Homework 8

Ethan Chang - ehc586

This homework is due on April 4, 2023 at 11:00pm. Please submit as a pdf file on Canvas.

Problem 1: (6 pts) The dataset `BA_degrees` contains information about the proportion of different degrees students receive, as a function of time.

```
head(BA_degrees)
```

```
## # A tibble: 6 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
```

Create a subset of the `BA_degrees` dataset that only considers the degree fields "Business", "Education", and "Psychology". Then make a single plot that satisfies these three criteria:

(a) Plot a time series of the proportion of degrees (colum `perc`) in each field over time and create a separate panel per degree field.
(b) Add a straight line fit to each panel.
(c) Order the panels by the difference between the maximum and the minimum proportion (i.e., the range of the data).

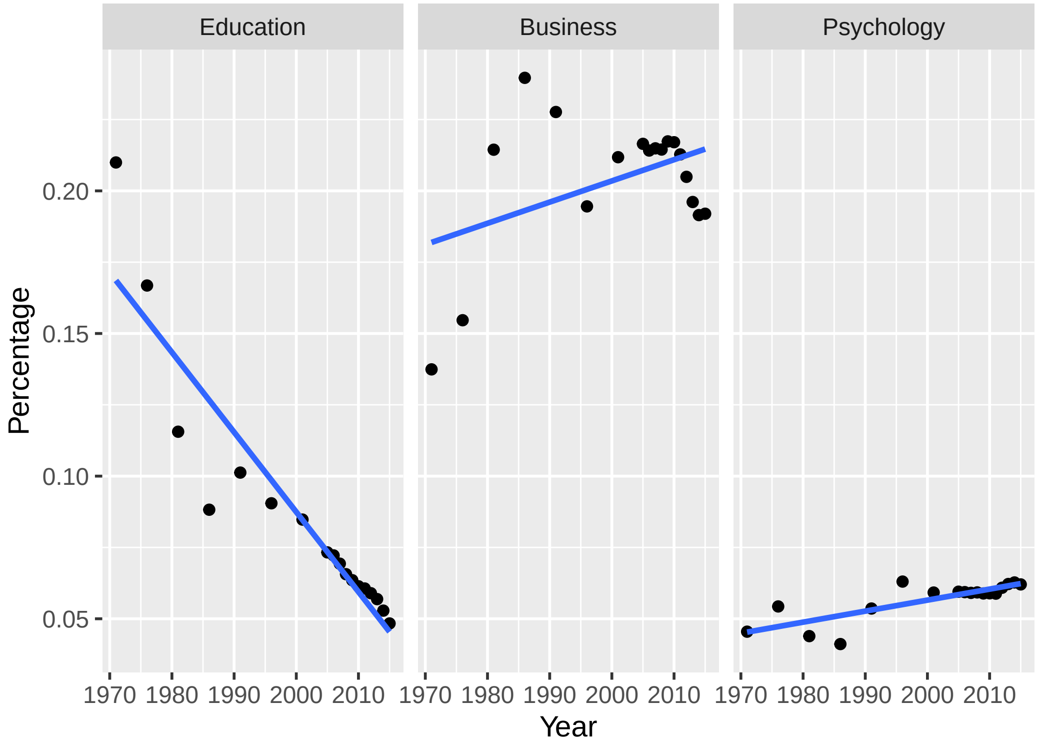
```
# your code goes here
BEP_degrees <- BA_degrees %>%
  filter(field %in% c("Business", "Education", "Psychology"))

reorder_range <- BEP_degrees %>%
  group_by(field) %>%
  summarize(range = diff(range(perc))) %>%
  arrange(-range)

reorder_fields <- reorder_range$field

ggplot(BEP_degrees, aes(year, perc)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~factor(field,levels = reorder_fields)) +
  labs(x = "Year", y = "Percentage")
```

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



Problem 2: (4 pts) Create a single pipeline that fits a linear model to each of the three fields from Problem 1 and outputs results in a tidy linear model summary table. The first column of the table should be `field` and the remaining columns should contain the linear model summary statistics such as `r.squared` for each field. Display the resulting table below.

```
# your code goes here
BEP_degrees %>%
  nest(data = -field) %>%
  mutate(
    fit = map(data, ~lm(perc ~ year, data = .x)),
    glance_out = map(fit, glance)) %>%
  select(field, glance_out) %>%
  unnest(cols = glance_out)
```

```
## # A tibble: 3 x 13
##   field      r.squa~1 adj.r~2   sigma stati~3 p.value    df logLik    AIC    BIC
##   <chr>         <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>  <dbl>  <dbl>  <dbl>
## 1 Business      0.177   0.126 0.0229     3.44 8.21e-2     1   43.5  -80.9  -78.3
## 2 Education     0.857   0.848 0.0163    96.0  3.63e-8     1   49.6  -93.1  -90.5
## 3 Psychology    0.655   0.633 0.00401   30.4  4.75e-5     1   74.8 -144.    -141.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>, and
## #   abbreviated variable names 1: r.squared, 2: adj.r.squared, 3: statistic
```