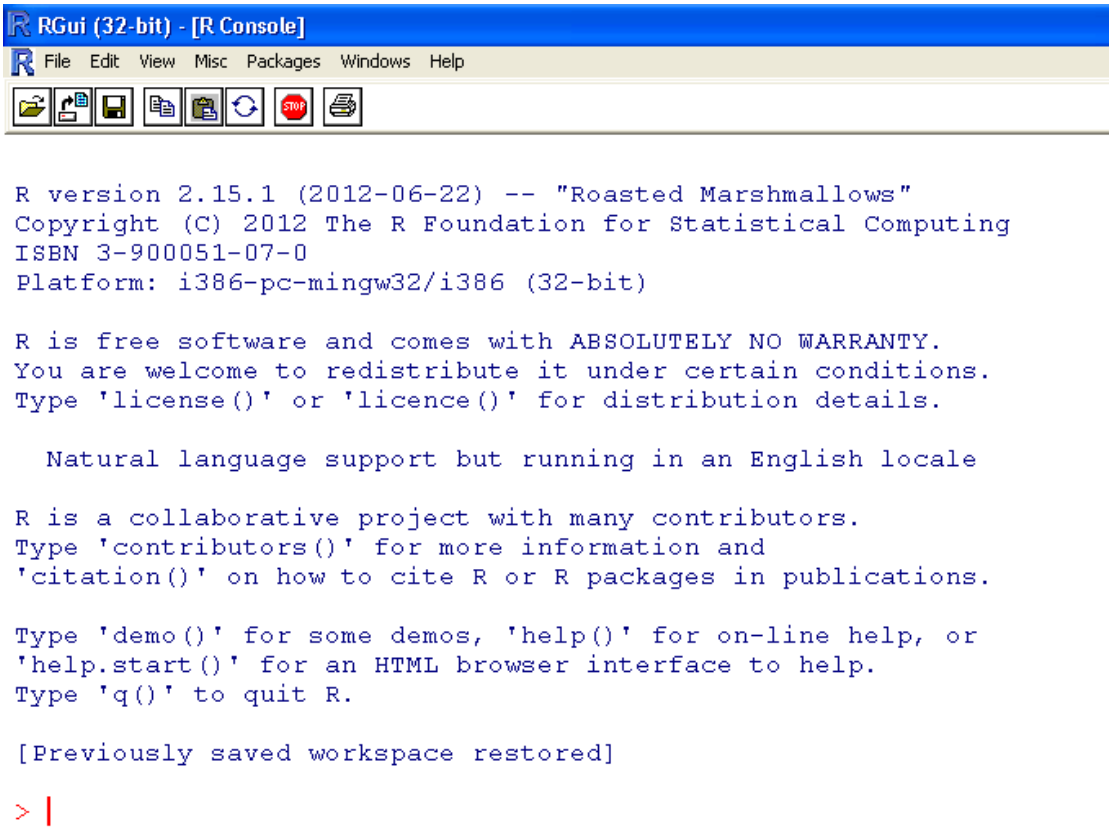
MA4605 Chemometrics 2012 - Laboratory A

Outcomes for Week 1 laboratories:

- Familiarisation with the ***R*** computing environment.
- Perform basic calculations using the command line interface.
- Using basic ***R*** command to determine the characteristics of a data set.
- Using basic ***R*** command to determine basic descriptive statistics of a data set.

Part 1: Familiarisation with the R Environment

When you open ***R*** (by clicking on the icon on the desktop), you should see something like this:



This text discusses many important matters relating to ***R***. For us, the key pieces of information relate to using help facilties, i.e. `help()` and `help.start()`.

By typing the following text as demonstrated below, run the command `help.start()` and `demo()`. The HTML help utility is a very useful resource for all ***R*** users. Also you can enter the command `data()` to find out the names of inbuilt data sets .

```
> demo()
> help.start()
> data()
```

We are going to use one of these inbuilt data sets “Iris” for this class. To access more information about Iris, we can use the help function as follows.

```
> help(iris)
```

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

Source

> *Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems.* Annals of Eugenics*, 7, Part II, 179–188.*
>
> *The data were collected by Anderson, Edgar (1935). The irises of the Gaspe Peninsula,* Bulletin of the American Iris Society*, 59, 2–5.*

Part 2: Performing Basic Calculations.

Before we continue, let us first see how ***R*** can be used to perform simple calculations. In your submission sheet, write out the result of each of the following commands in your submission sheets.

```
> 256/146
>
> pi * 4
>
> 3.14^2
>
> log(4.11)
>
> log(4.11,2)
>
> factorial(6)
```

Part 3: The Iris Data Set

The command **`head()`** is used to display the column names and the first six records of a dataset, thus allowing us to get a sense of the information contained in that data set.

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

The **`summary()`** command is a very versatile command that displays a statistical summary of data.

Writing the answers in your submission sheet, use the **`summary()`** command to get determine the following for each of the first four columns of the Iris data set:

- The mean
- The median
- The inter-quartile range. (you can leave it in the form Q_3-Q_1)

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Following the example of the **`summary()`** command, try out the following commands

- **`dim()`**
- **`names()`**
- **`nrow()`**
- **`ncol()`**

What do you think these commands are used for? (Hint: to find out what **`dim()`** does – type in the command **`help(dim)`** on the command line).

Write down the description and outputs on your submission sheet.

We are going create a new data set, using only the first four columns of the data set. We shall discuss the method for producing subsets of datasets in more detail at a later stage.

To subset the Iris data set, such that we are retain only numeric variables, we use the following command.

```
> iris.2 = iris[,1:4]
```

We know the mean and median for each of the four variables. Let us try to find out the variance and covariance of the variables.

```
> var(iris.2)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length      0.68569    -0.04243       1.2743      0.5163
Sepal.Width      -0.04243     0.18998      -0.3297     -0.1216
Petal.Length      1.27432    -0.32966       3.1163      1.2956
Petal.Width       0.51627    -0.12164       1.2956      0.5810
```

This command yields the "Variance Covariance" matrix of the data.

The diagonal elements of this matrix are the variances of the respective variables. The off-diagonal elements are the co-variances of the respective variables.

- The variance for `Sepal.Width` is 0.18998.
- The covariance for `Sepal.Width` and `Sepal.Length` is 0.04243

Remark: notice how the matrix is symmetric.

In the same manner as the `var()` function, compute the correlation matrix using the `cor()` function.

- Write down the matrix in your submission sheet.
- What do you notice about the diagonal elements? What do you think this means?
- Which two variables have the highest (positive) correlation?

Part 4: Descriptive Statistics for a Simple Data Set

The reproducibility of a method for the determination of selenium in foods was investigated by taking nine samples from a single batch of brown rice and determining the selenium concentration in each.

The following results were obtained:

0.07 0.07 0.08 0.07 0.07 0.08 0.08 0.09 0.08

(Morena-Dominguez, T., Garda-Moreno, C. and Marine-Font, A. 1983. Analyst 108: 505)

In this part of the exercise, we shall compute basic descriptive statistics for this data set, such as sample mean, median, variance etc.
But first, we must enter the data into the ***R*** environment before we can perform any such calculations.

We shall create a "vector" called `X` using the following piece of code.

We can check that it has been entered successfully by entering the name of the data set into the command line and pressing return. The contents of the vector should be printed to the screen.

```
> X = c(0.07, 0.07,0.08,0.07,0.07,0.08,0.08,0.09,0.08)
> X
[1] 0.07 0.07 0.08 0.07 0.07 0.08 0.08 0.09 0.08
```

We are now able to perform many useful statistical calculations on this data set.

Enter the first six outputs on your submission sheet (i.e. `mean` to `summary`)

- The mean value of the data set: `mean(X)`
- The standard Deviation `sd(X)`
- The median of the data set `median(X)`
- The length (number of elements) of the data set `length(X)`
- The sum of the elements `sum(X)`
- Statistical summary of the data set : `summary(X)`
- The variance of the data set : `var(X)`
- The product of the data set : `prod(X)`
- The inter-quartile range of the data set `IQR(X)`