BMIS2542 Fall 2020

Midterm Project Part-1: Data Wrangling and Exploratory Analysis

Use the “worldvalues-data.csv” and the corresponding data dictionary posted on *Canvas* for this project. **Read the data into Python as a Pandas DataFrame and use appropriate data wrangling techniques to answer the following questions. Use effective visualizations to supplement your answers.** You may use any libraries available in the Python ecosystem.

1. Comment on the trends of missing values in the data. Supplement your answers with suitable graphs whenever possible.

   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).

   a. Are the missing values from respondents increasing over the years (Survey Year-V262)?
   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	Country	Total count of respondents	Proportion of respondents with missing responses for more than five questions

2. Use respondent’s literate/illiterate status (V255) to answer the following questions. Supplement your answers with suitable graphs whenever possible.
   a. How does the proportion of respondents who are illiterate vary across countries?
   b. Are there differences in the religious beliefs[1] between literate and illiterate respondents? Does the extent of this difference vary across countries?
   c. Examine the data to identify other noticeable differences between literate and illiterate respondents.

3. Considering only United States data, answer the following questions. Supplement your answers with suitable graphs whenever possible.
   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. Job worries (V181)

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

b. Create a pivot table with U.S. State (V256B) as index and gender (V240) as the columns. The values in the table must be the average values of the overall satisfaction with life (V23). Query the pivot table 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)?

4. Use "SACSECVAL" to answer the following questions. **Visualizations are not required for this question.**

   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 **continents** in the world. Use the given country-continent.csv file to do this.

5. Based on this dataset, come up with an investigative idea of your own. Briefly explain why you find this question interesting (bullet points would do). Use data wrangling techniques to examine the data and answer your question. Illustrate with effective visualizations.