R for Empirical Economics Research Homework 2

Yichuan Zhang (47-216786)

Contents

Calculation of the state graduation rate

Set up library

```
Sys.setenv(LANG = "en")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
## v readr   2.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(griffen)
```

```
## Loading required package: magrittr
```

```
## 
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
## 
##     set_names
```

```
## The following object is masked from 'package:tidyr':
## 
##     extract
```

```
## Loading required package: lubridate
```

```
## 
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
## 
##     date, intersect, setdiff, union
```

To see what variables we have

```
options(width = 100)
# show what variables we have
names(cps)
```

```
##  [1] "age"                "year"               "wage"               "hours_lastweek"
##  [5] "employed"           "education_category" "educ_years"         "black"
##  [9] "white"              "female"             "married"            "single"
## [13] "divorced"           "state"              "region"             "sampling_weight"
```

dplyr:select

Select the variables we need

```
options(width = 100)
# we need to select the variables we need
new_df <- cps %>% select(state, education_category)
new_df
```

```
## # A tibble: 691,069 x 2
##    state        education_category
##    <chr>        <chr>
##  1 Ohio         highschool
##  2 Mississippi  highschool
##  3 Alaska       somecollege
##  4 North Dakota somecollege
##  5 Ohio         highschool
##  6 Kentucky     highschool
##  7 New Jersey   highschool
##  8 Michigan     somecollege
##  9 Delaware     highschool
## 10 Idaho        highschool
## # ... with 691,059 more rows
```

dplyr:group_by

Calculate the frequency of each categorical class for each state

```r
options(width = 100)

# group by the state and educatio_category and count all the categories

#count_df <- new_df %>% group_by(state, education_category) %>% summarise(n = n()).count()

count_df <- count(new_df %>% group_by(state, education_category))

count_df
```

```
## # A tibble: 153 x 3
## # Groups:   state, education_category [153]
##    state    education_category     n
##    <chr>    <chr>              <int>
##  1 Alabama  college             1355
##  2 Alabama  highschool          5117
##  3 Alabama  somecollege         2182
##  4 Alaska   college             2006
##  5 Alaska   highschool          4234
##  6 Alaska   somecollege         3431
##  7 Arizona  college             1768
##  8 Arizona  highschool          4414
##  9 Arizona  somecollege         2886
## 10 Arkansas college             1148
## # ... with 143 more rows
```

dplyr:filter; dplyr:summarise; dplyr:mutate

loop to get graduation rate for each state

```r
options(width = 100)
# get the unique states
unique_state <- unique(new_df["state"])[[1]]
```

```
# create an empty list
desired_length <- 1
graduation_rate <- rep(NA, desired_length)

for (i in unique_state) {
    # group by the dataframe
    group_df <- count_df %>% filter(state == i)
    # get the summation of all frequency of (college, others)
    total_number_df <- group_df %>% summarise(total_num = sum(n))
    # get the total population in this state
    total_number <- as.integer(total_number_df["total_num"] %>%
                    summarise(total_student = sum(total_num)))
    # get the precentage for each class
    graduation_rate_df <- count_df %>% filter(state == i) %>%
                    mutate(graduation_rate = n / total_number)
    # insert the graduation rate within in a list
    state_graduation_rate <- graduation_rate_df["graduation_rate"][[1]][1]
    graduation_rate <- c(graduation_rate, state_graduation_rate)
}

# remove the first na value in the list
graduation_rate <- graduation_rate[-1]
graduation_rate
```

```
##  [1] 0.1806519 0.1464154 0.2074243 0.2089841 0.1814624 0.2450819 0.1874676 0.2410952 0.1809235
## [10] 0.1947045 0.2203958 0.1825258 0.2241121 0.2103692 0.2706426 0.1565750 0.2456805 0.2159098
## [19] 0.2048812 0.2916188 0.3048440 0.1940942 0.2448820 0.2661879 0.1796601 0.1827148 0.4054960
## [28] 0.2118800 0.1809694 0.2202383 0.3026216 0.1734301 0.1405062 0.1742669 0.1997349 0.1949713
## [37] 0.2722555 0.1464659 0.2659961 0.2248695 0.2754799 0.3145783 0.1940869 0.1673764 0.1839734
## [46] 0.2495575 0.2138614 0.2235216 0.1615628 0.1829787 0.2144050
```

Make a dataframe for drawing the figure

used :factor

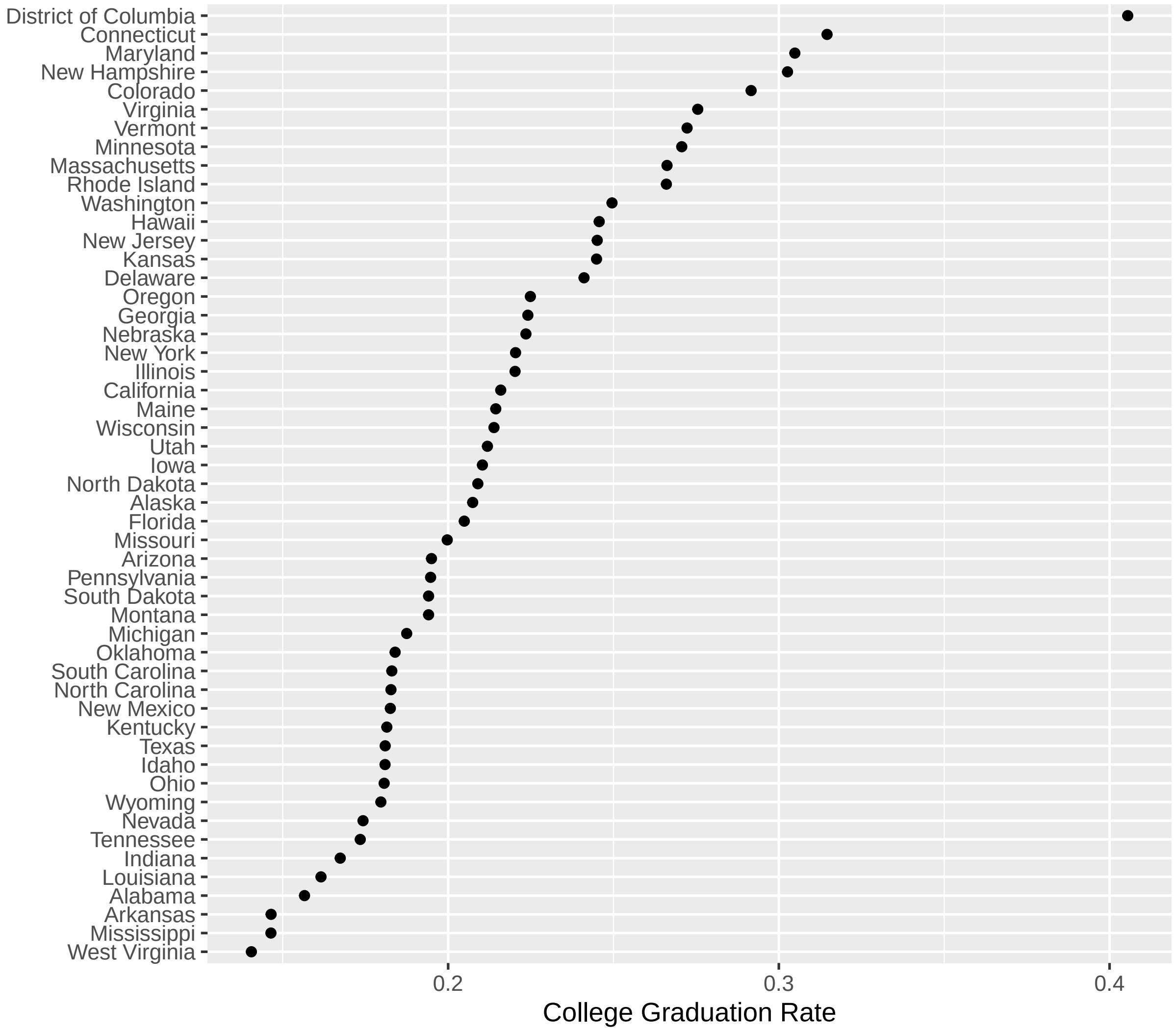
```
options(width = 100)
# make a new dataframe
final_df <- data.frame(unique_state, graduation_rate)
# order the dataframe
final_df <- final_df[order(graduation_rate,
 decreasing = FALSE),]
# rename the index
rownames(final_df) <- 1 : length(rownames(final_df))
# avoid the ggplot sort the geom_point automaticlly
final_df$unique_state <- factor(final_df$unique_state,
    levels = final_df$unique_state)
final_df
```

```
##             unique_state graduation_rate
## 1          West Virginia       0.1405062
## 2           Mississippi       0.1464154
## 3              Arkansas       0.1464659
## 4               Alabama       0.1565750
## 5             Louisiana       0.1615628
## 6               Indiana       0.1673764
## 7             Tennessee       0.1734301
## 8                Nevada       0.1742669
## 9               Wyoming       0.1796601
## 10                 Ohio       0.1806519
## 11                Idaho       0.1809235
## 12                Texas       0.1809694
## 13             Kentucky       0.1814624
## 14           New Mexico       0.1825258
## 15       North Carolina       0.1827148
## 16       South Carolina       0.1829787
## 17             Oklahoma       0.1839734
## 18             Michigan       0.1874676
## 19              Montana       0.1940869
## 20         South Dakota       0.1940942
## 21         Pennsylvania       0.1947045
## 22              Arizona       0.1949713
## 23             Missouri       0.1997349
## 24              Florida       0.2048812
## 25               Alaska       0.2074243
## 26         North Dakota       0.2089841
## 27                 Iowa       0.2103692
## 28                 Utah       0.2118800
## 29            Wisconsin       0.2138614
## 30                Maine       0.2144050
## 31           California       0.2159098
## 32             Illinois       0.2202383
## 33             New York       0.2203958
## 34             Nebraska       0.2235216
## 35              Georgia       0.2241121
## 36               Oregon       0.2248695
## 37             Delaware       0.2410952
## 38               Kansas       0.2448820
## 39           New Jersey       0.2450819
## 40               Hawaii       0.2456805
## 41           Washington       0.2495575
## 42         Rhode Island       0.2659961
## 43        Massachusetts       0.2661879
## 44            Minnesota       0.2706426
## 45              Vermont       0.2722555
## 46             Virginia       0.2754799
## 47             Colorado       0.2916188
## 48        New Hampshire       0.3026216
## 49             Maryland       0.3048440
## 50          Connecticut       0.3145783
## 51 District of Columbia       0.4054960
```

Draw the figure

used :fct_reorder

```
p <- ggplot(data = final_df,
 mapping = aes(x = fct_reorder(unique_state, graduation_rate, .desc = FALSE),
          y = graduation_rate)) +
  geom_point() + coord_flip() + labs(y = "College Graduation Rate", x = "")
p
```


District of Columbia
Connecticut
Maryland
New Hampshire
Colorado
Virginia
Vermont
Minnesota
Massachusetts
Rhode Island
Washington
Hawaii
New Jersey
Kansas
Delaware
Oregon
Georgia
Nebraska
New York
Illinois
California
Maine
Wisconsin
Utah
Iowa
North Dakota
Alaska
Florida
Missouri
Arizona
Pennsylvania
South Dakota
Montana
Michigan
Oklahoma
South Carolina
North Carolina
New Mexico
Kentucky
Texas
Idaho
Ohio
Wyoming
Nevada
Tennessee
Indiana
Louisiana
Alabama
Arkansas
Mississippi
West Virginia
College Graduation Rate