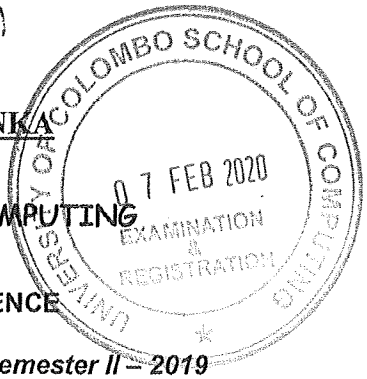
161



UNIVERSITY OF COLOMBO, SRI LANKA

UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

Academic Year 2016/2017 – Second Year Examination – Semester II – 2019

SCS2211– Laboratory II – Part A

TWO (2) HOURS (for A+B)

To be completed by the candidate

Examination Index No: .....

Important Instructions to candidates:

1. The medium of instruction and question is **English**.
2. **Write your answers** in **English.**
3. If a page or a part of this question paper is not printed, please inform the supervisor immediately.
4. Note that questions appear on both sides of the paper. If a page is not printed, please inform the supervisor immediately.
5. Write your index number on each and every page of the Question paper.
6. This paper has **04** questions and **18** pages.
7. Answer **ALL** questions. All questions carry equal marks (25 marks).
8. **This paper consists of two parts, Part A (Question No 1 and Question No 2) and Part B (Question No 3 and Question No 4) and submit separately.**
9. Any electronic device capable of storing and retrieving text including electronic dictionaries and mobile phones are **not allowed**.
10. **Non-Programmable** calculators are **allowed**.

For Examiner's use only	
Question No	**Marks**
1	
2	
3	
4	
Total	

Index No: .....

Part A

Question 1

(a) A student has recorded the number of runs scored by his favourite cricketer in each inning at one-day international cricket matches which were held last year.

(i) Identify the scale of measurement of the variable "the number of runs scored". [1 mark]

(ii) Specify whether the variable "the number of runs scored" is discrete or continuous. [1 mark]

The data recorded during 13 innings of the first 6 months of the year are as follows:

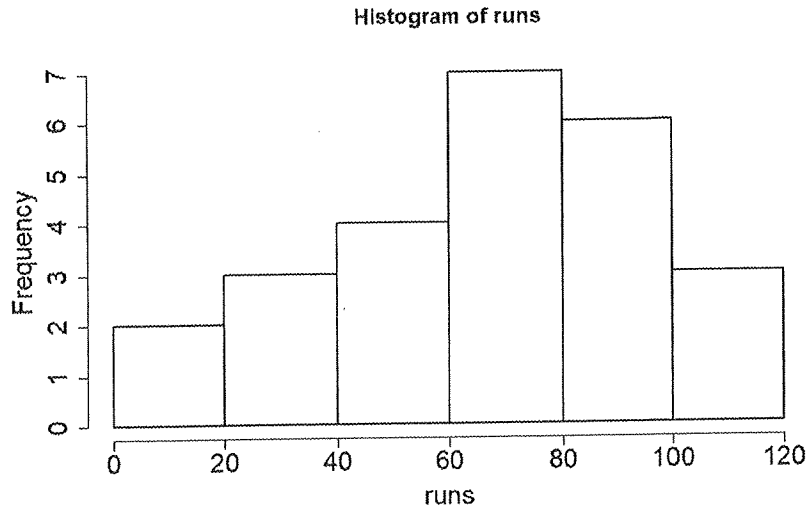
50, 40, 55, 60, 35, 65, 78, 45, 38, 64, 73, 66, 78

(iii) Arrange the above data in a stem and leaf plot. [3 marks]

Stem	Leaf

(iv) Find the median value of the above data. [2 marks]

The student's favourite cricketer has played for 25 international cricket matches during the entire year. The histogram drawn for the number of runs recorded for the entire year is given below.



(v) Briefly comment on the shape of the distribution of the number of runs. [2 marks]

.....

.....

.....

(vi) Which measure of central tendency is most suitable for this data? [1 mark]

.....

(vii) What is the class interval which will contain the measure of central tendency you suggested in part (vi)? [1 mark]

.....

(viii) Do you think the player has performed better in the latter 6 months of the year? Justify your answer using the previous calculations/findings. [2 marks]

(b) A group of students from a private medical institution in USA are interested in knowing whether the infant birth weight is different in white mothers compared to black mothers. They selected 2 random samples of 25 white mothers and 25 black mothers who got admitted to the central hospital for delivery and measured the birth weight (in grams) of the new born infants.

(i) State the null and alternative hypotheses to test the students' interest. (Define any notations used) [4 marks]

The following R-code and outputs were generated for the analysis. The vectors "white" and "black" contains the measured weights of the infants belonging to white and black mothers respectively.

```
> t.test(white,black,var.equal = T)

    Two Sample t-test

data:  white and black
t = 2.5119, df = 48, p-value = 0.01543

95 percent confidence interval:
  86.23408 778.08592
sample estimates:
mean of x mean of y
  3124.56   2692.40
```

(ii) State two important assumptions which should be satisfied in order to carry out the above analysis. [2 marks]

(iii) What is the value of the test statistic? [1 mark]

(iv) State the statistical conclusion of the test at 5% level of significance. [3 marks]

(v) State the general conclusion arising from the above test. [2 mark]

Question 2

(a) The dataset "all.mammals.milk" contains a list of animals and the constituents of their milk. It contains two variables; percentage of water *(water)* and percentage of protein *(protein)* and a researcher is interested in knowing whether there is any linear relationship between these two variables.

(i) Suggest a suitable graph which could reveal the above relationship. [2 marks]

(ii) Suppose the graph drawn indicated that there is a negative linear relationship between the two variables. Suggest a measure which can quantify the strength of this relationship. [2 marks]

(iii) The researcher next aims at predicting the percentage of protein in milk, when the percentage of water is given. What statistical technique should be used to obtain such predictions? [2 marks]

The following R code and outputs were obtained for the above-mentioned analysis.

```
> fit<-lm(protein~water)
> summary(fit)

Call:
lm(formula = protein ~ water)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2237 -1.8119  0.2098  1.7639  4.6237

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.04957    3.08110   7.481 1.33e-07 ***
water       -0.21536    0.03891  -5.535 1.25e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.443 on 23 degrees of freedom
Multiple R-squared:  0.5712,       Adjusted R-squared:  0.5525
F-statistic: 30.63 on 1 and 23 DF,  p-value: 1.25e-05
```

(iv) Write down the estimated equation which can be used to predict the percentage of protein. [3 marks]

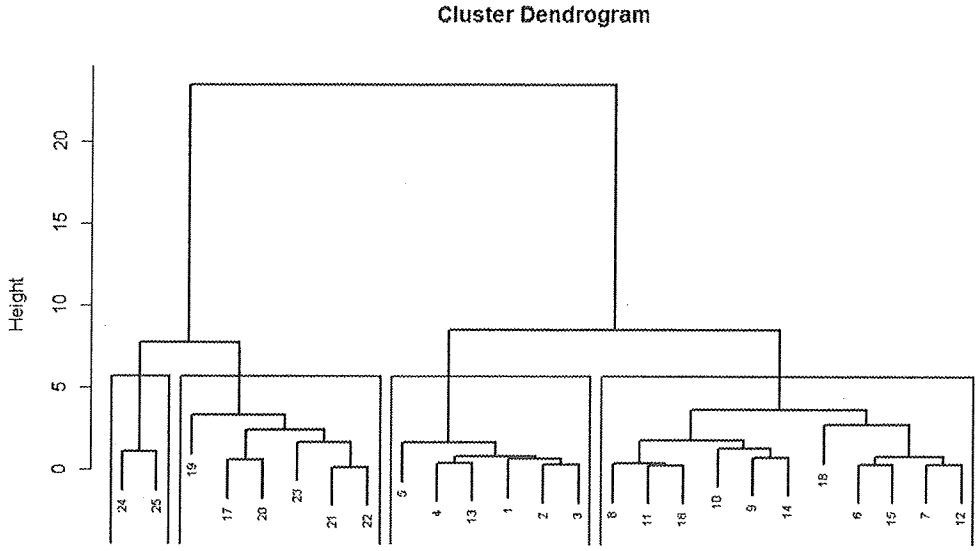
(v) Using an appropriate measure, comment on the goodness of fit of the model. [3 marks]

(vi) The researcher wishes to check whether the relationship between the two variables is significant using an F-test. Write down the null and alternative hypotheses for this test in standard notation.
[2 marks]

(vii) Using the correct p-value, carry out a test to determine whether the relationship between the two variables is significant at 5% level. [4 marks]

(b) The dataset also has the percentages of fat, lactose and ash in milk. The researcher wishes to classify the animals into clusters considering the percentages of each content. However, he has no prior information about how many clusters to construct.

(i) What is the type of clustering method that you can suggest for the researcher? [2 marks]



(ii) The above diagram illustrates the output from the clustering method he followed. How many clusters has he identified among the animals? [1 mark]

(iii) During the process he followed, he reached the above number of clusters in the second step of clustering. State whether he has used agglomerative method or divisive method. [2 marks]

(iv) If he already knew the number of clusters among the animals, what is the clustering method that he could have used to classify the animals? [2 marks]
