Week 3 Homework

Cam Nguyen

2022-04-20

1 Read in excel file

```
dat = read_excel('USMA_Progression.xlsx')
```

2 Recreating femalespeers, malespeers, and totpeople columns

```
dat <- dat %>%
  group_by(company_n,class,year) %>%
  mutate(femalespeers = ifelse(female == 0, sum(female), sum(female)-1), female = ifelse
(female == 0, 1, 0), malespeers = ifelse(female == 1, sum(female) - 1, sum(female)), fem
ale = ifelse(female == 1, 0, 1), totpeople = femalespeers + malespeers)
dat
```

```
## # A tibble: 17,223 × 8
## # Groups:   company_n, class, year [595]
##     year class female femalespeers malespeers continue_or_grad company_n
##    <dbl> <dbl>  <dbl>        <dbl>      <dbl>            <dbl> <chr>
##  1    78     1      0            0         33                1 A-1
##  2    78     1      0            0         33                1 A-1
##  3    78     1      0            0         33                1 A-1
##  4    78     1      0            0         33                1 A-1
##  5    78     1      0            0         33                1 A-1
##  6    78     1      0            0         33                1 A-1
##  7    78     1      0            0         33                1 A-1
##  8    78     1      0            0         33                1 A-1
##  9    78     1      0            0         33                1 A-1
## 10    78     1      0            0         33                1 A-1
## # … with 17,213 more rows, and 1 more variable: totpeople <dbl>
```

3 Investigation

My recreation does not match the original because the original accounts for everyone and not just peers. The number of peers does not change depending on whether the person is male or female. Another issue is that people not included in the same company, year, and class are not included as a peer even if they are in the same company and year. The description for the totpeople variable description says “peers total in company/class. Equal to femalespeers + malespeers”. The description is incorrect because femalespeers and malespeers are grouped by company/cohort/year. For the year 1978, the description says that there are only freshmen but there are also sophomores. The original has issues with the peers and totpeople columns, so I trust my calculations more.

4 Create new columns

```
dat <- dat %>%
  mutate(company = str_sub(company_n,1,1), division = str_sub(company_n,3,3))
dat
```

```
## # A tibble: 17,223 × 10
## # Groups:   company_n, class, year [595]
##     year class female femalespeers malespeers continue_or_grad company_n
##    <dbl> <dbl>  <dbl>        <dbl>      <dbl>            <dbl> <chr>
##  1    78     1      0            0         33                1 A-1
##  2    78     1      0            0         33                1 A-1
##  3    78     1      0            0         33                1 A-1
##  4    78     1      0            0         33                1 A-1
##  5    78     1      0            0         33                1 A-1
##  6    78     1      0            0         33                1 A-1
##  7    78     1      0            0         33                1 A-1
##  8    78     1      0            0         33                1 A-1
##  9    78     1      0            0         33                1 A-1
## 10    78     1      0            0         33                1 A-1
## # … with 17,213 more rows, and 3 more variables: totpeople <dbl>,
## #   company <chr>, division <chr>
```

5 Years with all 4 classes

```
filter(dat, year == 81)
```

```
## # A tibble: 4,290 × 10
## # Groups:   company_n, class, year [144]
##     year class female femalespeers malespeers continue_or_grad company_n
##    <dbl> <dbl>  <dbl>        <dbl>      <dbl>            <dbl> <chr>
##  1    81     1      0            3         30                0 A-1
##  2    81     1      0            3         30                0 A-1
##  3    81     1      0            3         30                0 A-1
##  4    81     1      0            3         30                0 A-1
##  5    81     1      0            3         30                0 A-1
##  6    81     1      0            3         30                1 A-1
##  7    81     1      0            3         30                1 A-1
##  8    81     1      1            2         31                1 A-1
##  9    81     1      0            3         30                1 A-1
## 10    81     1      0            3         30                1 A-1
## # … with 4,280 more rows, and 3 more variables: totpeople <dbl>, company <chr>,
## #   division <chr>
```

6

A) Top four companies with highest continue or grad rates

```
dat %>%
  select(continue_or_grad, company) %>%
  group_by(company) %>%
  summarize(rate_cont = mean(continue_or_grad)) %>%
  ungroup() %>%
  arrange(desc(rate_cont)) %>%
  slice(1:4)
```

```
## Adding missing grouping variables: `company_n`, `class`, `year`
```

```
## # A tibble: 4 × 2
##   company rate_cont
##   <chr>       <dbl>
## 1 D           0.926
## 2 A           0.918
## 3 H           0.915
## 4 B           0.914
```

B) Continue or Grad Rates by Class

```
dat %>%
  select(continue_or_grad, class) %>%
  group_by(class) %>%
  summarize(rate_cont = mean(continue_or_grad))
```

```
## Adding missing grouping variables: `company_n`, `year`
```

```
## # A tibble: 4 × 2
##   class rate_cont
##   <dbl>     <dbl>
## 1     1     0.848
## 2     2     0.889
## 3     3     0.957
## 4     4     0.971
```

C) Continue or Grad Rates of Women by Class

```
dat %>%
  filter(female == 1) %>%
  select(continue_or_grad, class) %>%
  group_by(class) %>%
  summarize(rate_cont = mean(continue_or_grad))
```

```
## Adding missing grouping variables: `company_n`, `year`
## # A tibble: 4 × 2
##   class rate_cont
##   <dbl>     <dbl>
## 1     1     0.811
## 2     2     0.792
## 3     3     0.954
## 4     4     0.985
```