R Self-Quiz

Please try to answer the following questions in under 1 hour.

- Download and install R from the Comprehensive R Archive Network. Make sure to choose a version that is appropriate for your computing platform (Windows, Mac, or Unix/Linux)
- Download the dataset available located on this web page and load it into R with the `read.csv` function. Assign the output of `read.csv` to an object named `dataset`.

```
## One way (easiest and fastest)
dataset <- read.csv("http://www.biostat.jhsph.edu/~rpeng/coursera/selfquiz/selfquiz-data.csv")

## You may want to store a local copy for later
download.file("http://www.biostat.jhsph.edu/~rpeng/coursera/selfquiz/selfquiz-data.csv",
    "selfquiz-data.csv")
dataset <- read.csv("selfquiz-data.csv")
```

- What are the column names of the data frame?

```
names(dataset)
```

```
## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
```

```
colnames(dataset)  ## also works
```

```
## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
```

- What are the row names of the data frame?

```
row.names(dataset)
```

```
##   [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"
##  [12] "12"  "13"  "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"
##  [23] "23"  "24"  "25"  "26"  "27"  "28"  "29"  "30"  "31"  "32"  "33"
##  [34] "34"  "35"  "36"  "37"  "38"  "39"  "40"  "41"  "42"  "43"  "44"
##  [45] "45"  "46"  "47"  "48"  "49"  "50"  "51"  "52"  "53"  "54"  "55"
##  [56] "56"  "57"  "58"  "59"  "60"  "61"  "62"  "63"  "64"  "65"  "66"
##  [67] "67"  "68"  "69"  "70"  "71"  "72"  "73"  "74"  "75"  "76"  "77"
##  [78] "78"  "79"  "80"  "81"  "82"  "83"  "84"  "85"  "86"  "87"  "88"
##  [89] "89"  "90"  "91"  "92"  "93"  "94"  "95"  "96"  "97"  "98"  "99"
## [100] "100" "101" "102" "103" "104" "105" "106" "107" "108" "109" "110"
## [111] "111" "112" "113" "114" "115" "116" "117" "118" "119" "120" "121"
## [122] "122" "123" "124" "125" "126" "127" "128" "129" "130" "131" "132"
## [133] "133" "134" "135" "136" "137" "138" "139" "140" "141" "142" "143"
## [144] "144" "145" "146" "147" "148" "149" "150" "151" "152" "153"
```

```
rownames(dataset)  ## also works
```

```
##   [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"
##  [12] "12"  "13"  "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"
##  [23] "23"  "24"  "25"  "26"  "27"  "28"  "29"  "30"  "31"  "32"  "33"
##  [34] "34"  "35"  "36"  "37"  "38"  "39"  "40"  "41"  "42"  "43"  "44"
##  [45] "45"  "46"  "47"  "48"  "49"  "50"  "51"  "52"  "53"  "54"  "55"
##  [56] "56"  "57"  "58"  "59"  "60"  "61"  "62"  "63"  "64"  "65"  "66"
##  [67] "67"  "68"  "69"  "70"  "71"  "72"  "73"  "74"  "75"  "76"  "77"
##  [78] "78"  "79"  "80"  "81"  "82"  "83"  "84"  "85"  "86"  "87"  "88"
##  [89] "89"  "90"  "91"  "92"  "93"  "94"  "95"  "96"  "97"  "98"  "99"
## [100] "100" "101" "102" "103" "104" "105" "106" "107" "108" "109" "110"
## [111] "111" "112" "113" "114" "115" "116" "117" "118" "119" "120" "121"
## [122] "122" "123" "124" "125" "126" "127" "128" "129" "130" "131" "132"
## [133] "133" "134" "135" "136" "137" "138" "139" "140" "141" "142" "143"
## [144] "144" "145" "146" "147" "148" "149" "150" "151" "152" "153"
```

- Extract the first 6 rows of the data frame and print them to the console

```
## One way
print(dataset[1:6, ])
```

```
##   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

# Alternatively
head(dataset, 6)

##   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
```

- How many observations (i.e. rows) are in this data frame?

```
nrow(dataset)

## [1] 153
```

- Extract the *last* 6 rows of the data frame and print them to the console

```
## One way
n <- nrow(dataset)
print(dataset[(n - 6 + 1):n, ])

##     Ozone Solar.R Wind Temp Month Day
## 148    14      20 16.6   63     9  25
## 149    30     193  6.9   70     9  26
## 150    NA     145 13.2   77     9  27
## 151    14     191 14.3   75     9  28
## 152    18     131  8.0   76     9  29
## 153    20     223 11.5   68     9  30

## Alternatively
tail(dataset)

##     Ozone Solar.R Wind Temp Month Day
## 148    14      20 16.6   63     9  25
## 149    30     193  6.9   70     9  26
## 150    NA     145 13.2   77     9  27
## 151    14     191 14.3   75     9  28
## 152    18     131  8.0   76     9  29
## 153    20     223 11.5   68     9  30
```

- How many missing values are in the "Ozone" column of this data frame?

```
miss <- is.na(dataset[, "Ozone"])  ## A vector of TRUE/FALSE
sum(miss)

## [1] 37
```

- What is the mean of the "Ozone" column in this dataset? Exclude missing values (coded as `NA`) from this calculation.

```
## Easy way
mean(dataset[, "Ozone"], na.rm = TRUE)

## [1] 42.13


## Hard way
use <- !is.na(dataset[, "Ozone"])  ## Find non-missing values
mean(dataset[use, "Ozone"])

## [1] 42.13
```

- Extract the subset of rows of the data frame where Ozone values are above 31 and Temp values are above 90.

```
## One way
subset(dataset, Ozone > 31 & Temp > 90)
```

```
##     Ozone Solar.R Wind Temp Month Day
## 69     97     267  6.3   92     7   8
## 70     97     272  5.7   92     7   9
## 120    76     203  9.7   97     8  28
## 121   118     225  2.3   94     8  29
## 122    84     237  6.3   96     8  30
## 123    85     188  6.3   94     8  31
## 124    96     167  6.9   91     9   1
## 125    78     197  5.1   92     9   2
## 126    73     183  2.8   93     9   3
## 127    91     189  4.6   93     9   4
```

- Use a `for` loop to create a vector of length 6 containing the mean of each column in the data frame (excluding all missing values).

```
m <- numeric(6)
for (i in 1:6) {
    m[i] <- mean(dataset[, i], na.rm = TRUE)
}
print(m)
```

```
## [1]  42.129 185.932   9.958  77.882   6.993  15.804
```

- Use the `apply` function to calculate the standard deviation of each column in the data frame (excluding all missing values).

```
s <- apply(dataset, 2, sd, na.rm = TRUE)
print(s)
```

```
##   Ozone Solar.R    Wind    Temp   Month     Day
##  32.988  90.058   3.523   9.465   1.417   8.865
```

- Calculate the mean of "Ozone" for each Month in the data frame and create a vector containing the monthly means (exclude all missing values).

```
tapply(dataset$Ozone, dataset$Month, mean, na.rm = TRUE)
```

```
##     5     6     7     8     9
## 23.62 29.44 59.12 59.96 31.45
```

- Draw a random sample of 5 rows from the data frame

```
set.seed(1)  ## Just so the answer is repeatable
dataset[sample(nrow(dataset), 5), ]
```

```
##     Ozone Solar.R Wind Temp Month Day
## 41     39     323 11.5   87     6  10
## 57     NA     127  8.0   78     6  26
## 87     20      81  8.6   82     7  26
## 137     9      24 10.9   71     9  14
## 31     37     279  7.4   76     5  31
```