빅데이터 처리 (화요일 (1:3교시))

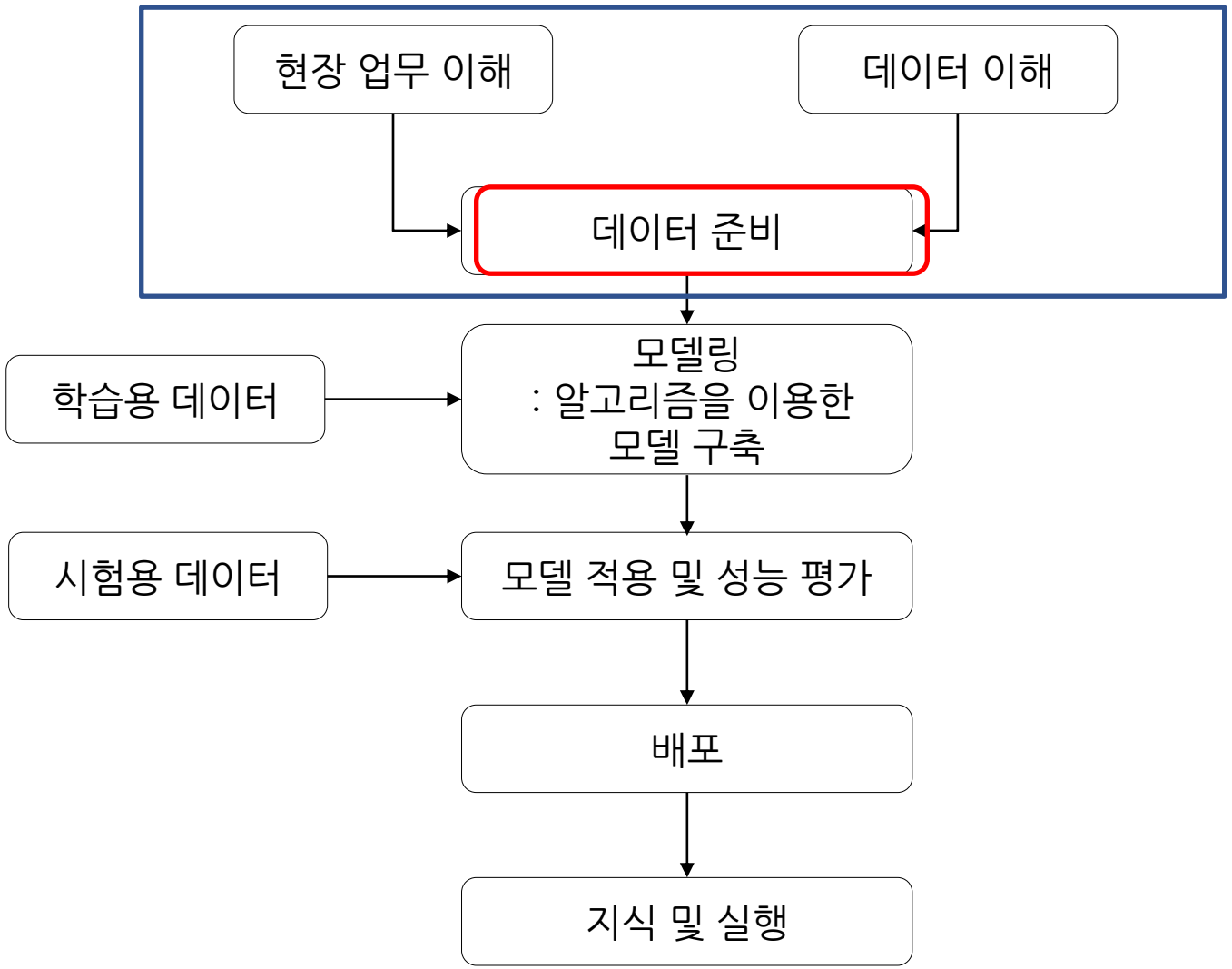
4주차 강의

데이터 탐색

2022.03.29

Institutor: JS LEE

분석 프로세스



* 전체 프로세스에서 시간이 가장 많이 소요되는 부분은?

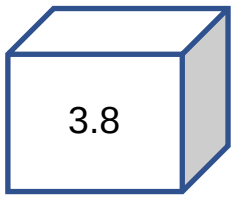
R 기본 – 데이터 구조

- 변수 (Variable)

학번	대학교	단과대학	결석	평점	등급
111111	삼육대학교	인문사회대학	0	4.3	A+
111112	삼육대학교	미래융합대학	0	3.4	B
111113	삼육대학교	인문사회대학	1	2.3	C
111114	삼육대학교	미래융합대학	2	4.5	A+
111115	삼육대학교	인문사회대학	1	3.7	B+
111116	삼육대학교	과학기술대학	3	2.4	C

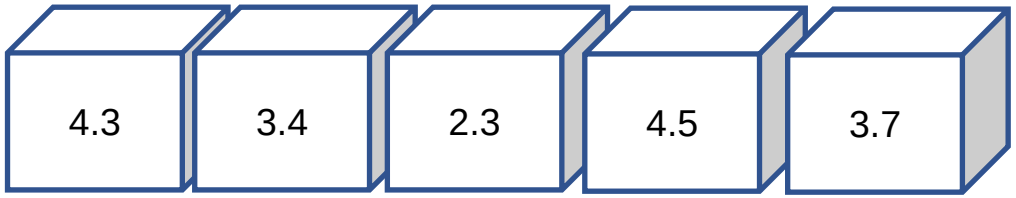
- 학점이라는 변수에 3.8이라는 데이터(값)이 있음

  “학점” 이라는 상자에 4.3 이라는 값을 넣음



- 벡터(Vector)

“학점” 이라는 상자에 여러 개의 값이 들어가 있음



- 같은 데이터 타입이어야 함
  (수치이면 모두 수치. 문자이면 모두 문자)
- R의 가장 기본적인 데이터 구조
- 일반적인 1차배열의 형태임
- 벡터는 숫자, 문자열, 논리(T/F), 팩터(범주형 데이터)의 데이터 타입을 다룰 수 있음
- 변수명, 벡터명 등은 영어로 명명할 것을 권장함

R 기본 – 데이터 구조

- Data Frame

  " 여러 데이터 유형을 한 번에 "

  분석 시 가장 많이 사용

Data Frame

학번	대학교	단과대학	결석	평점	등급
111111	삼육대학교	인문사회대학	0	4.3	A+
111112	삼육대학교	미래융합대학	0	3.4	B
111113	삼육대학교	인문사회대학	1	2.3	C
111114	삼육대학교	미래융합대학	2	4.5	A+
111115	삼육대학교	인문사회대학	1	3.7	B+
111116	삼육대학교	과학기술대학	3	2.4	C

각각의 벡터

알아둡시다

-Tip

- R 객체들은 모두 메모리 상에서 처리 되기에 갑자기 R를 다운되면 객체들이 사라지게 됨
- R을 저장하고 불러와서 사용하는 명령어를 알아둡시다.

기능	함수사용법
객체목록조회	ls()
객체 삭제	rm(객체, 객체. …)
특정 객체를 파일로 저장	save(객체,객체,…, file = " 파일명.rdata”)
모든 객체를 파일로 저장	save.image(“파일명.rdata”)
파일로부터 객체 불러오기	load(" 파일명.rdata”)

R 데이터 분석을 위한 기본

데이터 (Data)

- Data 구조
  (variable, Vector, Data Frame ..)
- Data 값의 유형
  (수치형, 명목형,..)

함수 (Function)

- 함수 : 데이터 값을 미리 정해 둔 공식에 따라 특정한 결과로 도출해 주는 기능
- 예> print(), min(), mean(),…..

패키지 (Package)

- 유사 기능의 함수들의 모아 둔 것
- 예> readxl, dplyr, ggplot….

패키지 설치:
Install.packages("패키지명")
패키지 로드하기
library(패키지명)

패키지명	용도
readxl	데이터 불러오기(엑셀)
psych	기술통계량
descr	빈도분석
dplyr	데이터 핸들링
ggplot	다양한 그래프(향후)

데이터 불러오기

데이터 불러오기 분석할 데이터 셋을 R 분석환경으로 가져오기(Loading)

- Excel 데이터 불러오기 위한 패키지 : readxl
- Install.packages(readxl)
- library(readxl)

1) 현재 지정되어 있는 디렉토리 확인

- 현재 디렉토리 위치는 **getwd()** 라는 함수를 이용
- 현재 지정되어 있는 디렉토리에 데이터 셋이 존재하는 경우 디렉토리 변경 하지 않음

2) 분석할 데이터 셋이 존재하는 디렉토리 지정

- setwd("디렉토리명") 단, /로 구분에 주의
- 예> setwd("D:/SW_DATA)

3) Excel 형식의 데이터 불러오기 (데이터 셋 :sample1.xlsx)

sample1 <- read_excel("sample1.xlsx")

<Tip>

- 바로 데이터가 존재하는 디렉토리를 read-excel에서 정의해도 됨
: sample1 <- read_excel("D:/SW_DATA/ sample1.xlsx")
- 엑셀 Sheet 가 여러 개 인 경우 sheet옵션에 해당되는 위치 입력
read_excel(" sample1.xlsx", sheet = 2) => 두번째 시트 가져오기

4) 데이터 확인하기

- [Environment] - [Data] 창에 sample1 존재여부 확인
- View(sample1) : 데이터셋 확인

ID	SEX	AGE	AREA	AMT17	Y17_CNT	AMT16	Y16_CNT
1	F	50	서울	1300000	50	100000	40
2	M	40	경기	450000	25	700000	30
3	F	28	제주	275000	10	50000	5
4	M	50	서울	400000	8	125000	3
5	M	27	서울	845000	30	760000	28
6	F	23	서울	42900	1	300000	6
7	F	56	경기	150000	2	130000	2
8	F	47	서울	570000	10	400000	7
9	M	20	인천	930000	4	250000	2
10	F	38	경기	520000	17	550000	16

데이터 불러오기

데이터 셋 파악하기

데이터 파악할 때 사용하는 함수

함수	기능
head()	데이터 앞부분 출력(변수가 많을 경우, 일부분 출력)
tail()	데이터 뒷부분 출력(변수가 많을 경우, 일부분 출력)
View()	뷰어 창에서 데이터 확인 (V : 대문자 주의)
dim()	데이터 차원 출력
str()	데이터 속성 출력

dim(sample1)

[1] 10 8

str(sample1)

```
$ ID     : num [1:10] 1 2 3 4 5 6 7 8 9 10
$ Gender : chr [1:10] "F" "M" "F" "M" ...
$ AGE    : num [1:10] 50 40 28 50 27 23 56 47 20 38
$ AREA   : chr [1:10] "서울" "경기" "제주" "서울" ...
$ AMT17  : num [1:10] 1300000 450000 275000 400000 845000 42900 150000 570000 930000 520000
$ Y17_CNT: num [1:10] 50 25 10 8 30 1 2 10 4 17
$ AMT16  : num [1:10] 100000 700000 50000 125000 760000 300000 130000 400000 250000 550000
$ Y16_CNT: num [1:10] 40 30 5 3 28 6 2 7 2 16
```

head(sample1)

	ID	Gender	AGE	AREA	AMT17	Y17_CNT	AMT16	Y16_CNT
	<dbl>	<chr>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	F	50	서울	1300000	50	100000	40
2	2	M	40	경기	450000	25	700000	30
3	3	F	28	제주	275000	10	50000	5
4	4	M	50	서울	400000	8	125000	3
5	5	M	27	서울	845000	30	760000	28
6	6	F	23	서울	42900	1	300000	6

tail(sample1)

	ID	Gender	AGE	AREA	AMT17	Y17_CNT	AMT16	Y16_CNT
	<dbl>	<chr>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	5	M	27	서울	845000	30	760000	28
2	6	F	23	서울	42900	1	300000	6
3	7	F	56	경기	150000	2	130000	2
4	8	F	47	서울	570000	10	400000	7
5	9	M	20	인천	930000	4	250000	2
6	10	F	38	경기	520000	17	550000	16

View(sample1)

ID	SEX	AGE	AREA	AMT17	Y17_CNT	AMT16	Y16_CNT
1	F	50	서울	1300000	50	100000	40
2	M	40	경기	450000	25	700000	30
3	F	28	제주	275000	10	50000	5
4	M	50	서울	400000	8	125000	3
5	M	27	서울	845000	30	760000	28
6	F	23	서울	42900	1	300000	6
7	F	56	경기	150000	2	130000	2
8	F	47	서울	570000	10	400000	7
9	M	20	인천	930000	4	250000	2
10	F	38	경기	520000	17	550000	16

데이터 탐색

- 변수(속성)를 대표하는 값은 ?
- 데이터 포인트가 대표 값과 차이는?
- 데이터셋에 특이 값이 존재하는가?
- 데이터 셋에 결츨치는 존재하는가?
- 속성 간의 상호작용을 파악 (예> 상관도가 높은가?)
- …..

1) 데이터 준비

데이터 마이닝 알고리즘을 적용하기 전에

- 유의미한 데이터 선별하고
- 이상 데이터(특이값/결측치 등)를 찾아내고,
- 중복되거나 높은 상관관계가 있는 속성들을 제거

(입력 속성들 간의 상관관계가 높을 경우 성능이 좋지 않을 수 있음)

2) 규칙 파악

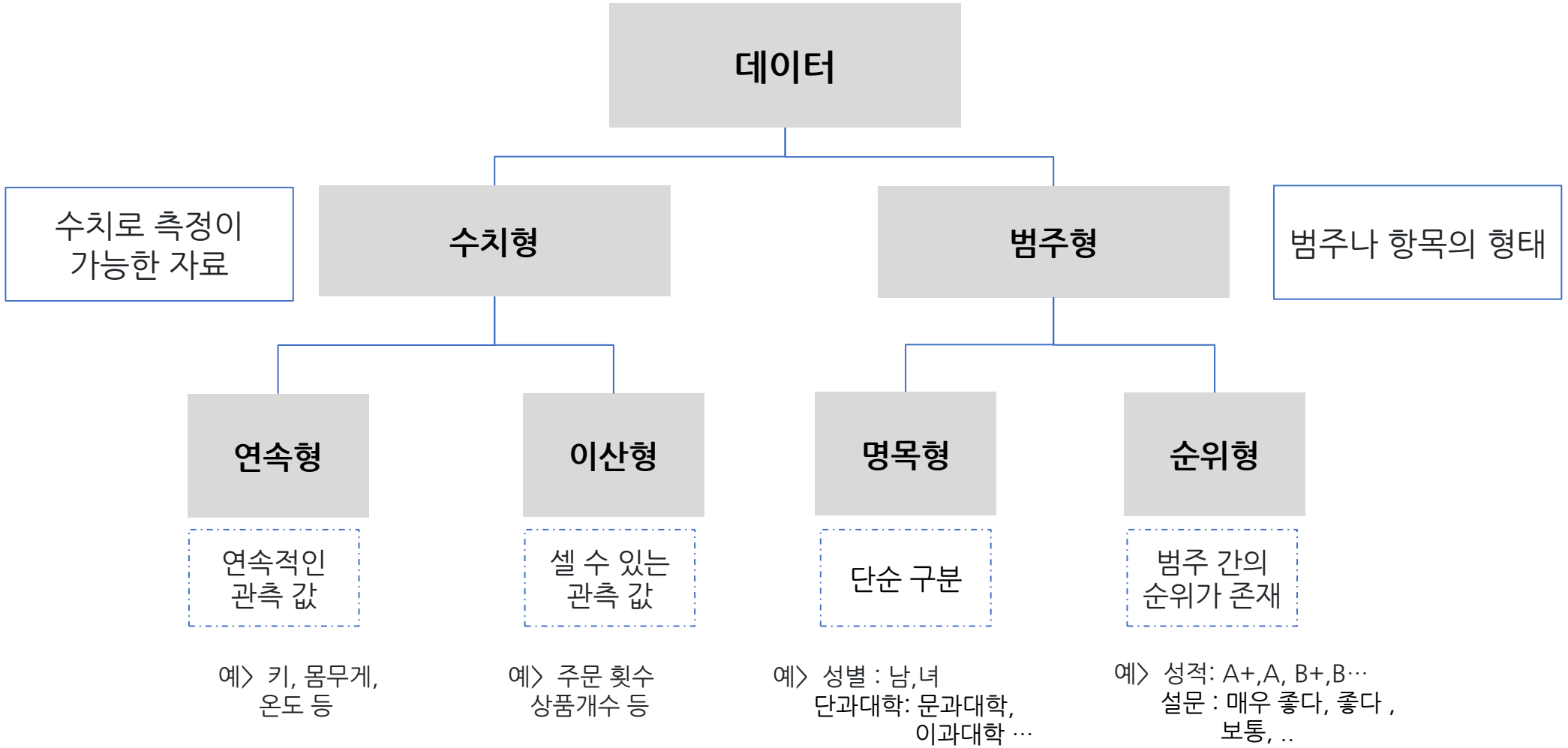
- 데이터 탐색 그 자체로 데이터 셋의 변수 간의 유의미한 규칙/패턴 등 파악

3) 결과 해석

- 데이터 마이닝 과정의 예측, 분류, 군집화의 결과를 해석

데이터 탐색

데이터 값의 형태

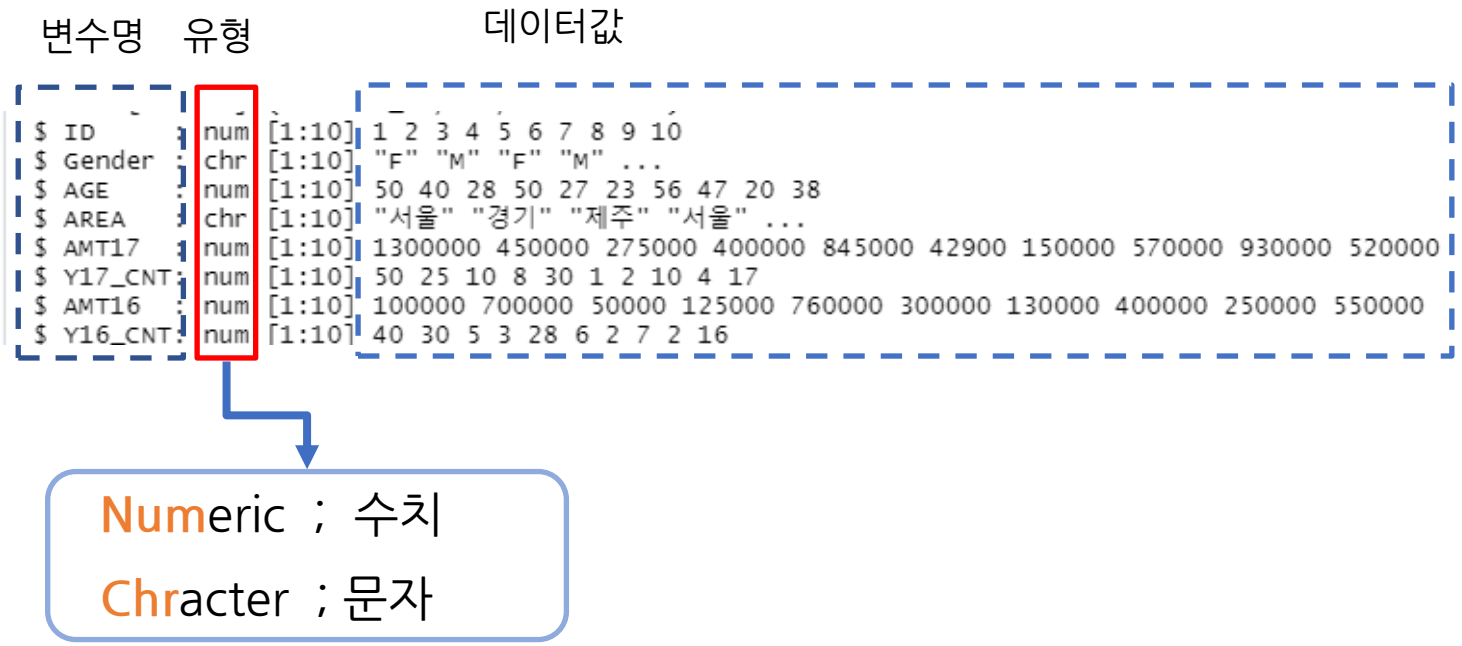


데이터 탐색

데이터 셋 살펴보기

변수명	변수 설명
ID	id
Gender	성별
AGE	연령
AREA	지역
AMT17	2017년 카드사용액
Y17_CNT	2017년 카드사용건수
AMT16	2016년 카드사용액
Y16_CNT	2016년 카드사용건수

데이터 출처 : 처음시작하는 R 데이터 분석



짚고 갑시다. ID 는 식별자이므로, 데이터 값은 수치형이나, character로 변경하기

```
sample$ID <- as.character(sample1$ID)
```

Gender와 AREA 는 범주형이므로,character에서 범주형(factor)로 변경하기

```
sample$Gender <-  as.factor(sample$Gender)
sample$AREA <-  as.factor(sample$AREA)
```

데이터 유형 바꾸기
as.유형()

* 변경 후, str(sample1)로 확인

데이터 탐색

변수유형	탐색 방법
수치형 (연봉,연령, 매출액 ..)	통계 : 기술통계 (대표값, 산포도, 최대,최소..) Pearson 상관분석 시각화 :Scatter Plot(산점도), Histogram(히스토그램), Boxplot…
범주형 (성별, 학력, 연령대 ..)	통계: 빈도분석, Spearman 상관분석 (Ordinal: 순위형) 시각화 :Bar chart(막대그래프), Pie chart(원 그래프, 파이차트), 면적 그래프 ……

데이터 탐색

평균키 180cm 병사들이 평균수심 150cm 강에 빠져죽은 건...

다. 여기서 말하는 우화는 다음과 같다. "100명의 군인들이 강을 건넌다. 군인들의 평균 키는 180㎝, 강의 평균 깊이는 150㎝다. 보고를 받은 장군이 도강을 명령했다. 강 언저리를 지나면서 물이 갑자기 깊어졌고 병사들이 빠져죽기 시작했다. 겁이 난 병사들이 뒤를 흘깃흘깃 쳐다봤지만 장군은 '돌격 앞으로'만 외쳤다. 물에 빠져죽는 병사가 속출하자 장군은 당황하며 그제야 회군을 명령했다. 하지만 이미 많은 군사를 잃은 뒤였다. 알고 보니 이 강의 최대 수심은 200㎝였고 군사 중 200㎝가 넘는 사람은 30명이 채 안 됐다." 연말정산 사태는 연봉 5500만 원 이상부터 세금이 오르는 것으로

예를 들어보자. 50가구가 사는 어느 작은 산골마을의 사례다. 이 마을의 이장은 "우리 마을의 가구당 평균 소득은 500만 원으로 매우 가난하다"고 주장하고 복덕방 영감은 "우리 마을의 가구당 평균 소득은 1억여 원으로 부자다"라고 반박한다. 사실을 알고 보니 50가구 중에 25가구는 가난한 농가로 연 소득이 500만 원에 불과하다. 다른 24가구는 500만 원에서 2000만 원 사이의 소득을 올리고 있다. 나머지 한 가구는 서울의 한 사업가가 물 좋고 공기 좋은 곳에 내려와 사는 집으로 이 가구의 연 소득은 50억 원에 달한다. 저소득농민을 위한 각종 정부지원을 기대하는 마을 이장은 최빈수를 사용해 연소득 평균이 500만 원밖에 안 되는 마을이라고 주장한다. 반면 복덕방 영감은 은퇴 후 시골에서 살려는 사람들을 유인하기 위해 산술평균을 사용해 평균 소득이 1억여 원인 부자마을 휴양지라고 선전한다.

Ref :http://kid.chosun.com/site/data/html_dir/2017/09/13/2017091302025.html

데이터 탐색

1) 수치형 변수 - 기술 통계량

- 데이터셋의 주요 특성을 수량화 하기 위해 평균, 표준편차, 분포 등과 같이 요약하는 통계적 방법
- 기술적 척도들은 데이터셋에 대한 이해에 도움을 줌
- (예) 연평균 수입, 주택가격 중앙값, 신용점수 범위 등

R 과 기술 통계량

- summary()와 describe() 함수를 이용하면, 기술 통계량을 한번에 확인 가능함
- describe()함수는 psych 패키지에 내장되어 있으므로, 먼저 psych 패키지 인스톨 후, 로딩해야 함
- describe() 함수가 더 다양한 기술 통계량을 포함하고 있음
- 수치형 변수만을 선택하여 위 두 함수를 이용하는 것을 권장함

데이터 탐색

2) 수치형 변수 - 기술 통계량

데이터셋의 특성	척도
데이터셋의 중심	평균, 중앙값, 최빈값
데이터셋의 산포도	범위, 분산, 표준편차
데이터셋의 분포 모양	왜도, 첨도

R함수 : summary()

출력 값	통계량	설명
min	최소값	가장 작은 값
Ist Qu.	1사분위수	25% 위치에 해당하는 값
Median	중앙값	중앙(50% 위치)에 해당하는 값
Mean	평균값	모든 값을 합한 후 데이터 값의 개수로 나눈 값
3rd Qu.	3사분위수	75% 위치에 해당하는 값
Max	최대값	가장 큰 값

R함수 : describe ()

출력 값	설명
n	관측값 개수
mean	평균
sd	표준편차 : 분산의 제곱근 (평균을 중심으로 얼마나 퍼져 있는지)
median	중앙값
trimmed	10% 절사평균
mad	중앙값 절대 편차
min	최소
max	최대
range	범위 (최대-최소)
skew	왜도 (대칭정도) 0 : 대칭 + : 오른쪽으로 꼬리가 김, -: 왼쪽으로 꼬리가 김
kurtosis	첨도 (뾰쪽함 정도) 3; 정규분포 3 < : outlier가 많음
se	표준오차

데이터 탐색

2) 수치형 변수 - 기술 통계량

0,0,14,14	0,6,8,14	6,6,8,8

평균은? 7

표준 편차는? 7 5 1

데이터 탐색

보고서. 1번 문제

1) 수치형 변수 - 기술 통계량

summary(sample1)

```
       ID                Gender      AGE              AREA       AMT17              Y17_CNT          AMT16              Y16_CNT
 Length:10          F:6    Min.   :20.00    경기:3   Min.   :  42900   Min.   : 1.0   Min.   : 50000   Min.   : 2.0
 Class :character   M:4    1st Qu.:27.25    서울:5   1st Qu.: 306250   1st Qu.: 5.0   1st Qu.:126250   1st Qu.: 3.5
 Mode  :character          Median :39.00    인천:1   Median : 485000   Median :10.0   Median :275000   Median : 6.5
                           Mean   :37.90    제주:1   Mean   : 548290   Mean   :15.7   Mean   :336500   Mean   :13.9
                           3rd Qu.:49.25             3rd Qu.: 776250   3rd Qu.:23.0   3rd Qu.:512500   3rd Qu.:25.0
                           Max.   :56.00             Max.   :1300000   Max.   :50.0   Max.   :760000   Max.   :40.0
```

- character 유형은 결과 값 도출 안됨
- Factor 유형은 범주별 빈도수 도출

데이터 탐색

보고서. 1번 문제

1) 수치형 변수 - 기술 통계량

sample1의 데이터 셋에서 3번째 변수, 5번째에서 8번째 변수만을 추출하여 함수에 적용한다.

```
describe(sample1[,c(3,5:8)])
```

```
        vars  n      mean        sd   median   trimmed       mad   min      max   range  skew kurtosis        se
AGE        1 10      37.9     12.77     39.0     37.88     16.31    20       56      36 -0.05    -1.75      4.04
AMT17      2 10 548290.0 383039.19 485000.0 517500.00 404008.50 42900 1300000 1257100  0.52    -0.93 121127.63
Y17_CNT    3 10      15.7     15.40     10.0     13.25     11.12     1       50      49  0.98    -0.26      4.87
AMT16      4 10 336500.0 257186.16 275000.0 319375.00 240922.50 50000  760000  710000  0.46    -1.50  81329.41
Y16_CNT    5 10      13.9     13.88      6.5     12.12      6.67     2       40      38  0.69    -1.28      4.39
```

- 2017년도의 카드사용액과 사용건수가 2016년도 보다 활발함

- 결측치 확인 방법

```
colSums(is.na(sample1))
```

각 변수의 결측치 개수

```
  ID Gender      AGE     AREA    AMT17 Y17_CNT    AMT16 Y16_CNT
   0      0        0        0        0       0        0       0
```

데이터 탐색

2) 수치형 변수 - 그래프(시각화)

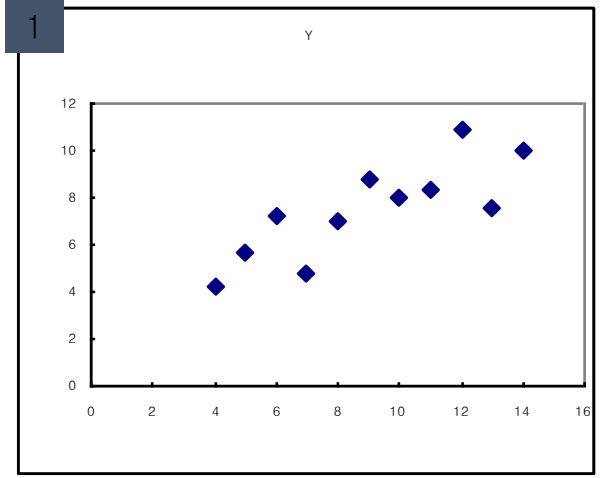
1		2		3		4	
X	Y	X	Y	X	Y	X	Y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	9.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

N=11
X 평균 = 9.0
Y평균=7.5
절편=3
기울기=0.5

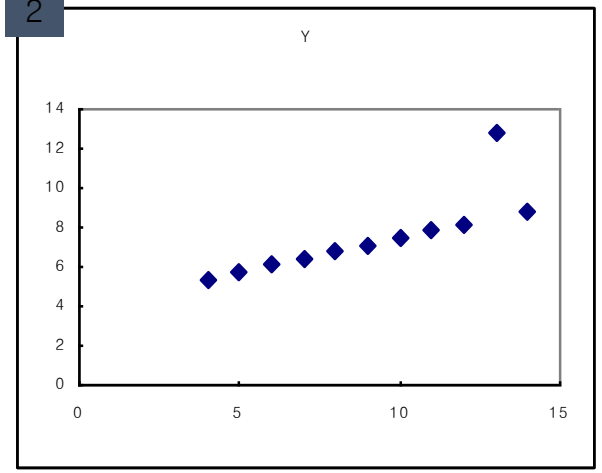
데이터 탐색

2) 수치형 변수 - 그래프(시각화)

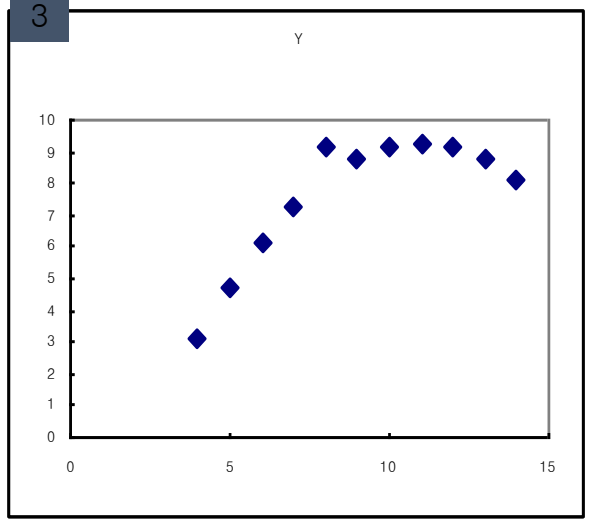
1



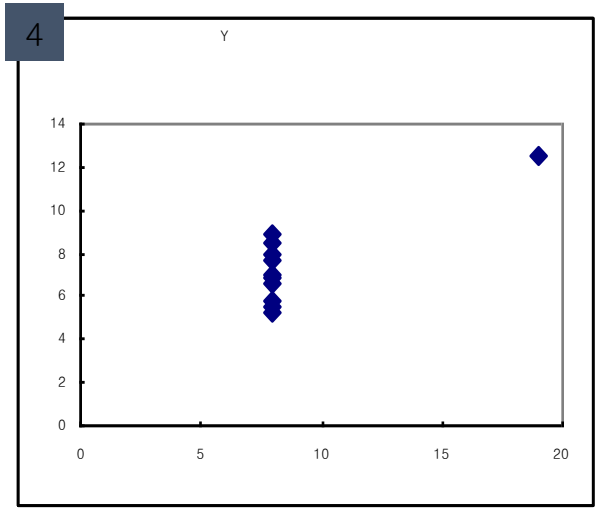
2



3



4



데이터 탐색

2) 수치형 변수 - 그래프(시각화)

데이터 시각화에 대한 필요성

1. 복잡한 정보의 이해
   - 간단한 시각화 차트라도 수천 개의 데이터 포인트들을 쉽게 포함
   - 시각적으로 살펴보면 패턴 뿐만 아니라 장기적인 추세도 파악 가능
   - ☞ 데이터를 단순히 숫자로 표현할 때에는 파악하기 곤란

2. 관계 파악
   - 데이터를 직교 좌표에 나타내면 변수들 간의 관계 파악 가능
   - 3차원을 초과하는 데이터를 시각적으로 표현하는 것은 불가능하지만 데이터 포인트의 크기, 색, 모양, 흐름도 등을 이용하면 가능

☞ 2차원 공간에서 각 데이터 포인트에 둘 이상의 속성을 사용

- R 을 활용한 그래프 생성 시 참고할 사이트
  https://r-charts.com/

데이터 탐색

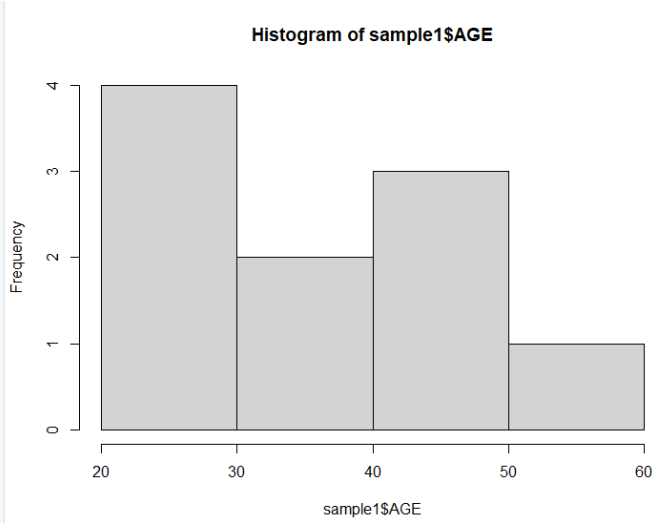
2) 수치형 변수 - 그래프 (시각화)

1) 히스토그램 (Histogram)

- 데이터 도수의 분포를 시각화
- 속성값들이 어떻게 분포를 하며 분포의 모양은 어떠한 가를 보여줌
- 값들의 발생 빈도를 이해하는 가장 기본적인 시각화 방법
- 가로축: 변수, 세로축: 발생 빈도
- 연속적인 수치형 변수에 대해서 일정구간의 값들을 모아서 그 구간의 범위나 구간 대푯값을 명시

(예) 키: 152.00과 152.99 사이에 있는 모든 값들을 152라는 이름으로 그룹화

- 대략적인 규칙: 구간의 개수 = 데이터 수의 제곱근이나 세제곱근



데이터 탐색

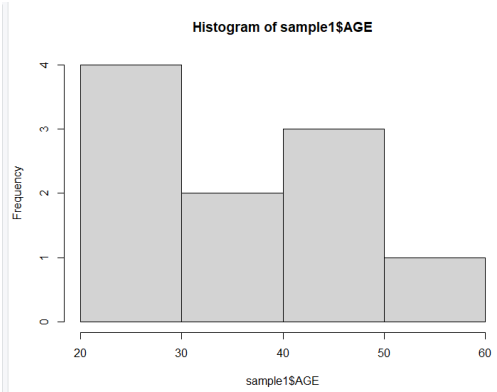
2) 수치형 변수 - 그래프 (시각화)

R 과 Histogram

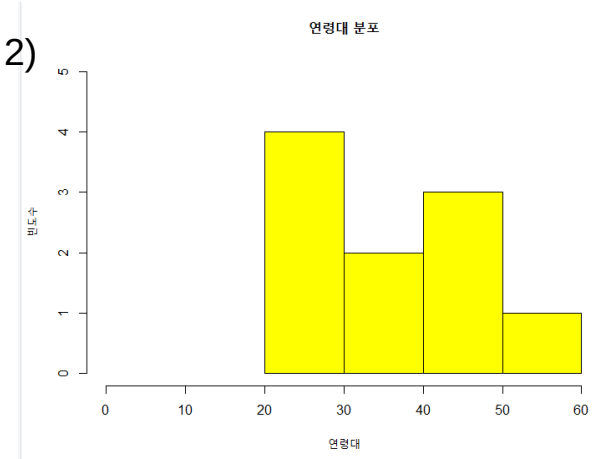
1) 기본: hist(sample1$AGE)
2) 옵션

```
hist(sample1$AGE,
     xlim=c(0,60) ,   # x축
     ylim=c(0,5),   # y축
     main = "연령대 분포",   # 히스토그램 제목
     xlab="연령대" ,              # x축명
     ylab="빈도수" ,              # y축명
     col ="Yellow"   # 색지정
     breaks =8   # 구간개수)
```

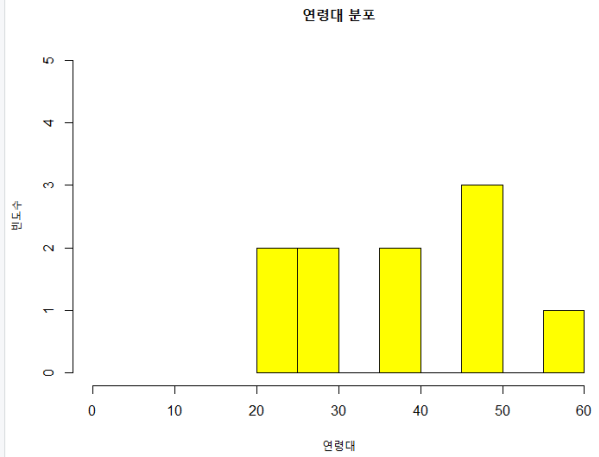
1)



2)



- 20-30대의 비중이 가장 높으며,
- 50-60대의 비중이 가장 낮음



- 46-50세의 비중이 가장 높으며,
- 20-30대, 35-40세의 비중은 동일함
- 41-45세, 31-35세, 51-55세는 카드 사용 안 함

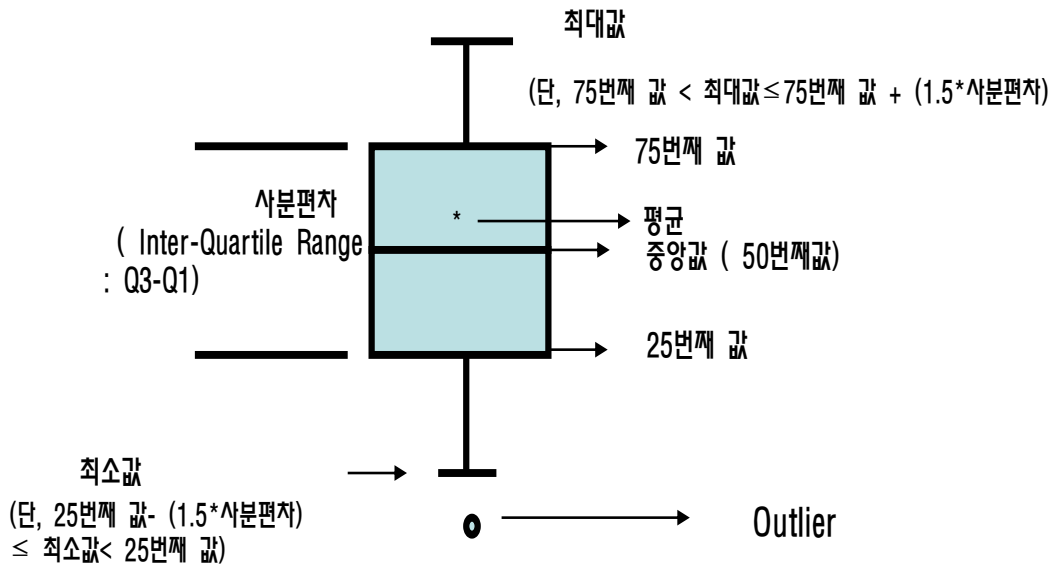
데이터 탐색

2) 수치형 변수 - 그래프 (시각화)

2) 상자그림 (Boxplot) - 계속

- 4분위수와 이상치(Outlier)를 시각화 하여, 데이터의 중심위치와 분포를 파악하는데 유용한 그래프

  * 4분위수 (25%,50%,75%,100%)의 위치에 해당하는 값



데이터 탐색

보고서. 2번 문제

2) 수치형 변수 - 그래프 (시각화)

2) 상자그림 (Boxplot) - 계속

R 과 BoxPlot

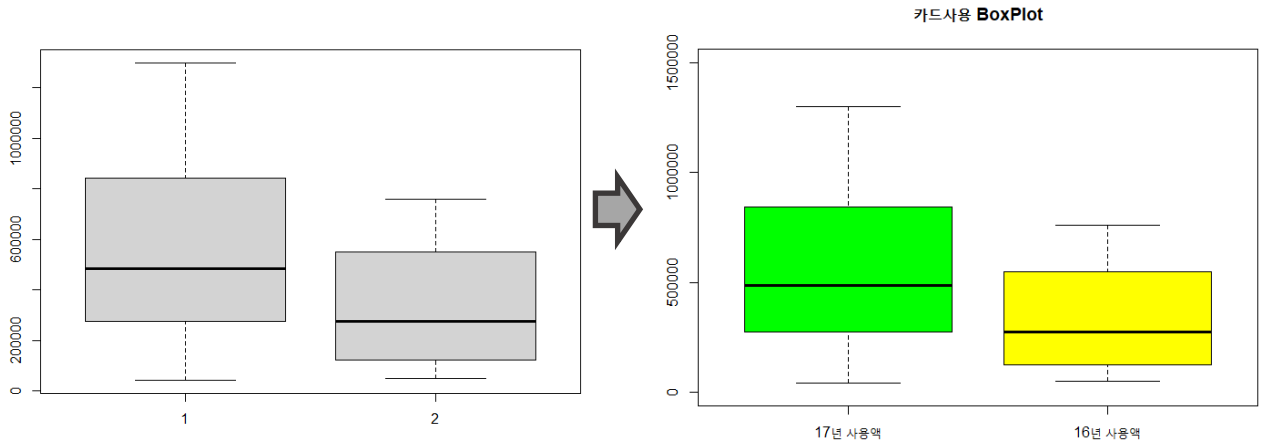
1. 사분위수 : quantile(sample1$AMT17)

2. 박스플롯 기본 :
boxplot (sample1$AMT17,sample1$AMT16)

3. 박스플롯 옵션 :
boxplot(sample1$AMT17, sample1$AMT16,
ylim =c(0,1500000), # y축조정
main = “카드사용 BoxPlot”, # “제목 "
names =c(“17년 사용액“,”16년 사용액”), # 박스플랏 명“
col = c(“green”,”yellow”)) # 박스프랏 색상

AMT17, AMT16의 4분위수와 박스플롯을 그려서 , 2017년 카드사용금액과 2016년 카드사용금액의 분포정도를 비교해보자

```
quantile(sample1$AMT17)
   0%     25%     50%     75%    100%
42900  306250  485000  776250 1300000
quantile(sample1$AMT16)
   0%     25%     50%     75%    100%
50000 126250 275000 512500 760000
```

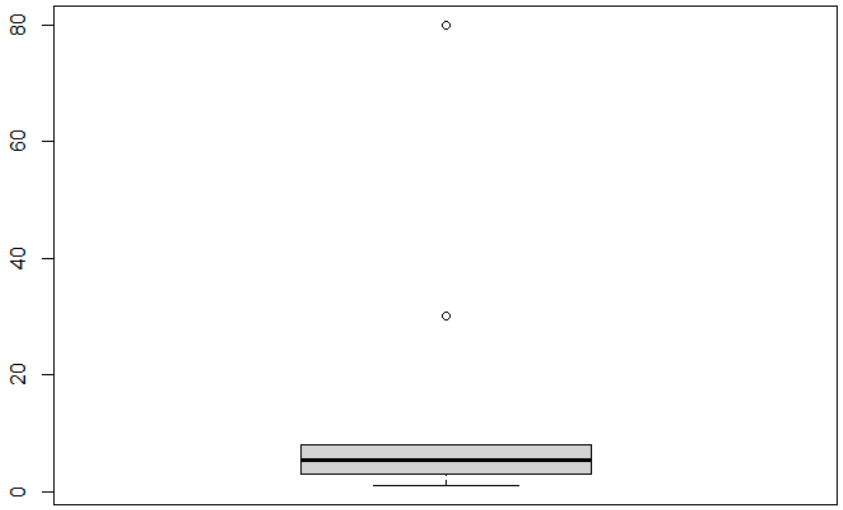


2017년이 2016년보다 카드 사용액이 높음을 알 수 있음

데이터 탐색

2) 수치형 변수 - 그래프 (시각화)

2) 상자그림 (Boxplot)



(1,2,3,4,5,6,7,8, 80,30) 일 때 , Boxplot

: 30,80을 outlier로 동그라미(○)로 표현 됨

데이터 탐색

3) 수치형 변수 (그룹별)- 기술 통계량

범주형 변수 (예> 성별, 지역)별로, 수치형변수의 기술통계량

R 과 그룹별 기술 통계량

describeBy()함수 이용

descrbeBy(수치형 변수들, group=범주형 변수)

=>describeBy(데이터셋명[,c(위치1,위치2)], group=범주형변수)

describeBy(수치형변수, group=범주형 변수)

예> 성별로 2016년 카드사용액

=> describeBy(sample1$AMT16, group=sample1$Gender)

```
 Descriptive statistics by group
group: F
   vars n   mean        sd median trimmed    mad   min    max  range skew kurtosis      se
X1    1 6 255000 195831.56 215000  255000 207564 50000 550000 500000 0.32     -1.8 79947.9
------------------------------------------------------------------------------------------------------------
group: M
   vars n   mean        sd median trimmed    mad    min    max  range  skew kurtosis        se
X1    1 4 458750 318286.43 475000  458750 378063 125000 760000 635000 -0.04    -2.37 159143.21
> |
```

여성보다 남성이 2016년 카드 사용액이 더 많음

보고서. 3번 문제

예> 지역별로 2016년 카드사용액,사용회수,2017카드사용액, 사용회수)

=> describeBy(sample1[,c(5:8)], sample$AREA)

또는

describeBy(sample1$AMT16,sample1$Y17_CNT,

sample1$AMT17,sample1$Y16_CNT , group=sample $AREA)

데이터 탐색

3) 수치형 변수 (그룹별)- 기술 통계량

예> 지역별로 2016년 카드사용액,사용회수,2017카드사용액, 사용회수)

=> describeBy(sample1[,c(5:8)], group=sample$AREA)

```
 Descriptive statistics by group
group: 경기
        vars n      mean        sd median   trimmed       mad    min    max  range  skew kurtosis        se
AMT17      1 3 373333.33 196553.64 450000 373333.33 103782.00 150000 520000 370000 -0.33    -2.33 113480.30
Y17_CNT    2 3     14.67     11.68     17     14.67     11.86      2     25     23 -0.19    -2.33      6.74
AMT16      3 3 460000.00 295465.73 550000 460000.00 222390.00 130000 700000 570000 -0.28    -2.33 170587.22
Y16_CNT    4 3     16.00     14.00     16     16.00     20.76      2     30     28  0.00    -2.33      8.08
------------------------------------------------------------------------------------------------------------
group: 서울
        vars n     mean        sd median  trimmed       mad    min     max   range skew kurtosis        se
AMT17      1 5 631580.0 473365.33 570000 631580.0 407715.00  42900 1300000 1257100 0.17    -1.71 211695.41
Y17_CNT    2 5     19.8     20.03     10     19.8     13.34      1      50      49 0.48    -1.76      8.96
AMT16      3 5 337000.0 267104.85 300000 337000.0 259455.00 100000  760000  660000 0.56    -1.54 119452.92
Y16_CNT    4 5     16.8     16.33      7     16.8      5.93      3      40      37 0.42    -1.98      7.30
------------------------------------------------------------------------------------------------------------
group: 인천
        vars n   mean sd median trimmed mad    min    max range skew kurtosis se
AMT17      1 1 930000 NA 930000  930000   0 930000 930000     0   NA       NA NA
Y17_CNT    2 1      4 NA      4       4   0      4      4     0   NA       NA NA
AMT16      3 1 250000 NA 250000  250000   0 250000 250000     0   NA       NA NA
Y16_CNT    4 1      2 NA      2       2   0      2      2     0   NA       NA NA
------------------------------------------------------------------------------------------------------------
group: 제주
        vars n   mean sd median trimmed mad    min    max range skew kurtosis se
AMT17      1 1 275000 NA 275000  275000   0 275000 275000     0   NA       NA NA
Y17_CNT    2 1     10 NA     10      10   0     10     10     0   NA       NA NA
AMT16      3 1  50000 NA  50000   50000   0  50000  50000     0   NA       NA NA
Y16_CNT    4 1      5 NA      5       5   0      5      5     0   NA       NA NA
```

서울지역의 카드사용자가 제일 많으며,
2016년에는 경기지역의 사용자의
평균 사용금액이 서울지역보다 높았음

데이터 탐색

보고서. 4번 문제

4) 수치형 변수 (그룹별)- 그래프

1) Histogram :계속

성별에 따라 2017년 카드사용액 구간별로 빈도수를 알고 싶다.

=> 성별 구분에 히스토그램 그리기

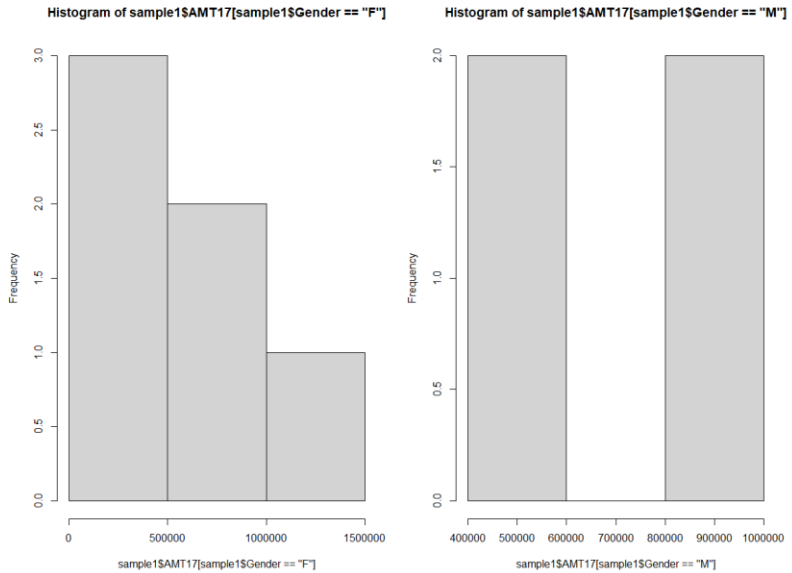
R 과 그룹별 히스토그램

- 성별 구분별로 동시에 한 창(윈도우)에서 출력되지 않음
- 한 위도우를 범주의 개수별로 구분함

  함수 : par(mfrow=c(행의 수, 열의 수))
- 각 범주별로 히스토그램 그리기

  함수 : hist(수치형변수[범주형변수 == “범주값“])

```
par(mfrow=c(1,2))
hist(sample1$AMT17[sample1$Gender=="F"])
hist(sample1$AMT17[sample1$Gender=="M"])
```



tip plot이 나온 창 위의 zoom 버튼 선택 시, 그래프 크게 볼 수 있음
Export 선택 시, 그래프 저장 할 수 있음

데이터 탐색

보고서. 4번 문제

4) 수치형 변수 (그룹별)- 그래프

1) Histogram

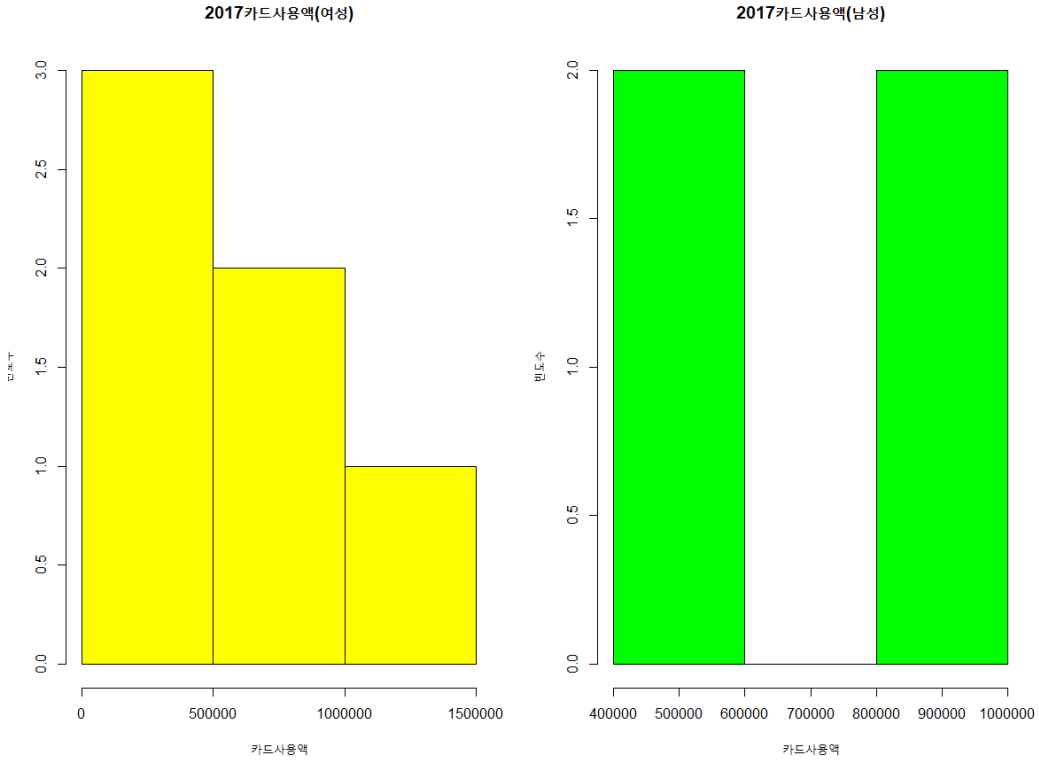
- 그래프의 가독성을 높이기 위해, histogram 옵션들을 활용
- ggplot을 이용하면 더욱 가독성 좋은 그래프 생성 가능(시각화에서 다루고자 함)

```
par(mfrow=c(1,2))

hist(sample1$AMT17[sample1$Gender=="F"],
     main = "2017카드사용액(여성)",   # 히스토그램 제목
     xlab= "카드사용액",              # x축명
     ylab= "빈도수" ,                 # y축명
     col =  "Yellow" ) # 색지정

hist(sample1$AMT17[sample1$Gender=="M"],
     main = "2017카드사용액(남성)",   # 히스토그램 제목
     xlab= "카드사용액",              # x축명
     ylab= "빈도수" ,                 # y축명
     col =  "green" ) # 색지정

 par(mfrow=c(1,1))  # 원래 하나의 창으로
```



데이터 탐색

보고서. 5번 문제

- boxplot (sample1$AMT17 ~ sample1$Gender)
- boxplot(sample1[, 5] ~ sample1[, 2])
  에러 발생 시 (요휴하지 않은 리스트…)
  sample1를 데이터프레임으로 지정한다
  sample1 <-as.data.frame(sample1)
- boxplot(AMT17 ~ Gender, data=sample1)

4) 수치형 변수 (그룹별)- 그래프

2) BoxPlot

성별에 따라 2017년 카드사용액의 데이터의 중심위치와 분포를 파악

=> 성별 구분에 boxplot 그리기

R 과 그룹별 BoxPlot

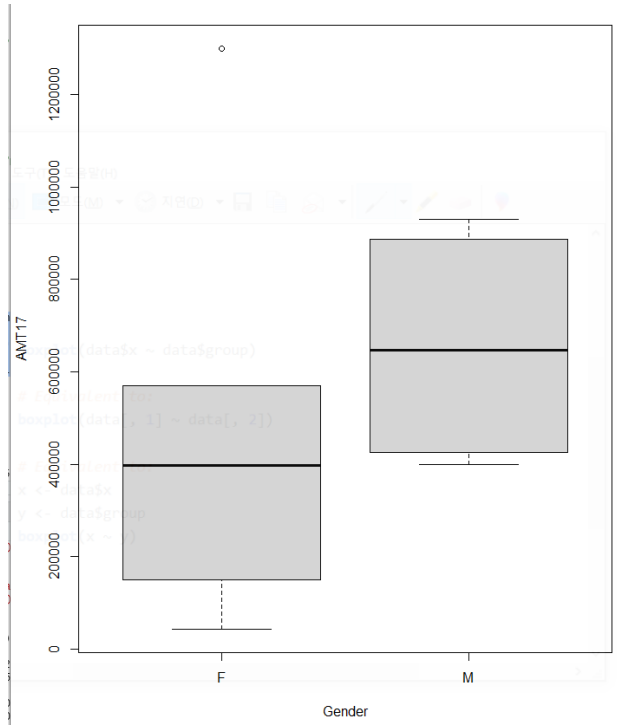
```r
boxplot(data$x ~ data$group)

# Equivalent to:
boxplot(data[, 1] ~ data[, 2])

# Equivalent to:
x <- data$x
y <- data$group
boxplot(x ~ y)
```

```r
boxplot(x ~ group, data = data)
```

refer:https://r-charts.com/distribution/box-plot-group/



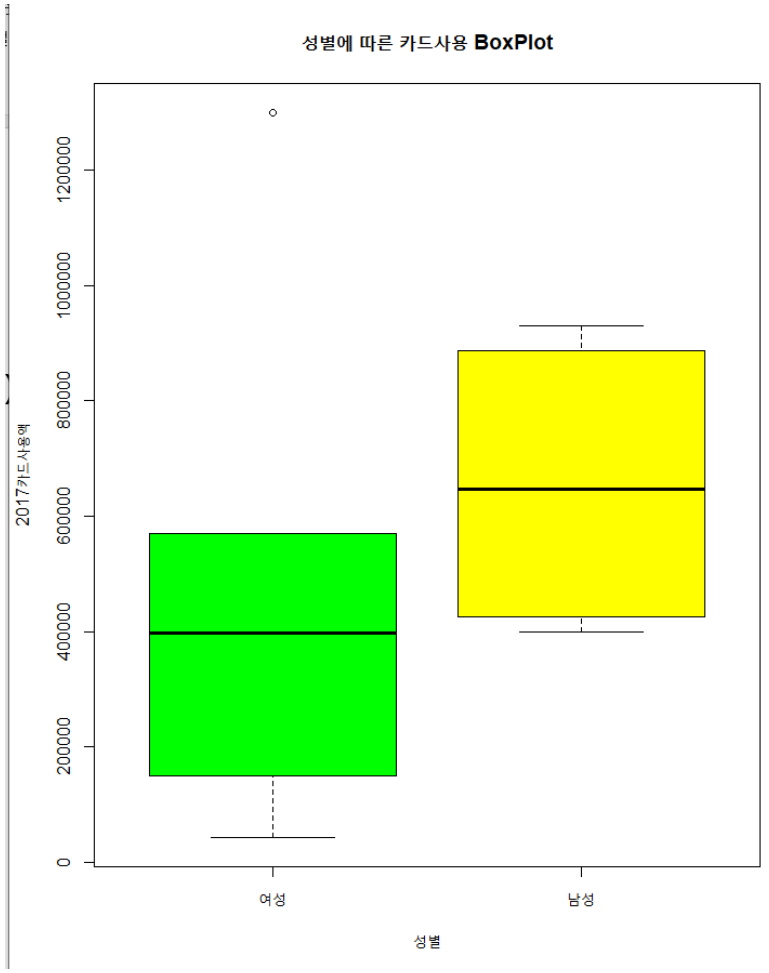
데이터 탐색

4) 수치형 변수 (그룹별)- 그래프

2) Boxplot

- 그래프의 가독성을 높이기 위해, Boxplot 옵션들을 활용

```
boxplot (sample1$AMT17 ~ sample1$Gender,
        # ylim =c(0,1500000), # y축조정
        main = "성별에 따른 카드사용 BoxPlot", # 제목
        xlab = "성별", # x축이름
        ylab = "2017카드사용액",  #y축이름
        names =c("여성","남성"), # 박스플랏 명
        col = c("green","yellow")) # 박스프랏 색상
```



데이터 탐색

5) 수치형 변수 간의 관계 파악

상관관계 분석(계속)

두 변수 간의 통계적인 관계, 특히 한 변수의 다른 변수에 대한 의존도

- 두 변수의 상관관계가 존재함은 한 변수가 변할 때 다른 변수도 같은 비율로 변함을 의미함(양 또는 음의 방향)
- -1 에서 1사이의 값을 가지며 절대값의 수가 클 수록 상관관계가 높음을 0은 상관관계가 없음을 의미함
- 음의 상관관계는 반대 방향으로 변함을 의미

  (예) 하루의 평균 기온과 아이스크림 매출이 양의 상관관계가 존재함

  - 두 변수는 서로 종속적이며, 한 변수(기온)를 이용하여 다른 변수(아이스크림 매출) 예측 가능
  - 기온이 올라갈 수록 아이스크림 매출도 올라감

- 두 변수 간의 상관관계가 인과관계를 뜻하는 것은 아님
  - 한쪽이 반드시 다른 쪽의 원인이 된다는 것은 아님

  (예) 아이스크림 매출과 상어공격 횟수는 상관관계가 높다고 하여 두 변수 사이에는 인과관계가 존재하는 것은 아님

데이터 탐색

5) 수치형 변수 간의 관계 파악

상관관계 분석(계속)

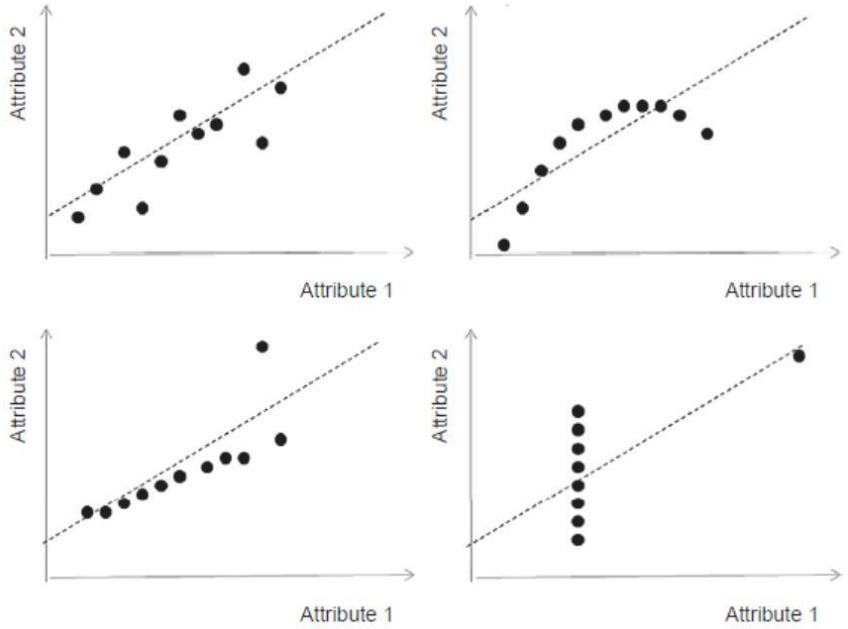
피어슨(Pearson) 상관계수(r)

- 선형 종속성의 정도를 측정 (-1에서 1사이의 값)

$$r_{XY} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N \times s_X \times s_Y}$$

- 상관계수의 한계점
  - 비선형의 관계를 나타내기 어려움
  - 특이값으로 인해 왜곡될 수 있음

  - => 산점도 그래프를 이용하여, 시각화를 통해 이러한 한계점을 보완 함
    :비선형 관계를 찾아내고 특이값을 보여줌



평균, 분산, 상관계수가 동일

데이터 탐색

보고서. 6번 문제

5) 변수간의 관계 파악 (수치형변수)

상관관계 분석

R 과 상관관계

cor() 함수 이용

cor(변수1,변수2,….,)

cor() 함수의 인자	
x	상관계수를 구할 데이터 혹은 변수1(벡터형, matrix형, 데이터 프레임)
y	상관계수를 구할 데이터 혹은 변수2(벡터형, matrix형, 데이터 프레임) 생략시에는 y =x가 되고 x에 데이터 프레임이면 지정할 필요가 없음
use	결측치 처리 방식 - "everything" : 변수에 결측값이 존재하는 경우, 상관계수를 계산하지 않고 NA 출력 - "all.obs" : 모든 관측값을 사용하려 하며, 만일 데이터에 결측값이 존재하면 오류 메시지를 출력함. - "complete.obs" : 결측치를 제거하고 상관계수를 계산함. 목록별 제거(두 변수의 상관계수를 구할 때에도 다른 3의 변수에 결측치가 존재하면 그 행은 제외하고 계산) - "pairwise" : 두 변수 중 결측치를 제거하고 상관계수를 계산함. 쌍별 제거(두 변수의 상관계수를 구할 때, 두 변수 중 결측치를 제외하고 계산)
method	상관계수 종류 ("pearson", "spearman", "kendall")

refer:
https://m.blog.naver.com/pmw9440/221530246292

- 2017,2016 사용금액과 사용건수의 상관관계를 알아보자
- cor(sample1[,c(5:8)])

```
              AMT17      Y17_CNT        AMT16      Y16_CNT
AMT17    1.0000000000 0.7548529319 0.0956722971 0.6320115076
Y17_CNT  0.7548529319 1.0000000000 0.2491221111 0.9571373893
AMT16    0.0956722971 0.2491221111 1.0000000000 0.4447047179
Y16_CNT  0.6320115076 0.9571373893 0.4447047179 1.0000000000
```

- 16,17년 카드 사용금액과는 상관도가 없음
- 그러나 16,17년 사용횟수와는 상관도가 높음
  => 16년에 카드 사용건수가 높을 수록 17년 사용건수는 높다
- 사용회수와 사용금액은 상관도가 높음

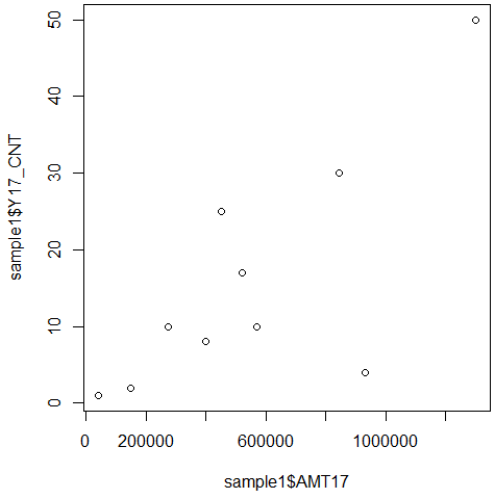
데이터 탐색

보고서. 7번 문제

5) 변수간의 관계 파악 (수치형변수)

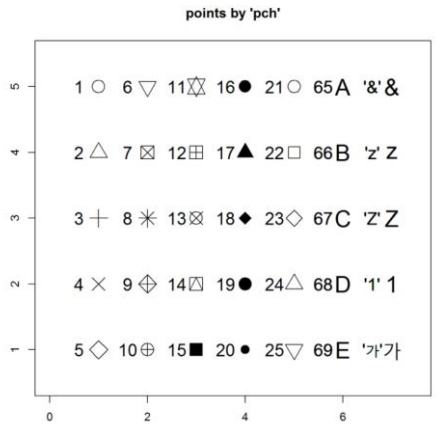
상관관계 분석 - 그래프(plot)

- 2017카드 사용건수와 사용금액의 산점도를 그려보자
- plot(sample1$AMT17,sample1$Y17_CNT)



R 과 plot

- plot() 함수 이용
- plot(변수명1, 변수명2) ; 기본
- plot(변수명1, 변수명2,
  main = “ “, : 제목
  xlim=c(,) : x축범위
  ylim=c(,) : y축범위
  xlab = “ “, : x축명
  ylab = “ ”, : y축명
  cex = , : 포인트크기
  pch= ,: 포인트 종류
  col= “” ,: 색상)



데이터 탐색

보고서. 8번 문제

5) 변수간의 관계 파악 (수치형변수)

상관관계 분석 - 그래프(plot)

- 2017,2016 카드 사용건수와 사용금액의 산점도를 그려보자
- plot(sample1[,c(5:8)])

* 파란색과 주황색영역은 동일한 관계

x축 : Y17_CNT
y축 : AMT 17

x축 : AMT 17
y축 : Y17_CNT

기울기가 1에 가까운 직선 형태 패턴

패턴 없음

	AMT17	Y17_CNT	AMT16	Y16_CNT
AMT17	1.0000000000	0.7548529319	0.0956722971	0.6320115076
Y17_CNT	0.7548529319	1.0000000000	0.2491221111	0.9571373893
AMT16	0.0956722971	0.2491221111	1.0000000000	0.4447047179
Y16_CNT	0.6320115076	0.9571373893	0.4447047179	1.0000000000

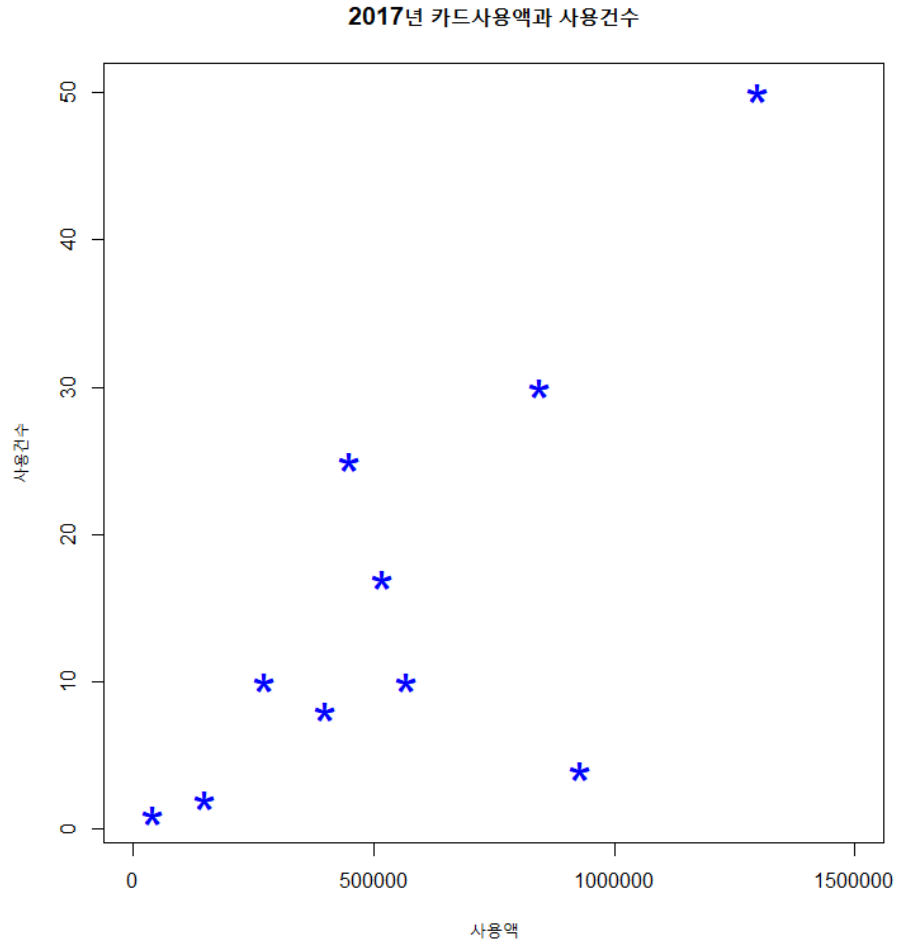
데이터 탐색

5) 변수간의 관계 파악 (수치형변수)

상관관계 분석 - 그래프(plot)

- 2017년 카드 사용건수와 사용금액의 산점도를 가독성 있게 그려봅시다

```
plot(sample1$AMT17,sample1$Y17_CNT,
    main = "2017년 카드사용액과 사용건수", # 제목
          xlab = "사용액",
          ylab = "사용건수",
          xlim= c(0,1500000),
          cex = 3,
          pch= "*",
          col= "blue")
```



데이터 탐색

6) 범주형 변수 - 빈도분석

범주형은 각 변수의 범주가 어떻게 구성되어 있는지로, 데이터의 특성을 파악
그 대표적인 방법으로 도수분포표를 이용한 빈도분석이 있음.

■ 도수분포표

- 계급, 도수 및 상대도수로 구성됨
- 계급(class) : 자료가 취하는 전체 범위를 몇 개의 소집단으로 나눈 것
- 도수(frequency) : 각 계급에 속하는 자료의 수
- 상대도수(relative frequency) : 도수를 전체 자료의 수, 즉 전체 도수로 나눈 비율

계급	도수	상대도수
남자	70	0.7
여자	30	0.3

R 과 빈도분석

- table(), freq()함수 이용
- describe()함수는 descr 패키지에 내장되어 있으므로, 먼저 descr 패키지 인스톨 후, 로딩해야 함
- describe()로 도수분포표와 막대그래프(bar chart) 생성가능
- table()함수를 이용한 결과를 barplot()함수를 이용하여 바 차트를 pie() 함수를 이용하여 파이차트(pie chart) 로 표현 가능

데이터 탐색

6) 범주형 변수 - 빈도분석

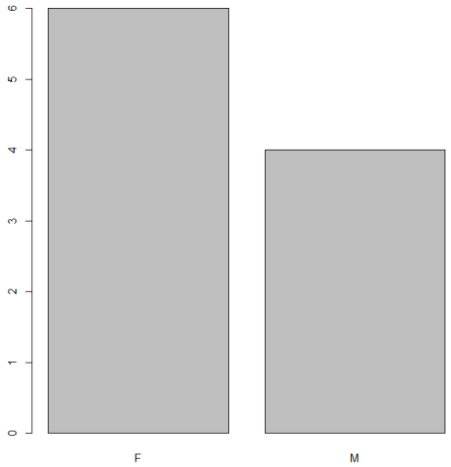
보고서. 10번 문제

- 성별과 지역의 범주별 빈도를 알아보자

- freq(sample1$Gender)
- freq(sample$AREA, plot=F) plot=F : barchart 생성 안하기

```
        Frequency Percent
F               6      60
M               4      40
Total          10     100
```

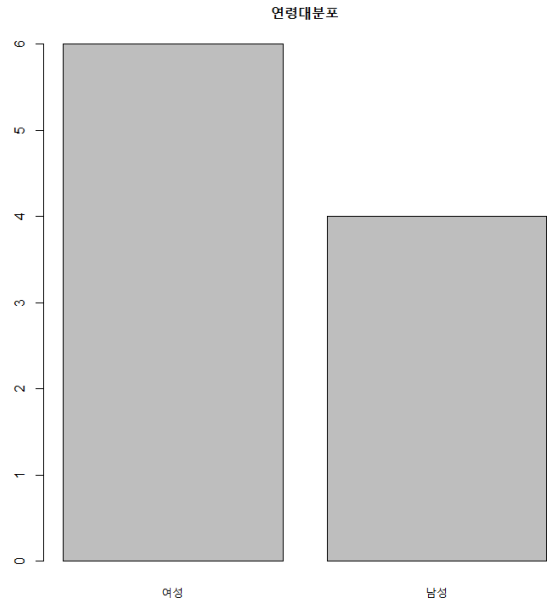
```
        Frequency Percent
경기            3      30
서울            5      50
인천            1      10
제주            1      10
Total          10     100
```



* table()함수 이용 시, 상대도수(%)는 표현 안됨

```
F M
6 4
```

- barplot(table(sample1$Gender),
  main = "연령대분포",
  names =c("여성","남성"))



데이터 탐색

6) 범주형 변수 - 빈도분석

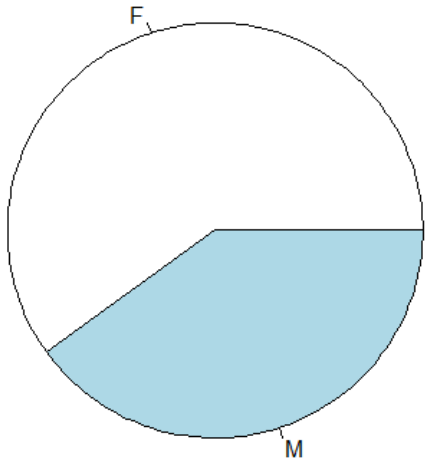
보고서. 10번 문제

파이차트(Pie Chart)

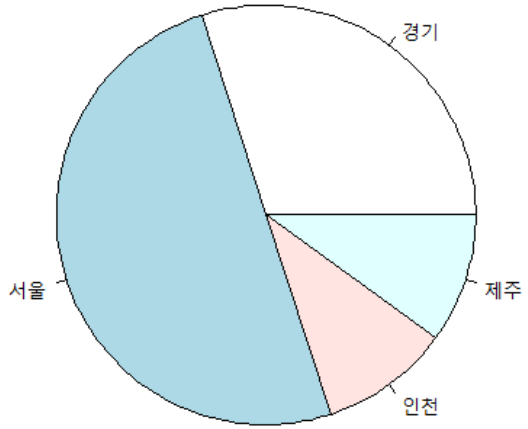
- 원형에 범주의 비율로 나눠 표현하는 그래프로, 전체대비 범주간의 비중을 비교할 때 유용하게 사용됨

- 성별과 지역의 파이차트

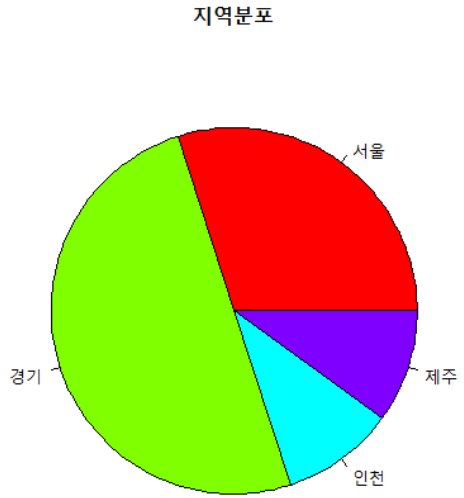
pie(table(sample1$Gender))



pie(table(sample1$AREA))




```
pie(table(sample1$AREA),
    main ="지역분포",  # 파이차트명
    labels=c( " 서울 " , " 경기 " , " 인천 " , " 제주 " ),  # label순서
    col=rainbow(4))  # 범주별 색상
```



데이터 가공

파생변수 생성하기

	ID	Gender	AGE	AREA	AMT17	Y17_CNT	AMT16	Y16_CNT
1	1	F	50	서울	1300000	50	100000	40
2	2	M	40	경기	450000	25	700000	30
3	3	F	28	제주	275000	10	50000	5
4	4	M	50	서울	400000	8	125000	3
5	5	M	27	서울	845000	30	760000	28
6	6	F	23	서울	42900	1	300000	6
7	7	F	56	경기	150000	2	130000	2
8	8	F	47	서울	570000	10	400000	7
9	9	M	20	인천	930000	4	250000	2
10	10	F	38	경기	520000	17	550000	16

분석을 좀 더 풍부하게 해 줄 수 있는 새로운 변수 생성
(기존변수를 활용하여)

새로운 변수	공식	변수명
2016과 2017년 전체 카드사용액	AMT17+AMT16	AMT
2016과 2017년 전체 카드건수	Y17_CNT + Y16_CNT	CNT
연평균 사용액	(AMT17+AMT16)/2	AVG_AMT
연평균 사용건수	(Y17_CNT+Y16_CNT)/2	AVG_CNT
건당 사용금액	(AMT17+AMT16)/(Y_17_CNT+Y16_CNT)	AMT_P_CNT

데이터 가공

데이터 가공 (전처리)에 유용한 패키지 dplyr

함수	기능
filter()	행 추출
select()	열(변수) 추출
arrange()	정렬
mutate()	변수 추가
summarise()	통계치 산출
group_by()	집단별로 나누기
left_join()	데이터 합치기(열)
bind_rows()	데이터 합치기(행)

- dplyr패키지는 %>% 기호를 이용하여 함수들을 나열하는 방식으로 코드 작성
- 데이터 셋 명을 계속 사용하지 않고, 변수명만 사용할 수 있음

filter의 기능

class	english	science
2	98	50
1	97	60
2	86	78
1	98	58
1	80	65
2	89	98

➡

class	english	science
1	97	60
1	98	58
1	80	65

select 의 기능

id	class	english	science
1	2	98	50
2	1	97	60
3	2	86	78
4	1	98	58
5	1	80	65
6	2	89	98

➡

class	english
2	98
1	97
2	86
1	98
1	80
2	89

ref:https://github.com/youngwoos/Doit_R/blob/master/Lecture/pdf/Doit_part06.pdf

데이터 가공

데이터 가공 (전처리)에 유용한 패키지 dplyr

서울지역만 가져와서 seoul이라 정의

```
seoul <- sample1 %>% filter(AREA=="서울")
```

%>% :shift+ctrl+m

	ID	Gender	AGE	AREA	AMT17	Y17_CNT	AMT16	Y16_CNT	AMT	CNT
1	1	F	50	서울	1300000	50	100000	40	1400000	90
2	4	M	50	서울	400000	8	125000	3	525000	11
3	5	M	27	서울	845000	30	760000	28	1605000	58
4	6	F	23	서울	42900	1	300000	6	342900	7
5	8	F	47	서울	570000	10	400000	7	970000	17

ID, Gender, AGE, AREA 변수만 선별하여 sample2 정의하기

```
sample2 <- sample1 %>% select(ID, Gender, AGE, AREA)
```

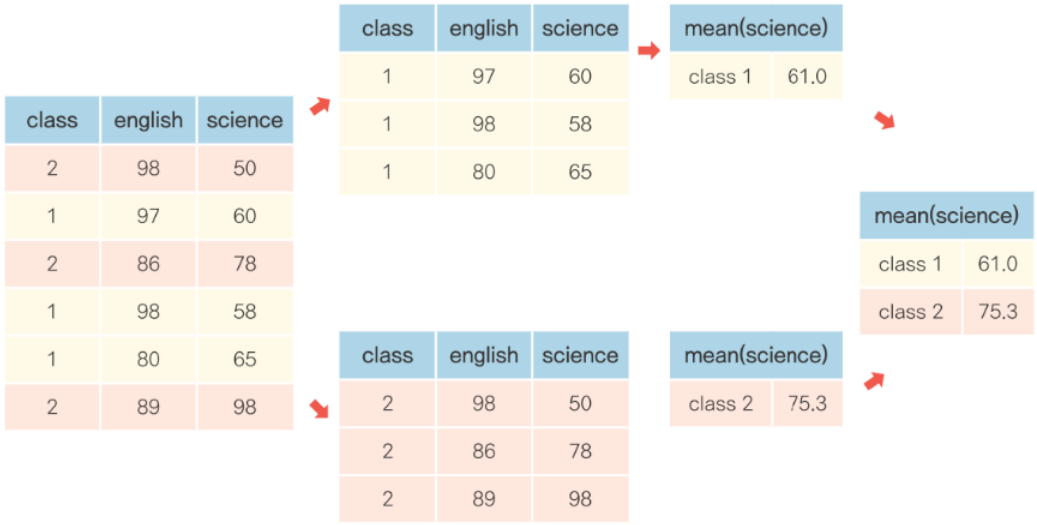
ID	Gender	AGE	AREA
1	F	50	서울
2	M	40	경기
3	F	28	제주
4	M	50	서울
5	M	27	서울
6	F	23	서울
7	F	56	경기
8	F	47	서울
9	M	20	인천
10	F	38	경기

데이터 가공

그룹별로 요약하기 (그룹별로 기술통계량 구하기)

보고서. 11번 문제

- group_by(), summarise()함수 활용
- 이 함수를 이용하여 요약표를 만들면 집단 간에 어떠한 차이가 있는 지를 쉽게 알 수 있음



- 성별별로, 2년간의 카드사용액 총합 (AMT)의 평균값, 합계, 최소값, 카드 사용자수(빈도), 최대값을 알아보자

```
sample1 %>%
  group_by(Gender) %>%
  summarise(Sum_AMT =sum(AMT),
            MEAN_AMT =mean(AMT),
            MIN_AMT =min(AMT),
            MAX_AMT=max(AMT),
             n=n())
```

```
Gender  Sum_AMT MEAN_AMT MIN_AMT MAX_AMT     n
<fct>     <dbl>    <dbl>   <dbl>   <dbl> <int>
F       4387900  731317.  280000 1400000     6
M       4460000 1115000   525000 1605000     4
```