비즈니스 모델링1 개인 프로젝트 레포트

60191434 배정연

목 차

1. 나이, 성별에 의한 연간 자원 봉사 활동 횟수 평균 분석 -
“나이와 성별 중 연간 자원 봉사 활동 횟수에 더 유의미한 영향을 끼치는 변수는 무엇일까?”

<분석의 절차>
1. 16차 웨이브 자료의 코딩북을 이용하여 데이터가 어떻게 저장되어 있는 지 확인한다.

자료의 p1605_6이 자원봉사활동 연간 횟수이고 코딩북의 정보를 통해서 모름/무응답에 해당하는 999는 결측치 처리를 코드에서 해야 함을 파악하였다.

h16_g4의 정보를 가져와서 생을 이용하여 나이를 추가하고 성별은 h16_g3을 가져오면 되는 것을 파악하였다.

2. 데이터 분석 준비 작업 & 변수 추가하고 전처리 하기

```r
install.packages("foreign")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("haven")
install.packages("readxl")
install.packages("xlsx")

library(foreign)
library(dplyr)
library(ggplot2)
library(haven)
library(readxl)
library(xlsx)

raw_data <- read_spss("Koweps_hpc16_2021_beta1.sav")
view(raw_data)



# 1. 나이, 성별에 의한 자원봉사활동 횟수 연평균 분석

volunteer_data <- raw_data
volunteer_data <- rename(volunteer_data, birth = h16_g4,gender = h16_g3, volunteer_activities = p1604_6)


# 모름/무응답 결측치 처리
volunteer_data$volunteer_activities <- ifelse(volunteer_data$volunteer_activities == 999, NA, volunteer_data$volunteer_activities)
# 태어난 연도를 통해 나이를 추가
volunteer_data$age <- 2021 - volunteer_data$birth + 1
```

필요한 패키지들을 로드하고 16차 웨이브 자료를 raw_data안에 담았다. 그리고 필요한 자료를 구하기 위해서 복사본인 volunteer_data를 만들고 필요한 정보들을 rename을 이용하여 이해할 수 있는 단어로 바꿔서 volunteer_data에 넣었다.
그리고 앞서 말한 모름/무응답에 해당하는 999는 결측치 처리를 해주었고 2021년도 자료임을 이용하여 태어난 연도를 나이로 환산하였다.

가구원 번호	
가구주와의 관계	가구주와의 관계코드표 참고
성별	1.남 2.여
태어난 연도	년

코드북을 통해 성별 1이 남성이고 2가 여성임을 확인하였으니
1과 2를 각각 남성과 여성으로 표시될 수 있게 한다.

```r
volunteer_data <- volunteer_data %>%
  mutate(gender_type = ifelse(volunteer_data$gender == 1, "남자", ifelse(volunteer_data$gender == 2, "여자", NA)))

age_household_volunteer <- volunteer_data %>%
  filter(!is.na(volunteer_activities)) %>%
  group_by(age, gender_type) %>%
  summarise(mean_volunteer = mean(volunteer_activities))
```

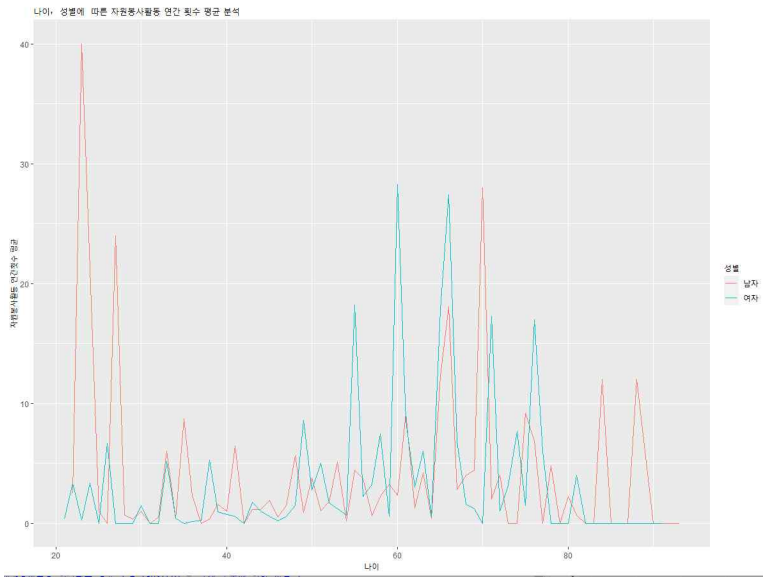
그리고 결측치를 filter해주고 age와 gender를 묶어서 groupby해주고 summarise를 이용해 요약된 평균을 만든다.

3. 최종 - 그래프 그리기

```
42
43 ggplot(data = age_household_volunteer, aes(x = age, y = mean_volunteer, col=gender_type)) + geom_line()+ ggtitle("나이, 성별에 따른
   자원봉사활동 연간 횟수 평균 분석") + xlab("나이") + ylab("자원봉사활동 연간횟수 평균") + labs(col = "성별")
44
45
```

최종적으로 x축은 나이, y축은 연간 자원봉사 평균, 색은 성별에 따라서 다르게 나오도록 그래프를 그리고 x축과 y축에 출력될 정보들을 각각 담아서 line그래프를 출력한다.

<결과 분석>



20~30대의 여성이 눈에 띄게 연간 자원봉사의 횟수가 높은 것을 알 수 있으며 성별에 상관없이 자원봉사의 횟수가 높은 나이는 50대 후반부터 80대까지인 것으로 알 수 있고 성별에 따른 그래프의 특징도 찾아보기 힘들고 나이에 따른 그래프의 경향성도 찾아보기 힘든 것을 통해 연간 자원봉사 횟수는 나이와 성별과는 유의미한 상관관계를 갖지 않음을 알 수 있다.

2. 지역별 집의 가격 - “지역에 따라서 집의 가격이 얼마나 차이 날까?”

<분석의 절차>

1.지역별 집의 가격을 알려면 지역 정보와 집의 가격정보를 먼저 데이터에서 찾아야 한다. 코딩북을 이용하여 관련 데이터가 어떤 이름으로 구성되어 있는지 파악하였다.

지역에 관련된 데이터들의 열은 2가지가 있었는데 h16_reg5의 경우에는 지역에 따른 특색을 보기에는 부적합한 것 같아서 h16_reg7을 택했다.

집의 가격과 같은 경우에는 h1606_6열의 데이터를 이용하면 됨을 파악하였다.

```r
# 2. 지역에 따른 집의 가격
region_data <- raw_data
region_data <- rename(region_data, region = h16_reg7, price = h1606_6)
region_data$price <- ifelse(region_data$price == 0 | region_data$price == 999999, NA, region_data$price)

region_data$region <- ifelse(region_data$region == 1, "서울", ifelse(region_data$region == 2, "인천,경기", ifelse(region_data$region == 3,
"부산,경남,울산", ifelse(region_data$region == 4, "대구.경북", ifelse(region_data$region == 5, "대전,충남", ifelse(region_data$region == 6,
"강원,충북",ifelse(region_data$region == 7, "광주,전북,전남,제주도", NA )))))))
```

2. 데이터 분석 준비 작업 & 변수 추가하고 전처리 하기

region_data에 data를 담고 각각의 열을 알기 쉬운 이름으로 rename 해주고 각각의 지역의 구분을 알수 있게 한글로 명명해주었다. 그리고 없음, 모름/ 무응답에 해당하는 결측치들을 제거해주었다.

```
region_price <- region_data %>%
  filter(!is.na(region) & !is.na(price)) %>%
  select(region, price)
```

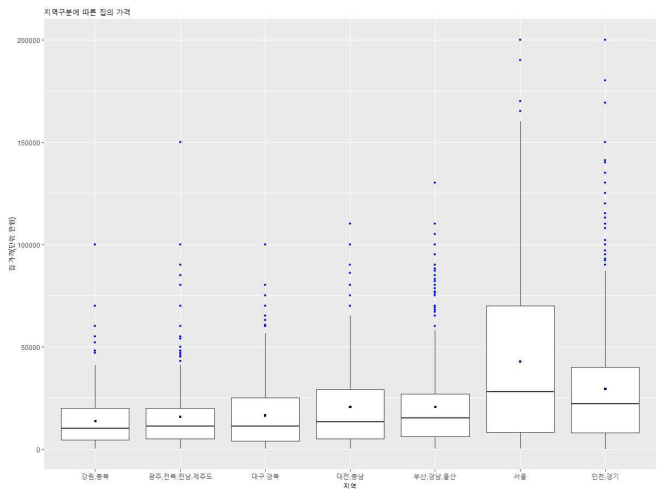
region_price에 region_data를 담고, 결측치를 filter를 이용하여 배제하였고, 지역정보와 가격정보를 선택하여 담았다.

3. 최종 - 그래프 그리기

```
ggplot(data = region_price, aes(x = region, y = price))+ geom_boxplot(width=0.8, outlier.size=1, outlier.shape=16, outlier.colour
="blue") + stat_summary(fun="mean", geom="point", shape=22, size=1, fill="blue") + ggtitle("지역구분에 따른 집의 가격") + xlab
("지역") + ylab("집 가격(단위: 만원)")  + scale_y_continuous(limits = c(0, 2e+05))
```

x좌표와 y좌표에 각각 지역과 가격을 담고 boxplot을 그렸다.

<결과 분석>



집값이 평균적으로도 가장 비싸고 편차가 큰곳은 서울이였고 하지만 서울과 인천 경기는 이상치의 가격은 비슷하였다. 집값의 이상치가 가장 적은 곳은 대구/경북이였고 평균적으로 집값이 제일 싼 곳은 강원, 충북이였다.

3. 장애의 종류에 따른 월급 차이 - “장애도 종류에 따라서 생계에 실질적으로 영향을 끼치는 정도의 차이가 심한데 정부에서 더 많이 경제적 지원을 해줘야 하는 장애는 어떤 것일까?”

<분석의 절차>

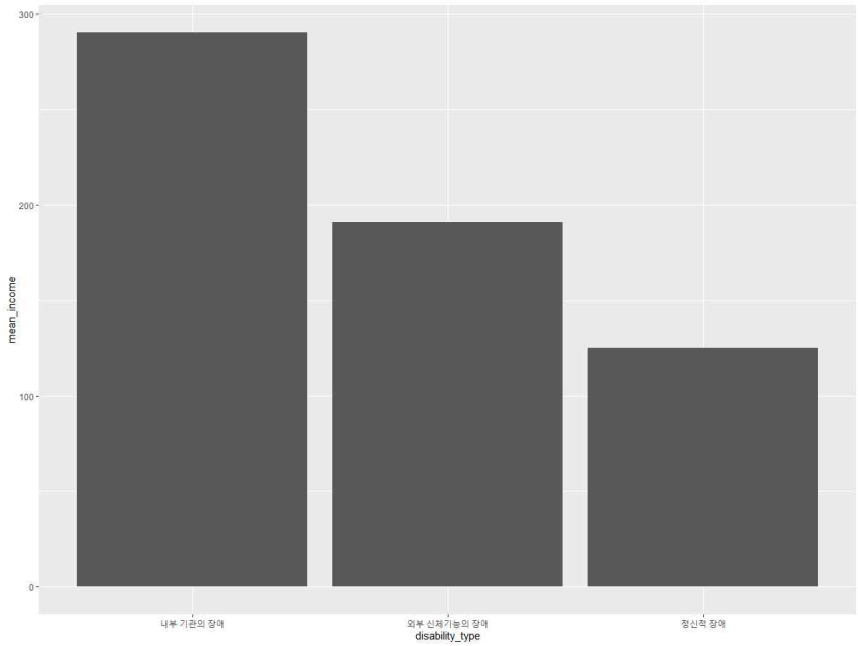
1. 앞선 두 방법들과 동일하게 코딩북에서 장애의 종류에 대한 열을 찾았더니 h16_g8 이였고 1~15까지는 장애의 종류였고 16은 장애의 종류를 파악하지 못한 자였고 0은 비장애인 이였다.
2. 그리고 월급을 분석하려고 p1602_8aq1열이 월 평균 임금이라서 해당열을 가져왔다.

```
# 3. 장애 종류에 따른 월급 분석
disability_data <- raw_data
disability_data <- rename(disability_data, disability_type = h16_g8, income = p1602_8aq1)
#결측치 제거(0은 비장애인임, 16은 비동류 장애인이라서 종류를 파악할 수 없어 제거)
disability_data$disability_type <- ifelse(disability_data$disability_type==0 |disability_data$disability_type==16,NA
,disability_data$disability_type)
#결측치 제거된 것 확인
table(disability_data$disability_type)
#이름부여 및 데이터 가공
disability_data$disability_type <-ifelse(disability_data$disability_type == 1 | disability_data$disability_type
==2|disability_data$disability_type ==3|disability_data$disability_type == 4| disability_data$disability_type == 5|
disability_data$disability_type == 13, "외부 신체기능의 장애", ifelse(disability_data$disability_type == 9|
disability_data$disability_type == 10| disability_data$disability_type == 11| disability_data$disability_type == 12|
disability_data$disability_type == 14| disability_data$disability_type == 15,"내부 기관의 장애", ifelse
(disability_data$disability_type == 6| disability_data$disability_type == 7| disability_data$disability_type == 8,"정신적 장애"
,NA)))
#이상치 결측 처리(월급이 1~9998 사이 나타나기 때문에 0이거나 9999를 이상치 처리(코드북에 따라 모름/무응답이 9999임))
disability_data$income <- ifelse(disability_data %in% c(0,9999), NA,disability_data$income)
#장애 종류에 따른 평균 표 만들
disability_income <- disability_data %>%
 filter(!is.na(disability_type)) %>%
 filter(!is.na(income)) %>%
 group_by(disability_type)%>%
 summarise(mean_income = mean(income))

disability_income
ggplot(data=disability_income,aes(x=disability_type,y=mean_income))+geom_col()
```

3. 결측치를 제거하고 장애의 종류가 많아서 외부 장애, 내부 장애, 정신장애로 묶었다. 장애 종류에 따라서 표를 만들고 표를 만들때 결측치를 각각 filter를 이용하여 제거하였다.

그리고 x축에는 장애의 종류, y축에는 평균 임금으로 그래프를 그렸다.



<결과 분석>

해당 표에 따르면 정신적 장애가 가장 낮은 평균임금을 가지므로 정신적 장애를 가장 국가에서 많이 지원해줘야하고 외부 신체기능의 장애를 그다음으로 지원해줘야하고 내부 기관의 장애와 정신적 장애의 임금수준이 2배 차이 남에 따라서 지원 가이드라인을 만들어야한다.