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 with your .Rmd file.

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 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
```

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)

It can be seen below that the `perc` column is a percentage of the degree out of all degrees *for that year*. The sum of all for one year is exactly **1**.

```
BA_degrees %>% filter(year == 1971) %>% pull(perc) %>% sum()
```

```
## [1] 1
```

Because the `perc` column is from **0** to **1**, I mutate the data to scale to true percentage; in other words, out of **100**. The data is also rounded to one decimal place for ease of display.

```
BA_degrees = BA_degrees %>% mutate(percentage = round(perc * 100, 1))
BA_degrees
```

```
## # A tibble: 594 x 5
##    field                                       year  count    perc percentage
##    <chr>                                      <dbl>  <dbl>   <dbl>      <dbl>
##  1 Agriculture and natural resources           1971  12672 1.51e-2        1.5
##  2 Architecture and related services           1971   5570 6.63e-3        0.7
##  3 Area, ethnic, cultural, gender, and group st~  1971   2579 3.07e-3        0.3
##  4 Biological and biomedical sciences          1971  35705 4.25e-2        4.3
##  5 Business                                    1971 115396 1.37e-1       13.7
##  6 Communication, journalism, and related progr~  1971  10324 1.23e-2        1.2
##  7 Communications technologies                 1971    478 5.69e-4        0.1
##  8 Computer and information sciences           1971   2388 2.84e-3        0.3
##  9 Education                                   1971 176307 2.10e-1       21
## 10 Engineering                                 1971  45034 5.36e-2        5.4
## # ... with 584 more rows
```

I have selected the six most numerous degrees in 2001 (arbitrarily), and plotted their percentage over time. By filtering by name in descending count order, the facets in the following graph will be ordered by count already.

```
BA_degrees %>% filter(year == 2001) %>% arrange(-count)
```

```
## # A tibble: 33 x 5
##    field                                       year  count   perc percentage
##    <chr>                                      <dbl>  <dbl>  <dbl>      <dbl>
##  1 Business                                    2001 263515 0.212        21.2
##  2 Social sciences and history                 2001 128036 0.103        10.3
##  3 Education                                   2001 105458 0.0848        8.5
##  4 Health professions and related programs     2001  75933 0.0610        6.1
##  5 Psychology                                  2001  73645 0.0592        5.9
##  6 Visual and performing arts                  2001  61148 0.0491        4.9
##  7 Biological and biomedical sciences          2001  60576 0.0487        4.9
##  8 Engineering                                 2001  58209 0.0468        4.7
##  9 Communication, journalism, and related progra~  2001  58013 0.0466        4.7
## 10 English language and literature/letters     2001  50569 0.0406        4.1
## # ... with 23 more rows
```

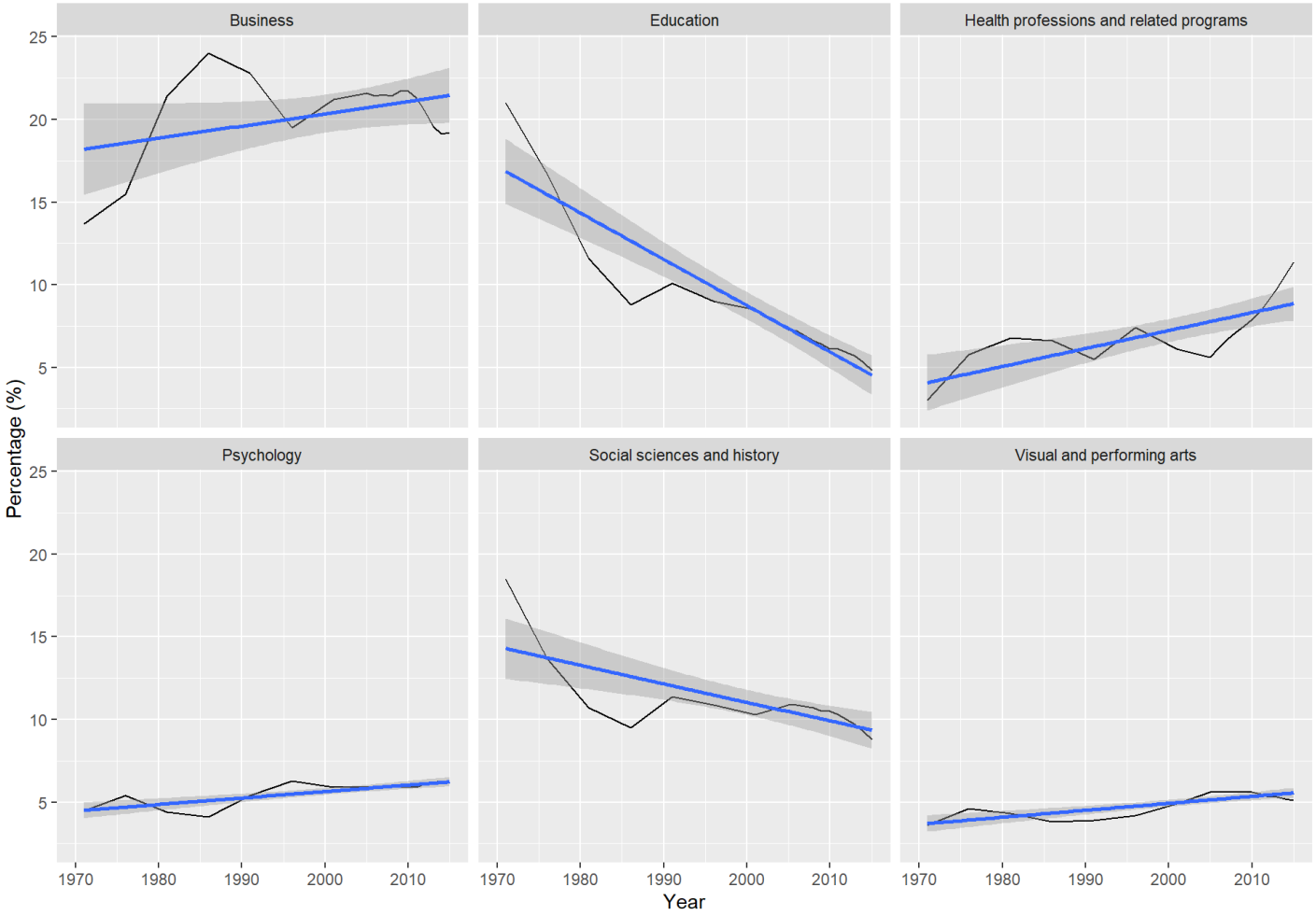
```
BA_degrees = BA_degrees %>% filter(
  field == 'Business' |
  field == 'Social sciences and history' |
  field == 'Education' |
  field == 'Health professions and related programs' |
  field == 'Psychology' |
  field == 'Visual and performing arts')
BA_degrees
```

```
## # A tibble: 108 x 5
##    field                                year  count   perc percentage
##    <chr>                               <dbl>  <dbl>  <dbl>      <dbl>
##  1 Business                             1971 115396 0.137        13.7
##  2 Education                            1971 176307 0.210        21
##  3 Health professions and related programs  1971  25223 0.0300        3
##  4 Psychology                           1971  38187 0.0455        4.5
##  5 Social sciences and history          1971 155324 0.185        18.5
##  6 Visual and performing arts           1971  30394 0.0362        3.6
##  7 Business                             1976 143171 0.155        15.5
##  8 Education                            1976 154437 0.167        16.7
##  9 Health professions and related programs  1976  53885 0.0582        5.8
## 10 Psychology                           1976  50278 0.0543        5.4
## # ... with 98 more rows
```

I now plot the data over time faceted to degree.

```
BA_degrees %>%
  ggplot(aes(year, percentage)) +
  geom_line() +
  stat_smooth(method='lm') +
  facet_wrap(vars(field)) +
  scale_x_continuous(
    name = "Year",
    limits = c(1970, 2015),
    breaks = c(1970, 1980, 1990, 2000, 2010)) +
  scale_y_continuous(
    name = "Percentage (%)"
  )
```

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



It can be see from the graphs that business and health degrees are becoming more common, social sciences have become less common, psychology and arts degrees have relatively stagnated, and perhaps **most** important is that education degrees are drastically fallen. One could presume that the degrees that are becoming more common are ones that have better earning potential, degrees that have fallen have lost earning potential, and degrees that have stagnated likely didn't not change in earning potential.

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 = nrow(txhousing_jan_2015))
```

```
## # A tibble: 46 x 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
## 11 El Paso            2015     1   406    6.27e7 135200     3995       7.7  2015
## 12 Bay Area           2015     1   401    8.12e7 172200     1910       2.9  2015
## 13 Arlington          2015     1   261    4.61e7 159700      552       1.3  2015
## 14 Tyler              2015     1   248    3.98e7 139400     2290       6.9  2015
## 15 Corpus Christi     2015     1   241    4.53e7 162800     1872       4.8  2015
## 16 Amarillo           2015     1   204    3.32e7 138500     1120       4.3  2015
## 17 Lubbock            2015     1   202    3.24e7 132400      979       3.1  2015
## 18 Killeen-Fort Hood  2015     1   188    2.44e7 114100     1372       6.1  2015
## 19 Bryan-College St~  2015     1   173    3.82e7 189300      988       3.8  2015
## 20 Abilene            2015     1   158    2.35e7 134100      801       4.4  2015
## 21 Beaumont           2015     1   151    2.17e7 122000     1558       7.3  2015
## 22 McAllen            2015     1   146    1.94e7 118300     2068      11.6  2015
## 23 Waco               2015     1   144    2.26e7 137500     1034       5    2015
## 24 Longview-Marshall  2015     1   134    1.82e7 131400     1766       9.1  2015
## 25 Garland            2015     1   114    1.76e7 135800      198       1.1  2015
## 26 Temple-Belton      2015     1   107    1.77e7 124500      727       4.9  2015
## 27 Midland            2015     1    91    2.41e7 235900      840       5    2015
## 28 Sherman-Denison    2015     1    88    1.22e7 121700      549       4.4  2015
## 29 Irving             2015     1    82    2.02e7 157800      278       1.9  2015
## 30 Laredo             2015     1    82    1.26e7 136200      512       5.2  2015
## 31 San Angelo         2015     1    82    1.38e7 138300      477       3.7  2015
## 32 Texarkana          2015     1    75    9.33e6 101400      317       3.7  2015
## 33 Harlingen          2015     1    74    7.53e6  85000     1560      18.7  2015
## 34 Wichita Falls      2015     1    71    7.52e6  82100      829       7.2  2015
## 35 Brazoria County    2015     1    69    1.04e7 146000      301       2.8  2015
## 36 Odessa             2015     1    63    1.00e7 156200      308       3    2015
## 37 Victoria           2015     1    54    1.04e7 172500      280       3.6  2015
## 38 Kerrville          2015     1    53    1.35e7 212500      643      11.4  2015
## 39 Galveston          2015     1    43    1.08e7 187500      575       5.8  2015
## 40 Brownsville        2015     1    41    5.40e6  97000      733      10.7  2015
## 41 Lufkin             2015     1    37    6.87e6 134000      404       7.6  2015
## 42 Port Arthur        2015     1    37    3.96e6  93800      558       7.8  2015
## 43 Paris              2015     1    25    3.61e6 123300      299       8.1  2015
## 44 South Padre Isla~  2015     1    22    4.89e6 180000      688      18.5  2015
## 45 Nacogdoches        2015     1    20    3.22e6 140000      284      10.5  2015
## 46 San Marcos         2015     1    18    3.38e6 150000       85       3.4  2015
```

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 would be best. Both pie charts and stacked bar charts are not good for direct comparison as it is hard to see the relative size against the other options. Pie charts are better for general overview without needing to compare, and stacked bars are not that useful in general; the beginning of each bar color is different per bar with the exception of the first and last, so it's difficult to make any valuable comparison.

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))

txhousing_lumped %>% ggplot() +
  aes(sales, "YY", fill = fct_reorder(city, -sales)) +
  geom_col() +
  coord_polar() +
  scale_x_continuous(
    name = NULL,
    breaks = NULL
  ) +
  scale_y_discrete(
    name = NULL,
    breaks = NULL
  ) +
  scale_fill_viridis_d(
    name = "City",
    option = "C"
  ) +
  theme(panel.background = element_blank()) +
  ggtitle("Number of House Sales in Texas by City for January 2015")
```

Number of House Sales in Texas by City for January 2015