Homework 05 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!**

Instructions

Please read this entire document completely before you begin.

To begin, create a HW05 enclosing folder, rename the necessary files as appropriate, and place these files into your enclosing folder. As you complete the code portions of the problems below, please also complete your corresponding written answers. Please note:

- You must ***show your work*** for all questions. This means answering questions using code, statistics, code *and* statistics, and so forth, and it means describing and synthesizing your answers.
- For any written answers supported with computations, be sure to ***refer to the specific filename(s) and line number(s)*** for the code that supports your written answer(s). For example, *"A is superior to be B except when C is true (results.py: lines 101-121)"*.
- Avoid waiting until the last minute to begin working on the assignment.

To submit:

1. Prepare your files and compress your HW05 working directory as per the "preparing and submitting your homework" slides. Make sure the zipped file includes your `HW05_[NETID].py` script file (properly renamed), your writeup, and any other files you may have generated while completing the assignment. Note that this assignment does not provide a filechecker.py. 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 `HW05_[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.)

Problem 1: 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 (see below) in (properly formatted and captioned) tables for this dataset, describing all the columns (variables) 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.

P1.3. Present graphically the probability distribution (PDF or PMF) of each variable in the data, as appropriate and by using the techniques discussed in class. When the variable allows it, supplement these graphics with plots of the ECDF, as also discussed in class. When computing ECDFs, be sure to use the proper technique discussed in class; no credit will be given for computations of the ECDF that do not show all available data. Order your figures within your write-up by variable name in the same order as the key file and ensure (with captions) that each plot clearly identifies which variable it illustrates.

Take care choosing your summary statistics as the appropriateness of your summary statistics will be a significant factor of this problem's grade. Either omitting appropriating statistics or including inappropriate statistics will incur points loss. Readability is also an important factor: correct but badly presented or difficult-to-understand results will also incur points loss.

Problem 2: 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 write-up 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.

Problem 3: Missing data imputation

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 distinguish the imputed values from values that did not need imputing. 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/Required]** 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 write-up, report your analysis and answer the following questions: Can you conclude if the data are MCAR? If so, why? If not, why not?