Data frames in R

Data frame subsets

There is a `subset()` function that allow us to retrieve a specific set of columns or drop a specific set of them. To illustrate it we will first create a simple Data frame.

```
df = data.frame(a =5:9, b = 6:10, c= 7:11)
print(df)
```

```
##   a  b  c
## 1 5  6  7
## 2 6  7  8
## 3 7  8  9
## 4 8  9 10
## 5 9 10 11
```

Then we can take for example columns "b" and "c" with the following code:

```
df1 = subset(df, select = c(b,c))
print(df1)
```

```
##    b  c
## 1  6  7
## 2  7  8
## 3  8  9
## 4  9 10
## 5 10 11
```

Another posible approach to obtain a data frame composed only of the columns "b" and "c" would to drop the column "a":

```
df2 = subset(df, select = -c(a))
print(df2)
```

```
##    b  c
## 1  6  7
## 2  7  8
## 3  8  9
## 4  9 10
## 5 10 11
```

Summary statistics

There are many packages that allow us to obtain the summary statistics of a data frame. An example would be the `fBasics` package, which has the `basicStats(dataFrame)` function.

Another possibility is to use the do.call functoin in combination with a set of specific functions such as `mean()`, `median()`, `sd()`..., to obtain a specific set of summary statistics:

```
ss = do.call(data.frame, list(
    mean = sapply(df, mean),
    sd = sapply(df, sd),
    median = sapply(df, median),
    min = sapply(df, min),
```

```
    max = sapply(df, max)
))
print(ss)
```

```
##   mean       sd median min max
## a    7 1.581139      7   5   9
## b    8 1.581139      8   6  10
## c    9 1.581139      9   7  11
```