1 Reading in a CSV file

To read in a csv file, it is convenient to save it to your ***working directory***, the default file directory that R uses. Once it is in the working directory, you can use the `read.csv()` command to load it into the R environment.

```
getwd()
#[1] "C:/Users/Computer5/Documents"

HW0 <- read.csv("hw0_data.csv",header = TRUE)
```

You can change the working directory using the `setwd()` to the one you require.

```
getwd()
#[1] "C:/Users/Computer5/Documents"

setwd("C:/Users/Computer5/Documents/R")

#  > getwd()
#  [1] "C:/Users/Computer5/Documents/R"
```

2 Inspecting the data set

2a) Dimensions

2b) Column names (i.e. variables names)

2c) structure of data frame

Lets compute the dimensions of the data frame ***iris***, and also the length of the ***Nile*** data set.

```
dim(iris)
nrow(iris)
ncol(iris)
length(Nile)
```

Data frames often have specifically named rows and columns. Lets find out the names of the variables (columns) and cases (rows) for the data sets ***iris*** and ***mtcars***.

```
names(iris)
colnames(iris)
rownames(iris) # simply the case numbers.

names(mtcars)
rownames(mtcars)
colnames(mtcars)
```

2.1 The summary() command

The summary() command can be used to extract a short statistical summary (if applicable) from each column of the data frame. If there are missing values, the frequency of missing values will also be listed for each column.

```
> 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
```

2.2 Types of Data in Dataframes

What is the data type of each column in the iris data set? To find out, we use the str() command.

```
str(iris)
```

The output of this command is given below. There are four numeric variables, and one factor (i.e. categorical) variable.

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
>
```

3 Reduction and Subsetting

3a) The `head()` and `tail()` function

3b) Accessing a particular row or set of rows

3c) subsetting with a relational condition.

3.1 The `head()` and `tail()` function

The head and tail functions can be used to access the first six and last six rows of a data frame. If a different number of rows is required, all you have to do is specify that number as an additional argument.

```
head(iris)     #First 6 rows
head(iris,2)   #First 2 rows

tail(iris)     #Last 6 rows
tail(iris,4)   #Last 4 rows
```

3.2 Accessing a particular row or set of rows

- Each value in a dataframe can be accessed directly by specifying the row and column i.e. `df[r,c]`.
- To access a particular row, simply specify the row number, while leaving the column number blank i.e. `df[r,]`
- To access a particular column, simple specify the column number, while leaving the row number blank i.e. `df[,c]`

```
iris[10,2]

iris[10,]

Formaldehyde[,2]
```

```
> iris[10,2]
[1] 3.1
> iris[10,]
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
10          4.9         3.1          1.5         0.1  setosa
```

```
>
> Formaldehyde[,2]
[1] 0.086 0.269 0.446 0.538 0.626 0.782
```

4 Missing data

(4a) Determining the number of missing data items

(4b) Performing statistical operations removing missing data

As stated previously, the `summary()` command can be used to determine the number of missing data items in a data frame. The additional argument `na.rm=TRUE` can also be used with certain functions (see the help file)

```
> X <- c(4,6,3,12,NA,8)
> mean(X)
[1] NA
>
> summary(X)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    3.0     4.0     6.0     6.6     8.0    12.0       1
>
> mean(X ,na.rm = TRUE)
[1] 6.6
```

5 Subsetting Data

Logical and Relational Operator

- AND - The logical operator is &
- OR - The logical operator is ||

Selection using the `subset()` Function

The `subset( )` function is the easiest way to select variables and observation.
Example 1
In the following example we will use the ***iris*** data set, we select all rows that have a value of sepal length of 6 or more, and determine how many observations there are, and then compute the mean of the petal lengths. (The answer is 5.263, from the summary output).

```
#call the subset iris.2

iris.2 = subset(iris,iris$Sepal.Length >= 6)

dim(iris.2)

summary(iris.2)
```

Example 2
There are three types of iris - setosa, versicoloir and virginica. Suppose we wish to compute the median of petal widths for the setosa irises only.

The equality operator is ==.

```
> iris.setosa =subset(iris,iris$Species=="setosa")
> summary(iris.setosa)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100
 1st Qu.:4.800   1st Qu.:3.200   1st Qu.:1.400   1st Qu.:0.200
 Median :5.000   Median :3.400   Median :1.500   Median :0.200
 Mean   :5.006   Mean   :3.428   Mean   :1.462   Mean   :0.246
 3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300
 Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600
       Species
 setosa    :50
 versicolor: 0
 virginica : 0
```

Example 3
In the following example we will use the ***iris*** data set, we select all rows that have a value of sepal length of 6 or more, but have sepal width is at least 2.5.

```
> iris.3 = subset(iris,(iris$Sepal.Length >= 6)&(iris$Sepal.Width >=2.5))
> summary(iris.3)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :6.000   Min.   :2.500   Min.   :4.000   Min.   :1.200
 1st Qu.:6.300   1st Qu.:2.800   1st Qu.:4.750   1st Qu.:1.500
 Median :6.500   Median :3.000   Median :5.300   Median :1.800
 Mean   :6.641   Mean   :3.013   Mean   :5.313   Mean   :1.848
 3rd Qu.:6.900   3rd Qu.:3.200   3rd Qu.:5.750   3rd Qu.:2.150
 Max.   :7.900   Max.   :3.800   Max.   :6.900   Max.   :2.500
       Species
 setosa    : 0
 versicolor:21
 virginica :42
```