SCHOOL OF SCIENCE & TECHNOLOGY
EEET2482 – SOFTWARE ENGINEERING DESIGN
COSC2082 – ADVANCED PROGRAMMING TECHNIQUE

ASSIGNMENT 1 – A C++ STATISTICS PROGRAM

SPECIFICATIONS

You are required to write a C++ program which performs a set of **descriptive** and **inferential** statistics. The program takes a data file name (e.g. *data.csv*) as input and automatically return a set of descriptive statistics and as described below.

A. Input dataset.

You are provided with a csv file (e.g. data.csv) that contains data for two variables (**x** and **y**). Data for each variable is in column-base wit the first line stores the name of the variables. Below is a sample set of data you might receive.

x,y
1,2
3,4
…

B. Descriptive Statistics. (60 marks)

Below are some descriptive statistics that one can compute given a variable. Note that your dataset has 2 variables (2 columns of data). You will need to compute the below statistics for both variables in your provided dataset.

1. *median* – Median is the middle value of a set of ordered numbers. Example: 2, 5, 9, 7, 5, 4, 3 – after reordering, the middle number is 5, so the median = 5. **(5 marks)**

   You program will output the median values for both variables x and y in the following format.

   ```
   median_x=__ - median_y=__
   ```

2. *mode* – Mode is the most frequently occurring element in a dataset. **(10 marks)**

   You program will output the mode values for both variables x and y in the following format. Note that you could have more than one most frequently occurred element in the dataset.

   ```
   mode_x={__} - mode_y={__}
   ```

3. *Variance and standard deviation* – The variance or standard deviation is the most common measure of the spread of a set of points and tell us how much the actual values differ from the mean. The formula for variance is below. **(10 marks)**

   $$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

   You program will output the variance and standard deviation values for both variables x and y in the following format.

   ```
   var_x=__ - var_y=__
   stdev_x=__ - stdev_y=__
   ```

4. *Mean Absolute Deviations (MAD)* – The mean absolute deviation of a dataset is the average distance between each data point and the mean. **(10 marks)**

   $$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

   You program will output the MAD values for both variables x and y in the following format.

   ```
   mad_x=__ - mad_y=__
   ```

5. *First quartile (Q1)* – is the median of the lower half of the data set. This means that about 25% of the numbers in the data set lie below Q1 and about 75% lie above Q1. **(10 marks)**

You program will output the Q1 values for both variables x and y in the following format.

```
q1_x=__ - q1_y=__
```

6. *Skewness(x)* – Returns an estimate of the skewness of a variable x. Skewness is a measure of the asymmetry of the distribution. A positively skewed distribution has a thicker upper tail than lower tail, while a negatively skewed distribution has a thicker lower tail than upper tail. A normal distribution has a skewness of zero. **(10 marks)**

Skewness(x)=$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i-\bar{x}}{s}\right]^3$

You program will output the skewness values for both variables x and y in the following format.

```
skew_x=__ - skew_y=__
```

7. *Kurtosis(x)* – Kurtosis is a measure of the peakedness of a distribution. A distribution with long thin tails has a positive kurtosis. A distribution with short tails and high shoulders, such as the uniform distribution, has a negative kurtosis. A normal distribution has zero kurtosis. A constant value (with no variation) has a kurtosis of -3. **(5 marks)**

kurtosis(x)=$(\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i-\bar{x}}{s}\right]^4 - 3$

You program will output the kurtosis values for both variables x and y in the following format.

```
kurt_x=__ - kurt_y=__
```

C. Inferential Statistics (40 marks)

1. *Covariance* – Covariance measures how much the movement in one variable predicts the movement in a corresponding variable. **(10 marks)**

You can find the formula to calculate the covariance between variable X against variable Y below.

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})\,(Y_i - \bar{Y})$$

You program will output the covariance value between variables x and y in the following format.

```
cov(x_y) =__
```

2. *Pearson correlation coefficient (bivariate correlation or just **correlation coefficient**)* – is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. **(10 marks)**

You can find the formula to calculate the correlation coefficient between variable X against variable Y below.

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{(N\sum X^2 - (\sum X)^2)(N\sum Y^2 - (\sum Y)^2)}}$$

Your program will output the correlation coefficient value between variables x and y in the following format.

```
r(x_y) =__
```

3. *Linear Regression* – Linear Regression is a Regression Analysis technique that models and analyzes the relationship between a dependent variable (x) and an independent variable (y). It is widely used for prediction and forecasting. **(20 marks)**

In this part of the assignment, you are required to compute a simple linear function of the form:

$y = ax + b + \varepsilon$

which describes a line with slope *a*, y-intercept *b*, and an error term ε. The goal is to find estimated values *a* and *b* which would provide the '**best fit**' in some sense for the data points in the provided dataset. Let:

- *mean_x* is the mean of variable X (which is computed in section B.1)
- *mean_y* is the mean of variable Y
- *stdev_x* is the standard deviation of X (which is computed in section B.4)
- *stdev_y* is the standard deviation of Y
- *r* is the Pearson correlation coefficient between X and Y (which is computed in section C.2)

Below are the formulas to compute a and b, respectively.

The slope (a) can be calculated as follows:

$$a = \frac{r \times stdev\_y}{stdev\_x}$$

and the intercept (b) can be calculated as

$$b = mean\_y - a \times mean\_x$$

Your program will output the linear regression fit between variables x and y in the following format.

```
y = ax + b
```

D. Program Operation.

From the command line, your statistics program can be executed as followed:

1. *assignment1_group<TT>.exe <filename>.csv*
    a. *where <TT> denotes your group number*
    b. *<filename> denotes the input file name for the statistics operations*
2. Once running, the program will automatically load the data from the csv file and output both descriptive statistics and inferential statistics. Output format can be found in sections above.
3. Before the program exits, each student ID string must be displayed to the console in the following form.
   ASSIGNMENT 1 GROUP<TT>
   sWWWWWWW, sWWWWWWW@rmit.edu.vn, FirstName, LastName
   sXXXXXXX, sXXXXXXX@rmit.edu.vn, FirstName, LastName
   sYYYYYYY, sYYYYYYY@rmit.edu.vn, FirstName, LastName
   sZZZZZZZ, sZZZZZZZ@rmit.edu.vn, FirstName, LastName

OTHER SPECIFICATIONS

1. Your program should at least be compilable in Microsoft Visual Studio 2017. If your program does not execute at all, **you will only be eligible for 50% of your laboratory mark**. The teaching staff will NOT be fixing code to make programs compile or for debugging issues during assessment.
2. Your group leader, as stated by Canvas, is responsible for submitting the group's work prior to the deadline. **Late submissions will incur a penalty of 10% per day**. **Submissions which are three days past the deadline will not be accepted and a grade of zero will be given**. A zip file of your C++ code and the exe file (i.e. *assignment1_group<TT>.zip,* and a word document of your report (i.e. *assignment1_group<TT>.pdf)*, will need to be submitted to Canvas for assessment, where <*TT*> denotes your group number. Your report will be checked through Turnitin to ensure academic integrity is maintained.
3. Follow the structure on the following page to write your report.
4. **No libraries, except for the followings <iostream>, <fstream>, <string>, and <math.h>** can be used – penalties will apply if other external libraries are used.

REPORT STRUCTURE

1. Introduction & body:
   - Brief of your team and team member.
   - Clearly specify which components of the assignment you have completed (and not completed)
   - Show how you have tested the finished components. This part should have captured outputs of your program in operation.
2. Flowcharts
   - Provide flowcharts which depicts the algorithms for at least 3 statistics functions required in this assignment.
   - Construct your flowcharts from the viewpoint that another person should be able to follow it and write their own software from it.

3. Conclusions
   - Discuss any issues you came across and how you solved them.
   - Should be no longer than 1 paragraph.
   - NOT a section for repeating your activities or re-writing/paraphrasing the laboratory notes.
4. References
   - Any references you used must be placed here.
   - IEEE referencing style must be followed.