UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING

SCS2211 - Laboratory II 2021

Group Assignment 1

1. A group should consist of a maximum of 5 members.
2. You should select a dataset from https://www.kaggle.com/datasets. There are 71000 + datasets at the moment. One dataset can be used only by one team. There will be a **40% penalty** if you use the same dataset.
3. You should analyse the dataset using R and produce a report. Use **IEEE referencing style** if required.
4. Your report should contain
    a. Observations about the Data set
    b. Appropriate Plots/Charts to describe data
    c. Minimum of one hypothesis
    d. Justification of hypothesis
    e. Plot your multivariate data
    f. Select two variables which depicts strongest relationship in your multivariate data plots and;
        i. Describe it
        ii. Find the correlation
        iii. Draw the least squares regression line.
        iv. Draw the residual plot
    g. Apply clustering techniques to your data to identify whether there is any natural grouping among data.
5. Deadline for submission is **27th March 2021 , 11.59 PM**.
6. Google sheet for Group Allocations is given below.

https://docs.google.com/spreadsheets/d/1u80U1rZqxYZKacoytMxs5bOYc3y6I4CjkYD6I-pJnt4/edit?usp=sharing

[Continued in page 2]

7. Google form for selection of data set is given below.

https://forms.gle/VfRjZD77ioWK1HWKA

8. Corresponding Google sheet for the above form is given below. Datasets are allocated on a **first come first served** basis. **Make sure you haven't selected a dataset, selected by another group.**

https://docs.google.com/spreadsheets/d/1VtWM9_4CPA1w3DsKe4Fyb63Jrj8g6f6sB9oBVTdKhXg/edit?usp=sharing

9. Include a page with individual contributions in the report.
10. Report name should be your group number **eg- Group_01.pdf**

Note : This assignment is conducted according to the request received from the students via student union. (Email dated 18.03.2021)