03 데이터 다루기(1) ¶

학습 내용

- 데이터 프레임 알아보기
- read.csv()에 대해 알아보기
- read._excel()에 대해 알아보기
- read.table()에 대해 알아보기
- rda 파일 활용하기

3-1 데이터 프레임

- 가장 많이 사용하는 데이터 형태로서 행과 열로 구성된 사각형 모양의 표이다.

성별	연령	키	한달 소비
남	26	175	3000만원
여	33	177	4000만원
여	11	154	50만원

- 행과 열로 구성된다.

데이터 프레임 만들기

이름	국어	영어	수학
김칠수	80	90	95
홍길동	80	80	100
박난희	90	80	70

In [1]:

```
kor <- c(80,80,90)
eng <- c(90,80,80)
math <- c(95,100,70)
```

In [2]:

```
print(kor)
print(eng)
print(math)
```

```
[1] 80 80 90
[1] 90 80 80
[1]  95 100  70
```

In [3]:

```
df_score <- data.frame(kor, eng, math)
df_score
```

kor	eng	math
80	90	95
80	80	100
90	80	70

In [4]:

```
### 평균 구하기
mean(df_score)
```

```
Warning message in mean.default(df_score):
"argument is not numeric or logical: returning NA"
```

<NA>

In [5]:

```
mean(df_score$kor)
```

83.3333333333333

데이터 프레임 만들기 2

In [6]:

```
df_score2 <- data.frame(kor = c(80,80,90), eng=c(90,80,80), math=c(95,100,70))
df_score2
```

kor	eng	math
80	90	95
80	80	100
90	80	70

(ex) 3-1 실습해보기

- 데이터 프레임을 만들어 출력해 보자.

제품	가격	판매량
사과	6000	10
딸기	8000	5
수박	12000	5

(더 해보기) 가격 평균을 구해보기.

3-2 외부 데이터 불러오기

- read_excel :: readxl => 엑셀 파일 불러오기
- reac.csv => csv파일 불러오기

In [7]:

```
install.packages("readxl")
```

```
Warning message:
"unable to access index for repository http://www.stats.ox.ac.uk/pub/RWin/bin/window
s/contrib/3.5: (http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5:)
  URL 'http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5/PACKAGES'를 열 수
 없습니다"

package 'readxl' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:₩Users₩WITHJS₩AppData₩Local₩Temp₩RtmpOeGcND₩downloaded_packages
```

In [3]:

```
library(readxl)
```

In [17]:

```
df_bike <- read_excel("D:₩₩dataset₩₩bike₩₩train_bike.xlsx") # 첫번째 줄은 변수명으로 인식
head(df_bike,10)
```

datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	r
2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0000	3	
2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0000	8	
2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0000	5	
2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0000	3	
2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0000	0	
2011-01-01 05:00:00	1	0	0	2	9.84	12.880	75	6.0032	0	
2011-01-01 06:00:00	1	0	0	1	9.02	13.635	80	0.0000	2	
2011-01-01 07:00:00	1	0	0	1	8.20	12.880	86	0.0000	1	
2011-01-01 08:00:00	1	0	0	1	9.84	14.395	75	0.0000	1	
2011-01-01 09:00:00	1	0	0	1	13.12	17.425	76	0.0000	8	

In [6]:

```
print(is(df_exam))
print(dim(df_exam))
print(summary(df_exam))
```

```
[1] "tbl_df"     "tbl"        "data.frame" "list"       "oldClass"  
[6] "vector"    
[1] 10886    12
    datetime                       season         holiday       
 Min.   :2011-01-01 00:00:00   Min.   :1.000   Min.   :0.00000  
 1st Qu.:2011-07-02 07:15:00   1st Qu.:2.000   1st Qu.:0.00000  
 Median :2012-01-01 20:30:00   Median :3.000   Median :0.00000  
 Mean   :2011-12-27 05:56:22   Mean   :2.507   Mean   :0.02857  
 3rd Qu.:2012-07-01 12:45:00   3rd Qu.:4.000   3rd Qu.:0.00000  
 Max.   :2012-12-19 23:00:00   Max.   :4.000   Max.   :1.00000  
   workingday        weather           temp           atemp      
 Min.   :0.0000   Min.   :1.000   Min.   : 0.82   Min.   : 0.76  
 1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:13.94   1st Qu.:16.66  
 Median :1.0000   Median :1.000   Median :20.50   Median :24.24  
 Mean   :0.6809   Mean   :1.418   Mean   :20.23   Mean   :23.66  
 3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:26.24   3rd Qu.:31.06  
 Max.   :1.0000   Max.   :4.000   Max.   :41.00   Max.   :45.45  
    humidity        windspeed          casual         registered   
 Min.   :  0.00   Min.   : 0.000   Min.   :  0.00   Min.   :  0.0  
 1st Qu.: 47.00   1st Qu.: 7.002   1st Qu.:  4.00   1st Qu.: 36.0  
 Median : 62.00   Median :12.998   Median : 17.00   Median :118.0  
 Mean   : 61.89   Mean   :12.799   Mean   : 36.02   Mean   :155.6  
 3rd Qu.: 77.00   3rd Qu.:16.998   3rd Qu.: 49.00   3rd Qu.:222.0  
 Max.   :100.00   Max.   :56.997   Max.   :367.00   Max.   :886.0  
     count      
 Min.   :  1.0  
 1st Qu.: 42.0  
 Median :145.0  
 Mean   :191.6  
 3rd Qu.:284.0  
 Max.   :977.0  
```

In [16]:

```
df_exam <- read_excel("D:₩₩dataset₩₩Bike₩₩test_notitle.xlsx") # 첫번째 줄은 변수명으로 인식
head(df_exam,10)
```

datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed
2011-01-20 00:00:00	1	0	1	1	10.66	11.365	56	26.0027
2011-01-20 01:00:00	1	0	1	1	10.66	13.635	56	0.0000
2011-01-20 02:00:00	1	0	1	1	10.66	13.635	56	0.0000
2011-01-20 03:00:00	1	0	1	1	10.66	12.880	56	11.0014
2011-01-20 04:00:00	1	0	1	1	10.66	12.880	56	11.0014
2011-01-20 05:00:00	1	0	1	1	9.84	11.365	60	15.0013
2011-01-20 06:00:00	1	0	1	1	9.02	10.605	60	15.0013
2011-01-20 07:00:00	1	0	1	1	9.02	10.605	55	15.0013
2011-01-20 08:00:00	1	0	1	1	9.02	10.605	55	19.0012
2011-01-20 09:00:00	1	0	1	2	9.84	11.365	52	15.0013

- col_names를 이용하여 첫번째 행을 변수명이 아닌 데이터로 인식해서 불러온다.
- 변수명은 'X_숫자' 로 자동 지정.

In [15]:

```
df_exam <- read_excel("D:₩₩dataset₩₩Bike₩₩test_notitle.xlsx", col_names=F) # 첫번째 줄은 변수명으로
head(df_exam,10)
```

X__1	X__2	X__3	X__4	X__5	X__6	X__7	X__8
datetime	season	holiday	workingday	weather	temp	atemp	humidity	wind
40563	1	0	1	1	10.66	11.365	56	26.0027000000
40563.041666666664	1	0	1	1	10.66	13.635	56	
40563.083333333336	1	0	1	1	10.66	13.635	56	
40563.125	1	0	1	1	10.66	12.88	56	11
40563.166666666664	1	0	1	1	10.66	12.88	56	11
40563.208333333336	1	0	1	1	9.84	11.365	60	15.0013000000
40563.25	1	0	1	1	9.02	10.605	60	15.0013000000
40563.291666666664	1	0	1	1	9.02	10.605	55	15.0013000000
40563.333333333336	1	0	1	1	9.02	10.605	55	19.0012000000

(ex) 3-2 실습해보기

- sheet=3을 이용하여 excel_exam_sheet.xlsx를 불러오기

In [14]:

```
df_csv_exam <- read.csv("D:₩₩dataset₩₩bike₩₩train.csv", header=F)
head(df_csv_exam,10)
```

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V
datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	register
2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0	3	
2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0	8	
2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0	5	
2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0	3	
2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0	0	
2011-01-01 05:00:00	1	0	0	2	9.84	12.88	75	6.0032	0	
2011-01-01 06:00:00	1	0	0	1	9.02	13.635	80	0	2	
2011-01-01 07:00:00	1	0	0	1	8.2	12.88	86	0	1	
2011-01-01 08:00:00	1	0	0	1	9.84	14.395	75	0	1	

In [13]:

```
df_csv_exam <- read.csv("D:₩₩dataset₩₩bike₩₩train.csv", header=T)
head(df_csv_exam,10)
```

datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	r
2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0000	3	
2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0000	8	
2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0000	5	
2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0000	3	
2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0000	0	
2011-01-01 05:00:00	1	0	0	2	9.84	12.880	75	6.0032	0	
2011-01-01 06:00:00	1	0	0	1	9.02	13.635	80	0.0000	2	
2011-01-01 07:00:00	1	0	0	1	8.20	12.880	86	0.0000	1	
2011-01-01 08:00:00	1	0	0	1	9.84	14.395	75	0.0000	1	
2011-01-01 09:00:00	1	0	0	1	13.12	17.425	76	0.0000	8	

3-3 데이터를 파일로 저장하기

In [15]:

```
df_score3 <- data.frame(kor, eng, math)
df_score3
```

kor	eng	math
80	90	95
80	80	100
90	80	70

In [16]:

```
write.csv(df_score3, file="df_score.csv")
```

3-4 RData 파일 활용하기

- save(데이터셋, file="파일명.rda")
- load("___.rda")

In [17]:

```
save(df_score3, file="df_score.rda")
```

In [18]:

```
rm(df_score3)
```

In [19]:

```
# 변수의 리스트 확인
ls.str()
```

```
df_csv_exam : 'data.frame':	20 obs. of  5 variables:
 $ id     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ class  : int  1 1 1 1 2 2 2 2 3 3 ...
 $ math   : int  50 60 45 30 25 50 80 90 20 50 ...
 $ english: int  98 97 86 98 80 89 90 78 98 98 ...
 $ science: int  50 60 78 58 65 98 45 25 15 45 ...
df_exam : Classes 'tbl_df', 'tbl' and 'data.frame':	8 obs. of  5 variables:
 $ X__1: num  1 2 3 4 5 6 7 8
 $ X__2: num  1 1 2 2 3 3 4 4
 $ X__3: num  50 60 25 50 20 50 46 48
 $ X__4: num  98 97 80 89 98 98 98 87
 $ X__5: num  50 60 65 98 15 45 65 12
df_score : 'data.frame':	3 obs. of  3 variables:
 $ kor : num  80 80 90
 $ eng : num  90 80 80
 $ math: num  95 100 70
df_score2 : 'data.frame':	3 obs. of  3 variables:
 $ kor : num  80 80 90
 $ eng : num  90 80 80
 $ math: num  95 100 70
eng :  num [1:3] 90 80 80
kor :  num [1:3] 80 80 90
math :  num [1:3] 95 100 70
```

In [20]:

```
## 불러오기
load("df_score.rda")
ls.str()
```

```
df_csv_exam : 'data.frame':	20 obs. of  5 variables:
 $ id     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ class  : int  1 1 1 1 2 2 2 2 3 3 ...
 $ math   : int  50 60 45 30 25 50 80 90 20 50 ...
 $ english: int  98 97 86 98 80 89 90 78 98 98 ...
 $ science: int  50 60 78 58 65 98 45 25 15 45 ...
df_exam : Classes 'tbl_df', 'tbl' and 'data.frame':	8 obs. of  5 variables:
 $ X__1: num  1 2 3 4 5 6 7 8
 $ X__2: num  1 1 2 2 3 3 4 4
 $ X__3: num  50 60 25 50 20 50 46 48
 $ X__4: num  98 97 80 89 98 98 98 87
 $ X__5: num  50 60 65 98 15 45 65 12
df_score : 'data.frame':	3 obs. of  3 variables:
 $ kor : num  80 80 90
 $ eng : num  90 80 80
 $ math: num  95 100 70
df_score2 : 'data.frame':	3 obs. of  3 variables:
 $ kor : num  80 80 90
 $ eng : num  90 80 80
 $ math: num  95 100 70
df_score3 : 'data.frame':	3 obs. of  3 variables:
 $ kor : num  80 80 90
 $ eng : num  90 80 80
 $ math: num  95 100 70
eng :  num [1:3] 90 80 80
kor :  num [1:3] 80 80 90
math :  num [1:3] 95 100 70
```