# Word count - 4000

# Estate Workers Productivity Data Challenge-2022

# Overview

Estate workers who are working in tea, rubber estates are contributing the Sri Lankan economy in many ways. This dataset contains 2015 estate workers productivity contribution during the period from 2012 to 2015, in four CSV files, extracted from the payroll system.

The **employee.csv** file contains 2,015 observations in 6 variables of worker related data. The **fields.csv** file contains 63 observations in 7 variables of estate, its division & crop cultivated related data. The **workdetails.csv** file contains 1,110,902 observations in 10 variables of estate worker’s productivity related data during the period from 2012 to 2015. The **workcode.csv** file contains 483 observations in 3 variables of work code information.

As a data scientist, your task will be to clean, normalise and transform these data into R compatible formats and undertake an extensive data mining using Machine Learning. The main objective of this data challenge is to develop Machine Learning model to identify various data patterns and forecast the production. Report on any interesting patterns, (for example, crop production/yield patterns), that you may reveal from the data analysis and possible visualizations.

In your discussion you will provide a critical synopsis of the challenges of data analysis, integration and visualisation you faced during this exercise. You will provide relevant assumptions you made with valid justifications during this exercise.

**Datasets**

4 csv files have been provided as explained above.

**Please note, NO any other information provided, on the data definitions or meaning of the fields. You may have to explore and identify the meaning and relationships with other datasets.**

**Assignment Tasks**

1. Provide detailed description of each datasets, their properties and relationships [10 marks]
2. Read data from csv files to R environment for processing [3 marks]
3. Clean any outliers, exceptional values from the datasets [4 marks]
4. Normalizations, Scaling [10 marks]
5. Merge the datasets [5 marks]
6. Create training and test datasets, if required [10 marks]
7. Training a model on the data [5 marks]
8. Apply different Machine Learning approaches and discuss [10 marks]
9. Accuracy of each different models [10 marks]
10. Alternative ways of normalizations, model building, and their performances [10 marks]
11. Patterns identified and their visualizations [6 marks]
12. Describe a detailed comparative analysis between the scaling, Machine Learning approaches – strengths, limitations, uniqueness [6 marks]
13. Comparative analysis should be in relation to integration, transformation, visualization and data mining [6 marks]
14. Provide a brief discussion about the knowledge gained [5 marks]

**Referencing Requirements**

All referencing should utilize the Harvard Style.

**REPORT STRUCTURE**

* Paper Size : A4
* Word Count : 4000 words
* Printing Margins : LHS; RHS: 1 Inch
* Binding Margin : ½ Inch
* Header and Footer : 1 Inch
* Printing : Single Sided
* Basic Font Size : 12
* Font Style : Arial/Times New Roman
* Presentation : Bound Document