Sammy Pardes

IST652

Homework 1

7/31/2020

**Introduction**

For those of use who love to travel, lately, our travel bugs have been squashed by the ongoing pandemic. While it’s not feasible to take a flight to the south of France right now, one way to feel as though you’ve emerged at a world-class vinyard, is by delving into a bottle of wine from around the world. Fortunately, with a little bit of Python, there is a lot of information to uncover that will help us determine where to find the best bottles of wine.

**About the data**

Data.world has put together a data set of a spectrum of wines from around the world. The Initial data set contained nearly 25,000 rows. For this assignment, the data set was re-saved as a .csv file and then cut down to a randomized 4,000 rows from the original 24,997.

There are ten attributes for each wine in the data set. The vintage is the year that the grapes were harvested. Each vintage begins on January 1st of its respective year. Country is the nation in which the winery is located. Similarly, the province field defines the location of the winery. It specifies the region of the country where the winery is. The county field is the growing region within the province. The designation field tells about the vineyard where the grapes were grown. The title field is the name of the wine. The price attribute is the price per bottle in USD. The variety attribute tells us the type of grape that was used to make the wine. The winery field is the name of the winery at which the wine was made.

Lastly, the points field is a metric to determine the quality of the wine. The value of the points field ranges from 0 to 100. Here, a score from 95-100 indicates a wine that is great. Outstanding wines score from 90-94. Any wine that scores between 85 and 89 is considered very good. A wine score of 80-84 is indicative of a good wine. In this data set, all wines have scored at least an 80.

**Data Cleaning and Exploration**

The first step in the data exploration process was to import the data. The original Excel file was saved as a .csv file into the same folder as the python program within Jupyter Notebook. An empty list was created and then each line of the file was read into the list using the append() function. The designation field was removed using a for loop and “del”, as it provided extraneous information and was not necessary for the analysis. The first row containing the column names was removed by slicing the wineList from the 1st element to the last, omitting the 0th element.

Next, any field with missing data was removed from the data set with another for loop. Since the target number of instances was between 500 and 4,000 rows, removing any column with missing data left plenty of wines with complete fields to analyze. Omitting rows with missing information shortened the data set to 19,494 rows. All full rows were appended to a wineClean list.

The sample() function from the random package was used to generate a list of 4,000 rows from the cleansed data set. The smaller list of wines was then converted into a pandas dataframe. From there, the column names were defined and stored in a list for later use.

To wrap up the data cleaning, the points field was set to numeric, and the vintage field was converted to a datetime, utilizing only the year. Finally, the “$” and “,” characters were removed from the price field and it was converted to a float data type. This cleansed dataframe of 4,000 rows is what the data analysis was performed on.

Below is the output of the first five rows of the cleansed pandas dataframe:



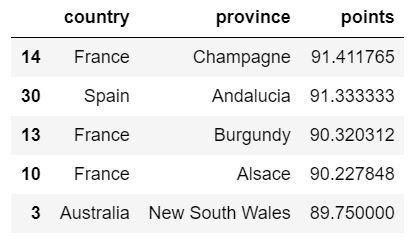
**Question 1: Program Steps and Output**

While we’re at home, it would be ideal to treat ourselves to the best wines in the world. So, what are the provinces of the world with the best average wine scores, and what are the top wines within those provinces?

The shortened dataframe contained wine from seven different countries. A new dataframe was set up where the wines were categorized by country and province and then sorted by average wine score using the groupby() and mean() functions.

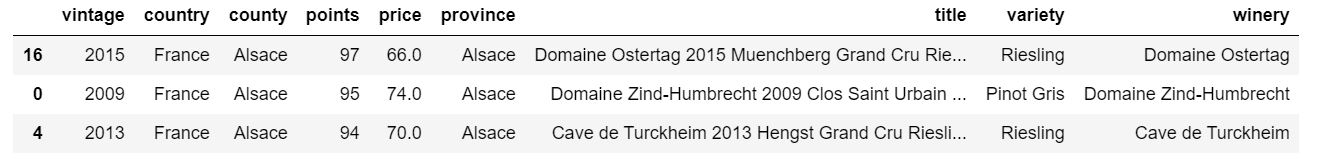
To determine the best wines in the world, the unit of analysis was the average number of points wines in a given region scored. For each province in the seven represented countries, the average wine score was computed with the nlargest() function.

As shown below, the top five rows of the output tells us that the best country/province combinations for wine are: 1.) Champagne, France, 2.) Andalucia, Spain, 3.) Burgundy, France, 4.) Alsace, France, and 5.) New South Wales, Australia.

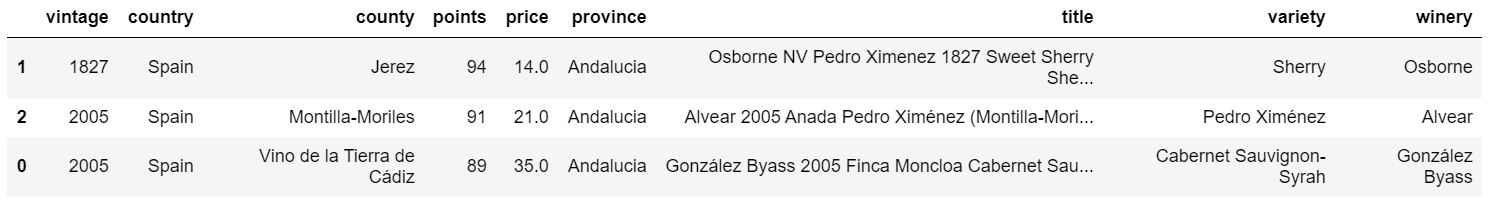


To further break down the best wine regions, the wine dataframe was converted back to a list. Each province ranking in the top five was then initialized as an empty list. Wines for each of the best provinces were distributed to their respective lists and turned back into dataframes.

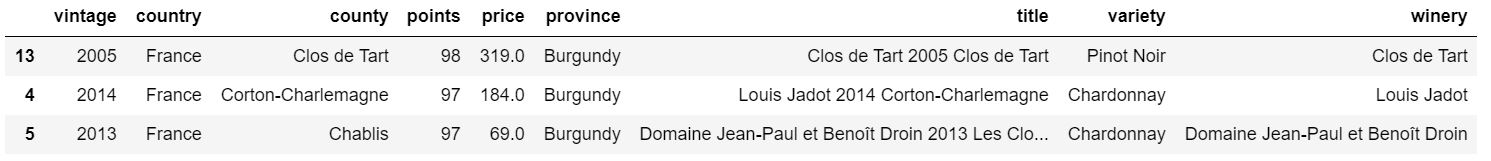
Using the nlargest() function again, the top three wines for each of the top five provinces was calculated. In Alsace, France, the best wine scored 97 points. This area must be known for its white pine as two of the top three wines in Alsace were Rieslings and the other is a Pinot Gris.



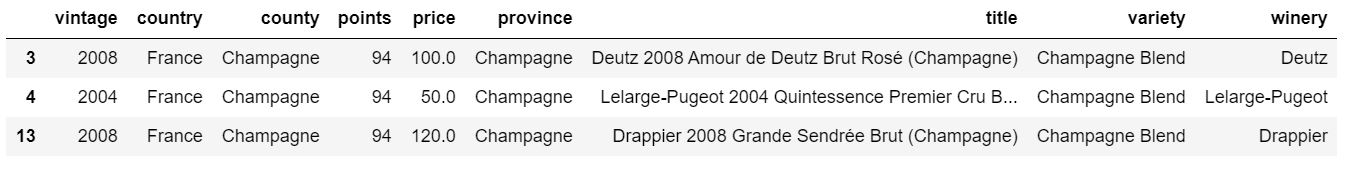
As the representative region from Spain, the best wine in the province of Andalucia scored 94 points. Interestingly, this vintage is from 1827, but is a reasonable $14.00. The next best wine in Andalucia was a Pedro Ximénez that scored 91 points. It seems 2005 was a great year for wines across Andalucia.



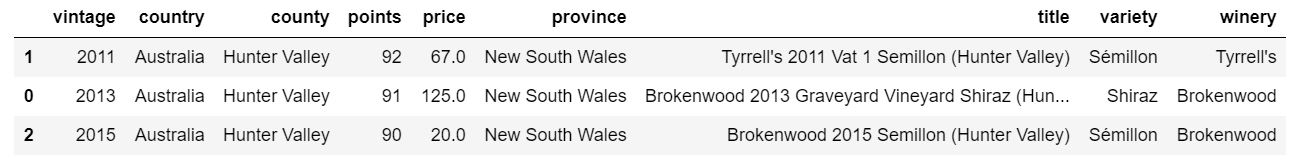
Burgundy, France, ranked third for average wine score. One Pinot Noir and two Chardonnays appear in the top three for this province. While each of the wines here have scored an impressive 97 or 98 points, they are very expensive at $319.00, $184.00, and $69.00 per bottle. Perhaps there are many other wines in this province that are bringing down the extremely high scores of the top three wines to give it an average of about 90 points.



It’s no surprise that the highest ranking province was Champagne, France and that all three of the best wines there are Champagne Blends. While not as pricey as the Pinot Noir from Burgundy, these wines have a premium price tag, no matter the winery.



Down under in New South Wales, Australia, the top three wines are all considered outstanding. Two wines from the Brokenwood winery appear on this list. For all three, they are from the county of Hunter Valley and the vintages are from the 2010s.



I would certainly place an order from any of these five provinces. Since I’m a fan of white wine, I am interested in the wines from Alsace and Champagne. I’m drawn to the price point of the wines from Andalucia. Were I to purchase wines from New South Wales, I would be sure to target the county of Hunter. For a very special occasion, I would consider the 98 point-scoring Pinot Noir from Burgundy.

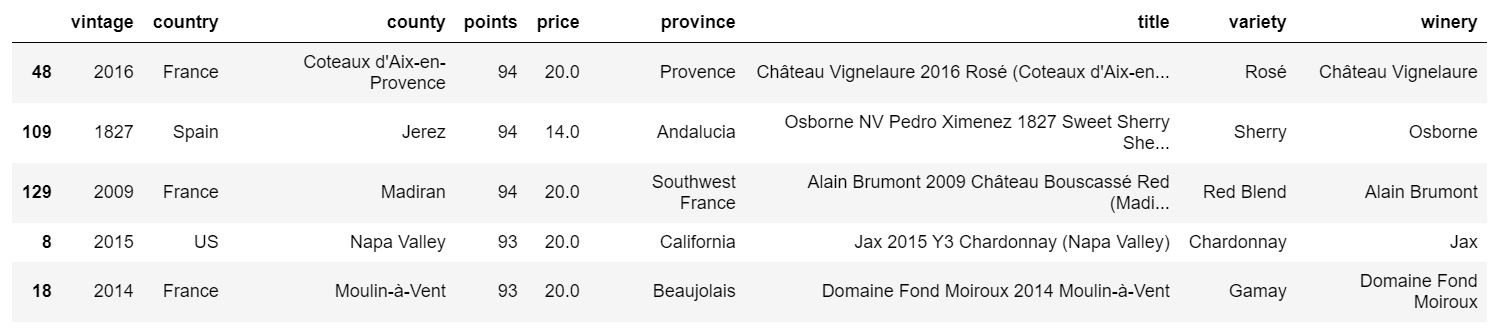
**Question 2: Program Steps and Output**

Of course, we want to drink the best wines, but some of these are extremely pricey. Not everyone has the luxury of being able to afford a $319.00 bottle of Pinot Noir or a $125.00 bottle of Shiraz. So, where can we find the best wines on a budget? Do outstanding wines (90-94 points) or better exist for $20.00 or less?

Luckily for us, creating an empty list called cheapWine where only wines that cost $20 or less and scored at least 90 points were added, yielded 139 results. Repeating the step of turning the list into a pandas dataframe and utilizing the nlargest() function, it’s easy to see where we can find the best, cheap wines using the unit of analysis of points.

If you’re into reds or whites, or even rosés, you can certainly enjoy a high-quality wine for $20.00 or less. Just looking at the top scoring wines for $20.00 or under, there are a wide range of varieties to choose from. The highest-scoring wines for $20.00 or less scored 94 points. Three wines hit this mark.

One could acquire three bottles of outstanding French wine for $60, nearly half the cost of one bottle of the best Champagne in Champagne, France, which scored the same or just one point higher.



**Conclusion**

It’s important to recognize the short-comings of this analysis. Because we have only observed 4,000 rows, there may be trends within this data set we are unable to recognize. Additionally, it’s possible that there are better provinces or higher-scoring cheap wines that were not factored into this subset. Therefore, running this program more than once may not yield identical results. Going forward, I look forward to diving deeper into this data set. Perhaps we can predict the wine scores of different bottles based on attributes like vintage, county, or variety.