



**Sri Lanka Institute of Information Technology**

**B. Sc (Hons) in Information Technology**

**Year 02 – Semester II – 2022**

**Probability and Statistics – IT2110**

**Lab Sheet 06**

The data set is about rainforest data. This is a data mining approach to predict forest fires using meteorological data. The variable description is given below.

- X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
- Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
- Month - month of the year: "jan" to "dec"
- Day - day of the week: "mon" to "sun"
- FPMC - FPMC index from the FWI system: 18.7 to 96.20
- DMC - DMC index from the FWI system: 1.1 to 291.3
- DC - DC index from the FWI system: 7.9 to 860.6
- ISI - ISI index from the FWI system: 0.0 to 56.10
- Temp - temperature in Celsius degrees: 2.2 to 33.30
- RH - relative humidity in %: 15.0 to 100
- Wind - wind speed in km/h: 0.40 to 9.40
- Rain - outside rain in mm/m<sup>2</sup> : 0.0 to 6.4
- Area - the burned area of the forest (in ha): 0.00 to 1090.84 (this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform)

- 1) Identify the variables and import the given data set into R.
- 2) Get the summary of the data set
- 3) How many observations are there?
- 4) What is the maximum and minimum wind speed of this data set?
- 5) Get five number summary of temperature
- 6) How many outliers are there in the wind variable?
- 7) According to the boxplot of wind what kind of a distribution it has?
- 8) What is the median of temperature?
- 9) What is the mean and standard variation of wind variable?
- 10) What is the interquartile range of wind variable?
- 11) How many observations have measured during Friday in August?
- 12) What is the average temperature, during September?
- 13) On which day have they measured most observations during month of July?