Homework Assignment #4

CS 3753/5163

Individual work

Submit one Jupyter notebook named as yourLastName_HW04 with appropriate code cell to solve the following questions. Make sure that each question is included in your notebook as a markdown cell above your answer. You must use only the basic Python, math module, and numpy library. Include any specific direction/instruction to run your script in comments.

Note: if you want to include your script as separate .py files, you may submit all source files with a Jupyter notebook in a zipped file (compressed). In the notebook, inside the code cell of a question, write %run Qx. Where x is the number of the question.

Q1) (40 points) Write a Python script to perform the tasks in Q1 and Q3 from homework assignment 3 using numpy. Your program should accept the path to the file to read (example: C:\Documents\random.txt). Don't submit the provided input file (random.txt) with your notebook.

Q2) (60 points) In this question we will be analyzing the Behavioral Risk Factor Surveillance System (BRFSS) weight vs. height data. The data can be found in the fixed-width ASCII file called CDBRFS08.ASC.gz. For this analysis, we are interested in five (5) pieces of data: age, current weight (cw), weight a year ago (waya), height, and gender. Please refer to the guideline of the given dataset at https://www.cdc.gov/brfss/annual_data/2008/pdf/codebook08.pdf. Based on the guideline, these data can be found in the following columns: age (101-102), current weight (119-122), weight a year ago (127-130), height (123-126), and gender (143,143).

Your program should accept the path to the file to read (example: C:\Documents\CDBRFS08.ASC.gz). Don't submit the provided input file (CDBRFS08.ASC.gz) with your notebook.

Use the following line code to open the file as zipped file: gzip.open(yourFile, 'rt'). Where 'rt' is used to open the file and read it as a text file. Import gzip to be able to use the command.

Create a numby array of five (5) columns to maintain the data.

Clean the data by removing any invalid or missing entry. Refer to the guideline for more information about the invalid/missing data. Delete all rows contain any invalid/missing data.

For example: The only valid entries for cw are the values 50-0999 and 9000-9998.

Value	Value Label
50 - 0999	Weight (pounds) Notes: 0 _ _ _ = weight in pounds
7777	Don't know/Not sure
9000 - 9998	Weight (kilograms) Notes: The initial '9' indicates this was a metric value.
9999	Refused
BLANK	Not asked or Missing

Convert weights to kg (lb/2.2) and round it up to 1 decimal point. Convert the height to centimeters (feet*30.48 + inches*2.54) and round it down to the integer number (<=0.5 truncate, otherwise round the fraction up).

a) (**10 points**) Your final cleaned/converted array will have 385,974 entries/rows. Here are the first 15 rows of the array for your reference. Print same information.

```
npdata[0:15], len(npdata), npdata.dtype

(array([[ 39. ,  88.6,  88.6, 180. ,   1. ],
        [ 64. ,  75. ,  84.5, 155. ,   2. ],
        [ 51. , 100. , 100. , 183. ,   1. ],
        [ 35. ,  63.6,  61.4, 170. ,   2. ],
        [ 62. ,  70.5,  70.5, 173. ,   2. ],
        [ 64. ,  63.6,  63.6, 157. ,   2. ],
        [ 55. ,  82.7,  82.7, 155. ,   2. ],
        [ 71. ,  59.1,  56.8, 155. ,   2. ],
        [ 21. ,  81.8,  86.4, 180. ,   1. ],
        [ 45. ,  90.9,  90.9, 165. ,   2. ],
        [ 53. ,  51.8,  51.8, 163. ,   2. ],
        [ 51. ,  56.8,  59.1, 155. ,   2. ],
        [ 70. , 105.5, 112.7, 170. ,   2. ],
        [ 59. ,  84.1,  84.1, 165. ,   2. ],
        [ 59. , 131.8, 129.5, 191. ,   1. ]]), 385974, dtype('float64'))
```

b) (**10 points**) Produce summary statistics for cw, waya, and height (mean, standard deviation, range, and median). Round the final answer to two (2) decimal places.

Sample output:

cw:

mean: 79.06 std: 19.51 range: 280.0 median: 77.3

waya:

mean: 79.8 std: 20.58 range: 319.6 median: 77.3

height:

mean: 169.01 std: 10.39 range: 175.0 median: 168.0

c) (**5 points**) How many entries are females younger (<) than 40?

Sample output: Number of females under 40: xxxx

d) (**5 points**) How many male is within 1 std (<=) in height from the mean of the entire set and from the mean of male entries?

Sample output:

Number of males within 1 std in height from the mean of entire set: xxxxx

Number of males within 1 std in height from the mean of males: xxxxx

e) (**5 points**) How many outlier entries we have for waya for females?

Sample output: Number of outlier female entries in waya: xxxxx

f) (**5 points**) Find the coefficient of variation for cw.

Sample output: Coefficient variation (CV) for cw: xx.xx%

g) (**5 points**) Which group has less variation in cw, male or female? Why? Support your answer with numbers.

Sample output: female/male group has less variation Print out the numbers you used to conclude.

h) (**5 points**) Do males tend to gain more weight when they are older than (>) 40 compared to those who are younger? Use cw to find out. Support your answer with numbers.

Sample output: yes/no. Because …… print out the numbers you used to conclude.

i) (**10 points**) Define weight change (delta_w) = cw – waya. Calculate correlation between delta_w and the following variables, and determine which one is most correlated (regardless of the sign of correlation) with delta_w:
a) cw, b) waya, c) height, and d) age.

Sample output: the most correlated variable is x because …… here are the numbers

Corr. Coef. delta and cw: x
Corr. Coef delta and waya: x
Corr. Coef. delta and height: x
Corr. Coef. delta and age: x

Due date: 03/28/2019 at 11:59PM

How to submit

Through blackboard. No hard copy is accepted. The system will close after 11:59PM and you will not be able to turn it in. No late submission is accepted unless you receive instructor's approval no less than two days before the due date.