Homework 3

This homework is due on the deadline posted on edX. Please submit a .pdf file of your output and upload a .zip file containing your .Rmd file. Do NOT include your name or EID in your filenames.

Problem 1: For this problem, we will work with the `BA_degrees` dataset. It contains the proportions of Bachelor's degrees awarded in the US between 1970 and 2015.

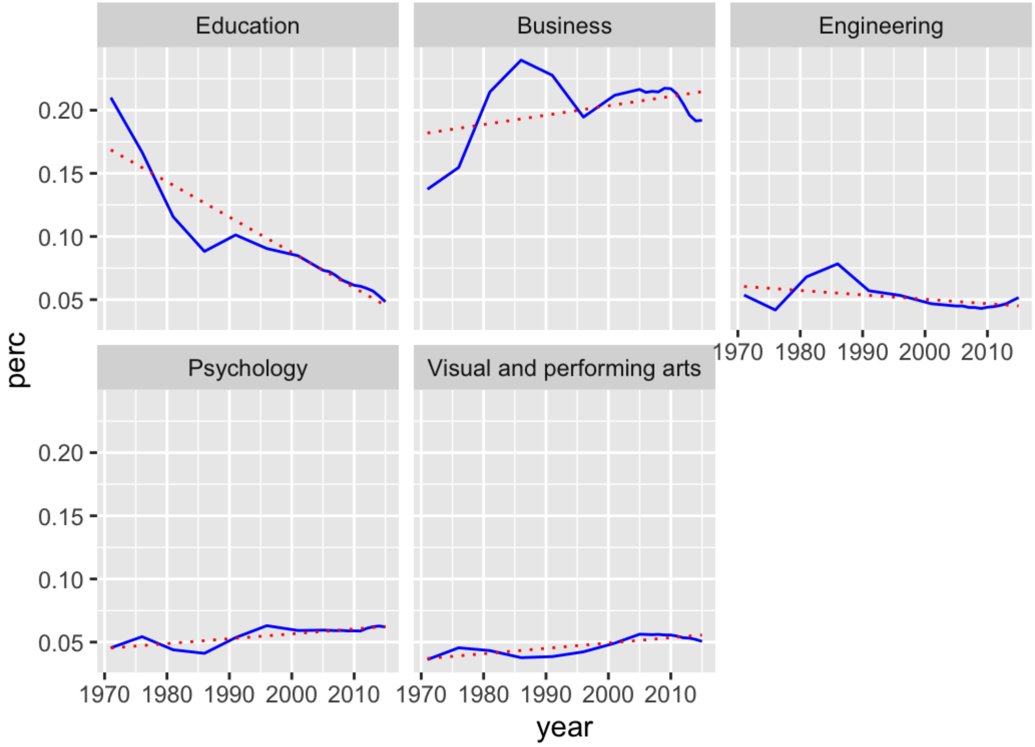
```
BA_degrees <- read_csv("https://wilkelab.org/SDS375/datasets/BA_degrees.csv")
BA_degrees
```

```
## # A tibble: 594 × 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
```

From the entire dataset, select a subset of 6 fields of study, using arbitrary criteria. Plot a time series of the proportion of degrees (column `perc`) in this field over time, using facets to show each field. Also plot a straight line fit to the data for each field. You should modify the order of facets to maximize figure appearance and memorability. What do you observe?

Hint: To get started, see slides 34 to 44 in the class on getting things into the right order: https://wilkelab.org/DSC385/slides/getting-things-in-order.html#34 (https://wilkelab.org/DSC385/slides/getting-things-in-order.html#34)

```
BA_degrees %>%
  filter(field %in% c("Business", "Engineering", "Education", "Psychology", "Visual and
 performing arts")) %>%
  mutate(field = fct_reorder(field, perc, function(x) { min(x) - max(x) })) %>%
  ggplot(aes(year, perc)) +
  geom_line(color='blue') +
  geom_smooth(method = "lm", se = FALSE, linetype='dotted', size=.5, color='red') +
  facet_wrap(~field)
```



Fields that have rapid growth: Business

Fields that are rapidly declining: Education

Fields that have stayed pretty much consistent: Engineering, Psychology, Visual and Performing Arts.

Problem 2: We will work the `txhousing` dataset provided by **ggplot2**. See here for details: https://ggplot2.tidyverse.org/reference/txhousing.html (https://ggplot2.tidyverse.org/reference/txhousing.html)

Consider the number of houses sold in January 2015. There are records for 46 different cities:

```
txhousing_jan_2015 <- txhousing %>%
  filter(year == 2015 & month == 1) %>%
  arrange(desc(sales))
```

If you wanted to visualize the relative proportion of sales in these different cities, 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: *Side by side bars since there are a large number of subsets in the dataset.*

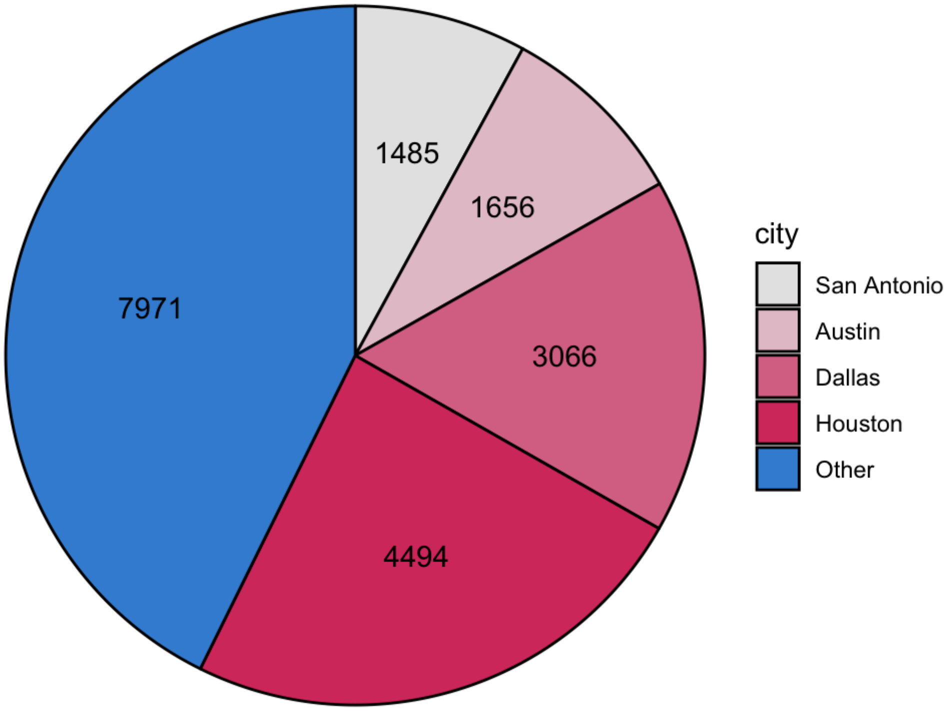
Problem 3: Now make a pie chart of the `txhousing_jan_2015` dataset, but show only the four cities with the most sales, plus all others lumped together into “Other”. (The code to prepare this lumped dataset has been provided for your convenience.) Make sure the pie slices are arranged in a reasonable order. Choose a reasonable color scale and a clean theme that avoids distracting visual elements.

```
# data preparation
top_four <- txhousing_jan_2015$sales[1:4]

txhousing_lumped <- txhousing_jan_2015 %>%
  mutate(city = ifelse(sales %in% top_four, city, "Other")) %>%
  group_by(city) %>%
  summarize(sales = sum(sales))

pie_data <- txhousing_lumped %>%
  arrange(sales) %>% # sort so pie slices end up sorted
  mutate(
    end_angle = 2*pi*cumsum(sales)/sum(sales),   # ending angle for each pie slice
    start_angle = lag(end_angle, default = 0),   # starting angle for each pie slice
    mid_angle = 0.5*(start_angle + end_angle),   # middle of each pie slice, for text la
bels
    # horizontal and vertical justifications for outer labels
    hjust = ifelse(mid_angle > pi, 1, 0),
    vjust = ifelse(mid_angle < pi/2 | mid_angle > 3*pi/2, 0, 1)
  ) %>%
  mutate(city = fct_reorder(city, sales, min))

ggplot(pie_data) +
  aes(
    x0 = 0, y0 = 0, r0 = 0, r = 1,
    start = start_angle, end = end_angle,
    fill = city
  ) +
  geom_arc_bar() +
  geom_text( # place amounts inside the pie
    aes(
      x = 0.6 * sin(mid_angle),
      y = 0.6 * cos(mid_angle),
      label = sales
    )
  ) +
  coord_fixed() +
  theme_void() +
  scale_fill_manual(values= c(
    '#E5E5E5',
    '#E4C5CE',
    '#DA7593',
    '#D73E6C',
    '#3E8FD7'
  )
  )
```


1485
1656
3066
4494
7971
city
San Antonio
Austin
Dallas
Houston
Other