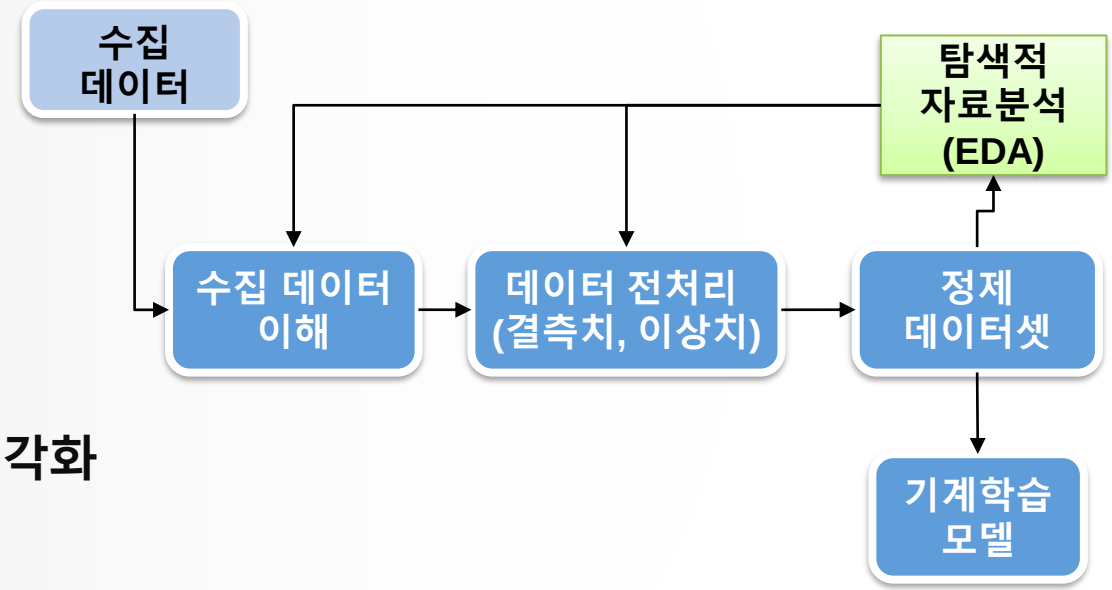
7. EDA & 데이터 전처리

chap07_EDA_Preprocessing 수업내용

1. EDA?
2. 수집 데이터 이해
3. 결측값(NA) 처리
4. 이상치(극단치) 처리
5. 코딩 변경
6. 탐색적 분석을 위한 시각화
7. 표본 샘플링



EDA?

1. EDA 란?

- 탐색적 데이터 분석(Exploratory Data Analysis)
- 수집 데이터를 다양한 각도에서 관찰하고 이해하는 과정
- 그래프나 통계적 방법으로 자료를 직관적으로 파악하는 과정

EDA?

2. EDA 필요성

1) 데이터의 분포와 통계를 파악하여 데이터가 갖고 있는 특성을 이해, 잠재적인 문제 발견

2) 분석 전에 발견이 어려운 문제를 다양한 문제점을 발견하고, 이를 바탕으로 기존 가설 수정 또는 새로운 가설 수립

EDA?

3. EDA 과정

1) 분석의 목적과 변수 특징 확인
   - ✓ 셀 수 없는 변수(Categorical), 셀 수 있는 변수(Numerical)
2) 데이터 셋 확인 및 전처리
   - ✓ 결측치, 이상치 확인 및 정제
3) 데이터 개별 변수 값 관찰
   - ✓ 통계, 시각화(이산변수, 연속변수)
4) 변수 간의 관계에 초점을 맞춰 변수 패턴 발견
   - ✓ 상관관계, 고급 시각화

단계1. 수집 데이터 이해

- 분석 전에 수집 데이터(명세서)를 면밀히 살펴보고, 이해하는 단계
- 자료구조, 관측치 길이, 변수 구성
- 각 변수 의미, 측정 방법, 척도 유형(명목,서열,비율,등간)
- 실제 명세서 또는 의미대로 코딩 되었는지 확인
- 데이터로부터 좋은 인사이트(insight)를 얻는 출발점

단계1. 수집 데이터 이해

[실습] 데이터 셋 특징

- **dataset # 데이터 셋 전체 보기**
- **View(dataset) # 별도의 데이터 뷰어창에서 출력됨**
- **head(dataset) # 앞부분 데이터 셋 6개**
- **tail(dataset) # 끝부분 데이터 셋 6개**
- **head(dataset, 10) # 앞부분 10개**
- **names(dataset)** # 변수명(컬럼)
- **attributes(dataset)** # names(열이름), class, row.names(행이름)
- **str(dataset)** # 데이터 구조 보기

단계1. 수집 데이터 이해

- **척도(Scale) 실습**

data Mart(dataset.csv)

resident	gender	job	age	position	price	survey
거주지역	성별	직업	나이	직위	구매가격	만족도
명목	명목	명목	비율	서열	비율	등간
1~5	1~ 2	1~3	20~69	1~5	?	5점척도

단계1. 수집 데이터 이해

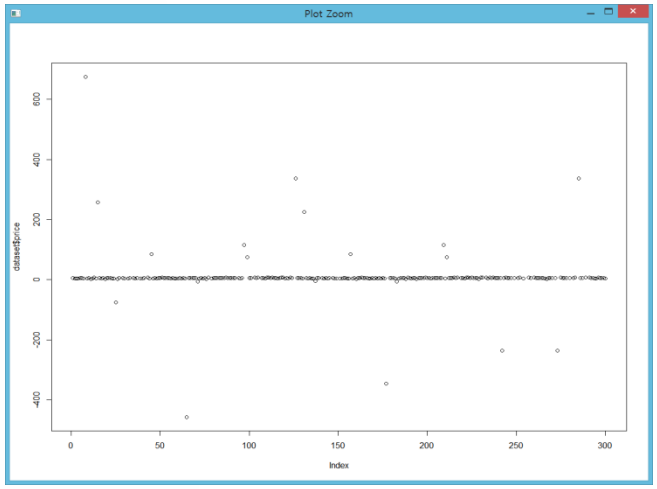
- **척도(Scale)**
  - **변수에 값을 부여하는 방법**
  - **변수 측정 단위(응답자가 선택할 수 있는 질문 항목)**

정성적-질적 척도(범주형 변수)		정량적-양적 척도(연속형 변수)	
명목척도	이름이나 범주를 대표하는 의미 없는 숫자 (예 : ① 남자 ② 여자)	**등간척도**	속성에 대한 각 수준 간의 간격이 동일한 숫자(가감산 연산) (예: 연소득이 어디에 해당되십니까?)
서열척도	측정 대상 간의 높고 낮음(서열), 순서에 대한 의미 없는 숫자 (예 : 좋아하는 순위를 표시하시오.)	**비율척도**	등간척도의 특성에 절대원점(0)이 존재하고, 비율계산이 가능한 숫자(사칙연산) (예 : 나이가 몇 세 입니까?)

단계1. 수집 데이터 이해

- 데이터 셋 조회

```
#dataset 데이터 중 특정변수 조회
dataset$age # [1] [27] [148] age값의 색인
dataset$resident
length(dataset$age) # data 수-300개

x <- dataset$gender # 결과를 변수에 저장
y <- dataset$price
x;y

plot(dataset$price) # 산점도 형태 전반적인 가격분포 보기

# $기호 대신 [""]기호를 이용한 특정변수 조회
dataset["gender"]
dataset["price"]
```



단계1. 수집 데이터 이해

```
# dataset 데이터 중 변수를 2개 이상 조회하는 경우
dataset[c("job","price")]
dataset[c(2,6)] # gender, price

dataset[c(1,2,3)] #resident,gender,age
dataset[c(1:3)] #resident,gender,age
dataset[c(2,4:6,3,1)] #gender,age,position,price,job,resident

# dataset 데이터 중 특정 행/열을 지정해 조회
dataset[,c(2:4)] #2~4열(gender job age) 전체 -> test[c(2:4)]과 동일
dataset[c(2:4),] #2~4행 전체
dataset[-c(1:100),] # 1~100행 제외
```

2. 결측치(NA) 발견과 처리

- 결측치(NA) 처리

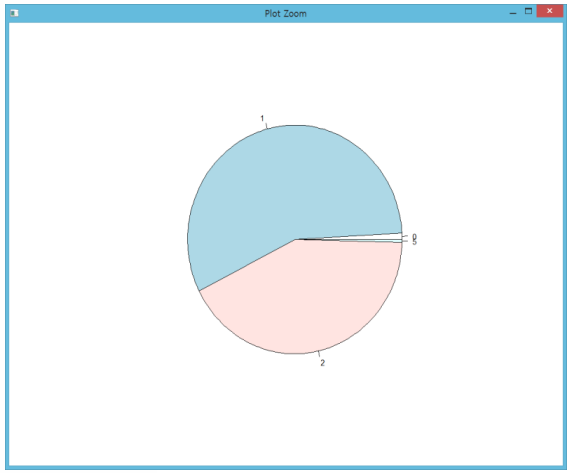
```
summary(dataset$price) # 결측치 확인 -> NA's - 30개
sum(dataset$price) # NA 출력

# 결측데이터 제거 방법1
sum(dataset$price, na.rm=T) # 2362.9

# 결측데이터 제거 방법2
price2 <- na.omit(dataset$price) # price에 있는 모든 NA 제거
sum(price2) # 2362.9
length(price2) # 270 -> 30개 제거
```

3. 이상치 발견과 정제

- 이상치(극단치) 발견과 정제

(1) 범주형 변수 극단치 처리

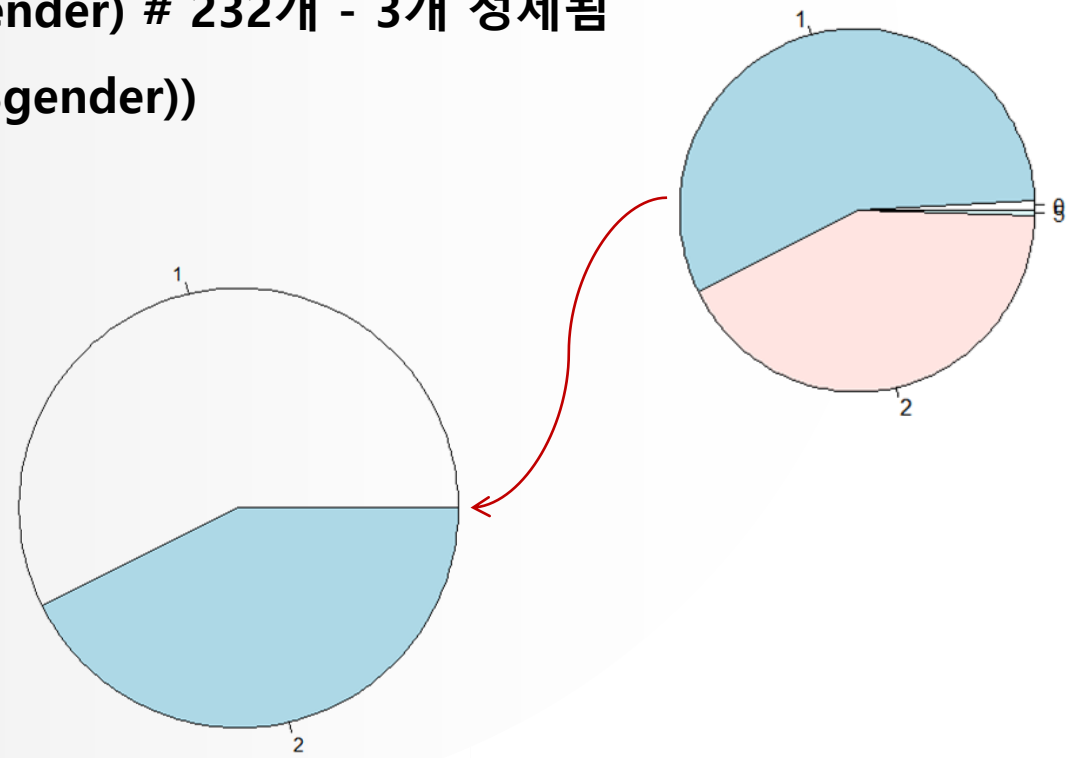
```
# gender 변수 outlier 확인
gender <- dataset$gender
hist(gender) # 히스토그램으로 outlier 확인
table(gender) # 빈도수로 outlier 확인
pie(table(gender)) # 파이 차트로 outlier 확인
```



3. 이상치 발견과 정제

- 성별 데이터 정제 - subset() 함수 이용

```
data <- subset(data,data$gender == 1 | data$gender == 2)
data # gender변수 데이터 정제
length(data$gender) # 232개 - 3개 정제됨
pie(table(data$gender))
```



3. 이상치 발견과 정체

2) 연속형 변수 극단치 처리

```
# price outlier 확인
dataset$price # 세부 데이터 보기
length(dataset$price) #300개(NA포함)
plot(dataset$price) # 산점도 형태 전반적인 가격분포 보기
summary(dataset$price) # -457~675 범위확인
```

3. 이상치 발견과 정체

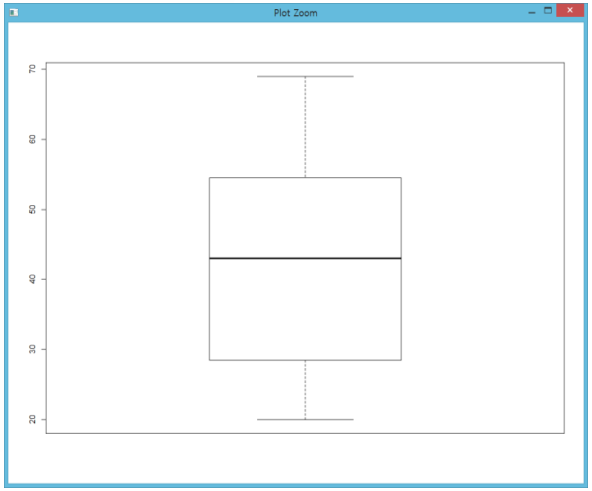
 # price변수 정제(2~8)

```
data <- subset(dataset, dataset$price >= 2 & dataset$price <= 8)

length(data$price) #251개(49개 정제)

stem(data$price) # 줄기와 잎 도표보기
```

```
The decimal point is at the |

  2 | 133
  2 |
  3 | 000003344
  3 | 555589
  4 | 00000000000000111111111222333334444
  4 | 566666777889999
  5 | 0000000000000000001111111111222222222333334444444
  5 | 5555555556677777788899
  6 | 000000000011111122222222222223333333333333444444444
  6 | 5555777777788889999
  7 | 000111122
  7 | 7799
```

3. 이상치 발견과 정체

```
# age 변수 NA 발견
summary(data$age) # Min(20), Max(69), NA(16)
length(data$age) # 251

# age 변수 정제(20~69)
data <- subset(data,data$age >= 20 & data$age <= 69)
length(data$age) # 235개(16 정제)
# box 플로팅으로 평균연령 분석
boxplot(data$age) # 45대 중반 평균 연령
```



4. 코딩 변경

- **코딩변경 - 변수변환 : 리코딩 하기**

  - 데이터의 가독성, 척도 변경, 최초 코딩 내용 변경을 목적으로 수행

```
# 문자열로 리코딩(청년층, 중년층, 장년층)
data$age2[data$age <= 30] <-"청년층"
data$age2[data$age > 30 & data$age <=45] <-"중년층"
data$age2[data$age > 45] <-"장년층"

head(data) # data 테이블 전체 - age와 age2 비교
head(data[c("age","age2")]) # 2개만 지정
# age  age2
# 26  청년층
head(data) # dataset 테이블 전체 - age2 컬럼 생성
head(data[c("age","age2")]) # 2개만 지정
# age  age2  age3
# 26  청년층  1
```

4. 코딩 변경

- **코딩변경 : 역코딩 : 긍정순서(1 -> 5, 5 -> 1)**

```
data$survey
survey <- data$survey
csurvey <- 6-survey
csurvey
survey  # 역코딩 결과와 비교
data$survey <- csurvey # data set에 survery 변수 수정
head(data) # survey 결과 확인
```

```
 resident gender job age position price survey
1        1      1   1  26        2   5.1      1
2        2      1   2  54        5   4.2      2
3       NA      1   2  41        4   4.7      4
4        4      2  NA  45        4   3.5      2
5        5      1   3  62        5   5.0      1
6        3      1   2  57       NA   5.4      2
```

5. 정제 데이터 저장

- 정제된 데이터 저장

```
setwd("C:/Rwork/Part-II")

# (1) 정제된 데이터 저장
write.csv(dataset2,"cleanData.csv", quote=F, row.names=F)

# (2) 저장된 파일 불러오기/확인
new_data <- read.csv("cleanData.csv", header=TRUE)
new_data
dim(new_data) # 248  13
```

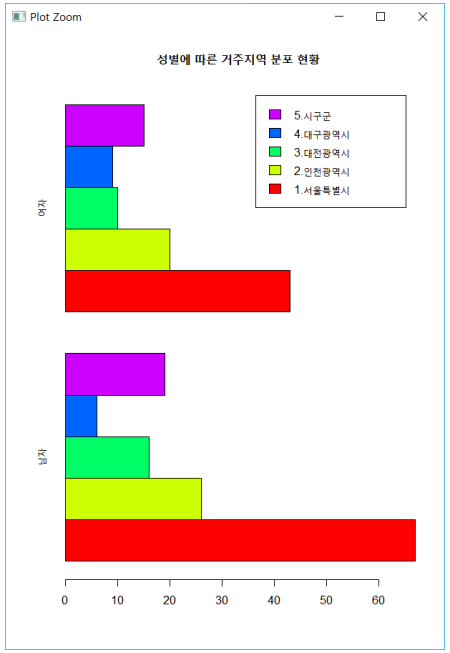
6. 변수 간의 관계 분석

- **변수 간의 관계 분석을 위한 시각화**

1) 명목척도(범주/서열) : 명목척도(범주/서열) - 거주지역 : 성별

- 거주지역과 성별 칼럼 시각화

```
resident_gender <- table(new_data$resident2, new_data$gender2)
gender_resident <- table(new_data$gender2, new_data$resident2)

# 성별에 따른 거주지역 분포 현황
barplot(resident_gender, beside=T, horiz=T,
        col = rainbow(5),
        legend = row.names(resident_gender),
        main = '성별에 따른 거주지역 분포 현황')
# row.names(resident_gender) # 행 이름
```

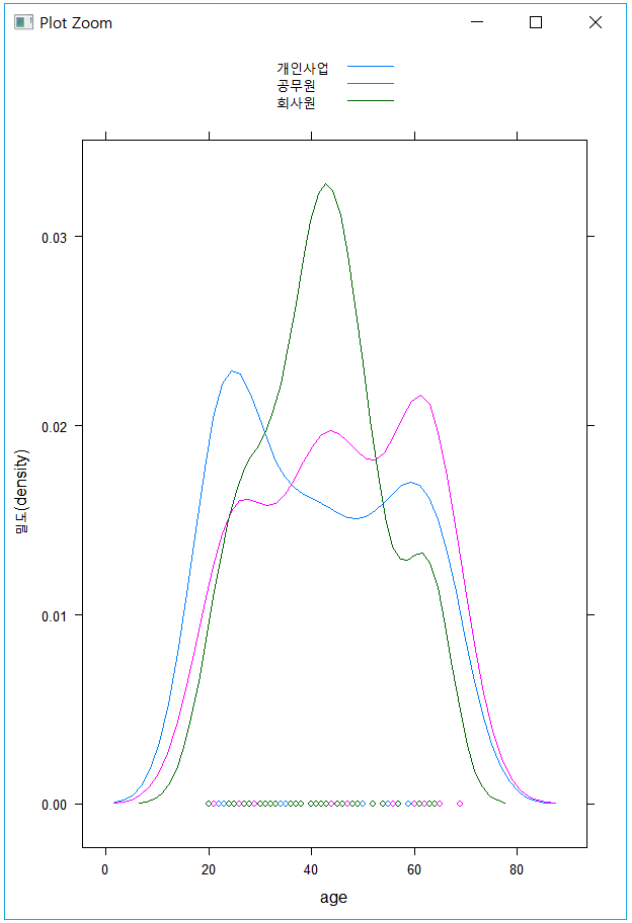


6. 변수 간의 관계 분석

2) 비율척도(연속) : 명목척도(범주/서열)

- 나이와 직업유형에 따른 시각화

```
install.packages("lattice")
library(lattice)

# 나이와 직업유형 데이터 분포
densityplot( ~ age, data=new_data, groups = job2,
             plot.points=T, auto.key = T)
# plot.points=T : 밀도, auto.key = T : 범례
```

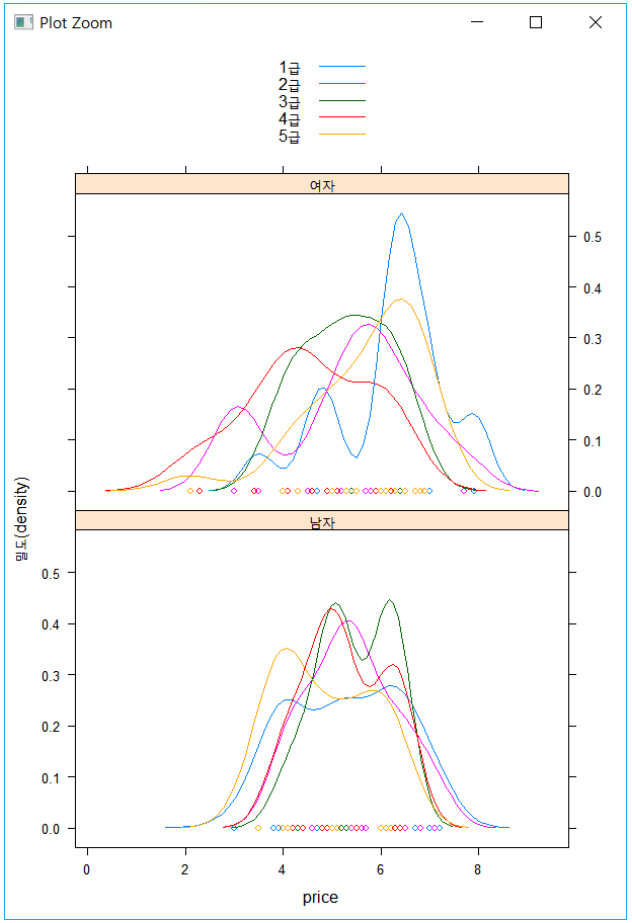


6. 변수 간의 관계 분석

3) 비율(연속),명목(범주/서열) : 명목(범주/서열)

- 구매비용(연속) : x칼럼 , 직급(서열):조건,
  성별(범주):그룹

```
densityplot(~ price | factor(gender2),
    data=new_data,
        groups = position2, plot.points=T,
    auto.key = T)
# groups = gender2 : 한 격자에 나타날 그룹 칼럼
    지정
```



7. 표본 샘플링

❖ 표본 샘플링 예

```
# 정제 데이터 대상 셈플링하기
nrow(data) # 235개 : 행수 구하기 -> Number of Rows

# 235개 중 30개 무작위 추출
choice1 <- sample(nrow(data), 30)
choice1 # 추출된 행 번호 출력

# 50~235 사이에서 30개 무작위 추출
choice2 <- sample(50:nrow(data), 30)
choice2

# 50~100 사이에서 30개 무작위 추출
choice3 <- sample(c(50:100), 30)
choice3
```

Sampling 결과는 데이터 셋에서 선택된 레코드 번호를 의미