03 데이터 다루기(1) ¶

학습 내용

- 데이터 프레임 알아보기
- read.csv()에 대해 알아보기
- read._excel()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 [8]:

```
library(readxl)
```

In [9]:

```
df_exam <- read_excel("D:WWdatasetWWR_DoitWWexcel_exam.xlsx") # 첫번째 줄은 변수명으로 인식
df_exam
```

id	class	math	english	science
1	1	50	98	50
2	1	60	97	60
3	1	45	86	78
4	1	30	98	58
5	2	25	80	65
6	2	50	89	98
7	2	80	90	45
8	2	90	78	25
9	3	20	98	15
10	3	50	98	45
11	3	65	65	65
12	3	45	85	32
13	4	46	98	65
14	4	48	87	12
15	4	75	56	78
16	4	58	98	65
17	5	65	68	98
18	5	80	78	90
19	5	89	68	87
20	5	78	83	58

In [10]:

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

```
[1] "tbl_df"     "tbl"        "data.frame" "list"       "oldClass"
[6] "vector"
[1] 20  5
       id            class        math          english         science
 Min.   : 1.00   Min.   :1   Min.   :20.00   Min.   :56.0   Min.   :12.00
 1st Qu.: 5.75   1st Qu.:2   1st Qu.:45.75   1st Qu.:78.0   1st Qu.:45.00
 Median :10.50   Median :3   Median :54.00   Median :86.5   Median :62.50
 Mean   :10.50   Mean   :3   Mean   :57.45   Mean   :84.9   Mean   :59.45
 3rd Qu.:15.25   3rd Qu.:4   3rd Qu.:75.75   3rd Qu.:98.0   3rd Qu.:78.00
 Max.   :20.00   Max.   :5   Max.   :90.00   Max.   :98.0   Max.   :98.00
```

In [11]:

```
df_exam <- read_excel("D:WWdatasetWWR_DoitWWexcel_exam_novar.xlsx") # 첫번째 줄은 변수명으로 인식
df_exam
```

1	1__1	50	98	50__1
2	1	60	97	60
3	2	25	80	65
4	2	50	89	98
5	3	20	98	15
6	3	50	98	45
7	4	46	98	65
8	4	48	87	12

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

In [12]:

```
df_exam <- read_excel("D:WWdatasetWWR_DoitWWexcel_exam_novar.xlsx", col_names=F) # 첫번째 줄은 변수
df_exam
```

X__1	X__2	X__3	X__4	X__5
1	1	50	98	50
2	1	60	97	60
3	2	25	80	65
4	2	50	89	98
5	3	20	98	15
6	3	50	98	45
7	4	46	98	65
8	4	48	87	12

(ex) 3-2 실습해보기

- sheet=3을 이용하여 excel_exam_sheet.xlsx를 불러오기

In [13]:

```
df_csv_exam <- read.csv("D:WWdatasetWWR_DoitWWcsv_exam.csv", header=F)
df_csv_exam
```

V1	V2	V3	V4	V5
id	class	math	english	science
1	1	50	98	50
2	1	60	97	60
3	1	45	86	78
4	1	30	98	58
5	2	25	80	65
6	2	50	89	98
7	2	80	90	45
8	2	90	78	25
9	3	20	98	15
10	3	50	98	45
11	3	65	65	65
12	3	45	85	32
13	4	46	98	65
14	4	48	87	12
15	4	75	56	78
16	4	58	98	65
17	5	65	68	98
18	5	80	78	90
19	5	89	68	87
20	5	78	83	58

In [14]:

```
df_csv_exam <- read.csv("D:WWdatasetWWR_DoitWWcsv_exam.csv", header=T)
df_csv_exam
```

id	class	math	english	science
1	1	50	98	50
2	1	60	97	60
3	1	45	86	78
4	1	30	98	58
5	2	25	80	65
6	2	50	89	98
7	2	80	90	45
8	2	90	78	25
9	3	20	98	15
10	3	50	98	45
11	3	65	65	65
12	3	45	85	32
13	4	46	98	65
14	4	48	87	12
15	4	75	56	78
16	4	58	98	65
17	5	65	68	98
18	5	80	78	90
19	5	89	68	87
20	5	78	83	58

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
```