ANLY 500 Final Project

Data Cleaning

- MERGED2012_13_PP.csv is too large for git repo, please copy it to this directory manually.
- Choose 12-13 data because later data do not have salary information.
- `mrc_table10` is data from Mobility Report Card data, they give each school a tier. Looks more useful than payscale school type. For detail, please see Codebook-MRC-Table-6.pdf.

Read Data

```
payscale_college_type = read.csv('salaries-by-college-type.csv')
payscale_college_type = payscale_college_type[, c(1,2)]
payscale_region = read.csv('salaries-by-region.csv')
payscale_region = payscale_region[, c(1,2)]

mrc_table10 = read.csv('mrc_table10.csv')
mrc_table10 = mrc_table10[, c(1, 12)]

# Treat "NULL" and "PrivacySuppressed" as NA when read
college_scorecard = read.csv('MERGED2012_13_PP.csv', na=c("NULL", "PrivacySuppressed"))
college_scorecard = college_scorecard[, c(3, 4,6,17,377,379,380,1638,1639,1640,1642,1643,1645,1646,1647
```

Merge Data

Process School Name

Normalize all school name for merging.

```
process_school_name = function(data) {
  data = sub(" \\(.*\\)", "", data)
  data = sub(" - ", "-", data)
  data = sub(", ", "-", data)
  data = sub("\\.", "", data)
  data = sub(" & ", " and ", data)
  data = sub("&", " and ", data)
  data = sub("St ", "Saint ", data)
  data
}

# 269 rows
payscale_college_type$School.Name = process_school_name(payscale_college_type$School.Name)
# 320 rows
payscale_region$School.Name = process_school_name(payscale_region$School.Name)
# 7793 rows
college_scorecard$INSTNM = process_school_name(college_scorecard$INSTNM)
```

Merge Payscale data

```
payscale = merge(payscale_college_type, payscale_region, by="School.Name", all = FALSE)

payscale %>% group_by(School.Name) %>% filter(n() > 1)

payscale_party = payscale %>%
  filter(School.Type == "Party")

# 248 rows
payscale = payscale %>%
  filter(School.Type != "Party") %>%
  mutate(Is.Party = School.Name %in% payscale_party$School.Name)
```

All the duplicate rows in Payscale data is because it duplicates all Party Schools. So we split whether or not is a party school into a separate column.

Merge with College Scorecard

```
scorecard_payscale = merge(payscale, college_scorecard, by.y = "INSTNM", by.x = "School.Name")

scorecard_payscale %>% group_by(School.Name) %>% filter(n() > 1)

# 181 rows
scorecard_payscale = scorecard_payscale %>%
  filter( !((School.Name == 'Union College' & STABBR != 'NY') |
          (School.Name == 'Wentworth Institute of Technology' & is.na(COSTT4_A))))
```

Two schools are duplicate after merged with College Scorecard data. After some search, we only keep Union College in New York because that is the only one in northwest. And we only keep one Wentworth Institute of Technology because the others do not have data.

Merge with MRC

```
# 147 rows
data = merge(scorecard_payscale, mrc_table10, by.x = "OPEID6", by.y = "super_opeid")

data %>% group_by(School.Name) %>% filter(n() > 1)
```

No duplicate in this step

Other processing

```
names(data)[names(data) == 'tier_name'] <- 'Tier'
earning_colnames = c("COUNT_WNE_P6", "MN_EARN_WNE_P6", "MD_EARN_WNE_P6", "PCT25_EARN_WNE_P6",
"PCT75_EARN_WNE_P6", "SD_EARN_WNE_P6", "COUNT_WNE_INC1_P6", "COUNT_WNE_INC2_P6",
"COUNT_WNE_INC3_P6", "COUNT_WNE_MALE0_P6", "COUNT_WNE_MALE1_P6",
"MN_EARN_WNE_INC1_P6", "MN_EARN_WNE_INC2_P6", "MN_EARN_WNE_INC3_P6",
"MN_EARN_WNE_MALE0_P6", "MN_EARN_WNE_MALE1_P6", "COUNT_WNE_P8",
"MD_EARN_WNE_P8", "COUNT_WNE_P10", "MN_EARN_WNE_P10", "MD_EARN_WNE_P10", "PCT25_EARN_WNE_P10",
"PCT75_EARN_WNE_P10", "SD_EARN_WNE_P10", "COUNT_WNE_INC1_P10",
"COUNT_WNE_INC2_P10", "COUNT_WNE_INC3_P10", "COUNT_WNE_MALE0_P10",
"COUNT_WNE_MALE1_P10", "MN_EARN_WNE_INC1_P10", "MN_EARN_WNE_INC2_P10",
"MN_EARN_WNE_INC3_P10", "MN_EARN_WNE_MALE0_P10", "MN_EARN_WNE_MALE1_P10")
cost_colnames = c("COSTT4_A", "TUITIONFEE_IN", "TUITIONFEE_OUT")
data = data[c("School.Name", "School.Type", "Region", "Is.Party",
"STABBR", "CONTROL", "Tier", cost_colnames, earning_colnames)]

data$CONTROL = factor(data$CONTROL, levels = c(1,2), labels = c("Public", "Private nonprofit"))

write_csv(data, "data_cleaned.csv")
```

We reorder the column for easy inspection. And convert CONTROL into factor.

Accuracy and Outlier

Accurary & Missing value

```
summary(data)

percentmiss <- function(x){length(x[is.na(x)])/length(x)*100}

# process column first will get more records left
missing_col = apply(data, 2, percentmiss)
missing_col
delete <- which(missing_col > 5)
replace_col = data[,-delete]
dont_col = data[,delete]

missing_row = apply(replace_col, 1, percentmiss)
missing_row[missing_row > 5]
replace_row = subset(replace_col, missing_row <= 5)
dont_row = subset(replace_col, missing_row > 5)

# change to "cart" to avoid error, increase iteration to get reliable result
temp_no_miss = mice(replace_row, maxit=100, method='cart', seed=500)
no_miss = complete(temp_no_miss,1)

# combine data back
all_rows = rbind(dont_row, no_miss)
all_col = cbind(dont_col, all_rows)
```

There is no accuracy problem in the data. We use mice to complete the missing value for data meet 5% rule.

Outlier

```
# pass tolerance to prevent mahalanobis think it is singular matrix
mahal <- mahalanobis(no_miss[-c(1:7)],
                     colMeans(no_miss[-c(1:7)],na.rm=TRUE),
                     cov(no_miss[-c(1:7)], use = "pairwise.complete.obs"),
                     tol=1e-30)
```

```r
cutoff = qchisq(1-.001,ncol(no_miss[-c(1:7)]))
print(cutoff)
```

```
## [1] 63.8701
```

```r
summary(mahal < cutoff)
```

```
##    Mode   FALSE    TRUE 
## logical      11     131
```

```r
noout = subset(no_miss, mahal < cutoff)

no_miss[mahal >= cutoff, c("School.Name", "COSTT4_A", "TUITIONFEE_IN", "TUITIONFEE_OUT", "MN_EARN_WNE_P
```

```
##                                        School.Name COSTT4_A TUITIONFEE_IN
## 11                      San Diego State University    19560          6578
## 30                                 Yale University    58250         42300
## 57                                 Amherst College    56898         44610
## 59           Massachusetts Institute of Technology    55270         42050
## 75                               Dartmouth College    58638         45042
## 76                            Princeton University    53934         39537
## 78  New Mexico Institute of Mining and Technology    17172          5496
## 90                                Davidson College    52498         40809
## 113                             Swarthmore College    55895         43080
## 114                     University of Pennsylvania    57360         43738
## 141                  University of Central Florida    18213          6247
##     TUITIONFEE_OUT MN_EARN_WNE_P6 MD_EARN_WNE_P6
## 11           18236          38300          35200
## 30           42300          75500          60100
## 57           44610          44700          35800
## 59           42050         102700          77900
## 75           45042          69900          54600
## 76           39537          75600          52400
## 78           16367          46200          39700
## 90           40809          48300          41500
## 113          43080          40900          33700
## 114          43738          89000          67200
## 141          22345          36600          34100
```

We get 11 schools as outliers. Need more analysis about them.