

Homework 6

Enter your name and EID here Jeremy Ulfohn // jau392

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: Since for instance precision production's count is way, way less than 1% that of business, a pie chart would be useless and so would a stacked bar since you wouldn't even be able to see that these less popular degrees even exist (except in the key, if one is even shown). Therefore a side-by-side bar chart with the relative proportions on the y axis, which more emphatically displays/labels the existence of all 32 areas, would be best. That aside, 32 areas is far too many for a reasonable pie chart anyway.

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

```
library(forcats)
library(scales)

##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##   discard

## The following object is masked from 'package:readr':
##
##   col_factor

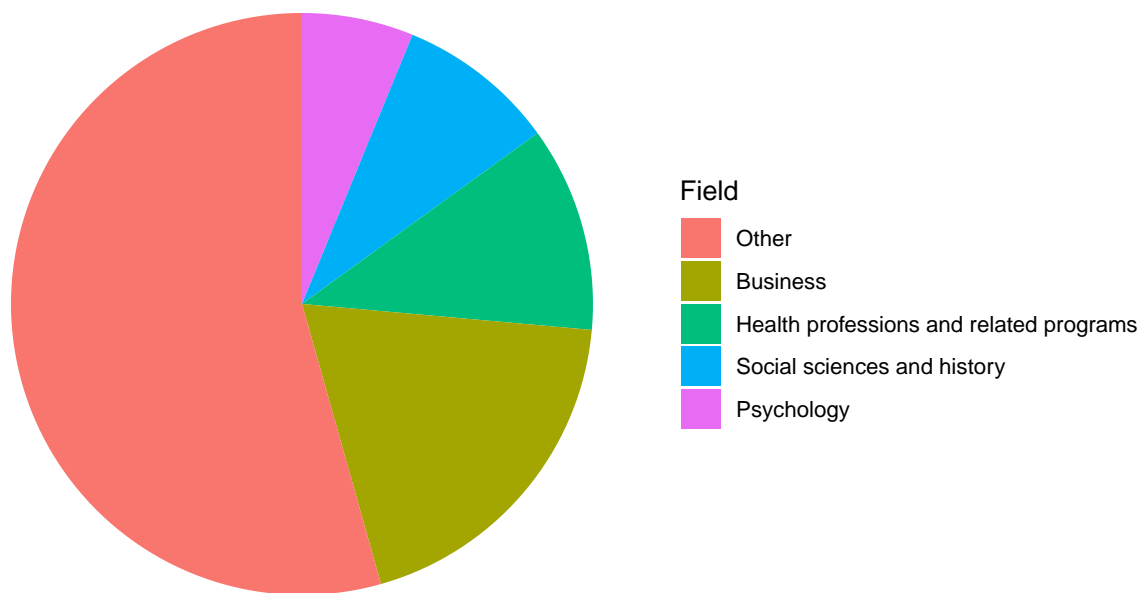
# data preparation
top_four <- BA_degrees_2015$field[1:4] # works because we sorted by perc in Problem 1
BA_degrees_lumped <- BA_degrees_2015 %>%
  mutate(field = ifelse(field %in% top_four, field, "Other")) %>%
  group_by(field) %>%
```

```
arrange(count) %>%
summarize(perc = sum(perc))
```

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

```
# your code goes here
# Bar chart in polar coordinates
ggplot(BA_degrees_lumped[order(BA_degrees_lumped$perc), ]) +
  aes(perc, "YY", fill = reorder(field, -perc)) +
  geom_col() +
  coord_polar() +
  scale_x_continuous(name = NULL, breaks = NULL) +
  scale_y_discrete(name = NULL, breaks = NULL) +
  scale_fill_discrete(name = "Field") +
  theme_void() +
  ggtitle("Most Popular 2015 BA Degrees")
```

Most Popular 2015 BA Degrees



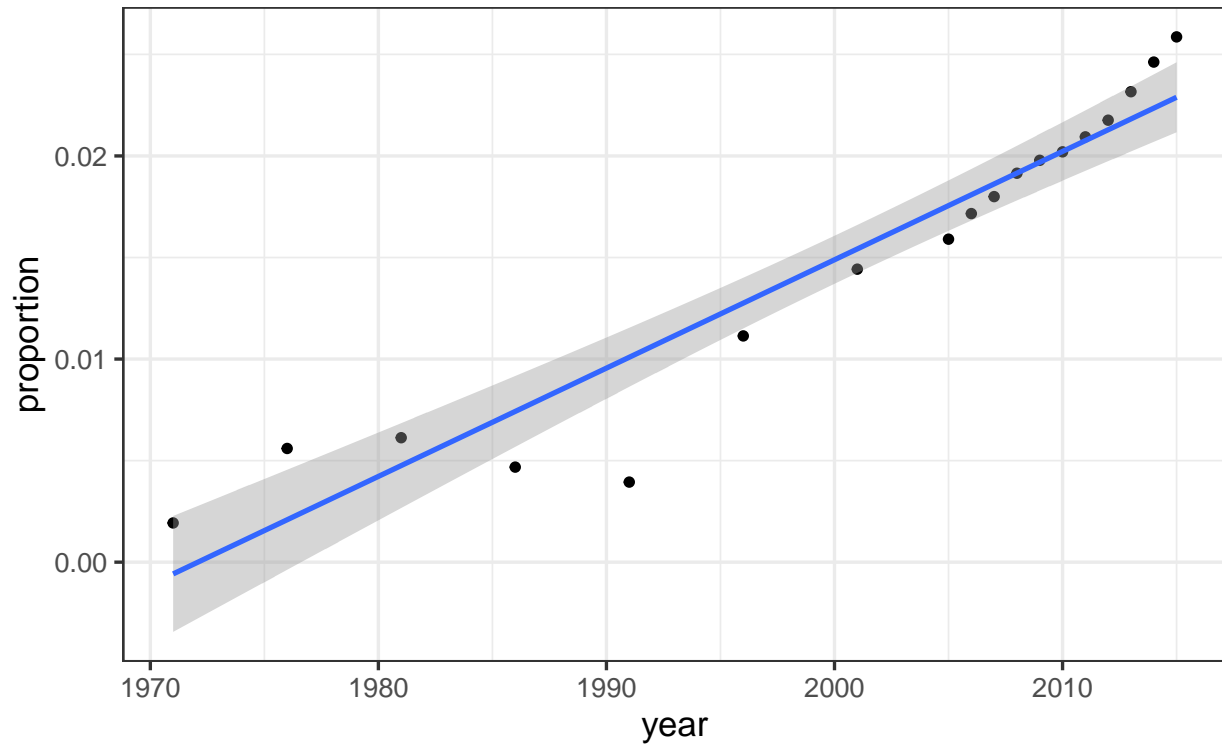
Problem 3: (2 pts) Now go back to the entire dataset `BA_degrees` but focus only on the degree field "Parks, recreation, leisure, and fitness studies". Plot a time series of the proportion of degrees (column `perc`) in this field over time. Also plot a straight line fit to the data. What do you observe?

```
# Filter the dataset down to only this field
parks_etc <- BA_degrees %>%
  filter(field == "Parks, recreation, leisure, and fitness studies")

# your code goes here
ggplot(parks_etc) + # Concentrate time series on this one field
  aes(year, perc) +
  geom_point() +
```

```
scale_y_continuous(name = "proportion") +
theme_bw(14) +
geom_smooth(
  method = "lm"
)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



I observe that though there was a dip in popularity between approx. 1982 and 1992, this degree field has overall been on the rise both before and after this period. The rate at which it has been rising has fluctuated even in modern years, but it has been positive nonetheless. This trend is further cooberated by the fact that there has been a data point ever year starting in approx. 2005.