BMIS2542-Data Programming Essentials with Python

AY 2019-20 Spring

Assignment-1: Data Wrangling

Notes:

- *Due Date: Feb. 12, 2020 at 5:30pm. Submit your solution Notebook through Canvas.*
- *Pitt Academic Integrity rules apply for this assignment. Follow the honor system and answer the questions by yourself (or with a partner when working in a group of two students).*

Use the "worldvalues-data.csv" and the corresponding data dictionary posted on Canvas for this assignment. Read the data into Python as a Pandas dataframe, use appropriate data wrangling techniques, and answer the following questions. You may use any libraries available in the Python ecosystem.

1. Comment on the trends of missing values[1] in the data. For example,
    a. Are the missing values from respondents increasing over the years?
    b. How do the proportion of records with missing values vary over the different countries?
    c. Create a new Pandas dataframe with the following columns, populate the dataframe with appropriate values, and write it out as a CSV file.

Survey year	Continent	Country	Total count of respondents	Proportion of respondents with missing responses for more than five questions

    d. Identify and comment on any other missing value patterns that stand out to you

2. Use respondent's literate/illiterate status (V255) to answer the following questions:
    a. How does the proportion of respondents who are illiterate vary across countries?
    b. Are there differences in the religious beliefs[2] between literate and illiterate respondents? Does the extent of this difference vary across countries? Across continents?
    c. Examine the data to identify other noticeable differences between literate and illiterate respondents.

3. Considering only United States data, answer the following questions:
    a. Derive the absolute rank and the rank in percentile terms for the different U.S. states featured in the dataset (V256B) according to:
        i. Overall satisfaction with life (V23)
        ii. Confidence in the federal government (V115)
        iii. Job worries (V181)
        iv. Abortion belief (V204)
        v. Computer use (V225)
    b. Create a pivot table with U.S region (V256), U.S. State (V256B), and the above five variables in Q3a as the indices. The values in the table must be the average values of the

---

[1] In this context "missing values" refers to both NaN values and responses to survey questions that have been coded as "Missing" (or its numerical code, for example, "-5" for the question V4).

[2] Look through the data dictionary to identify appropriate variables that correspond to religious beliefs.

respective variables. Query the pivot table for values of (1) all “Middle Atlantic States” and (2) for Pennsylvania.

c. Are there significant differences in the respondents’ values regarding environment (e.g., V30, V78, V80, V81, V83, V122) across people living in the different regions of the U.S. (see V256 and V256B)?
d. Identify and comment on the differences between the groups of respondents who identified themselves as “I am born in this country” and those who identified as “I am an immigrant to this country” (V245)?
e. Identify and comment on the differences between the groups of respondents who identified themselves as voting for the Republican party and those who identified as voting for the Democrat party (V228)?

4. Use “SACSECVAL” to answer the following questions:
    a. Insert a column and call it “secular_category”. Populate the values for the new column as following:

secular_category	SACSECVAL
Low	0 to 0.3
Medium	Greater than 0.3 and less than <0.7
High	Equal to and greater than 0.7

    b. List the countries where the proportion of respondents in the “Low” secular_category is greater than the “Medium” and “High” categories.
    c. Check if there are differences in the distribution of respondents in the secular_category across different regions of the world: across continents, east vs. mid-east vs. west, and northern vs. southern hemispheres.

5. Use the gender of the respondents (V240) to identify if there are significant gender differences in the values and beliefs regarding “gender equality.” Is there heterogeneity in those differences across other socio-economic (e.g., age, income, religious beliefs) or geo-political variables (e.g., country, democracy)?