준비는 모두 끝났다. 실전투입!

실전 분석 미션 "한국인의 삶을 파악하라!" - 한국복지패널 데이터

분석 목표

- 분석1: 성별에 따른 소득
- 분석2: 나이와 소득의 관계
- 분석3: 연령대에 따른 소득
- 분석4: 연령대 및 성별에 따른 소득

준비하기

foreign 패키지 설치

```
install.packages("foreign")
```

패키지 로드

```
library(foreign)
library(dplyr)
library(ggplot2)
```

데이터 불러오기

```
# 복지패널데이터 로드
raw_welfare <- read.spss("data_spss_Koweps2014.sav", to.data.frame = T)

# 데이터 copy
welfare <- raw_welfare
```

데이터 검토

```
dim(welfare)
str(welfare)
head(welfare)
summary(welfare)
View(welfare)
```

변수명

```
welfare <- rename(welfare,
                  sex = h0901_4,     # 성별
                  birth = h0901_5,   # 태어난 연도
                  income = h09_din)  # 소득
```

분석1: 성별에 따른 소득

절차

1.변수 검토 및 정제 - 성별

- 1-1.변수 검토, 수정
- 1-2.정제 - 이상치 확인 및 결측처리

2.변수 검토 및 정제 - 소득

- 2-1.변수 검토, 수정
- 2-2.정제 - 이상치 확인 및 결측처리

3.성별 소득 평균 분석

- 성별 소득 평균표 생성
- 그래프 생성

1.변수 검토 및 정제- 성별

1-1.변수 검토, 수정

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

```
## [1] "numeric"
```

```
summary(welfare$sex)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.309   2.000   2.000
```

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

```
##
##    1    2
## 4873 2175
```

1-2.정제 - 이상치 확인 및 결측처리

- 성별 이상치 : 모름/무응답=9

```
# 이상치 확인
table(welfare$sex)
```

```
##
##    1    2
## 4873 2175
```

```
# 이상치 결측 처리
welfare$sex <- ifelse(welfare$sex == 9, NA, welfare$sex)

# 결측치 확인
table(is.na(welfare$sex))
```

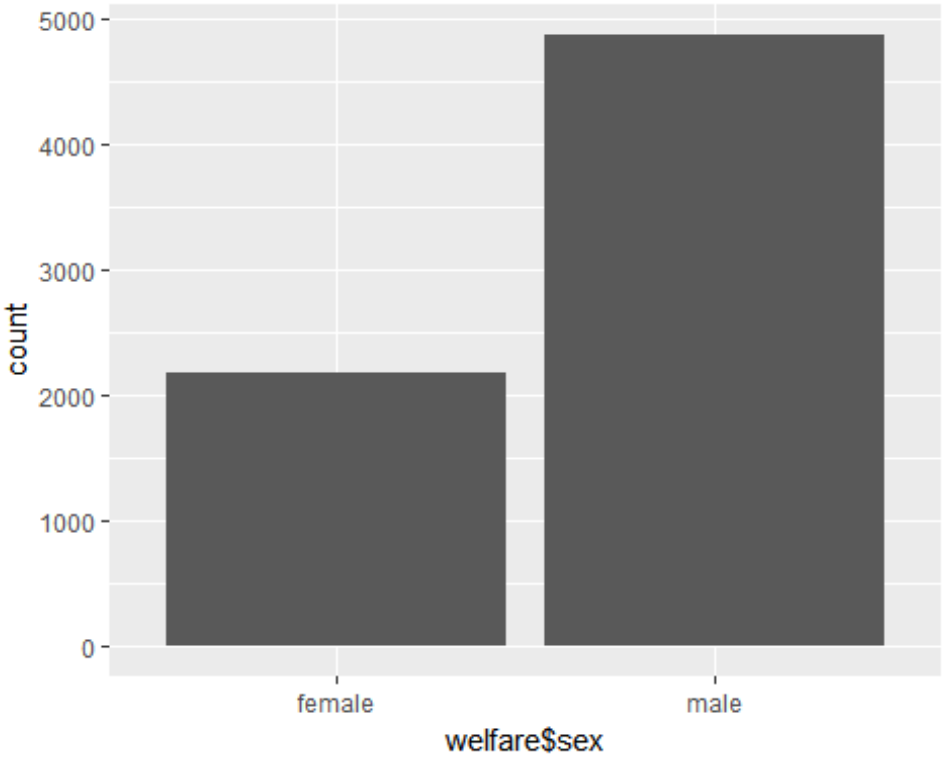
```
##
## FALSE
##  7048
```

변수 값 변경

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

```
##
## female   male
##   2175   4873
```

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


5000
4000
3000
2000
1000
0
count
female
male
welfare$sex

2.변수 검토 및 정제- 소득

2-1.변수 검토, 수정

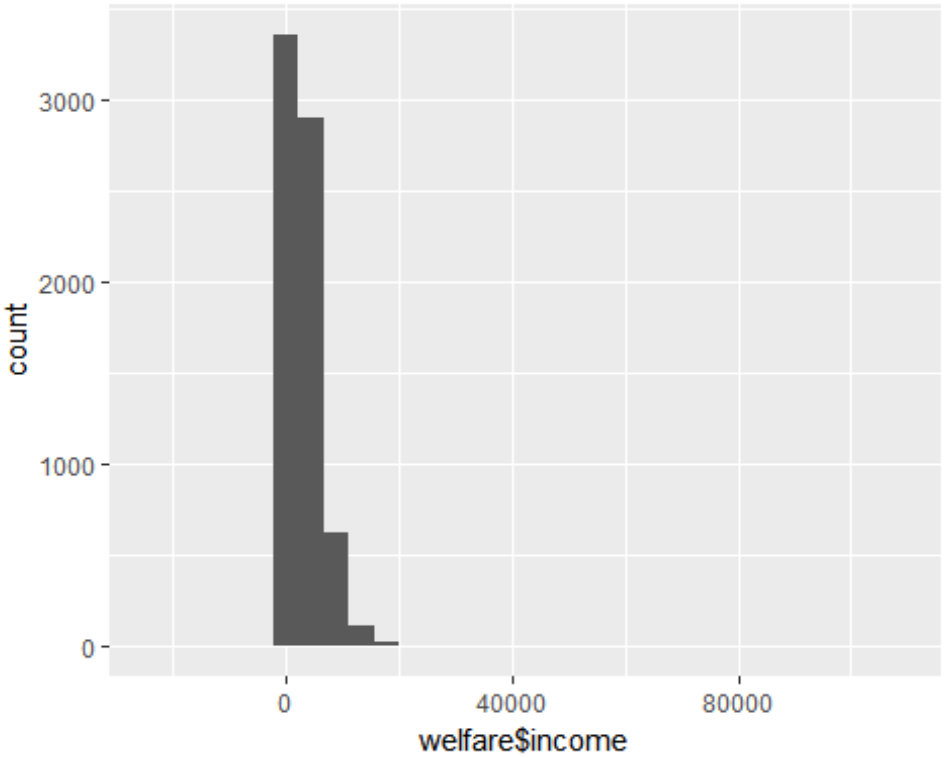
```
class(welfare$income)
```

```
## [1] "numeric"
```

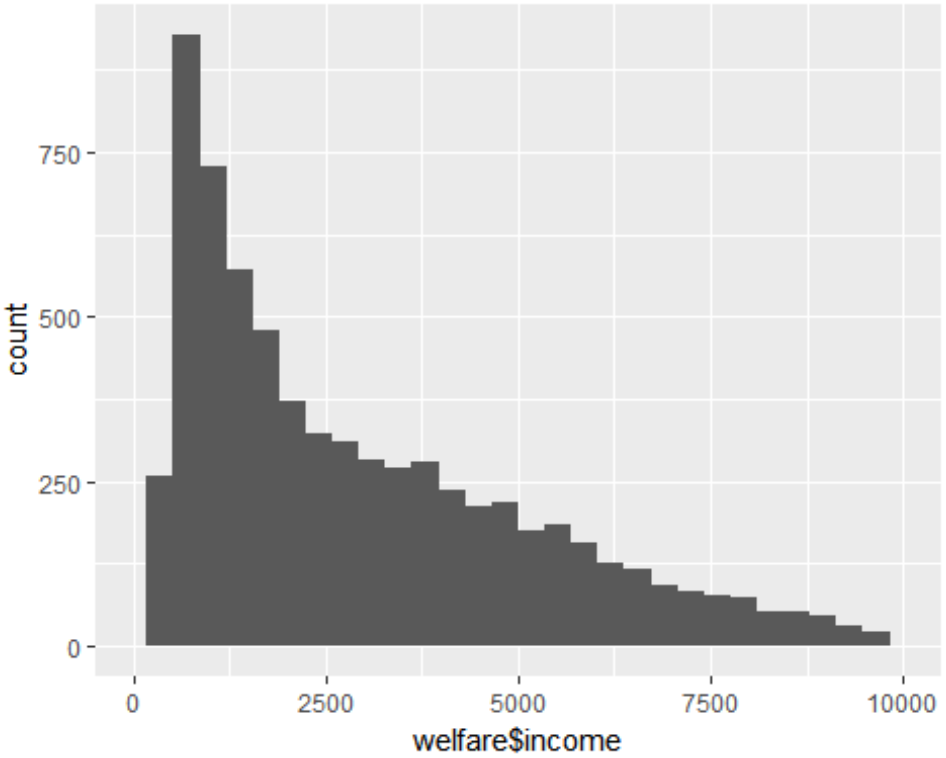
```
summary(welfare$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -20516    1108    2404    3336    4642  108888
```

```
qplot(welfare$income)
```



```r
qplot(welfare$income) + xlim(0, 10000)  # x 축 설정
```



2-2.정제 - 이상치 확인 및 결측처리

- 소득 이상치 : 모름/무응답 없음

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

```
##
## FALSE
##  7048
```

3.성별 소득 평균 분석

성별 소득 평균표 생성

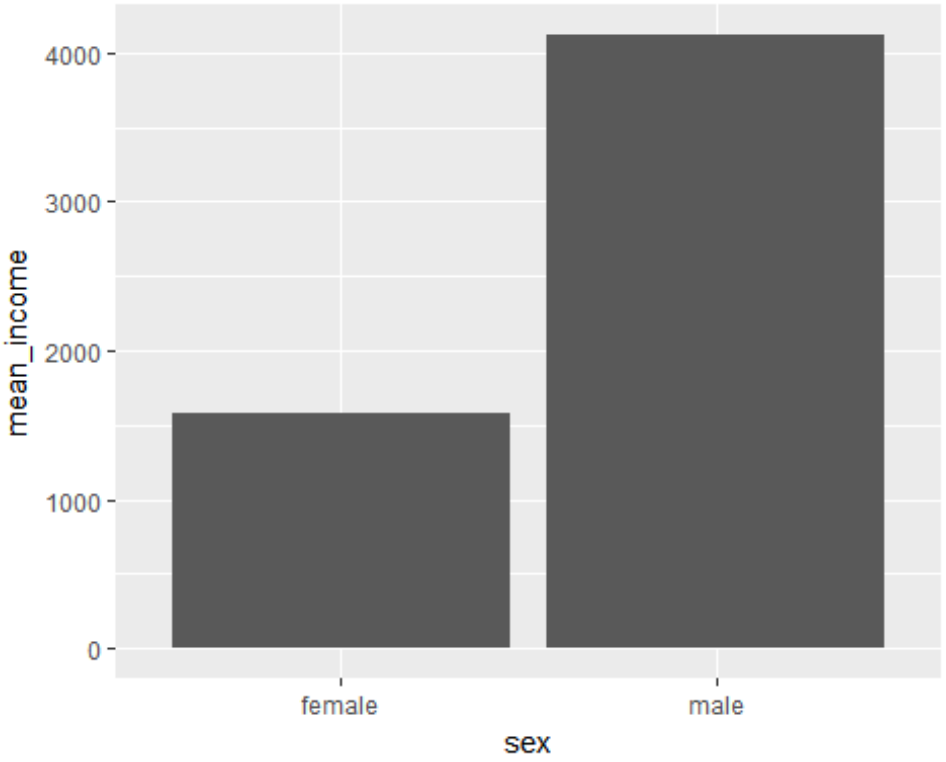
```
sex_income <- welfare %>%
  group_by(sex) %>%
  summarise(mean_income = mean(income))

sex_income
```

```
## # A tibble: 2 x 2
##      sex mean_income
##    <chr>       <dbl>
## 1 female    1581.255
## 2   male    4118.903
```

그래프 생성

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



분석2: 나이와 소득의 관계

절차

1.변수 검토 및 정제 - 나이

- 1-1.태어난 연도 변수 검토
- 1-2.정제 - 이상치 확인 및 결측처리
- 1-3.나이 변수 생성

2.변수 검토 및 정제 - 소득

- 앞에서 완료됨

3.나이별 소득 평균 분석

- 나이별 소득 평균표 생성
- 그래프 생성

1.변수 검토 및 정제- 나이

1-1.태어난 연도 변수 검토

1.변수 검토 및 정제- 나이

1-1.태어난 연도 변수 검토

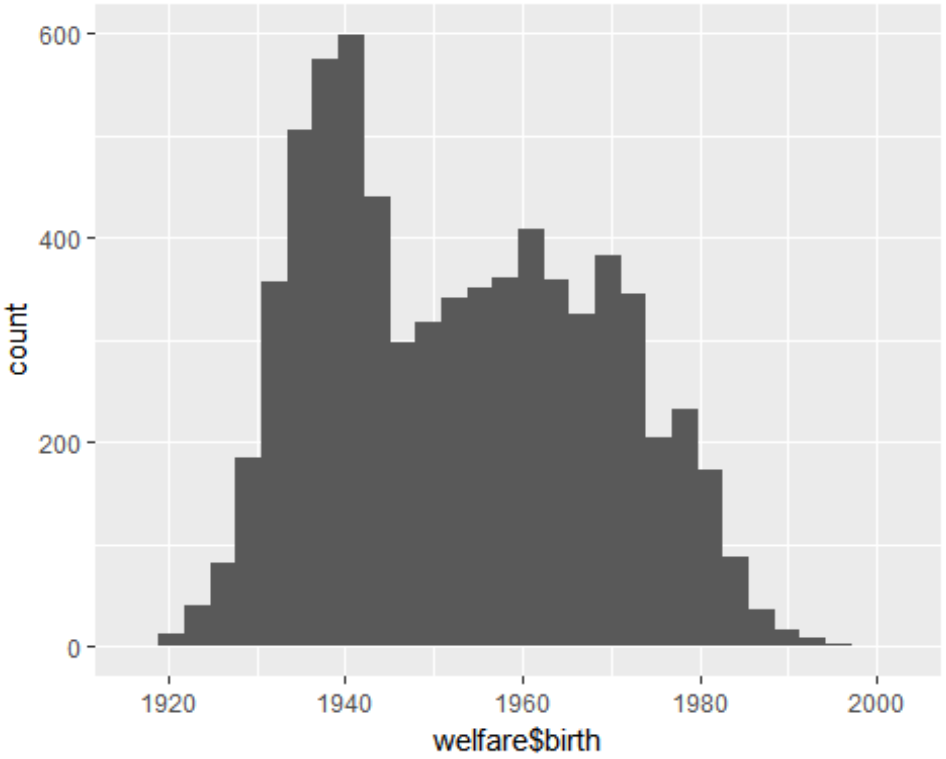
```
class(welfare$birth)
```

```
## [1] "numeric"
```

```
summary(welfare$birth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1918    1940    1952    1953    1966    2002
```

```
qplot(welfare$birth)
```


600
400
200
0
count
1920
1940
1960
1980
2000
welfare$birth

1-2.정제 - 이상치 확인 및 결측처리

- 태어난 연도 이상치 : 모름/무응답=9999
  - (1)이상치 확인, 결측처리
  - (2)결측치 확인

1-2.정제 - 이상치 확인 및 결측처리

- 태어난 연도 이상치 : 모름/무응답=9999

```
# 이상치 확인
summary(welfare$birth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1918    1940    1952    1953    1966    2002
```

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

# 결측치 확인
table(is.na(welfare$birth))
```

```
##
## FALSE
##  7048
```

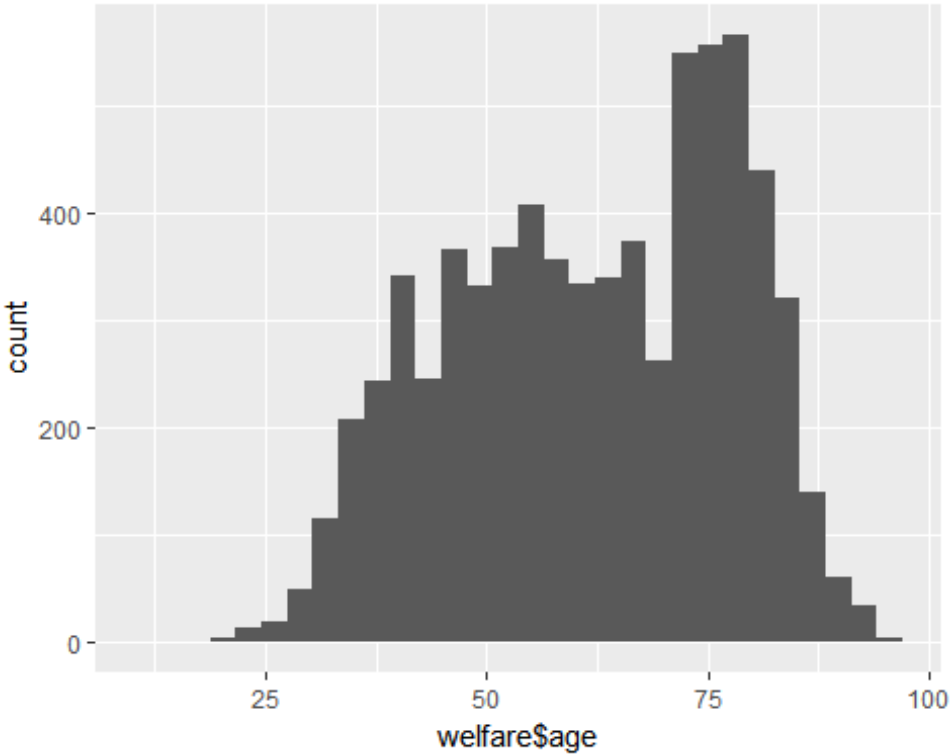
1-3.나이 변수 생성

1-3.나이 변수 생성

```
welfare$age <- 2014-welfare$birth+1
summary(welfare$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   49.00   63.00   62.01   75.00   97.00
```

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



2.변수 검토 및 정제- (2)소득

- 앞에서 완료됨

3.나이별 소득 평균 분석

나이별 소득 평균표 생성

3.나이별 소득 평균 분석

나이별 소득 평균표 생성

```
age_income <- welfare %>%
  group_by(age) %>%
  summarise(mean_income = mean(income))

age_income
```

```
## # A tibble: 79 x 2
##      age mean_income
##    <dbl>       <dbl>
## 1     13    252.0000
## 2     20   1094.9000
## 3     21   2117.6000
## 4     22   2656.0000
## 5     23   1748.2500
## 6     24   5429.6000
## 7     25   2310.4000
## 8     26   5273.3714
## 9     27   3394.9800
## 10    28   3061.2222
## 11    29   6700.5000
## 12    30   3829.3478
```

```
## 13    31   4631.0200
## 14    32   4120.4977
## 15    33   4602.2392
## 16    34   4890.4436
## 17    35   6498.3254
## 18    36   5183.6235
## 19    37   5245.9724
## 20    38   5339.9528
## 21    39   4935.6589
## 22    40   5451.0591
## 23    41   5600.6653
## 24    42   5028.2800
## 25    43   5611.1456
## 26    44   5915.0214
## 27    45   4902.6651
## 28    46   5151.2195
## 29    47   4536.9596
## 30    48   5095.5735
## 31    49   5410.0626
## 32    50   5025.2876
## 33    51   4442.5982
## 34    52   5274.7871
## 35    53   4968.7415
## 36    54   4796.8304
## 37    55   4810.4013
## 38    56   4764.4008
```

```
## 39    57   4923.5692
## 40    58   4136.8440
## 41    59   4286.6234
## 42    60   4261.5271
## 43    61   3471.1073
## 44    62   3729.9290
## 45    63   3783.8135
## 46    64   2702.1936
## 47    65   3225.8067
## 48    66   3010.4055
## 49    67   2599.9184
## 50    68   2480.3602
## 51    69   2541.8000
## 52    70   2450.2841
## 53    71   2225.1972
## 54    72   1929.1064
## 55    73   1791.0443
## 56    74   1823.9989
## 57    75   1840.8915
## 58    76   1619.7357
## 59    77   1472.0482
## 60    78   1594.6102
## 61    79   1393.5847
## 62    80   1286.4084
## 63    81   1307.8524
## 64    82   1262.2224
```

```
## 65    83   1294.1739
## 66    84   1027.0374
## 67    85   1169.5734
## 68    86   1278.5464
## 69    87   1158.0653
## 70    88   1022.4857
## 71    89   1150.9250
## 72    90    974.3478
## 73    91   1270.5857
## 74    92    713.3385
## 75    93    696.6308
## 76    94   1545.5250
## 77    95    828.5000
## 78    96   2041.0000
## 79    97   1109.0000
```

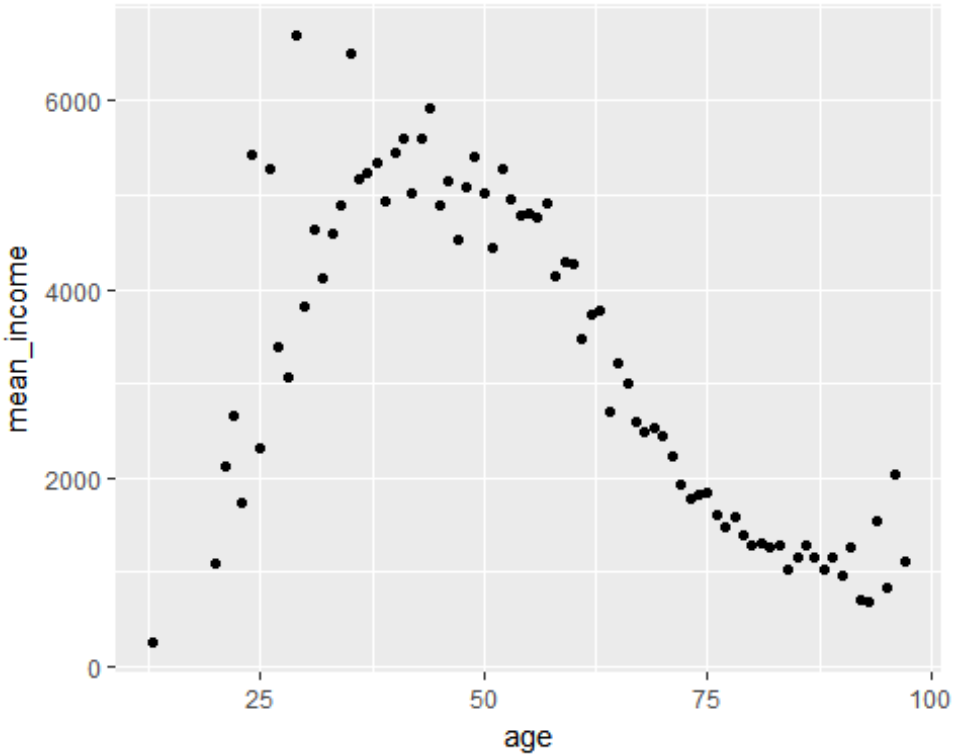
3.나이별 소득 평균 분석

그래프 생성 - 산점도

3.나이별 소득 평균 분석

그래프 생성 - 산점도

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



분석3: 연령대에 따른 소득

절차

1.변수 검토 및 정제 - 연령대

- 1-1.연령대 변수 생성

2.변수 검토 및 정제 - 소득

- 앞에서 완료됨

3.연령대별 소득 평균 분석

- 연령대별 소득 평균표 생성
- 그래프 생성

1.변수 검토 및 정제 - 연령대

1-1.연령대 변수 생성

범주	기준
초년	30세 미만
중년	30~59세
노년	60세 이상

1.변수 검토 및 정제 - 연령대

1-1.연령대 변수 생성

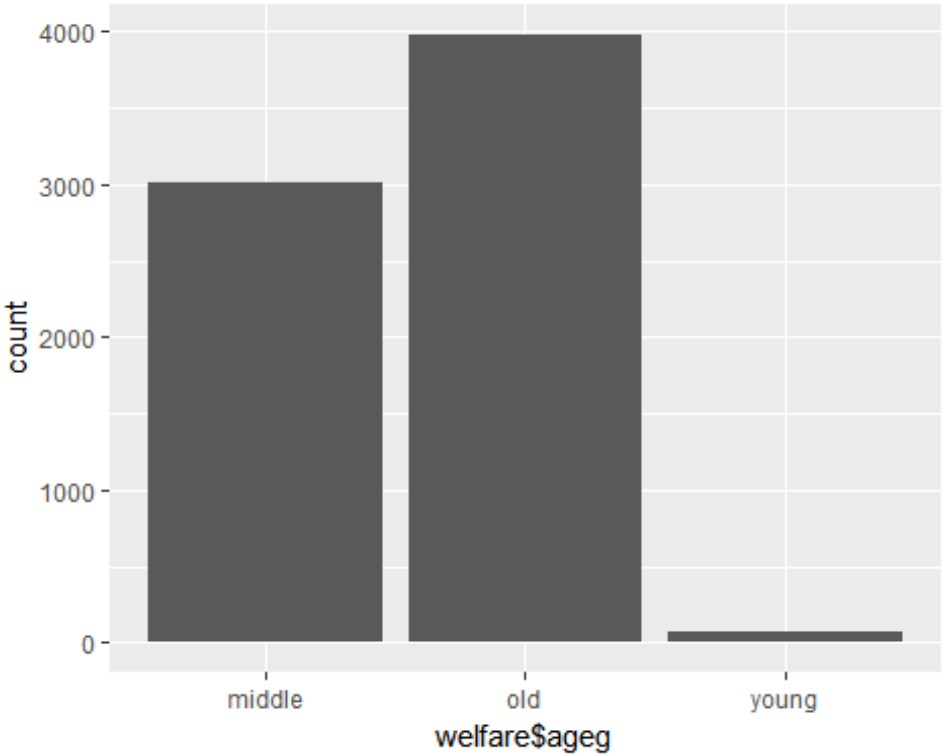
범주	기준
초년	30세 미만
중년	30~59세
노년	60세 이상

```
welfare <- welfare %>%
  mutate(ageg = ifelse(age < 30, "young",
                       ifelse(age <= 59, "middle", "old")))


table(welfare$ageg)
```

```
##
## middle    old  young
##   3004   3979     65
```

```
qplot(welfare$ageg)
```


4000
3000
2000
1000
0
count
middle
old
young
welfare$ageg

2.변수 검토 및 정제 - 소득

- 앞에서 완료됨

3.연령대별 소득 평균 분석

연령대별 소득 평균표 생성

- 초년 빈도 적으므로 제외

3.연령대별 소득 평균 분석

연령대별 소득 평균표 생성

- 초년 빈도 적으므로 제외

```
welfare_income <- welfare %>%
  filter(ageg != "young") %>%
  group_by(ageg) %>%
  summarise(mean_income = mean(income))

welfare_income
```

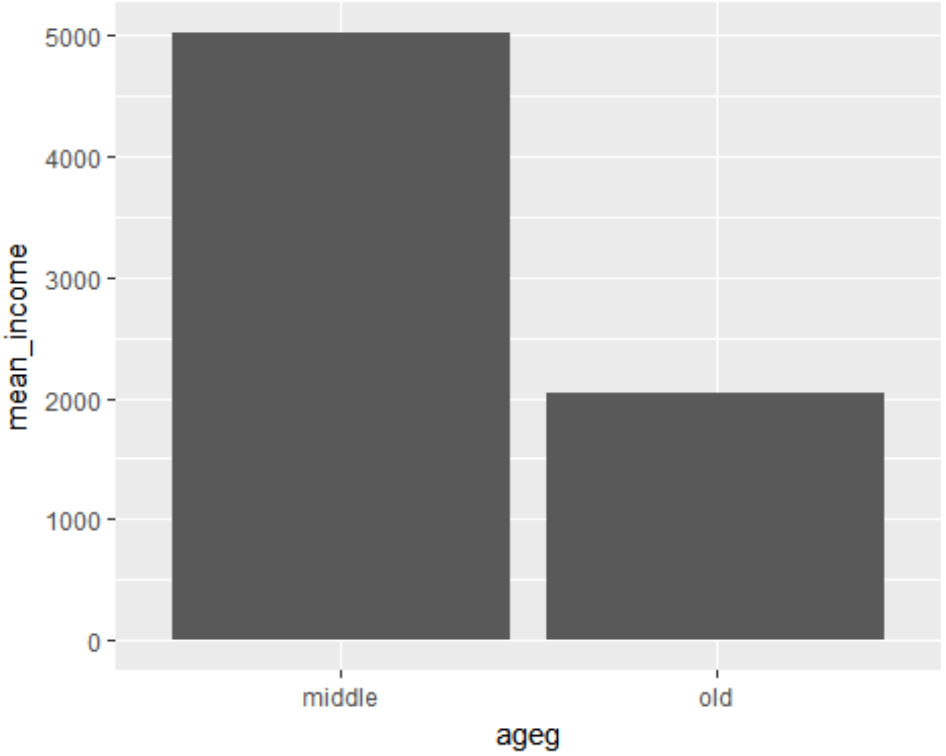
```
## # A tibble: 2 x 2
##     ageg mean_income
##    <chr>       <dbl>
## 1 middle    5017.822
## 2    old    2049.348
```

3.연령대별 소득 평균 분석

그래프 만들기

그래프 만들기

```
ggplot(data = welfare_income, aes(x = ageg, y = mean_income)) + geom_col()
```



분석4: 연령대 및 성별에 따른 소득

절차

1.연령대 및 성별 소득 평균표 생성

2.그래프 만들기

1.연령대 및 성별 소득 평균표 생성

- 초년 제외

1.연령대 및 성별 소득 평균표 생성

- 초년 제외

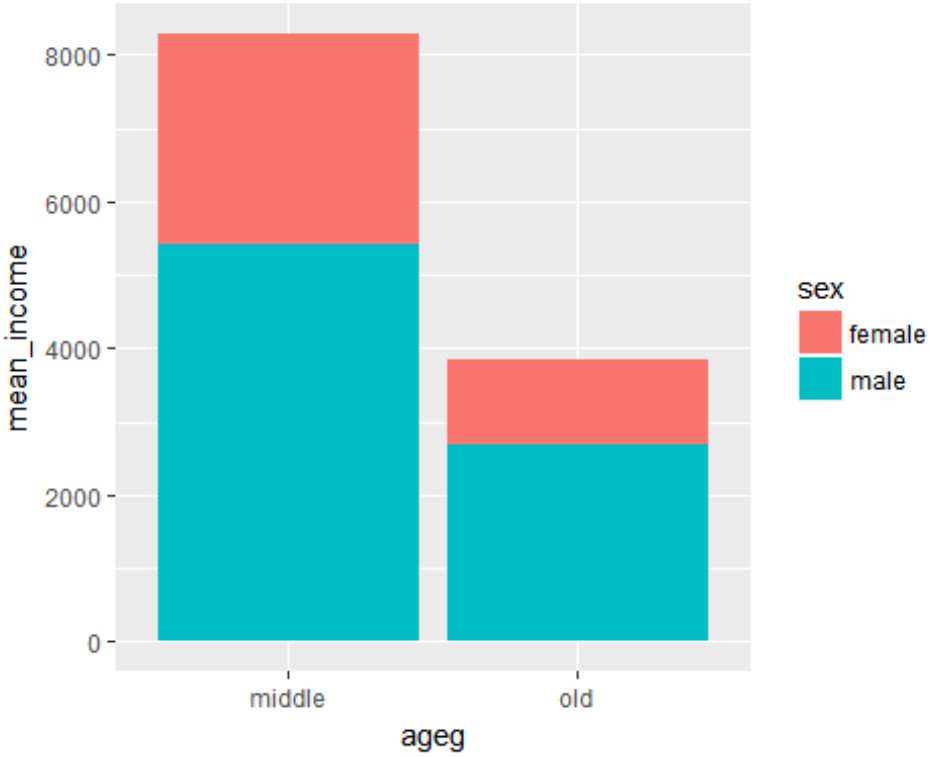
```
sex_income <- welfare %>%
  filter(ageg != "young") %>%
  group_by(ageg, sex) %>%
  summarise(mean_income = mean(income))

sex_income
```

```
## Source: local data frame [4 x 3]
## Groups: ageg [?]
##
##     ageg    sex mean_income
##    <chr>  <chr>       <dbl>
## 1 middle female    2868.804
## 2 middle   male    5420.444
## 3    old female    1179.192
## 4    old   male    2674.162
```

2.그래프 만들기

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



2.그래프 만들기

```
ggplot(data = sex_income, aes(x = ageg, y = mean_income, fill = sex)) +
  geom_col(position = "dodge")  # position 변경(기본값 = "stack")
```