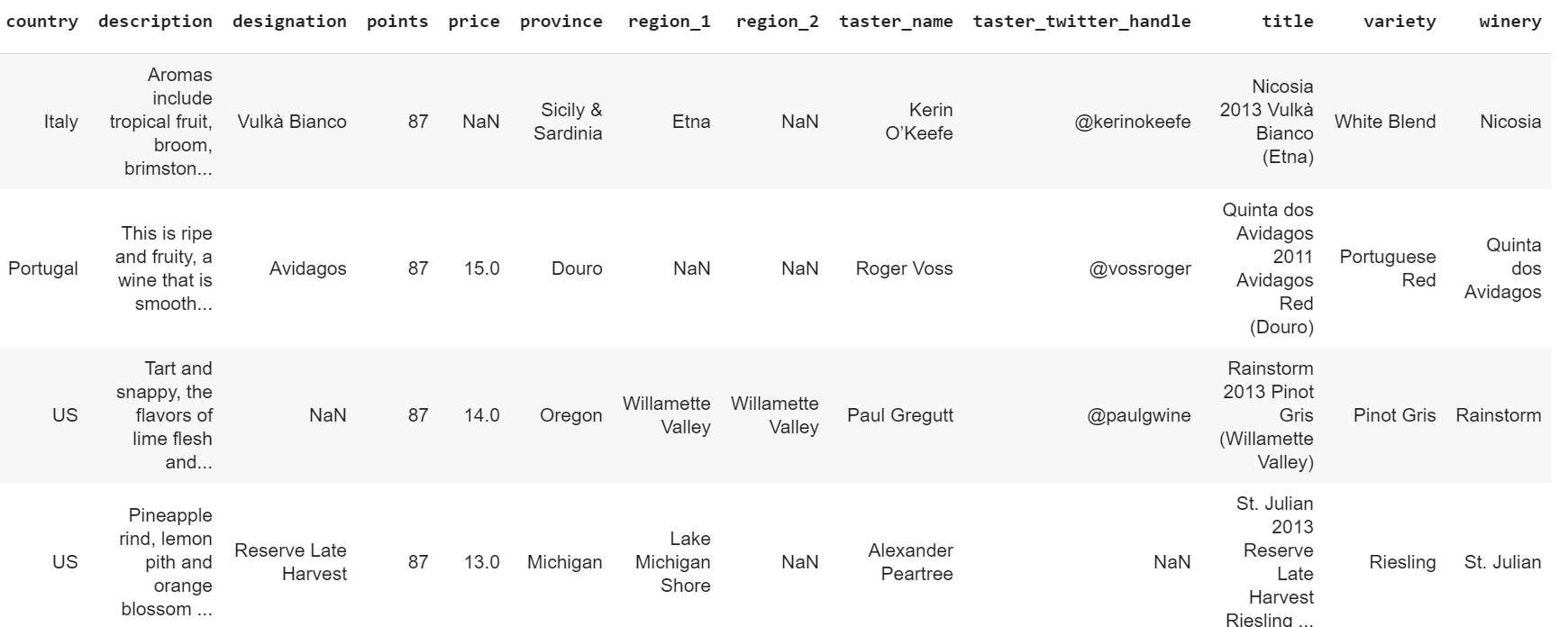
1. Country : The name of the country where wine is from
2. Description : Few comments from the wine steward about its smell, taste, colour etc.
3. Designation : the wine producers within the vineyard where the grapes that made the wine are from.
4. Points : The number of points wine lovers assessed the wine on scale on 1-100.
5. Price : Estimated price of that particular wine.
6. Province : The state from the wine is.
7. Region\_1 : state of wine growing area.
8. Region\_2 : Specific regions from the state wine growing area.
9. Taster\_name : The name of person who taste the wine.
10. Taster\_twitter\_handle : twitter handle of tasters.
11. Title : Title of wine.
12. Variety : Type of grape used to make wine.
13. Winery : winery name who made the wine.

After downloading the data from Kaggle we loaded it using pandas in python.

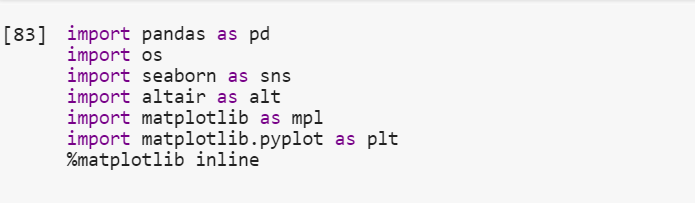
Following are the commands used to load file in python:



This is how dataset looks like :

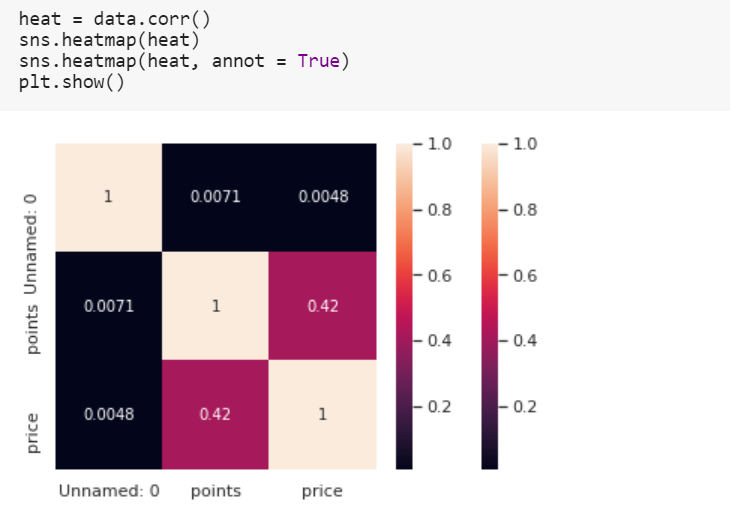


Importing the necessary libraries :



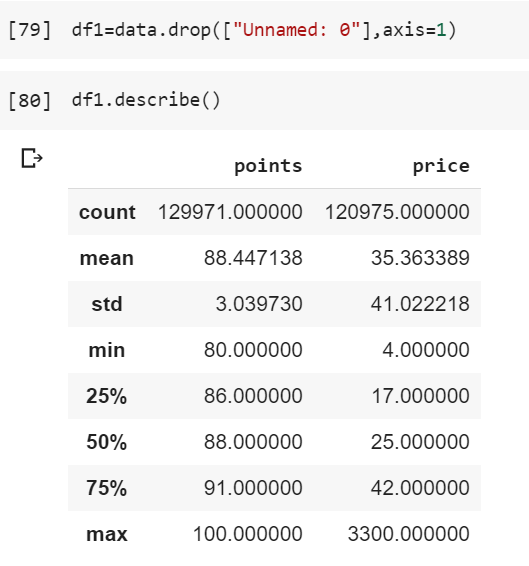
Cleaning of Data :

To check if there are any NULL values in data, following heatmap is created :



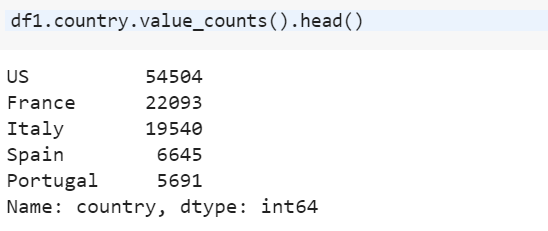
There are some NULL or UNMANED values in the dataset. Column Price and Points contains the NULL values.

So, we simply replaced NULL values with ‘0’.



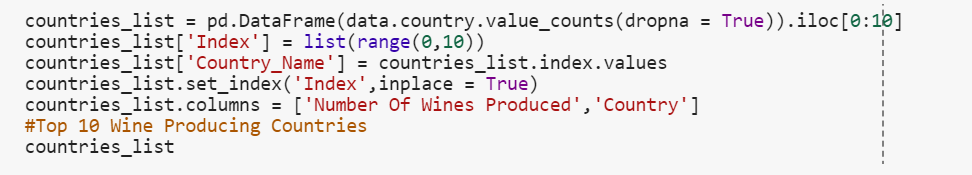
Insights of Data :

First five country with highest wine review :

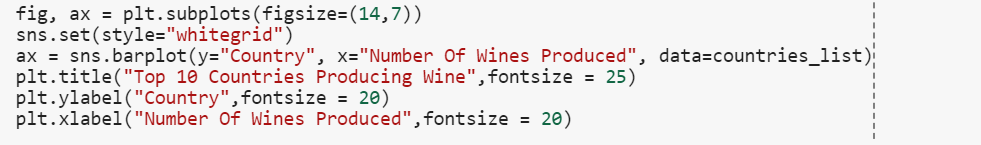


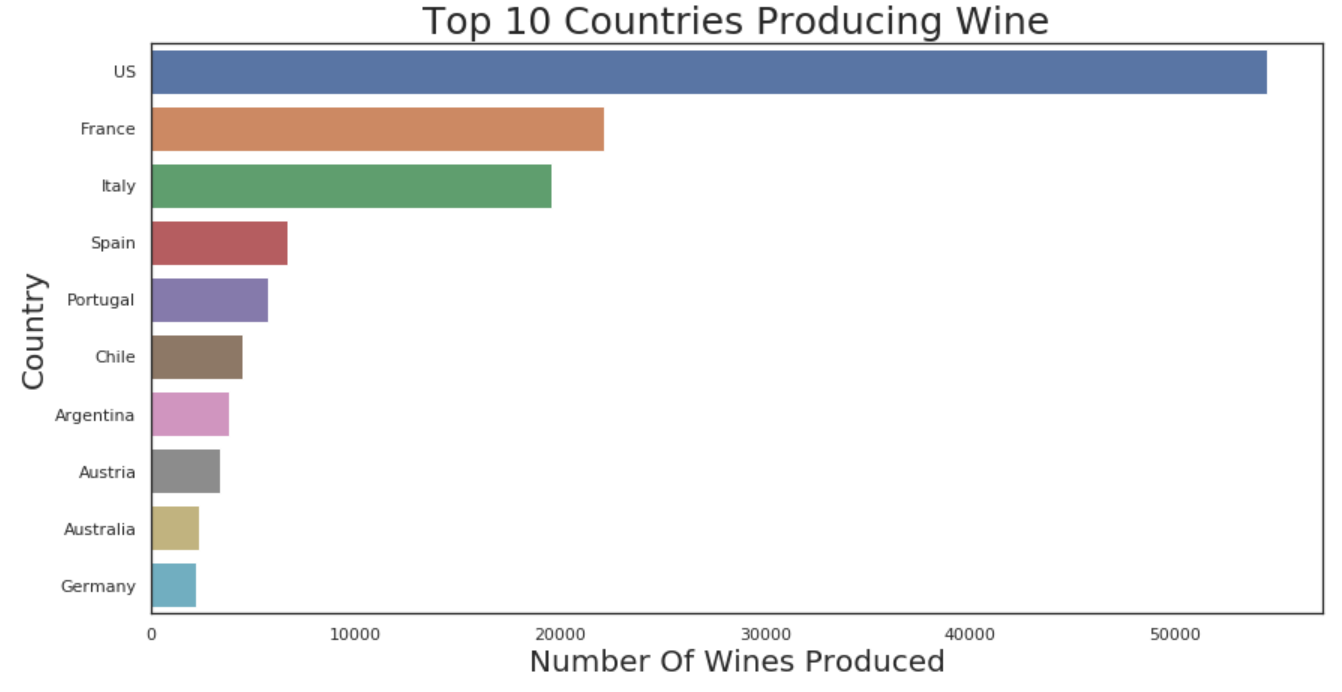
As we can see US(54504), France(22093), Italy(19540), Spain(6645) and Portugal(5691) are the top 5 countries with highest reviews.

Top 10 Countries Producing Wine :

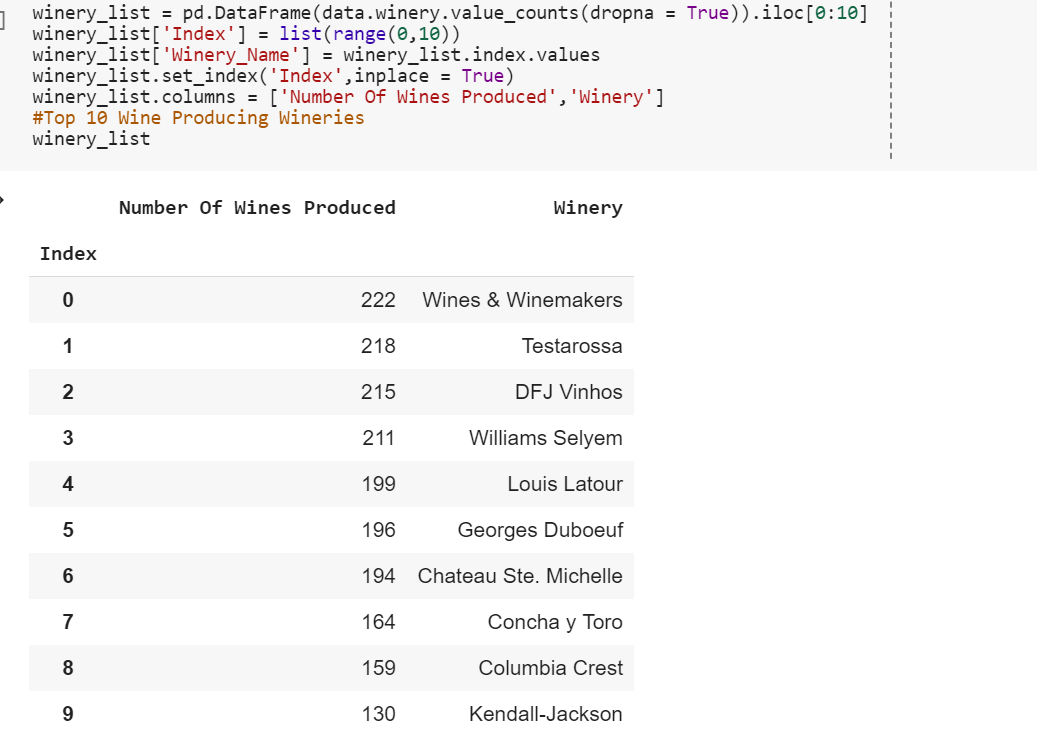


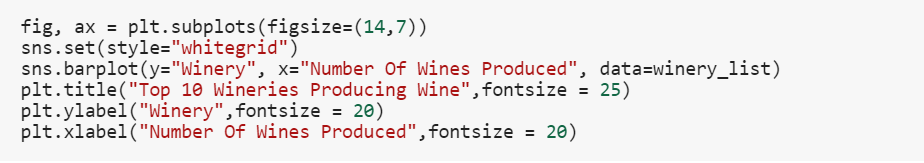


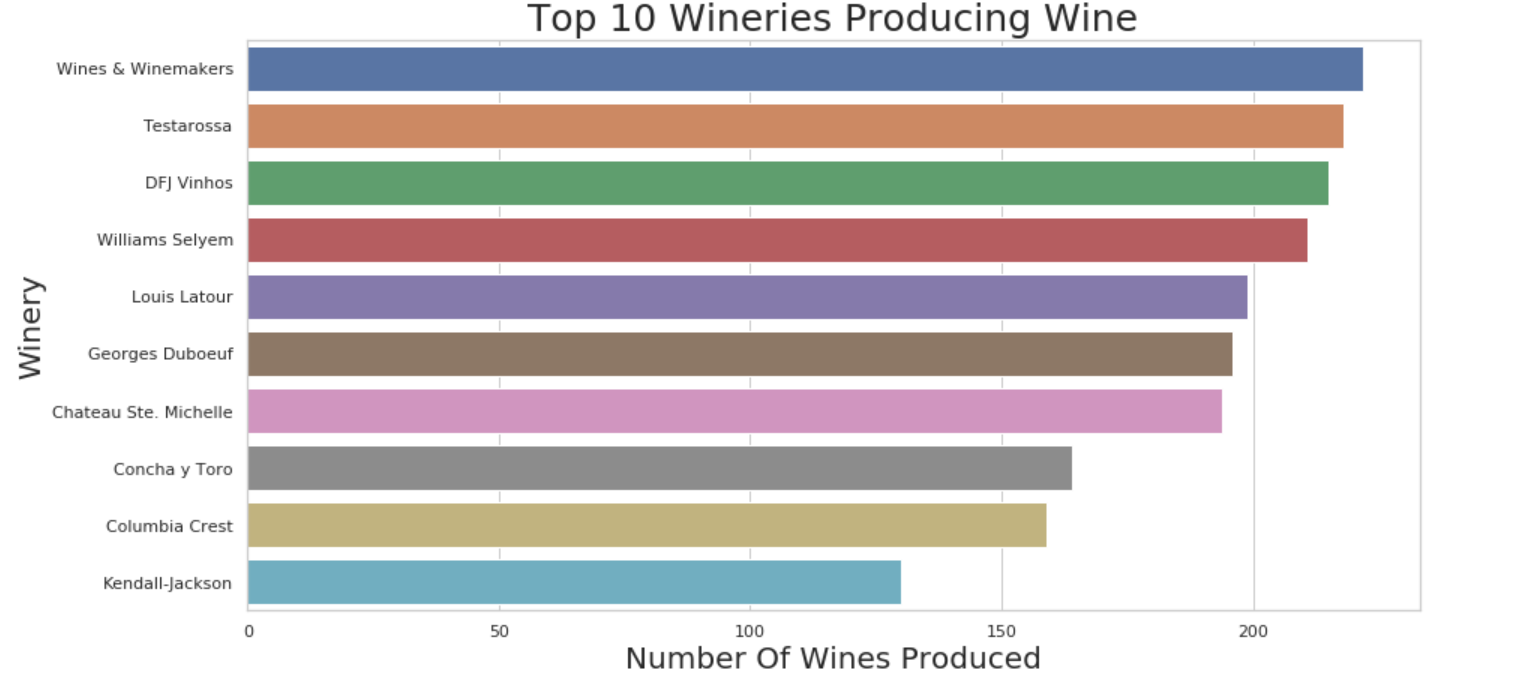




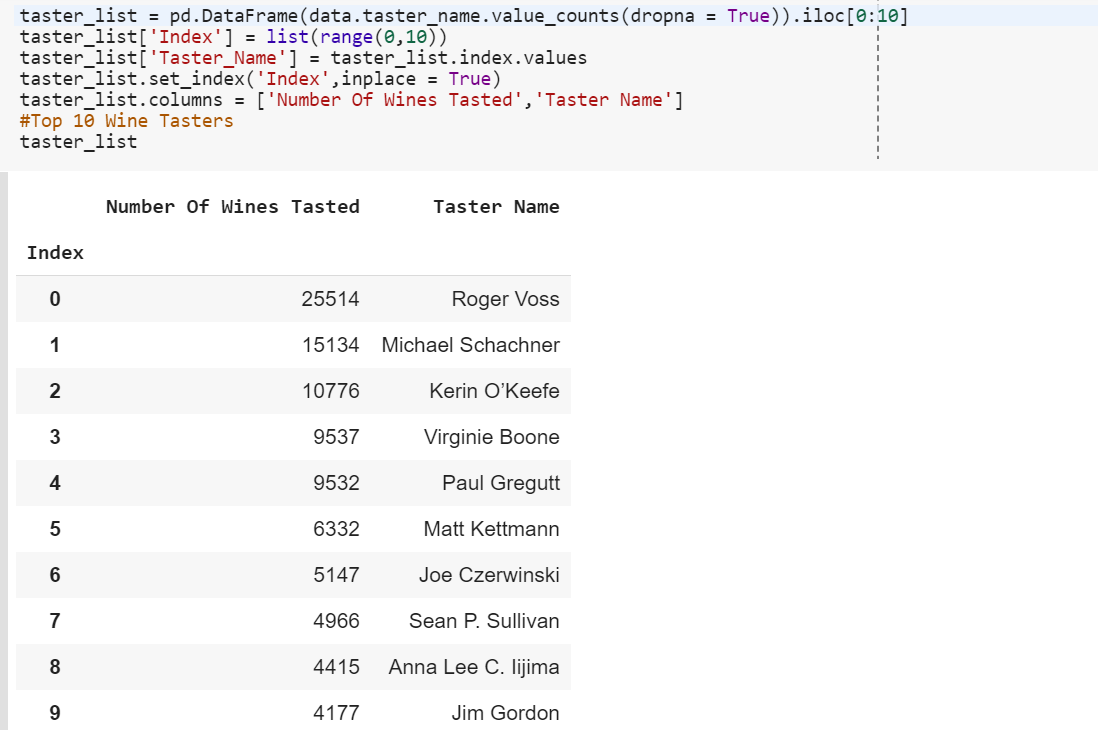
Top 10 Wineries Producing Countries - Number Of Wines Produced:

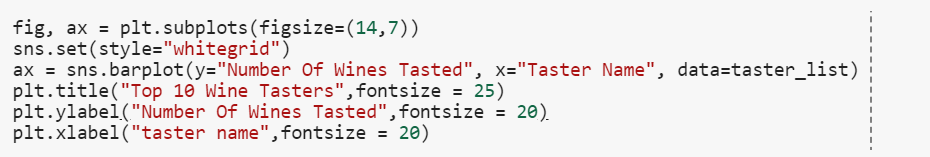


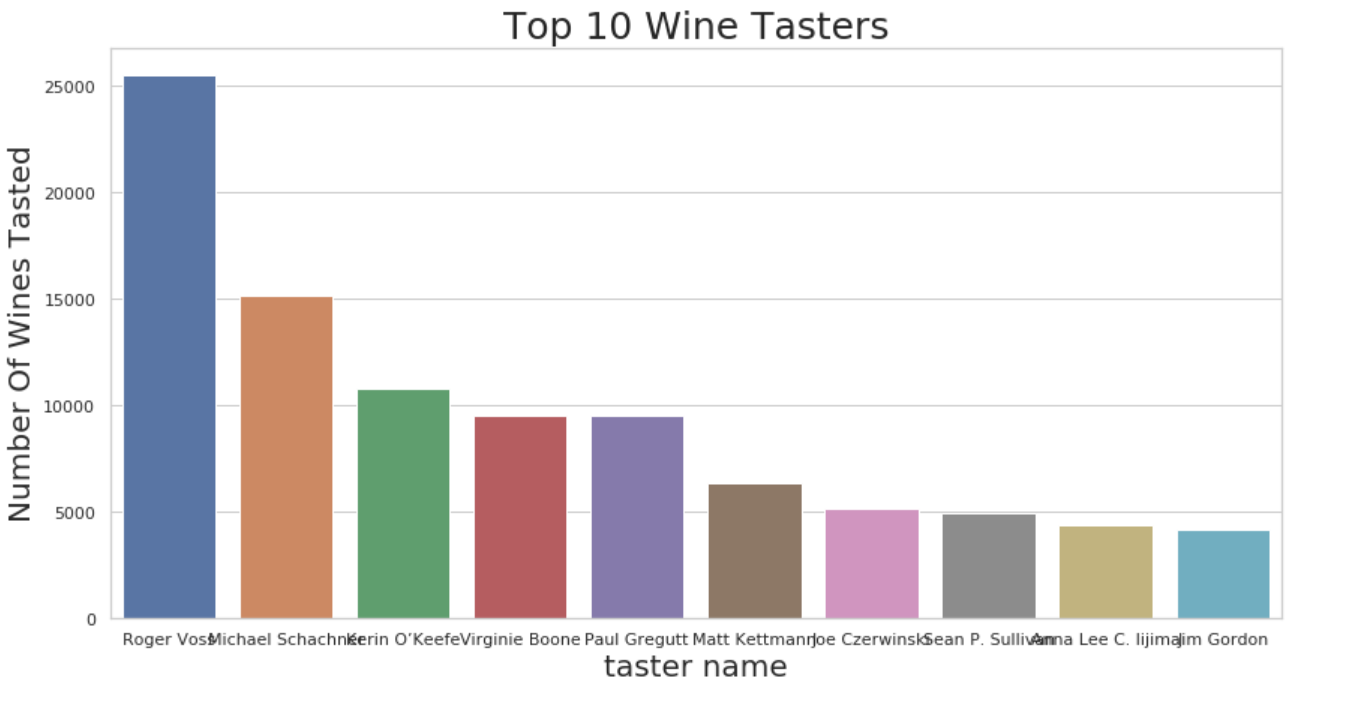




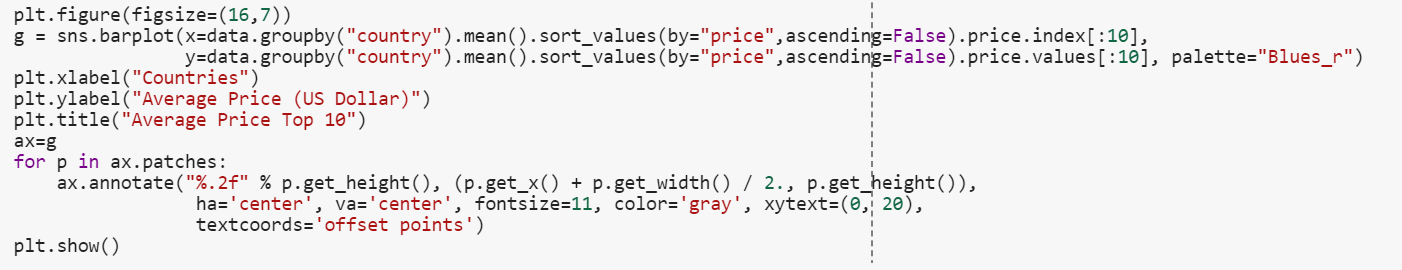
## Top 10 Wine Tasters - Number Of Wines Tasted :

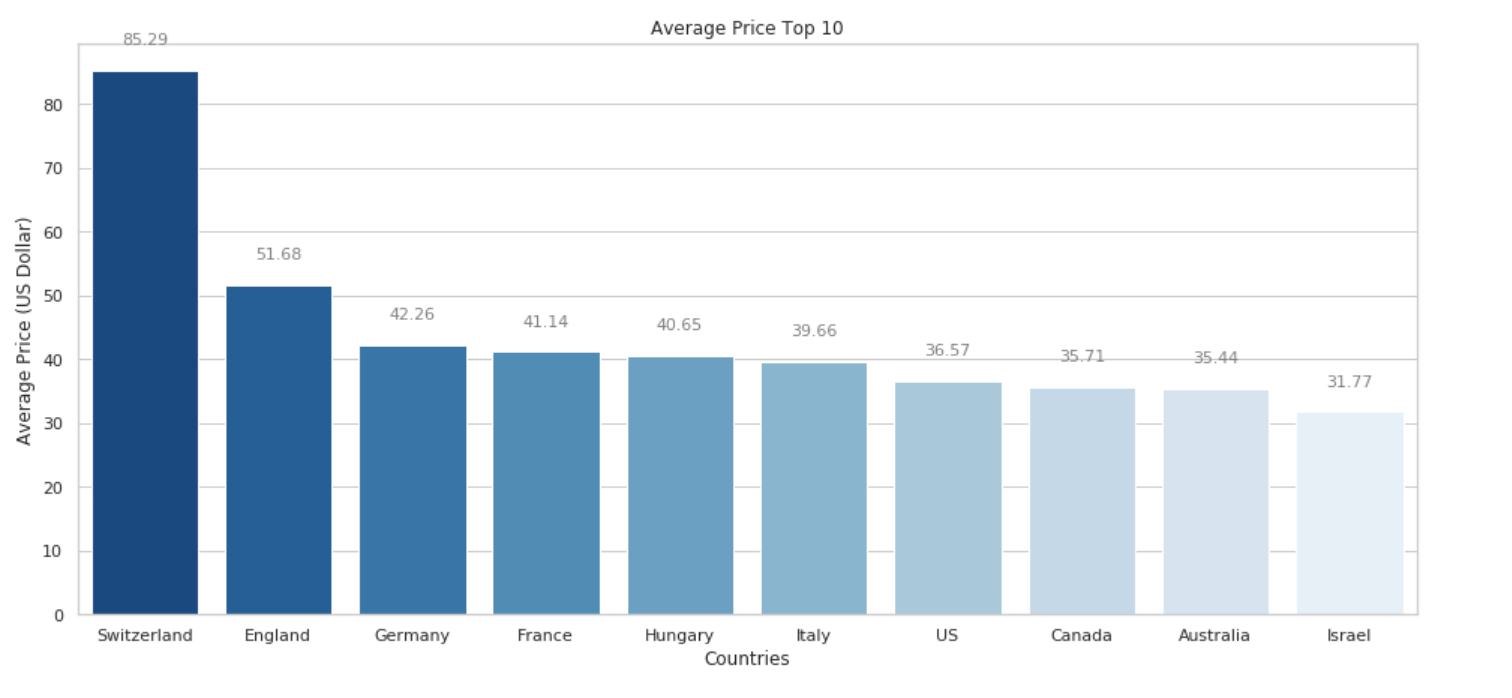






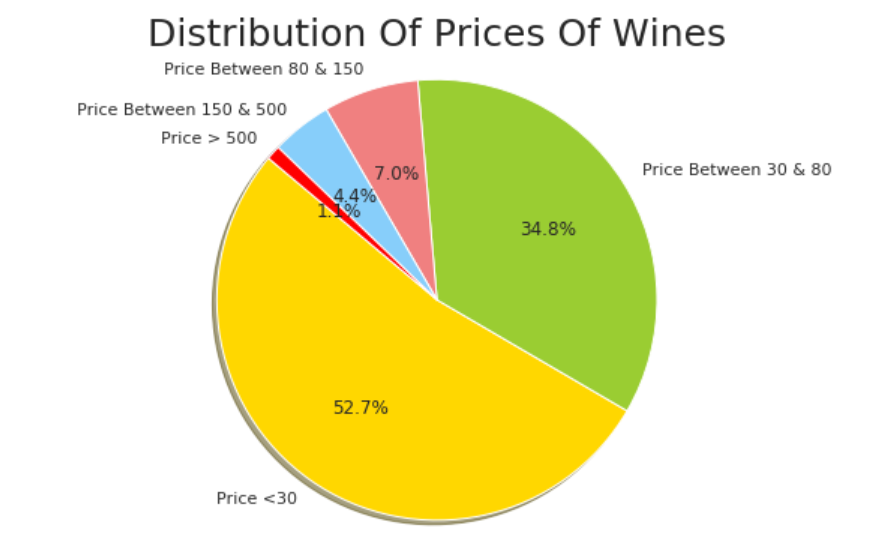
Average Price Of Top 10 Wines :





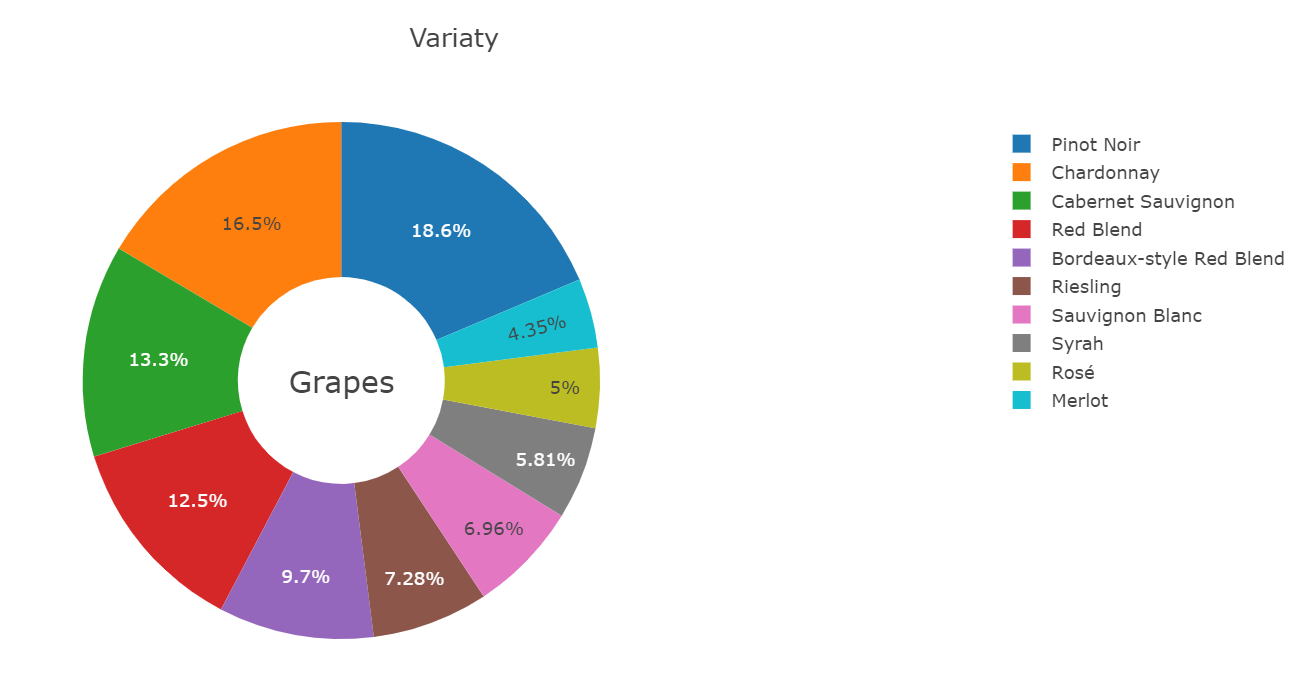
# Visualising Distribution Of Wines In Various Price Groups:





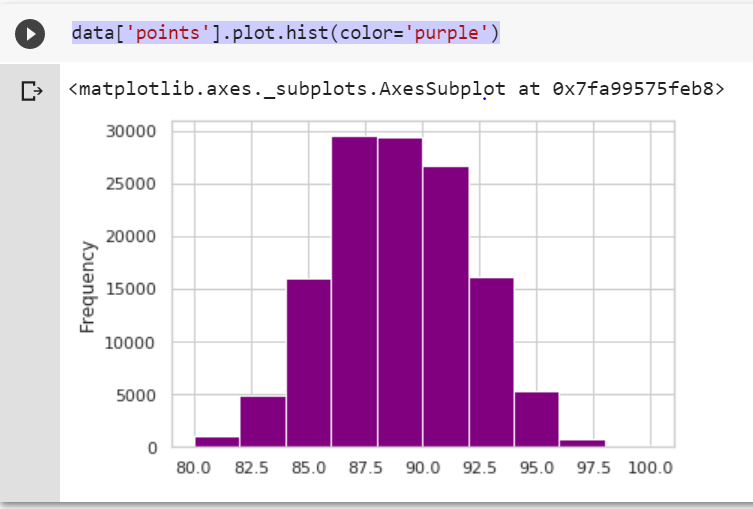
**The Type of Grapes Used to Make The Wine**(Top 10) :





Frequencies of wine price :

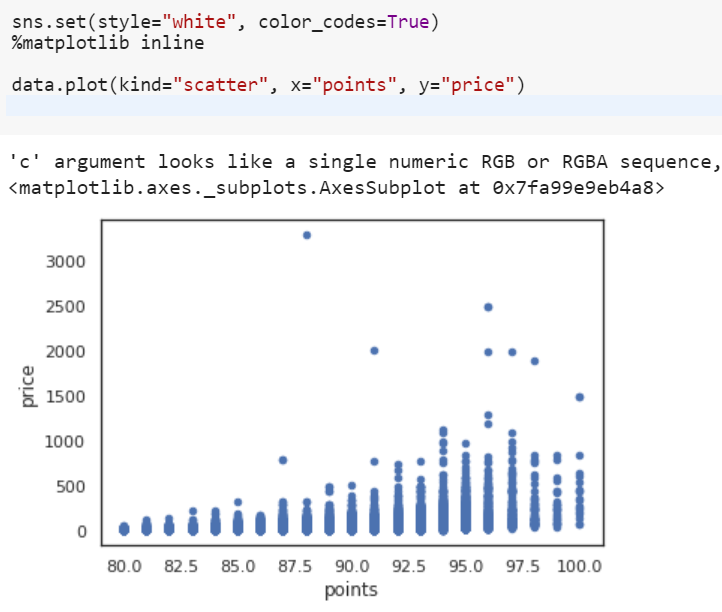
A histogram looks, trivially, like a bar plot. And it basically is! In fact, a histogram is special kind of bar plot that splits your data into even intervals and displays how many rows are in each interval with bars. The only analytical difference is that instead of each bar representing a single value, it represents a range of values.



This histogram shows us the frequencies of wine price. we are only looking for the wines frequencies which has lower price than 200. The maximum price of wine is less than 25. Now we were trying to find out the maximum points given to the wins. As we can see in above histogram, maximum points are between 85 to 92.5. And 87.5 points are given most of the wines.

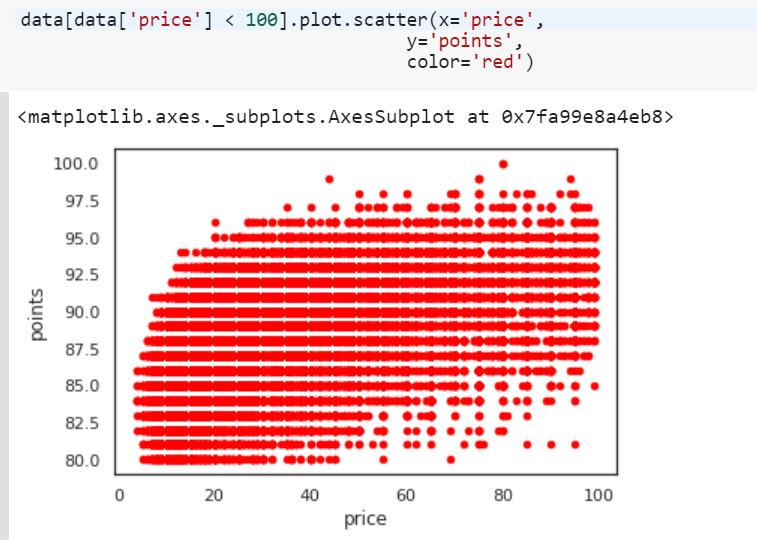
Correlation between wine price and ratings:

The simplest bivariate plot is the lowly scatter plot. A simple scatter plot simply maps each variable of interest to a point in two-dimensional space.



There are very few high-priced wines in the dataset. There are some low-cost wines for which reviewers have given a 100-point rating - we might look at the reviewer’s density for them.

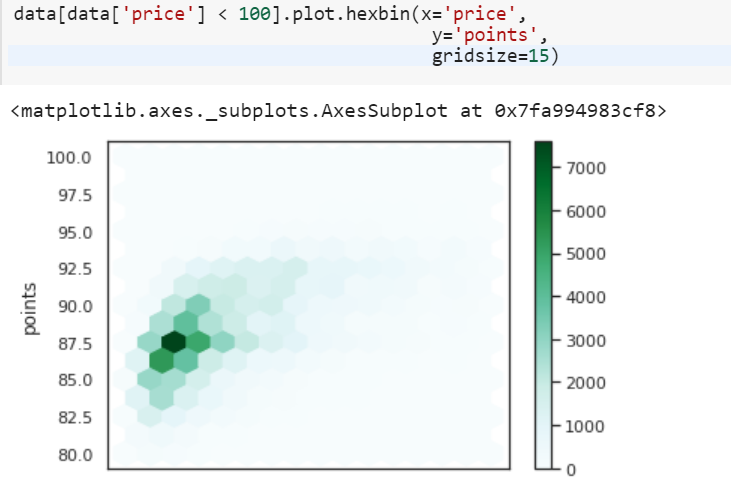
Correlation between Points and Price:



This plot shows us that price and points are weakly correlated: that is, that more expensive wines do generally earn more points when reviewed.

Hex Plot showing Points and Price relationship :

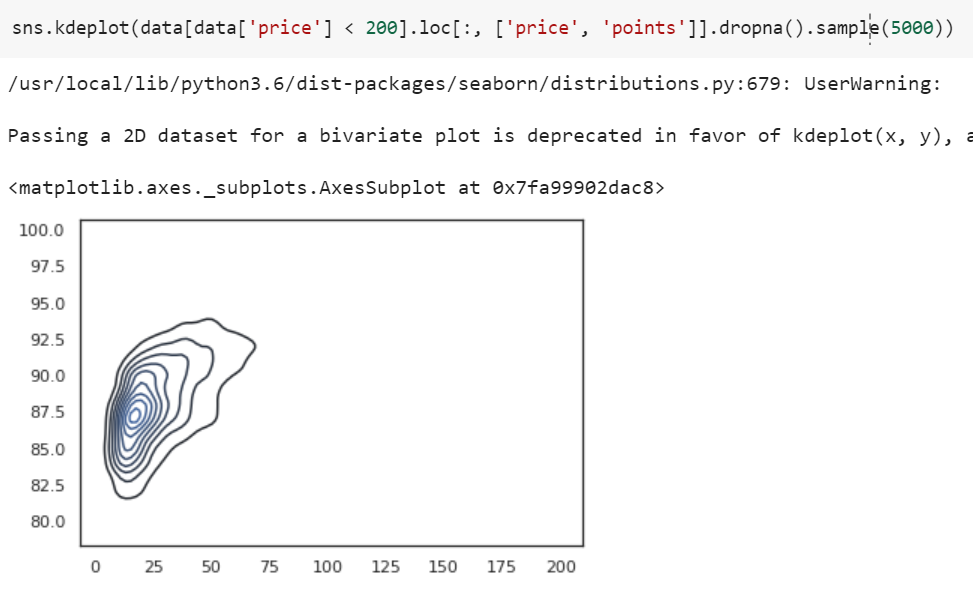
A **hex plot** aggregates points in space into hexagons, and then colours those hexagons based on the values within them:



## Distribution Of Points & Price Of Wine < 100 :

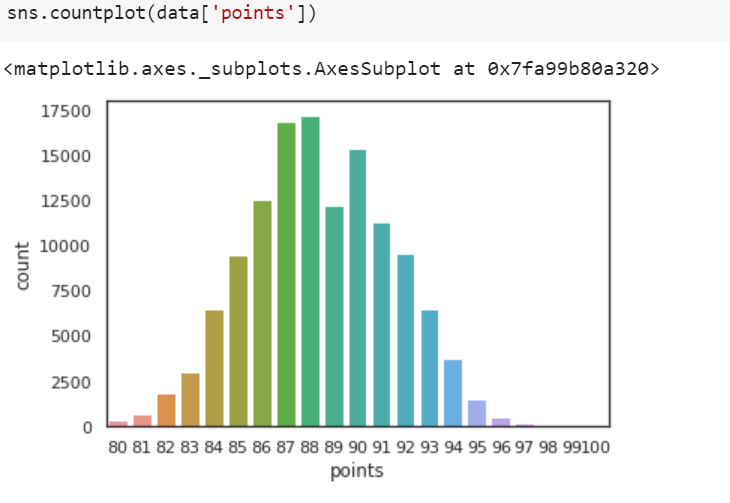
## 

A KDE plot is better than a line chart for getting the "true shape" of interval data. In fact, I recommend always using it instead of a line chart for such data. However, it's a worse choice for ordinal categorical data.



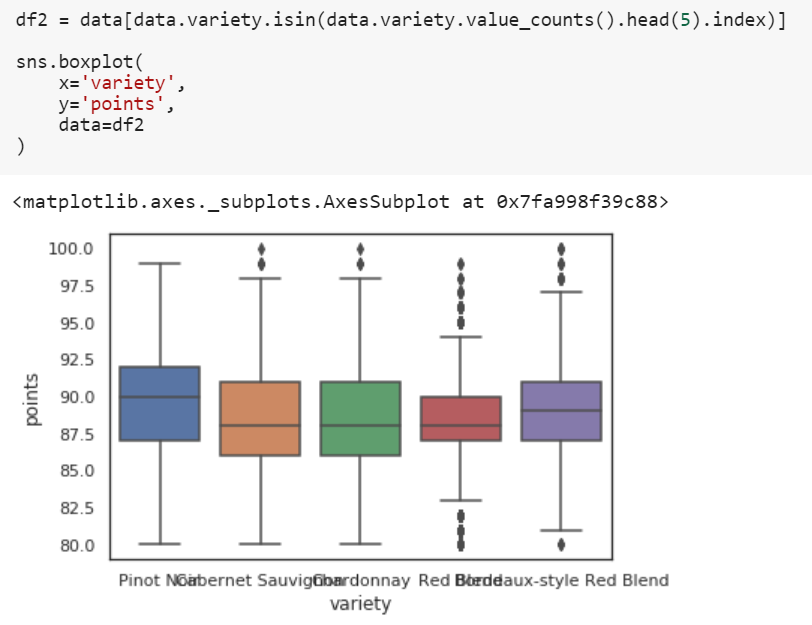
A KDE plot expects that if there are 200 wine rated 85 and 400 rated 86, then the values in between, like 85.5, should smooth out to somewhere in between (say, 300). However, if the value in between can't occur (wine ratings of 85.5 are not allowed), then the KDE plot is fitting to something that doesn't exist. In these cases, use a line chart instead. KDE plots can also be used in two dimensions.

Count plot :The pandas bar chart becomes a seaborn count plot. We are counting the Points of wine.

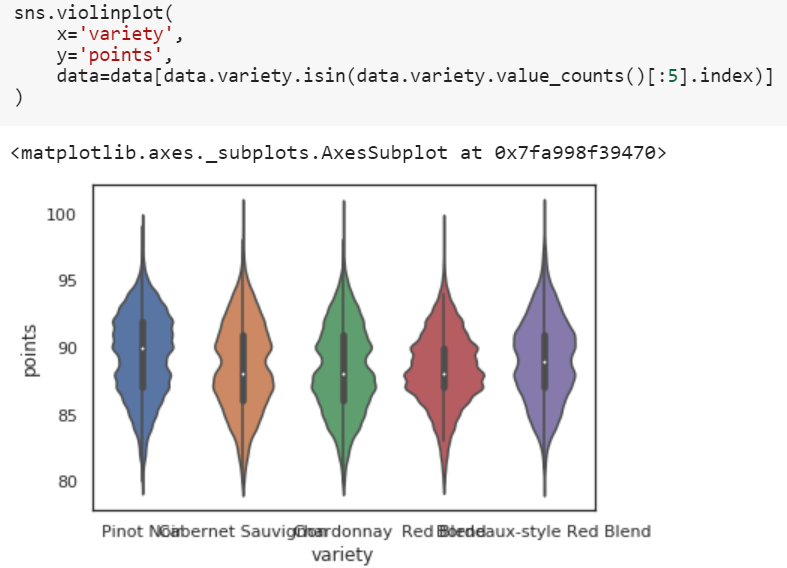


## Boxplot and violin plot:

seaborn provides a boxplot function. It creates a statistically useful plot that looks like this:

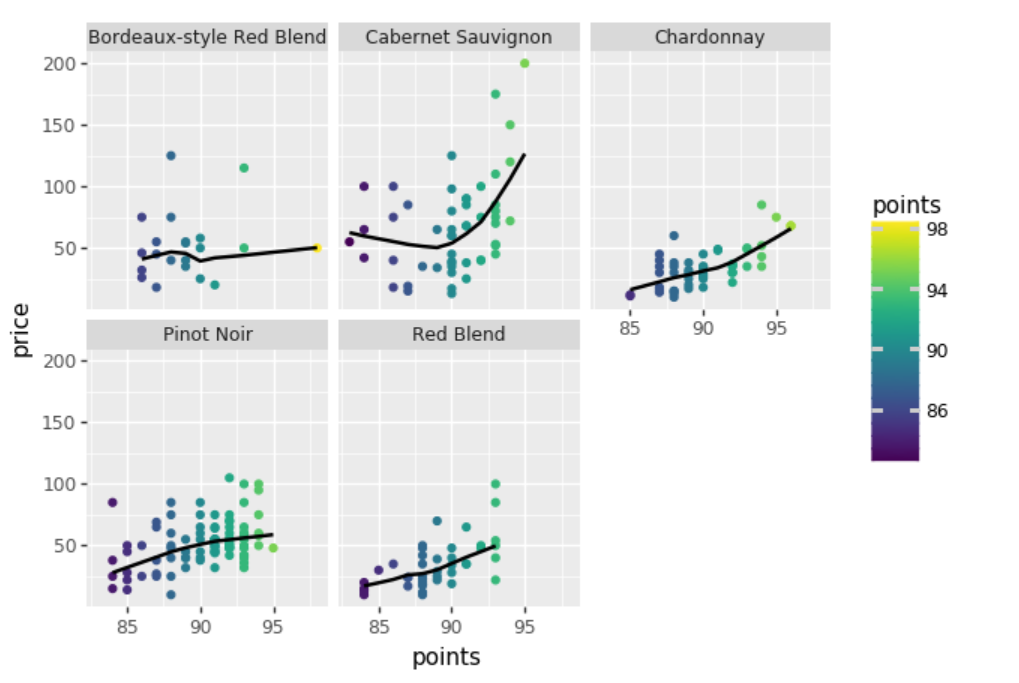


The centre of the distributions shown above is the "box" in boxplot. The top of the box is the 75th percentile, while the bottom is the 25th percentile. In other words, half of the data is distributed within the box! The green line in the middle is the median. The other part of the plot, the "whiskers", shows the extent of the points beyond the centre of the distribution. Individual circles beyond that are outliers. This boxplot shows us that although all five wines receive broadly similar ratings, Bordeaux-style wines tend to be rated a little higher than a Chardonnay.



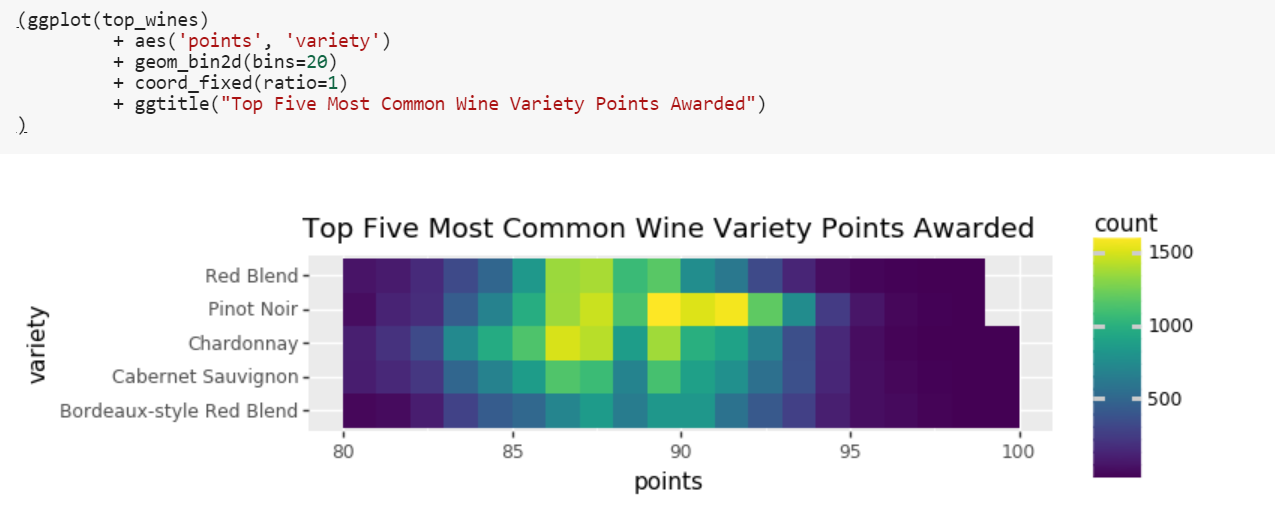
Comparison of variety of wines by Price and Points :





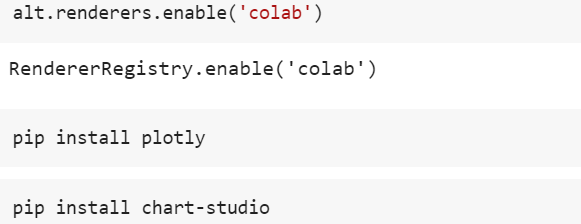
As shown in the graph cabernet sauvignon has the highest price and even peak in points as compared other wines. The lowest points and price are marked in red blend wine . Moderate price and points are noticed in pinot noir.

Top Five Most Common Variety Points Awarded :

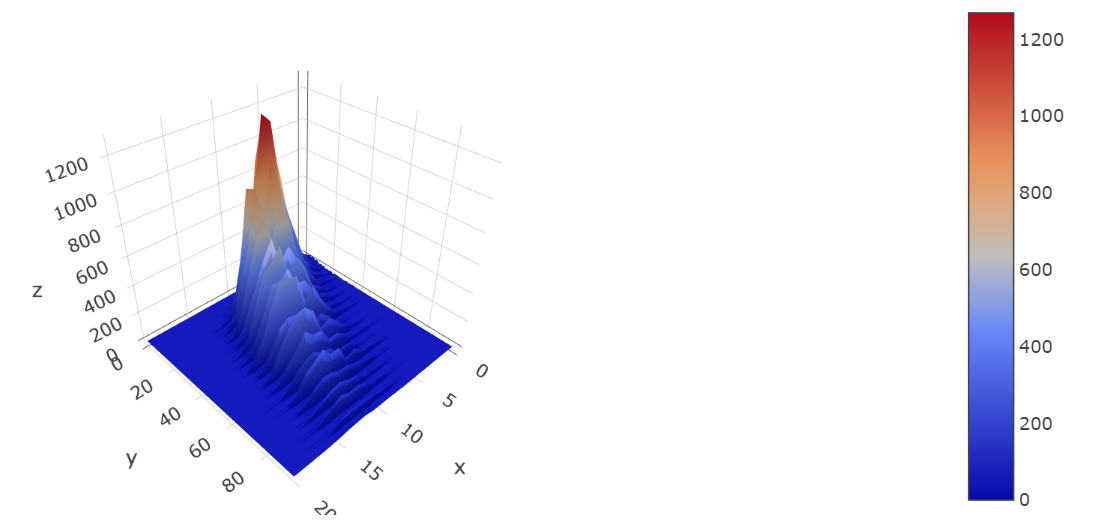


The above graphical representation shows the five top varieties of wine with their maximum points and highly awarded amongst them. Red blend, Piont Noir, Chardonnay, Cabernet Sauvignon and Bordeaux-style Red Blend.

Following are the Important libraries need to install before performing plotly function in Python :





The thThe above three-dimensional interactive presentation displays that ploty function is used to visualize the correlation between two attributes that are price and points of the wines. In this three-dimensional presentation we can even zoom in and out for more attractive display.