CS3002 – Artificial Intelligence: Introduction to R and RStudio

R is an open source suite of software facilities for data manipulation, calculation and graphical display. Among other things it has

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display
- a simple and effective programming language (called 'S') which includes conditionals, loops, user defined recursive functions and input and output facilities.

R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly and has been extended by a large collection of *packages*.

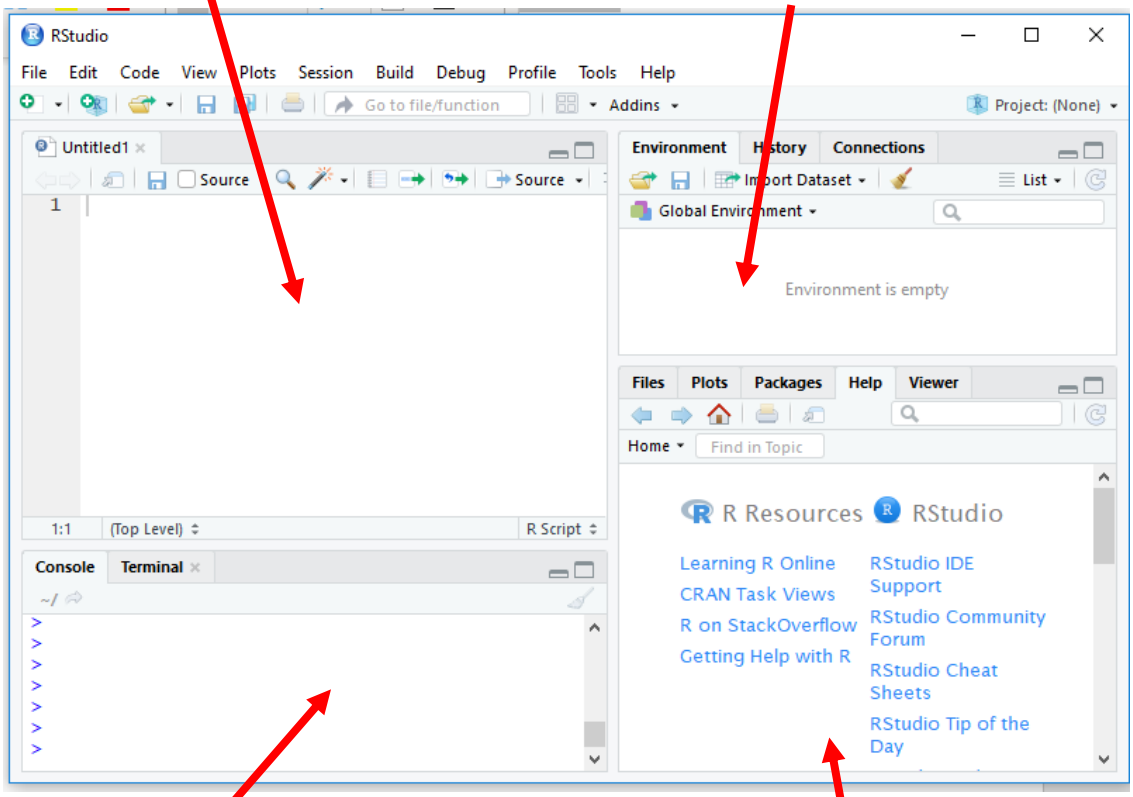
It can be installed from here: https://cran.r-project.org/

RStudio is an interface for developing your own R functions and scripts and can be installed from here: https://www.rstudio.com/products/rstudio/

Start up RStudio from the windows menu (search for “RStudio”) and begin to explore the GUI environment:

Where you load / save / write your scripts / functions

Where you can see all variables / data structures



Where you can run commands and observe outputs

Where you can view help files and plots

Read through the following Tutorials that are available on Blackboard:

“A (very) short introduction to R” by Torfs and Brauer

“A Quick Introduction to R and RStudio” by Vlad Krotof

Also refer to the “Cheat Sheets” and the “Introduction to R” reference manual if necessary

Create a new script and save it as “123456.R” where you use your student ID as the name.

The script should read in the following data file into a data frame: “forestfires.csv”:

```
mydata = read.csv('J:\\Data\\Teaching\\2019-2020\\CS3002\\NEWLAB_SCRIPT\\Data\\forestfires.csv', sep=",")
```

This file contains data on forest fires in Portugal (from the UCI repository for Machine Learning: https://archive.ics.uci.edu/ml/datasets/Forest+Fires).

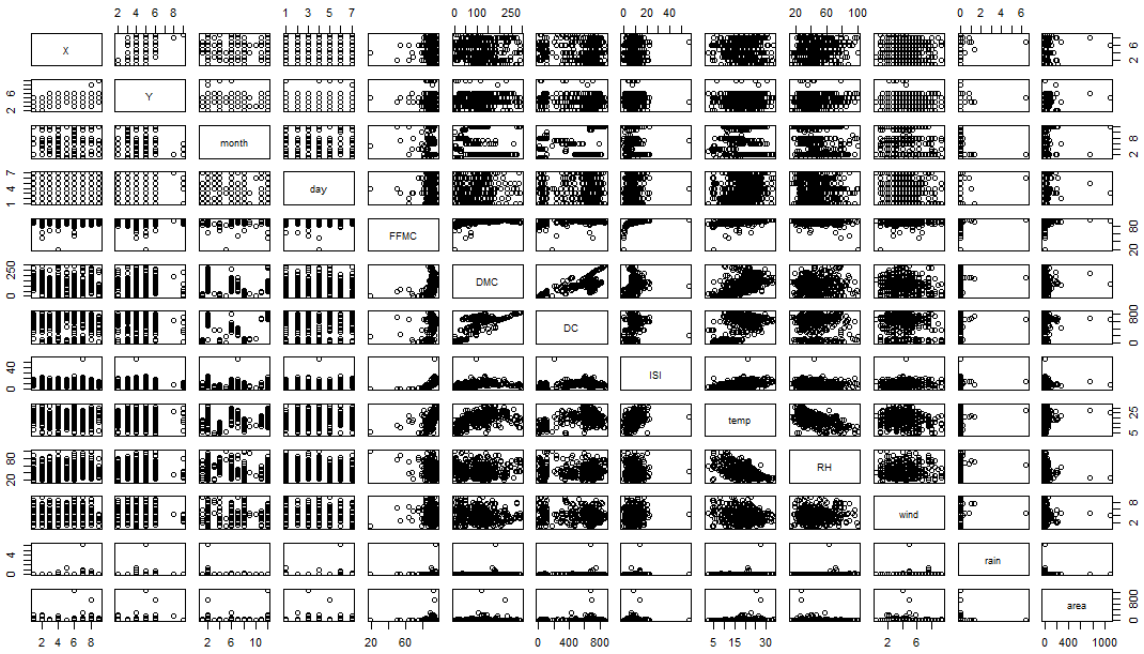
The variables are described below:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84

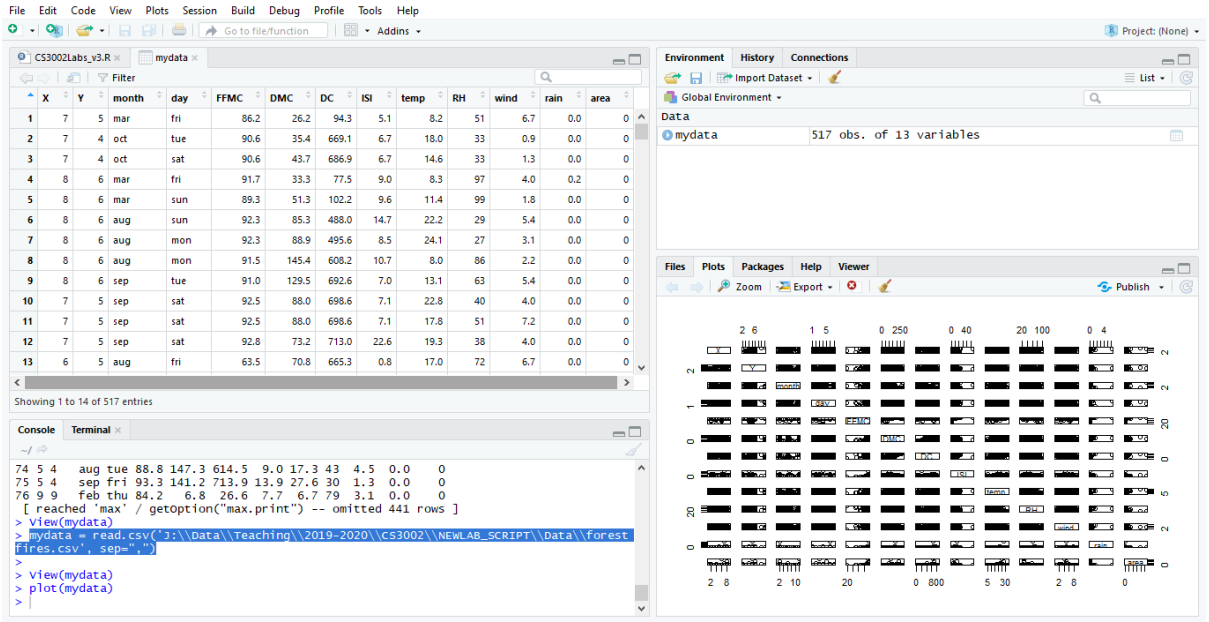
(this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

You can plot all of the variables in one large summary (the plot appears in the bottom right panel but you can expand it by clicking on *zoom*):

```
plot(mydata)
```



Or click on the data frame in the top right panel which will produce the table of data in the top left panel (typing `view(mydata)` also does this):



You can also create scatterplots of individual columns using the column name with a $ sign:

```
plot(mydata$temp, mydata$wind)
```

or by referring to the column:

```
plot(mydata[,9], mydata[,11])
```

Other plots include histograms:

```
hist(mydata$temp)
```

and line plots:

```
plot(mydata$temp, type="l")
```

You can also plot colours. For example, here are the X and Y coordinates plotted on a scatter colour codes according to the temperature:

```
plot(mydata$X, mydata$Y, col=mydata$temp)
```

Now calculate some statistics the mean, median, maximum and minimum values for each column

```
meantemp = mean(mydata$temp)
```

and then write them to a new csv file called “Output.csv”

```
write.csv(meantemp, file = "Output.csv")
```

or build a linear model using regression:

```
plot(mydata$temp,mydata$ISI)
lmfire=line(mydata$ISI~mydata$temp)
abline(coef(lmfire))
```

Now explore other examples from the two tutorials. All assessed labs will be using R and you will be expected to submit R scripts.