Solution to exercise 3

Data exploration

Explore the `temperature` dataset

i) Load the dataset `temperature.csv` from the `01_Data` folder and assign it to an object with a meaningful name (e.g. `temperature`)

```
# Set the working directory
setwd("~/R_Basic_Introduction/01_Data") # replace with your path to the folder "01_Data"
```

```
# Load data
temperature <- read.csv(file = "temperature.csv")
```

ii) Get an overview of the dataset:
View the first six rows of the dataset

```
head(temperature)
```

```
##     site       temp day month
## 1 Zurich -2.6652164   6     1
## 2 Zurich -1.1469265   7     1
## 3 Zurich  1.9932443   8     1
## 4 Zurich  0.9122417   9     1
## 5 Zurich -4.1277218  10     1
## 6 Zurich -3.5909123  11     1
```

How many rows does the dataset have?

```
nrow(temperature) # returns the number or rows
```

```
## [1] 180
```

```
dim(temperature) # returns the number of rows and columns
```

```
## [1] 180   4
```

How many columns does the dataset have?

```
ncol(temperature) # returns the number or columns
```

```
## [1] 4
```

```r
dim(temperature) # returns the number of rows and columns
```

```
## [1] 180   4
```

What class do the columns have? Can you guess?

```r
class(temperature$site)
```

```
## [1] "factor"
```

```r
class(temperature$temp)
```

```
## [1] "numeric"
```

```r
class(temperature$day)
```

```
## [1] "integer"
```

```r
class(temperature$month)
```

```
## [1] "integer"
```

All in one: Sturcture of the dataset

```r
str(temperature)
```

```
## 'data.frame':    180 obs. of  4 variables:
##  $ site : Factor w/ 2 levels "Bern","Zurich": 2 2 2 2 2 2 2 2 2 2 ...
##  $ temp : num  -2.665 -1.147 1.993 0.912 -4.128 ...
##  $ day  : int  6 7 8 9 10 11 21 22 23 24 ...
##  $ month: int  1 1 1 1 1 1 1 1 1 1 ...
```

iii) Calculate the mean temperature

```
# select column 'temp'
temp <- temperature$temp # select with name
temp <- temperature[, 2] # select with number

# calculate the mean
mean(temp) # result is NA because 'temp' contains NA's
```

```
## [1] NA
```

```
mean(temp, na.rm = TRUE)
```

```
## [1] -2.092522
```

```
# or combined in one line
mean(temperature$temp, na.rm = TRUE)
```

```
## [1] -2.092522
```

iv) In which months were the measurements taken?

```
# extract unique values of column 'month'
month_measure <- unique(temperature$month)
month_measure
```

```
## [1] 1 2 3 4
```

v) What month and day was the maximum temperature measured?

```
# select column 'temp'
temp <- temperature$temp

# option 1
which.max(temp) # row number of maximum
```

```
## [1] 180
```

```
temperature[180, ] # select the row with the maximum temperature measurement
```

```
##     site     temp day month
## 180 Bern 13.36245  14     4
```

```
# option 2
temperature[which.max(temp), ]
```

```
##     site     temp day month
## 180 Bern 13.36245  14     4
```

```
# option 3
# select the row with the maximum temperature measurement and the columns
# 'Month' and 'Day'
temperature[180, 3:4]
```

```
##     day month
## 180  14     4
```

vi) Load the internal dataset `airquality` and caluclate the Pearson correlation between `Wind` and `Temp`. Do you expect a positive or negative correlation?

```
# load internal dataset airquality
data(airquality)

# get an overview of the dataset
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

```
str(airquality)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
# calculate the Pearson correlation
cor_wind_temp <- cor(airquality$Wind, airquality$Temp, method = "pearson")
cor_wind_temp
```

```
## [1] -0.4579879
```