Homework 6

Sookja Kang, sk26949

This homework is due on March 29, 2021 at 11:00pm. Please submit as a pdf file on Canvas.

For all problems in this homework, we will work with the `BA_degrees` dataset. It contains the proportions of Bachelor's degrees awarded in the US between 1970 and 2015.

```
BA_degrees <- read_csv("https://wilkelab.org/SDS375/datasets/BA_degrees.csv")
BA_degrees
```

```
## # A tibble: 594 x 4
##    field                                           year  count     perc
##    <chr>                                          <dbl>  <dbl>    <dbl>
##  1 Agriculture and natural resources               1971  12672 0.0151
##  2 Architecture and related services               1971   5570 0.00663
##  3 Area, ethnic, cultural, gender, and group studies 1971  2579 0.00307
##  4 Biological and biomedical sciences              1971  35705 0.0425
##  5 Business                                        1971 115396 0.137
##  6 Communication, journalism, and related programs 1971  10324 0.0123
##  7 Communications technologies                     1971    478 0.000569
##  8 Computer and information sciences               1971   2388 0.00284
##  9 Education                                       1971 176307 0.210
## 10 Engineering                                     1971  45034 0.0536
## # ... with 584 more rows
```

Problem 1: (3 pts)

Consider the Bachelor's degrees awarded in 2015. There are 32 different areas:

```
BA_degrees_2015 <- BA_degrees %>%
  filter(year == 2015) %>%
  arrange(desc(perc))

print(BA_degrees_2015, n = nrow(BA_degrees_2015))
```

```
## # A tibble: 33 x 4
##    field                                               year  count    perc
##    <chr>                                              <dbl>  <dbl>   <dbl>
##  1 Business                                            2015 363799 1.92e-1
##  2 Health professions and related programs             2015 216228 1.14e-1
##  3 Social sciences and history                         2015 166944 8.81e-2
##  4 Psychology                                          2015 117557 6.20e-2
##  5 Biological and biomedical sciences                  2015 109896 5.80e-2
##  6 Engineering                                         2015  97858 5.16e-2
##  7 Visual and performing arts                          2015  95832 5.06e-2
##  8 Education                                           2015  91623 4.84e-2
##  9 Communication, journalism, and related programs     2015  90650 4.78e-2
## 10 Homeland security, law enforcement, and firefighting 2015 62723 3.31e-2
## 11 Computer and information sciences                   2015  59581 3.14e-2
## 12 Parks, recreation, leisure, and fitness studies     2015  49006 2.59e-2
```

```
## 13 Multi/interdisciplinary studies                          2015  47556  2.51e-2
## 14 English language and literature/letters                  2015  45847  2.42e-2
## 15 Liberal arts and sciences, general studies, and humani~   2015  43647  2.30e-2
## 16 Agriculture and natural resources                        2015  36277  1.91e-2
## 17 Public administration and social services                2015  34363  1.81e-2
## 18 Physical sciences and science technologies               2015  30038  1.59e-2
## 19 Family and consumer sciences/human sciences              2015  24584  1.30e-2
## 20 Mathematics and statistics                               2015  21853  1.15e-2
## 21 Foreign languages, literatures, and linguistics          2015  19493  1.03e-2
## 22 Engineering technologies                                 2015  17238  9.10e-3
## 23 Philosophy and religious studies                         2015  11072  5.84e-3
## 24 Theology and religious vocations                         2015   9708  5.12e-3
## 25 Architecture and related services                        2015   9090  4.80e-3
## 26 Area, ethnic, cultural, gender, and group studies        2015   7782  4.11e-3
## 27 Communications technologies                              2015   5135  2.71e-3
## 28 Transportation and materials moving                      2015   4711  2.49e-3
## 29 Legal professions and studies                            2015   4420  2.33e-3
## 30 Military technologies and applied sciences               2015    276  1.46e-4
## 31 Library science                                          2015     99  5.22e-5
## 32 Precision production                                     2015     48  2.53e-5
## 33 Not classified by field of study                         2015      0  0.
```

If you wanted to visualize the relative proportion of these different degree areas, which plot would be most appropriate? A pie chart, a stacked bar chart, or side-by-side bars? Please explain your reasoning. You do not have to make the chart.

Answer: *There are 32 different fields in BA_degrees_2015, which will result in many broken pieces in a pie chart or a stacked bar chart. Using a pie chart or a stacked bar chart with many pieces (different degree areas) can be unclear to visualize the relative proportion. Side-by-side bars can easily visualize relative proportion across the different degree areas that can show a trend of the dataset.*

Problem 2: (5 pts) Now make a pie chart of the `BA_degrees_2015` dataset, but show only the four most common areas, plus all others lumped together into "Other". (The code to prepare this lumped dataset has been provided for your convenience.) Make sure the pie slices are arranged in a reasonable order. Choose a reasonable color scale and a clean theme that avoids distracting visual elements.

Grading rubric: 3 pts for making the right plot, 2 pts for visual design

```r
# data preparation
top_four <- BA_degrees_2015$field[1:4]  # works because we sorted by perc in Problem 1
BA_degrees_lumped <- BA_degrees_2015 %>%
  mutate(field = ifelse(field %in% top_four, field, "Other")) %>%
  group_by(field) %>%
  summarize(perc = sum(perc)) %>%
  arrange(-desc(perc)) %>%
  mutate(field = fct_reorder(field, perc))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
pie_data <- BA_degrees_lumped %>%
  mutate(end_angle = 2*pi*cumsum(perc)/sum(perc),
         start_angle = lag(end_angle, default = 0),
         mid_angle = 0.5*(start_angle + end_angle),
         hjust = ifelse(mid_angle > pi, 1, 0),
         vjust = ifelse(mid_angle < pi/2 | mid_angle > 3*pi/2, 0, 1)
  )
```

```
pie_data
```

```
## # A tibble: 5 x 7
##   field                       perc end_angle start_angle mid_angle hjust vjust
##   <fct>                      <dbl>     <dbl>       <dbl>     <dbl> <dbl> <dbl>
## 1 Psychology                0.0620     0.390       0         0.195     0     0
## 2 Social sciences and history 0.0881   0.943       0.390     0.667     0     0
## 3 Health professions and rel~ 0.114    1.66        0.943     1.30      0     0
## 4 Business                  0.192      2.87        1.66      2.26      0     1
## 5 Other                     0.544      6.28        2.87      4.57      1     1
```

```r
# your code goes here
ggplot(pie_data) +
  aes(
    x0 = 0, y0 = 0, r0 = 0, r = 1,
    start = start_angle, end = end_angle,
    fill = field
  ) +
  geom_arc_bar() +
  geom_text(
    aes(
      x = 0.7 * sin(mid_angle),
      y = 0.7 * cos(mid_angle),
      label = paste0(signif(perc*100, 2), "%")
    )
  )+
  coord_fixed() +
  ggtitle("Four most common BA degrees awarded in 2015") +
  theme_void() +
  scale_fill_brewer(palette = "YlGn")
```

Four most common BA degrees awarded in 2015
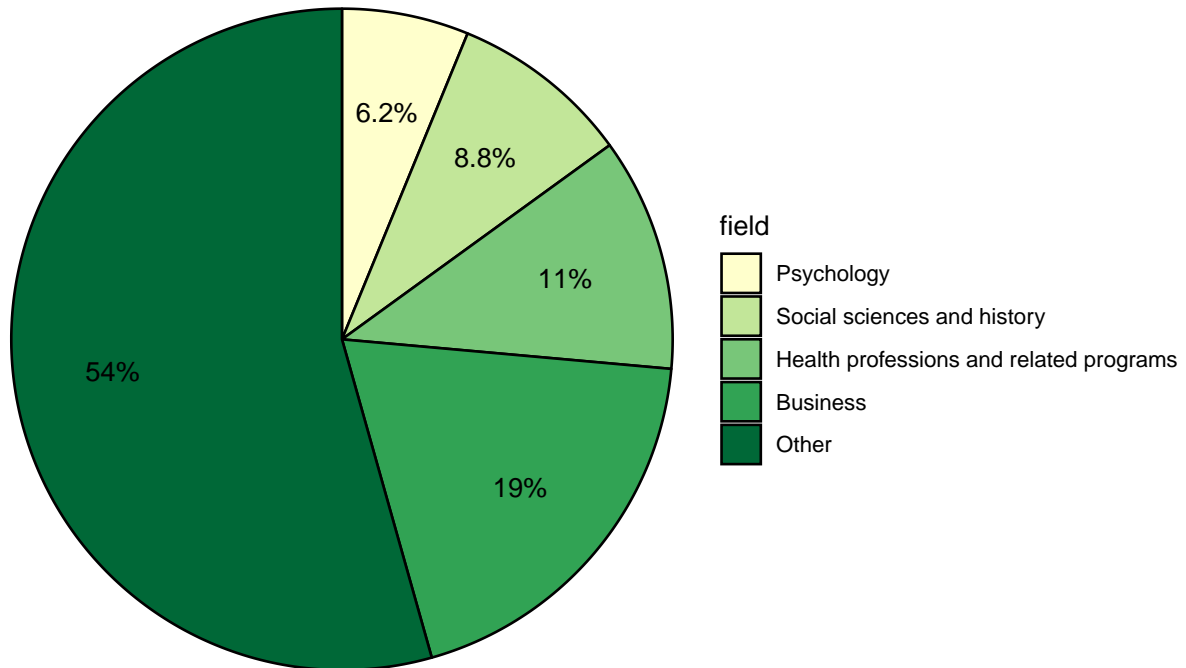
Problem 3: (2 pts) Now go back to the entire dataset `BA_degrees` but focus only on the degree field
`"Parks, recreation, leisure, and fitness studies"`. Plot a time series of the proportion of degrees
(colum `perc`) in this field over time. Also plot a straight line fit to the data. What do you observe?

```
# your code goes here
BA_degrees %>%
  filter(field == "Parks, recreation, leisure, and fitness studies") %>%
  ggplot(aes(year, perc)) +
  theme_bw(14) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```
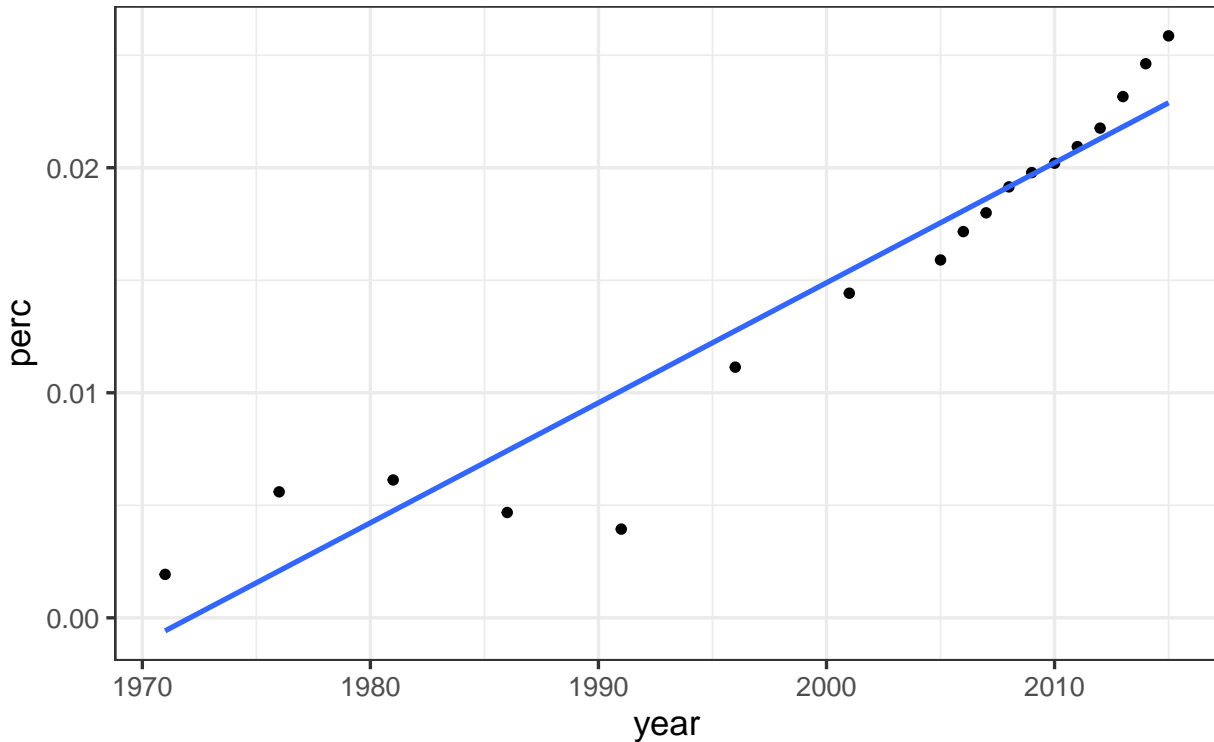
```
#BA_degrees_subset <- filter(BA_degrees, field == "Parks, recreation, leisure, and fitness studies")
#lm_out <- lm(perc ~ year, data = BA_degrees_subset)
#summary(lm_out)
```

This plot shows the proportion changes of Parks, recreation, leisure, and fitness studies Bachelor's degree awarded between 1970 and 2015. In the chart, the proportion has been gradually increasing throughout the time period except the time between 1985 and 1995. After adding the straight line that fits data, there is an approximately linear relationship between years and the proportion changes of Parks, recreation, leisure, and fitness studies.