경제 활동 조사 보고서 분석 및 시각화

오정우 201513408

2019 12 18

knitr::opts\_chunk$set(echo = TRUE)  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## 경제 활동 조사 보고서 분석 및 시각화

raw = read.csv("raw\_welfare.csv")

### 데이터 검토

str(raw)

## 'data.frame': 16664 obs. of 958 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ h10\_id : int 1 2 3 4 4 6 6 6 6 6 ...  
## $ h10\_ind : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ h10\_sn : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ h10\_merkey : int 10101 20101 30101 40101 40101 60101 60101 60101 60101 60101 ...  
## $ h\_new : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ h10\_cobf : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h10\_reg5 : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ h10\_reg7 : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ h10\_din : num 864 600 1571 3579 3579 ...  
## $ h10\_cin : num 864 600 1619 3687 3687 ...  
## $ h10\_flag : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ p10\_wgl : num 777 960 1059 1012 1075 ...  
## $ p10\_wsl : num 0.257 0.317 0.35 0.334 0.355 ...  
## $ p10\_wgc : num 764 949 1048 992 1057 ...  
## $ p10\_wsc : num 0.252 0.314 0.346 0.328 0.349 ...  
## $ h10\_hc : int 2 2 1 1 1 1 1 1 1 1 ...  
## $ nh1001\_1 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ nh1001\_2 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h1001\_1 : int 1 1 1 2 2 5 5 5 5 5 ...  
## $ h10\_pind : int 1 1 1 1 4 1 1 1 1 1 ...  
## $ h10\_pid : int 101 201 301 401 402 601 602 603 604 605 ...  
## $ h10\_g1 : int 1 1 1 1 2 1 2 3 4 5 ...  
## $ h10\_g2 : int 10 10 10 10 2 10 20 11 1 2 ...  
## $ h10\_g3 : int 2 2 1 1 2 1 2 2 1 2 ...  
## $ h10\_g4 : int 1936 1945 1948 1942 1923 1962 1963 2003 1927 1934 ...  
## $ h10\_g6 : int 2 4 3 7 2 6 5 3 4 4 ...  
## $ h10\_g7 : int 0 5 5 3 0 5 5 1 5 5 ...  
## $ h10\_g8 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ h10\_g9 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ h10\_g10 : int 2 2 2 3 2 1 1 0 1 1 ...  
## $ h10\_g11 : int 2 2 2 1 1 1 1 1 1 1 ...  
## $ h10\_g12 : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ h1001\_110 : int 1 1 1 5 5 5 5 5 5 5 ...  
## $ h1001\_5aq1 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ h1001\_5aq2 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ h1001\_5aq3 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ h1001\_5aq4 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ h10\_med1 : int 1 1 1 1 2 1 2 3 4 5 ...  
## $ h10\_med2 : int 3 4 3 3 4 3 5 2 3 3 ...  
## $ h10\_med3 : int 60 28 12 3 6 5 0 3 0 14 ...  
## $ h10\_med4 : int 0 0 0 0 0 0 5 0 0 0 ...  
## $ h10\_med5 : int 0 0 0 0 0 0 23 0 0 0 ...  
## $ h10\_med6 : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ h10\_med7 : int 3 2 2 1 1 2 1 2 0 2 ...  
## $ h10\_med8 : int 0 1 0 1 1 1 1 0 0 1 ...  
## $ h10\_g9\_1 : int 3 3 3 3 3 3 3 0 0 3 ...  
## $ h10\_med9 : int 8 5 7 3 15 23 1 0 0 6 ...  
## $ h10\_med10 : int 0 0 0 0 0 1 1 1 0 0 ...  
## $ h10\_eco1 : int 1 1 1 1 2 1 2 3 4 5 ...  
## $ h10\_eco2 : int 2 3 1 1 3 1 1 0 3 3 ...  
## $ h10\_eco3 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h10\_eco4 : int 9 9 2 2 9 6 9 NA 9 9 ...  
## $ h10\_eco4\_1 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h10\_eco5\_1 : int NA NA 1 1 NA NA NA NA NA NA ...  
## $ h10\_eco6 : int NA NA 2 2 NA NA NA NA NA NA ...  
## $ h10\_eco\_7\_1 : int NA NA 1 1 NA NA NA NA NA NA ...  
## $ h10\_eco\_7\_2 : int NA NA 2 2 NA NA NA NA NA NA ...  
## $ h10\_eco\_7\_3 : int NA NA 1 3 NA NA NA NA NA NA ...  
## $ h10\_eco8 : int NA NA 75 42 NA 46 NA NA NA NA ...  
## $ h10\_eco9 : int NA NA 942 762 NA 530 NA NA NA NA ...  
## $ h10\_eco10 : int NA NA 3 2 NA 1 NA NA NA NA ...  
## $ h10\_eco11 : int 10 10 NA NA 10 NA 6 NA 10 10 ...  
## $ h10\_soc1 : int 1 1 1 1 2 1 2 3 4 5 ...  
## $ h10\_soc\_2 : int 0 0 0 1 0 2 0 0 0 0 ...  
## $ h10\_soc\_3 : int NA NA NA NA NA 1 NA NA NA NA ...  
## $ h10\_soc\_4 : int NA NA NA NA NA 2 NA NA NA NA ...  
## $ h10\_soc\_5 : int NA NA NA NA NA 1 NA NA NA NA ...  
## $ h10\_soc\_6 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h10\_soc\_7 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h10\_soc\_8 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h10\_soc\_9 : int NA NA NA NA NA 0 NA NA NA NA ...  
## $ h10\_soc\_10 : int NA NA NA NA NA 0 NA NA NA NA ...  
## $ h10\_soc\_11 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h10\_soc8 : int 0 0 1 2 0 0 0 0 0 0 ...  
## $ h10\_soc9 : int 0 0 1 2 0 0 0 0 0 0 ...  
## $ h10\_soc11 : int 0 0 2 2 0 0 0 0 0 0 ...  
## $ h10\_soc10 : int 0 0 2 2 0 0 0 0 0 0 ...  
## $ h10\_soc\_12 : int 4 4 4 4 4 4 4 4 4 4 ...  
## $ h10\_soc\_13 : int 4 4 1 3 4 3 4 4 4 4 ...  
## $ h1005\_1 : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ h1005\_3aq1 : int 2 2 1 2 2 2 2 2 2 2 ...  
## $ h1005\_2 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h1005\_3 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h1005\_4 : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ h1005\_5 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h1005\_6 : int NA NA NA NA NA NA NA NA NA NA ...  
## $ h1005\_7 : int 0 0 0 0 0 1 1 1 1 1 ...  
## $ nh1005\_8 : int 4 3 3 3 3 3 3 3 3 3 ...  
## $ nh1005\_9 : int 5 4 3 3 3 4 4 4 4 4 ...  
## $ h1005\_3aq2 : num 0 0 0 0 0 11 11 11 11 11 ...  
## $ h1006\_aq1 : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ h1006\_1 : int 2 2 2 1 1 2 2 2 2 2 ...  
## $ h1006\_2 : int 3 3 1 3 3 3 3 3 3 3 ...  
## $ h1006\_4 : int 2 3 1 3 3 3 3 3 3 3 ...  
## $ h1006\_5 : int 33 198 23 73 73 82 82 82 82 82 ...  
## $ h1006\_3 : int 2 1 3 1 1 1 1 1 1 1 ...  
## $ h1006\_6 : num 5000 60000 200 20000 20000 50700 50700 50700 50700 50700 ...  
## $ h1006\_8 : int 1 1 1 1 1 1 1 1 1 1 ...  
## [list output truncated]

### 변수명 변경

###### 데이터 컬럼이 너무 많기 때문에 필요한 컬럼만 추출하고 컬럼명 변경

raw = rename(raw, Gender=h10\_g3,   
 Birth=h10\_g4,   
 Income=p1002\_8aq1,  
 Code\_job=h10\_eco9)  
dat = raw %>% select(Gender, Birth, Income, Code\_job)  
str(dat)

## 'data.frame': 16664 obs. of 4 variables:  
## $ Gender : int 2 2 1 1 2 1 2 2 1 2 ...  
## $ Birth : int 1936 1945 1948 1942 1923 1962 1963 2003 1927 1934 ...  
## $ Income : num NA NA 120 200 NA NA NA NA NA NA ...  
## $ Code\_job: int NA NA 942 762 NA 530 NA NA NA NA ...

str(dat)

## 'data.frame': 16664 obs. of 4 variables:  
## $ Gender : int 2 2 1 1 2 1 2 2 1 2 ...  
## $ Birth : int 1936 1945 1948 1942 1923 1962 1963 2003 1927 1934 ...  
## $ Income : num NA NA 120 200 NA NA NA NA NA NA ...  
## $ Code\_job: int NA NA 942 762 NA 530 NA NA NA NA ...

### 이상치 및 결측치 보정

###### 성별 결측치는 없다.

dat[is.na(dat[,"Gender"]),]

## [1] Gender Birth Income Code\_job  
## <0 rows> (or 0-length row.names)

###### 태어난 연도 결측치도 없다.

dat[is.na(dat[,"Birth"]),]

## [1] Gender Birth Income Code\_job  
## <0 rows> (or 0-length row.names)

###### 수입은 평균으로 결측치를 보정해 주었다.

dat[is.na(dat[,"Income"]),"Income"] = mean(dat$Income, na.rm=T)  
dat[is.na(dat[,"Income"]),]

## [1] Gender Birth Income Code\_job  
## <0 rows> (or 0-length row.names)

###### Koweps\_Codebook.xlsx를 보면 학생의 직종코드가 없다. 따라서 데이터에는 학생, 측 19세 이하는 직종코드가 NA로 나온다. 때문에 모든 학생들을 새로운 학생 코드 9999을 만들어 할당한다.

dat[is.na(dat[,"Code\_job"]),"Code\_job"] = 9999  
dat[is.na(dat[,"Code\_job"]),]

## [1] Gender Birth Income Code\_job  
## <0 rows> (or 0-length row.names)

### 성별 연령대별 수입의 평균 분석

###### 연령대에 따라 Age 컬럼을 지정해준다.

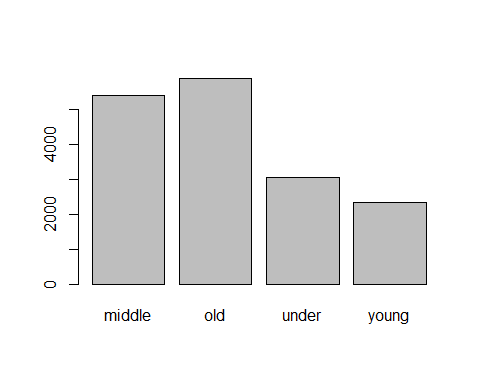
head(dat$Birth)

## [1] 1936 1945 1948 1942 1923 1962

dat$Age = ifelse(dat$Birth >= 1996, "under",  
 ifelse(dat$Birth >= 1980, "young",  
 ifelse(dat$Birth >= 1955, "middle", "old")))  
dat$Age = as.factor(dat$Age)

###### Age컬럼을 팩터로 바꾸고 plot()을 그려본다.

dat$Age = as.factor(dat$Age)  
plot(dat$Age)



위 그래프를 보면 35세 이상의 인구인 middle과 old가 다른 두 집합보다 상대적으로 많은 양을 차지하고 있는 것을 볼 수 있다. 특히 young과 middle은 같은 나이 범위를 갖음에도 불구하고 차이가 확연한 것을 확인할 수 있다. 이는 이 데이터를 가진 집단이 **고령화 사회**로 접어들었을을 짐작해볼 수 있다.

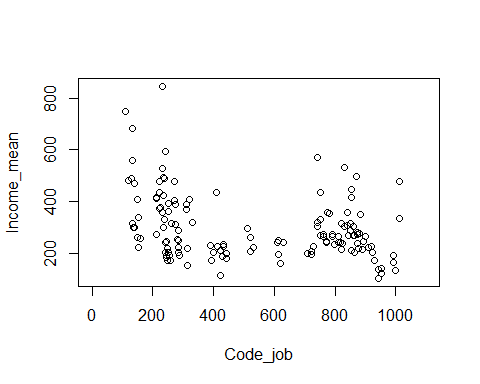
### 직업별 수입 평균을 구하고, 수입이 많은 상위 10개 직업과 하위 10개 직업 탐색

###### 직업별 수입 평균

Income\_mean = dat %>%   
 group\_by(Code\_job) %>%   
 summarise(Income\_mean = mean(Income))  
summary(Income\_mean)

## Code\_job Income\_mean   
## Min. : 111.0 Min. :103.8   
## 1st Qu.: 248.8 1st Qu.:214.0   
## Median : 515.5 Median :265.4   
## Mean : 605.0 Mean :298.4   
## 3rd Qu.: 838.8 3rd Qu.:357.7   
## Max. :9999.0 Max. :845.1

plot(Income\_mean, xlim=c(0, 1100))



###### 수입 상위 10개 직업

Income\_mean %>%   
 arrange(desc(Income\_mean)) %>%   
 head(10)

## # A tibble: 10 x 2  
## Code\_job Income\_mean  
## <dbl> <dbl>  
## 1 233 845.  
## 2 111 750   
## 3 132 682.  
## 4 241 593.  
## 5 742 572.  
## 6 134 557.  
## 7 831 532.  
## 8 232 528.  
## 9 871 496.  
## 10 235 492.

수입이 제일 많은 직업은 금속 재료 공학 기술자 및 시험원이다.

###### 수입 하위 10개 직업

Income\_mean %>%   
 arrange(Income\_mean) %>%   
 head(10)

## # A tibble: 10 x 2  
## Code\_job Income\_mean  
## <dbl> <dbl>  
## 1 941 104.  
## 2 421 114.  
## 3 951 124.  
## 4 999 134.  
## 5 942 138.  
## 6 952 141.  
## 7 953 141.  
## 8 314 155.  
## 9 620 162.  
## 10 992 163.

수입이 제일 적은 직업은 청소원 및 환경 미화원이다.