Homework assignment #4

2021320322 / Minseo Yoon

November 29, 2022

1

```r
library(tidyverse)
library(dplyr)
library(ggplot2)
```

19.6.1

```r
library(dslabs)
data("research_funding_rates")
research_funding_rates
```

```
##            discipline applications_total applications_men applications_women
## 1   Chemical sciences                122               83                 39
## 2   Physical sciences                174              135                 39
## 3             Physics                 76               67                  9
## 4          Humanities                396              230                166
## 5  Technical sciences                251              189                 62
## 6   Interdisciplinary                183              105                 78
## 7 Earth/life sciences                282              156                126
## 8     Social sciences                834              425                409
## 9    Medical sciences                505              245                260
##   awards_total awards_men awards_women success_rates_total success_rates_men
## 1           32         22           10                26.2              26.5
## 2           35         26            9                20.1              19.3
## 3           20         18            2                26.3              26.9
## 4           65         33           32                16.4              14.3
## 5           43         30           13                17.1              15.9
## 6           29         12           17                15.8              11.4
## 7           56         38           18                19.9              24.4
## 8          112         65           47                13.4              15.3
## 9           75         46           29                14.9              18.8
##   success_rates_women
## 1                25.6
## 2                23.1
## 3                22.2
## 4                19.3
## 5                21.0
## 6                21.8
## 7                14.3
## 8                11.5
## 9                11.2
```

```
tab <- research_funding_rates %>%
  summarize(awards_men = sum(awards_men),
            awards_women = sum(awards_women),
            nonawards_men = sum(applications_men) - awards_men,
            nonawards_women = sum(applications_women) - awards_women) %>%
  pivot_longer(c('awards_men', 'awards_women', 'nonawards_men', 'nonawards_women'),
               names_to = 'awarded', values_to = 'num') %>%
  separate(awarded, c('awarded', 'gender')) %>%
  pivot_wider(names_from = gender, values_from = num)
tab
```

```
## # A tibble: 2 x 3
##   awarded     men women
##   <chr>     <dbl> <dbl>
## 1 awards      290   177
## 2 nonawards  1345  1011
```

19.6.2

```
tab %>%
  mutate(prop_men = men / sum(men) * 100,
         prop_women = women / sum(women) * 100) %>%
  filter(awarded == 'awards') %>%
  select(prop_men, prop_women)
```

```
## # A tibble: 1 x 2
##   prop_men prop_women
##      <dbl>      <dbl>
## 1     17.7       14.9
```

Answer: The percentage of men who is awarded among applications is about 2.8 percent higher than that of women who is awarded.

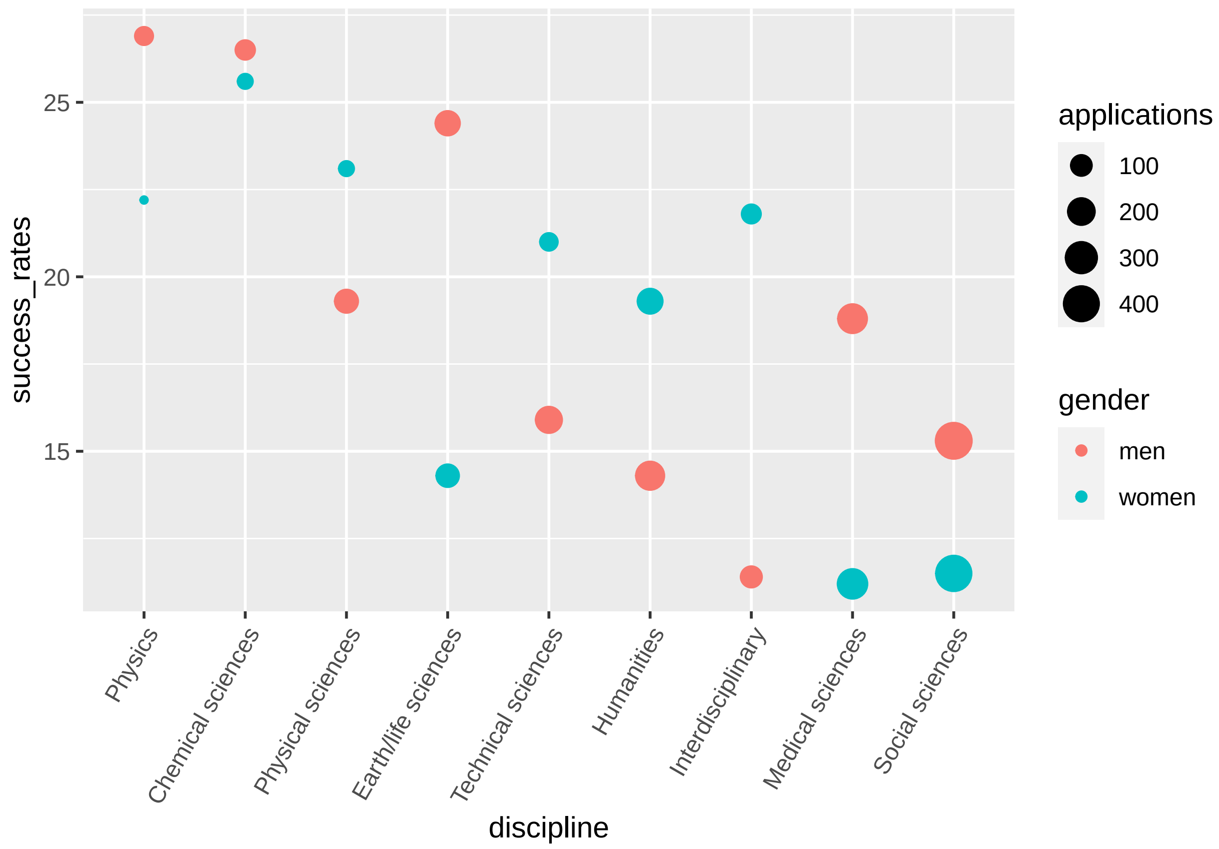
19.6.4

```
tab.2 <- research_funding_rates %>%
  mutate(discipline = reorder(discipline, -success_rates_total)) %>%
  select(-applications_total, -awards_total, -success_rates_total) %>%
  rename(rates_men = success_rates_men,
         rates_women = success_rates_women) %>%
  pivot_longer(c('applications_men', 'applications_women',
                 'awards_men', 'awards_women',
                 'rates_men', 'rates_women'),
               names_to = 'name', values_to = 'num') %>%
  separate(name, c('variable_name', 'gender')) %>%
  pivot_wider(names_from = variable_name, values_from = num) %>%
  rename(success_rates = rates)

tab.2
```

```
## # A tibble: 18 x 5
##    discipline          gender applications awards success_rates
##    <fct>               <chr>         <dbl>  <dbl>         <dbl>
##  1 Chemical sciences   men              83     22          26.5
##  2 Chemical sciences   women            39     10          25.6
##  3 Physical sciences   men             135     26          19.3
##  4 Physical sciences   women            39      9          23.1
##  5 Physics             men              67     18          26.9
##  6 Physics             women             9      2          22.2
##  7 Humanities          men             230     33          14.3
##  8 Humanities          women           166     32          19.3
##  9 Technical sciences  men             189     30          15.9
## 10 Technical sciences  women            62     13          21
## 11 Interdisciplinary   men             105     12          11.4
## 12 Interdisciplinary   women            78     17          21.8
## 13 Earth/life sciences men             156     38          24.4
## 14 Earth/life sciences women           126     18          14.3
## 15 Social sciences     men             425     65          15.3
## 16 Social sciences     women           409     47          11.5
## 17 Medical sciences    men             245     46          18.8
## 18 Medical sciences    women           260     29          11.2
```

19.6.5

```
tab.2 %>%
  ggplot(aes(x = discipline, y = success_rates)) +
  geom_point(aes(size = applications, color = gender)) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



2

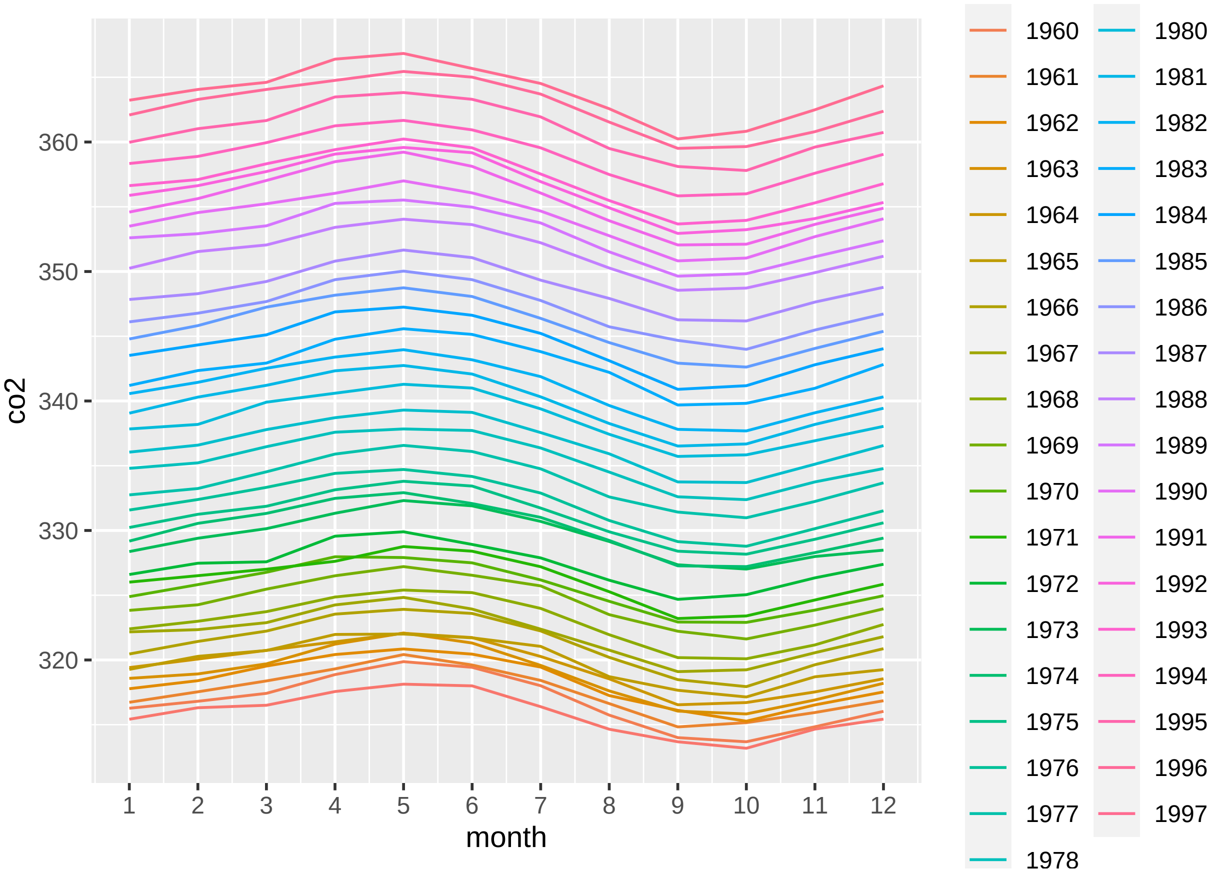
21.5.1

```
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
```

```
## # A tibble: 468 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
##  7 1959  7       316.
##  8 1959  8       315.
##  9 1959  9       314.
## 10 1959  10      313.
## # ... with 458 more rows
```

21.5.2

```
co2_tidy <- co2_wide %>%
  pivot_longer(`1`:`12`, names_to = 'month', values_to = 'co2',
               names_transform = list(month = as.integer))

co2_tidy %>%
  ggplot(aes(month, co2, color = year)) +
  geom_line() +
  scale_x_continuous(breaks = 1:12)
```



21.5.3

Answer: b. CO2 measures are higher in the summer and the yearly average increased from 1959 to 1997.

21.5.4

```
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
```

21.5.5

```
tmp <- admissions %>%
  pivot_longer(c(admitted, applicants), names_to = 'name', values_to = 'value')
tmp
```

```
## # A tibble: 24 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
##  7 D     men    admitted      33
##  8 D     men    applicants   417
##  9 E     men    admitted      28
## 10 E     men    applicants   191
## # ... with 14 more rows
```

21.5.6

```
tmp <- tmp %>%
  unite(column_name, name, gender)
tmp
```

```
## # A tibble: 24 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
##  7 D     admitted_men      33
##  8 D     applicants_men   417
##  9 E     admitted_men      28
## 10 E     applicants_men   191
## # ... with 14 more rows
```

21.5.7

```
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
```

21.5.8

```
admissions %>%
  pivot_longer(c(admitted, applicants), 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
```

3

22.4.1

```
library(Lahman)

top <- Batting %>%
  filter(yearID == 2016) %>%
  arrange(desc(HR)) %>%
  slice(1:10)

top %>% as_tibble()
```

```
## # A tibble: 10 x 22
##    playerID  yearID stint teamID lgID      G    AB     R     H   X2B   X3B    HR
##    <chr>      <int> <int> <fct>  <fct> <int> <int> <int> <int> <int> <int> <int>
##  1 trumbma01   2016     1 BAL    AL      159   613    94   157    27     1    47
##  2 cruzne02    2016     1 SEA    AL      155   589    96   169    27     1    43
##  3 daviskh01   2016     1 OAK    AL      150   555    85   137    24     2    42
##  4 doziebr01   2016     1 MIN    AL      155   615   104   165    35     5    42
##  5 encared01   2016     1 TOR    AL      160   601    99   158    34     0    42
##  6 arenano01   2016     1 COL    NL      160   618   116   182    35     6    41
##  7 cartech02   2016     1 MIL    NL      160   549    84   122    27     1    41
##  8 frazito01   2016     1 CHA    AL      158   590    89   133    21     0    40
##  9 bryankr01   2016     1 CHN    NL      155   603   121   176    35     3    39
## 10 canoro01    2016     1 SEA    AL      161   655   107   195    33     2    39
## # ... with 10 more variables: RBI <int>, SB <int>, CS <int>, BB <int>,
## #   SO <int>, IBB <int>, HBP <int>, SH <int>, SF <int>, GIDP <int>
```

```
People %>% as_tibble() %>% head(10)
```

```
## # A tibble: 10 x 26
##    playerID  birthYear birthMonth birthDay birthCountry birthState birthCity
##    <chr>         <int>      <int>    <int> <chr>        <chr>      <chr>
##  1 aardsda01      1981          12        27 USA          CO          Denver
##  2 aaronha01      1934           2         5 USA          AL          Mobile
##  3 aaronto01      1939           8         5 USA          AL          Mobile
##  4 aasedo01       1954           9         8 USA          CA          Orange
##  5 abadan01       1972           8        25 USA          FL          Palm Beach
##  6 abadfe01       1985          12        17 D.R.         La Romana  La Romana
##  7 abadijo01      1850          11         4 USA          PA          Philadelphia
##  8 abbated01      1877           4        15 USA          PA          Latrobe
##  9 abbeybe01      1869          11        11 USA          VT          Essex
## 10 abbeych01      1866          10        14 USA          NE          Falls City
## # ... with 19 more variables: deathYear <int>, deathMonth <int>,
## #   deathDay <int>, deathCountry <chr>, deathState <chr>, deathCity <chr>,
## #   nameFirst <chr>, nameLast <chr>, nameGiven <chr>, weight <int>,
## #   height <int>, bats <fct>, throws <fct>, debut <chr>, finalGame <chr>,
## #   retroID <chr>, bbrefID <chr>, deathDate <date>, birthDate <date>
```

```
top <- left_join(top, People, by = 'playerID') %>%
  select(playerID, nameFirst, nameLast, HR)
top
```

```
##     playerID nameFirst    nameLast HR
## 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
```

22.4.2

```
Salaries <- Salaries %>%
  filter(yearID == 2016)

right_join(Salaries, top, 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
```

22.4.3

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

yearly_avg <- co2_wide %>%
  group_by(year) %>%
  summarize(avg_co2 = mean(co2))

yearly_avg
```

```
## # A tibble: 39 x 2
##     year avg_co2
##    <int>   <dbl>
##  1  1959    316.
##  2  1960    317.
##  3  1961    317.
##  4  1962    318.
##  5  1963    319.
##  6  1964    319.
##  7  1965    320.
##  8  1966    321.
##  9  1967    322.
## 10  1968    323.
## # ... with 29 more rows
```

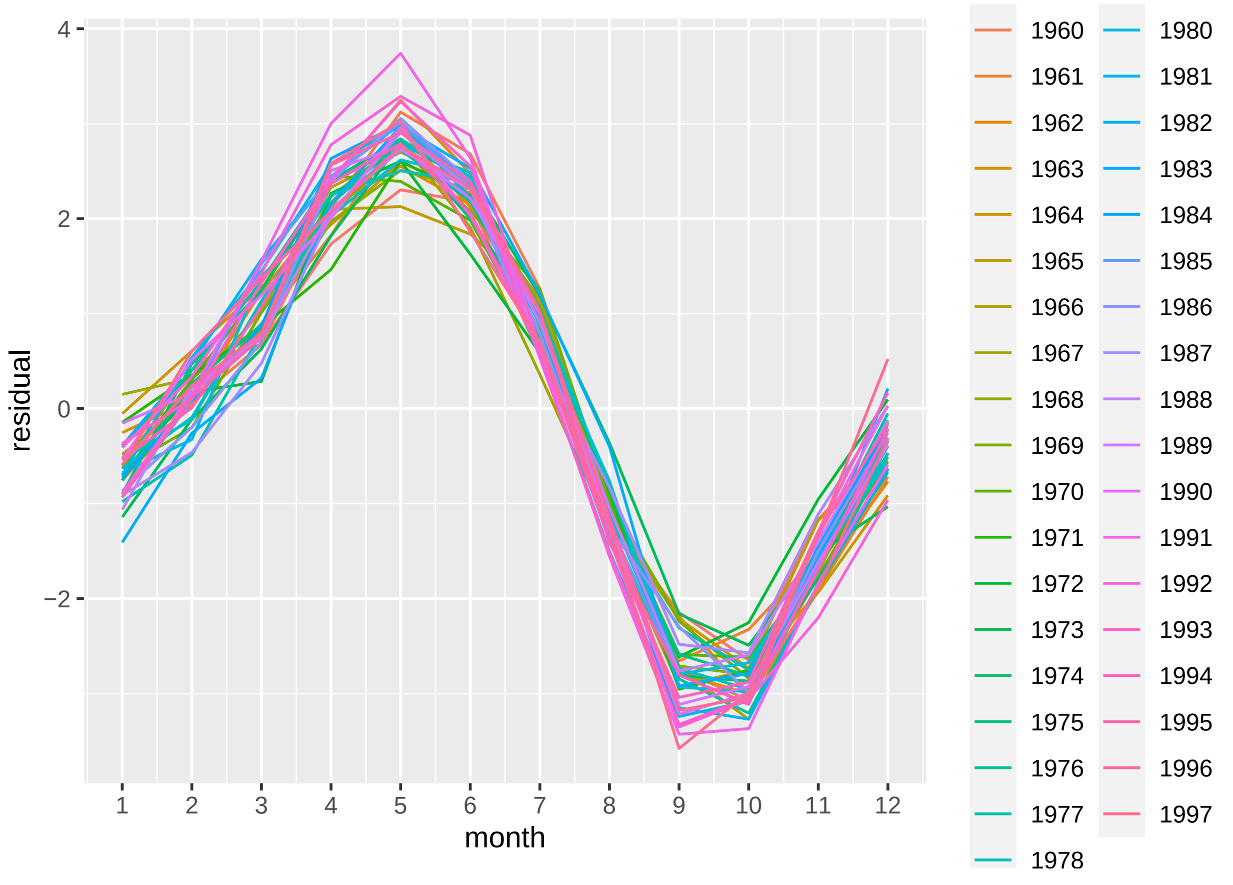
22.4.4

```
co2_wide <- left_join(co2_wide, yearly_avg, by = 'year') %>%
  mutate(residual = co2 - avg_co2)

co2_wide
```

```
## # A tibble: 468 x 5
##     year month   co2 avg_co2 residual
##    <int> <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
##  6  1959     6  318     316.    2.17
##  7  1959     7  316.    316.    0.564
##  8  1959     8  315.    316.   -1.18
##  9  1959     9  314.    316.   -2.15
## 10  1959    10  313.    316.   -2.65
## # ... with 458 more rows
```

22.4.5

```
co2_wide %>%
  ggplot(aes(x = month, y = residual, color = factor(year))) +
  geom_line() +
  scale_x_continuous(breaks = 1:12)
```


residual
month