Q1

rhadoop

2019년 3월 27일

#### 데이터 LOAD

h11\_g3 성별(sex)

h11\_g4 태어난 연도(birth)

h11\_g10 혼인상태(marriage)

h11\_g11 종교(religion)

h11\_eco9 직종(code\_job)

p1102\_8aq1 일한달의 월 평균 임금(income)

h11\_reg7 7개 권역별 지역구분(code\_region)

library(tidyverse)

## ─ Attaching packages ──────────────────────────────────────── tidyverse 1.2.1 ─

## ✔ ggplot2 3.1.0 ✔ purrr 0.3.2   
## ✔ tibble 2.1.1 ✔ dplyr 0.8.0.1  
## ✔ tidyr 0.8.3 ✔ stringr 1.4.0   
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ─ Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ─  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(foreign)  
  
read.dta("koweps/Koweps\_hpwc11\_2016\_beta2.dta") %>%   
 rename(sex=h11\_g3) %>%   
 rename(birth=h11\_g4) %>%   
 rename(marriage=h11\_g10) %>%   
 rename(income=p1102\_8aq1) %>%  
 rename(religion=h11\_g11) %>%   
 rename(code\_job=h11\_eco9) %>%   
 rename(code\_region=h11\_reg7) ->welfare

### 데이터 전처리

#### 1.남/여 데이터확인

table(welfaresex = ifelse(welfare$sex==1, "male", "female") table(is.na(welfare$sex))

#### 2. income 확인

welfare %>% filter(!is.na(income)) %>% select(income) %>% arrange(income) %>% head(20)

welfare %>% filter(!is.na(income)) %>% select(income) %>% arrange(desc(income)) %>% head(20)

#### 3. extream value NA 처리

welfare$income =ifelse(welfare$income==0, NA, welfare$income)

welfare %<>% mutate(sex=ifelse(sex==1, "male","female")) %>%   
 mutate(income=ifelse(income==0, NA, income)) %>%   
 mutate(age = 2016-welfare$birth+1) -> welfare

#### 4. 데이터 저장

save(welfare,file="koweps/welfare\_2016.rda")

### 성별에 따른 월급차이(incomebysex) - 성별에 따라 월급이 다를까?

load("koweps/welfare\_2016.rda")

#### 1.성별 월급 평균기

welfare %>% group\_by(sex) %>%  
 summarise(mean\_income=mean(income,na.rm = T))->incomebysex

#### 2. 그래프 그리기

ggplot(data = incomebysex, aes(x = sex, y = mean\_income)) + geom\_col()

