**Introduction :**

R has plenty of built-in data sets. The assessment is designed with data() in R. The coursework requires you to document your analysis and critically discuss your analysis approaches and your results in a business report suitable for a non-technical audience. You should not include code within the report.

You should not include code within the report , You may include any technical details that support your arguments in a set of appendices, which will not count against the word limit. You must ensure the main text is readable and that your argument is coherent without needing to consult the appendices. All parts of the text supported by an appendix must cross-reference directly to the relevant part.

**Important:** Demonstrates a deep understanding of the subject matter that indicates you already thinking at a high level.

Employs sophisticated reasoning and shows evidence of understanding complex concepts.

Excellent presentation of materials organisation, analysis procedures and interpretation of results.

Outstanding command of critical analysis and judgement.

**Assignment:**

1. Using the data set cars, which refers to pairs of the speed and stopping distances of cars, the data are arranged in a data frame. You are asked to do the following:
   1. Draw a scatterplot between the two variables that are contained in the data frame cars. Discuss the relationship between the two variables. (2 marks)
   2. Calculate the correlation coefficient and come up with the conclusion that if any relationship exists. (2 marks)
   3. Draw a histogram of each of the two variables. Comment on the shape of the two distributions. (4 marks)
   4. Draw a box-plot for each of the two variables and explain the plot. (4 marks)
2. Consider a population for which the height follows a normal distribution with mean 170 cm and standard deviation of 9 cm.  
   1. Simulate the height of 1000 people and save in a vector called heights. (2 marks)
   2. Plot the simulated heights against the theoretical density function for this distribution, and comment on the distribution. (2 marks)
   3. What is the minimum height required so that one belongs to the tallest 10% of this population? (2 marks)
   4. What percentage of the population is shorter than 164 cm? (2 marks)
3. Consider the built-in data set cars which consists of two variables, speed and breaking distance.
   1. Fit a simple linear regression model of distance (dependent variable) on speed (independent variable) and check its summary. (4 marks)
   2. Graphically explore the four assumptions of the linear regression model, perform relevant tests, and discuss the conclusions.  (8 marks)
4. Install the datarium package, which contains the data set marketing. This data set consist of four variables and 200 records, advertising expenditure in three media (YouTube, Facebook, and Newspaper) and the sales (which is the dependent variable). With the marketing data:
   1. Discuss the correlation relationship among variables and explore the top ten rows of the data set. (2 marks)
   2. Split the data into an in-sample (that contains the first 150 records of the data) and an out-of-sample (that contains the last 50 records of the data). (2 marks)
   3. Using the in-sample data, build the simple linear regression models with one independent variable "youtube". Report and discuss the summary of the model. (6 marks)
   4. Based on the results above, build the multiple linear regression model with more than one independent variable. Compare and discuss the model results and choose the best model with justifications. (8 marks)
   5. Use the chosen model to predict the sales for the expenditures of the out-of-sample data. (2 marks)
5. In this task, you will be working with the “Default of Credit Card Clients” data set, for which you can find full details regarding the variables [here](https://code.datasciencedojo.com/datasciencedojo/datasets/tree/master/Default%20of%20Credit%20Card%20Clients). For convenience, the data is also attached below, named "default of credit card clients.xls". Once you load the data:
   1. Randomly split the data into two sets: in-sample (75% of the data) and out-of-sample (the remaining 25%). (2 marks)
   2. Using the in-sample data, fit a suitable logistic regression model (note that the solution might not be unique). (8 marks)
   3. Produce the classification table (confusion matrix). (4 marks)
   4. Use the fitted model and the out-of-sample data to make predictions. Produce the classification table (confusion matrix) for these predictions. (4 marks)
6. Regression does not provide proof of a causal link between the dependent variable and an independent variable. Explain why this is the case.  (10 marks)
7. Write a short essay (500 words), that discusses your understanding of how business statistics could be beneficial in economic recovery post COVID, with statements and justifications. (20 marks)