

EDA와 고급 시각화

Sangkon Han(sangkon@pusan.ac.kr)

2023-03-24

선택된 연습문제

- 아래 선택된 연습문제를 practice_20130317.R에 저장해서 R파일만 제출하시면 됩니다. 연습문제 중 해결이 힘들거나 어려운 문제는 해당 파일에 주석으로 문제점이나 궁금증을 남겨주세요.
- 연습문제 해결시 꼭 책을 참고하시고, 다같이 모여서 해결하시고 코드보다는 절차를 익히는데 집중하세요.
- 선택된 연습문제
 - 7장 연습문제 2번, 3번, 4번
 - 8장 연습문제 1번, 2번

7장 연습문제 풀이

```
install.packages("latticeExtra", repos = "https://cran.us.r-project.org")
```

```
## 패키지 'latticeExtra'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다
##
## 다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
## C:\Users\sigma\AppData\Local\Temp\RtmpYfqXiT\downloaded_packages
```

```
library(lattice)
library(latticeExtra)
```

7장 2번 연습문제 풀이

```
dataset <- read.csv("./data/dataset.csv", header = T)
dataset2 <- subset(dataset, price >= 2 & price <= 8)
pos <- dataset2$position
cpos <- 6 - pos
dataset2$position <- cpos
dataset2$position2[dataset2$position==1] <- '1급'
dataset2$position2[dataset2$position==2] <- '2급'
dataset2$position2[dataset2$position==3] <- '3급'
dataset2$position2[dataset2$position==4] <- '4급'
dataset2$position2[dataset2$position==5] <- '5급'
range(dataset2$resident, na.rm = T) # 1 5
```

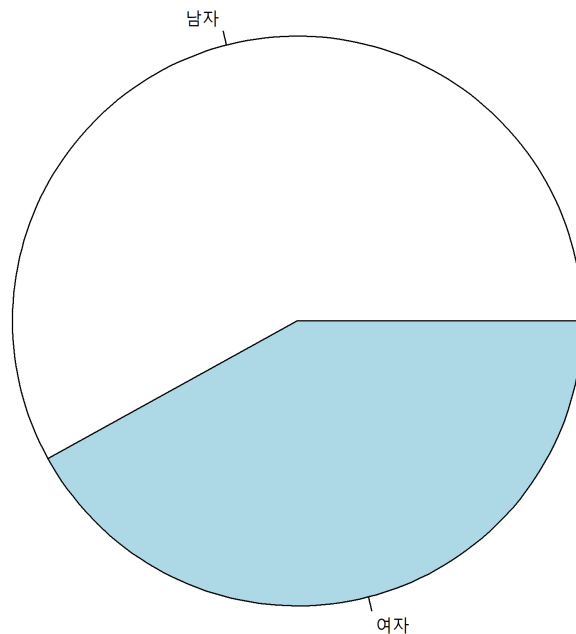
```
## [1] 1 5
```

```
dataset2 <- subset(dataset2, !is.na(dataset2$resident))  
head(dataset2)
```

```
##   resident gender job age position price survey position2  
## 1         1     1   1  26         4   5.1       1         4급  
## 2         2     1   2  54         1   4.2       2         1급  
## 4         4     2  NA  45         2   3.5       2         2급  
## 5         5     1   3  62         1   5.0       1         1급  
## 6         3     1   2  57        NA   5.4       2        <NA>  
## 7         2     2   1  36         3   4.1       4         3급
```

7장 3번 연습문제 풀이

```
dataset2$gender2[dataset2$gender == 1] <- '남자'  
dataset2$gender2[dataset2$gender == 2] <- '여자'  
pie(table(dataset2$gender2))
```



7장 4번 연습문제 풀이

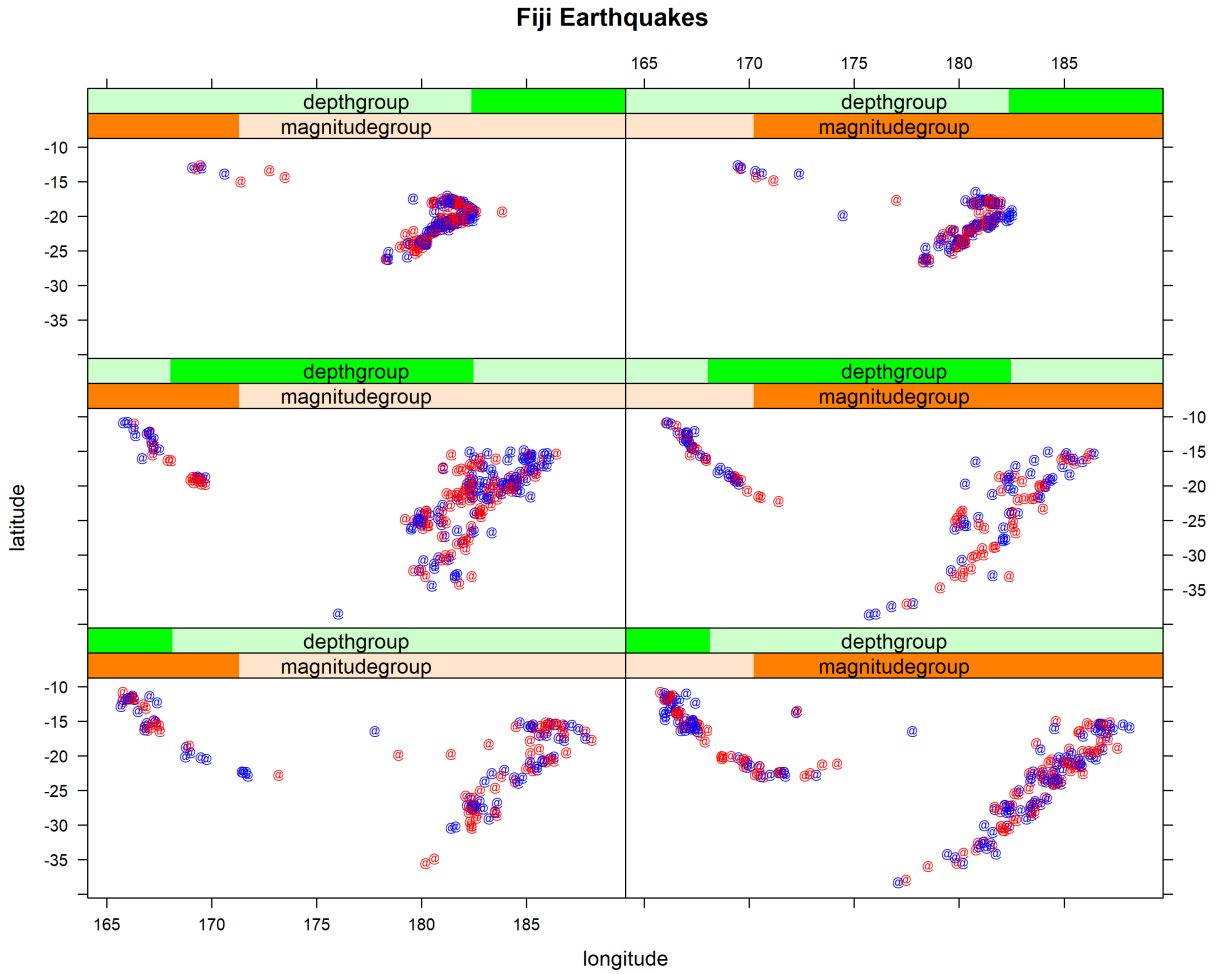
```
dataset2$age2[dataset2$age <= 30] <- "청년층"
dataset2$age2[dataset2$age > 30 & dataset2$age <= 55] <- "중년층"
dataset2$age2[dataset2$age > 55 ] <- "장년층"
dataset2$age3[dataset2$age <= 30] <-1
dataset2$age3[dataset2$age > 30 & dataset2$age <= 55] <-2
dataset2$age3[dataset2$age > 55] <-3
head(dataset2[c('age', 'age2', 'age3')])
```

```
##   age  age2 age3
## 1  26 청년층   1
## 2  54 중년층   2
## 4  45 중년층   2
## 5  62 장년층   3
## 6  57 장년층   3
## 7  36 중년층   2
```

8장 연습문제 풀이

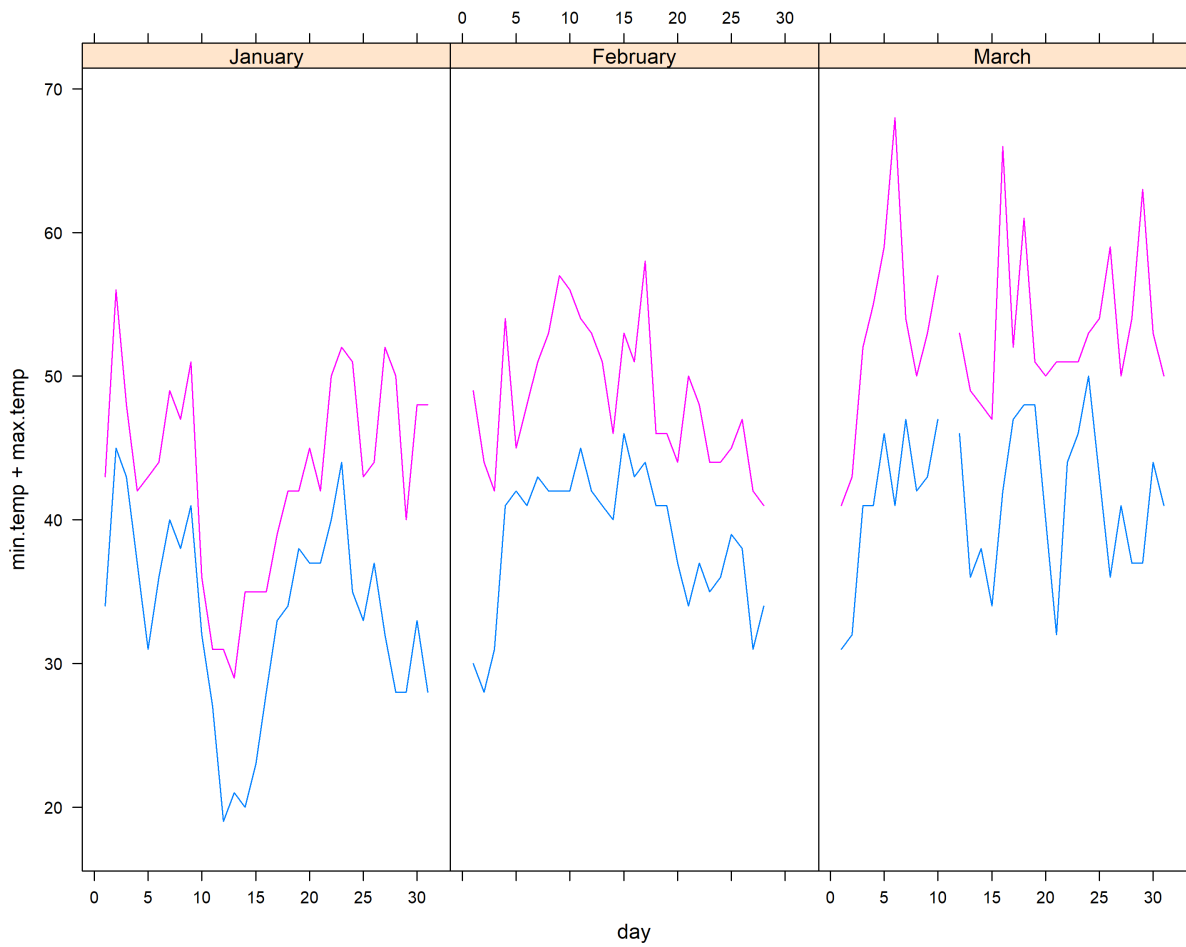
8장 1번 연습문제 풀이

```
depthgroup<-equal.count(quakes$depth, number=3, overlap=0)
magnitudegroup<-equal.count(quakes$mag, number=2, overlap=0)
xyplot(lat ~ long | magnitudegroup*depthgroup, data=quakes,
       main="Fiji Earthquakes", ylab="latitude", xlab="longitude",
       pch="@", col=c("red", "blue"))
```



8장 2번 연습문제 풀이

```
xyplot(min.temp + max.temp ~ day | month, data=SeatacWeather,
       type="l", layout=c(3,1))
```



2023년 3월 17일 강의 내용 요약 및 정리

EDA 5단계

- 1. 데이터 확인
- 2. 결측치 제거
- 3. 코딩(숫자 -> 문자)
- 4. 시각화
- 5. 보고서 작성

0. 프로젝트 구성

- 경로를 확인하세요.

```
getwd()
```

```
## [1] "C:/Users/sigma/works/practice-r"
```

- 필요한 패키지/라이브러리를 설치하세요.

```
suppressPackageStartupMessages({  
  install.packages("ggplot2", repos = "https://cran.us.r-project.org")  
  install.packages("lattice", repos = "https://cran.us.r-project.org")  
})
```

```
## 패키지 'ggplot2'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다  
##
```

```
## 다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
```

```
## C:\Users\sigma\AppData\Local\Temp\RtmpGKyTbj\downloaded_packages
```

```
## 패키지 'lattice'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다  
##
```

```
## 다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
```

```
## C:\Users\sigma\AppData\Local\Temp\RtmpGKyTbj\downloaded_packages
```

- 설치된 패키지 중에서 사용할 것을 불러오세요.

```
library(ggplot2)  
library(lattice)
```

1. 데이터 확인

- `str()` 명령어를 사용해서 데이터를 확인하도록 합니다. 특히 자료형에 주의하세요!
- 데이터는 '컬럼(열)'을 기준으로 선택됩니다. 데이터에서 컬럼과 열을 선택하는 방법은 과제1-1을 통해서 연습하도록 합니다.

```
dataset <- read.csv("./data/dataset.csv", header = T)  
str(dataset)
```

```
## 'data.frame': 300 obs. of 7 variables:  
## $ resident: int 1 2 NA 4 5 3 2 5 NA 2 ...  
## $ gender : int 1 1 1 2 1 1 2 1 1 1 ...  
## $ job : int 1 2 2 NA 3 2 1 2 1 2 ...  
## $ age : int 26 54 41 45 62 57 36 NA 56 37 ...  
## $ position: int 2 5 4 4 5 NA 3 3 5 3 ...  
## $ price : num 5.1 4.2 4.7 3.5 5 5.4 4.1 675 4.4 4.9 ...  
## $ survey : int 1 2 4 2 1 2 4 4 3 3 ...
```

2. 결측치 제거

2.1 결측치 확인 및 제거

결측치 제거를 위해선 기본적으로 결측치가 얼마나 있는지 확인해야 합니다.

```
summary(dataset$price)
```

```
##      Min.   1st Qu.   Median     Mean  3rd Qu.    Max.     NA's
## -457.200    4.425    5.400    8.752    6.300   675.000     30
```

```
sum(dataset$price)
```

```
## [1] NA
```

- 결측치 제거를 위한 함수 사용법은 아래와 같습니다.

```
price2 <- na.omit(dataset$price)
sum(price2)
```

```
## [1] 2362.9
```

2.2 결측치 및 이상치 제거

- 데이터 분석에 필요한 데이터를 기준으로 데이터 정제를 진행
- 기존 데이터와 별의 데이터를 생성할 필요가 있음

```
# 실습: price 변수의 데이터 정제와 시각화
dataset2 <- subset(dataset, price >= 2 & price <= 7.9)
dataset2 <- subset(dataset2, age >= 20 & age <= 69)
dataset2 <- na.omit(dataset2)
summary(dataset2)
```

```
##      resident      gender      job      age      position
## Min.   :1.000   Min.   :0.000   Min.   :1.0   Min.   :20.00   Min.   :1.000
## 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.0   1st Qu.:29.00   1st Qu.:2.000
## Median :2.000   Median :1.000   Median :2.0   Median :42.00   Median :4.000
## Mean   :2.189   Mean   :1.423   Mean   :2.1   Mean   :42.45   Mean   :3.333
## 3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.0   3rd Qu.:54.00   3rd Qu.:5.000
## Max.   :5.000   Max.   :5.000   Max.   :3.0   Max.   :69.00   Max.   :5.000
##      price      survey
## Min.   :2.100   Min.   :1.000
## 1st Qu.:4.600   1st Qu.:2.000
## Median :5.300   Median :3.000
## Mean   :5.373   Mean   :2.662
## 3rd Qu.:6.200   3rd Qu.:3.000
## Max.   :7.900   Max.   :5.000
```

3. 코딩(숫자 -> 문자)

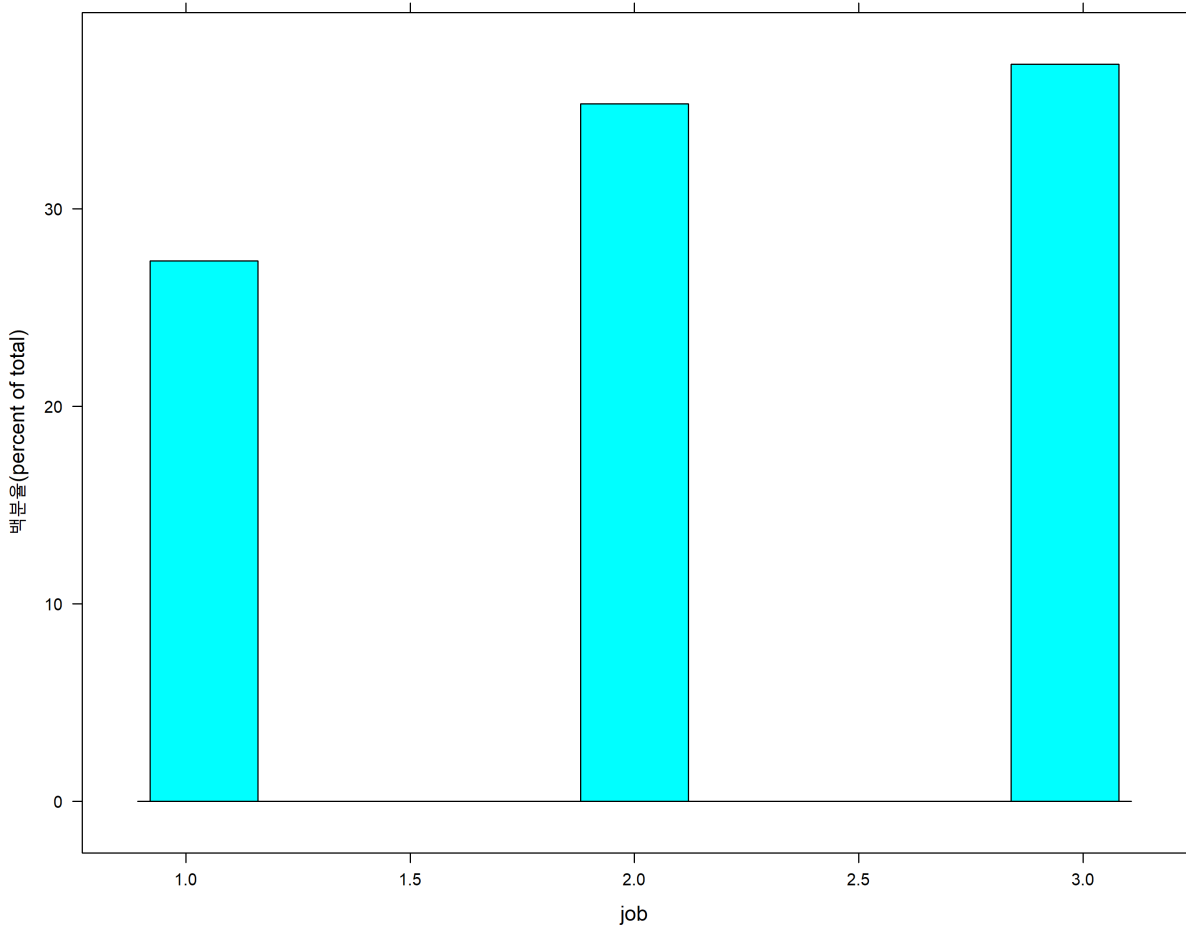
```
dataset2$job2[dataset2$job == 1] <- '공무원'
dataset2$job2[dataset2$job == 2] <- '회사원'
dataset2$job2[dataset2$job == 3] <- '개인사업'
head(dataset2[c("job", "job2")])
```

```
##      job      job2
## 1      1      공무원
## 2      2      회사원
## 5      3      개인사업
## 7      1      공무원
## 10     2      회사원
## 12     3      개인사업
```

4. 시각화(lattice)

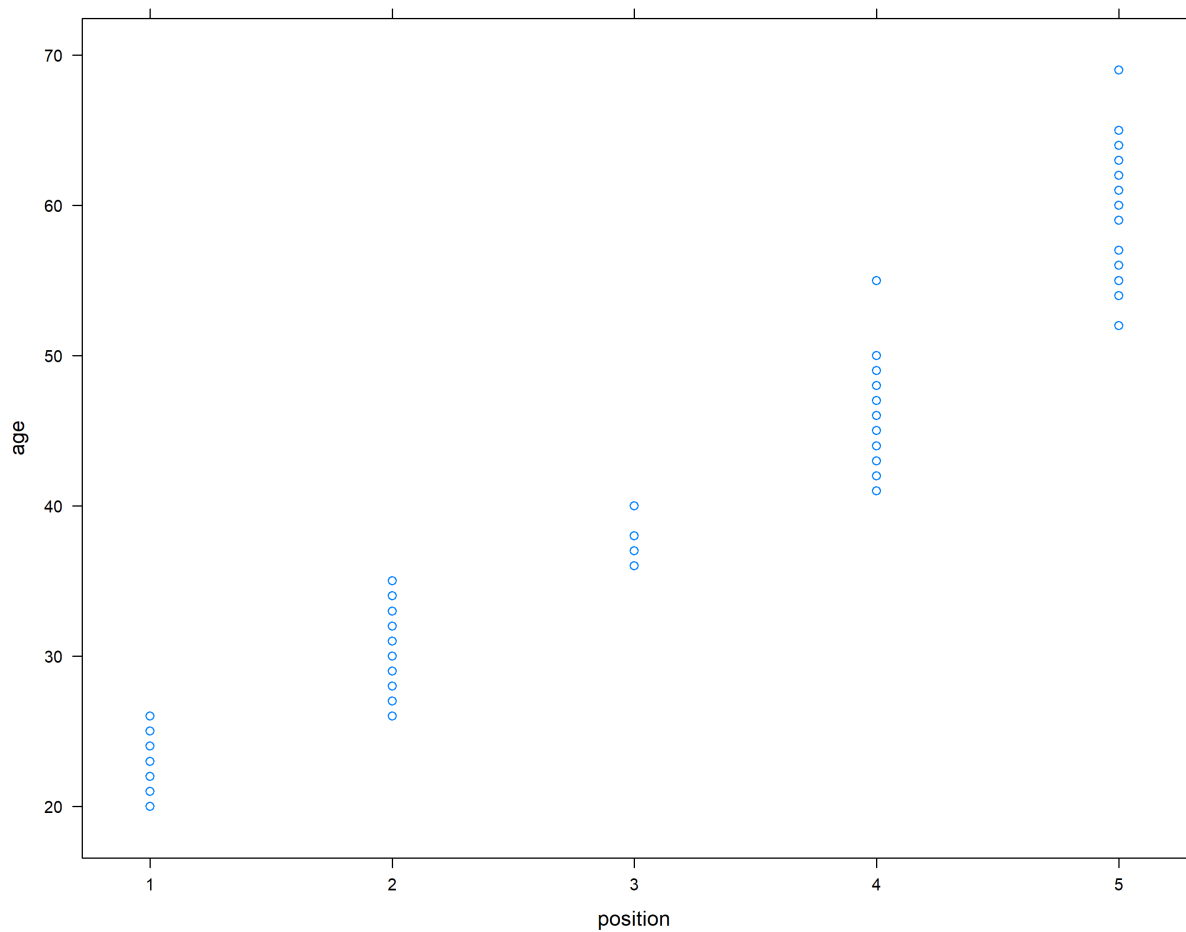
- 히스토그램

```
histogram(~job, data = dataset2)
```



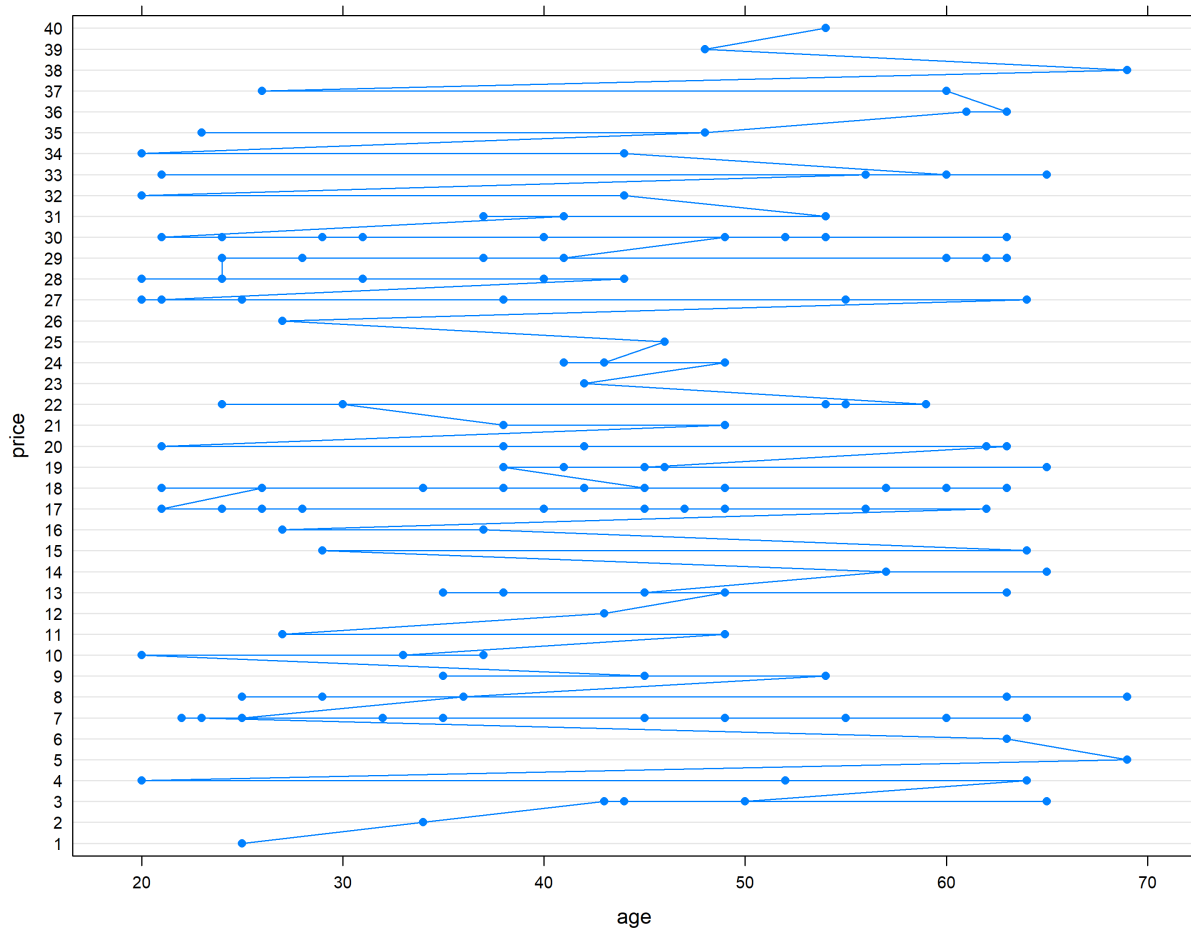
- 산포도

```
xyplot(age~position, data = dataset2)
```

- 간단한 선 그래프

```
dotplot(price~age, data=dataset2, type="o")
```



5. 보고서 작성

생략