ANLY 500 Final Project

Data Cleaning

- Unzip MERGED2014_15_PP.csv.zip, it is from College Scorecard webiste, not Kaggle. Since Kaggle do not have latest data.
- `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('MERGED2014_15_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
}

payscale_college_type$School.Name = process_school_name(payscale_college_type$School.Name)
nrow(payscale_college_type)
```

```
## [1] 269
```

```
payscale_region$School.Name = process_school_name(payscale_region$School.Name)
nrow(payscale_region)
```

```
## [1] 320
```

```r
college_scorecard$INSTNM = process_school_name(college_scorecard$INSTNM)
nrow(college_scorecard)
```

```
## [1] 7703
```

Merge Payscale data

```r
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")

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

nrow(payscale)
```

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

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

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

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

nrow(scorecard_payscale)
```

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

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

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

nrow(data)
```

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

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

```
## [1] 59.70306
```

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

```
##    Mode   FALSE    TRUE
## logical      10     132
```

```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
## 16                          Pomona College    59730         45832
## 30                         Yale University    61620         45800
## 56                           Colby College    59110         47350
## 57                         Amherst College    61544         48526
## 59  Massachusetts Institute of Technology    59020         45016
## 61                        Williams College    61850         48310
## 76                    Princeton University    57400         41820
## 107             Carnegie Mellon University    61990         49022
## 114             University of Pennsylvania    61800         47668
## 124                     University of Utah    18931          7835
##     TUITIONFEE_OUT MN_EARN_WNE_P6 MD_EARN_WNE_P6 MN_EARN_WNE_P10
## 16           45832          51200          41100           77300
## 30           45800          67800          56600          124400
## 56           47350          50200          42700           71000
## 57           48526          61600          44100           83300
## 59           45016          99600          82200          153600
## 61           48310          51400          42600           89800
## 76           41820          73600          60800          116300
## 107          49022          84000          69800          103000
## 114          47668          91200          71600          131600
## 124          25057          47200          40800           63500
##     MD_EARN_WNE_P10
## 16            58100
## 30            83200
## 56            58100
## 57            65000
## 59           104700
## 61            59000
## 76            74700
## 107           83600
## 114           85900
## 124           53000
```

We get 10 schools as outliers. Need more analysis about them.