기상청

[STEP2 데이터 탐색]

통계량에 근거한 데이터 탐색 방법

기상청

기상기후
빅데이터 분석 플랫폼

[분석교육] 기술통계량의 종류

1. 통계량 요약 함수

[분석 교육] 기술통계량의 종류(1/4)

기술통계란 측정이나 실험에서 수집한 자료의 정리, 표현, 요약, 해석 등을 통해 자료의 특성을 규명하는 통계적 방법이다. 자료의 특성을 표현하는 지표로 대푯값(평균값, 중앙값, 최빈값), 산포도(분산, 표준편차, 범위, 사분위수, 평균편차, 표준오차, 변이계수), 왜도 및 첨도가 있다.

● 대푯값

- 주어진 자료를 대표하는 특정 값을 그 자료의 대푯값이라 한다. 대푯값은 자료의 중심적인 경향이나 자료분포의 중심 위치를 나타내는데, 일반적으로 사용되는 대푯값에는 평균(Mean), 중앙값(Median), 최빈값(Mode)가 있다.

- **평균(Mean)**
  - 일반적으로 평균이라고 하는 것은 산술 평균을 의미한다.(기하평균, 조화평균과는 구별됨)
  - 산술 평균 : 자료의 모든 측정값을 합산하여 전체 자료의 수로 나눈 값
  - 기하 평균 : 합이 아닌 곱이 쓰이는 경우의 평균
  - 조화 평균 : 산술평균의 역수로 정의되며, 속력처럼 상대적인 비를 갖는 단위의 평균을 계산하는데 유용
  - 가중 산술 평균 : 같은 모집단에서 표본을 서로 다른 개수로 뽑을 때 유용

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$ 〈산술 평균〉

$$\bar{x} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$ 〈기하 평균〉

$$\bar{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$ 〈조화 평균〉

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$ 〈가중 산술 평균〉

- **중앙값(Median)**
  - 어떤 주어진 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값을 의미한다.
  - 예를 들어 1, 2, 100의 세 값이 있을 때, 2가 중앙값이다.
  - 값이 짝수개일 때에는 중앙값이 유일하지 않고 두개가 될 수 있다. 이 경우, 두 값의 평균을 중앙값으로 취한다.

[분석 교육] 기술통계량의 종류(2/4)

- **최빈값(Mode)**
  - 가장 많이 관측되는 수, 즉 주어진 값 중에서 가장 자주 나오는 값이다.
  - 예를 들어, {1, 3, 6, 6, 6, 7, 7, 12, 12, 17}의 최빈값은 6이다.
  - 최빈값은 유일한 값이 아닐 수도 있다.
  - 주어진 자료에서 평균이나 중앙값을 구하기 어려운 경우에 유용하다.

● 산포도(Statistical Dispersion)

- 산포도란 대푯값을 중심으로 자료들이 흩어져 있는 정도를 의미한다. 이는 하나의 수치로서 표현되며 수치가 작을수록 자료들이 대푯값에 밀집되어 있고, 클수록 자료들이 대푯값을 중심으로 멀리 흩어져 있다.

- **분산(Variance)**
  - 그 확률변수가 기댓값으로부터 얼마나 떨어진 곳에 분포하는지를 나타내는 값이다.
  - 분산보다는 분산의 제곱근인 표준편차를 더 자주 사용한다.
  - μ =E(X)가 확률변수 X의 기댓값(혹은 평균)일 때, 분산 var(X)는 다음과 같이 계산한다.

$$\operatorname{var}(X) = \mathrm{E}((X-\mu)^2)$$

- **표준편차(Standard Deviation)**
  - 분산은 편차를 제곱한 값이기 때문에 실제 편차보다 수치가 크게 나오는 특징이 있다.
  - 실제 편차와 근접하기 위해 분산에 제곱근을 하여 계산한다.
  - 표준편차가 작을수록 평균값에서 변량들의 거리가 가깝다.
  - 확률변수 X의 표준편차 σ는 다음과 같이 계산한다.(표본의 표준편차는 S로 나타냄)

$$\sigma = \sqrt{\mathrm{E}(X-\mathrm{E}(X))^2} = \sqrt{\mathrm{E}(X^2) - (\mathrm{E}(X))^2}$$

[분석 교육] 기술통계량의 종류(3/4)

- **범위(Range)**
  - 자료의 최댓값과 최솟값의 차이이다.
  - 자료가 변하는 정도를 잘 파악할 수 있으나 극단치의 영향을 많이 받아 잘 사용되지는 않는다.
  - R에서는 보통 최댓값과 최솟값을 각각 사용하며, 극단치를 확인하는 용도로 많이 쓰인다.

- **사분위수(Quartile)**
  - 분위수의 일종으로 크기가 작은 것이나 큰 것부터 나열하여 25%, 50%, 75% 되는 위치를 말한다.
  - 사분위 범위(Interquartile Range)는 25% 되는 위치(제 1사분위수)와 75% 되는 위치(제 3사분위수) 사이를 말한다.
  - 예를 들어, {1, 2, 3, 4, 5, 6, 10, 15, 30}의 제 1사분위수는 3이고, 제 3사분위수는 10이다.

- **평균편차(Mean Deviation)**
  - 절대편차(Absolute Deviation)이라고도 하며, 평균과 개별 관측치 사이 거리의 평균이다.
  - 각 측정치에서 전체 평균을 뺀 절댓값으로 계산된다.
  - 매우 크거나 작은 어느 하나의 값인 이상치로 인한 문제점을 보완하는 데 유용하다.

- **표준 오차 (Standard Error : SE)**
  - 모집단이 정규분포라는 가정하에, 여러 표본집단의 평균이 모집단 평균과 얼마나 떨어져 있는지를 측정하는 지표이다.
  - 표준오차가 작을수록 표본집단이 모집단에 근접할 가능성이 크다.
  - 표본평균의 표준오차는 표본표준편차를 표본크기의 제곱근으로 나누어 추정한다.

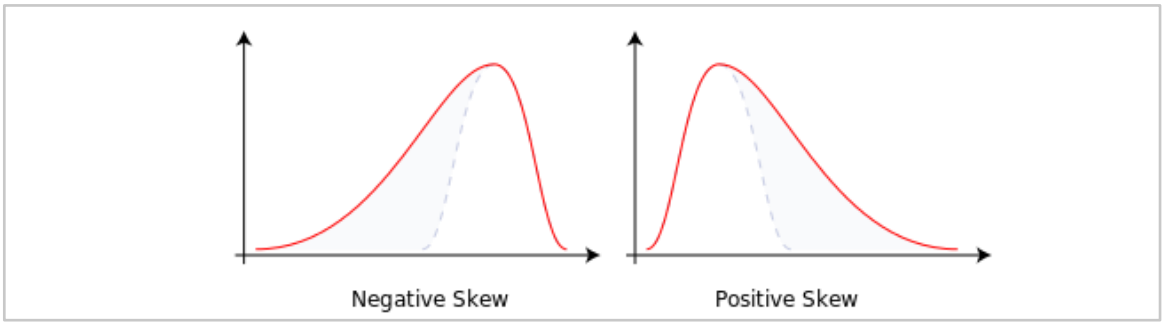
- **변이계수(Coefficient of Variation: CV)**
  - 변동계수라고도 하며 표준편차(σ)를 산술평균($\bar{x}$)으로 나눈 것이다.
  - 측정단위가 서로 다른 자료를 비교하고자 할 때 쓰인다.
  - 변동 계수의 값이 클수록 상대적인 차이가 크다는 것을 의미한다.

$$CV = \frac{\sigma}{\bar{x}}$$

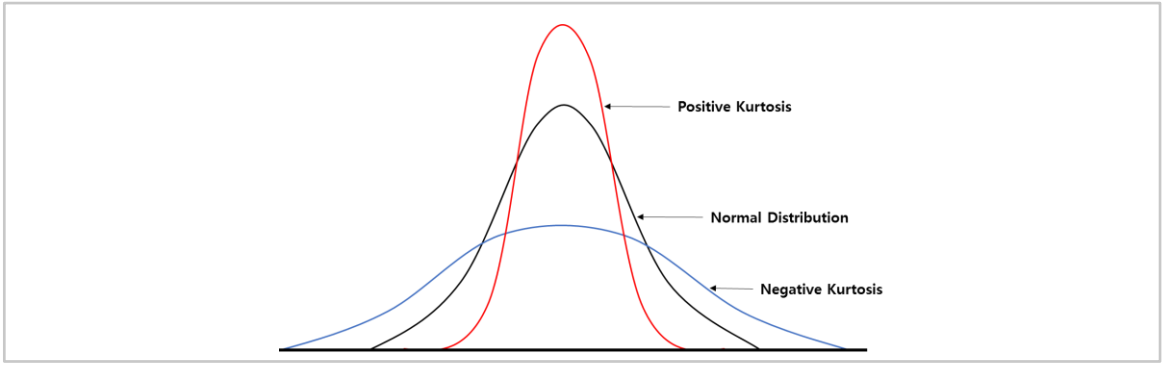
[분석 교육] 기술통계량의 종류(4/4)

왜도(Skewness) 및 첨도(Kurtosis)

- **왜도(Skewness)**
  - 왜도는 실수값 확률변수의 확률 분포 비대칭성을 나타내는 지표이다.
  - 왜도가 음수일 경우에는 확률밀도함수의 왼쪽 부분에 긴 꼬리를 가지며 중앙값을 포함한 자료가 오른쪽에 더 많이 분포해 있다.
  - 왜도가 양수일 때는 확률밀도 함수의 오른쪽 부분에 긴 꼬리를 가지며 자료가 왼쪽에 더 많이 분포해 있다.
  - 평균과 중앙값이 같으면 왜도는 0이 된다.



- **첨도(Kurtosis)**
  - 첨도는 확률분포의 뾰족한 정도를 나타내는 척도이다.
  - 관측치들이 어느 정도 집중적으로 중심에 몰려 있는가를 측정할 때 사용된다.
  - 첨도값이 0에 가까우면 산포도가 정규분포에 가깝다.
  - K < 0 의 경우, 정규분포보다 더 완만하게 납작한 분포이다.
  - K > 0 의 경우, 정규분포보다 더 뾰족한 분포이다.



참고 자료 : 위키백과, 통계청

1. 통계량 요약 함수(1/3)

- **summary**
  - 간략한 통계 요약을 보여줌
  - 수치형(Numeric) 데이터의 경우, 최솟값(Min), 1사분위수(1st Qu.), 중앙값(Median), 평균(Mean), 3사분위수(3st Qu.), 최댓값(Max) 제공
  - 요인형(Factor) 데이터의 경우, 각 레벨(수준)의 값 개수를 제공
  - 문자형(Character) 데이터의 경우, 문자형 데이터의 총 개수를 제공

▪ **Usage**

summary(object)

- object : 요약할 객체

▪ **Examples**

summary(onion.area) #팩터형 및 수치형 데이터 예시
summary(DATA) # 수치형 데이터 예시

```
> summary(onion.area)
     STN_ID    region_1    region_2        area_id        area
 129    : 1   경남:4   고령   : 1   4400000000:2   Min.   :  83
 170    : 1   경북:6   고흥   : 1   4500000000:3   1st Qu.: 325
 184    : 1   전남:9   군위   : 1   4600000000:9   Median : 523
 189    : 1   전북:3   김천   : 1   4700000000:6   Mean   : 760
 243    : 1   제주:2   남해   : 1   4800000000:4   3rd Qu.:1040
 260    : 1   충남:2   무안   : 1   5000000000:2   Max.   :3355
 (Other):20            (Other):20                  NA's   :8
       w
 Min.   :0.000
 1st Qu.:0.054
 Median :0.174
 Mean   :0.231
 3rd Qu.:0.404
 Max.   :0.577
> summary(DATA)
     STN_ID            TM              TA_AVG           TA_MAX
 Min.   : 95.0   Min.   :20060102   Min.   :-18.77   Min.   :-11.40
 1st Qu.:112.0   1st Qu.:20080626   1st Qu.:  5.07   1st Qu.:  9.20
 Median :131.0   Median :20101220   Median : 14.39   Median : 19.00
 Mean   :131.8   Mean   :20105384   Mean   : 13.20   Mean   : 17.43
 3rd Qu.:159.0   3rd Qu.:20130614   3rd Qu.: 21.59   3rd Qu.: 25.70
 Max.   :184.0   Max.   :20151231   Max.   : 32.72   Max.   : 37.80

     TA_MIN            SUM_RN           WS_AVG           WS_MAX
 Min.   :-26.700   Min.   :  0.10   Min.   : 0.040   Min.   : 0.300
 1st Qu.:  1.400   1st Qu.:  0.80   1st Qu.: 1.450   1st Qu.: 3.300
 Median : 10.500   Median :  3.60   Median : 2.100   Median : 4.400
 Mean   :  9.657   Mean   : 11.76   Mean   : 2.463   Mean   : 4.801
 3rd Qu.: 18.400   3rd Qu.: 13.10   3rd Qu.: 3.080   3rd Qu.: 5.800
 Max.   : 30.300   Max.   :372.50   Max.   :16.650   Max.   :27.200
                   NA's   :39750
```

1. 통계량 요약 함수(2/3)

- **psych::describe**
  - summary보다 자세한 기술통계량을 보여줌
  - 데이터수(n), 평균(mean), 표준편차(sd), 중앙값(median), 절사평균(trimmed), 중위수 절대편차(mad), 최솟값(min), 최댓값(max), 범위(range), 왜도(skew), 첨도(kurtosis), 표준오차(se) 제공

▪ **Usage**

describe(x, na.rm=TRUE, interp=FALSE, skew=TRUE, ranges=TRUE,
trim=.1, check=TRUE, …)

- x : 요약할 데이터 프레임이나 매트릭스 형태의 객체
- check : TRUE면 카테고리나 요인형 데이터를 수치형 데이터로 변환하여 통계량 계산
- na.rm : NA 제거 여부
- interp : 중앙값이 표준인지 채워야되는지 여부
- skew : 왜도와 첨도를 계산해야하는지 여부

▪ **Examples**

describe(DATA)

```
> describe(DATA)
           vars     n        mean       sd      median     trimmed      mad         min         max    range  skew kurtosis     se
STN_ID        1 58005      131.81    25.62      131.00      130.30    28.17       95.00      184.00    89.00  0.47    -0.86   0.11
TM            2 58005 20105383.92 28582.21 20101220.00 20105379.07 31291.76 20060102.00 20151231.00 91129.00  0.01    -1.22 118.68
TA_AVG        3 58005       13.20     9.72       14.39       13.58    11.95      -18.77       32.72    51.49 -0.29    -0.96   0.04
TA_MAX        4 58005       17.43     9.76       19.00       17.85    11.71      -11.40       37.80    49.20 -0.32    -0.97   0.04
TA_MIN        5 58005        9.66    10.12       10.50       10.03    12.45      -26.70       30.30    57.00 -0.27    -0.88   0.04
SUM_RN        6 18255       11.76    21.63        3.60        6.91     4.89        0.10      372.50   372.40  4.61    34.42   0.16
WS_AVG        7 58005        2.46     1.45        2.10        2.26     1.13        0.04       16.65    16.61  1.59     3.72   0.01
WS_MAX        8 58005        4.80     2.24        4.40        4.54     1.78        0.30       27.20    26.90  1.42     3.60   0.01
HM_AVG        9 58005       67.73    15.95       69.04       68.38    16.99        9.92      100.00    90.08 -0.35    -0.40   0.07
HM_MIN       10 58005       47.17    18.94       46.00       46.53    20.76        4.00      100.00    96.00  0.28    -0.58   0.08
SUM_SS       11 53563        6.35     3.47        7.00        6.43     3.85        0.01       14.20    14.19 -0.24    -1.05   0.01
SUM_SI       12 57962       13.21     6.69       12.49       13.00     7.58        0.05       32.70    32.65  0.25    -0.83   0.03
TD_AVG       13 58005        6.58    11.23        7.17        7.00    13.80      -27.57       27.93    55.50 -0.25    -0.90   0.05
PV_AVG       14 58005       12.39     8.25       10.27       11.68     8.70        0.64       37.69    37.05  0.61    -0.78   0.03
PA_AVG       15 58005     1006.86     9.86     1007.24     1007.21    10.13      967.53     1035.27    67.74 -0.34    -0.07   0.04
PS_AVG       16 58005     1016.14     7.91     1016.21     1016.15     9.09      989.62     1038.71    49.09 -0.01    -0.70   0.03
PS_MAX       17 58005     1018.83     8.07     1019.20     1018.86     9.49      994.60     1040.80    46.20 -0.04    -0.81   0.03
PS_MIN       18 58005     1013.79     8.06     1013.80     1013.82     9.04      975.50     1037.00    61.50 -0.05    -0.50   0.03
SD_HR3_MAX   19  2358        1.99     2.64        1.10        1.46     1.33        0.10       30.50    30.40  3.46    18.67   0.05
SD_TOT_MAX   20  4182        7.81    13.53        3.50        4.63     4.00        0.10      133.00   132.90  4.00    20.41   0.21
CA_TOT_AVG   21 58005        5.21     3.09        5.22        5.24     3.96        0.00       10.00    10.00 -0.05    -1.22   0.01
CA_MID_AVG   22 58005        3.35     2.47        3.22        3.26     3.29        0.00       10.00    10.00  0.21    -1.17   0.01
TS_AVG       23 58005       14.74    10.75       15.68       14.84    13.97      -13.98       42.68    56.66 -0.10    -1.21   0.04
TS_MAX       24 58005       22.84    13.26       23.80       22.74    15.27       -4.60       64.00    68.60  0.04    -0.79   0.06
TS_MIN       25 58005        9.96     9.75       10.10       10.02    13.49      -19.80       30.10    49.90 -0.04    -1.22   0.04
SUM_SML_EV   26 58005        3.13     1.99        2.80        2.98     2.08        0.00       22.50    22.50  0.74     0.57   0.01
HT           27 58005       77.79    51.01       68.94       69.71    25.00       20.45      222.80   202.35  1.50     1.75   0.21
```

1. 통계량 요약 함수(3/3)

- **psych::describeBy**
  - 특정 컬럼을 기준으로 그룹을 묶은 후 요약 값 계산

▪ **Usage**

describeBy(x, group = NULL, mat = FALSE, …)

- x : 요약을 수행할 데이터 프레임 또는 매트릭스
- group : 그룹으로 묶을 컬럼
- mat : 결과 출력 형태 설정, TRUE면 매트릭스 형태로, FALSE면 리스트형태로 결과 출력

▪ **Examples**

describeBy(DATA, group = "STN_ID")

```
> describeBy(DATA, group = "STN_ID")
$`95`
           vars    n        mean       sd      median     trimmed      mad         min         max    range  skew kurtosis     se
STN_ID        1 3284       95.00     0.00       95.00       95.00     0.00       95.00       95.00     0.00   NaN      NaN   0.00
TM            2 3284 20100704.49 25806.60 20100703.50 20100713.06 30095.30 20060102.00 20141231.00 81129.00  0.00    -1.23 450.33
TA_AVG        3 3284       11.82    10.67       13.82       12.56    12.54      -18.77       28.83    47.60 -0.49    -0.81   0.19
TA_MAX        4 3284       17.54    10.40       20.10       18.30    11.27      -11.40       35.40    46.80 -0.54    -0.83   0.18
TA_MIN        5 3284        6.73    11.47        8.10        7.40    13.94      -26.70       26.40    53.10 -0.40    -0.79   0.20
SUM_RN        6 1062       13.10    24.49        3.50        7.54     4.60        0.10      245.80   245.70  4.19    24.63   0.75
WS_AVG        7 3284        1.84     0.98        1.60        1.71     0.82        0.15        6.93     6.78  1.34     2.18   0.02
WS_MAX        8 3284        4.23     1.56        4.10        4.14     1.63        0.80       10.30     9.50  0.57     0.18   0.03
HM_AVG        9 3284       69.52    13.75       70.98       70.05    14.46       25.00       98.46    73.46 -0.35    -0.44   0.24
HM_MIN       10 3284       43.20    18.98       41.00       42.09    20.76        7.00       95.00    88.00  0.48    -0.55   0.33
SUM_SS       11 3054        6.37     3.45        7.10        6.48     3.85        0.01       12.60    12.59 -0.31    -1.09   0.06
SUM_SI       12 3284       12.89     6.69       12.54       12.78     8.13        0.65       27.84    27.19  0.12    -0.99   0.12
TD_AVG       13 3284        4.31    12.25        4.74        4.74    15.18      -26.09       25.41    51.50 -0.21    -1.01   0.21
PV_AVG       14 3284       11.13     8.09        8.64       10.36     8.12        0.74       32.51    31.77  0.66    -0.79   0.14
PA_AVG       15 3284      997.83     7.48      997.93      997.88     8.59      976.93     1017.87    40.94 -0.05    -0.70   0.13
PS_AVG       16 3284     1016.42     8.27     1016.52     1016.42     9.67      994.86     1038.71    43.85 -0.01    -0.77   0.14
PS_MAX       17 3284     1019.32     8.52     1019.75     1019.36    10.16      997.80     1040.80    43.00 -0.04    -0.85   0.15
PS_MIN       18 3284     1013.76     8.32     1013.70     1013.75     9.49      986.50     1036.50    50.00  0.00    -0.67   0.15
SD_HR3_MAX   19  173        1.54     1.60        1.00        1.29     1.19        0.10       12.10    12.00  2.25     9.61   0.12
SD_TOT_MAX   20  363        4.63     3.84        4.10        4.18     4.30        0.10       18.30    18.20  0.94     0.57   0.20
CA_TOT_AVG   21 3284        5.15     3.18        5.22        5.18     4.12        0.00       10.00    10.00 -0.05    -1.26   0.06
CA_MID_AVG   22 3284        3.49     2.61        3.33        3.39     3.47        0.00        9.56     9.56  0.19    -1.28   0.05
TS_AVG       23 3284       12.26    11.48       12.98       12.35    16.07       -9.76       36.96    46.72 -0.06    -1.35   0.20
TS_MAX       24 3284       21.48    14.49       22.55       21.24    17.72       -3.90       59.20    63.10  0.05    -1.01   0.25
TS_MIN       25 3284        7.31    10.63        7.10        7.39    13.79      -16.50       26.10    42.60  0.00    -1.25   0.19
SUM_SML_EV   26 3284        2.84     1.89        2.40        2.68     1.93        0.10        9.90     9.80  0.69    -0.42   0.03
HT           27 3284      153.70     0.00      153.70      153.70     0.00      153.70      153.70     0.00   NaN      NaN   0.00

$`101`
           vars    n        mean       sd      median     trimmed      mad         min         max    range  skew kurtosis     se
STN_ID        1 3649      101.00     0.00      101.00      101.00     0.00      101.00      101.00     0.00   NaN      NaN   0.00
TM            2 3649 20105664.21 28721.05 20101231.00 20105662.79 31324.37 20060102.00 20151231.00 91129.00  0.00    -1.22 475.46
TA_AVG        3 3649       11.45    10.93       12.79       11.90    14.17      -16.13       30.06    46.19 -0.26    -1.12   0.18
TA_MAX        4 3649       17.07    10.85       19.00       17.54    13.64       -8.70       36.70    45.40 -0.30    -1.15   0.18
TA_MIN        5 3649        6.74    11.50        7.50        7.13    14.68      -23.10       26.70    49.80 -0.21    -1.06   0.19
SUM_RN        6 1139       12.32    24.23        3.50        6.76     4.74        0.10      262.50   262.40  4.55    28.68   0.72
WS_AVG        7 3649        1.18     0.60        1.08        1.12     0.55        0.10        4.45     4.35  1.03     1.58   0.01
WS_MAX        8 3649        2.86     1.17        2.70        2.79     1.19        0.50        8.50     8.00  0.68     0.45   0.02
HM_AVG        9 3649       70.22    13.93       71.88       70.72    14.57       24.67      100.00    75.33 -0.33    -0.51   0.23
HM_MIN       10 3649       42.75    18.23       41.00       41.59    19.27        7.00      100.00    93.00  0.50    -0.35   0.30
SUM_SS       11 3362        6.11     3.32        6.70        6.15     3.41        0.10       13.20    13.10 -0.18    -0.91   0.06
SUM_SI       12 3648       13.44     6.71       12.61       13.22     7.39        0.05       31.29    31.24  0.28    -0.80   0.11
TD_AVG       13 3649        5.26    11.84        5.97        5.66    14.66      -24.47       26.32    50.79 -0.21    -1.04   0.20
PV_AVG       14 3649       11.65     8.19        9.43       10.91     8.60        0.86       34.29    33.43  0.62    -0.81   0.14
PA_AVG       15 3649     1007.13     7.77     1007.24     1007.15     9.00      984.48     1027.10    42.62 -0.03    -0.71   0.13
PS_AVG       16 3649     1016.43     8.16     1016.54     1016.43     9.55      993.14     1037.32    44.18 -0.01    -0.74   0.14
PS_MAX       17 3649     1019.58     8.42     1020.10     1019.62    10.08      997.50     1039.80    42.30 -0.05    -0.85   0.14
PS_MIN       18 3649     1013.69     8.23     1013.70     1013.68     9.49      987.90     1035.30    47.40  0.00    -0.63   0.14
SD_HR3_MAX   19  179        1.76     2.17        0.80        1.35     0.89        0.10       15.10    15.00  2.67    10.24   0.16
SD_TOT_MAX   20  406        5.43     5.04        4.20        4.73     4.45        0.10       30.00    29.90  1.38     2.36   0.25
CA_TOT_AVG   21 3649        5.27     3.07        5.44        5.33     3.94        0.00       10.00    10.00 -0.09    -1.20   0.05
CA_MID_AVG   22 3649        3.66     2.58        3.56        3.60     3.45        0.00       10.00    10.00  0.11    -1.24   0.04
TS_AVG       23 3649       13.27    11.66       14.06       13.35    16.25      -13.98       37.28    51.26 -0.07    -1.35   0.19
TS_MAX       24 3649       22.08    14.61       23.20       21.90    17.79       -3.90       60.70    64.60  0.03    -1.04   0.24
TS_MIN       25 3649        8.22    10.50        8.20        8.30    13.64      -19.80       28.30    48.10 -0.01    -1.25   0.17
SUM_SML_EV   26 3649        2.82     2.03        2.30        2.63     2.08        0.10       14.30    14.20  0.79     0.02   0.03
HT           27 3649       77.71     0.00       77.71       77.71     0.00       77.71       77.71     0.00   NaN      NaN   0.00
```

<결과 일부 생략>

참고 자료 : R Documentation

본 문서의 내용은 기상청의 날씨마루(http://big.kma.go.kr) 내
R 프로그래밍 교육 자료입니다.