

데이터 분석 프로젝트



분석 준비하기

이해하기

❖ 한국복지 패널데이터

- 한국보건사회연구원 발간
- 가구의 경제활동을 연구해 정책 지원에 반영할 목적
- 2006~2015년까지 전국에서 7000여 가구를 선정해 매년 추적 조사
- 경제활동, 생활실태, 복지욕구 등 수천 개 변수에 대한 정보로 구성

❖ SPSS

- 통계 분석 프로그램

데이터 준비하기

❖ <https://www.koweps.re.kr:442/data/data/list.do>

The screenshot shows the KOWEPS website interface. At the top, there is a navigation bar with the KOWEPS logo and the text '한국복지패널' (Korea Welfare Panel Study) and '조사설계' (Survey Design). Below the navigation bar is a banner image featuring a book with the KOWEPS logo and a rainbow. To the left of the banner is a sidebar menu with the following items: '데이터 & 설문지' (Data & Questionnaire), '데이터' (Data), '설문지' (Questionnaire), '유저가이드' (User Guide), and '코딩북' (Coding Book). The main content area is titled '14차 웨이브' (14th Wave) and contains a list of data download links for the 2019 14th Korea Welfare Panel Survey. The links are: (2019년 14차 한국복지패널조사) 데이터 (beta1)_stata.zip, (2019년 14차 한국복지패널조사) 데이터 (beta1)_spss.zip, and (2019년 14차 한국복지패널조사) 데이터 (beta1)_sas.zip. The second link, (2019년 14차 한국복지패널조사) 데이터 (beta1)_spss.zip, is highlighted with an orange border.

한국복지패널
Korea Welfare Panel Study

한국복지패널
Korea We

데이터 & 설문지

데이터

설문지

유저가이드

코딩북

14차 웨이브

- (2019년 14차 한국복지패널조사) 데이터 (beta1)_stata.zip
- (2019년 14차 한국복지패널조사) 데이터 (beta1)_spss.zip
- (2019년 14차 한국복지패널조사) 데이터 (beta1)_sas.zip

분석

❖ 분석 주제

- 성별에 따른 월급 차이
- 나이와 월급의 관계
- 연령대에 따른 월급 차이
- 연령대 및 성별 월급 차이
- 직업별 월급 차이
- 성별 직업 빈도
- 종교 유무에 따른 이혼율
- 지역별 연령대 비율

❖ 필요 데이터

- 성별, 나이, 월급, 직업, 종교, 지역, 혼인, 직업, 지역

데이터 분석 준비하기

❖ 패키지 준비하기

- `install.packages("foreign")` # foreign 패키지 설치
- `library(foreign)` # SPSS 파일 로드
- `library(dplyr)` # 전처리
- `library(ggplot2)` # 시각화
- `library(readxl)` # 엑셀 파일 불러오기

❖ 데이터 준비하기

- # 데이터 불러오기
- `raw_welfare <- read.spss(file = "Koweps_hpc10_2015_beta1.sav",`
▪ `to.data.frame = T)`
- # 복사본 만들기
- `welfare <- raw_welfare`

조사 설계서를 이용한 제목 변경



데이터&설문지

데이터

설문지

유저가이드

코딩북

10차 웨이브



- (2015년 10차 한국복지패널조사) 조사설계서(beta5).zip

10차 - 가구용, 가구원용, 아동 머지 데이터 파일	
구분	변수명
개인 일반 가중치(모수추정, 종단면분석)_추가표본 미포함	p10_wgl
5개 권역별 지역구분	h10_reg5
7개 권역별 지역구분	h10_reg7
가처분소득	h10_din
경상소득	h10_cin
균등화소득에 따른 가구구분_추가표본 미포함	h10_hc
균등화소득에 따른 가구구분_추가표본 포함	h10_hc_all
가구원진입차수	h10_pind
개인 패널 ID	h10_pid
가구원 번호	h10_g1
가구주와의 관계	h10_g2
성별	h10_g3
태어난 연도	h10_g4
교육수준1	h10_g6
교육수준2	h10_g7
장애종류	h10_g8
장애등급	h10_g9
-	
혼인상태	h10_g10
종교	h10_g11
동거여부	h10_g12

변수명 변경

```
welfare <- rename(welfare,  
  sex = h10_g3,          # 성별  
  birth = h10_g4,        # 태어난 연도  
  marriage = h10_g10,    # 혼인 상태  
  religion = h10_g11,    # 종교  
  income = p1002_8aq1,   # 월급  
  code_job = h10_eco9,    # 직종 코드  
  code_region = h10_reg7) # 지역 코드
```


데이터 분석 절차

1단계 변수 검토 및 전처리

전처리

sex	income		sex	income
2	270		2	270
3	210	→	1	350
1	350		1	430
2	0		2	320
1	430			
2	320			

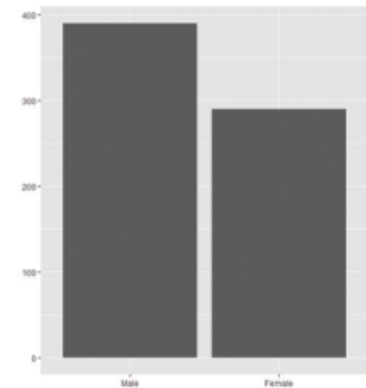


2단계 변수 간 관계 분석

요약표 만들기

sex	income
1	390
2	295

그래프 만들기





성별에 따른 월급 차이



분석 절차

1. 변수 검토 및 전처리

- 성별
- 월급

2. 변수 간 관계 분석

- 성별 월급 평균표 만들기
- 그래프 만들기

성별 변수 검토 및 전처리

#변수 종류 확인

```
class(welfare$sex)
```

#이상치 확인

```
table(welfare$sex)
```

이상치 결측 처리

```
welfare$sex <- ifelse(welfare$sex == 5, NA, welfare$sex)
```

결측치 확인

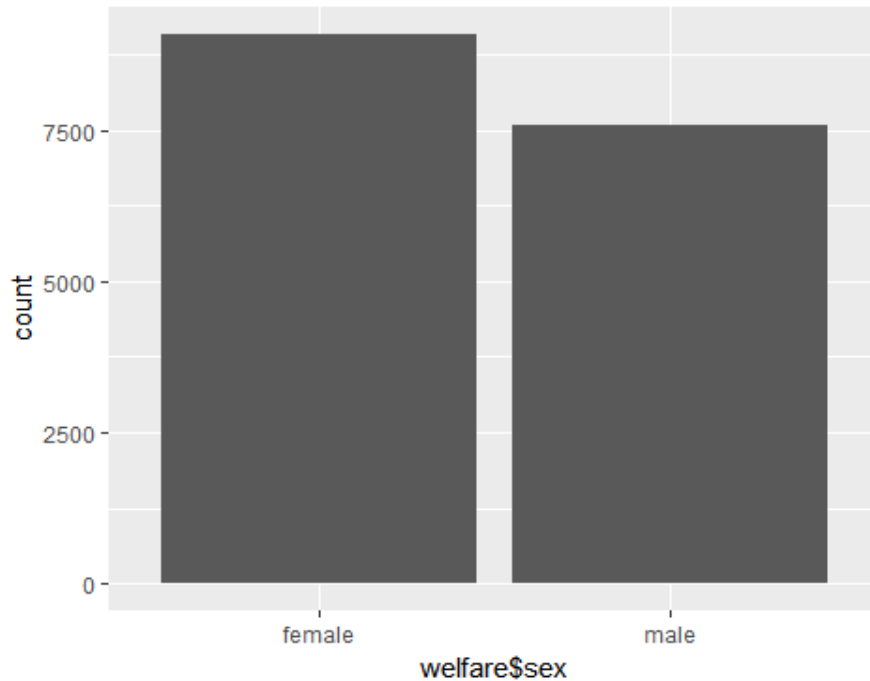
```
table(is.na(welfare$sex))
```

성별 변경

성별 항목 이름 부여

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

```
qplot(welfare$sex)
```



월급 변수 처리

#자료형 확인

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

#상태 확인

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

```
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

실행 결과 결측치 발생

잘못된 결측치 적용

```
test.welfare = welfare
```

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

```
attach(test.welfare)
```

```
class(income)
```

```
summary(income)
```

```
na.mean = mean(income, na.rm = T)
```

```
income = ifelse(is.na(income), na.mean, income)
```

```
summary(income)
```

```
detach(test.welfare)
```

```
summary(income)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	241.6	241.6	241.6	241.6	2400.0

너무나 많은 결측치를
평균을 적용하녀 4분위수가 망가짐

이상치 제거

#이상치 확인

```
attach(welfare)
```

```
summary(income)
```

```
boxplot(income)
```

```
boxplot(income)$stat
```

```
detach(welfare)
```

#이상치 변경

```
welfare$income = ifelse(welfare$income==0 | welfare$income>608, NA,  
welfare$income)
```

#결과 확인

```
welfare$income
```

결측치 확인

```
table(is.na(welfare$income))
```


성별에 따른 월급 차이 분석

#성별에 따른 월급

```
sex_income <- welfare %>%  
  filter(!is.na(income)) %>%  
  group_by(sex) %>%  
  summarise(mean_income = mean(income))
```

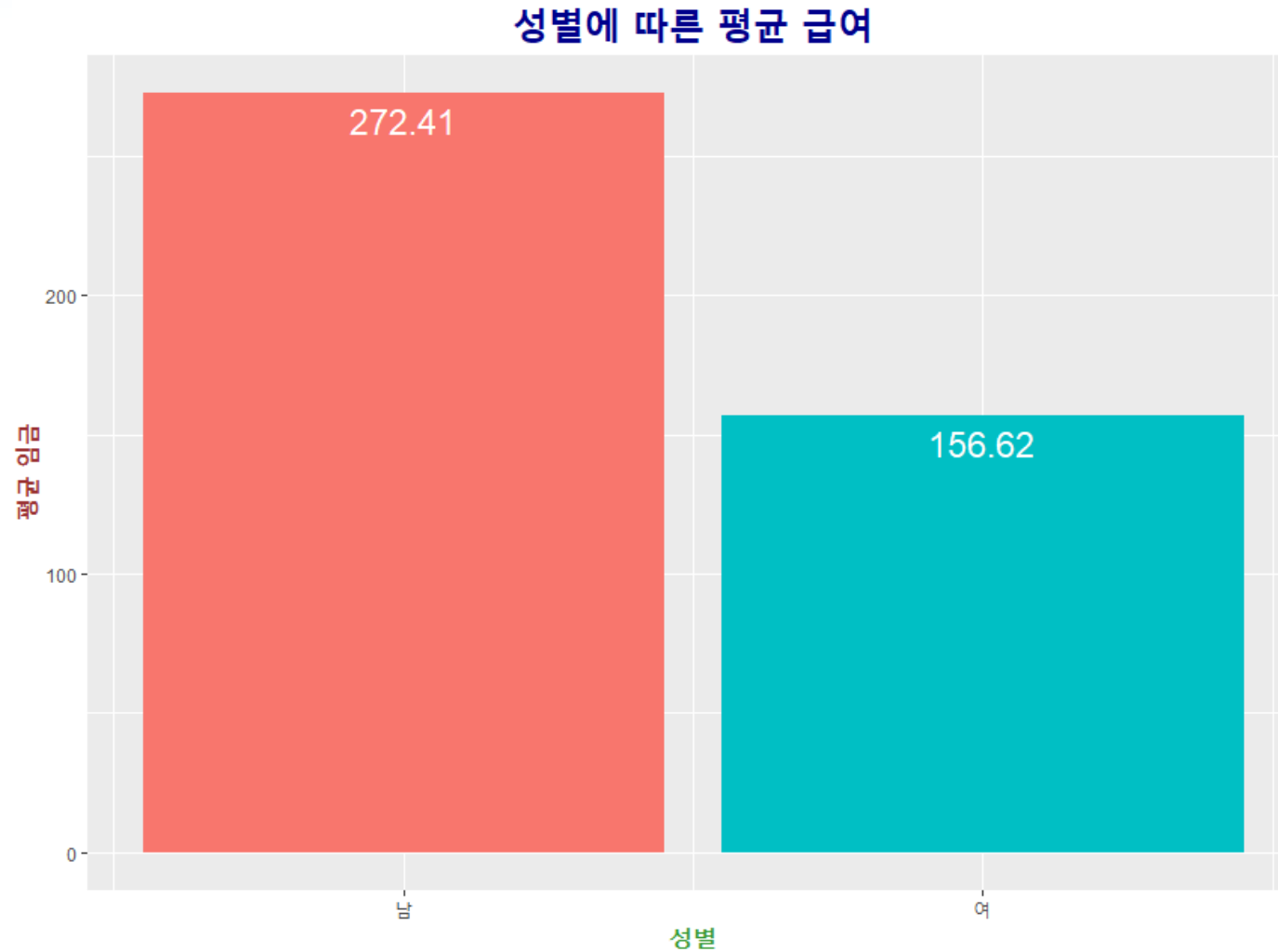
#결과 확인

```
sex_income
```

#막대 그래프

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

Quiz : 다음과 같이 그래프를 수정하시오



나이와 월급의 관계

분석 절차

1. 변수 검토 및 전처리

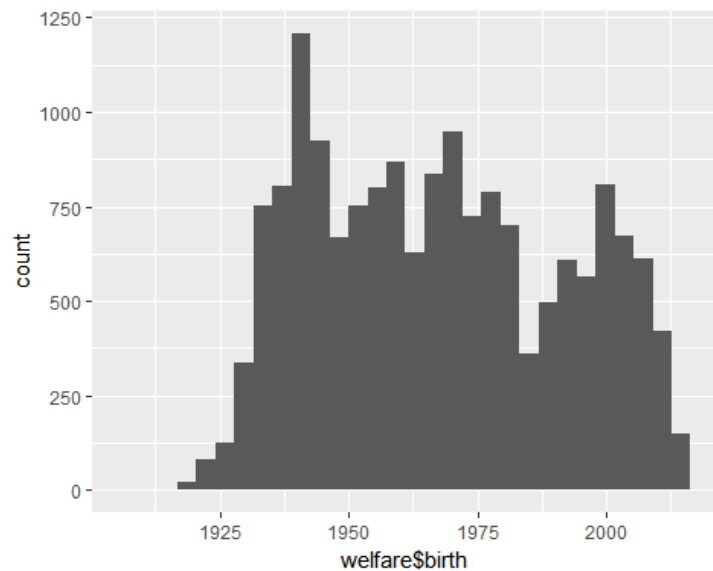
- 나이
- 월급

2. 변수 간 관계 분석

- 나이에 따른 월급 평균표 만들기
- 그래프 만들기

검토하기

```
class(welfare$birth)
## [1] "numeric"
summary(welfare$birth)
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1907  1946   1966   1968   1988   2014
qplot(welfare$birth)
```



전처리

결측치 확인

```
table(is.na(welfare$birth))
```

```
##
```

```
## FALSE
```

```
## 16664
```

이상치 결측 처리

```
welfare$birth <- ifelse(welfare$birth == 9999, NA, welfare$birth)
```

```
table(is.na(welfare$birth))
```

```
##
```

```
## FALSE
```

```
## 16664
```

파생변수 만들기

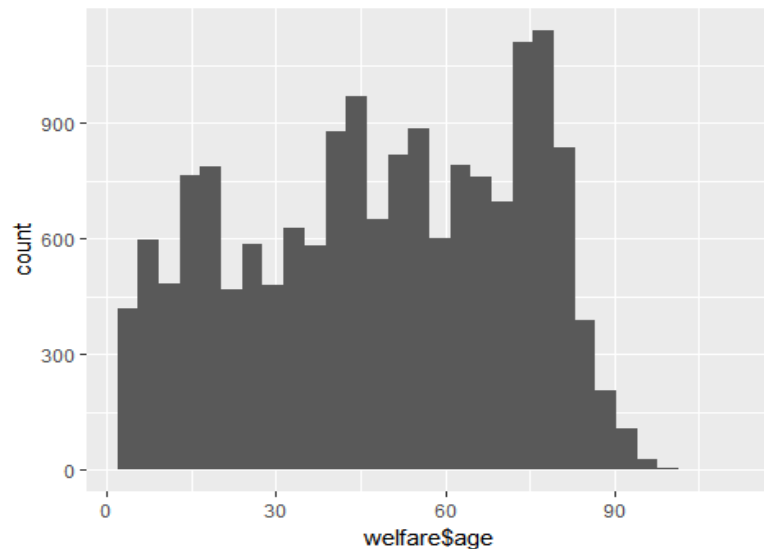
```
welfare$age <- 2015 - welfare$birth + 1
```

```
summary(welfare$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    2.00  28.00   50.00   48.43   70.00  109.00
```

```
qplot(welfare$age)
```



나이와 급여의 관계 분석

```
age_income <- welfare %>%  
  filter(!is.na(income)) %>%  
  group_by(age) %>%  
  summarise(mean_income = mean(income))
```

```
head(age_income)
```

```
## # A tibble: 6 x 2  
##   age mean_income  
##   <dbl>     <dbl>  
## 1    20    121.3000  
## 2    21    105.5185  
## 3    22    130.0923  
## 4    23    141.7157  
## 5    24    134.0877  
## 6    25    144.6559
```


그래프 만들기

```
ggplot(data = age_income, aes(x = age, y = mean_income)) + geom_line()
```

