Description of the Data files

The data used in this Project are stored in an EXCEL worksheet named **ProjectData& 2020.3**. '& stands for your data set number' It can be downloaded from vUWS.

These data comprise of the following variables:

Question 1 worksheet	
Column1 (Age)	Age (in years) of customers who ordered a product advertised on television
Column2 (Gender)	Gender (Male or Female) of the customers
Question 2 worksheet	
Column1 (Household)	Selected Sydney household
Column2 (Transport)	Annual total transport costs per household (in $1000s)
Question 3 worksheet	
Column1 (Online)	Final marks in Statistics received by students who underwent fully online.
Column2 (Blended)	Final marks in Statistics received by students who underwent Blended Learning method
Question 4 worksheet	
Column1 (Rating)	Rating on a scale from 1 to 10 on the service and products purchased
Column2 (Amount Spent)	Amount (in dollars) spent by customers in the department store

NOTE: The data was randomly created for the sole purpose of this assignment.

Question 1 (6 marks) **Marks**

The data can be found in the **worksheet named Question 1** in the data file **ProjectData& 2020.3** .

The product manager of a certain company is interested in determining the typical age of customers who ordered their product advertised on television during the previous month. A sample of 60 customers was randomly selected and their age together with their gender were recorded.

The data can be found in the **worksheet named Question1** in the data file **ProjectData&**.

a) Using RStudio, obtain a Descriptive Statistics output of the age in years from a random sample of 60 customers who ordered a product advertised on television during the previous month. (Mean, Median, Range and Standard Deviation). * [2]

b) Using RStudio, construct 2 histograms (with 5 classes) showing the distribution of the customers' ages, one for male customers and one for female customers.

[2]

c) Using the information from the descriptive statistics output and the graphical display, write one paragraph describing the data set. You must include information about the mean, mode, range, standard deviation and the shape of the distribution, of the data in your answer. [1]

d) Using RStudio, construct a 90% confidence interval for the population mean age of the customers who ordered a product advertised on television during the previous month and interpret the result. [1]

<u>* Evidence of work in RStudio is required</u>

Question 2 (7 marks)

The data can be found in the **worksheet named Question2** in the data file **ProjectData& 2020.3**.

You are about to test the following hypothesis:

According to a Money Magazine, Budget Direct, the average annual total transport costs per household in Sydney is $23.15 (in $'000s). A private researcher surveyed a sample of 80 Sydney households in an attempt to determine whether the population mean annual total transport costs differed from the claimed national mean.

a) Using RStudio, obtain an output for a hypothesis test at 1% level of significance. Assume that the mean annual total transport costs per household are normally distributed* [2]

b) Using the information in the output, conduct a hypothesis test in the space below. Make sure that you include the null and the alternative hypothesis, the decision rule and the value of the test statistics. Explain your decision and write a conclusion. [5]

<u>* Evidence of work in RStudio is required</u>

Question 3 (7 marks) **Marks**

The data can be found in the **worksheet named Question 3** in the data file **ProjectData& 2020.3**.

You are about to test the following hypothesis:

During the previous term in a certain university, the final marks received by students in Statistics were collected and analysed to find out which between Fully Online and Blended Learning method is more effective for students to learn Statistics. Is there a significant difference in the average final marks received between the two groups of students?

a) Using RStudio, obtain an output for a hypothesis test at 5% level of significance. Assume that the paired differences are normally distributed. * [2]

b) Using the information in the output, conduct a hypothesis test in the space below. Make sure that you include the null and the alternative hypothesis, the decision rule and the value of the test statistics. Explain your decision and write a conclusion. [5]

<u>* Evidence of work in RStudio is required</u>

Question 4 (5 marks) **Marks**

The data can be found in the **worksheet named Question4** in the data file **ProjectData& 2020.3**.

Customers at a large department store were asked to rate the service and products purchased on a scale from 1 to 10, with 10 being the highest. The amount that they spent was also recorded. Is there an evidence of linear relationship between the customer rating and the amount that they spent?

(a) Using Excel, obtain a linear regression output.* [2]
(Include graphical display and statistical analysis summary)

(b) Determine the linear regression equation that may be used to predict the amount spent by the customers at department store based on their ratings on service and products purchased. [1]

.

(c) Interpret the slope of the regression line. [1]

(d) State and interpret the coefficient of determination. [1]

<u>* Evidence of work in RStudio is required</u>

END OF PROJECT