College Scorecard Analysis

Jaan Choudhri

5/16/2021

R Codebook for College Scorecard Analysis

```
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(ggplot2)
library(scales)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.1.1     v purrr   0.3.4
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
```

```
data <- read_csv("Most-Recent-Cohorts-All-Data-Elements.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   UNITID = col_double(),
##   HCM2 = col_double(),
##   MAIN = col_double(),
##   NUMBRANCH = col_double(),
##   PREDDEG = col_double(),
##   HIGHDEG = col_double(),
##   CONTROL = col_double(),
##   ST_FIPS = col_double(),
```

```
##   REGION = col_double(),
##   LOCALE = col_double(),
##   LATITUDE = col_double(),
##   LONGITUDE = col_double(),
##   CCBASIC = col_double(),
##   CCUGPROF = col_double(),
##   CCSIZSET = col_double(),
##   MENONLY = col_double(),
##   WOMENONLY = col_double(),
##   DISTANCEONLY = col_double(),
##   CURROPER = col_double(),
##   ICLEVEL = col_double()
##   # ... with 1 more columns
## )
## i Use `spec()` for the full column specifications.
```

```
## Warning: 4275 parsing failures.
##  row       col expected actual                                        file
## 6332 LOCALE    a double   NULL 'Most-Recent-Cohorts-All-Data-Elements.csv'
## 6332 LATITUDE  a double   NULL 'Most-Recent-Cohorts-All-Data-Elements.csv'
## 6332 LONGITUDE a double   NULL 'Most-Recent-Cohorts-All-Data-Elements.csv'
## 6332 CCBASIC   a double   NULL 'Most-Recent-Cohorts-All-Data-Elements.csv'
## 6332 CCUGPROF  a double   NULL 'Most-Recent-Cohorts-All-Data-Elements.csv'
## .... ......... ........ ...... ........................................
## See problems(...) for more details.
```

We have about 1800 schools with over 1700 variables.

We are only interested in a few categorical and quantitative variables. Here's how we can slim down our dataset to only focus on our desired variables: - award bachelor's degrees - are not for-profit institutions - are currently operating. - are in the 50 states

```
newData <- data %>%
  filter(PREDDEG == 3,
         CONTROL != 3,
         CURROPER == 1,
         ST_FIPS <= 56
         ) %>%
  select(school = INSTNM, institutionType = CONTROL, admit = ADM_RATE, med_income = MD_EARN_WNE_P10,
         med_fam_inc = MD_FAMINC, NPT4_PUB, NPT4_PRIV)
```

In order to properly evaluate our variables, we have to convert the quantitative variables to doubles instead of strings.

```
newData$NPT4_PUB <- as.double(newData$NPT4_PUB)
```

```
## Warning: NAs introduced by coercion
```

```
newData$NPT4_PRIV <- as.double(newData$NPT4_PRIV)
```

```
## Warning: NAs introduced by coercion
```

```
newData$med_income <- as.double(newData$med_income)
```

```
## Warning: NAs introduced by coercion
```

```
newData$med_fam_inc <- as.double(newData$med_fam_inc)
```

```
## Warning: NAs introduced by coercion
```

```
newData$admit <- as.double(newData$admit)
```

```
## Warning: NAs introduced by coercion
```

```
newData <- newData %>%
  rowwise %>%
  mutate(cost = sum(NPT4_PUB, NPT4_PRIV, na.rm = TRUE)) %>%
  select(-c(NPT4_PUB, NPT4_PRIV))
newData["cost"][newData["cost"] == 0] <- NA
```

```
head(newData)
```

```
## # A tibble: 6 x 6
## # Rowwise:
##   school                       institutionType  admit med_income med_fam_inc  cost
##   <chr>                                  <dbl>  <dbl>      <dbl>       <dbl> <dbl>
## 1 Alabama A & M University                   1  0.899      31000       23553  14444
## 2 University of Alabama at ~                1  0.921      41200       34489  17005
## 3 Amridge University                         2 NA          39600       15034. 15322
## 4 University of Alabama in ~                1  0.809      46700       44787  20909
## 5 Alabama State University                   1  0.977      27700       22080. 13043
## 6 The University of Alabama                  1  0.591      44500       66734. 22232
```

A 1 for funding indicates a public school, while 2s indicate a private school. Admit is the rate of admission, med_earnings is the median income of a student following graduation, med_fam_inc is the students' family's median income, and price is the cost of attendance of the institution.

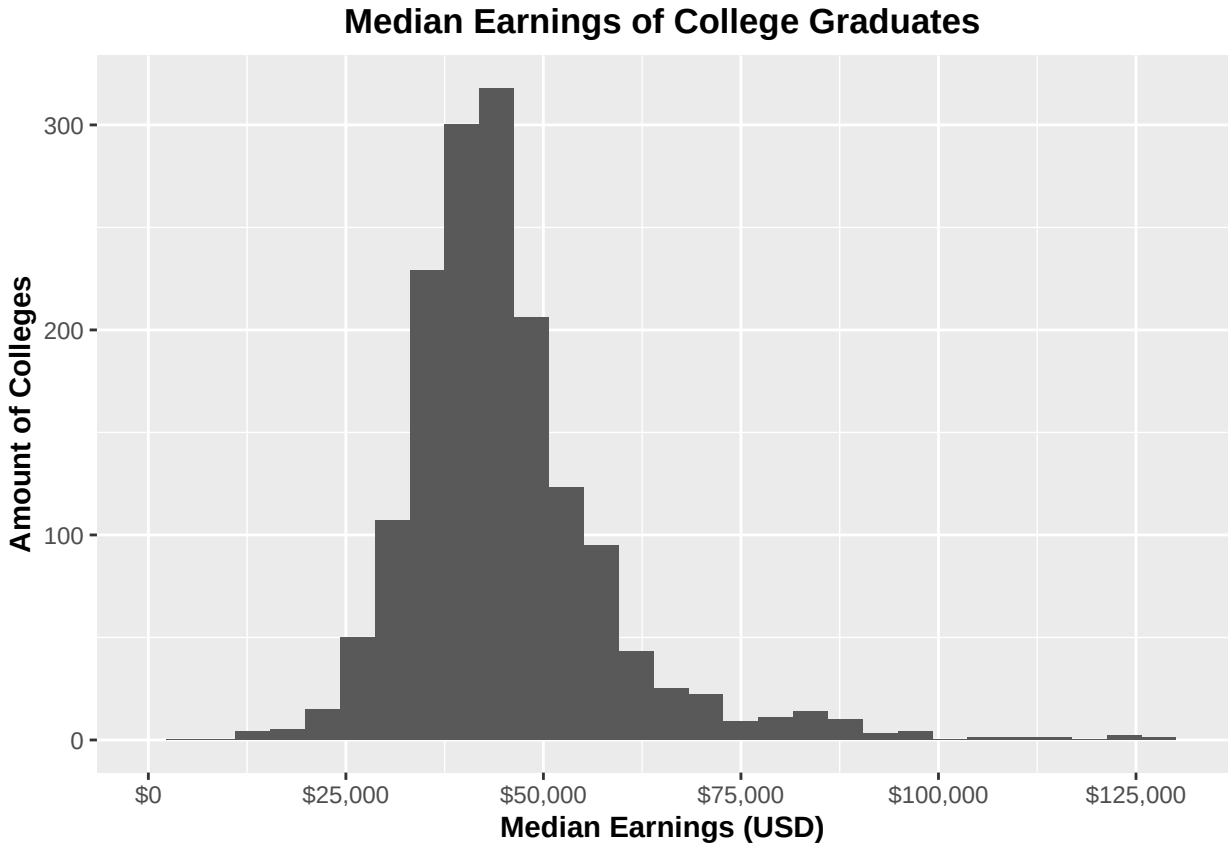
```
xdollar <- c(scale_x_continuous(labels = dollar,
                                breaks = seq(0, 130000, 25000),
                                limits = c(0, NA)))

ydollar <- c(scale_y_continuous(labels = dollar,
                                        breaks = seq(0, 130000, 25000),
                                        limits = c(0, NA)))

titling <- theme(plot.title = element_text(hjust = 0.5,
                                           face = "bold"),
                 axis.title.x = element_text(face = "bold"),
                 axis.title.y = element_text(face = "bold"))
```

```
ggplot(data = newData) +
  geom_histogram(mapping = aes(x = med_income)) +
  labs(title = "Median Earnings of College Graduates",
       x = "Median Earnings (USD)",
       y = "Amount of Colleges") +
  xdollar +
  titling
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 201 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

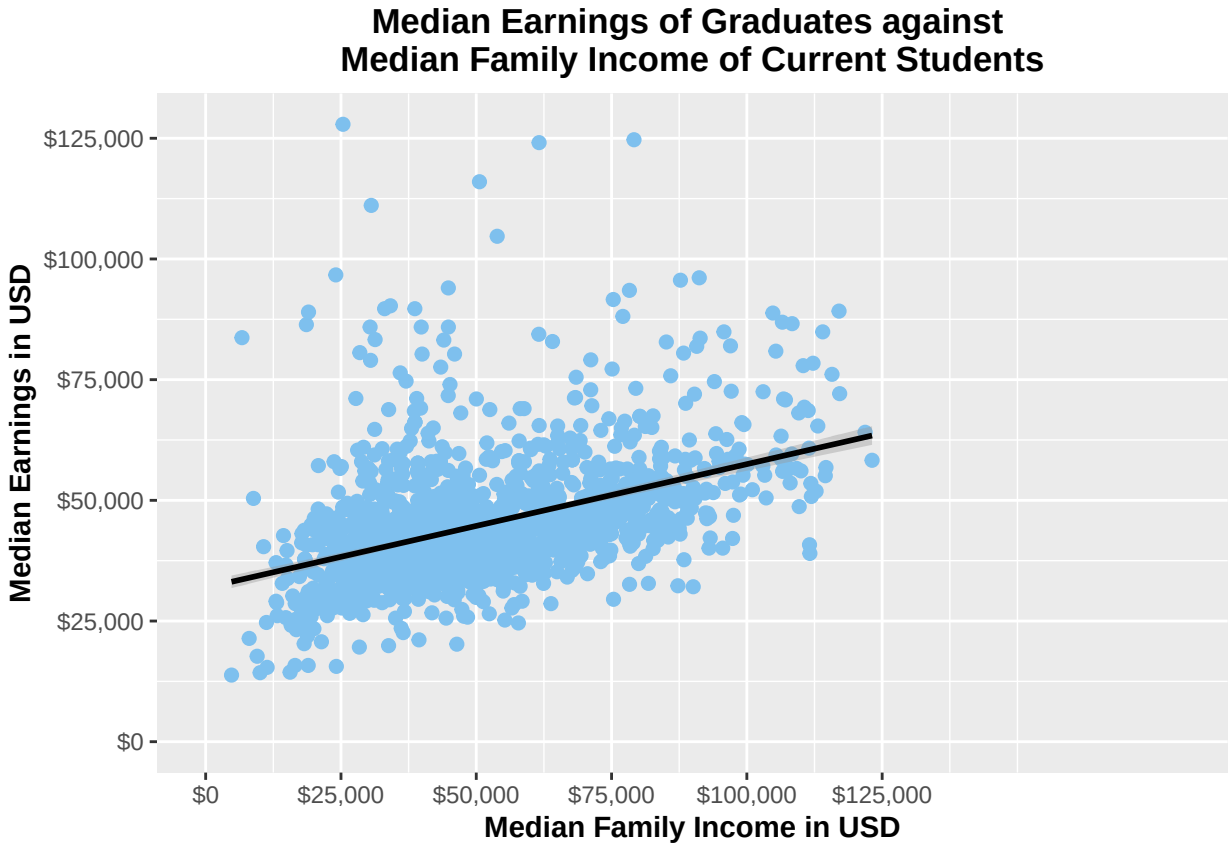



```
ggplot(data = newData,
       mapping = aes(x = med_fam_inc,
                     y = med_income)) +
  geom_point(size = 2,
             color = "skyblue2") +
  geom_smooth(method = "lm",
              color = "black") +
  labs(title = "Median Earnings of Graduates against \nMedian Family Income of Current Students",
       x = "Median Family Income in USD",
       y = "Median Earnings in USD") +
  xdollar +
  ydollar +
  titling
```

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

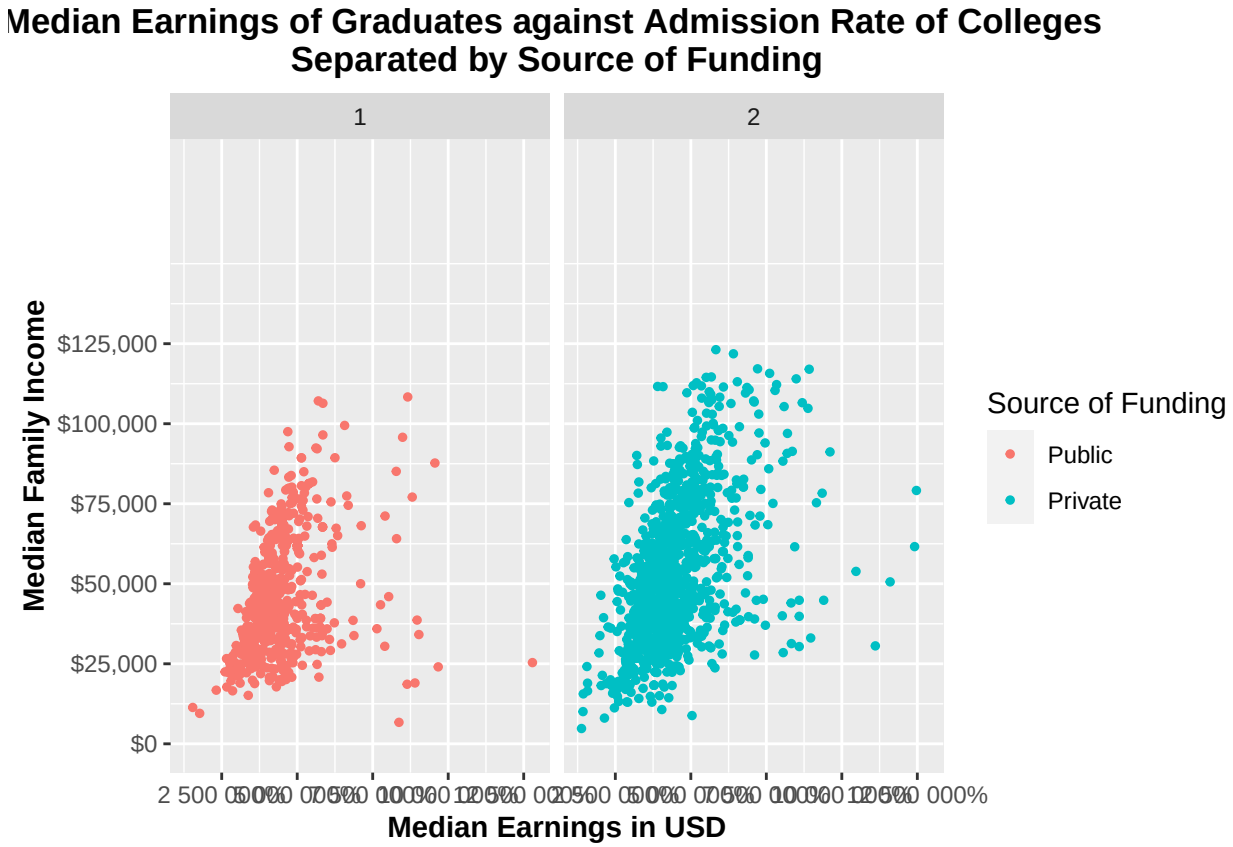
```
## Warning: Removed 202 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 202 rows containing missing values (geom_point).
```




```
point_theme <- c(scale_x_continuous(labels = percent),
                 ydollar,
                 scale_color_manual(labels = c("Public","Private"),
                                    values = c("#F8766D", "#00BFC4")))

scatter <- ggplot(data = newData,
                  mapping = aes(x = med_income,
                                y = med_fam_inc)) +
  geom_point(mapping = aes(color = factor(institutionType)),
             size = 1) +
  point_theme +
  titling
scatter +
  labs(title = "Median Earnings of Graduates against Admission Rate of Colleges \nSeparated by Source o
       x = "Median Earnings in USD",
       y = "Median Family Income",
       color = "Source of Funding") +
  facet_wrap(~ institutionType)
```

```
## Warning: Removed 202 rows containing missing values (geom_point).
```




```
fit2 <- lm(data = newData,
           med_income ~ med_fam_inc + institutionType + med_fam_inc*institutionType)
summary(fit2)
```

```
##
## Call:
## lm(formula = med_income ~ med_fam_inc + institutionType + med_fam_inc *
##     institutionType, data = newData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23079  -6950  -2210   3723  87507
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.984e+04  2.698e+03  14.764  < 2e-16 ***
## med_fam_inc                   1.529e-01  5.402e-02   2.830  0.00471 **
## institutionType              -4.935e+03  1.555e+03  -3.173  0.00154 **
## med_fam_inc:institutionType   6.348e-02  3.008e-02   2.111  0.03496 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11270 on 1594 degrees of freedom
##   (202 observations deleted due to missingness)
## Multiple R-squared:  0.2027, Adjusted R-squared:  0.2012
## F-statistic: 135.1 on 3 and 1594 DF,  p-value: < 2.2e-16
```