Assessment 3: Data processing and exploration project

**Assessment type:**Jupyter Notebook and project report

**Individual/group assessment:**Individual

**Word count/time limit:**Do not exceed 2000 words

### **Assessment description**

This final assessment is a small individual project involving data processing, analysis and visualisation. The project requires students to demonstrate python data science skills and techniques in data processing and exploration on real-world datasets with consideration of data ethics.

In this project, you will be given a dataset for data processing, analysis and visualisation.

We will be analysing the BRFSS (brfss.csv) weight vs height data. The Behavioral Risk Factor Surveillance System (BRFSS) is the nation’s premier system of health-related telephone surveys that collect state data about US residents regarding their health-related risk behaviours, chronic health conditions, and use of preventive services. Weight and height have been queried in a telephone interview.

**The data for this project: see Filename Assessment\_3\_data\_bfss.csv – USE PANDA to view**

The six columns in the data represent: age, current\_weight (kg), weight\_a\_year\_ago (kg), current\_weight with two decimals(kg), height (cm), and gender, where gender == 1 represents male and 2 represents female.

In this project, you are required to have insightful discovery about the data via initial exploratory and visualisation with the learned skills from this unit.

### **Assessment details**

Attempt the tasks below with the given dataset, at the same time, reflect on the development and applications of data science while ensuring the respect of human rights and of the values shaping open, pluralistic and tolerant information societies.

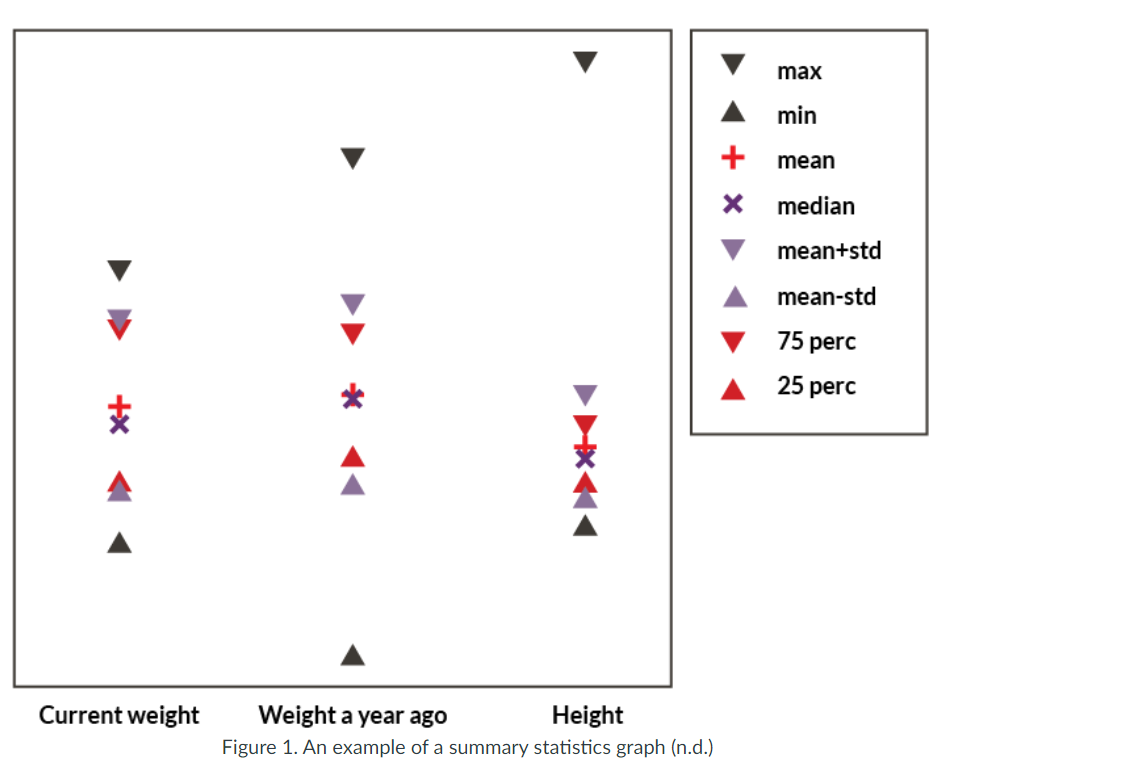
#### **Prepare a Jupyter Notebook for Tasks 1-3 of this project**

The structure of the Jupyter Notebook should alternate texts and python codes and cover topics listed in the following specific tasks. You may use this template to complete and submit these tasks 1-3 of this assessment:

**See FILENAME: Assessment 3.ipynb** **- FORMAT TO FOLLOW**

#### **Task 1 (10 marks)**

Produce a summary statistics graph on current\_weight, weight\_a\_year\_ago, and height. **[Hint: similar to figure 1 below]**

  
Figure 1. An example of a summary statistics graph (n.d.)

#### **Task 2 (10 marks): Calculate correlation**

Define weight\_change = (current\_weight – weight\_a\_year\_ago). Calculate the correlation between weight\_change and the following variables, and determine which one is most correlated (regardless of signs of correlation) with weight\_change. Use **scatter plots** to support your conclusion.

i. current\_weight  
ii. weight\_a\_year\_ago  
iii. age

**[Hint: One scatter plot for each variable.]**

#### **Task 3 (5 marks)**

Use t-test to check the significant difference

3.1(1 mark) Use t-test to test whether there is a significant difference between the weight\_change of male and female.

3.2 (1 mark) Randomly split the subjects into two groups of roughly equal sizes, and use t-test to test whether there is a significant difference between the weight\_change of the two groups.

3.3 (1 mark) Repeat the process in 3.2 1000 times and plot the distributions of the -log10(p-value) of the t-test results.

**[Hint: the x-axis is the number of experiments from 1 to 1000, and the y-axis is -log10(p-value). There should be two distributions: One for each group. Use seaborn.displot method.]  
What can you say about the difference between male and female in terms of their weight\_change? (Consider both the p-value and the absolute differences between the two means.)**

3.4 (1 mark) Define weight\_height \_ratio as current\_weight/height. Use t-test to test whether there is a significant difference between the weight\_height\_ratio of male and female.

3.5 (1 mark) Also, repeat the analysis you did in 3.4, but replace weight\_height\_ratio with weight\_change in your analysis.

**[Hint: use**[***t-test (Links to an external site.)***](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html)**here]**

[***t-test (Links to an external site.)***](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html)

#### **Prepare a report**

#### **Task 4 (15 marks)**

Write a report summarising insightful discoveries about the output and figures of the source code. The report should contain at least three sections, each should address a discussion of one of the above tasks. An extra conclusion/summary section could also be included. The ‘Structure’ section below gives an example of the report skeleton you could follow. Use Microsoft Word to create this report, save as a PDF, and submit as requested below.

|  |
| --- |
| **Title: Exploring BRFSS data****Introduction:** A brief introduction about the dataset, e.g., the background of the data (search from the web) and how many rows and columns it has. **Section 1: Summary statistics analysis** In this section, describe the statistics you used in your code and then attach the figure you’ve obtained from the visualization. Then, do a comparison of the statistics for the three variables. **Section 2: Correlations analysis** In this section, describe the formula you used to calculate the correlation in your code and present the numbers and scatter plots you obtain for the correlations between weight change and the three factors mentioned as an example in Task 2. Then do a quick comparison for the three factors. **Section 3: Significant difference analysis** In this section, summarise the results you obtain for each task in Task 3. **Section 4: Conclusion** Sum up your overall insights about this dataset. |

#### **Some helpful websites and resources**

* [Anaconda (Links to an external site.)](https://docs.anaconda.com/anaconda/) environment
* [Python (Links to an external site.)](https://pandas.pydata.org/) official website

Useful Python packages:

* [NumPy (Links to an external site.)](https://numpy.org/)
* [Pandas (Links to an external site.)](https://pandas.pydata.org/)
* [Matplotlib (Links to an external site.)](https://matplotlib.org/)

#### **Unit learning outcomes**

This assessment is linked to the following learning outcomes:

* **Learning outcome 1:** Appraise the use of data processing, analysis and visualisation techniques and tools to solve real-world data science problems.
* **Learning outcome 2:** Examine data science ethical issues as they impact human dignity and privacy.

#### **Graduate attributes**

* GA 3: Apply ethical perspectives in informed decision-making
* GA 4: Think critically and reflectively
* GA 5: Demonstrate values, kno