Q2

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(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

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

# 

# 나이와 월급의 관계 incomebyage

load("koweps/welfare.rda")

# 

### 변수: birth

##### 1) NA 확인

table(is.na(welfare$birth))

##   
## FALSE   
## 15989

## 

##### 2) outlier 확인

welfare[order(welfare$birth), 'birth'] %>% head

## [1] 1907 1911 1914 1919 1919 1919

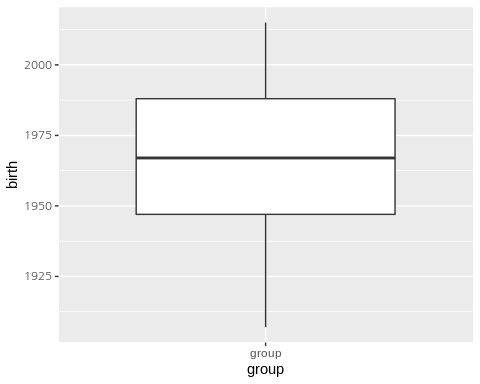
welfare[order(-welfare$birth), 'birth'] %>% head

## [1] 2015 2015 2015 2015 2015 2015

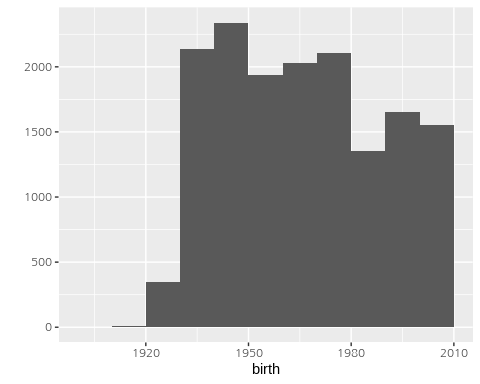
## 

##### 3) boxplot, histogram

qplot(x="group",y=birth, data=welfare, geom="boxplot")



qplot(x=birth, data=welfare, geom="histogram", breaks=seq(1900,2015,by=10))



## 

## 

## age 파생변수 만들기

welfare$age = 2016-welfare$birth+1  
summary(welfare$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.00 29.00 50.00 48.95 70.00 110.00

table(is.na(welfare$age))

##   
## FALSE   
## 15989

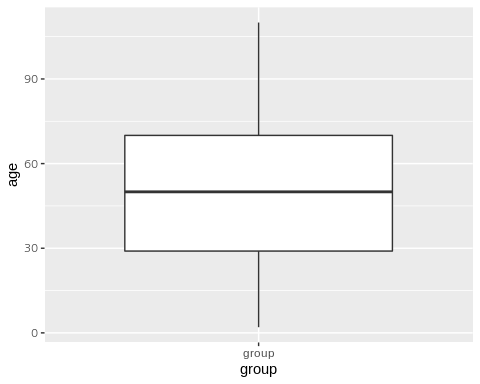
welfare[order(welfare$age), 'age'] %>% head

## [1] 2 2 2 2 2 2

welfare[order(-welfare$age), 'age'] %>% head

## [1] 110 106 103 98 98 98

qplot(x="group",y=age, data=welfare, geom="boxplot")



qplot(x=age, data=welfare, geom="histogram", breaks=seq(0,110,by=10))



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

# 

# income by age

load("koweps/welfare.rda")

## 

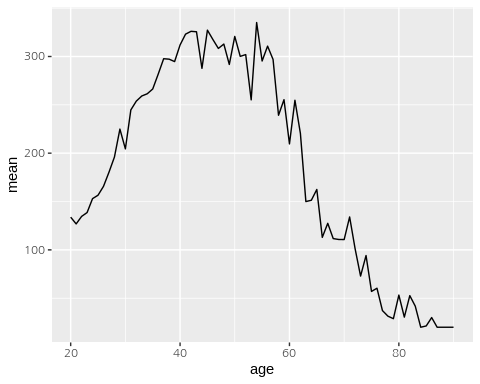
##### ㄱ. 나이별 평균월급표 요약표 작성

welfare %>% filter(!is.na(income)) %>%   
 group\_by(age) %>%   
 summarise(mean=mean(income)) -> incomebyage

## 

##### ㄴ. 라인차트 그리기

ggplot(incomebyage,aes(age,mean))+geom\_line()



ggplot(incomebyage,aes(age,mean))+geom\_point()

