**CAP 4784 – Introduction to Data Analytics**

Lab 4 – Analyzing BRFSS Data Assignment

**Submit Python script for the below-specified assignment description.**

In this lab assignment, we will be analyzing the BRFSS weight vs. height data. Download data file from Canvas. The data file has five data attributes: age, current\_weight (lbs), weight\_a\_year\_ago (lbs), height (inches), and gender, where gender == 1 represents male and 2 represents female.

To complete this assignment, you need to perform the following tasks using Python:

1. Load the data file to create a NumPy array
2. Display the first five rows and shape of the array
3. Calculate weight change. Weight change is defined as the difference between current weight and weight a year ago.
4. Calculate and display mean, median, standard deviation, and interquartile range for the weight change
5. Concatenate weight change array with the main data array
6. Display the five rows and shape of the concatenated array
7. Split the concatenated array based on the gender column
8. Display the five rows and shape of the array relevant to males data
9. Calculate and display mean, median, standard deviation, and interquartile range for the data relevant to males
10. Display the five rows and shape of the array relevant to females data
11. Calculate and display mean, median, standard deviation, and interquartile range for the data relevant to females

Use the sample output provided below to design your Python program for the assignment. After the sample output section, I have provided some notes that should help complete some of the above tasks.

# Sample Output

First Five Rows of the Data:

[[ 1 77 175 175 70 1]

[ 2 33 125 132 64 2]

[ 3 49 105 105 60 2]

[ 4 42 132 136 66 2]

[ 5 55 150 157 61 2]]

Shape of the data: (20000, 6)

Descriptive Statistics for Weight Change Data:

Mean: -7.38

Median: -3.0

Standard Deviation: 14.68

Interquartile Range: 10.0

First Five Rows of the Data with Weight Changes:

[[ 1 77 175 175 70 1 0]

[ 2 33 125 132 64 2 -7]

[ 3 49 105 105 60 2 0]

[ 4 42 132 136 66 2 -4]

[ 5 55 150 157 61 2 -7]]

Shape of the data: (20000, 7)

First Five Rows of the Data relevant to Males:

[[ 1 77 175 175 70 1 0]

[ 7 31 194 203 71 1 -9]

[ 8 45 170 177 67 1 -7]

[ 10 44 180 183 70 1 -3]

[ 11 46 186 188 69 1 -2]]

Shape of the data: (9569, 7)

Descriptive Statistics for Data relevant to Males:

Mean: 1487.24

Median: 71.0

Standard Deviation: 4074.41

Interquartile Range: 195.0

First Five Rows of the Data relevant to Females:

[[ 2 33 125 132 64 2 -7]

[ 3 49 105 105 60 2 0]

[ 4 42 132 136 66 2 -4]

[ 5 55 150 157 61 2 -7]

[ 6 55 114 114 64 2 0]]

Shape of the data: (10431, 7)

Descriptive Statistics for Data relevant to Females:

Mean: 1499.01

Median: 66.0

Standard Deviation: 4127.76

Interquartile Range: 158.0

# Notes

## About Data File

The data file used in this lab assignment is relevant to the Behavioral Risk Factor Surveillance System (BRFSS). The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States. As its name implies, the BRFSS is designed to identify risk factors in the adult population and report emerging health trends. The BRFSS Web site (<http://www.cdc.gov/brfss>) contains a complete description of the survey, including the research questions that motivate the study and many interesting results derived from the data.

We will focus on a random sample of 20,000 people from the BRFSS survey conducted in 2000. While there are over 200 variables in this data set, we will work with a small subset of five variables.

## Loading Data File

To read CSV data into a record array in NumPy, you can use the genfromtxt() function. In this function’s argument, you need to set the delimiter to a comma to read a CSV file. You can set skiprows=1 so that the first line with header information is not recorded into the NumPy array. genfromtxt() function returns a NumPy array.

Assuming that you have a CSV file named myfile in the same folder as the Python file, the following would be the code to read the CSV file and create a NumPy array.

import numpy as np

my\_data = np.genfromtxt('myfile.csv', delimiter=',', skip\_header=1)

Review genfromtxt function documentation for more information on other parameters. For example, you can read the below docs link to learn how to set the resulting array’s data type.

NumPy doc for genfromtxt function:

<https://numpy.org/doc/stable/reference/generated/numpy.genfromtxt.html>

## Concatenating NumPy Arrays

For concatenating two arrays, refer to numpy.concatenate() function: <https://numpy.org/doc/stable/reference/generated/numpy.concatenate.html>

numpy.concatenate function accepts a sequence of arrays that are of the same shape. When you want to concatenate a 1-D array with a 2-D array, consider using numpy.column\_stack() function. numpy.column\_stack() function takes a sequence of 2-D array and 1-D array and concatenates 1-D array as a column to the 2-D array. Review the sample code below to understand how to use the column\_stack function.

Code:

import numpy as np

a = np.array([[1, 2], [3, 4], [5, 6]])

b = np.array([7, 8, 9])

print("Two dimensional array:")

print(a)

print()

print("One dimensional array:")

print(b)

print()

c = np.column\_stack((a, b))

print("One dimensional array column wise concatenated to Two dimensional array:")

print(c)

Output:

Two dimensional array:

[[1 2]

[3 4]

[5 6]]

One dimensional array:

[7 8 9]

One dimensional array concatenated to Two dimensional array:

[[1 2 7]

[3 4 8]

[5 6 9]]

## Splitting NumPy Array based on a Column value

To split a NumPy Array into subarrays based on values of a column, you will find the below code useful:

y = [x[x[:,1]==k] for k in np.unique(x[:,1])]

In the above code, on the rightmost end, you can note that we are using the unique function to get unique values in the second column of array x, which are stored in variable k, which is then used to slice the array x based on the k matching values of the second column.

Review below sample code implementation and output to understand how a NumPy array can be split into subarrays based on values of a column:

Code:

import numpy as np

a = [1, 1], [5, 2], [3, 1], [4, 2], [2, 1]

print('list of list as input:')

print(a)

print()

x = np.array(a)

print('array that needs to be splitted based on second column values')

print(x)

print()

y = [x[x[:,1]==k] for k in np.unique(x[:,1])]

print('splitted array as array of arrays')

print(y)

print()

print('array with value 1')

print(y[0])

print()

print('array with value 2')

print(y[1])

Output:

list of list as input:

([1, 1], [5, 2], [3, 1], [4, 2], [2, 1])

array that needs to be splitted based on second column values

[[1 1]

[5 2]

[3 1]

[4 2]

[2 1]]

splitted array as array of arrays

[array([[1, 1],

[3, 1],

[2, 1]]), array([[5, 2],

[4, 2]])]

array with value 1

[[1 1]

[3 1]

[2 1]]

array with value 2

[[5 2]

[4 2]]