Homework 3

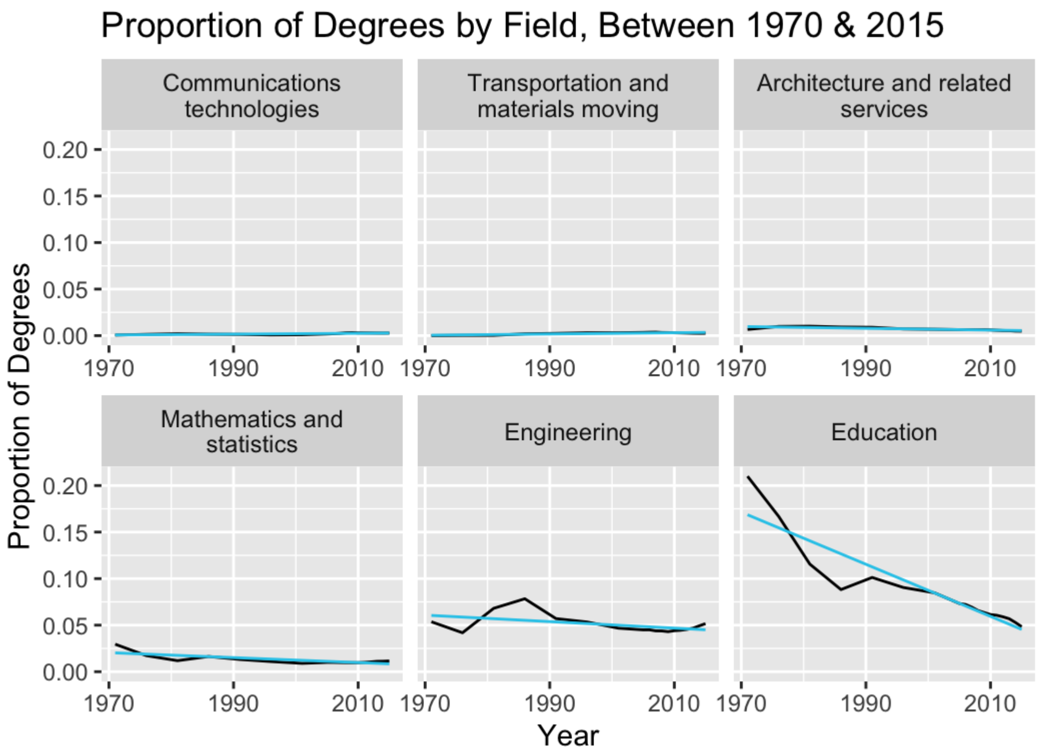
This homework is due on the deadline posted on edX. Please submit a .pdf file of your output and upload a .zip file containing your .Rmd file. Do NOT include your name or EID in your filenames.

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.

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)

```r
plot <- BA_degrees |>
  #Pull in 6 fields
  filter(field %in% c("Mathematics and statistics", "Architecture and related services", "Engineering", "Communications technologies", "Transportation and materials moving", "Education")) |>

  #Order by percentage median
  mutate(field = fct_reorder(field, perc)) |>

  #Plot line
  ggplot(aes(year, perc)) + geom_line() +

  #Format x axis
  scale_x_continuous(
    name = "Year",
    breaks = c(1970, 1990, 2010)) +

  #Format y axis
  scale_y_continuous(
    name = "Proportion of Degrees") +

  #Add facets
  facet_wrap(vars(field), ncol = 3, scales="free_x", labeller = label_wrap_gen()) +

  #Add theme
  theme(
    strip.text = element_text(size = 9)
  ) +

  #Add Line of Best fit
  geom_smooth(method = "lm", se = FALSE, color = "#2acaea", alpha=.3, linewidth=.5) +

  #Add Title
  labs(title = "Proportion of Degrees by Field, Between 1970 & 2015")

suppressMessages(print(plot))
```



When observing the facets part of the plot above, it is interesting to note that all fields, except “Education”, appear relatively stable since 1970, noted by the flat blue line of best fit in each facet. However, it is apparent that the proportion of degrees for “Education” have decreased fairly significantly since 1970, which may suggest that the degree itself is not as lucrative as it once once. I was also surprised by the low proportion of degrees belonging to “Mathematics and statistics”, but this may suggest there are substitute degrees that are more attainable (i.e. Economics / Finance)

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))

print(txhousing_jan_2015, n = 10)
```

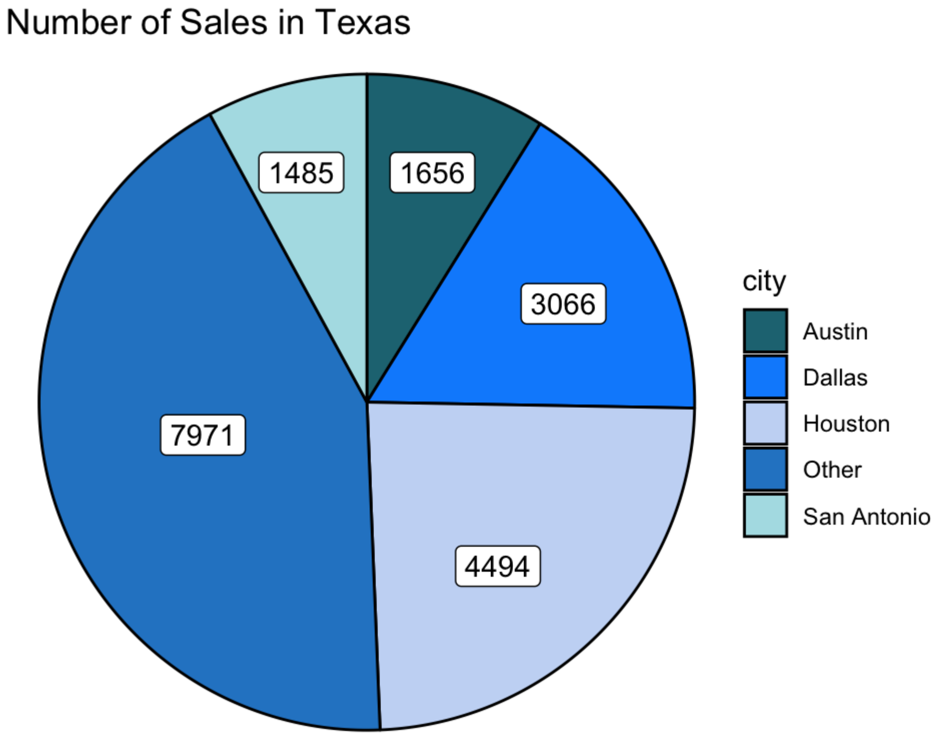
```
## # A tibble: 46 × 9
##    city               year month sales    volume median listings inventory  date
##    <chr>             <int> <int> <dbl>     <dbl>  <dbl>    <dbl>     <dbl> <dbl>
##  1 Houston            2015     1  4494    1.16e9 189300    18649       2.7  2015
##  2 Dallas             2015     1  3066    7.74e8 203300     9063       1.8  2015
##  3 Austin             2015     1  1656    5.12e8 237500     5567       2.2  2015
##  4 San Antonio        2015     1  1485    3.12e8 175900     7717       3.6  2015
##  5 Collin County      2015     1   776    2.42e8 268000     1780       1.3  2015
##  6 Fort Bend          2015     1   686    2.04e8 260300     2414       2.3  2015
##  7 Fort Worth         2015     1   658    1.12e8 143300     2089       2.1  2015
##  8 Montgomery County  2015     1   487    1.47e8 213200     2507       3.3  2015
##  9 NE Tarrant County  2015     1   482    1.27e8 204000     1093       1.4  2015
## 10 Denton County      2015     1   477    1.22e8 216100     1151       1.4  2015
## # ℹ 36 more rows
```

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: *The best chart for this particular use case would be a side-by-side bar chart. The reason being that this particular dataset we are using contains 46 different cities from home sales that occurred in 2015. A side-by-side bar chart would be best to compare the relative proportions of sales for each city, while using either a pie chart or stacked bar chart would not be ideal as the number of cities within this dataset would overcrowd those charts and it would be hard to gain any insight into the comparison of sales for each city.*

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.

```r
# 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))

plot <- ggplot(txhousing_lumped) +
  aes(
    x0 = 0, y0 = 0, # position of pie center
    r0 = 0, r = 1,  # inner and outer radius
    amount = sales,
    fill = city
  ) +
  geom_arc_bar(stat = "pie") +
  geom_label(x = c(0.2, 0.6, 0.4, -0.5, -0.2), y = c(0.7, 0.3, -0.5, -0.1, 0.7), aes(label = sales), fill="white") +
  theme_void() +
  coord_fixed() +
  scale_fill_manual(values = c(Houston = "#c8d9f5", Austin = "#207482", Dallas = "#088efc", `San Antonio` = "#b0e0e6", Other = "#2986CC")) +
  labs(title = "Number of Sales in Texas")
print(plot)
```


Number of Sales in Texas
city
Austin
Dallas
Houston
Other
San Antonio