문제 1

In [1]:

```
library(tidyverse)
load("../input/welfare2016.rda")
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
✔ ggplot2 3.1.0.9000     ✔ purrr   0.3.1
✔ tibble  2.0.1          ✔ dplyr   0.8.0.1
✔ tidyr   0.8.3          ✔ stringr 1.4.0
✔ readr   1.3.1          ✔ forcats 0.4.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

In [2]:

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

'numeric'

```
   1    2
7659 8330


FALSE
15989


< table of extent 0 >
```

In [3]:

```
welfare %<>% mutate(sex=ifelse(sex==1, "male", "female")) %>%
             mutate(income=ifelse(income==0, NA, income)) %>%
             mutate(age = 2016-birth+1) %>%
             mutate(ageg = ifelse(age < 30, 'young',
                                  ifelse(age < 60, 'middle', 'old'))) %>%
             mutate(code_job = as.character(code_job)) %>%
             mutate(code_job = ifelse(str_length(code_job)==3,
                                      str_c("0", code_job),
                                      code_job))%>%
             mutate(religion=ifelse(religion==1, "yes", "no")) %>%
             mutate(group_marriage=(ifelse(marriage==1, "marriage",
                                    ifelse(marriage==3, "divorce", NA)))) %>%

             select(sex,
                    birth,
                    age,
                    ageg,
                    marriage,
                    group_marriage,
                    religion,
                    income,
                    code_job,
                    code_region)
```

In [4]:

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

'character'

```
  no  yes
8330 7659


FALSE
15989


 divorce marriage
     701     8058
```

1) group by

In [5]:

```
welfare %>% filter(!is.na(group_marriage)) %>%
    filter(!is.na(religion)) %>%
    group_by(religion, group_marriage) %>%
    summarise(n = n()) %>%
    mutate(tot=sum(n)) %>%
    mutate(ratio=n/tot*100) %>%
    filter(group_marriage=='divorce') -> divorceratiobyreligion1

divorceratiobyreligion1
divorceratiobyreligion1 %>%
    ggplot(aes(religion, ratio)) + geom_col()
```

religion	group_marriage	n	tot	ratio
no	divorce	379	4382	8.649019
yes	divorce	322	4377	7.356637

2) count

In [6]:

```
welfare %>% filter(!is.na(group_marriage)) %>%
    filter(!is.na(religion)) %>%
    count(religion, group_marriage) %>%
    group_by(religion) %>%
    mutate(tot=sum(n)) %>%
    mutate(ratio=n/tot*100) %>%
    filter(group_marriage=='divorce') -> divorceratiobyreligion2

divorceratiobyreligion2
```

religion	group_marriage	n	tot	ratio
no	divorce	379	4382	8.649019
yes	divorce	322	4377	7.356637

3) SQL

In [7]:

```
library(sqldf)

sqldf("
      select x.religion, x.group_marriage, n, tot, 1.*n/tot*100 as ratio
      from
      (select religion, group_marriage, count(*) as n
      from welfare
      where group_marriage is not null
      group by religion,group_marriage) x,
      (select religion, group_marriage, count(*) as tot
      from welfare
      where religion is not null and
      group_marriage is not null
      group by religion) y
      where x.religion=y.religion and
      x.group_marriage == 'divorce'
      ") -> divorceratiobyreligion3

divorceratiobyreligion3
```

```
Loading required package: gsubfn
Loading required package: proto
Warning message:
"no DISPLAY variable so Tk is not available"Loading required package: RSQLite
```

religion	group_marriage	n	tot	ratio
no	divorce	379	4382	8.649019
yes	divorce	322	4377	7.356637

In [8]:

```
all(divorceratiobyreligion1==divorceratiobyreligion2)
all(divorceratiobyreligion1==divorceratiobyreligion3)
```

TRUE

TRUE

문제 2

1) group by

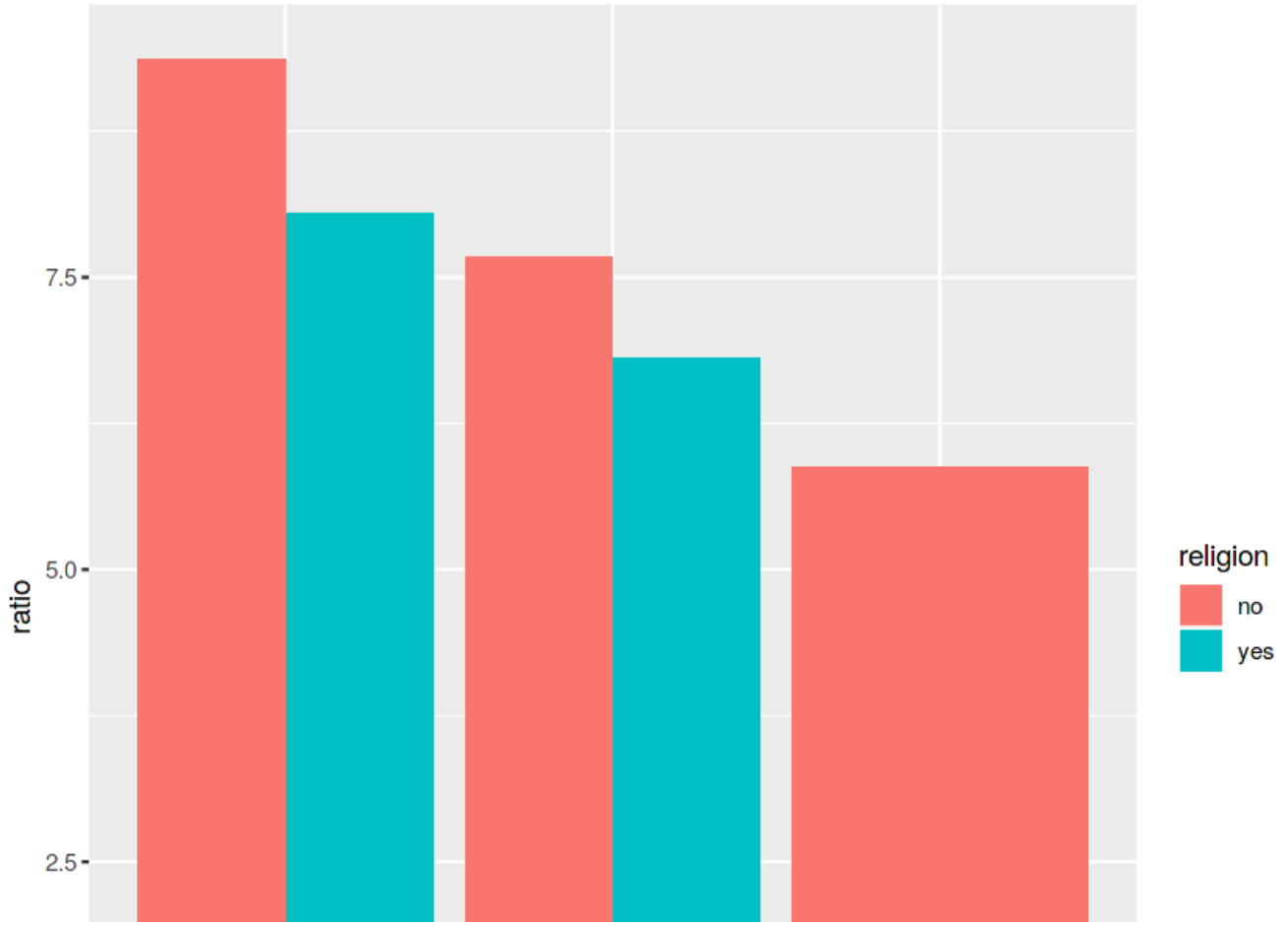
In [9]:

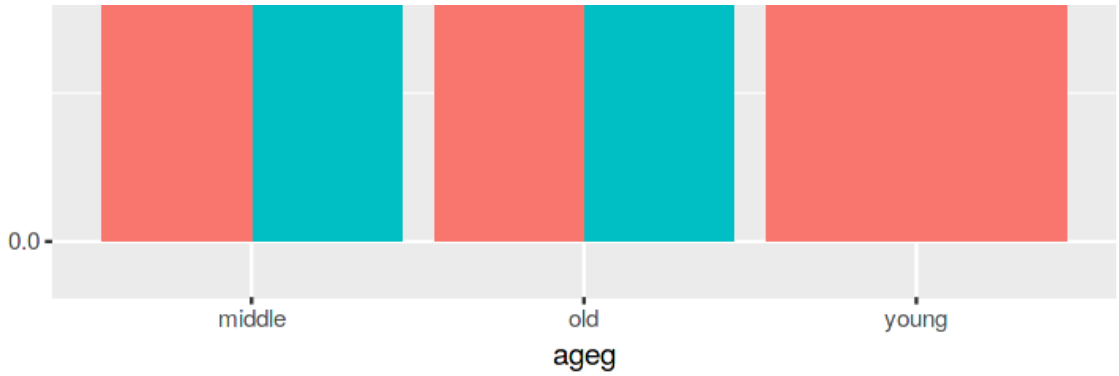
```
welfare %>% filter(!is.na(religion)) %>%
    filter(!is.na(ageg)) %>%
    filter(!is.na(group_marriage)) %>%
    group_by(ageg,religion,group_marriage) %>%
    summarise(n=n()) %>%
    mutate(tot=sum(n)) %>%
    mutate(ratio=n/tot*100) %>%
    filter(group_marriage=="divorce") -> divorceratiobyreligionbyageg1

divorceratiobyreligionbyageg1

divorceratiobyreligionbyageg1  %>% ggplot(aes(x=ageg,y=ratio, fill=religion))+ geom_col(position =
"dodge")
```

ageg	religion	group_marriage	n	tot	ratio
middle	no	divorce	241	2573	9.366498
middle	yes	divorce	165	2049	8.052709
old	no	divorce	135	1758	7.679181
old	yes	divorce	157	2308	6.802426
young	no	divorce	3	51	5.882353





2) count

In [10]:

```
welfare %>% filter(!is.na(religion)) %>%
    filter(!is.na(ageg)) %>%
    filter(!is.na(group_marriage)) %>%
    count(ageg, religion, group_marriage) %>%
    group_by(ageg, religion)  %>%
    mutate(tot=sum(n)) %>%
    mutate(ratio=n/tot*100) %>%
    filter(group_marriage=="divorce") -> divorceratiobyreligionbyageg2

divorceratiobyreligionbyageg2
```

ageg	religion	group_marriage	n	tot	ratio
middle	no	divorce	241	2573	9.366498
middle	yes	divorce	165	2049	8.052709
old	no	divorce	135	1758	7.679181
old	yes	divorce	157	2308	6.802426
young	no	divorce	3	51	5.882353

3) SQL

In [11]:

```
sqldf("select x.ageg, x.religion, x.group_marriage ,n,tot, 1.*n/tot*100 as ratio
      from
      (
      select ageg, religion, group_marriage, count(group_marriage) as n
      from welfare
      where group_marriage is not null
      group by ageg,religion,group_marriage
      ) x,
      (
      select ageg, religion, group_marriage ,count(*) as tot
      from welfare
      where group_marriage is not null
      group by ageg,religion
      ) y
      where x.ageg=y.ageg and
      x.religion=y.religion and
      x.group_marriage == 'divorce'
      ") ->divorceratiobyreligionbyageg3

divorceratiobyreligionbyageg3
```

ageg	religion	group_marriage	n	tot	ratio
middle	no	divorce	241	2573	9.366498
middle	yes	divorce	165	2049	8.052709
old	no	divorce	135	1758	7.679181

ageg	religion	group_marriage	n	tot	ratio
old	yes	divorce	157	2308	6.802426
young	no	divorce	3	51	5.882353

In [12]:

```
all(divorceratiobyreligionbyageg1==divorceratiobyreligionbyageg2)
all(divorceratiobyreligionbyageg1==divorceratiobyreligionbyageg3)
```

TRUE

TRUE

문제 3

In [13]:

```
region7 = data.frame(code_region=c(1,2,3,4,5,6,7),
                     region=c("서울",
                              "수도권(인천/경기)",
                              "부산/경남/울산",
                              "대구/경북",
                              "대전/충남",
                              "강원/충북",
                              "광주/전남/전북/제주도"),stringsAsFactors = F)


 region7
```

code_region	region
1	서울
2	수도권(인천/경기)
3	부산/경남/울산
4	대구/경북
5	대전/충남
6	강원/충북
7	광주/전남/전북/제주도

1) dplyr

In [14]:

```
welfare %>% filter(!is.na(ageg)) %>%
            filter(!is.na(code_region)) %>%
            count(code_region,ageg) %>%
            left_join(region7, by="code_region") %>%
            group_by(region) %>%
            mutate(tot=sum(n)) %>%
            mutate(ratio=n/tot*100) %>%
            select(region,ageg,ratio) -> ratiobyagegbyregion


ratiobyagegbyregion %>% filter(ageg=='old') %>% arrange(ratio) -> sort1
sort1
```

region	ageg	ratio
수도권(인천/경기)	old	30.29461
서울	old	33.23224
대전/충남	old	36.98332
부산/경남/울산	old	41.55796
광주/전남/전북/제주도	old	42.69351

region	ageg	ratio
강원/충북	old	45.57709
대구/경북	old	46.28776

2) SQL

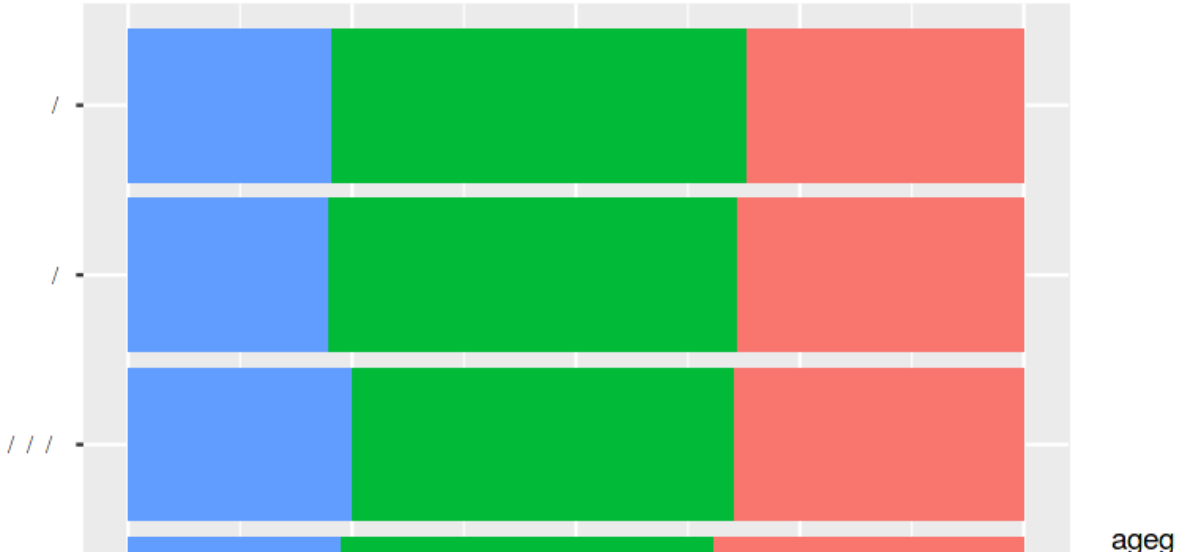
In [15]:

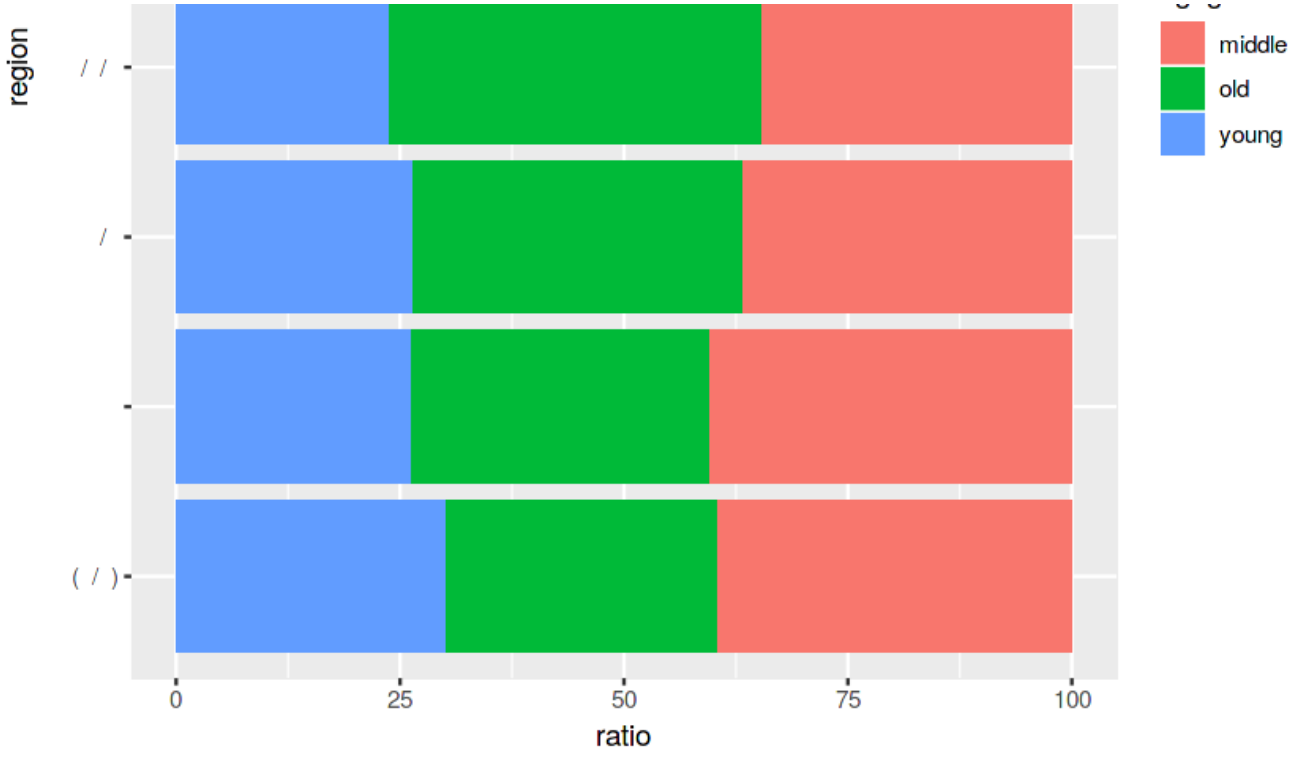
```
sqldf("select x.region, x.ageg, 1.*n/tot*100 as ratio
       from
      (
      select region, ageg, count(*) as n
              from welfare w, region7 r
             where ageg is not null and
            region is not null and
            w.code_region = r.code_region
             group by region, ageg
      ) x,
      (
     select region, count(*) as tot
              from welfare w, region7 r
             where ageg is not null and
            region is not null and
            w.code_region = r.code_region
             group by region
      ) y
     where ageg == 'old' and
    x.region = y.region
    order by ratio
     ")
```

region	ageg	ratio
수도권(인천/경기)	old	30.29461
서울	old	33.23224
대전/충남	old	36.98332
부산/경남/울산	old	41.55796
광주/전남/전북/제주도	old	42.69351
강원/충북	old	45.57709
대구/경북	old	46.28776

In [16]:

```
ratiobyagegbyregion %>% ggplot(aes(region,ratio,fill=ageg))+
                        geom_col()+
                        coord_flip() +
                        scale_x_discrete(limits=sort1$region)
```




region
/ /
/
(/)
0
25
50
75
100
ratio
middle
old
young