공공자료를 이용한 데이터 분석

환경생태데이터사이언스 실습 December 3, 2019

오늘의 학습 목표

오늘의 학습 목표

1. 자료를 이용한 한국인의 삶 분석

- 김영우 (2017) 쉽게 배우는 R 데이터 분석 - 셋째마당, 실전! 데이터 분석 프로젝트 발췌

‘공공자료를 통해 본 한국인의 삶’

‘한국복지패널데이터’ 분석 준비

- 쉽게 배우는 R 데이터 분석 자료 저장소 `bit.ly/doit_rb`
  - 2016년 발간된 복지패널데이터 이용
  - 6,914가구, 16,664명에 대한 정보
  - `Koweps_hpc10_2015_beta1.sav` (spss) 파일 다운로드

자료 읽기

```
#install.packages("foreign") # to read spss type file
library(foreign)
library(dplyr)
library(ggplot2)
library(readxl)
raw_welfare <- read.spss(file="./Data/Koweps_hpc10_2015_beta1.sav",
                         to.data.frame = T, reencode="euc-kr")

## Warning in read.spss(file = "./Data/
## Koweps_hpc10_2015_beta1.sav", to.data.frame = T, : ./Data/
## Koweps_hpc10_2015_beta1.sav: Compression bias (0) is not the
## usual value of 100

## re-encoding from euc-kr

welfare <- raw_welfare
```

자료 검토하기

```r
head(welfare) # Check the first six rows of the data
tail(welfare) # Check the last six rows of the data
# View (welfare) # Display data
dim(welfare) # Check the dimension of the data
str(welfare) # Check the basic structure of the data
summary(welfare) # Check basic descriptive statistics of the data.
```

변수명 바꾸기

- 복지패널데이터의 열 이름에 있는 코드를 쉬운 단어로 치환
  - 성별, 출생년도, 혼인상태, 종교, 수입, 직업 코드, 지역 코드

```
welfare <- welfare %>% rename(sex         = h10_g3,
                              birth       = h10_g4,
                              marriage    = h10_g10,
                              religion    = h10_g11,
                              income      = p1002_8aq1,
                              code_job    = h10_eco9,
                              code_region = h10_reg7)
```

변수 간 관계 분석 1

- 성별에 따라 월급이 다를까?
- 복지패널데이터 성별 코드
    - 1: 남자, 2: 여자, 9: 모름/무응답

```
# 1. Check the variable (sex)
class(welfare$sex) # Check the variable type
```

```
## [1] "numeric"
```

```
table(welfare$sex) # Check number of items included in each factor
```

```
##
##    1    2
## 7578 9086
```

변수 간 관계 분석 1

```r
# If there is factor with 9 which means NAs, then change 9 to NA.
welfare$sex <- ifelse(welfare$sex == 9, NA, welfare$sex)
# Check how many items are classified as NAs.
table(is.na(welfare$sex))
```

```
##
## FALSE
## 16664
```

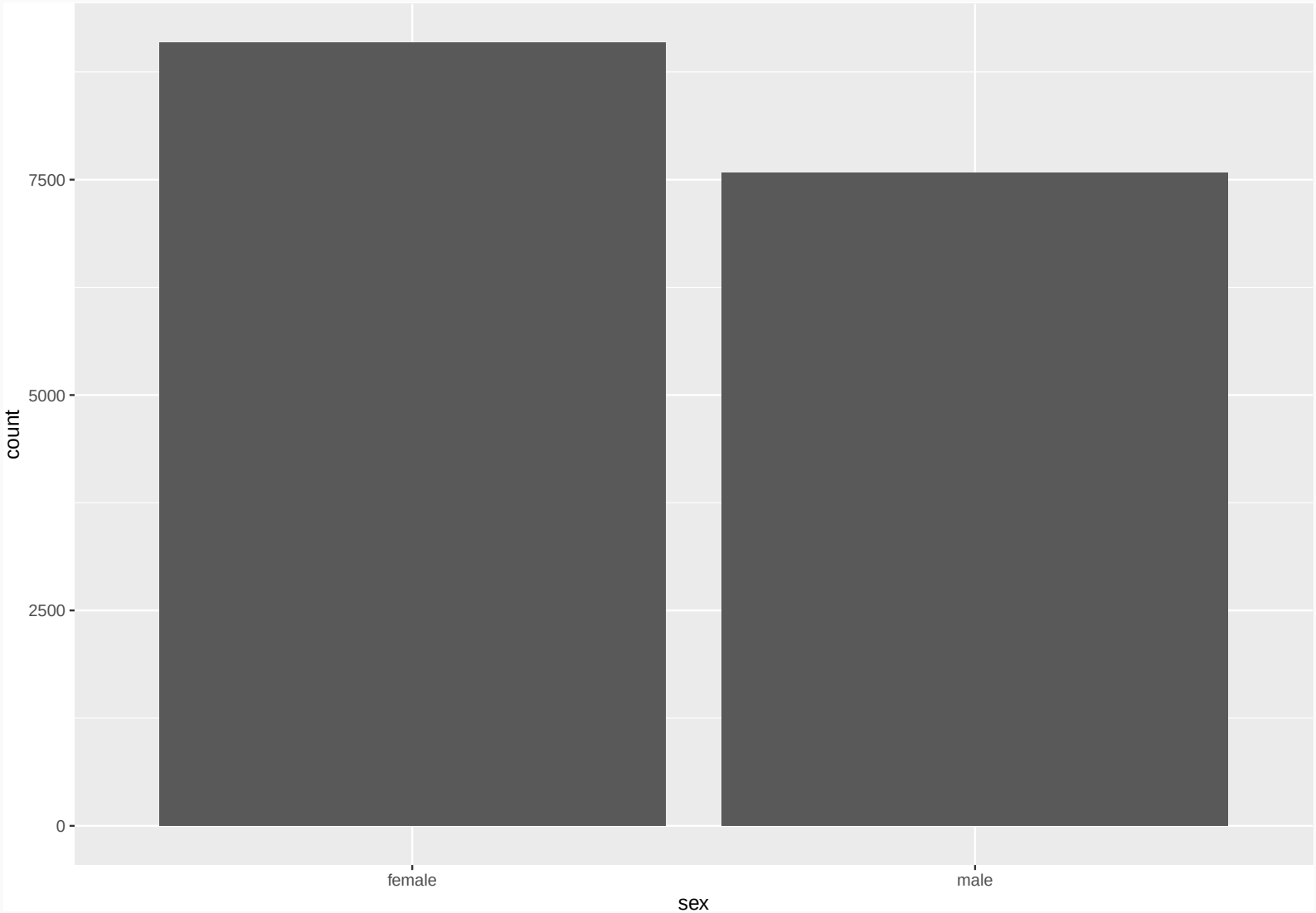
```r
welfare$sex <- ifelse(welfare$sex == 1, "male", "female")
table(welfare$sex)
```

```
##
## female   male
##   9086   7578
```

```r
NP_sex <- ggplot(data=welfare, aes(x=sex)) + geom_bar()
```

변수 간 관계 분석 1

```
NP_sex
```



변수 간 관계 분석 1

- 월급 변수 검토 및 전처리

```
class(welfare$income)
```

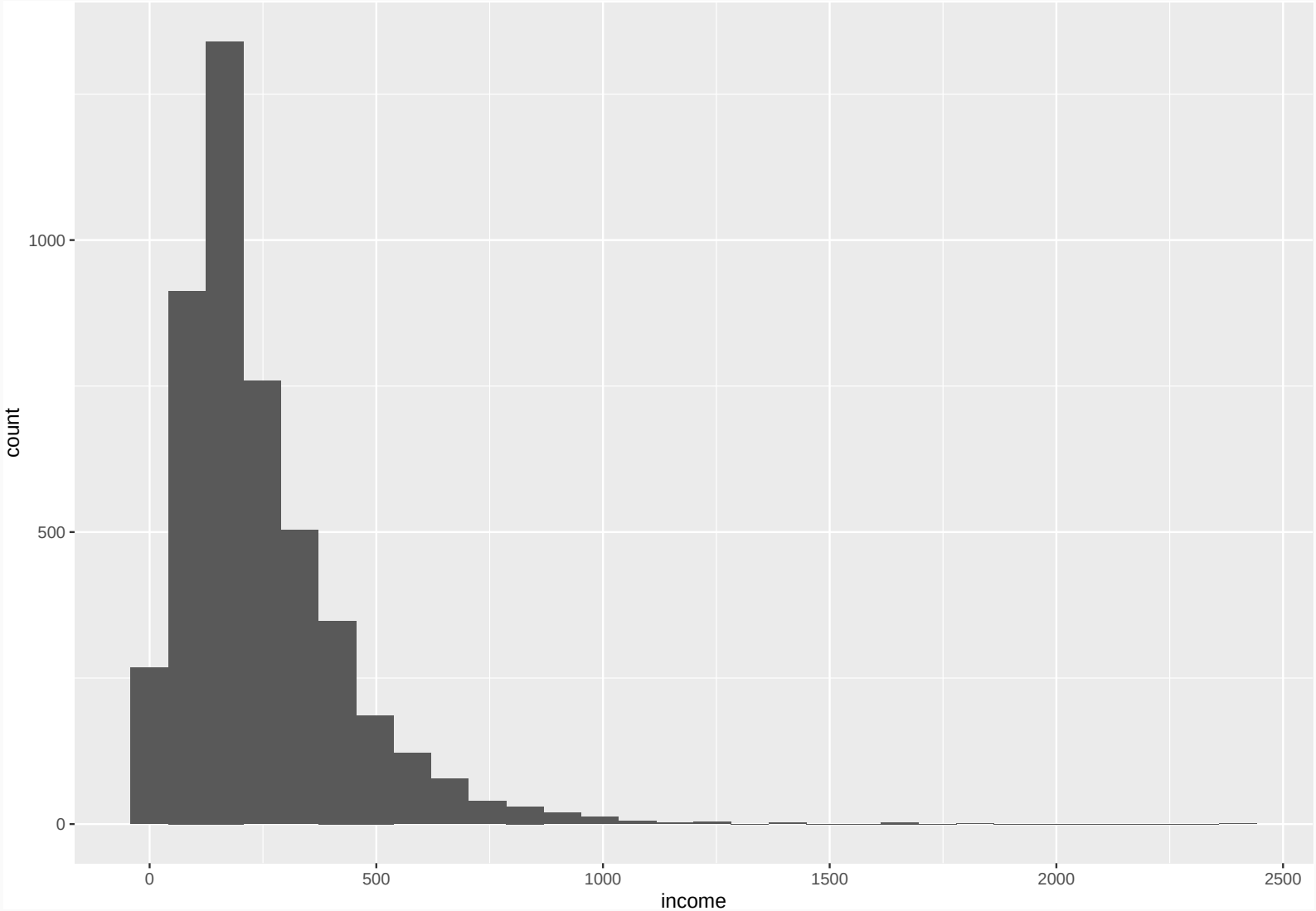
```
## [1] "numeric"
```

```
summary(welfare$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   122.0   192.5   241.6   316.6  2400.0   12030
```

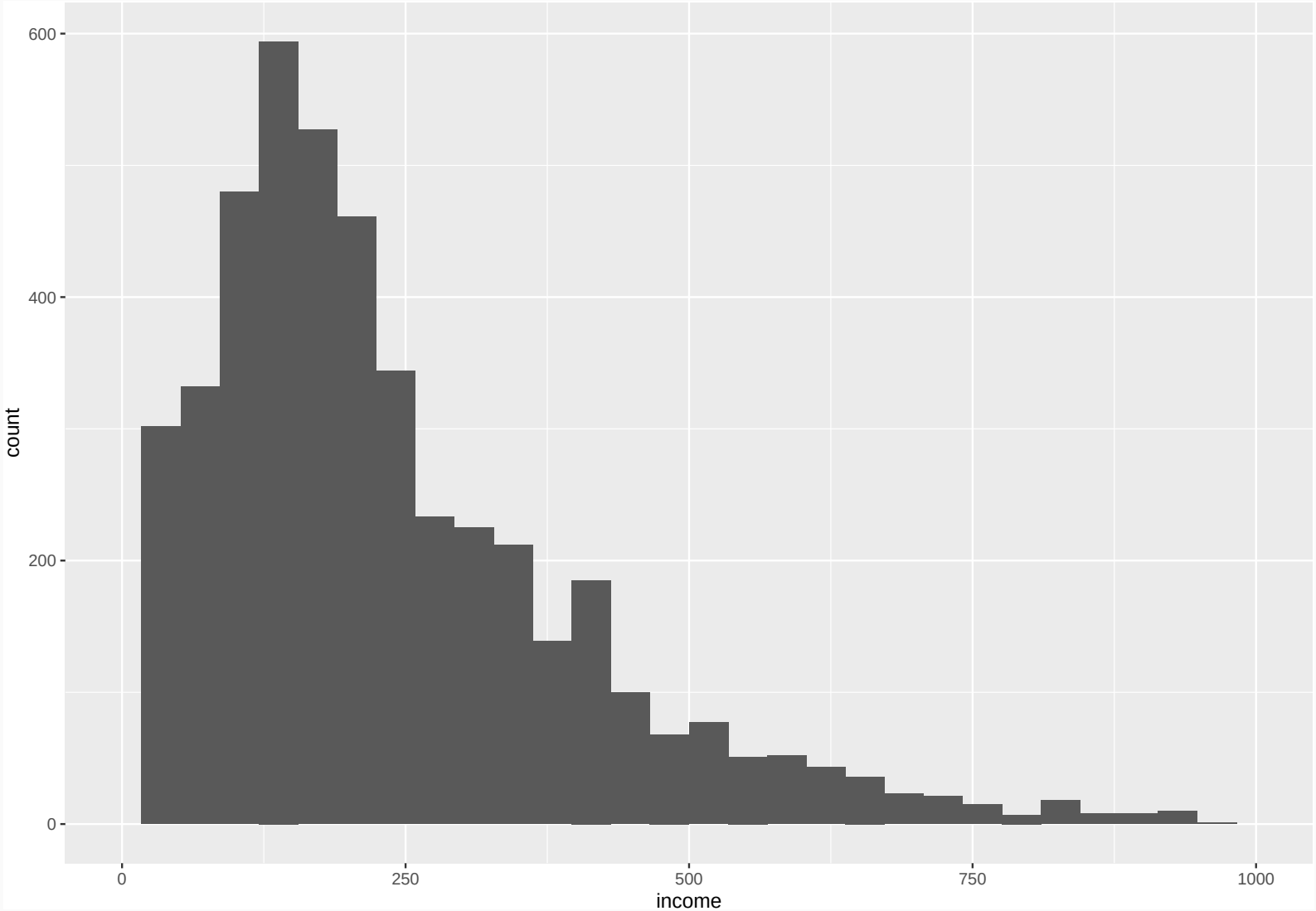
변수 간 관계 분석 1

```
ggplot(data=welfare, aes(x=income)) + geom_histogram()
```



변수 간 관계 분석 1

```
ggplot(data=welfare, aes(x=income)) + geom_histogram() + xlim(0,1000)
```



변수 간 관계 분석 1

- 이상치 및 결측 처리
  - 9999는 모름/무응답을 의미
  - 소득이 0인 경우 이상치로 간주

```
welfare$income <- ifelse(welfare$income %in% c(0, 9999),
                         NA, welfare$income)
table(is.na(welfare$income))
```

```
##
## FALSE  TRUE
##  4620 12044
```

변수 간 관계 분석 1

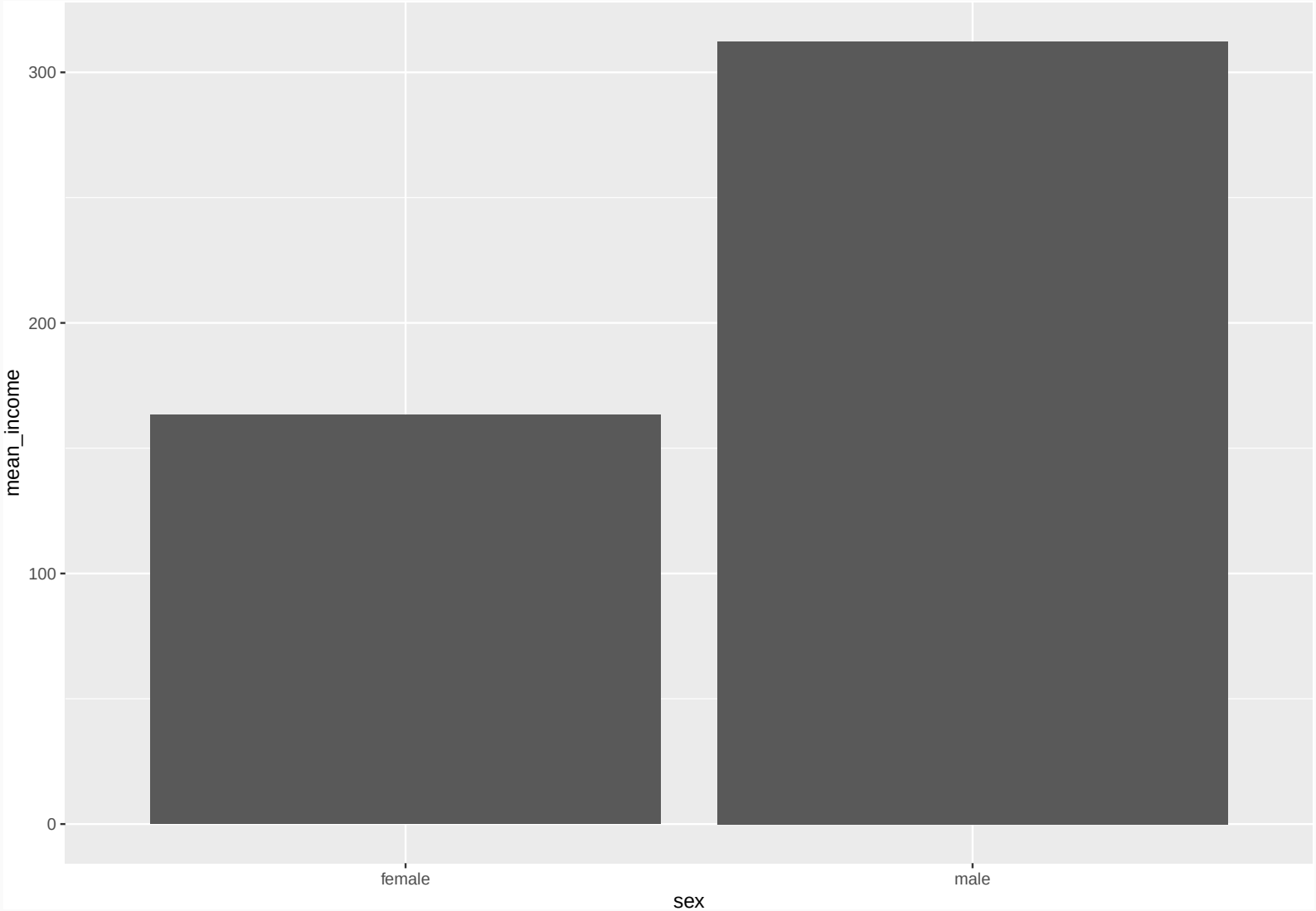
- 성별 월급 평균표 작성

```
income_sex <-   welfare %>%
                filter(!is.na(income)) %>%
                group_by(sex) %>%
                summarize(mean_income = mean(income, na.rm=T))
income_sex
```

```
## # A tibble: 2 x 2
##   sex    mean_income
##   <chr>        <dbl>
## 1 female        163.
## 2 male          312.
```

변수 간 관계 분석 1

```
ggplot(data = income_sex, aes(x=sex, y=mean_income)) + geom_col()
```



변수 간 관계 분석 1

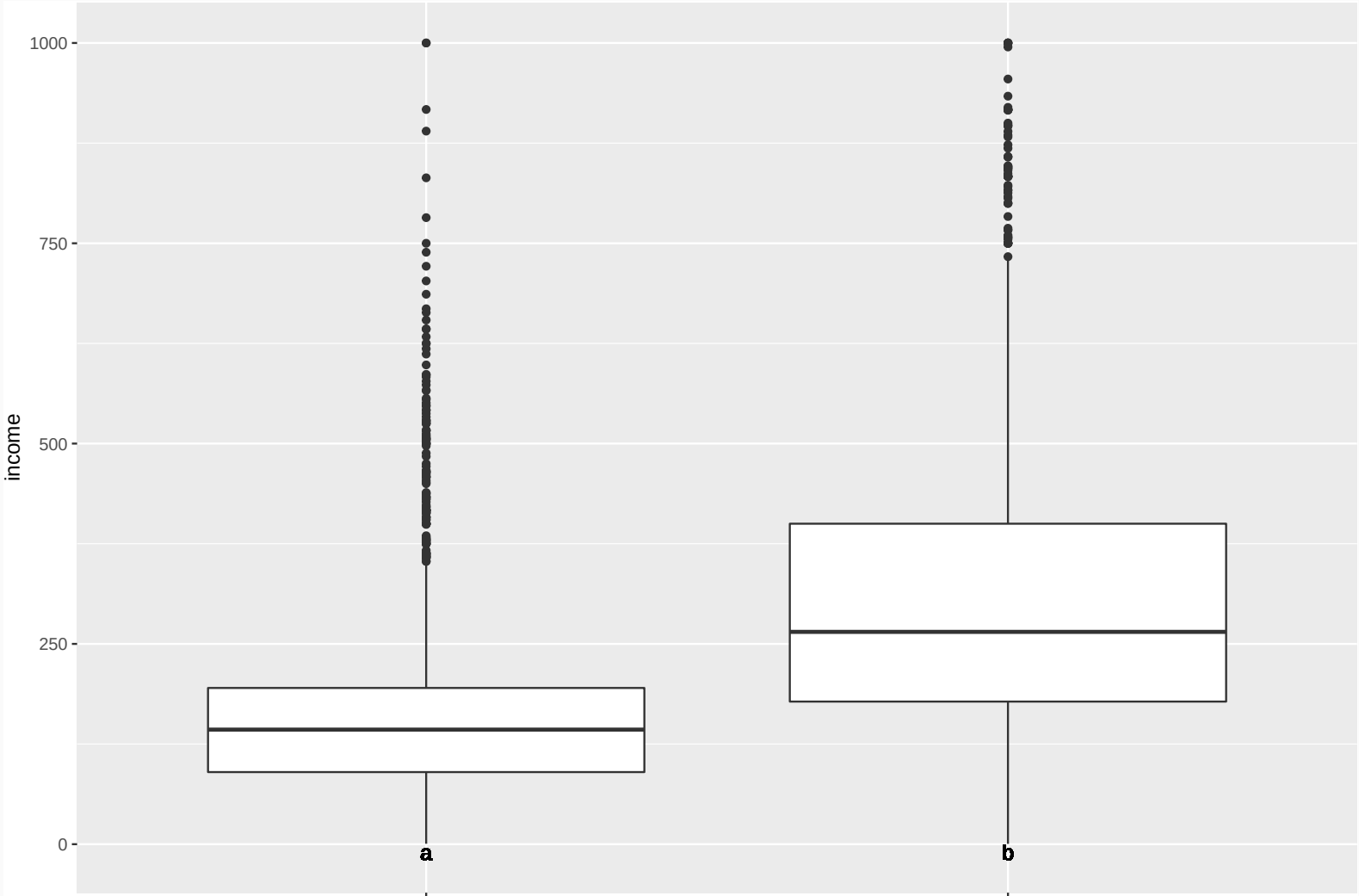
- 통계적 유의성 검증

```
t.test(income ~ sex, data= welfare%>%filter(!is.na(income)))
```

```
##
##  Welch Two Sample t-test
##
## data:  income by sex
## t = -30.792, df = 4204.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -158.5360 -139.5562
## sample estimates:
## mean in group female   mean in group male
##             163.2471             312.2932
```

변수 간 관계 분석 1

```
ggplot( data = welfare %>% filter(!is.na(income)),
        aes(x = as.factor(sex), y = income)) + geom_boxplot() + ylim(-10,100
        geom_text(x=1, y=-10, label="a") + geom_text(x=2, y=-10, label="b")
```



변수 간 관계 분석 2

- 남녀 간 월급 차이는 세대에 따라 개선되고 있을까?
  - 연령 구분을 청년 (30세 미만), 장년 (30 - 59세), 노년 (60세 이상)으로 구분

```
# calculate age (in 2015)
welfare <- welfare %>% mutate(age = 2015 - birth + 1)
# define age_class
welfare <- welfare %>% mutate(age_class =
                                ifelse(age < 30, "young",
                                       ifelse(age <= 59, "middle", "o
```

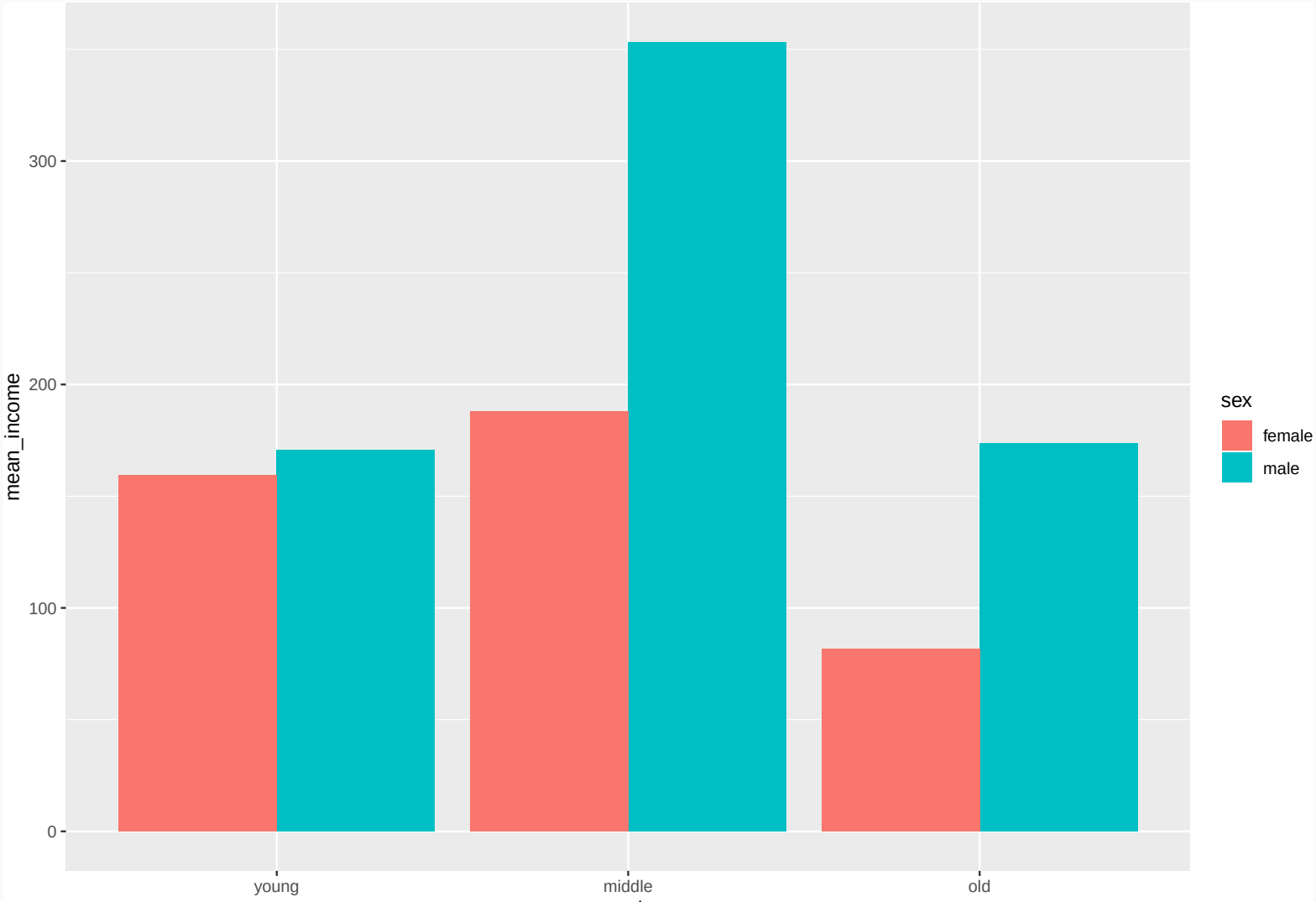
변수 간 관계 분석 2

```
income_sex_age <- welfare %>% filter(!is.na(income)) %>%
                    group_by(age_class, sex) %>% summarize(mean_incom
income_sex_age
```

```
## # A tibble: 6 x 3
## # Groups:   age_class [3]
##   age_class sex    mean_income
##   <chr>     <chr>        <dbl>
## 1 middle    female       188.
## 2 middle    male         353.
## 3 old       female        81.5
## 4 old       male         174.
## 5 young     female       160.
## 6 young     male         171.
```

변수 간 관계 분석 2

```
ggplot(data= income_sex_age, aes(x=age_class, y=mean_income, fill=sex)) +
    geom_col(position="dodge") +
    scale_x_discrete(limits = c("young", "middle", "old"))
```



변수 간 관계 분석 2

- 통계적 분석
  - 각 연령대별로 남여 임금에 차이가 있는지 분석.

```
# young class
t.test(income ~ sex, data=welfare %>% filter(!is.na(income) & age_cla
# middle class
t.test(income ~ sex, data=welfare %>% filter(!is.na(income) & age_cla
# old class
t.test(income ~ sex, data=welfare %>% filter(!is.na(income) & age_cla
```

변수 간 관계 분석 2

```r
# young
ggplot(data = welfare %>% filter(!is.na(income) & age_class == "young"),
       aes(x = sex, y = income) + geom_boxplot()
# middle
ggplot(data = welfare %>% filter(!is.na(income) & age_class == "middle"),
       aes(x = sex, y = income) + geom_boxplot()
# old
ggplot(data = welfare %>% filter(!is.na(income) & age_class == "old"),
       aes(x = sex, y = income) + geom_boxplot()
```

변수 간 관계 분석 2

- 그럼 나이 및 성별에 따른 월급 변화는?

```
income_sex_age_2 <- welfare %>% filter(!is.na(income)) %>%
                    group_by(age, sex) %>%
                    summarize(mean_income  = mean(income))
head(income_sex_age_2)
```

```
## # A tibble: 6 x 3
## # Groups:   age [3]
##     age sex    mean_income
##   <dbl> <chr>        <dbl>
## 1    20 female        147.
## 2    20 male           69
## 3    21 female        107.
## 4    21 male          102.
## 5    22 female        140.
## 6    22 male          118.
```

변수 간 관계 분석 2

```
ggplot(data=income_sex_age_2, aes(x = age, y = mean_income, col = sex)) +
    geom_line()
```