Homework 04 Instructions STAT/CS 287

Congratulations! You have just been contracted by a non-profit organization dedicated to novel health sciences. You have been given a newline-delimited JSON dataset containing information on a collection of individuals participating in a diabetes study. This non-profit intends to use this study to **develop a smartphone app to detect and diagnose diabetes mellitus!**

Please read these instructions carefully before you begin!

Show your work!

This means (i) provide the code you have produced, (ii) as part of any answers in your writeup include a statement indicating relevant portion(s) within your code. For example, parenthetical statements such as, *"(See plot_timeseries.py, lines 100-110.)"* may be acceptable indicators, using appropriate filenames as needed, of course.

To submit

1. Prepare your files and compress your HW04 working directory as per the "preparing and submitting your homework" slides. Make sure the zipped file includes your `HW04_[NETID].py` script file (properly renamed), your writeup, and any other files you may have generated while completing the assignment. Please do not include the original data in your submission.
2. Upload your zipped file to Blackboard. No other files should be submitted.

Please follow all instructions and address all the comments in the `HW04_[NETID].py` file.

Dataset and code constraints

Unfortunately, the circumstances of this job impose some constraints on your work. Because we want your code to work on smartphones, which may have limited computing resources, you may **only use the modules already imported in the provided Python script**. Further, as the non-profit *does not have extensive legal resources*, we do not want any patent or copyright liability on your code. Therefore, to fulfill your contract, you must provide 100% new, self-written code for all tasks given to you: ***No online code resources may be used***.

(Your assignment will be returned *ungraded* if these conditions are not met. Please ask questions if you need clarifications on these constraints.)

P1. Data acquisition and summarization

- P1.1 - Create a function `load_data()` that reads this file into python. Describe how to use this function in a docstring and make sure the code is clear and concise. (Choosing an appropriate data structure here will make the rest of the assignment must easier.) In your writeup, please provide a narrative describing how this function works (not how to use it).
- P1.2 - Provide summary statistics (as you see fit) in tables for this dataset, describing all the columns and what you may or may not understand about them. The "key" file is useful here as well. Please order your answers for each column in the same order as the key file, and use either a numbered list or subsections to organize your answer within this section of the writeup.

P2. Missing data detection

Your analysis for Problem 1 may reveal an important concern: data are missing!

- P2.1 - Find all the missing data, as best as possible given the information available about the data. Report in your writeup on how many observations are missing for each variable. Create a function `flag_missing_values()` which reads in the original dataset and returns a **copy** delineating each missing entry with a Python "`None`".
- P2.2 - Create a function `listwise_deletion()` that reads in the "flagged" dataset (created by `flag_missing_values()`), and returns a copy sanitized via the listwise deletion method. Report summary statistics (as in Problem 1) on this dataset and, specifically, contrast it with the full dataset (what is similar, what is different, etc.). Please order your answers here to match those of P1.2. Report the (Pearson) correlation coefficient between each pair of variables on the sanitized dataset.

P3. Missing data imputation

(You may use `matplotlib.pyplot` for P3 only.)

- P3.1 - Perform marginal mean imputation on the missing values for each observation. Make scatterplots for each pair of variables showing their associations and use different colors/symbols to highlight the imputed values compared with the original. Interpret this method of imputation. Report the Pearson correlation coefficient between variable pairs in the imputed dataset, and compare with the correlation coefficients measured in Problem 2.2.
- P3.2 - **(Bonus for undergraduates; required for graduate students)** An important aspect of missing data is **patterns** in the missingness. Perform an analysis, statistical, visual, etc. to see if the presence or absence of a variable having a missing value is associated with the values of other variables that are not missing. In your writeup, report your analysis and answer the following questions: Can you conclude if the data are MCAR? If so, why? If not, why not?