# Stat346_HW4_Statistical Data Science

2017100057/ 이영노

2022-11-30

## Ch.19

### Question 1

```
library(dslabs)
```

```
## Warning: 패키지 'dslabs'는 R 버전 4.1.3에서 작성되었습니다
```

```
library(tidyverse)
```

```
## Warning: 패키지 'tidyverse'는 R 버전 4.1.3에서 작성되었습니다
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.7      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.2
```

```
## Warning: 패키지 'ggplot2'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'tibble'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'tidyr'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'readr'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'purrr'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'dplyr'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'stringr'는 R 버전 4.1.3에서 작성되었습니다
```

```
## Warning: 패키지 'forcats'는 R 버전 4.1.3에서 작성되었습니다
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidyr)
data("research_funding_rates")

tab <- research_funding_rates %>%
  select(awards_men, awards_women, applications_men, applications_women) %>%
```

```
  summarize_all(funs(sum)) %>%
  summarize(y_men = awards_men,
            n_men = applications_men - awards_men,
            y_women = awards_women,
            n_women = applications_women - awards_women)%>%
  gather%>%mutate(gender=c('men','men','female','female'),
                  award=c('yes','no','yes','no'))
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
xtabs(value~gender+award,tab)
```

```
##         award
## gender     no  yes
##   female 1011  177
##   men    1345  290
```

### Question 2

```
xtabs(value~gender+award,tab)%>%prop.table(1)
```

```
##         award
## gender          no       yes
##   female 0.8510101 0.1489899
##   men    0.8226300 0.1773700
```

### Question 4

```
index=research_funding_rates%>%arrange(desc(success_rates_total))%>%
  pull(discipline)%>%as.vector()

mat = research_funding_rates%>%
  rename(success_total=success_rates_total,
         success_men=success_rates_men,
         success_women=success_rates_women)%>%
  gather(key,value,-discipline)%>%
```

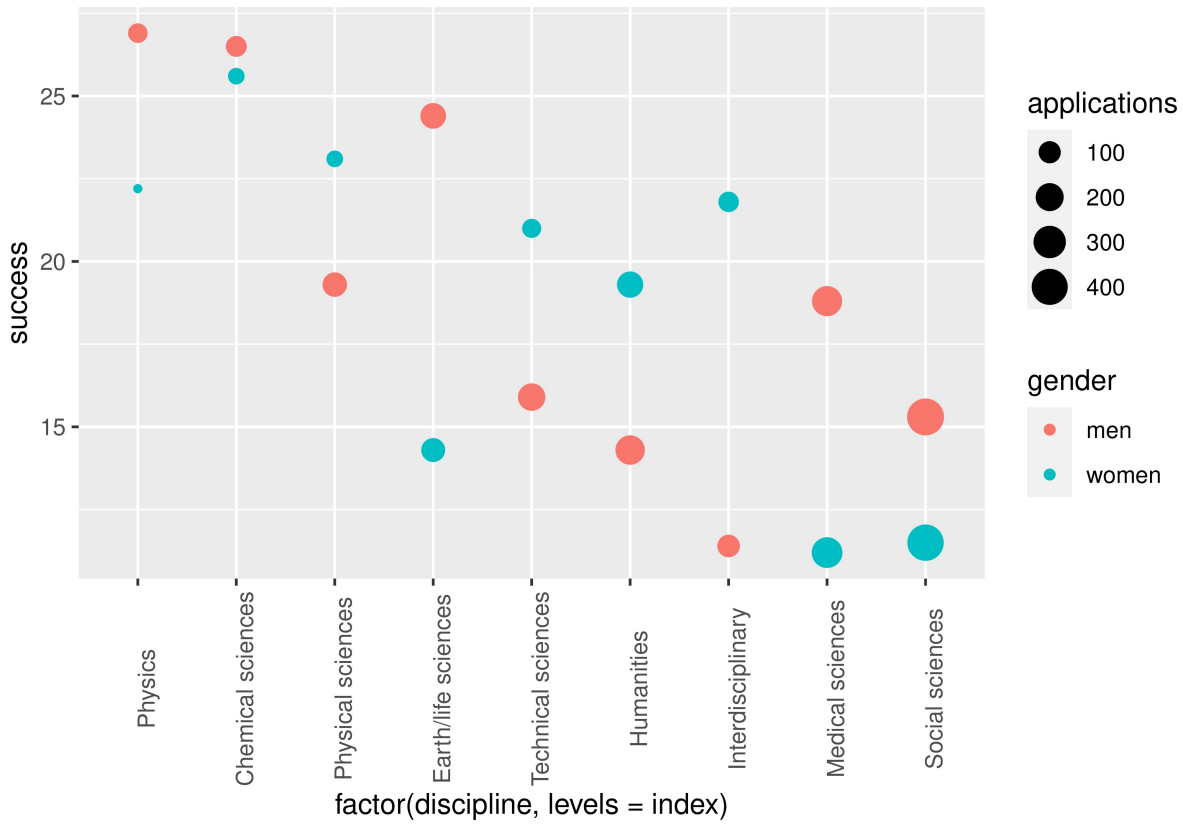
```
  separate(key,c('type','gender'))%>%filter(gender != 'total')%>%
  spread(type,value)%>%
  arrange(factor(discipline,levels=index))

print(mat)
```

```
##             discipline gender applications awards success
## 1              Physics    men           67     18    26.9
## 2              Physics  women            9      2    22.2
## 3    Chemical sciences    men           83     22    26.5
## 4    Chemical sciences  women           39     10    25.6
## 5    Physical sciences    men          135     26    19.3
## 6    Physical sciences  women           39      9    23.1
## 7  Earth/life sciences    men          156     38    24.4
## 8  Earth/life sciences  women          126     18    14.3
## 9   Technical sciences    men          189     30    15.9
## 10  Technical sciences  women           62     13    21.0
## 11          Humanities    men          230     33    14.3
## 12          Humanities  women          166     32    19.3
## 13   Interdisciplinary    men          105     12    11.4
## 14   Interdisciplinary  women           78     17    21.8
## 15    Medical sciences    men          245     46    18.8
## 16    Medical sciences  women          260     29    11.2
## 17     Social sciences    men          425     65    15.3
## 18     Social sciences  women          409     47    11.5
```

### Question 5

```
mat %>%
  ggplot(aes(factor(discipline,levels=index),success,size=applications,col=gender))+
  geom_point()+
  theme(axis.text.x=element_text(angle=90))
```




# Ch.21

## Question 1

```
co2_wide <- data.frame(matrix(co2, ncol = 12, byrow = TRUE)) |>
  setNames(1:12) |>
  mutate(year = as.character(1959:1997))

co2_tidy=co2_wide%>%pivot_longer(1:12,names_to="month",values_to="co2")
co2_tidy%>%head()
```

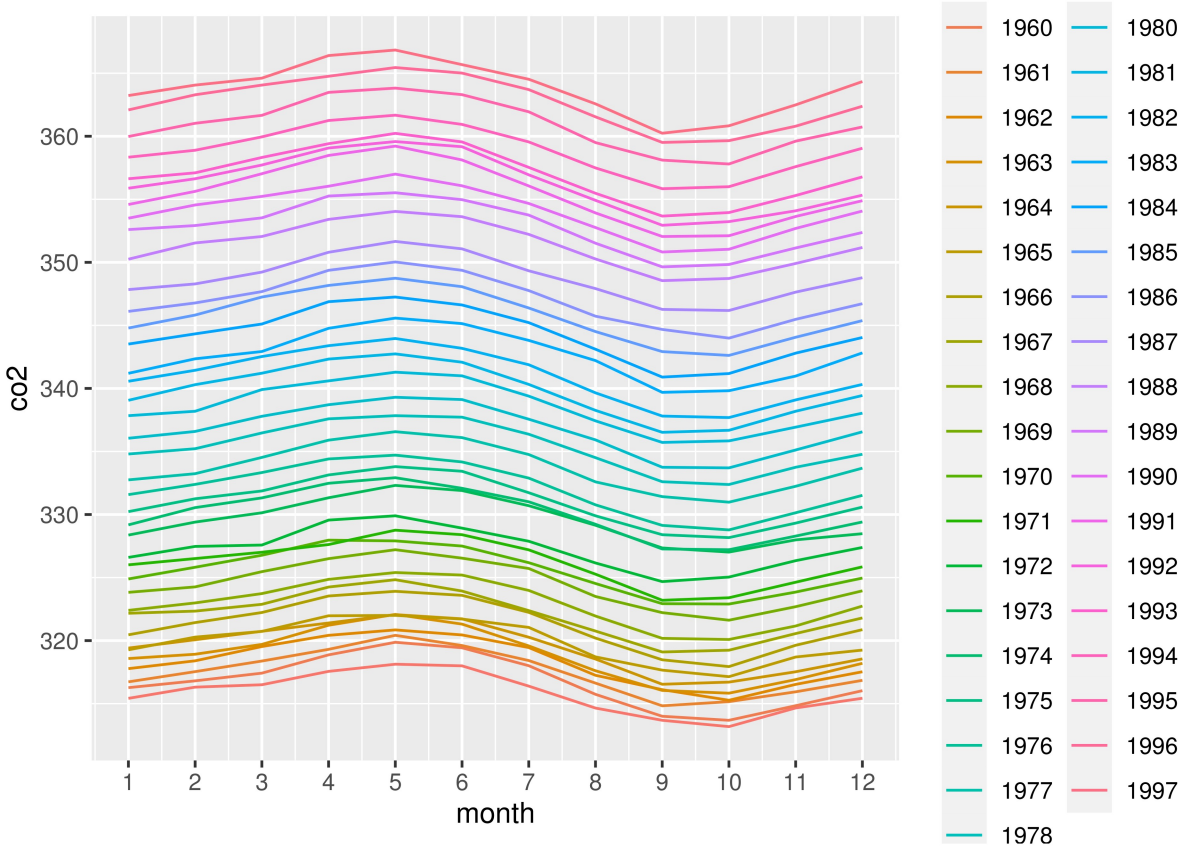
```
## # A tibble: 6 x 3
##   year  month   co2
##   <chr> <chr> <dbl>
## 1 1959  1      315.
## 2 1959  2      316.
## 3 1959  3      316.
## 4 1959  4      318.
## 5 1959  5      318.
## 6 1959  6      318
```

## Question 2

```
library(purrr)
class(co2_tidy$month)
```

```
## [1] "character"
```

```
co2_tidy[,'month']=as.numeric(co2_tidy$month)
co2_tidy %>% ggplot(aes(month, co2, color = year)) + geom_line()+
  scale_x_continuous(breaks=1:12)
```




## Question 3

a,b : True

c: False. co2 measures change over month and year.

d: False. co2 measures increase in summer and decrease in winter.

## Question 4

```
data(admissions)
dat <- admissions |> select(-applicants)
dat%>%pivot_wider(names_from=gender,values_from=admitted)
```

```
## # A tibble: 6 x 3
##   major   men women
##   <chr> <dbl> <dbl>
## 1 A        62    82
## 2 B        63    68
## 3 C        37    34
## 4 D        33    35
## 5 E        28    24
## 6 F         6     7
```

### Question 5

```
tmp= admissions%>%pivot_longer(3:4,names_to='name',values_to='value') # long 하게 만들걸 앞에 co
tmp%>%head()
```

```
## # A tibble: 6 x 4
##   major gender name       value
##   <chr> <chr>  <chr>      <dbl>
## 1 A     men    admitted      62
## 2 A     men    applicants   825
## 3 B     men    admitted      63
## 4 B     men    applicants   560
## 5 C     men    admitted      37
## 6 C     men    applicants   325
```

### Question 6

```
tmp=tmp%>%unite('column_name',name,gender)
tmp%>%head()
```

```
## # A tibble: 6 x 3
##   major column_name    value
##   <chr> <chr>          <dbl>
## 1 A     admitted_men      62
## 2 A     applicants_men   825
## 3 B     admitted_men      63
## 4 B     applicants_men   560
## 5 C     admitted_men      37
## 6 C     applicants_men   325
```

### Question 7

```
tmp%>%pivot_wider(names_from=column_name, values_from='value')
```

```
## # A tibble: 6 x 5
##   major admitted_men applicants_men admitted_women applicants_women
##   <chr>        <dbl>          <dbl>          <dbl>            <dbl>
## 1 A               62            825             82              108
```

```
## 2 B                63            560              68               25
## 3 C                37            325              34              593
## 4 D                33            417              35              375
## 5 E                28            191              24              393
## 6 F                 6            373               7              341
```

### Question 8

```
admissions%>%pivot_longer(3:4,names_to='name',values_to='value')%>%
  unite('column_name',name,gender)%>%
  pivot_wider(names_from=column_name, values_from='value')
```

```
## # A tibble: 6 x 5
##   major admitted_men applicants_men admitted_women applicants_women
##   <chr>        <dbl>          <dbl>          <dbl>            <dbl>
## 1 A               62            825             82              108
## 2 B               63            560             68               25
## 3 C               37            325             34              593
## 4 D               33            417             35              375
## 5 E               28            191             24              393
## 6 F                6            373              7              341
```

## Ch.22

### Question 1

```
library(Lahman)
```

```
## Warning: 패키지 'Lahman'는 R 버전 4.1.3에서 작성되었습니다
```

```
top <- Batting |>
  filter(yearID == 2016) |>
  arrange(desc(HR)) |>
  slice(1:10)%>% as_tibble()
people=People |> as_tibble()

HR=top%>%left_join(people,by="playerID")%>%select(playerID,nameFirst,nameLast,HR)
print(HR)
```

```
## # A tibble: 10 x 4
##    playerID  nameFirst nameLast       HR
##    <chr>     <chr>     <chr>       <int>
##  1 trumbma01 Mark      Trumbo         47
##  2 cruzne02  Nelson    Cruz           43
##  3 daviskh01 Khris     Davis          42
##  4 doziebr01 Brian     Dozier         42
##  5 encared01 Edwin     Encarnacion    42
##  6 arenano01 Nolan     Arenado        41
```

```
##  7 cartech02 Chris     Carter          41
##  8 frazito01 Todd      Frazier         40
##  9 bryankr01 Kris      Bryant          39
## 10 canoro01  Robinson  Cano            39
```

### Question 2

```
Salaries%>%filter(yearID==2016)%>%right_join(HR,by="playerID")%>%
  select(nameFirst,nameLast,teamID,HR,salary)
```

```
##    nameFirst    nameLast teamID HR   salary
## 1       Mark      Trumbo    BAL 47  9150000
## 2       Kris      Bryant    CHN 39   652000
## 3       Todd     Frazier    CHA 40  8250000
## 4      Nolan     Arenado    COL 41  5000000
## 5      Chris      Carter    MIL 41  2500000
## 6      Brian      Dozier    MIN 42  3000000
## 7      Khris       Davis    OAK 42   524500
## 8   Robinson        Cano    SEA 39 24000000
## 9     Nelson        Cruz    SEA 43 14250000
## 10     Edwin Encarnacion    TOR 42 10000000
```

### Question 3

```
co2_wide <- data.frame(matrix(co2, ncol = 12, byrow = TRUE)) |>
  setNames(1:12) |>
  mutate(year = as.character(1959:1997)) |>
  pivot_longer(-year, names_to = "month", values_to = "co2") |>
  mutate(month = as.numeric(month))
yearly_avg=co2_wide%>%group_by(year)%>%summarize(average=mean(co2))
yearly_avg%>%head()
```

```
## # A tibble: 6 x 2
##   year  average
##   <chr>   <dbl>
## 1 1959     316.
## 2 1960     317.
## 3 1961     317.
## 4 1962     318.
## 5 1963     319.
## 6 1964     319.
```

### Question 4

```
res=co2_wide%>%left_join(yearly_avg,by='year')%>%mutate(residual=co2-average)
res%>%head(5)
```

```
## # A tibble: 5 x 5
```

```
##   year  month   co2 average residual
##   <chr> <dbl> <dbl>   <dbl>    <dbl>
## 1 1959      1  315.    316.   -0.406
## 2 1959      2  316.    316.    0.484
## 3 1959      3  316.    316.    0.674
## 4 1959      4  318.    316.    1.73
## 5 1959      5  318.    316.    2.30
```

```
residual_vector=res%>%pull(residual)
residual_vector%>%head(5)
```

```
## [1] -0.4058333  0.4841667  0.6741667  1.7341667  2.3041667
```

### Question 5

```
res %>% ggplot(aes(month, residual, color = year)) + geom_line()+
  scale_x_continuous(breaks=1:12)
```