Sammy Pardes

IST652

Homework 2

8/18/2020

**Introduction**

There is a lot of information on a wine bottle and it can be difficult to determine how a wine will taste based on the label alone. While vintage, ABV, and region surely all play a part in the quality of a wine, I wonder if the description of a wine is a good indicator of how many points a wine will score. Using Python and a data set obtained from Kaggle, can we determine the caliber of a wine based on specific language within the description?

**About the data**

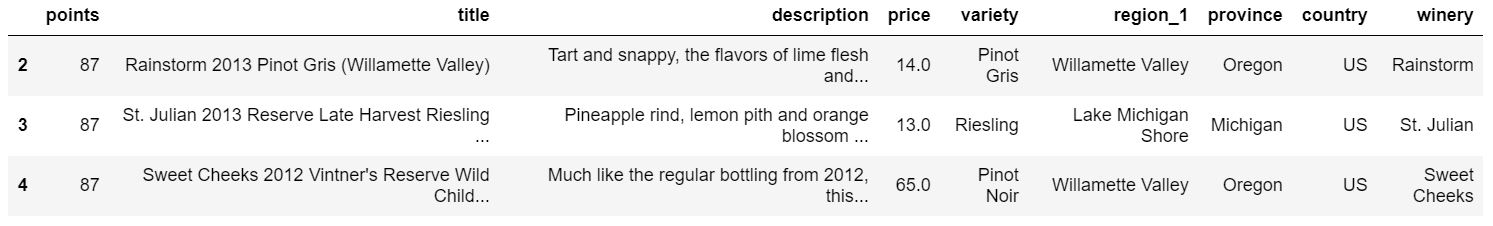
Available on Kaggle is a data set of 130,000 wine reviews gathered from WineEnthusiast in 2017. This information is available as a .csv or a .json file. For this assignment, the .json version of the document was used. For each wine in the data set, there are 13 attributes.

The country attribute defines the country that the wine is from. The province field is the province or state that the wine comes from. More specifically, region\_1 indicates the wine growing area within the province/state. Region\_2 is an even more precise location within the growing region. Winery is the name of the winery where the wine is made. Designation is the vineyard within the winery where the grapes in the wine are from. Variety specifies the type of grapes used to make the wine. The title column denotes the name of the wine. Price is the cost per bottle in USD. Points is the score the wine received from WineEnthusiast. It is on a scale of 1-100, however, all wines in this data set have scored at least 80 points. Taster\_name and taster\_twitter\_handle tell us who reviewed the wine and the Twitter user name of the reviewer. Description is the column of interest. It is the text from the WineEnthusiast review, describing the taste of the wine.

**Data Cleaning and Exploration**

129,971 wines were loaded into a pandas data frame using the read\_json() function. The "designation", "region\_2", "taster\_name", and "taster\_twitter\_handle" columns were deemed unnecessary and removed from the data frame. After these fields were discarded, all rows with missing data were excluded from analysis. The points field was converted to the int data type and the price attribute was set as a float.

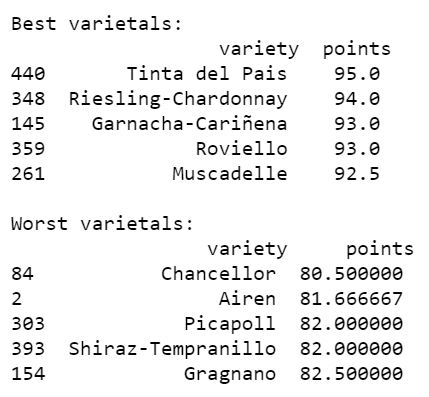
After all of the data cleaning steps, 101,400 rows of wine remained in the data frame. Below is a preview of the first few rows of the cleansed data frame.



**Questions: What are the highest rated and the lowest rated varieties of wine based on average points? What are the most common words used in the descriptions of the best wines and the worst wines?**

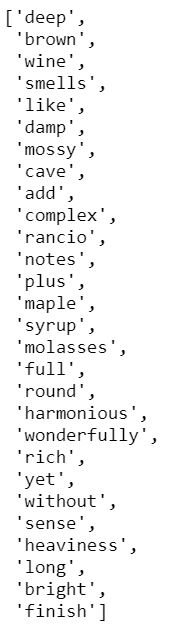
To determine what words were most commonly found in the best and worst wine descriptions, there were a number of steps taken to convert the text from each review into a format in which word frequency could be obtained. In total, there were 509 unique varieties of wine. This analysis focused on approximately the top and bottom 5% of varieties.

First, the wines were sorted by variety and average point score using groupby() and mean(). Next, two data frames were created, one for the highest-scoring wines, and one for the lowest. The top and bottom 25 varieties were stored in their respective data frames. Below are the top five highest-ranking varieties and the bottom five lowest-scoring varieties.

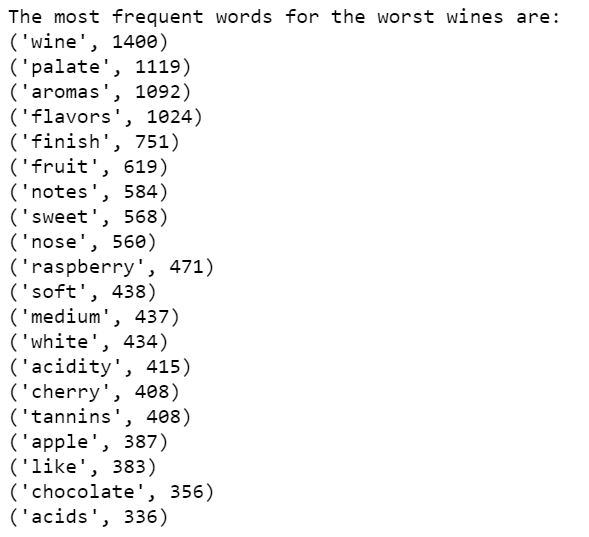
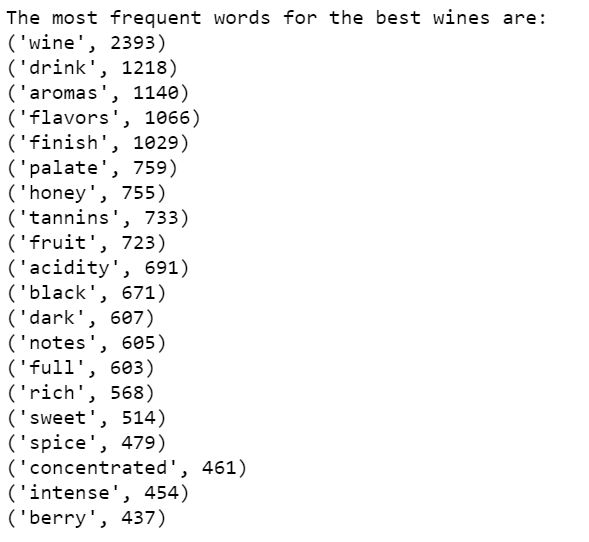


Each data frame was next converted to a list. Then, for only wines with varieties found within the best/worst lists, two new lists were created. These lists contained all the attributes for all the wines that had varieties considered to be the best or worst. These lists were turned back into data frames for further evaluation.

Utilizing the RegexpTokenizer() function, separate lists were created containing only the descriptions from the best and worst wine data frames. The reviews were converted to lowercase and tokenized, meaning each word was broken down and stored as its own component. Additionally, using the stopwords from nltk, the english stopwords were removed from the description. This was done so that non-meaningful words like “and”, “is”, and “it” did not appear as the most common words. Below is the final version of the first review in the best list.



The words from all of the best reviews were combined and stored in a new list called bestWords. LIkewise, the words from all of the worst reviews were combined into a list called worstWords. Finally, the Counter() function was used to determine the highest frequency words in both the best and worst lists. The top 20 words for the best reviewed wines and the worst reviewed wines were gathered using most\_common(). The final output is below.



Unsurprisingly, the most frequent word in both lists is “wine”. Words such as “flavors”, “finish”, “palate”, “tannins”, “fruit” appear in both lists as well. It makes sense that there is a lot of overlap because each list is describing the same thing, wine. Independent of score, it is likely difficult to review and describe the taste of a wine without using the word “flavor”! Additionally, all of these wines are considered to be good wines, as they each scored at least 80 points, so none of them are truly bad wines and the differences may be subtle.

While half the words are identical in both lists, there are some notable differences. Reviews from the top 25 best varieties often contained the words “black”, “intense”, “dark”, “rich”, “concentrated”, and “full”. All of these words are indicative of bold flavor. In contrast, 25 lowest ranking varieties commonly were described with “white”, “soft”, and “medium”. These words are much less powerful descriptors. The lower-scoring wines frequently described specific flavors like “raspberry”, “cherry”, and “apple”, while the list of common words used to describe top-tier wines simply included “berry”.

**Conclusion**

Although wine preference varies from person to person, it seems there are distinct differences in word choice when it comes to good versus excellent wines. When searching for a truly outstanding wine, look for strong words such as “intense” and “rich”.

It’s important to recognize that only about 10% of the varieties in this data set were analyzed and that the entire data set only contained wines that scored at least 80 points. With further analysis, perhaps a predictive model could be created to determine the score of a wine based solely on the description.

Moving forward with the final project, I would like to combine information from the additional wine data set and explore if there are regional differences in word choice. I would also like to create visualizations to better display the final results.