dsmath-practice

1 6 correlation

1. Download a data set from the Internet and turn it straightaway into a data frame using `read.csv`. For the `file`, use the URL, and set `header` and `stringsAsFactors` to `TRUE`.

```
df <- read.csv(
  file="https://tinyurl.com/494vdr56",
  header=TRUE,
  stringsAsFactors=TRUE)
```

2. Check the data out: what's the structure?

```
str(df)
```

```
'data.frame':   10 obs. of  4 variables:
 $ Weight: int  55 85 75 42 93 63 58 75 89 67
 $ Height: int  161 185 174 154 188 178 170 167 181 178
 $ Sex   : Factor w/ 2 levels "female","male": 1 2 2 1 2 2 1 2 2 1
 $ Name  : Factor w/ 10 levels "Carl","Carla",..: 7 8 9 2 1 3 6 4 5 10
```

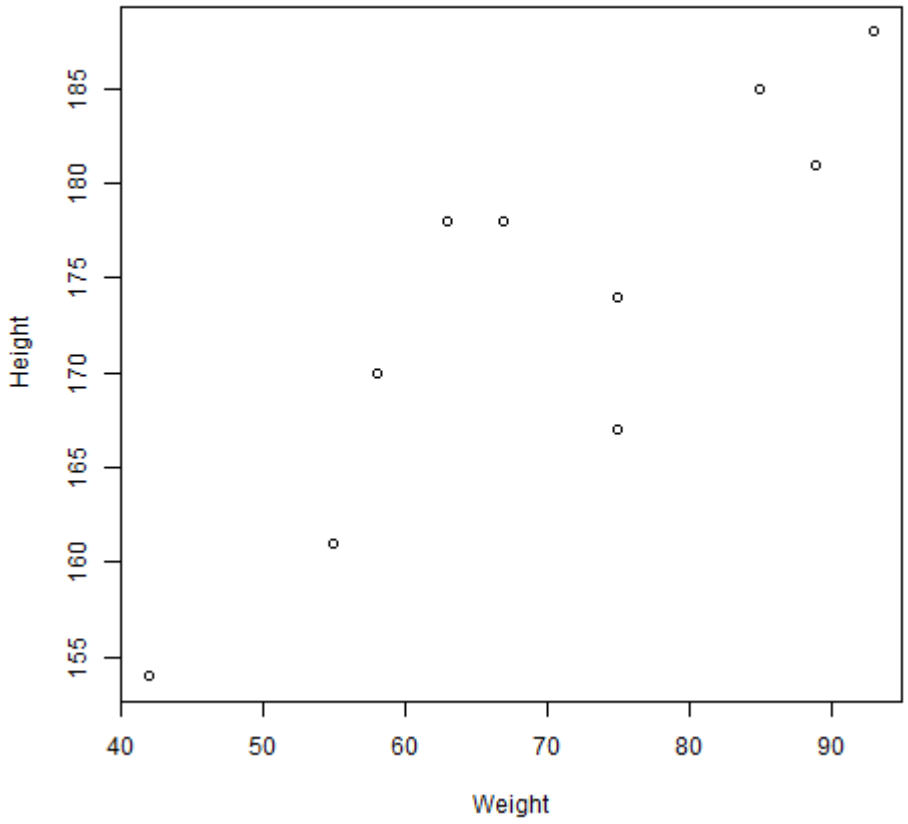
How does the data frame look like? Print it.

```
df
```

```
   Weight Height    Sex  Name
1      55    161 female  Jane
2      85    185   male   Jim
3      75    174   male   Joe
4      42    154 female Carla
5      93    188   male  Carl
6      63    178   male Chris
7      58    170 female  Dora
8      75    167   male  Dave
9      89    181   male Derek
10     67    178 female Lucia
```

3. Create a plot of `Height` vs. `Weight`.

```
plot(data=df, Height ~ Weight)  # plot(x=df$Weight, y=df$Height)
```



4. Compute the correlation coefficient of `Height` with `Weight`.

```
cor(df$Height, df$Weight)
```

```
[1] 0.8621007
```

5. What do you conclude regarding the correlation of these features?

   My conclusion: height and weight are strongly positively correlated - people who are tall are also heavier.

6. Inspect the built-in data set `mtcars` and look at the `help`, too. Identify two variables:

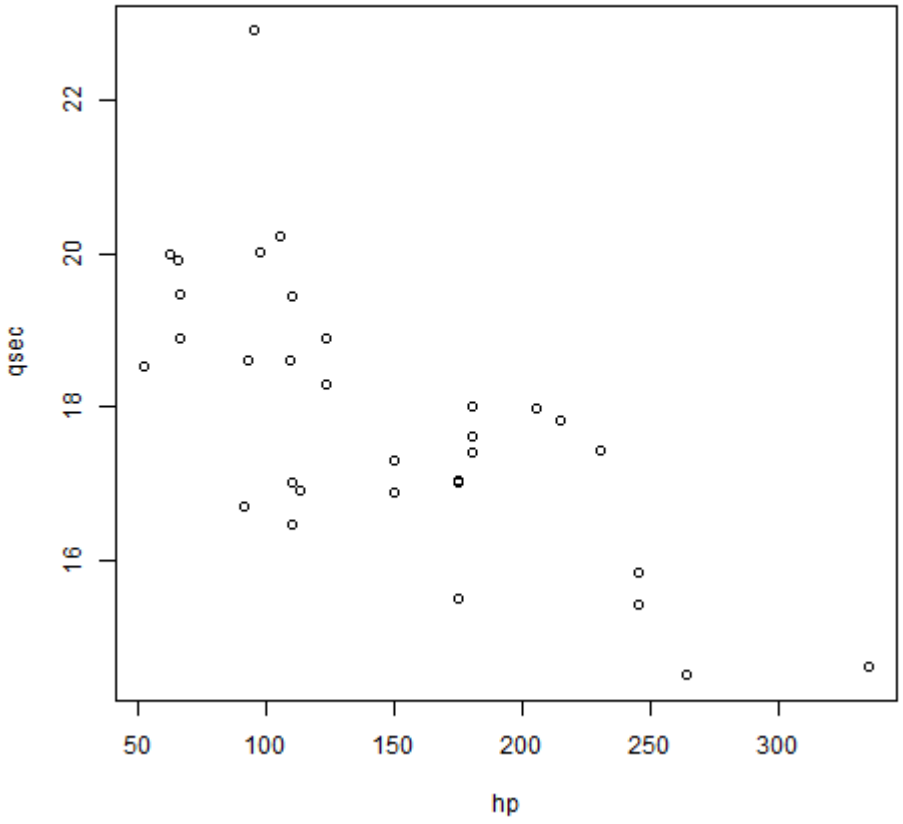
   - The vehicle's horsepower (in hp)
   - The shortest time taken to travel a quarter-mile distance (in sec)

```
str(mtcars)
```

```
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

7. Plot these last two variables with horsepower on the x-axis.

```
plot(data=mtcars, qsec ~ hp)
```



8. Compute the correlation coefficient for these last two variables.

```
cor(mtcars$qsec,mtcars$hp)
```

```
[1] -0.7082234
```

9. Compute the correlation coefficient after removing the two outliers visible in the plot:

```
qsec <- mtcars$qsec
outlier_qsec <- which(qsec==max(qsec))
qsec[outlier_qsec]
hp <- mtcars$hp
outlier_hp <- which(hp==max(hp))
hp[outlier_hp]
cor(qsec[-outlier_qsec],hp[-outlier_hp])
```

```
[1] 22.9
[1] 335
[1] -0.3748354
```

Created: 2022-11-03 Thu 09:18