Q1

rhadoop

2019년 3월 27일

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)  
library(plotly)

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

# 

# 성별에 따른 월급 차이 (incomebysex)

read.dta("koweps/Koweps\_hpwc11\_2016\_beta2.dta") %>%  
 rename(sex = h11\_g3) %>%   
 rename(birth = h11\_g4) %>%   
 rename(marriage = h11\_g10) %>%   
 rename(religion = h11\_g11) %>%   
 rename(code\_job = h11\_eco9) %>%   
 rename(income = p1102\_8aq1) %>%   
 rename(code\_region = h11\_reg7) -> welfare   
  
  
welfare %<>% mutate(sex=ifelse(sex==1, "male", "female")) %>%   
 mutate(income=ifelse(income==0, NA, income)) %>%   
 select(sex,   
 birth,   
 marriage,   
 religion,   
 income,   
 code\_job,   
 code\_region)   
  
save(welfare, file="koweps/welfare.rda")

load("koweps/welfare.rda")

# 

### 변수: sex

##### 1) NA 확인

table(is.na(welfare$sex))

##   
## FALSE   
## 15989

## 

##### 2) outlier, 빈도 확인

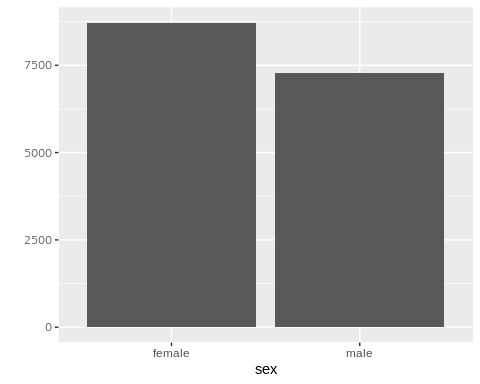
table(welfare$sex)

##   
## female male   
## 8724 7265

## 

##### 3) 빈도 막대 그래프

qplot(sex, data=welfare)



# 

### 변수: income

##### 1) NA 확인

table(is.na(welfare$income))

##   
## FALSE TRUE   
## 4528 11461

##### 2) 이상치 확인

welfare %>% filter(!is.na(income)) %>% select(income) %>% arrange(income) %>% head(10)

## income  
## 1 2.0  
## 2 8.0  
## 3 10.0  
## 4 10.0  
## 5 10.0  
## 6 12.0  
## 7 15.0  
## 8 16.2  
## 9 17.0  
## 10 18.0

welfare %>% filter(!is.na(income)) %>% select(income) %>% arrange(desc(income)) %>% head(10)

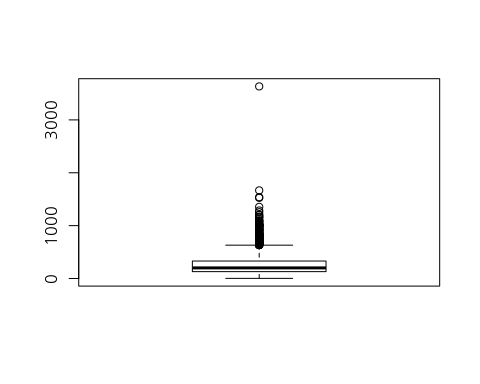
## income  
## 1 3633.30  
## 2 1666.60  
## 3 1538.00  
## 4 1525.00  
## 5 1354.75  
## 6 1286.30  
## 7 1254.50  
## 8 1208.30  
## 9 1200.00  
## 10 1168.00

# 

##### 3) boxplot

##### r base

boxplot(welfare$income)

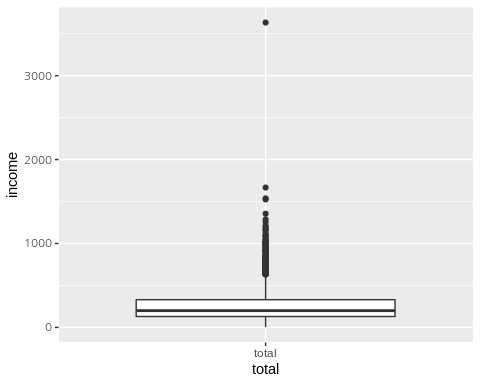


## 

##### qplot

qplot(x="total",y=income,data=welfare,geom="boxplot")

## Warning: Removed 11461 rows containing non-finite values (stat\_boxplot).

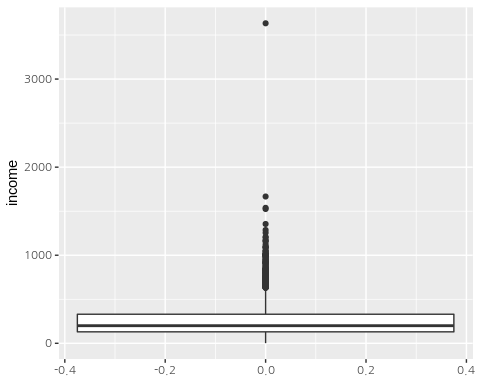


## 

##### ggplot

ggplot(welfare,aes(y=income))+geom\_boxplot()

## Warning: Removed 11461 rows containing non-finite values (stat\_boxplot).

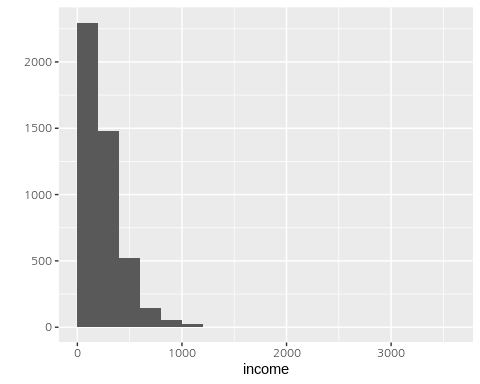


# 

##### 4) histogram - 200만원 단위

qplot(income,data=welfare,breaks=seq(0,3600, by=200))

## Warning: Removed 11461 rows containing non-finite values (stat\_bin).



# 

# 

# 

# income by sex

### SQL

sqldf("select sex, avg(income) as mean from welfare group by sex")

## sex mean  
## 1 female 170.6918  
## 2 male 324.0590

# 

### dplyr

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

