Time-series with a straightline fit using ggforce, and Piechart

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("http://wilkelab.org/SDS375/datasets/BA_degrees.csv")
BA_degrees
```

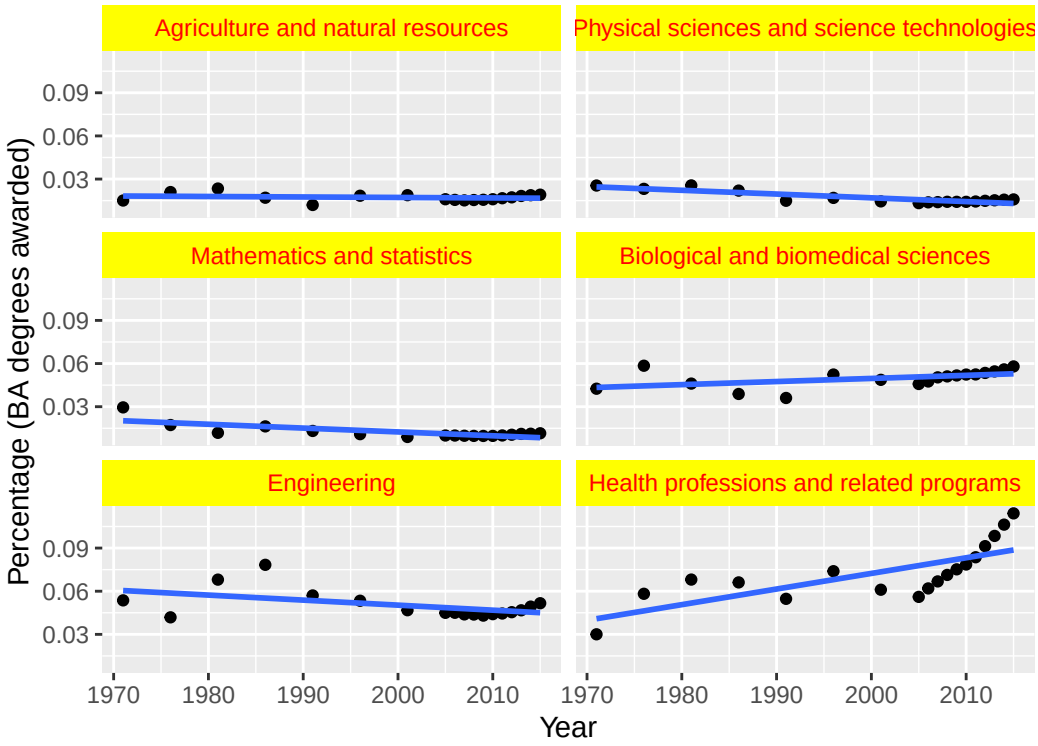
```
## # A tibble: 594 x 4
##    field                                              year  count     perc
##    <chr>                                             <dbl>  <dbl>    <dbl>
##  1 Agriculture and natural resources                  1971  12672 0.0151
##  2 Architecture and related services                  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
## # i 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

```
# time series of the proportion of degrees over time
BA_degrees %>%
  filter(field %in% c("Agriculture and natural resources", "Biological and biomedical sciences", "Engin
  mutate(field = fct_reorder(field, perc, function(x) { max(x) - min(x) })) %>%
  ggplot(aes(year, perc)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(vars(field), nrow = 3) +
  scale_y_continuous(
    name = "Percentage (BA degrees awarded)"
  ) +
  scale_x_continuous(
    name = "Year"
  ) +
  theme(strip.background = element_rect(fill = "yellow"),
        strip.text = element_text(color = "red")
        )
```

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



I picked the following academic fields to make the below observations based on the above faceted graphs:

- ***Agriculture and natural resources:*** *This field has not seen much difference in proportion between 1970 and 2010.*
- ***Biological and biomedical sciences:*** *This field has a clear upward trend in the past few years.*
- ***Engineering:*** *This field had downward between mid 1980's and late 2000's, however, we can see slight increase in the trend year-over-year in the past few years.*
- ***Health professions and related programs:*** *This is the fastest growing academic field with steep upward trend in the past few years.*
- ***Mathematics and statistics:*** *This field is showing slow downward trend over the past decades, however, there are more BA graduates in the past few years.*
- ***Physical sciences and science technologies:*** *Similar to mathematics, this field also shows downward trend year-over-year.*

Problem 2: We will work the `txhousing` dataset provided by **ggplot2**. See here for details: 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 Stati~  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 Island    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: *The dataset contains the housing sales data from 46 cities in Texas. Not matter which plot we use (stacked bar, side-by-side bar, pie-chart), the graphs generated would not be visually appealing and easy to understand. It would make sense to show sales of top N cities and lump together the rest of the sales data from different cities. We could then use pie-chart or side-by-side chart to show relative proportion of sales.*

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))

# ggplot for pie-chart
ggplot(txhousing_lumped) +
  aes(
    x0 = 0, y0 = 0,
    r0 = 0, r = 1,
    amount = sales,
    fill = city
  ) +
  geom_arc_bar(stat = "pie") +
  coord_fixed() +
  xlim(-1.0, 1.0) +
  ylim(-1.1, 1.4) +
  theme_void() +
  theme(
    strip.text = element_blank(),
    strip.background = element_blank()
  )
```