# 데이터분석 프로젝트

수학과 2018103585
김서연

# 목차

- 1. 지역별 평균 월급 – 지역에 따라 평균 월급은 어떻게 다를까?
- 2. 성별에 따른 종교유무 – 어떤 성별이 종교를 많이 가질까?

# 데이터 분석 준비하기

```
#패키지 설치&불러오기
Install.pacakages("foreign")
library(foreign)
library(dplyr)
library(readxl)
library(ggplot2)

#데이터 준비하기
raw_welfare <- read.spss(file="C:/r_data/easy_r/Koweps_hpc10_2015_beta1.sav",
to.data.frame=T) # to.data.frame은 데이터를 리스트가 아닌 데이터프레임 형태로 변환

#복사본 준비
welfare <- raw_welfare
```

# 변수명 바꾸기

```
welfare <- rename(welfare,
                  sex=h10_g3,
                  birth=h10_g4,
                  marriage=h10_g10,
                  religion=h10_g11,
                  income=p1002_8aq1,
                  code_job=h10_eco9,
                  code_region=h10_reg7)
```

# 1. 지역별 평균 월급

Q. 지역에 따라 평균 월급은 어떻게 다를까?

1)지역 변수 검토

```
class(welfare$code_region) #numeric
table(welfare$code_region)
```

```
class(welfare$income) #numeric
summary(welfare$income)
```

# 2. 전처리

```
#이상치 결측 처리
welfare$code_region <- ifelse(welfare$code_region ==9999, NA, welfare$code_region)

table(is.na(welfare$code_region)) #False만 나옴, 즉 결측치 없음.

#지역코드 목록 만들기
list_region <- data.frame(code_region = c(1:7),
                          region = c("서울","수도권(인천/경기)",
                                     "부산/경남/울산",
                                     "대구/경북",
                                     "대전/충남",
                                     "강원/충북",
                                     "광주/전남/전북/제주도"))
list_region
```

# 2.전처리

#이상치 결측처리

welfare$income <- ifelse(welfare$income %in% c(0,9999), NA, welfare$income)

#결측치 확인

table(is.na(welfare$income)) #False 4620, True 12044 . 결측치 존재.

# 3. 지역명 변수 추가

```
welfare <- left_join(welfare, list_region, id="code_region") #code_region을 기준으로 합침

welfare %>%
  filter(!is.na(code_region)) %>%
  select(code_region, region) %>%
  head(10)

region_income <- welfare %>%
  filter(!is.na(income)) %>%  #income에만 결측치 존재
  group_by(region) %>%
  summarise(mean_income=mean(income))
```

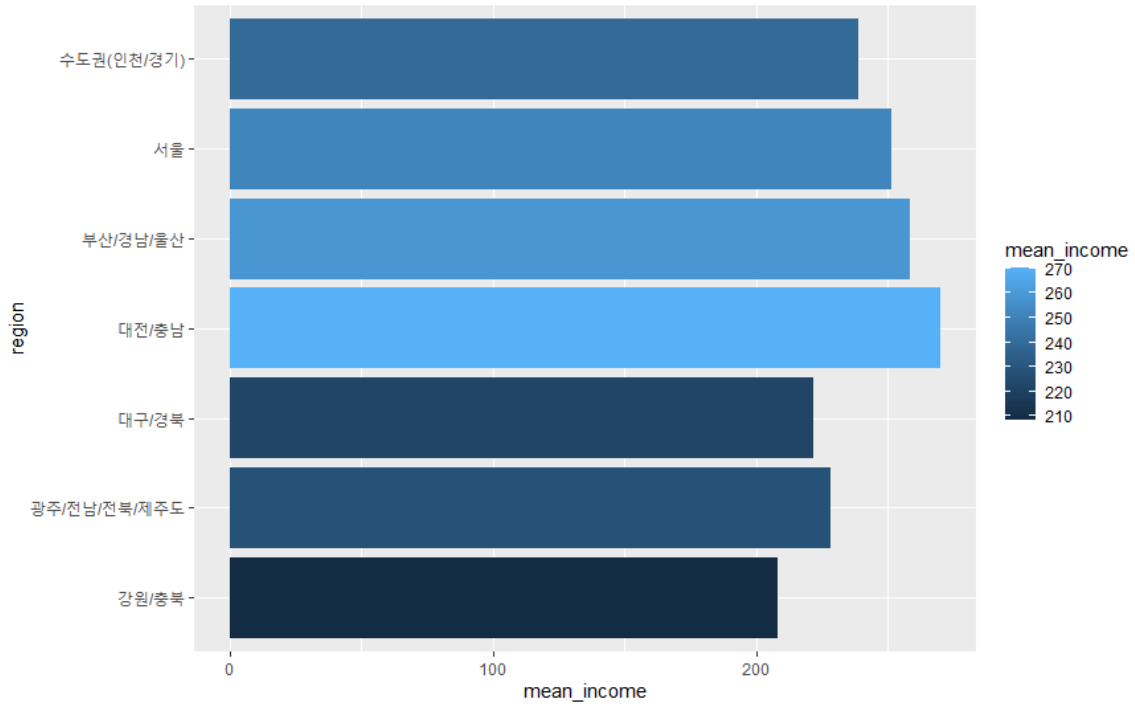
# 4. 그래프 만들기

ggplot(data = region_income, aes(x=region, y=mean_income, fill=mean_income))

+geom_col() + coord_flip()

#막대그래프를 오른쪽으로 90도 회전

#월급에 따라 다른색으로 표현



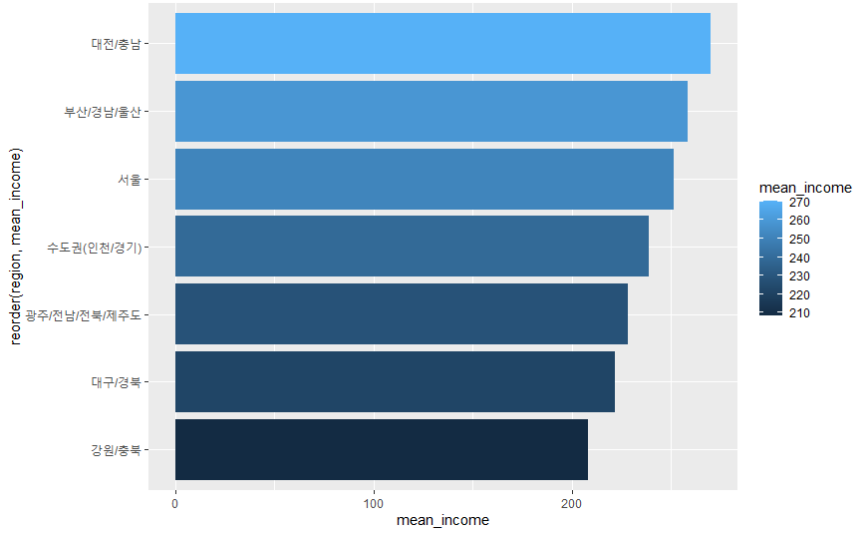
# 5. 그래프 정렬

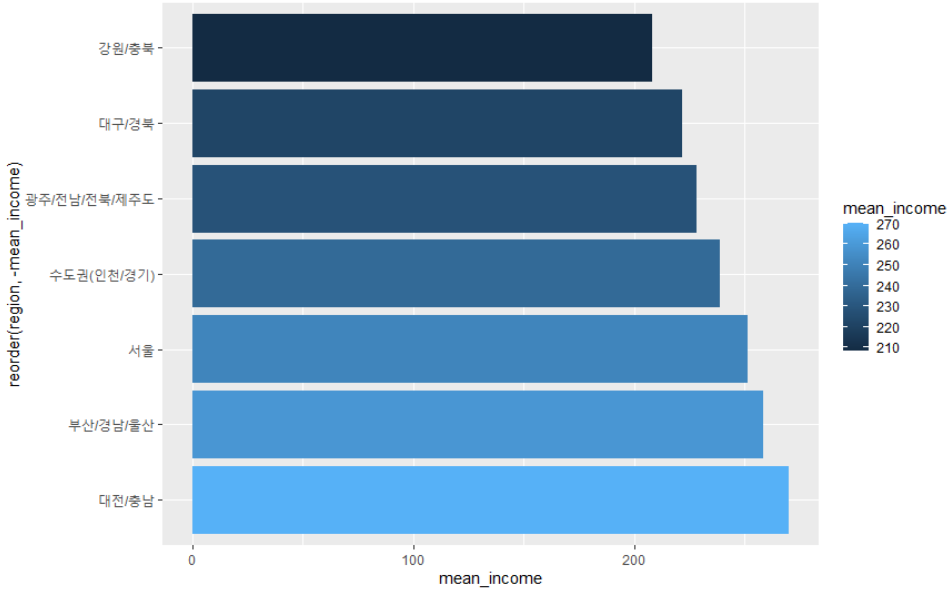
#월급 높은순

```
ggplot(data = region_income, aes(x=reorder(region,mean_income),
y=mean_income,fill=mean_income))+geom_col() + coord_flip()
```

#월급 낮은순

```
ggplot(data = region_income, aes(x=reorder(region,-mean_income),
y=mean_income,fill=mean_income))+geom_col() + coord_flip()
```



# 2. 성별에 따른 종교유무

Q. 유교와 무교의 성비는 어떻게 될까?

## 1) 지역변수 검토

class(welfare$religion) #numeric
table(welfare$religion) #1은 유교, 2는 무교

class(welfare$sex) #numeric
table(welfare$sex) #1은 남성, 2는 여성

# 2)전처리

```
#결측치 처리 & 확인
welfare$religion <- ifelse(welfare$religion==9, NA, welfare$religion)
table(is.na(welfare$religion)) #False만 존재, 결측치 없음

welfare$religion <- ifelse(welfare$religion==1, "Yes", "No")
table(welfare$religion) #1과 2가 각각 Yes, No로 바뀜

welfare$sex <- ifelse(welfare$sex==9, NA, welfare$sex)
table(is.na(welfare$sex)) #False만 존재, 결측치 없음


welfare$sex <- ifelse(welfare$sex==1, "Male", "Female")
table(welfare$sex) #1과 2가 각각 Male Female로 바뀜
```

# 3) 성별 종교 유무 표 만들기

```
religion_sex <- welfare %>%
  group_by(religion,sex) %>%
  summarise(n=n()) %>% #위의 첫번째 변수인 religion의 빈도수 계산해줌
  mutate(tot_group = sum(n)) %>%
  mutate(pct = round(n/tot_group*100),1) #종교 유무 및 성비 표현; round()를 통해 소수점 첫
                                                                  째 자리까지 표현
```

Religion_sex

| ## Religion | Sex | n | tot_group | pct |
|---|---|---|---|---|
| ## No | Female | 4256 | 8617 | 49 |
| ## No | Male | 4361 | 8617 | 51 |
| ## Yes | Female | 4830 | 8047 | 60 |
| ## Yes | Male | 3217 | 8047 | 40 |

*Male과 Female 각각의 총합은 약간 다르지만, 그냥 진행

# 4) 그래프 만들기

```
ggplot(data=religion_sex, aes(x=religion, y=pct, fill=sex)) + geom_col(position="dodge")
+scale_x_discrete(limits=c("Yes", "No"))
```

#fill=sex는 성별에 따라 다른색으로 표현

#position="dodge"를 통해, 막대를 분리

#scale_x_discrete을 통해, 유교, 무교 순으로 x축 정렬

#유교는 여성의 비율이 더 높고, 무교는 성비가 비슷함.