Recitation1

```
knitr::opts_chunk$set(echo = TRUE)
```

Marathon Dataset

The *summary* function provides min, max, mean, median, 1st quantile and 3rd quantile information for each column in the dataframe.

```
marathon <- read.csv('marathon.csv');
summary(marathon)
```

```
##       Year           Time
##  Min.   :1971   Min.   :2.411
##  1st Qu.:1978   1st Qu.:2.450
##  Median :1985   Median :2.469
##  Mean   :1985   Mean   :2.566
##  3rd Qu.:1992   3rd Qu.:2.542
##  Max.   :1999   Max.   :3.145
```

We see that the Marathon dataset contains 2 columns (Year and Time). When the dataset is large, we usually skim the dataset by examing the first few rows. This can be achieved with the function *head*, where the second input indicates the number of rows to be displayed. Similarly, *tail* function is used to examine the last few rows.

```
head(marathon,10)
```

```
##    Year    Time
## 1  1971 2.92278
## 2  1972 3.14472
## 3  1973 2.95194
## 4  1974 3.12472
## 5  1975 2.77056
## 6  1976 2.65306
## 7  1977 2.71944
## 8  1978 2.54167
## 9  1979 2.45917
## 10 1980 2.42833
```

To access a particular column, say the Time column, we can use the command *marathon$Time* or *marathon["Time"]*. The former outputs a vector, and the latter outputs another dataframe.

Compare

```
marathon$Time
```

```
##  [1] 2.92278 3.14472 2.95194 3.12472 2.77056 2.65306 2.71944 2.54167
##  [9] 2.45917 2.42833 2.42472 2.45389 2.45000 2.49167 2.47611 2.46833
## [17] 2.50472 2.46861 2.42500 2.51250 2.45889 2.41111 2.44000 2.46028
## [25] 2.46833 2.47167 2.47833 2.42139 2.41833
```

with

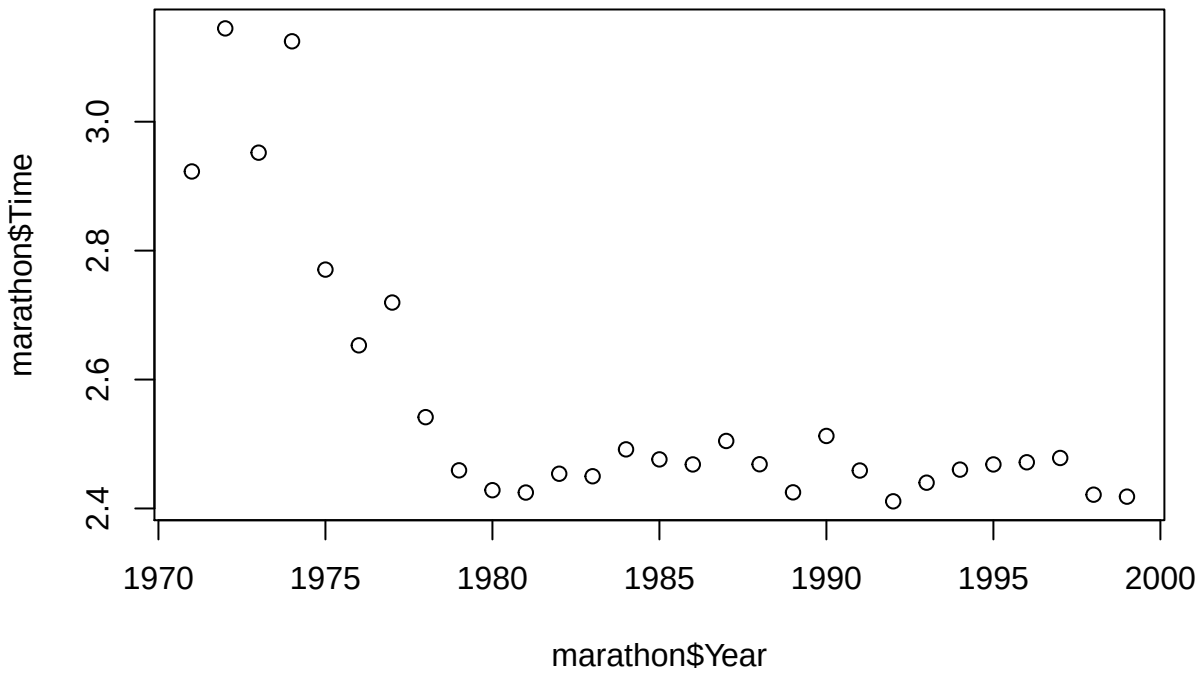
```
marathon['Time']
```

```
##       Time
```

```
## 1  2.92278
## 2  3.14472
## 3  2.95194
## 4  3.12472
## 5  2.77056
## 6  2.65306
## 7  2.71944
## 8  2.54167
## 9  2.45917
## 10 2.42833
## 11 2.42472
## 12 2.45389
## 13 2.45000
## 14 2.49167
## 15 2.47611
## 16 2.46833
## 17 2.50472
## 18 2.46861
## 19 2.42500
## 20 2.51250
## 21 2.45889
## 22 2.41111
## 23 2.44000
## 24 2.46028
## 25 2.46833
## 26 2.47167
## 27 2.47833
## 28 2.42139
## 29 2.41833
```

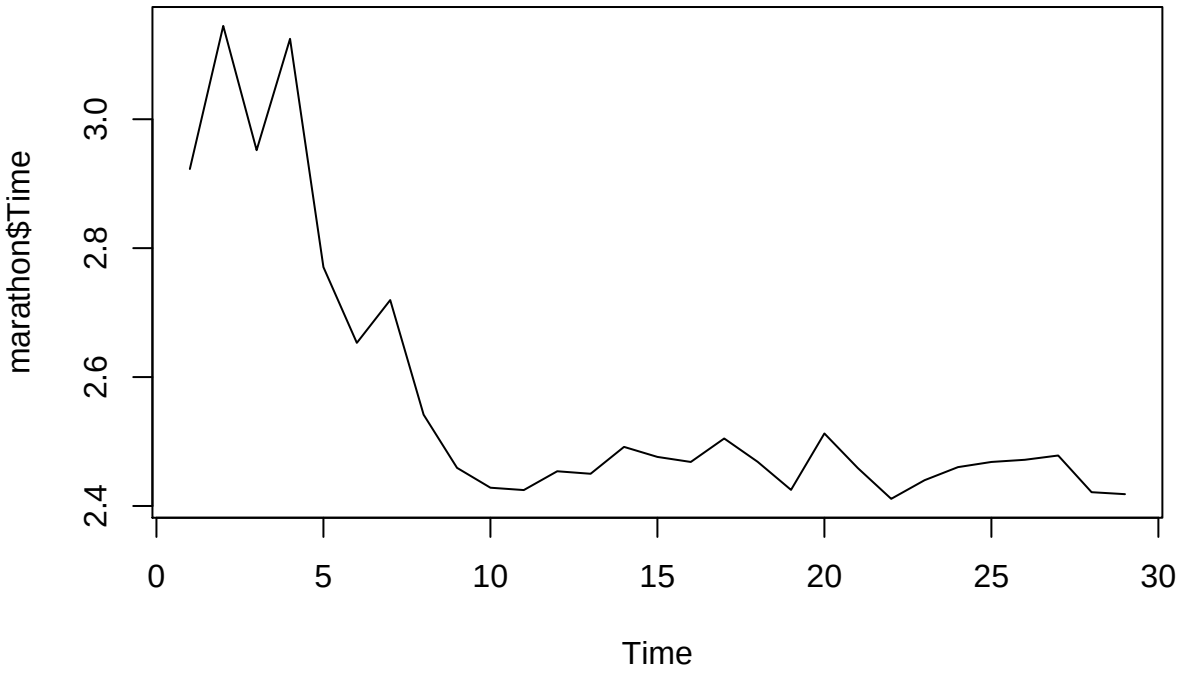
Finally, let us plot Time versus Year of the Marathon dataset. We can have either scatter plot

```
plot(marathon$Year, marathon$Time)
```



or the time series plot:

```
ts.plot(marathon$Time)
```



Sleep Dataset

The Sleep dataset is a built-in dataset in R. Let us first gather some information about this dataset.

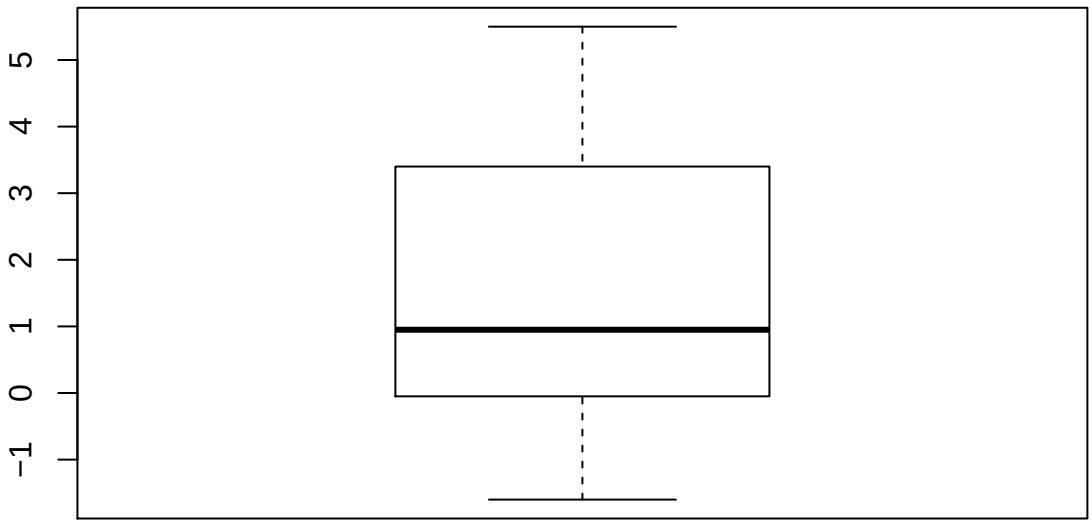
```
attributes(sleep)
```

```
## $names
## [1] "extra" "group" "ID"
##
## $class
## [1] "data.frame"
##
## $row.names
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

It has three columns, extra, group and ID.

Boxplot is a nice way to visualize the data, let us try for the extra column: the increase in the number of hours of sleep resulting from either drug.

```
boxplot(sleep['extra'])
```



In some data analysis, we need to focus on a subset of the data. For instance, we want to separate group=1 and group=2. This can be achieved using the following command:

```
group1 <- subset(sleep,group==1);
group2 <- subset(sleep,group==2)
```

We can check group1 and group 2:

```
group1
```

```
##    extra group ID
## 1    0.7     1  1
## 2   -1.6     1  2
## 3   -0.2     1  3
## 4   -1.2     1  4
## 5   -0.1     1  5
## 6    3.4     1  6
## 7    3.7     1  7
## 8    0.8     1  8
## 9    0.0     1  9
## 10   2.0     1 10
```

```
group2
```

```
##    extra group ID
## 11   1.9     2  1
## 12   0.8     2  2
## 13   1.1     2  3
## 14   0.1     2  4
## 15  -0.1     2  5
## 16   4.4     2  6
## 17   5.5     2  7
## 18   1.6     2  8
## 19   4.6     2  9
## 20   3.4     2 10
```

Now we can do a side by side comparison for the extra column between two groups in order to understand the effects of these two types of drugs.

```r
boxplot(group1$extra,group2$extra)
```