

# 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.

```
library(Amelia)
```

```
## Loading required package: Rcpp

## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.5, built: 2018-05-07)
## ## Copyright (C) 2005-2019 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
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 tidyverse 1.0.0      v stringr  1.4.0
## v readr    1.3.1      vforcats  0.4.0
```

```

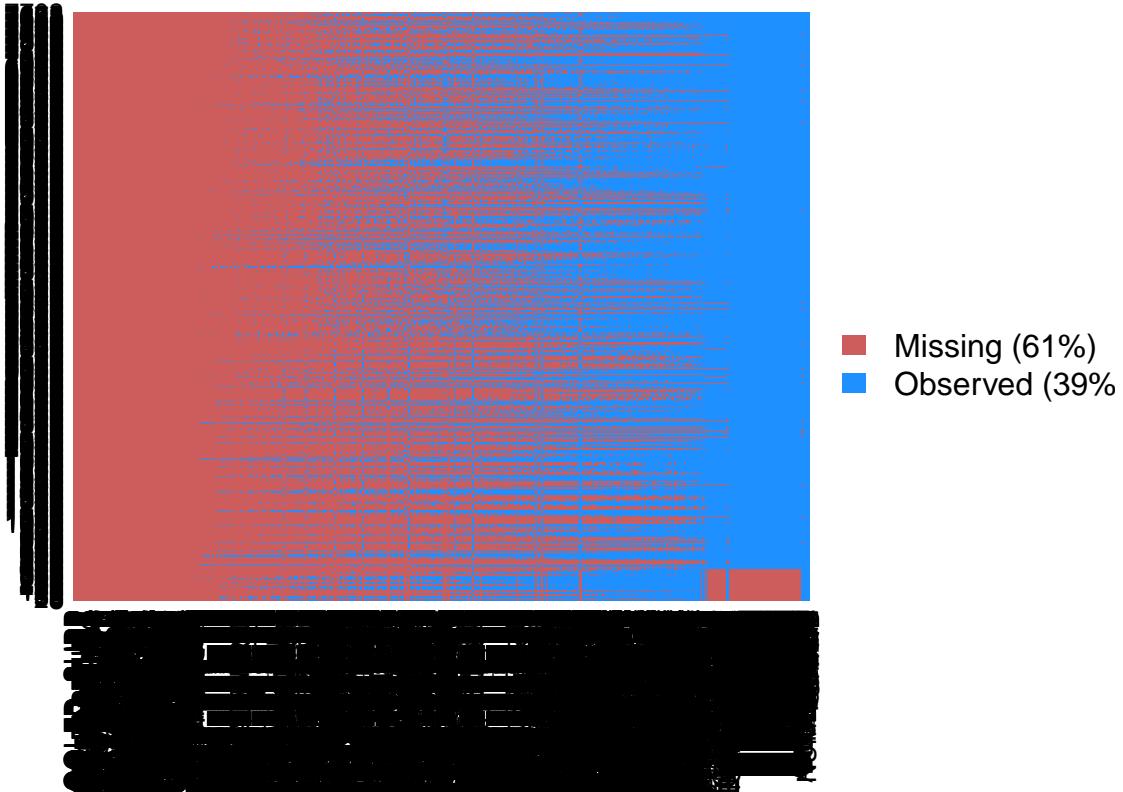
## -- Conflicts -----
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

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

missmap(college.data)

```

## Missingness Map



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

```

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)

```

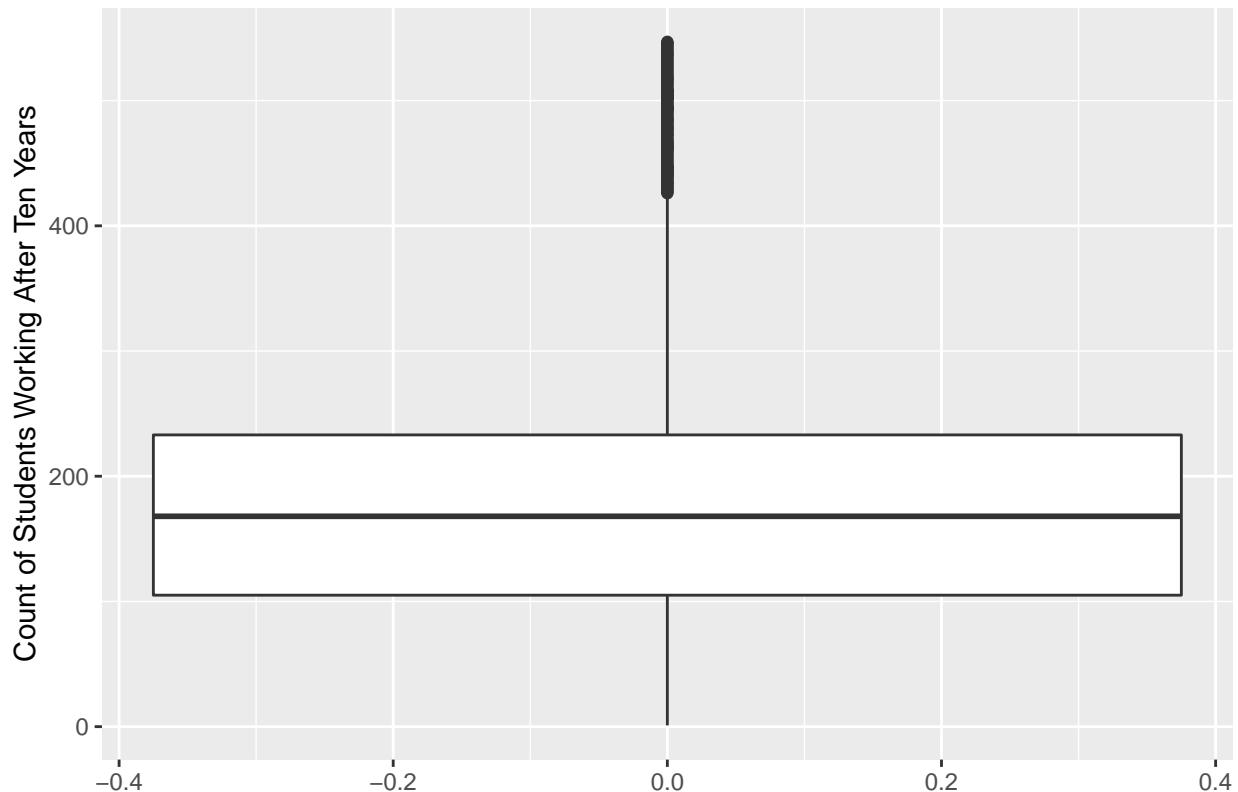
```
# Make all columns numeric
college.data.section[, c(1:ncol(college.data.section))] <- sapply(college.data.section[, c(1:ncol(college
```

### Univariate Analysis of Response

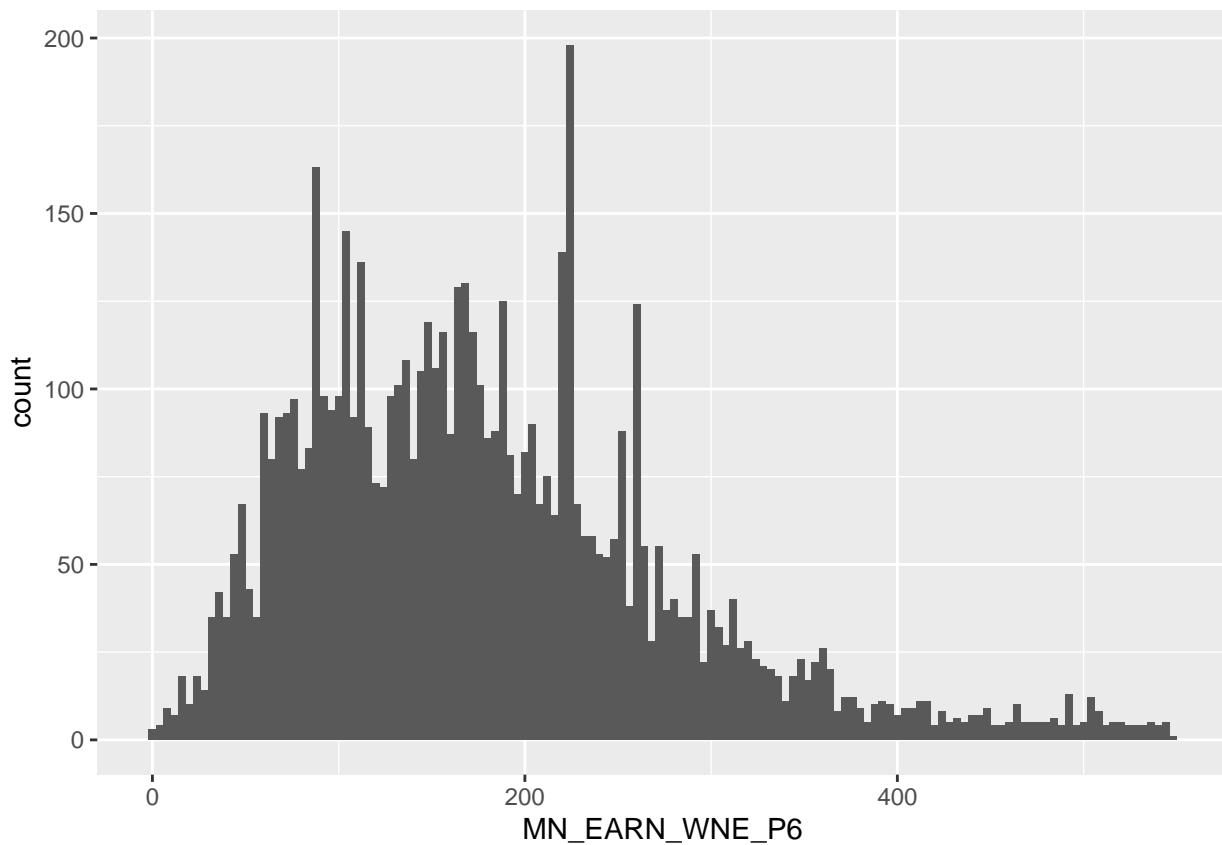
So now let's examine our response variable, MN\_EARN\_WNE\_P6, which is the number of students working and not enrolled 10 years after entry.

```
library(ggplot2)
ggplot(data = college.data.section, mapping = aes(y = MN_EARN_WNE_P6)) + geom_boxplot() + labs(y = "Count of Students Working After Ten Years")
```

Boxplot of Response Variable



```
ggplot(college.data.section, aes(x=MN_EARN_WNE_P6)) + geom_histogram(binwidth = 4)
```

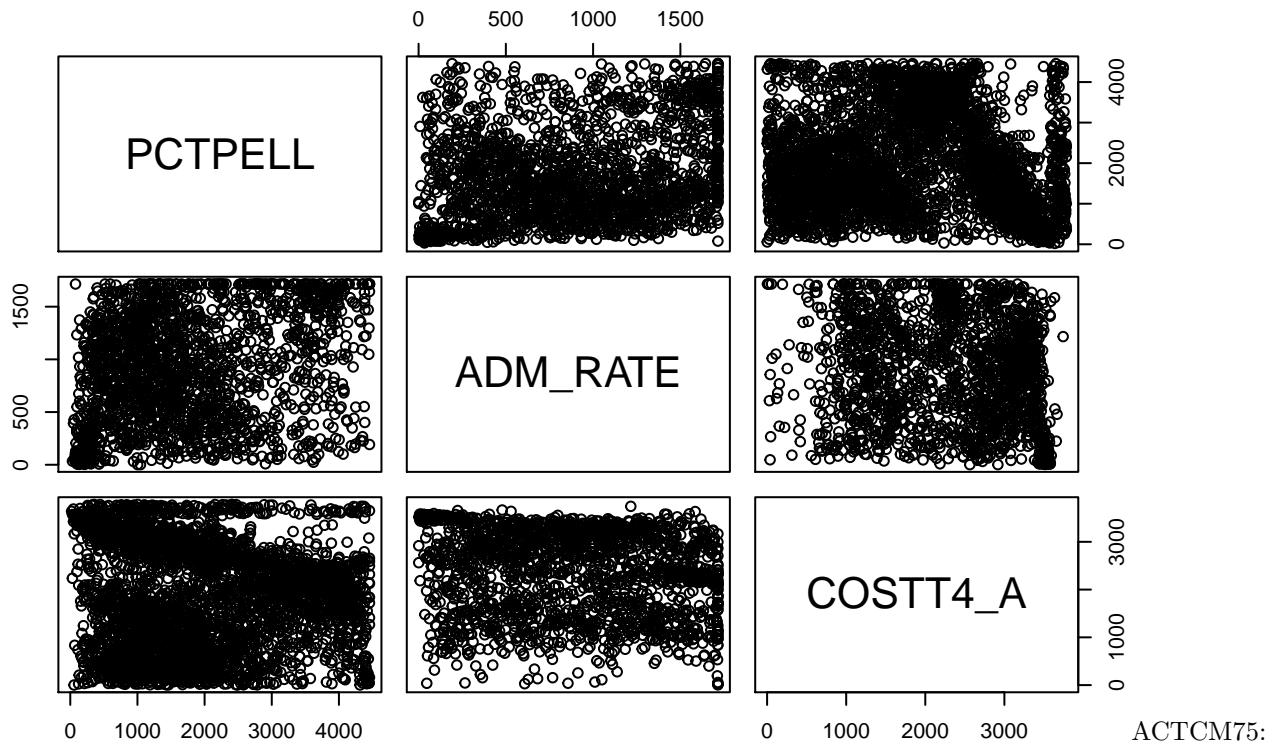


### Visualizing Multicollinearity

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:

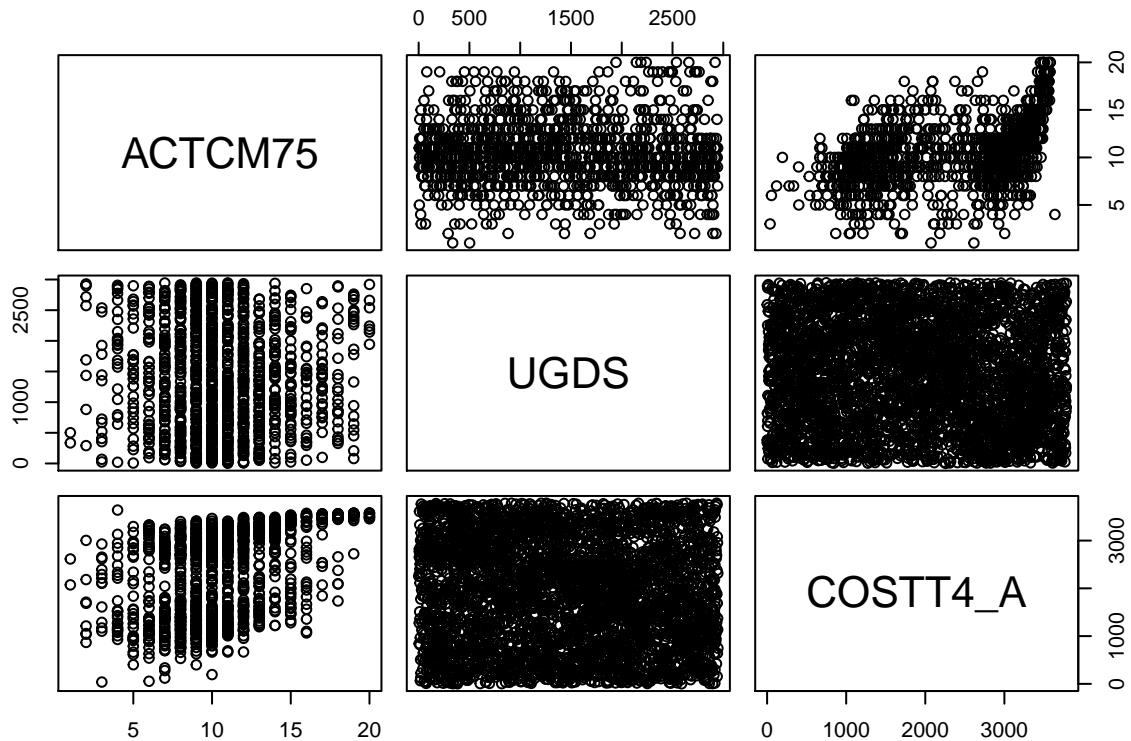
PCTPELL: percentage of students receiving pell grants  
ADM\_RATE: admission rate  
COSTT4\_A: Average cost of attendance (academic year institutions)

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



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

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

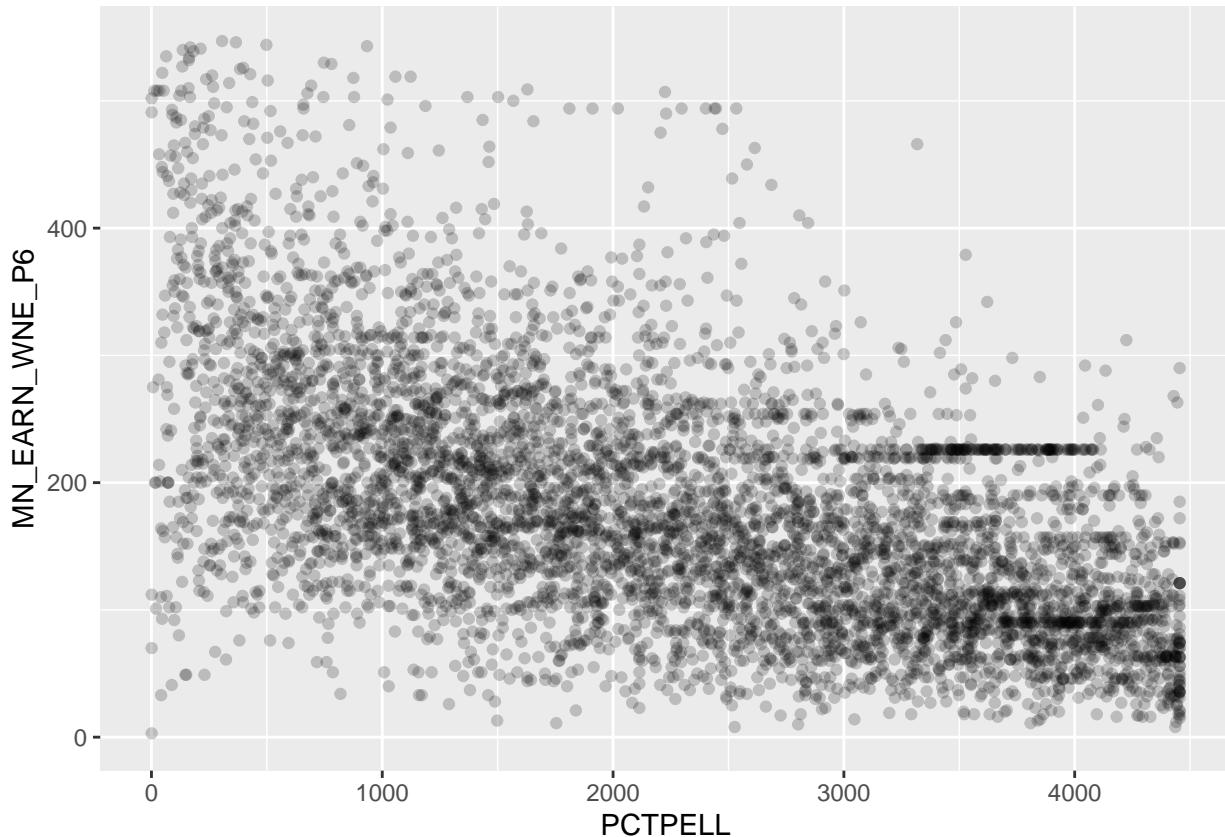


## Bivariate Visual Representations

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)
```

## Warning: Removed 461 rows containing missing values (geom\_point).



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

## Warning: Removed 2499 rows containing missing values (geom\_point).

