ISCG8026 Introduction to Data Science

Semester 1 – 2020

Lab Exercises - Part B

Due Date: 30 May 2020, 13:00

Weight: 20%

School of Computing & Information Technology

For this set of lab exercises, you will be working individually. It is worth **20%** of your final grade and it must be submitted on Moodle by the due time mentioned above.

Late Submission of Assignments

Assignments submitted after the due date and time without having received an extension through Special Assessment Circumstances (SAC) will be penalised according to the following:

- 10% of marks deducted if submitted within 24hrs of the deadline
- 20% of marks deducted if submitted after 24hrs and up to 48hrs of the deadline
- No grade will be awarded for an assignment that is submitted later than 48hrs after the deadline

Assignments submitted in more than 48 hours late will not be marked unless Special Assessment Circumstances apply. So, it is better to submit an incomplete assignment on time.

Special Assessment Circumstances

A student, who due to circumstances beyond his or her control, misses a test, final exam or an assignment deadline or considers his or her performance in a test, final exam or an assignment to have been adversely affected, should complete the Special Assessment Circumstances (SAC) form available from Student Central. Within any semester, a student may have only one SAC per course. When requesting an SAC for an assignment, the SAC application form must be submitted (along with the work completed to date) within the time frame of the extension requested; i.e. if the Doctor's certificate is for one (1) day, then the SAC application form and work completed must be submitted within one (1) day.

Assistance to other Students

Students themselves can be an excellent resource to assist the learning of fellow students, but there are issues that arise in assessments that relate to the type and amount of assistance given by students to other students. It is important to recognise what types of assistance are beneficial to another's learning and also what types of assistance are unacceptable in an assessment.

Beneficial Assistance

- Study Groups.
- Discussion.
- Sharing reading material.
- Testing another student's programming work using the executable code and giving them the results of that testing.

Unacceptable Assistance

- Working together on one copy of the assessment and submitting it as own work.
- Giving another student your work.
- Copying someone else's work. This includes work done by someone not on the course.
- Changing or correcting another student's work.
- Copying from books, Internet etc. and submitting it as own work. Anything taken directly from another source must be acknowledged correctly: show the source alongside the quotation.

Do you want to do the best that you can do on this assignment and improve your grades? You could:

• Talk it over with your lecturer
• Visit Student Success and Achievement for learning advice and support (in Te Puna)
• Visit the Centre for Pacific Development and Support

Instructions

For this set of lab exercises, you will be working individually. It is worth **20%** of your final grade and it must be submitted on Moodle by the due time mentioned above.

Question 1 [3.8 Marks]

Develop a R Script file that can perform the following tasks on the data set given in "Data Set 1a.csv":

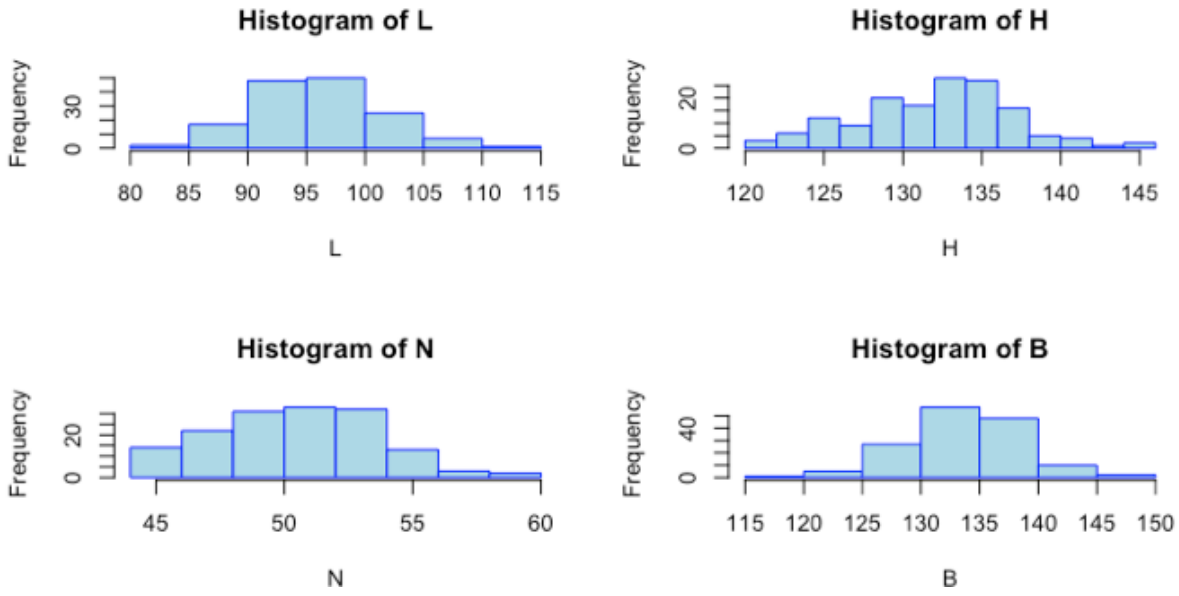
1. Omitting all the transactions made before 1 Jan 2010 **[1.4 Marks]**
2. Calculating monthly total sale **[2 Marks]**
3. Visualizing the results (monthly total sale) by using appropriate charts/plots. **[0.4 Mark]**

Pleas create a R Script file named B1.R, and insert the code and result in your report.

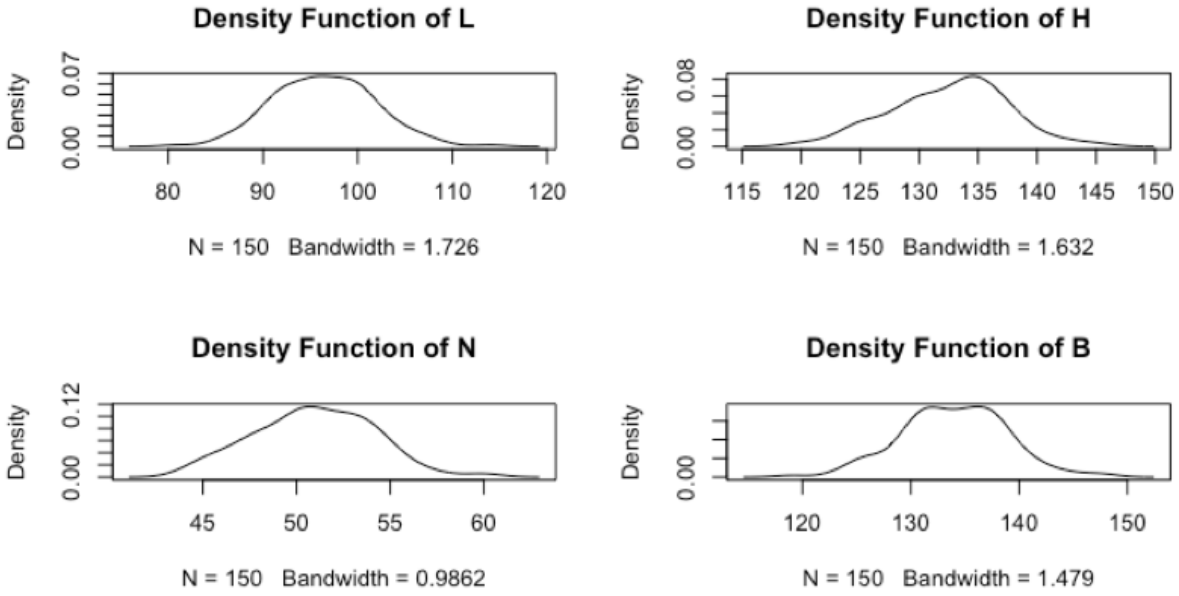
Question 2 [4 Marks]

Develop a R Script file that can perform the following tasks on the Skulls data set given in "Data Set 3.csv".

1. Plot the histograms of L, H, N, B in one page. The output should be similar to the following figure. **[1 Mark]**



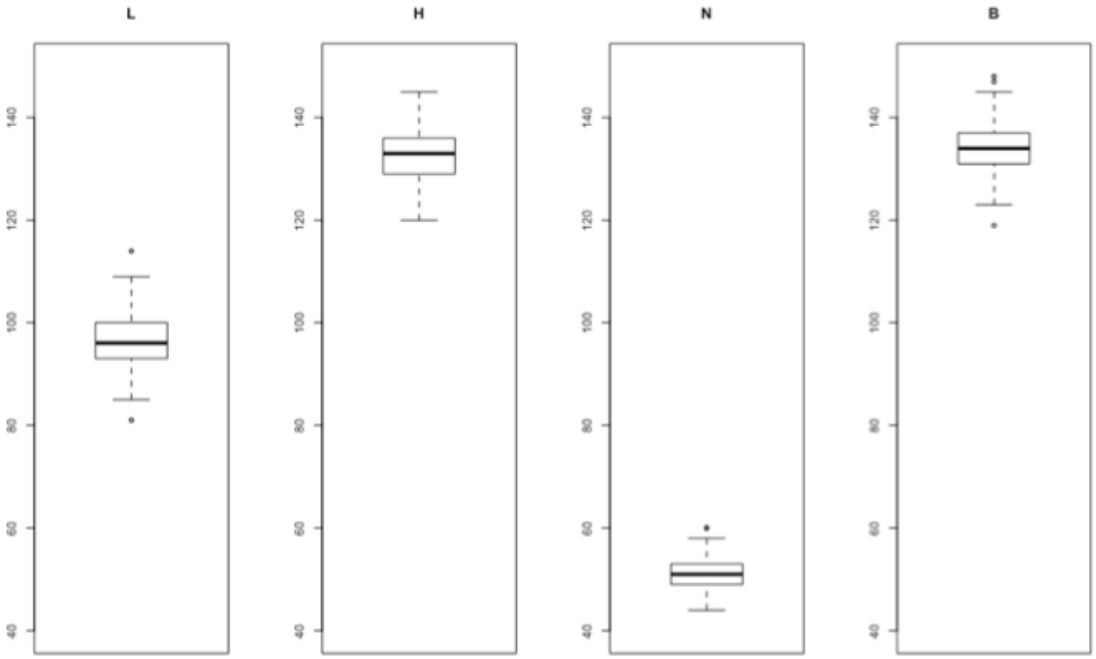
2. Plot the density functions of L, H, N, B in one page. The output should be similar to the following figure. **[1 Mark]**



3. Compare the density functions against a normal density function. Comments on the symmetry and sharpness of the density functions. **[0.4 Marks]**
   · *Hint: You may need to calculate shape measures and complete the following table.*

	L	H	N	B
Skewness				
Kurtosis				

4. Create the boxplots of L, H, N and B using a similar scale. The output should be similar to the following figure. **[1 Mark]**



5. Calculate the mean, variance and standard deviation of L, H, N and B and complete the following table. **[0.6 Marks]**

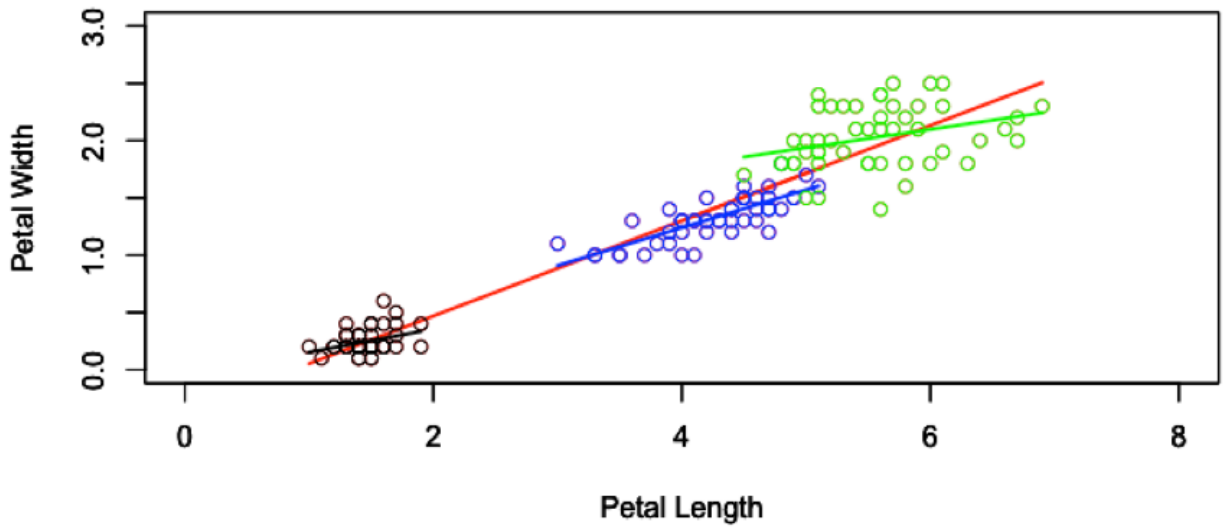
	L	H	N	B
Mean				
Variance				
Standard deviation				

Please create a R Script file named B2.R, and insert the code and result in your report.

Question 3 [4 Marks]

Develop a R Script file that can perform the following tasks on the Iris data set given in "Data Set 4.csv".

1. Create a scatter plot for petal length and width variables. **[0.1 Marks]**
2. Calculate a liner model between petal length and width and show it in the scatter plot. **[0.2 Marks]**
3. Based on the Species data, subdivide the iris dataset into three separate subsets (for each species).
4. Repeat steps 1 and 2 for each subsets. **[1.5 Marks]**
5. Plot all the results including scatter plots and linear models in one plot. The outcome should be similar to the following plot.



6. Show that the linear models obtained for the three subsets can describe the relationship between petal length and width more accurately. **[0.2 Marks]**
   - Hint: the accuracy of each model can be measured by using Mean Square Error, which can be calculated by the following R scripts:

```
Model <- lm(y~x)
error = Model$fitted.values - y
MSE = mean(error^2)
```

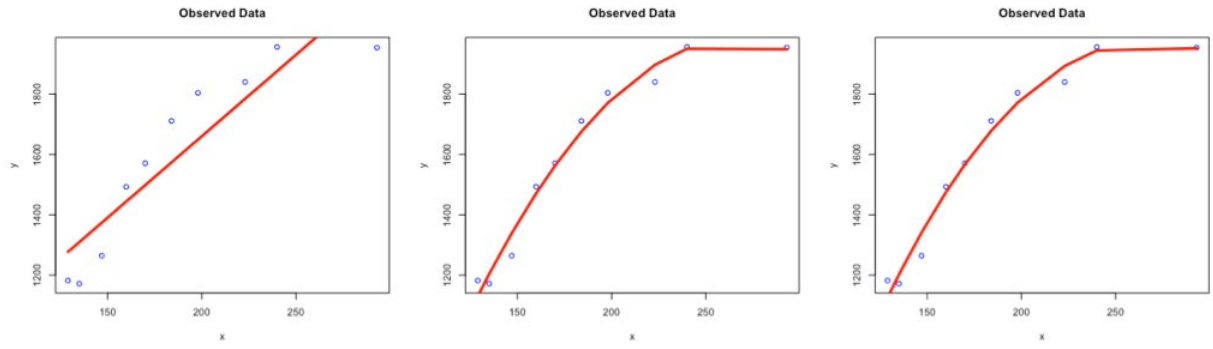
7. Repeat the above steps for sepal length and sepal width data. **[2 Marks]**

Please create a R Script file named B3.R, and Insert the code and result in your report.

Question 4 [3.2 Marks]

Develop a R Script file that can perform the following tasks on the Power Consumption data set given in "Data Set 5.csv". (The data are the electricity consumptions in kilowatt-hours per month from different houses and the areas in square meter of those houses)

1. Create a scatter plot for the two variables. **[0.4 Marks]**
2. Calculate a linear regression model. **[0.2 Marks]**
3. Calculate polynomial regression models of order 2 and 3. **[0.4 Marks]**
4. Plot the regression models (the result should be similar to the following figure) **[1 Mark]**



5. Compare the accuracy of the three models **[1.2 Marks]**

Please create a R Script file named B4.R, and Insert the code and result in your report.

Question 5 [2 Marks]

Two machines are used for filling plastic bottles with a net volume of 16.0 ounces. The filling process can be assumed to be normal. The quality engineering department suspects that both machines fill to the same net volume. An experiment is performed by taking a random sample from the output of each machine. Would you reject or accept the quality engineering department hypothesis? Perform your calculation by using R. (alpha = 0.05)

Machine 1		Machine 2	
16.03	16.01	16.02	16.03
16.04	15.96	15.97	16.04
16.05	15.98	15.96	16.02
16.05	16.02	16.01	16.01
16.02	15.99	15.99	16.00

Pleas create a R Script file named B5.R, and insert the code and result in your report.

Question 6 [3 Marks]

In the population, the average IQ is 100 with a standard deviation of 10. A team of scientists wants to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 110. Did the medication affect intelligence, using alpha = 0.01? Find the confidence interval?

Please create a R Script file named B6.R, and insert the code and result in your report.