Exploratory Data Analysis

Alice Chang
Ben Thomas
Olek Wojcik

11/19/2019

Data description

Our group is trying to explore why students of different colleges/universities have different employment outcomes. Specifically, we're interested in both (1) students' employment rate after graduation and (2) students' salaries after graduation. To approach this problem, we're using data from the U.S. Department of Education: the "College Scorecard."

The College Scorecard (found here: https://collegescorecard.ed.gov/data/) has an absolutely huge amount of data, across many years. We've started our analysis with the 2014-15 dataset, which is more complete than more recent datasets. Before cleaning the data, we have 1,977 predictor variables and 7,703 observations, which represent individual colleges and universities. After cleaning the data (below), we have 1,593 predictor variables and 5,492 observations. This large decrease is primarily due to (1) removing the variables which only contained `NA` values and (2) removing rows for which our dependent variable (`MN_EARN_WNE_P6`) was `NA`.

Data exploration

Before diving into the exploration, we first need to actually get the data.

```
college.data <- read.csv("2014_2015_college_data.csv", header = TRUE)
```

Now, let's take a look at this data. How much is missing? To do this, I'll use the `Amelia` package. Note that I've changed the "NULL" and "PrivacySuppressed" values to "NA" so that they show up in this visualization. Note also that instead of the real output, I've attached a screenshot of the output, as the actual pdf file with the output included is *gigantic*.

```
library(Amelia)
library(tidyverse)

college.data <- college.data %>%
  replace(.=="NULL", NA) %>%
  replace(.== "PrivacySuppressed", NA)

missmap(college.data)
```

61% of the data in this dataset is missing; not ideal. Let's now clean this data and get it in workable form.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.1     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

Figure 1:

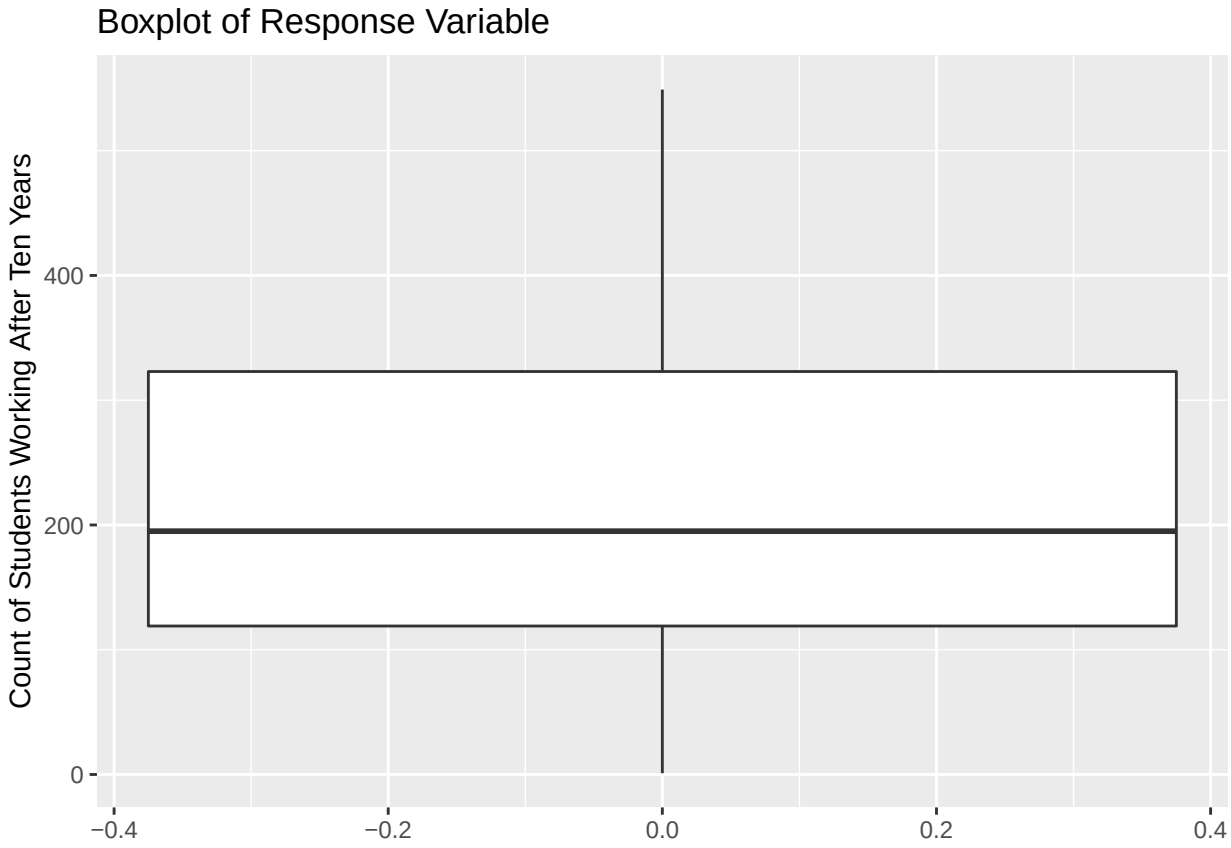
```
## -- Conflicts ------------------------------------------------------------------------------------------ t
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
not_all_na <- function(x) any(!is.na(x))

# Remove intro variables and other earnings variables
college.data.section <- college.data %>%
  select(MAIN:OMENRUP_PARTTIME_POOLED_SUPP)

# Drop NAs in response
college.data.section <- college.data.section %>%
  drop_na(MN_EARN_WNE_P6)

# Remove columns with only NA values
college.data.section <- college.data.section %>%
  select_if(not_all_na)

# Make all columns numeric
college.data.section[, c(1:ncol(college.data.section))] <- sapply(college.data.section[, c(1:ncol(colleg
```

Univariate Analysis of Response

So now let's examine our response variable, MN_EARN_WNE_P6, which is the number of students working and not enrolled 10 years after entry.

```
library(ggplot2)
ggplot(data = college.data.section, mapping = aes(y = MN_EARN_WNE_P6)) + geom_boxplot() + labs(y = "Cou
```

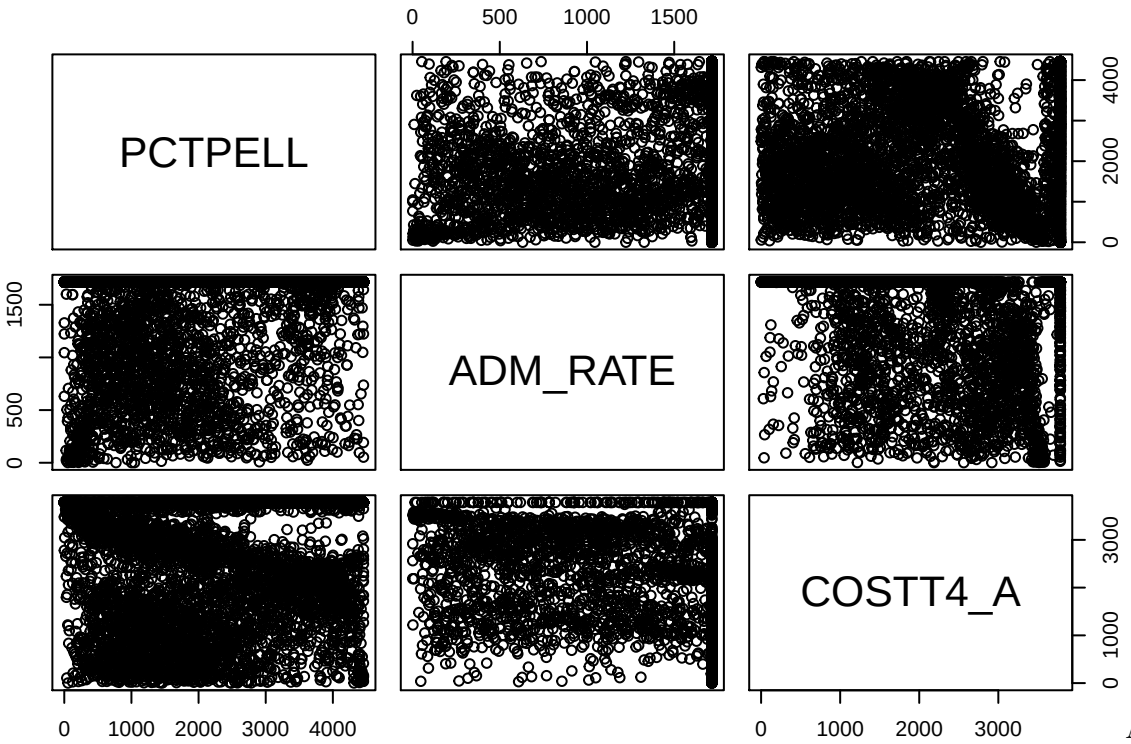


Visualizing Multicollinearity

Because of the large amount of missing data still left, it's difficult to calculate the exact correlations between a few variables that we assume are important in this case. Here are some scatter plots of a few potentially important predictors:

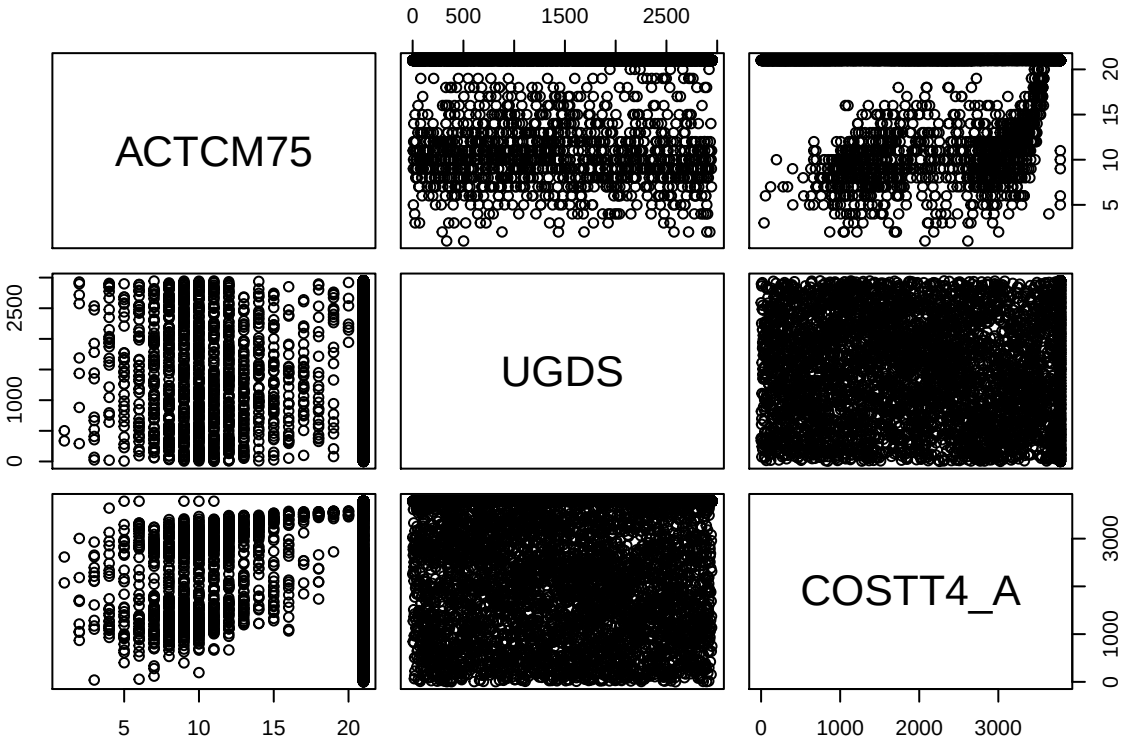
PCTPELL: percentage of students receiving pell grants ADM_RATE: admission rate COSTT4_A: Average cost of attendance (academic year institutions)

```
pairs(~PCTPELL + ADM_RATE + COSTT4_A, college.data.section)
```



ACTCM75: 75th percentile of the ACT cumulative score UGDS: Enrollment of undergraduate certificate/degree-seeking students

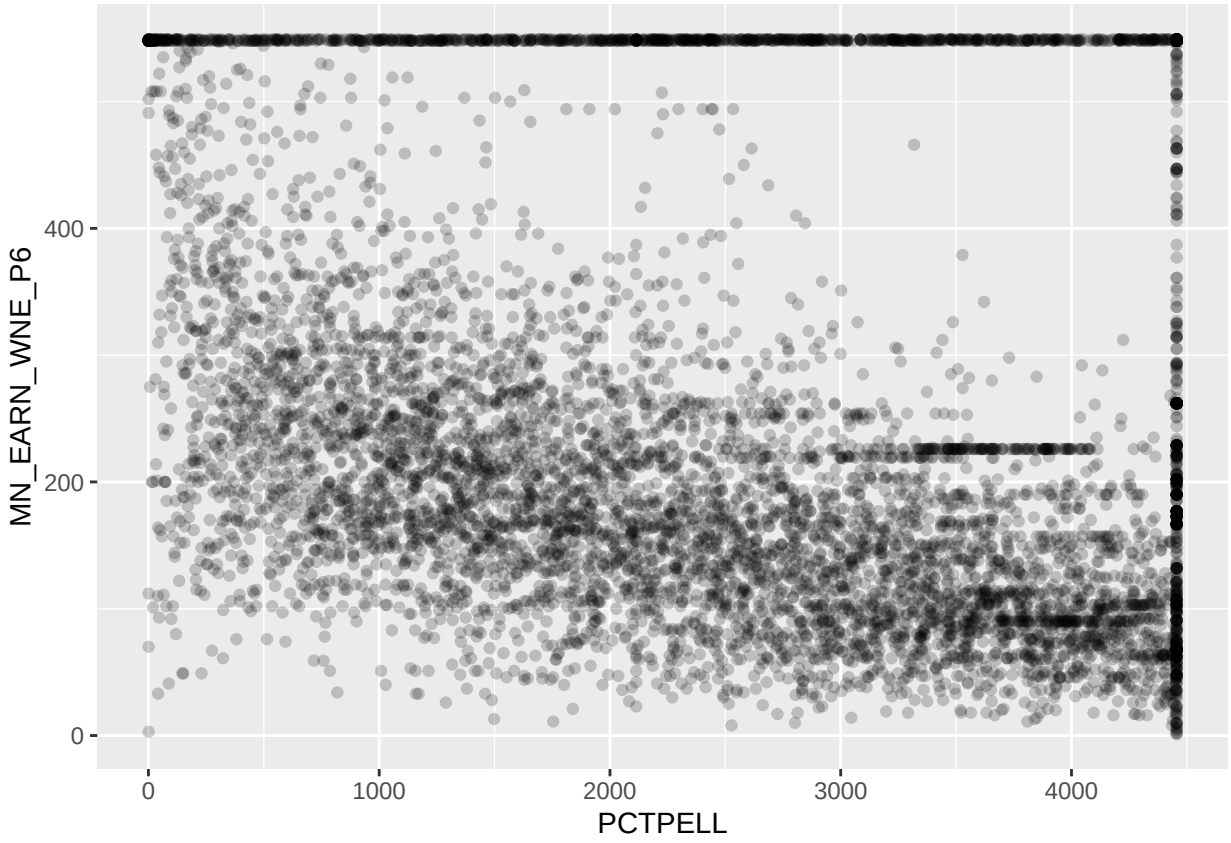
```
pairs(~ACTCM75 + UGDS + COSTT4_A, college.data.section)
```



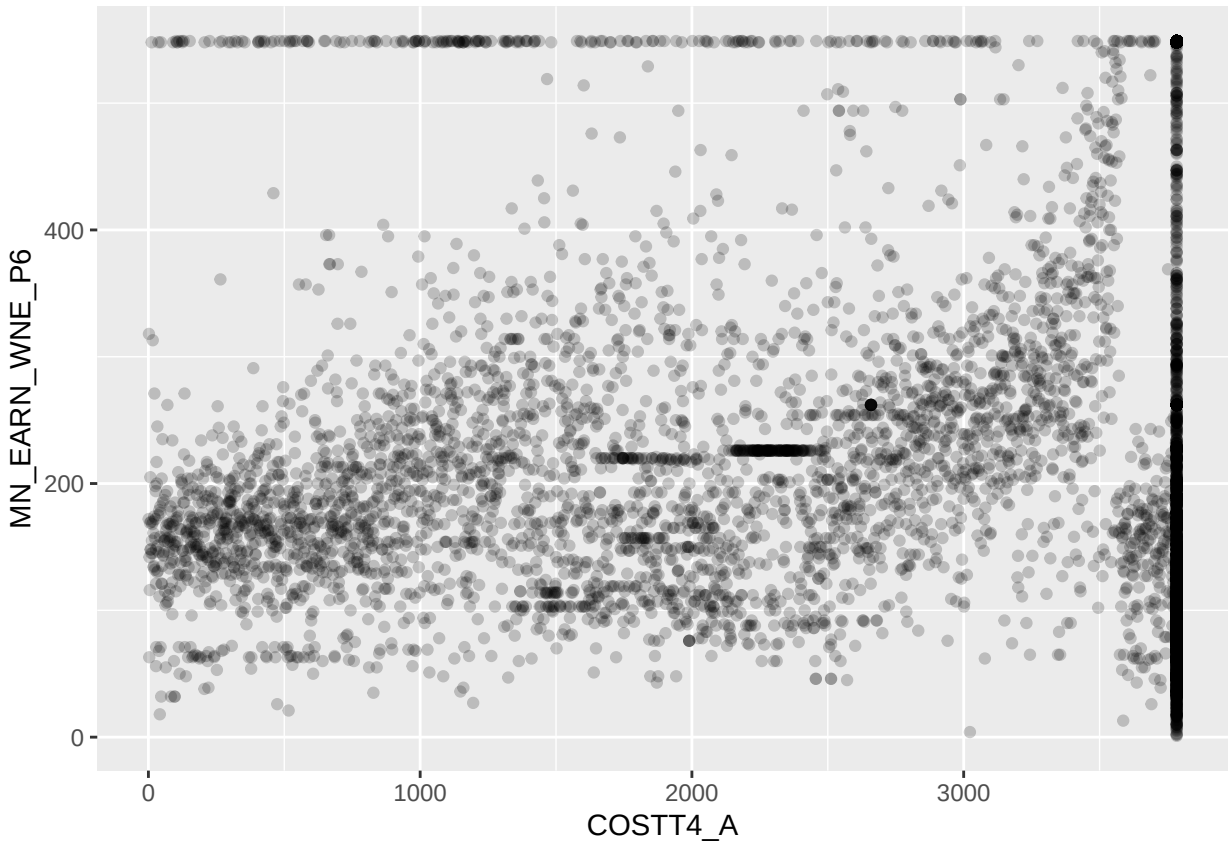
Bivariate Visual Representations

We ran a random forest model on the data, which said that LO_INC_YR8_N has the highest importance in terms of both MSE and node purity. It is the number of low-income (less than $30,000 in nominal family income) students in overall 8-year completion cohort. Below, we've graphed both `LO_INC_YR8_N` and `PCTPELL` with our response variable.

```
ggplot(college.data.section, aes(x=PCTPELL , y=MN_EARN_WNE_P6)) +
  geom_point(alpha = 0.2)
```




```
ggplot(college.data.section, aes(x= COSTT4_A, y=MN_EARN_WNE_P6)) +
  geom_point(alpha = 0.2)
```


MN_EARN_WNE_P6
400
200
0
0
1000
2000
3000
COSTT4_A