

데이터과학을 위한 **R**프로그래밍

5주차. 데이터탐색



이혜선 교수

포항공과대학교 산업경영공학과



목차

5주차. 데이터탐색

1차시

데이터 다루기(결합, 분할)

2차시

데이터탐색과 기술통계치

3차시

데이터시각화를 이용한 데이터탐색



5주차

1차시

데이터 다루기 (결합, 분할)

데이터 다루기

✓ dplyr 함수

dplyr 함수명	내용	비교
inner_join() left_join() right_join() full_join()	데이터 결합	merge()
bind_rows()	행 기준 데이터 결합	rbind()
arrange()	데이터 정렬	order(), sort()
filter()	조건식에 맞는 데이터 추출	subset()
select()	열의 추출	data[,c("name")]
summarize()	요약 통계치	aggregate

데이터 다루기

✓ 데이터 결합 방법 1 : `merge(data1, data2, by="ID")`

➤ data1과 data2는 아래와 같이 식별변수 ID를 기준으로 결합

➤ data1 : 게임장르, 나이, 성별 data2 : 주당게임시간, 음주경험, 흡연경험

data1.csv

A	B	C	D
ID	age	gender	game
111	16	F	RTS
112	17	F	FPS
113	15	M	Sport
114	18	M	MMORPG
115	14	F	MMORPG
116	15	F	FPS
117	13	M	Sport
118	19	F	FPS
119	17	M	Sport
120	18	F	RTS

data2.csv

A	B	C	D
ID	hourwk	alcohol	smoke
111	10	yes	yes
112	8	no	no
113	4	no	no
114	10	no	no
115	2	no	yes
116	10	yes	yes
117	12	yes	yes
118	8	no	no
119	6	no	no
120	4	no	no



	ID	age	gender	game	hourwk	alcohol	smoke
1	111	16	F	RTS	10	yes	yes
2	112	17	F	FPS	8	no	no
3	113	15	M	Sport	4	no	no
4	114	18	M	MMORPG	10	no	no
5	115	14	F	MMORPG	2	no	yes
6	116	15	F	FPS	10	yes	yes
7	117	13	M	Sport	12	yes	yes
8	118	19	F	FPS	8	no	no
9	119	17	M	Sport	6	no	no
10	120	18	F	RTS	4	no	no

데이터 다루기

☑ 데이터 결합 방법 2 : `inner_join(data1, data2, by="ID")`

➤ data1과 data2는 아래와 같이 식별변수 ID를 기준으로 결합

➤ data1 : 게임장르, 나이, 성별 data2 : 주당게임시간, 음주경험, 흡연경험

data1.csv

A	B	C	D
ID	age	gender	game
111	16	F	RTS
112	17	F	FPS
113	15	M	Sport
114	18	M	MMORPG
115	14	F	MMORPG
116	15	F	FPS
117	13	M	Sport
118	19	F	FPS
119	17	M	Sport
120	18	F	RTS

data2.csv

A	B	C	D
ID	hourwk	alcohol	smoke
111	10	yes	yes
112	8	no	no
113	4	no	no
114	10	no	no
115	2	no	yes
116	10	yes	yes
117	12	yes	yes
118	8	no	no
119	6	no	no
120	4	no	no



	ID	age	gender	game	hourwk	alcohol	smoke
1	111	16	F	RTS	10	yes	yes
2	112	17	F	FPS	8	no	no
3	113	15	M	Sport	4	no	no
4	114	18	M	MMORPG	10	no	no
5	115	14	F	MMORPG	2	no	yes
6	116	15	F	FPS	10	yes	yes
7	117	13	M	Sport	12	yes	yes
8	118	19	F	FPS	8	no	no
9	119	17	M	Sport	6	no	no
10	120	18	F	RTS	4	no	no

데이터 다루기

✓ 데이터 결합 : `inner_join(data1, data2, by="ID")`

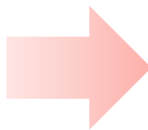
➤ dat1과 dat2를 ID를 기준으로 결합 (관측치수는 동일함, 변수들의 정보가 추가됨)

```
# install.packages(dplyr)
library(dplyr)

# set working directory
setwd("D:/tempstore/moocr")

# practice data with dplyr
dat1<-read.csv(file="data1.csv")
dat2<-read.csv(file="data2.csv")

# data merging
# dat12<-merge(dat1, dat2, by="ID")
dat12<-inner_join(dat1,dat2, by="ID")
```



	ID	age	gender	game	hourwk	alcohol	smoke
1	111	16	F	RTS	10	yes	yes
2	112	17	F	FPS	8	no	no
3	113	15	M	Sport	4	no	no
4	114	18	M	MMORPG	10	no	no
5	115	14	F	MMORPG	2	no	yes
6	116	15	F	FPS	10	yes	yes
7	117	13	M	Sport	12	yes	yes
8	118	19	F	FPS	8	no	no
9	119	17	M	Sport	6	no	no
10	120	18	F	RTS	4	no	no

데이터 다루기

✓ 데이터 결합 : `inner_join(data1, data2, by="ID")`

✓ (dplyr)의 join 함수

함수명	내용
<code>inner_join()</code>	data1, data2의 겹치는 ID만 결합 , 나머지 삭제
<code>left_join()</code>	data1의 ID 기준, data2에 겹치는 ID 결합 , data2에 없는 경우 <NA> 적힘
<code>right_join()</code>	data2의 ID 기준, data1에 겹치는 ID 결합 , data1에 없는 경우 <NA> 적힘
<code>full_join()</code>	data1, data2 전체 결합

데이터 다루기

✓ 데이터 결합 : `rbind(data3, data4)`

▶ data3과 data4가 동일한 변수들을 갖고 있을 때 두개 데이터를 행(row)으로 결합

dat12

	ID	age	gender	game	hourwk	alcohol	smoke
1	111	16	F	RTS	10	yes	yes
2	112	17	F	FPS	8	no	no
3	113	15	M	Sport	4	no	no
4	114	18	M	MMORPG	10	no	no
5	115	14	F	MMORPG	2	no	yes
6	116	15	F	FPS	10	yes	yes
7	117	13	M	Sport	12	yes	yes
8	118	19	F	FPS	8	no	no
9	119	17	M	Sport	6	no	no
10	120	18	F	RTS	4	no	no

data3

ID	age	gender	game	hourwk	alcohol	smoke
121	20	F	RTS	10	yes	yes
122	21	F	FPS	8	no	no
123	20	M	Sport	12	no	no

```
# add more data (combine in a row)
# dat123<-rbind(dat12, dat3)
dat3<-read.csv(file="data3.csv")
dat123<-rbind(dat12, dat3)
dat123
```

```
# using dplyr function
# dat123<-bind_rows(dat12,dat3)
```

```
> dat123
  ID age gender  game hourwk alcohol smoke
1 111 16    F   RTS     10     yes   yes
2 112 17    F   FPS      8     no    no
3 113 15    M Sport      4     no    no
4 114 18    M MMORPG    10     no    no
5 115 14    F MMORPG      2     no   yes
6 116 15    F   FPS     10    yes   yes
7 117 13    M Sport     12    yes   yes
8 118 19    F   FPS      8     no    no
9 119 17    M Sport      6     no    no
10 120 18    F   RTS      4     no    no
11 121 20    F   RTS     10    yes   yes
12 122 21    F   FPS      8     no    no
13 123 20    M Sport     12     no    no
```


데이터 다루기

☑ 데이터 정렬 : `arrange(데이터이름, 변수1, 변수2)`

➤ 변수1로 먼저 정렬을 하고 그 다음 변수2로 정렬

```
# data sorting
# dats1<-dat12[order(dat12$age),]
# dats2<-dat12[order(dat12$gender, dat12$age), ]
dats1<-arrange(dat12, age)
dats1
dats2<-arrange(dat12, gender, age)
dats2
```

연령별(age)로 정렬



>	dats1						
	ID	age	gender	game	hourwk	alcohol	smoke
7	117	13	M	Sport	12	yes	yes
5	115	14	F	MMORPG	2	no	yes
3	113	15	M	Sport	4	no	no
6	116	15	F	FPS	10	yes	yes
1	111	16	F	RTS	10	yes	yes
2	112	17	F	FPS	8	no	no
9	119	17	M	Sport	6	no	no
4	114	18	M	MMORPG	10	no	no
10	120	18	F	RTS	4	no	no
8	118	19	F	FPS	8	no	no

데이터 다루기

✓ 데이터 정렬 : `arrange(데이터이름, 변수1, 변수2)`

```
# data sorting
# dats1<-dat12[order(dat12$age),]
# dats2<-dat12[order(dat12$gender, dat12$age), ]
dats1<-arrange(dat12, age)
dats1
dats2<-arrange(dat12, gender, age)
dats2
```

성별(gender)로 정렬한 다음
연령별(age)로 정렬

> dats2


	ID	age	gender	game	hourwk	alcohol	smoke
5	115	14	F	MMORPG	2	no	yes
6	116	15	F	FPS	10	yes	yes
1	111	16	F	RTS	10	yes	yes
2	112	17	F	FPS	8	no	no
10	120	18	F	RTS	4	no	no
8	118	19	F	FPS	8	no	no
7	117	13	M	Sport	12	yes	yes
3	113	15	M	Sport	4	no	no
9	119	17	M	Sport	6	no	no
4	114	18	M	MMORPG	10	no	no

데이터 다루기

✓ 데이터 추출 - filter(데이터이름, 조건1 & 조건2)

➤ dat12에서 gender=F이고 age>15이상인 데이터를 newdat라는 이름의 데이터로 저장

```
# data subset (selecting data)
# newdat<-dat12[which(dat12$gender=="F" & dat12$age>15),]
# newdat<-subset(dat12, dat12$gender=="F" & dat12$age>15)
newdat<-filter(dat12, dat12$gender=="F" & dat12$age>15)
newdat
```



```
> newdat
```

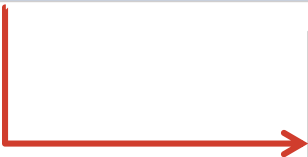
	ID	age	gender	game	hourwk	alcohol	smoke
1	111	16	F	RTS	10	yes	yes
2	112	17	F	FPS	8	no	no
8	118	19	F	FPS	8	no	no
10	120	18	F	RTS	4	no	no

데이터 다루기

☑ 데이터에서 일부변수 제거하기 : `select[데이터이름, -c("변수1", "변수2")]`

➤ `dat12`에서 `age`와 `gender`를 제외하고 `exdat`라는 이름의 데이터로 저장 (!는 not을 의미)

```
# excluding variables  
# exdat<-dat12[!names(dat12) %in% c("age", "gender")]  
exdat<-select(dat12, -c("age", "gender"))  
exdat
```



```
> exdat  
  ID game hourwk alcohol smoke  
1 111  RTS     10    yes   yes  
2 112  FPS      8     no    no  
3 113 Sport     4     no    no  
4 114 MMORPG    10     no    no  
5 115 MMORPG     2     no   yes  
6 116  FPS     10    yes   yes  
7 117 Sport    12    yes   yes  
8 118  FPS      8     no    no  
9 119 Sport     6     no    no  
10 120  RTS      4     no    no
```

데이터분석 : 데이터 사이언티스트

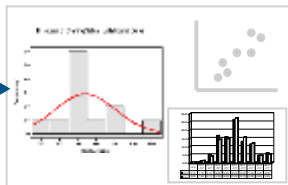
✓ 데이터 핸들링 -> 데이터 탐색 -> 통계적 모델링(통계모형, 기계학습, 인공지능)

탐색적 데이터분석

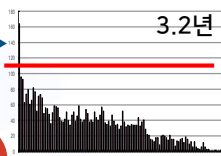
예 기술통계량 (평균, 빈도)
고객의 연령, 성별, 주거형태,
직업, 거주지

```
> describe(stud[vars])
vars  n  mean  sd  median
G1    1 395 10.91 3.32    11
G2    2 395 10.71 3.76    11
G3    3 395 10.42 4.58    11
```

예 히스토그램, 산점도, 파레토 그래프
연령대별, 제품가격대별,
구매수단별, 서비스, RFM



예 구매주기 - 제품교체주기 파
2회 이상구매자들의재구매시점을
계산 히스토그램 및 평균으로 분석



where
we are!!

통계적 분석기법

예 상관분석
(X,Y 모두 continuous variable)

일반적으로 0.7이상
이면 높다고 보지만
절대적 기준은 없다.

예 카이제곱분석 - 범주형 변수간
상관관계 (X,Y 모두 범주형 변수)
유의수준 0.1, 0.05에서 판단

```
R: Regression analysis: Tobit
> al <- asymlval ~ drug + age
> posthoc <- TobitModelTest("drug", model=al, 0.05)
> posthoc
Tobit multiple comparisons of means
    858 family-wise confidence interval

Fit: asymlformula = asyml ~ drug + age

            (Intercept)      drug      age      p-val
log-likelihood    -3.550000    -9.843248    -1.155712    0.0041976
p-value=1000    -3.333333    -9.074622    -4.990085    0.0000029
p-value=500     -3.333333    -9.174823    -3.490085    0.0000050
```

예 분산분석(ANOVA)
매장평수별 판매금액, 횟수의
차이, 그룹간 유의한 차이는
0.05, 0.1에서 결정

구매 중요 요인 도출 (마케팅)
불량 요인 도출 (제조업)
위험요인 도출 (금융업)