Financial Big Data HW1

QF24 109090046呂庭羽

2023-02-21

### Question 1

The variable ‘Temp’ of the dataset ‘airquality’ contains temperature in degrees Fahrenheit. Please add another variable (column) to ‘airquality’ by the name ‘TempC’ that contains degrees Celsius converted from ‘Temp’. Use the ‘str()’ function to check the resulting ‘airquality’ and copy the result to your answer sheet.

# Load the airquality dataset  
data(airquality)  
  
# Create a new variable TempC   
airquality$TempC <- (airquality$Temp - 32) \* (5/9)  
  
# Check the structure of updated dataset  
str(airquality)

## 'data.frame': 153 obs. of 7 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 ...  
## $ TempC : num 19.4 22.2 23.3 16.7 13.3 ...

### Question 2

Install a new package ‘skimr’ (use help to find out what functions it contains). Apply the ‘skim’ function from this package to the ‘airquality’ dataset and then apply the base function ‘summary()’ to the ‘airquality’ dataset. Copy all the results to your answer sheet and explain the difference between the outputs of these two functions.

# Install skimr  
#install.packages("skimr")  
  
# Use `help` to find out what it contains  
# help(package = "skimr")  
  
# Apply the `skim()` function from skimr  
library(skimr)  
skim(airquality)

Data summary

|  |  |
| --- | --- |
| Name | airquality |
| Number of rows | 153 |
| Number of columns | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ozone | 37 | 0.76 | 42.13 | 32.99 | 1.00 | 18.00 | 31.50 | 63.25 | 168.00 | ▇▃▂▁▁ |
| Solar.R | 7 | 0.95 | 185.93 | 90.06 | 7.00 | 115.75 | 205.00 | 258.75 | 334.00 | ▅▃▅▇▅ |
| Wind | 0 | 1.00 | 9.96 | 3.52 | 1.70 | 7.40 | 9.70 | 11.50 | 20.70 | ▂▇▇▃▁ |
| Temp | 0 | 1.00 | 77.88 | 9.47 | 56.00 | 72.00 | 79.00 | 85.00 | 97.00 | ▂▃▇▇▃ |
| Month | 0 | 1.00 | 6.99 | 1.42 | 5.00 | 6.00 | 7.00 | 8.00 | 9.00 | ▇▇▇▇▇ |
| Day | 0 | 1.00 | 15.80 | 8.86 | 1.00 | 8.00 | 16.00 | 23.00 | 31.00 | ▇▇▇▇▆ |
| TempC | 0 | 1.00 | 25.49 | 5.26 | 13.33 | 22.22 | 26.11 | 29.44 | 36.11 | ▂▃▇▇▃ |

It reports the type of variable, the number of observations, the number of missing values and complete rate. And the mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum values. It also provides the histogram.

# Apply base function `summary()`  
summary(airquality)

## Ozone Solar.R Wind Temp   
## Min. : 1.00 Min. : 7.0 Min. : 1.700 Min. :56.00   
## 1st Qu.: 18.00 1st Qu.:115.8 1st Qu.: 7.400 1st Qu.:72.00   
## Median : 31.50 Median :205.0 Median : 9.700 Median :79.00   
## Mean : 42.13 Mean :185.9 Mean : 9.958 Mean :77.88   
## 3rd Qu.: 63.25 3rd Qu.:258.8 3rd Qu.:11.500 3rd Qu.:85.00   
## Max. :168.00 Max. :334.0 Max. :20.700 Max. :97.00   
## NA's :37 NA's :7   
## Month Day TempC   
## Min. :5.000 Min. : 1.0 Min. :13.33   
## 1st Qu.:6.000 1st Qu.: 8.0 1st Qu.:22.22   
## Median :7.000 Median :16.0 Median :26.11   
## Mean :6.993 Mean :15.8 Mean :25.49   
## 3rd Qu.:8.000 3rd Qu.:23.0 3rd Qu.:29.44   
## Max. :9.000 Max. :31.0 Max. :36.11   
##

It only provides minimum, 25th percentile, median, 75th percentile, maximum values and the number of missing values.

In summary, summary() functions provide a quick overview of the data., while the skim() function provides a more detailed and informative summary that includes more descriptive statistics.

### Question 3

Please use the ‘apply()’ function together with the ‘mean()’, ‘median()’, ‘sd()’ and ‘quantile()’ functions, respectively, to each variable of the ‘airquality’ dataset. Please include the ‘na.rm’ option with both ‘F’ and ‘T’ values to see how this option affects the results. Write down the commands and copy all the results to your answer sheet.

# means without removing NAs  
apply(airquality, 2, mean, na.rm = FALSE)

## Ozone Solar.R Wind Temp Month Day TempC   
## NA NA 9.957516 77.882353 6.993464 15.803922 25.490196

# means with removing NAs  
apply(airquality, 2, mean, na.rm = TRUE)

## Ozone Solar.R Wind Temp Month Day TempC   
## 42.129310 185.931507 9.957516 77.882353 6.993464 15.803922 25.490196

# medians without removing NAs  
apply(airquality, 2, median, na.rm = FALSE)

## Ozone Solar.R Wind Temp Month Day TempC   
## NA NA 9.70000 79.00000 7.00000 16.00000 26.11111

# medians with removing NAs  
apply(airquality, 2, median, na.rm = TRUE)

## Ozone Solar.R Wind Temp Month Day TempC   
## 31.50000 205.00000 9.70000 79.00000 7.00000 16.00000 26.11111

# standard deviations without removing NAs  
apply(airquality, 2, sd, na.rm = FALSE)

## Ozone Solar.R Wind Temp Month Day TempC   
## NA NA 3.523001 9.465270 1.416522 8.864520 5.258483

# standard deviations with removing NAs  
apply(airquality, 2, sd, na.rm = TRUE)

## Ozone Solar.R Wind Temp Month Day TempC   
## 32.987885 90.058422 3.523001 9.465270 1.416522 8.864520 5.258483

# quartiles without removing NAs can't be calculated  
# apply(airquality, 2, quantile, na.rm = FALSE)   
  
# quartiles with removing NAs  
apply(airquality, 2, quantile, na.rm = TRUE)

## Ozone Solar.R Wind Temp Month Day TempC  
## 0% 1.00 7.00 1.7 56 5 1 13.33333  
## 25% 18.00 115.75 7.4 72 6 8 22.22222  
## 50% 31.50 205.00 9.7 79 7 16 26.11111  
## 75% 63.25 258.75 11.5 85 8 23 29.44444  
## 100% 168.00 334.00 20.7 97 9 31 36.11111