Project 6: Wine Study

Corrections and Clarifications

- Clarification: if the tests say `unexpected EOF while parsing`, you're probably using print instead of sending your output to an `Out[N]` box
- Mar 2: test.py has been fixed for questions 13-15

Introduction

This project is a wine connoisseurs' delight! Data Science can help us understand people's drinking habits around the world. For example, take a look at Mona Chalabi's analysis here: Where Do People Drink The Most Beer, Wine And Spirits?

For our part, we will be exploring a modified subset (the first 1501 rows) of the Kaggle wine reviews dataset;
you will be using various string manipulation functions that come with Python as well as rolling some of your own to solve the problems posed. Happy coding, and remember the Ballmer Peak is nothing but a myth!

Directions

Be sure to do lab 6 before starting this project; otherwise you probably won't get far.

Begin by downloading `wine.csv` and `test.py`. Create a `main.ipynb` file to start answering the following questions, and remember to run `test.py` often. There is no `project.py` this week, though we'll suggest some code in the lab that you can use to access the data. Use the `#qN` format as you have previously.

Q1: which country names are listed in the `country` column of the dataset?

Your output should be in the form of a Python list containing the country names. The tests don't care about the order, but there should should be no duplicate entries in the lists. Also, some country names are missing in the dataset (real-life data is often messy, unfortunately!). Missing values are represented as `None`, but you should make sure `None` does not appear in your answer list.

Now is a good time to run the tests with `python test.py`. If you did Q1 correctly, it should look like this:

```
Summary:
  Test 1: PASS
  Test 2: not found
```

```
  Test 3: not found
  Test 4: not found
  Test 5: not found
  Test 6: not found
  Test 7: not found
  Test 8: not found
  Test 9: not found
  Test 10: not found
  Test 11: not found
  Test 12: not found
  Test 13: not found
  Test 14: not found
  Test 15: not found
  Test 16: not found
  Test 17: not found
  Test 18: not found
  Test 19: not found
  Test 20: not found

TOTAL SCORE: 5.00%
```

Q2: what is the average wine price?

Be careful! There may be missing price information for some rows, so it's best to skip those.

Q3: which wine varieties are produced by Spain?

Answer in the form of a list containing no duplicates (for this and future questions).

Q4: which wineries make wines containing the phrase "beef" in the description?

This shoud match anything containing beef (in any case), regardless of spacing.

Q5: which wineries make wines containing the phrase "zesty" in the description?

Q6: which wineries make wines containing the phrase "black-fruit aroma" in the description?

Q7: which wine varieties are anagrams of the phrase "antibus governance"?

If you liked Professor Langdon's adventures in Da Vinci Code, you'll have fun with this one. :)

An anagram is a word or phrase formed by rearranging the letters of a different word or phrase, using all the original letters exactly once. (Read more here: https://en.wikipedia.org/wiki/Anagram). For our purposes, we'll ignore case and spaces when considering whether two

words are anagrams of each other.

Hint: although you'll need to loop over all the names to check for anagrams, checking whether a single word is an anagram of another word does not require writing a loop. So if you're writing something complicated, review the string methods and sequence operations to see if there is a short, clever solution.

Consider writing a function to solve Q7 and Q8 with the same code.

Q8: which wine varieties are anagrams of the phrase "Banned Petrol Mill".

Q9: what is the highest-rated wine variety made in "US"?

The rating is represented by the `points` column in the dataset.

Your answer should be in the form of a Python list. If there is a single best, that list should contain that single best variety. If multiple varieties tie for best, the list should contain all that tie.

Consider writing a function to solve Q9 and Q10 with the same code.

Q10: what is the highest-rated wine variety made in "Spain"?

Q11: what is the average points-per-dollar (PPD) ratio of the "Heitz" winery?

In this question, we're trying to find the best value using the `points` (the rating) and `price` (cost in dollars) columns.

Be careful! You need to compute the ratio for each wine of the given winery, then take the average of those ratios. Simply dividing the sum of all points by the sum of all prices will calculates the wrong answer.

Q12: what is the average PPD of the "Ponzi" winery?

Q13: which winery in New Zealand has the highest average PPD?

Consider writing a function to answer this and Q14 and Q15 with the same code.

Q14: which winery in Australia has the highest average PPD?

Q15: which winery in Canada has the highest average PPD?

Q16: which wine varieties are produced by the "Quinta Nova de Nossa Senhora do Carmo" winery?

Produce a Python list with no duplicates.

Q17: which wine varieties are produced by the "Adega Cooperativa de Borba" winery?

Q18: which wine varieties are produced by the "Global Wines" winery?

Q19: what percentage of the varieties produced by "Quinta Nova de Nossa Senhora do Carmo" are also produced by "Adega Cooperativa de Borba"?

Quinta Nova wants to better understand their competition, so they
hired a savvy data scientist (you!) to keep an eye on the competition.

Q20: what percentage of the varieties produced by "Quinta Nova de Nossa Senhora do Carmo" are also produced by "Global Wines"?

Cheers!