Korean Welfare

This project is aimed to understand the income circumstances of the people living in Korea. Dataset that I'm going to use is published by Korea Institute for Health and Social Affairs in 2015, which contains the record of 7000 families from 2006 to 2015. Each individuals were recorded by around 1000 factors, some of the examples are sex, income and religion.

Topics of this project

1. Incomes differ by sex?
2. At which age do people make the most income?
3. Which age group earns more?
4. Incomes by sex differ by age group?
5. People with religion have lower divorce rates?
6. Divorce rate by religion differ by age group?
7. Areas with more elders?

```
library(foreign)
library(readxl)
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
```

```
raw_data <- read.spss(file = "Koweps_hpc10_2015_beta1.sav",
                      to.data.frame = T)
```

```
## Warning in read.spss(file = "Koweps_hpc10_2015_beta1.sav", to.data.frame
## = T): Koweps_hpc10_2015_beta1.sav: Compression bias (0) is not the usual
## value of 100
```

```
welfare <- raw_data

dim(welfare)
```

```
## [1] 16664   957
```

As I mentioned above, there are about 1000 factors (957) available and this is way too much for my analysis.

```
colnames(welfare)[1:10]
```

```
##  [1] "h10_id"     "h10_ind"    "h10_sn"     "h10_merkey" "h_new"
##  [6] "h10_cobf"   "h10_reg5"   "h10_reg7"   "h10_din"    "h10_cin"
```

And the column names are written in codes. To make them more readable, I should better change the name of the factors that I'm going to use by reading the codebook. Here I will focus on these six factors: sex, birth, marriage, religion, income and code_region.

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

welfare <- welfare %>%
  select(sex, birth, marriage, religion, income, code_region)

head(welfare)
```

```
##   sex birth marriage religion income code_region
## 1   2  1936        2        2     NA           1
## 2   2  1945        2        2     NA           1
## 3   1  1948        2        2    120           1
## 4   1  1942        3        1    200           1
## 5   2  1923        2        1     NA           1
## 6   1  1962        1        1     NA           1
```

1. Incomes differ by sex?

Using factors: sex, income Aim is to preprocess these two factors and analyse their relationship to find out the income difference between the males and the females.

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

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

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

```
##
##    1    2
## 7578 9086
```

There are two sex groups, male and female, and male is recorded as 1 and female is recorded as 2 according to the codebook. No outliers. Let's replace these with proper words.

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

```
##
## female   male
##   9086   7578
```

Preprocessing of sex is finished. Now move on to income.

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

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

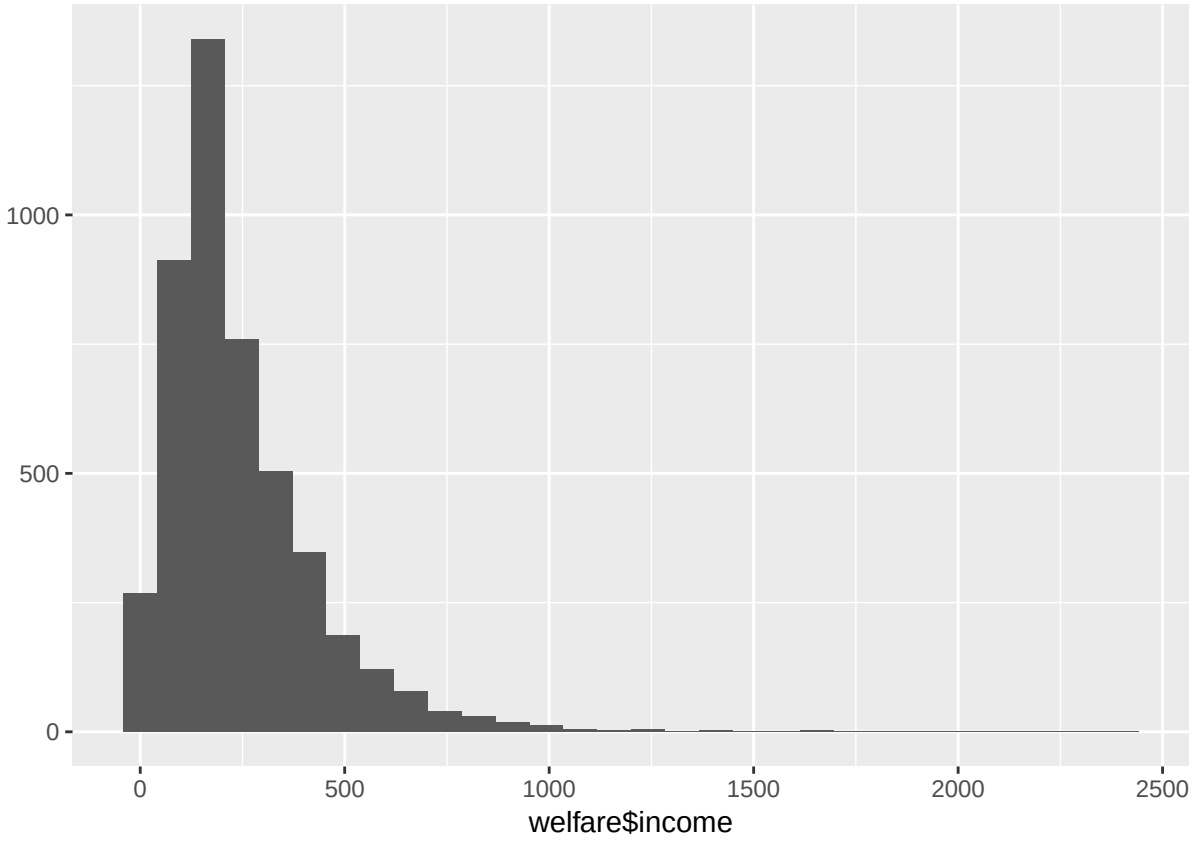
```r
summary(welfare$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   122.0   192.5   241.6   316.6  2400.0   12030
```

Variation is from 0 to 2400 million KRW per month (about \$0~\$30000 AUD) Most people earns between 122 to 316 but the median is 192, which is closer to the 1st quadrant. This puts more emphasis on the lower values.

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

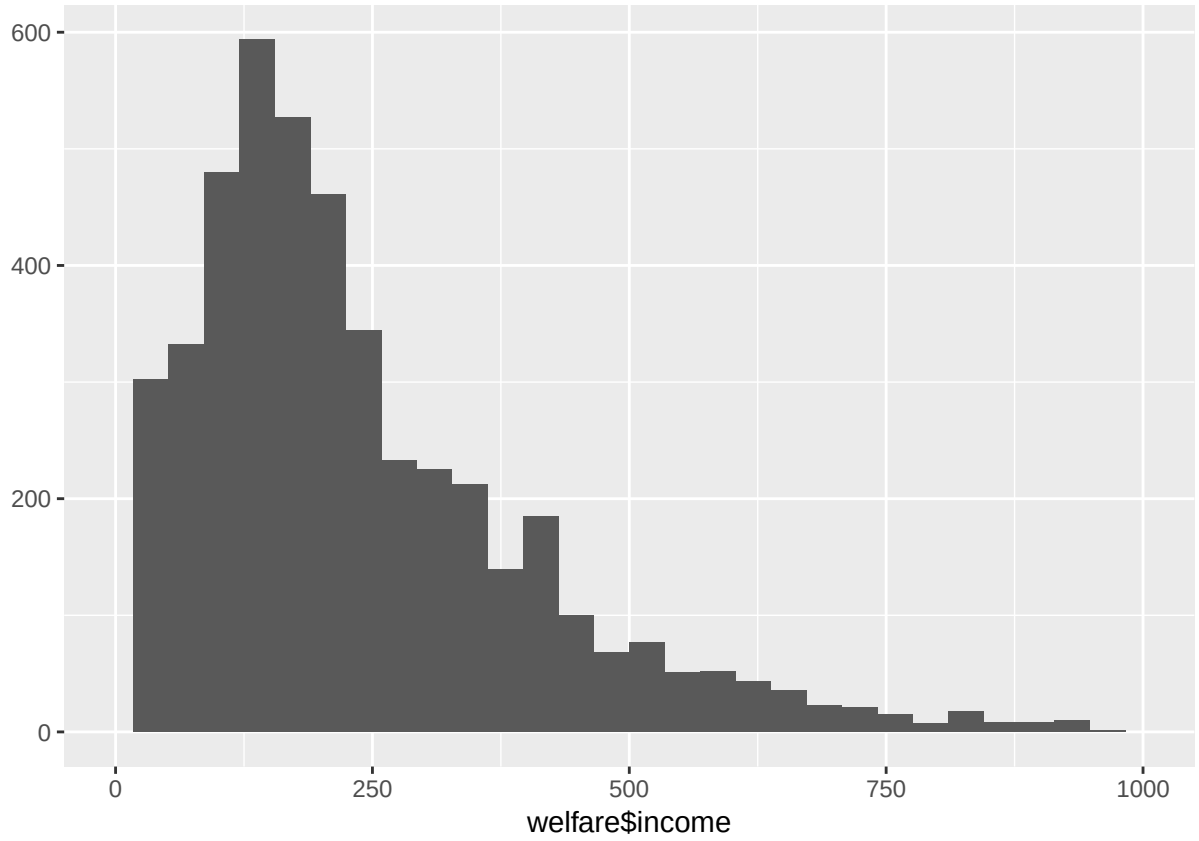
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 12030 rows containing non-finite values (stat_bin).
```



Above 1000 is extremely rare, so let's focus on 0~1000 region.

```
qplot(welfare$income) +
  xlim(0, 1000)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 12051 rows containing non-finite values (stat_bin).
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Majority of people have income between 0~250. Now, there are 12030 NA values. This is obvious because not everyone can earn money. (e.g. children) One thing that's weird is 0s. If we are stating "no income" as NA, then 0 values cannot exist. And the codebook also says that the income values can range from 1 to 9998, so I will replace all the values outside of this range as NA.

```
welfare$income <- ifelse(welfare$income < 1 | welfare$income >9998, NA, welfare$income)
table(is.na(welfare$income))
```

```
##
## FALSE  TRUE
##  4618 12046
```

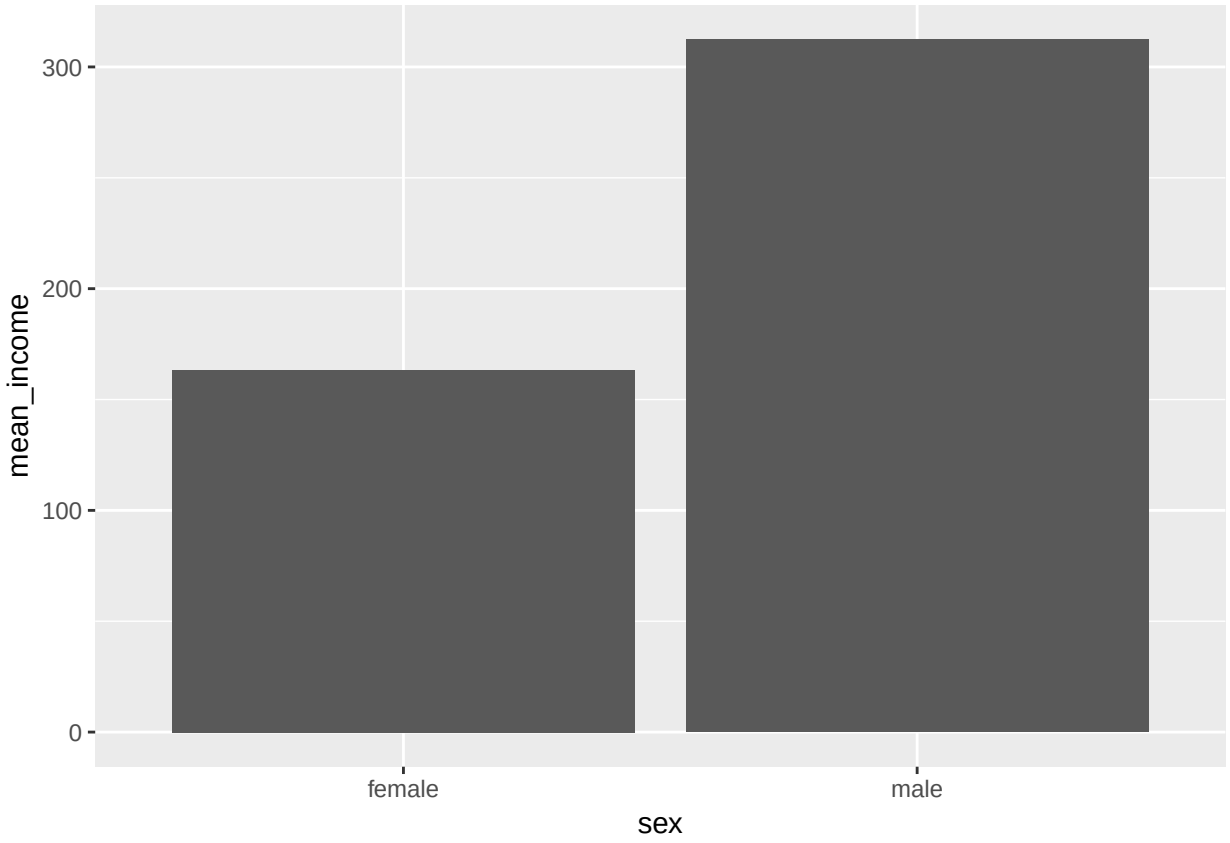
Preprocessing of sex and income are done. Let's create a table of average income per sex and its graph.

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

sex_income
```

```
## # A tibble: 2 x 2
##   sex    mean_income
##   <chr>        <dbl>
## 1 female        163.
## 2 male          312.
```

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



This tells me that, there exists a "huge" difference in amount of income received by females and males. Average income of females is 163 whereas average income of males is 312, so females are receiving 52 per cent of males' income.

2. At which age do people make the most income?

Using factors: income, birth Observation and preprocessing of income factor is unnecessary, so let's dive into age. Birth is the only factor that's relevant to age so this is where I should start.

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

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

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1907    1946    1966    1968    1988    2014
```

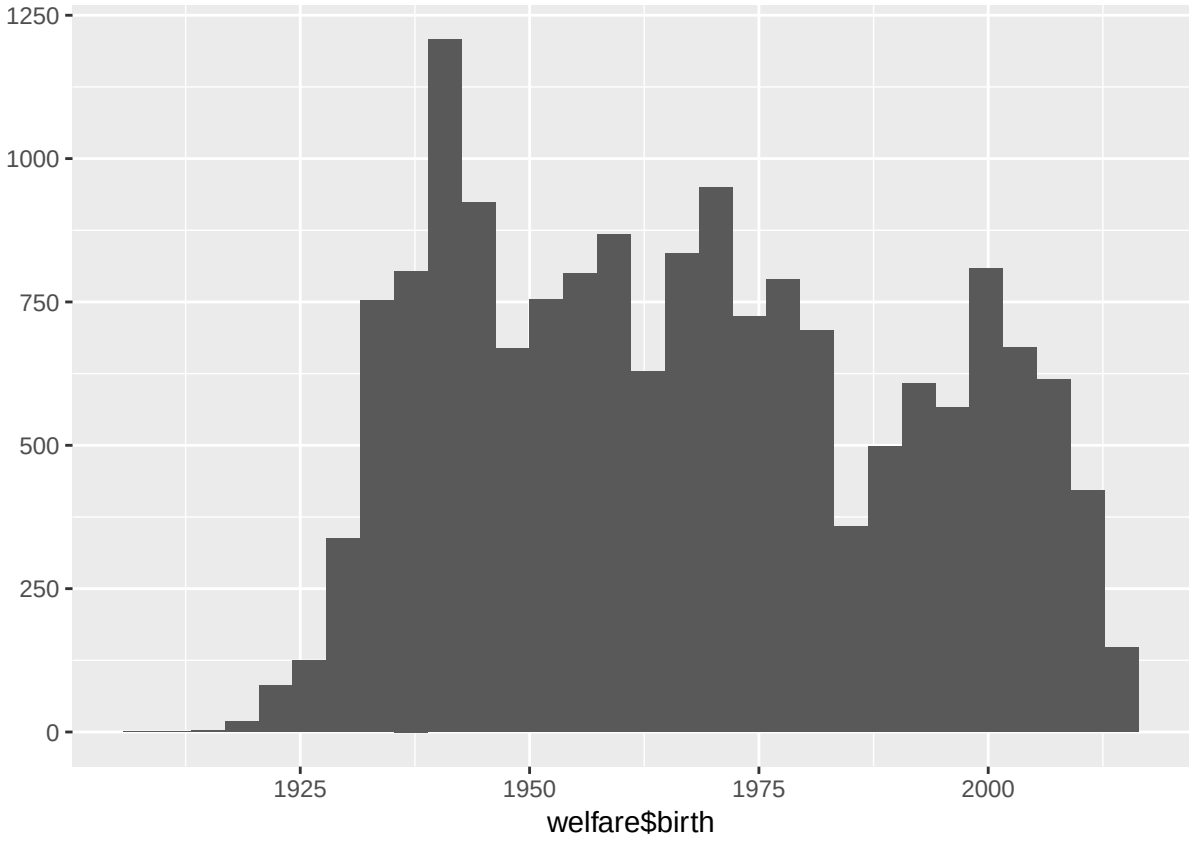
```
table(is.na(welfare$birth))
```

```
##
## FALSE
## 16664
```

According to the codebook, the range should be 1900~2014 and NA is for no answer. It seems like there is no NA value and all the values are within the range.

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



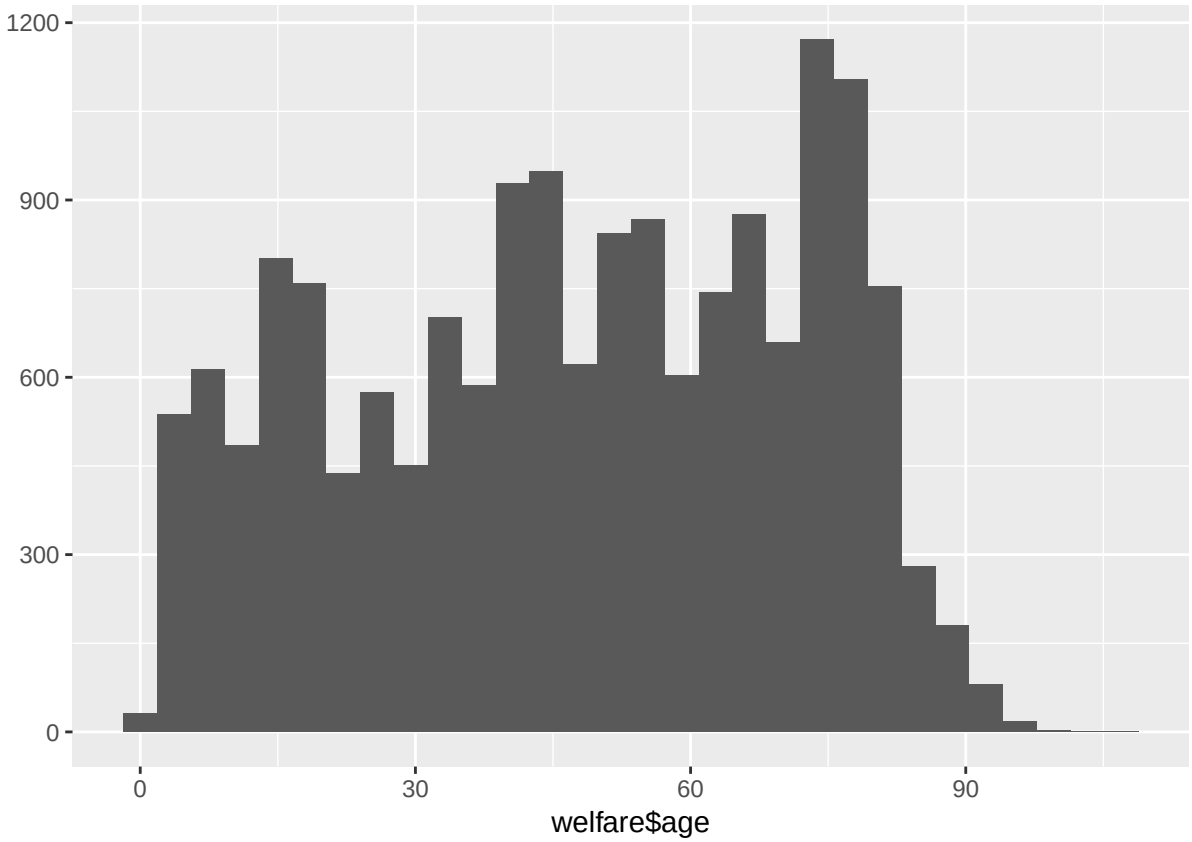
Birth factor has no problems. But I want the age of people, not their year of birth. Since this dataset was published at 2015, I will find out their ages at 2015 for accuracy.

```
welfare$age <- 2015-welfare$birth
summary(welfare$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   27.00   49.00   47.43   69.00  108.00
```

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



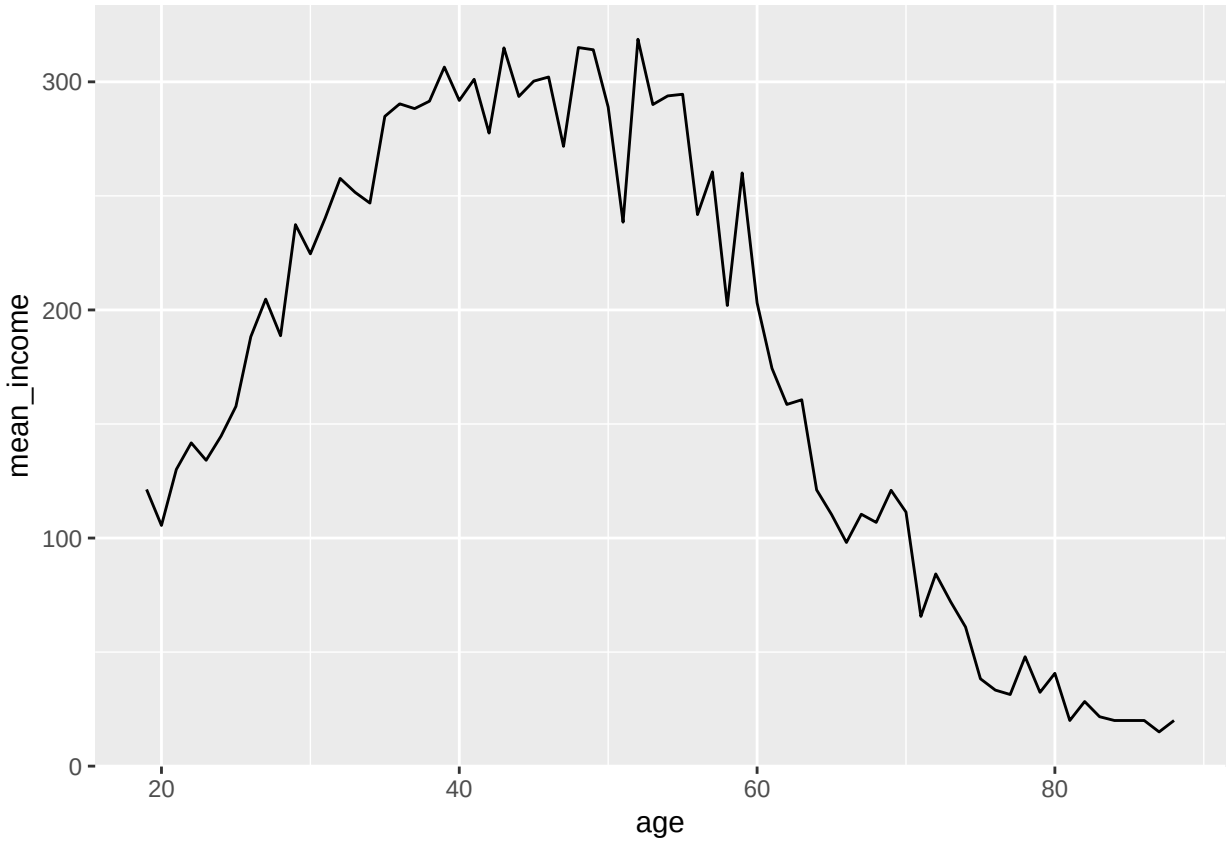
Preprocessing is done. Let's create the table of average income per age and its graph.

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

age_income
```

```
## # A tibble: 69 x 2
##      age mean_income
##    <dbl>       <dbl>
##  1    19        121.
##  2    20        106.
##  3    21        130.
##  4    22        142.
##  5    23        134.
##  6    24        145.
##  7    25        158.
##  8    26        188.
##  9    27        205.
## 10    28        189.
## # ... with 59 more rows
```

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



This shows that - People starts to earn money at the age of 19 in Korea - 20s generally earn 100 to 200 - 30s generally earn 200 to 300 - People earn the most money at 40s and early 50s - Amount of income decreases rapidly from late 50s - From late 60s, people earn less than 20s

3. Which age group earns more?

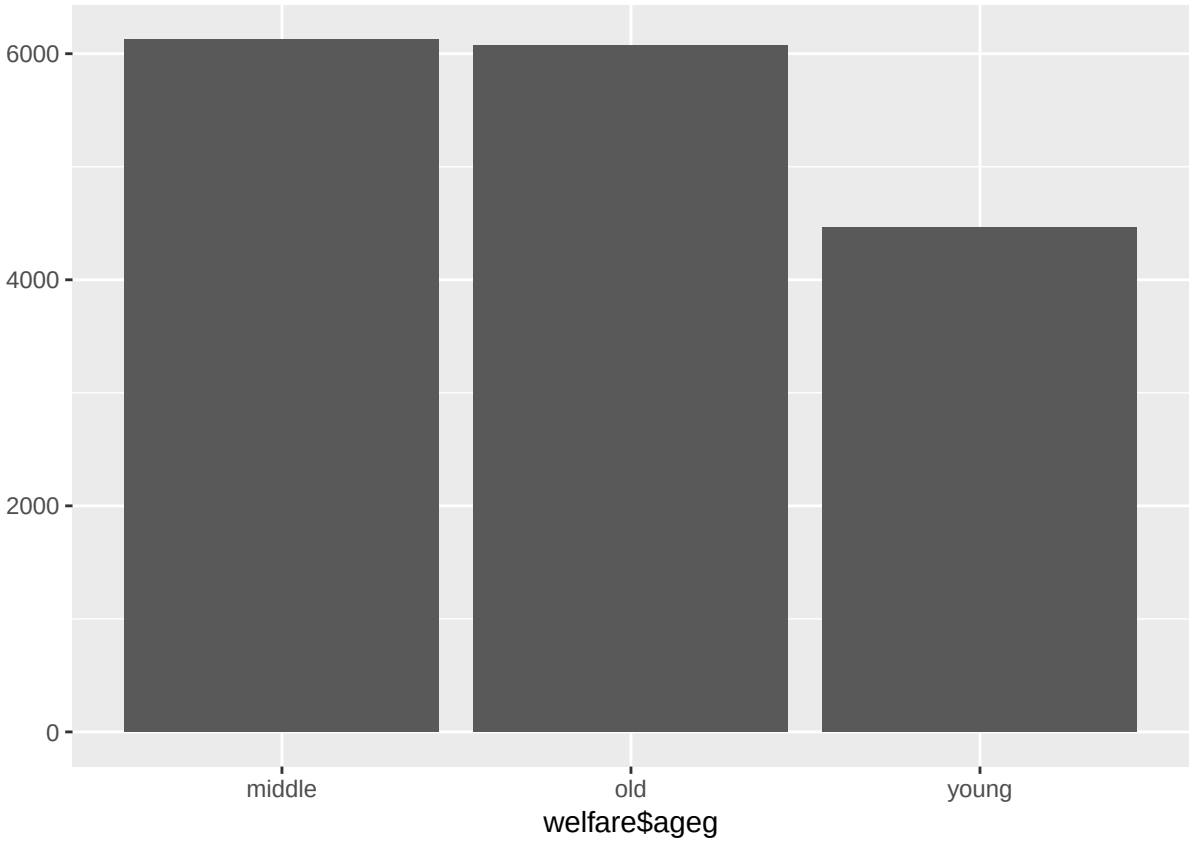
Using factors: income, age It is similar to section 2, but this time I would like to group the ages into three groups. - Young: < 30 - Middle: 30~59 - Old: >= 60

```
welfare <- welfare %>%
  mutate(ageg = ifelse(age < 30, "young",
                       ifelse(age >= 60, "old", "middle")))

table(welfare$ageg)
```

```
##
## middle    old  young
##   6128   6072   4464
```

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



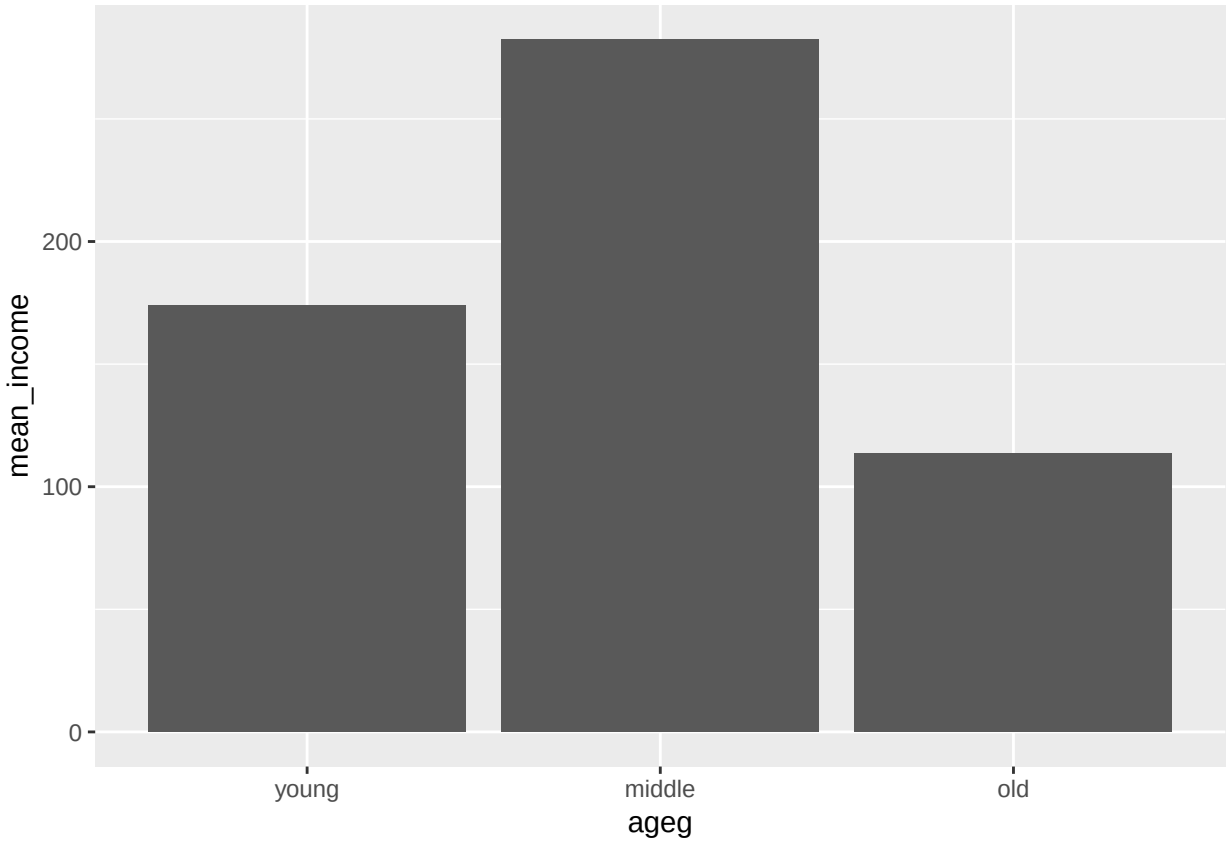
Middle and Old occupy the high ratio than young. This illustrates the Aging Society of Korea.

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

ageg_income
```

```
## # A tibble: 3 x 2
##   ageg   mean_income
##   <chr>        <dbl>
## 1 middle        282.
## 2 old           114.
## 3 young         174.
```

```
ggplot(data = ageg_income, aes(x=ageg, y=mean_income)) +
  geom_col() +
  scale_x_discrete(limits = c("young", "middle", "old"))
```



This result coorelates to my analysis on section 2, and people between 30~59 make the best profit.

4. Incomes by sex differ by age group?

Using factors: income, ageg, sex Aim is to upgrade section 3 by adding the sex factor, find out the average income of females and males on different age groups.

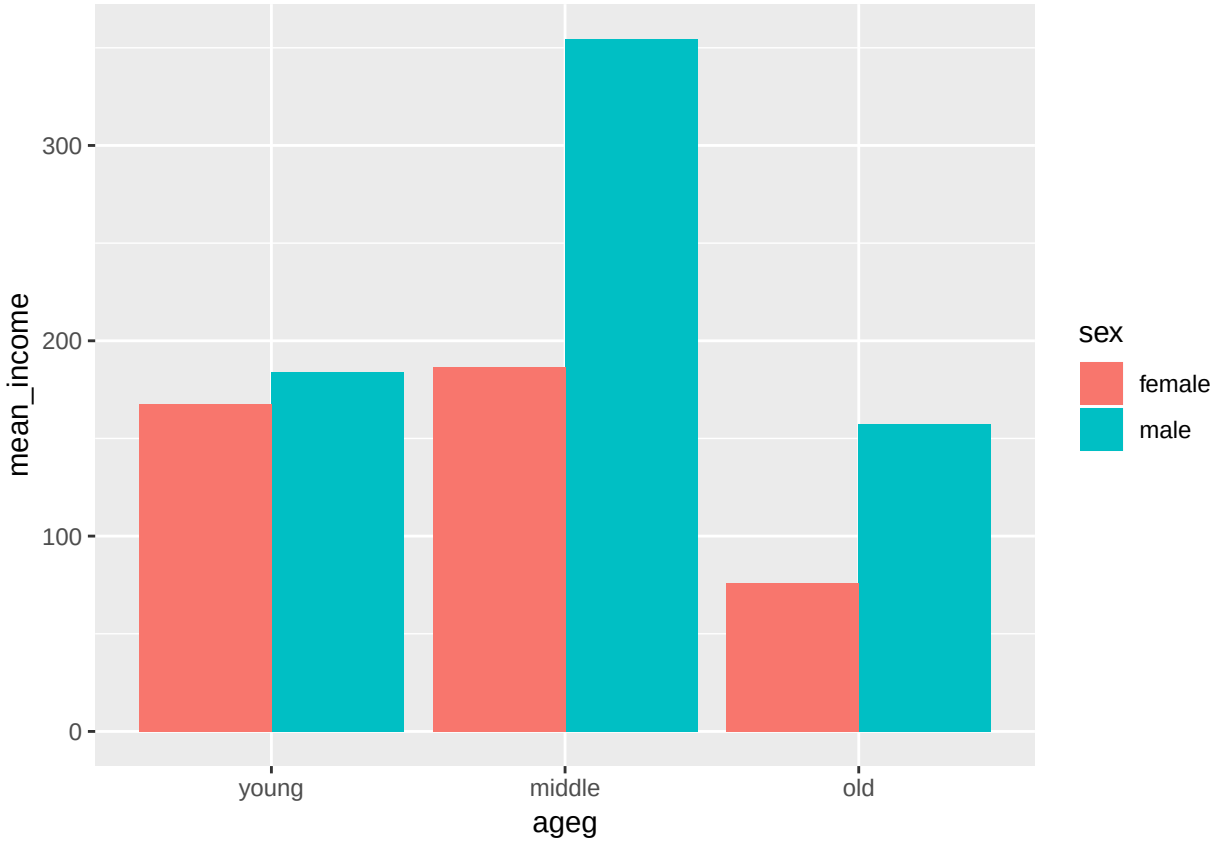
```
sex_income <- welfare %>%
  filter(!is.na(income)) %>%
  group_by(ageg, sex) %>%
  summarise(mean_income = mean(income))

sex_income
```

```
## # A tibble: 6 x 3
## # Groups:   ageg [?]
##   ageg   sex    mean_income
##   <chr>  <chr>        <dbl>
## 1 middle female       186.
## 2 middle male         355.
## 3 old    female        75.9
## 4 old    male         157.
## 5 young  female       168.
## 6 young  male         184.
```

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

```
  scale_x_discrete(limits = c("young", "middle", "old"))
```



- Young group does not have much difference between the males and the females
- The difference gets huge in Middle and Old group, about double in amount
- Not much difference in income between Young and Middle female groups compare to male groups
- Old group earns less than Young group

5. People with religion have lower divorce rates?

Using factors: religion, marriage Aim is to find out whether the people with religion has lower divorce rates than the people with no religion. Both religion and marriage factors need to be examined and proprocessed before analysing.

```
class(welfare$religion)
```

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

```
table(welfare$religion)
```

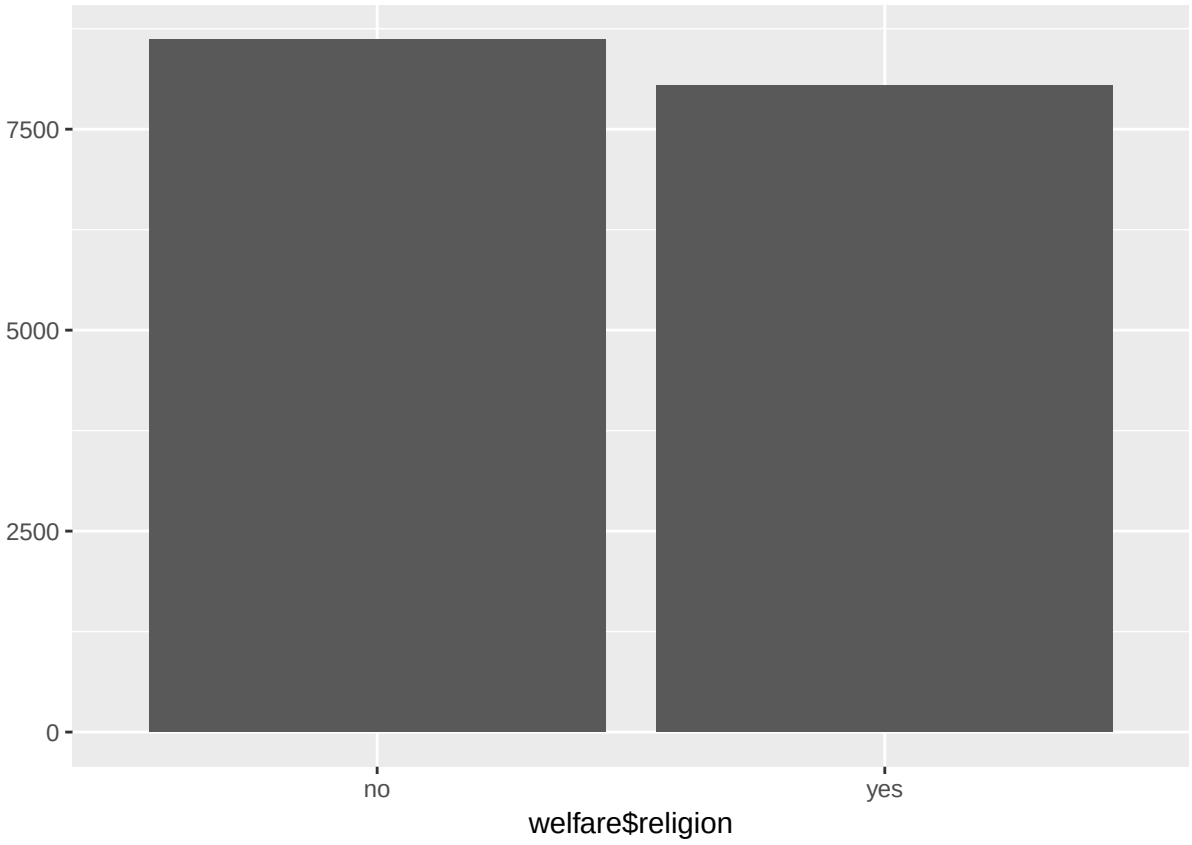
```
##
##    1    2
## 8047 8617
```

1 is for "has religion" and 2 is for "no religion" according to the codebook. Let's give yes and no for convenience.

```
welfare$religion <- ifelse(welfare$religion == 1, "yes", "no")
table(welfare$religion)
```

```
##
##   no  yes
## 8617 8047
```

```
qplot(welfare$religion)
```



Evenly distributed, the difference is less than 600.

```
class(welfare$marriage)
```

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

```
table(welfare$marriage)
```

```
##
##    0    1    2    3    4    5    6
## 2861 8431 2117  712   84 2433   26
```

According to the codebook, - 0: less than 18 - 1: married - 2: separation by death - 3: divorced - 4: living separately - 5: not married (over 18) - 6: etc

Among these, I only need 1 and 3.

```
welfare$group_marriage <- ifelse(welfare$marriage == 1, "married",
                                 ifelse(welfare$marriage == 3, "divorced", NA))

table(welfare$group_marriage)
```

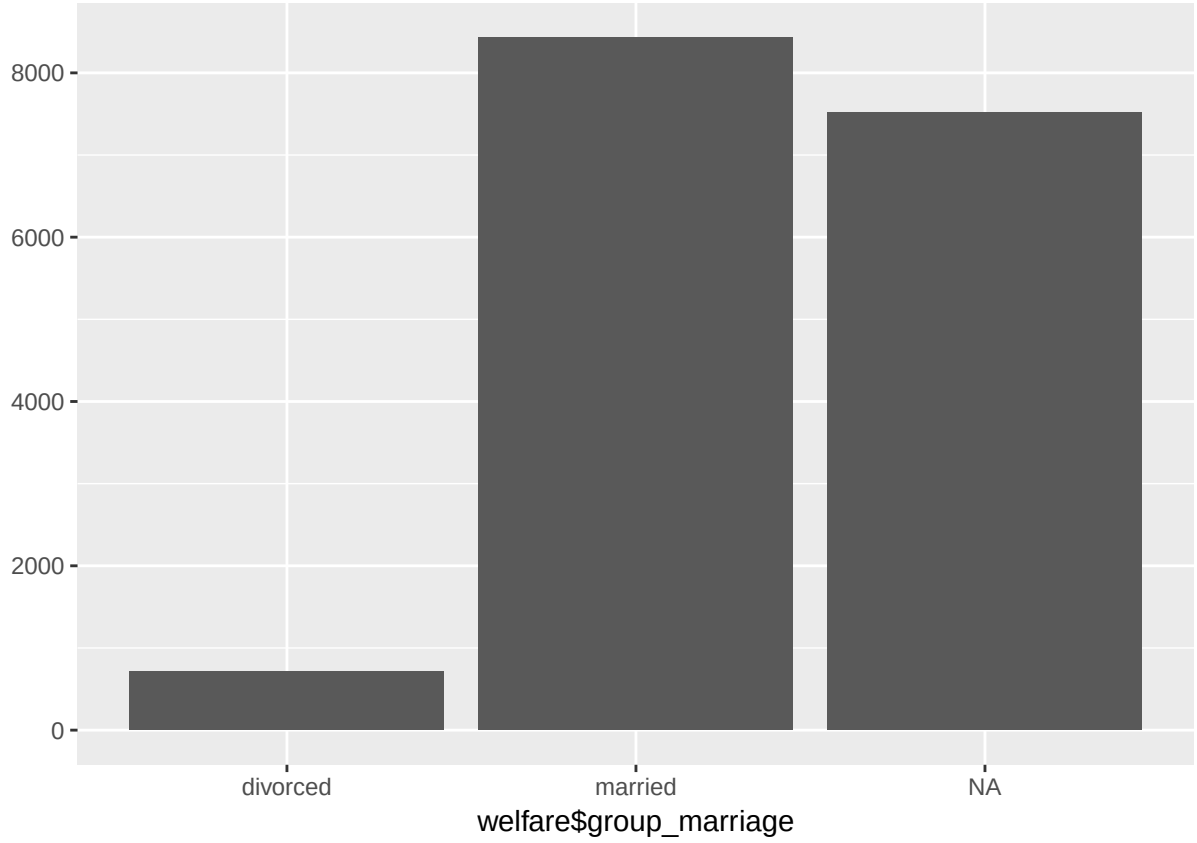
```
##
## divorced  married
```

```
##      712     8431
```

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

```
##
## FALSE  TRUE
##  9143  7521
```

```
qplot(welfare$group_marriage)
```



There exists 7521 NA values. These values will be discarded in the analysis.

```
religion_marriage <- welfare %>%
  filter(!is.na(group_marriage)) %>%
  group_by(religion, group_marriage) %>%
  summarise(n=n()) %>%
  mutate(tot_group = sum(n)) %>%
  mutate(pct = round(n/tot_group*100, 1))

religion_marriage
```

```
## # A tibble: 4 x 5
## # Groups:   religion [2]
##   religion group_marriage     n tot_group   pct
##   <chr>    <chr>          <int>     <int> <dbl>
## 1 no       divorced         384      4602   8.3
## 2 no       married         4218      4602  91.7
## 3 yes      divorced         328      4541   7.2
```

```
## 4 yes      married         4213      4541  92.8
```

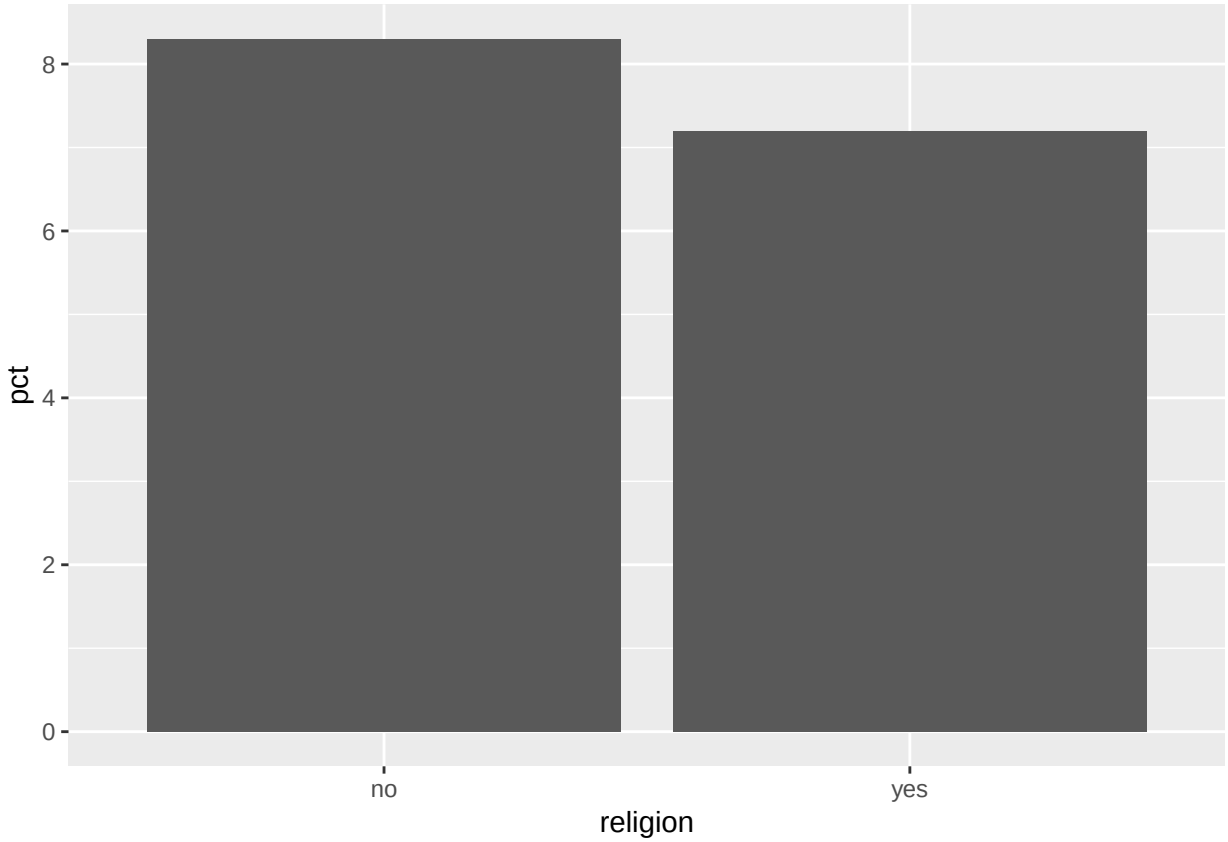
Extracting "divorced" state only

```
divorced <- religion_marriage %>%
  filter(group_marriage == "divorced") %>%
  select(religion, pct)

divorced
```

```
## # A tibble: 2 x 2
## # Groups:   religion [2]
##   religion   pct
##   <chr>    <dbl>
## 1 no         8.3
## 2 yes        7.2
```

```
ggplot(data = divorced, aes(x=religion, y=pct)) +
  geom_col()
```



- With religion: divorce rate is 7.2%
- No religion: divorce rate is 8.3% I can conclude that the people with religion has lower divorce rate.

6. Divorce rate by religion differ by age group?

Using factors: ageg, religion, group_marriage Firstly I will investigate the divorce rate by age group.

```
ageg_marriage <- welfare %>%
  filter(!is.na(group_marriage)) %>%
```

```
  group_by(ageg, group_marriage) %>%
  summarise(n=n()) %>%
  mutate(tot_group = sum(n)) %>%
  mutate(pct = round(n/tot_group*100, 1))

ageg_marriage
```

```
## # A tibble: 6 x 5
## # Groups:   ageg [3]
##   ageg   group_marriage     n tot_group   pct
##   <chr>  <chr>          <int>     <int> <dbl>
## 1 middle divorced         469      5067   9.3
## 2 middle married         4598      5067  90.7
## 3 old    divorced         241      3981   6.1
## 4 old    married         3740      3981  93.9
## 5 young  divorced           2        95   2.1
## 6 young  married           93        95  97.9
```

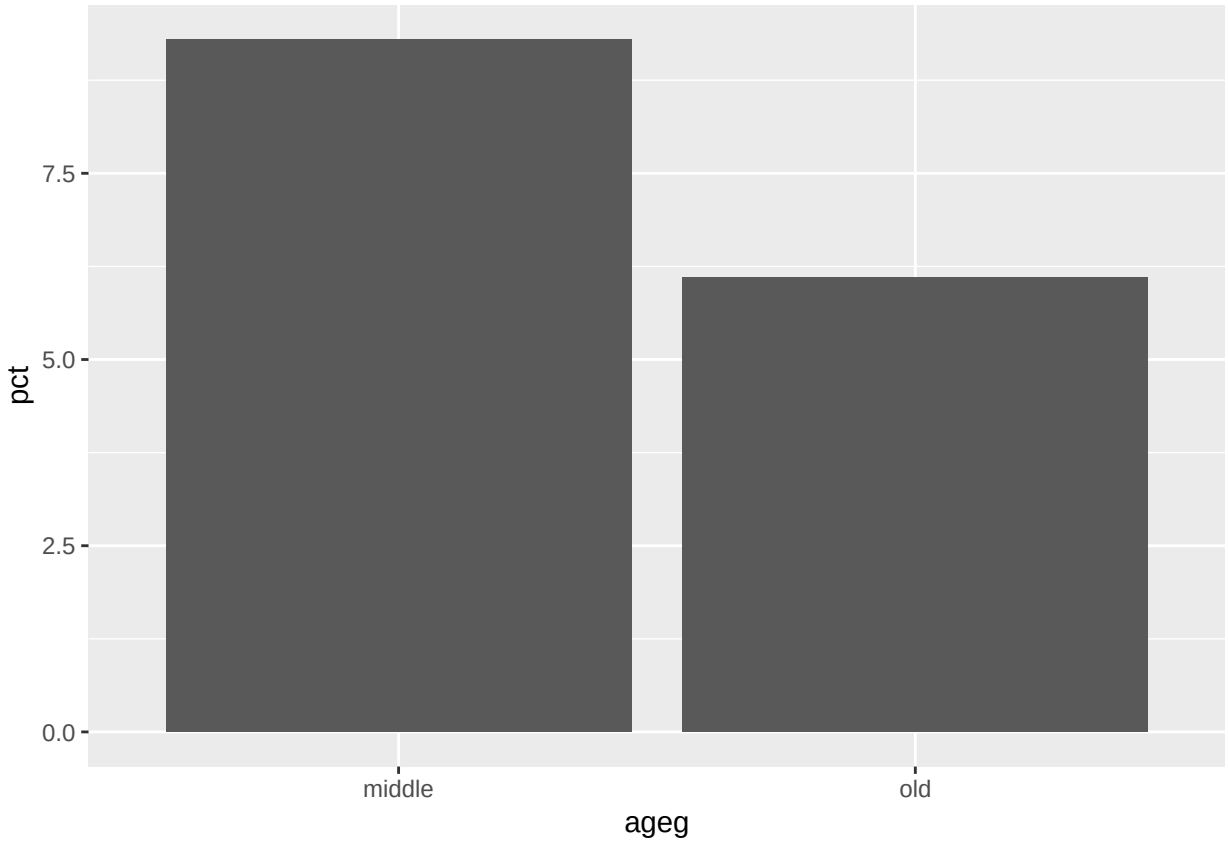
It's visible that Young group doesn't have much incidents compare to the other groups and thus let's omit this group for this analysis.

```
ageg_divorced <- ageg_marriage %>%
  filter(ageg != "young" & group_marriage == "divorced") %>%
  select(ageg, pct)

ageg_divorced
```

```
## # A tibble: 2 x 2
## # Groups:   ageg [2]
##   ageg     pct
##   <chr>  <dbl>
## 1 middle   9.3
## 2 old      6.1
```

```
ggplot(data = ageg_divorced, aes(x=ageg, y=pct)) +
  geom_col()
```



People in Middle group have higher divorce rate. This result is predictable as married young couples tend to get separated easily than the elders.

And now we are combining all three factors together to do the analysis.

```
ageg_religion_marriage <- welfare %>%
  filter(!is.na(group_marriage) & ageg != "young") %>%
  group_by(ageg, religion, group_marriage) %>%
  summarise(n=n()) %>%
  mutate(tot_group = sum(n)) %>%
  mutate(pct = round(n/tot_group*100, 1))

ageg_religion_marriage
```

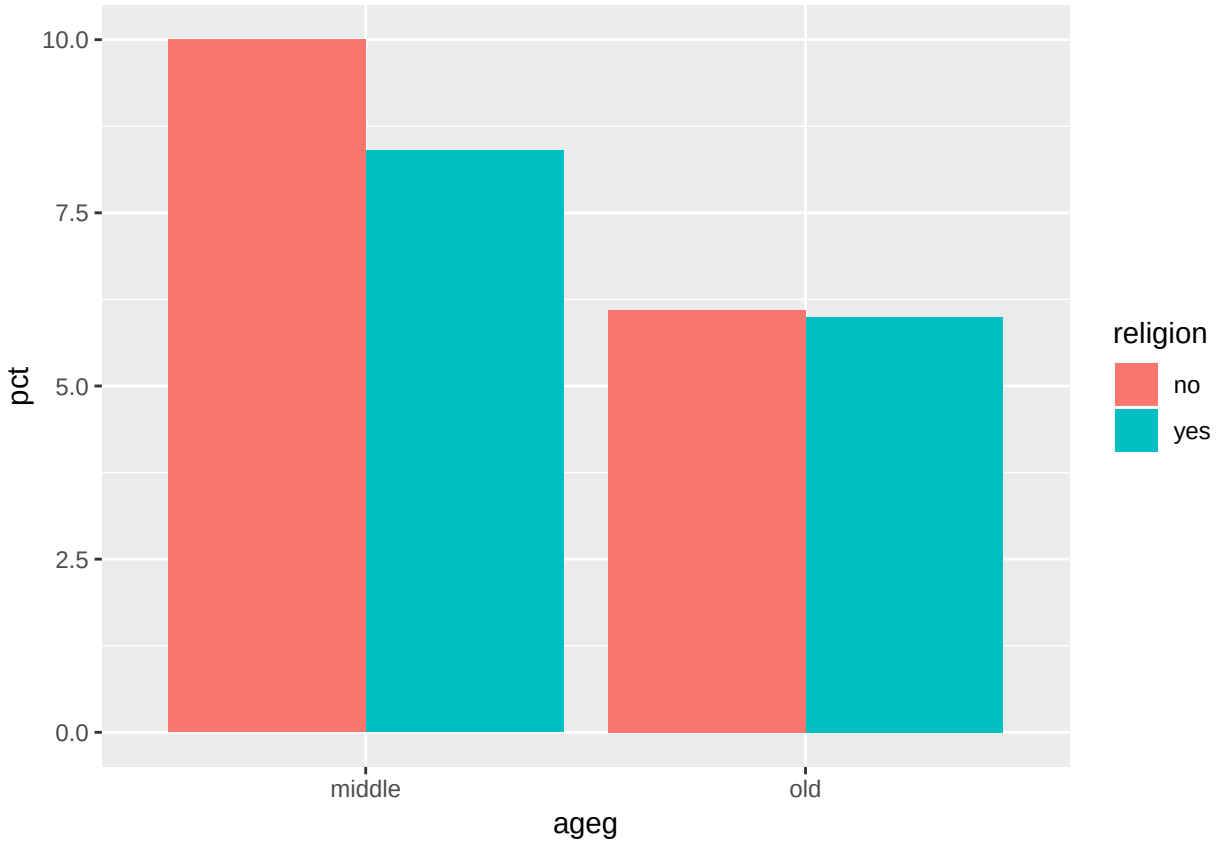
```
## # A tibble: 8 x 6
## # Groups:   ageg, religion [4]
##   ageg   religion group_marriage     n tot_group   pct
##   <chr>  <chr>    <chr>          <int>     <int> <dbl>
## 1 middle no       divorced         274      2747  10
## 2 middle no       married         2473      2747  90
## 3 middle yes      divorced         195      2320   8.4
## 4 middle yes      married         2125      2320  91.6
## 5 old    no       divorced         109      1795   6.1
## 6 old    no       married         1686      1795  93.9
## 7 old    yes      divorced         132      2186   6
## 8 old    yes      married         2054      2186  94
```

```
df_divorced <- ageg_religion_marriage %>%
  filter(group_marriage == "divorced") %>%
  select(ageg, religion, pct)

df_divorced
```

```
## # A tibble: 4 x 3
## # Groups:   ageg, religion [4]
##   ageg   religion   pct
##   <chr>  <chr>    <dbl>
## 1 middle no        10
## 2 middle yes        8.4
## 3 old    no         6.1
## 4 old    yes        6
```

```
ggplot(data = df_divorced, aes(x=ageg, y=pct, fill=religion)) +
  geom_col(position = "dodge")
```



- In Old group, religion doesn't really affect the divorce rate (less than 0.1% difference)
- In Middle group, difference is greater than 2.5%, people with no religion have higher divorce rate.

7. Areas with more elders?

Using factors: ageg, code_region Aim is to find out the regions with higher ratio of elder population.

```
class(welfare$code_region)
```

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

```
table(welfare$code_region)
```

```
##
##    1    2    3    4    5    6    7
## 2486 3711 2785 2036 1467 1257 2922
```

According to the codebook, - 1: Seoul - 2: Incheon/Gyeongi - 3: Busan/Gyeongnam/Ulsan - 4: Degu/Gyeongbuk - 5: Dejun/Chungnam - 6: Gangwon/Chungbuk - 7: Gwangju/Jeonnam/Jeonbuk/Jeju

```
list_region <- data.frame(code_region = c(1:7),
                          region = c("Seoul",
                                     "Incheon/Gyeongi",
                                     "Busan/Gyeongnam/Ulsan",
                                     "Degu/Gyeongbuk",
                                     "Dejun/Chungnam",
                                     "Gangwon/Chungbuk",
                                     "Gwangju/Jeonnam/Jeonbuk/Jeju"))

list_region
```

```
##   code_region                      region
## 1           1                       Seoul
## 2           2              Incheon/Gyeongi
## 3           3        Busan/Gyeongnam/Ulsan
## 4           4               Degu/Gyeongbuk
## 5           5               Dejun/Chungnam
## 6           6             Gangwon/Chungbuk
## 7           7 Gwangju/Jeonnam/Jeonbuk/Jeju
```

This information will be left-joined with our dataset.

```
welfare <- left_join(welfare, list_region, by="code_region")

welfare %>%
  select(code_region, region) %>%
  head()
```

```
##   code_region region
## 1           1  Seoul
## 2           1  Seoul
## 3           1  Seoul
## 4           1  Seoul
## 5           1  Seoul
## 6           1  Seoul
```

Let's create the table that contains the ratio of age groups by regions.

```
region_ageg <- welfare %>%
  group_by(region, ageg) %>%
  summarise(n=n()) %>%
  mutate(tot_group = sum(n)) %>%
  mutate(pct = round(n/tot_group*100, 1))

region_ageg
```

```
## # A tibble: 21 x 5
## # Groups:   region [7]
```

```
##    region                ageg       n tot_group   pct
##    <fct>                 <chr>  <int>     <int> <dbl>
##  1 Busan/Gyeongnam/Ulsan middle   992      2785  35.6
##  2 Busan/Gyeongnam/Ulsan old     1094      2785  39.3
##  3 Busan/Gyeongnam/Ulsan young    699      2785  25.1
##  4 Degu/Gyeongbuk        middle   651      2036  32
##  5 Degu/Gyeongbuk        old      904      2036  44.4
##  6 Degu/Gyeongbuk        young    481      2036  23.6
##  7 Dejun/Chungnam        middle   557      1467  38
##  8 Dejun/Chungnam        old      509      1467  34.7
##  9 Dejun/Chungnam        young    401      1467  27.3
## 10 Gangwon/Chungbuk      middle   431      1257  34.3
## # ... with 11 more rows
```

```
list_order_old <- region_ageg %>%
  filter(ageg == "old") %>%
  arrange(pct)

list_order_old
```

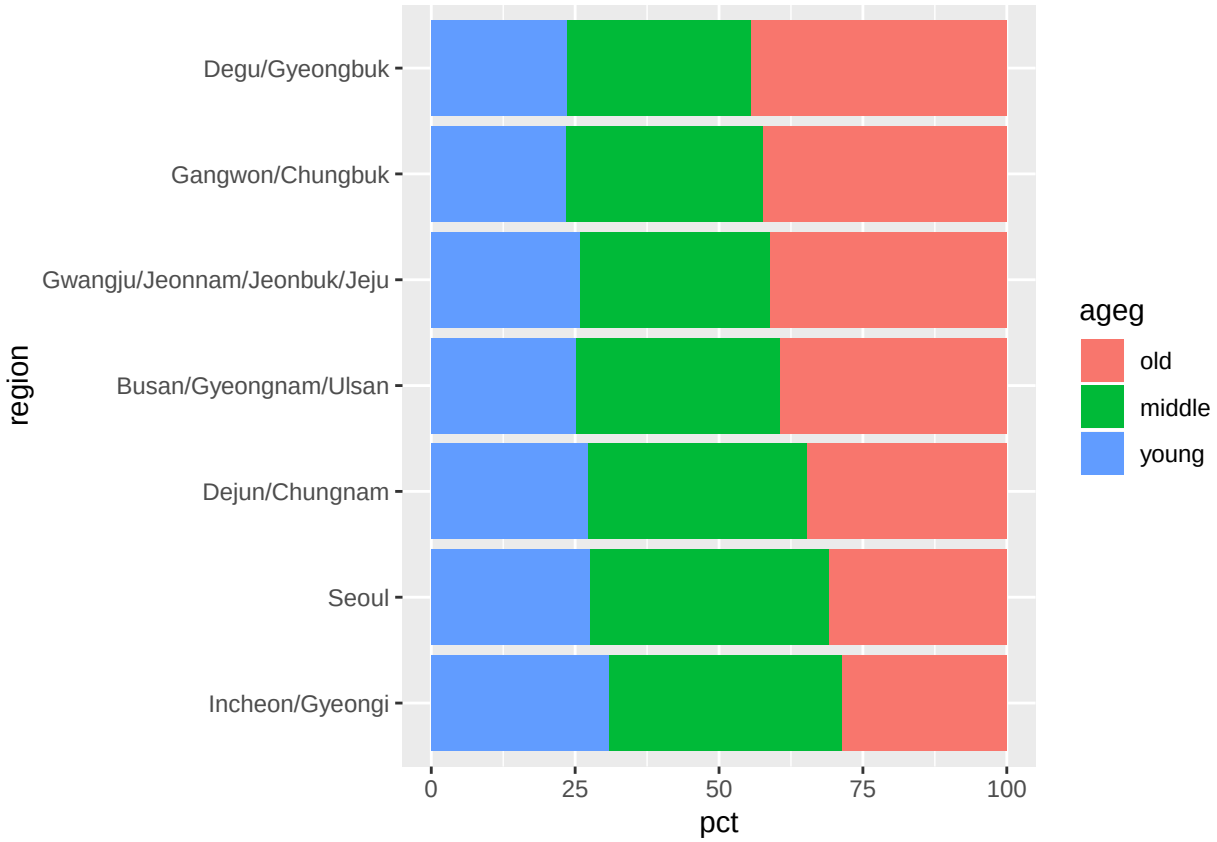
```
## # A tibble: 7 x 5
## # Groups:   region [7]
##   region                       ageg      n tot_group   pct
##   <fct>                        <chr> <int>     <int> <dbl>
## 1 Incheon/Gyeongi              old    1063      3711  28.6
## 2 Seoul                        old     769      2486  30.9
## 3 Dejun/Chungnam               old     509      1467  34.7
## 4 Busan/Gyeongnam/Ulsan        old    1094      2785  39.3
## 5 Gwangju/Jeonnam/Jeonbuk/Jeju old    1201      2922  41.1
## 6 Gangwon/Chungbuk             old     532      1257  42.3
## 7 Degu/Gyeongbuk               old     904      2036  44.4
```

```
order <- list_order_old$region
order
```

```
## [1] Incheon/Gyeongi              Seoul
## [3] Dejun/Chungnam               Busan/Gyeongnam/Ulsan
## [5] Gwangju/Jeonnam/Jeonbuk/Jeju Gangwon/Chungbuk
## [7] Degu/Gyeongbuk
## 7 Levels: Busan/Gyeongnam/Ulsan Degu/Gyeongbuk ... Seoul
```

```
region_ageg$ageg <- factor(region_ageg$ageg,
                           level = c("old", "middle", "young"))

ggplot(data = region_ageg, aes(x=region, y=pct, fill=ageg)) +
  geom_col() +
  coord_flip() +
  scale_x_discrete(limits = order)
```



- Degu/Gyeongbuk area has the highest ratio of elder population
- Seoul and Inchoen/Gyeongi areas are the capital and the major cities of Korea, and they have higher ratio of younger people.