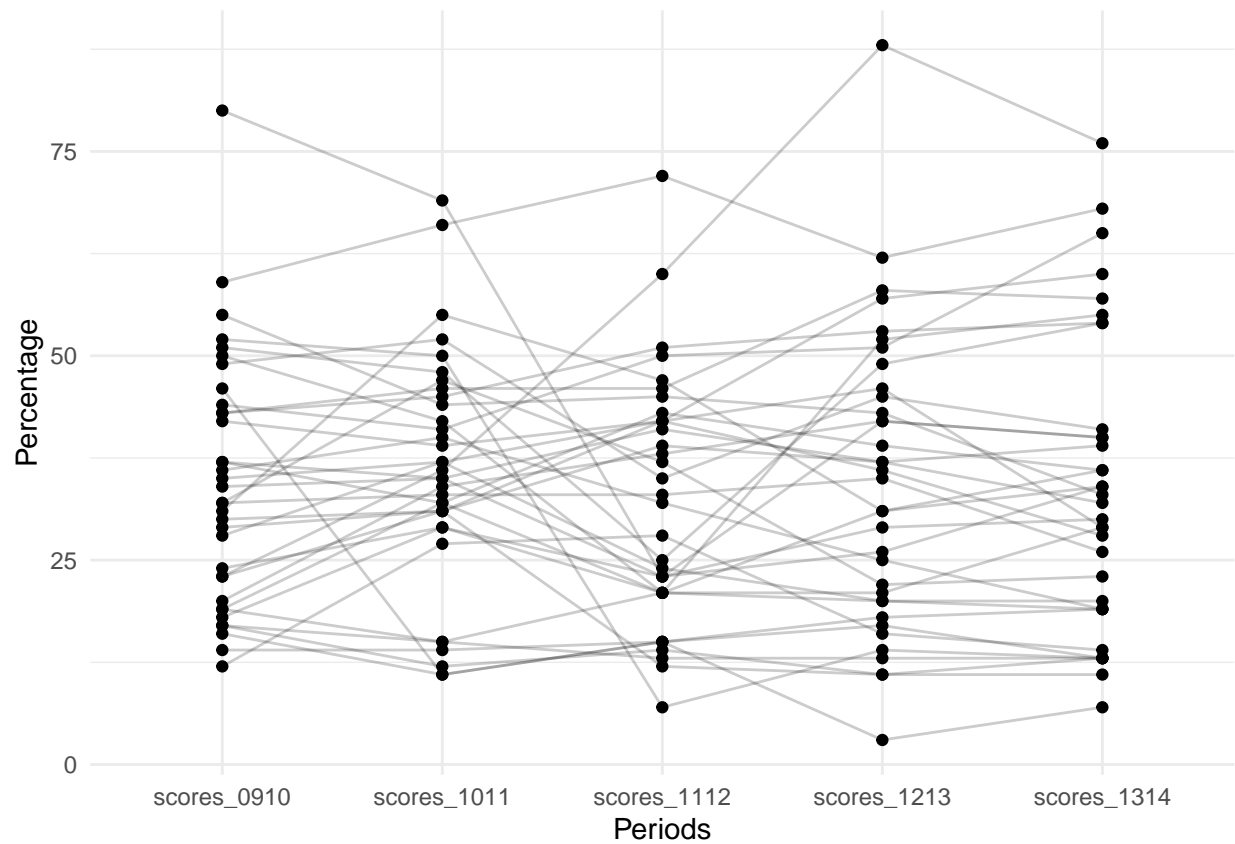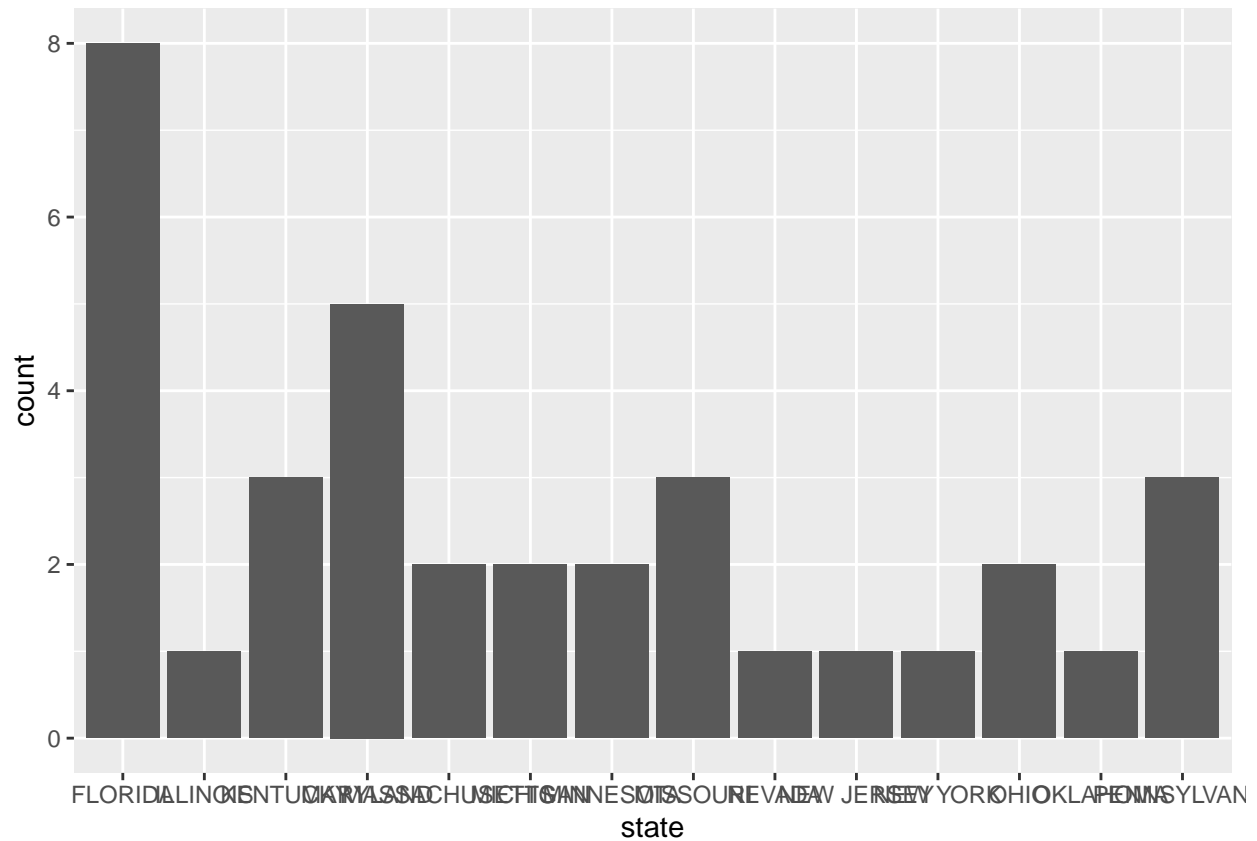# Education in USA

Ke-li Chiu, Sharon Allman & Diego Mamanche Castellanos

11/03/2020

TREATED GROUP BY COHORTS
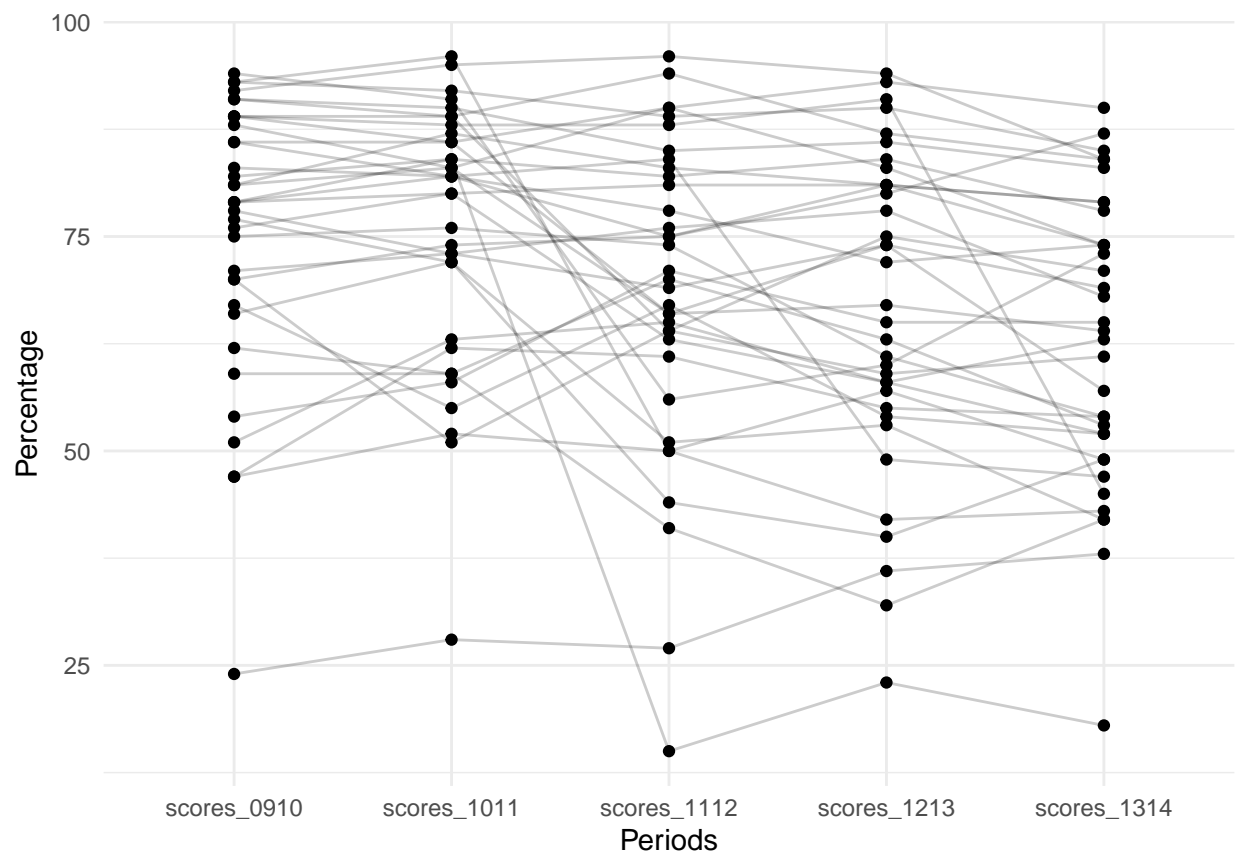
```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```
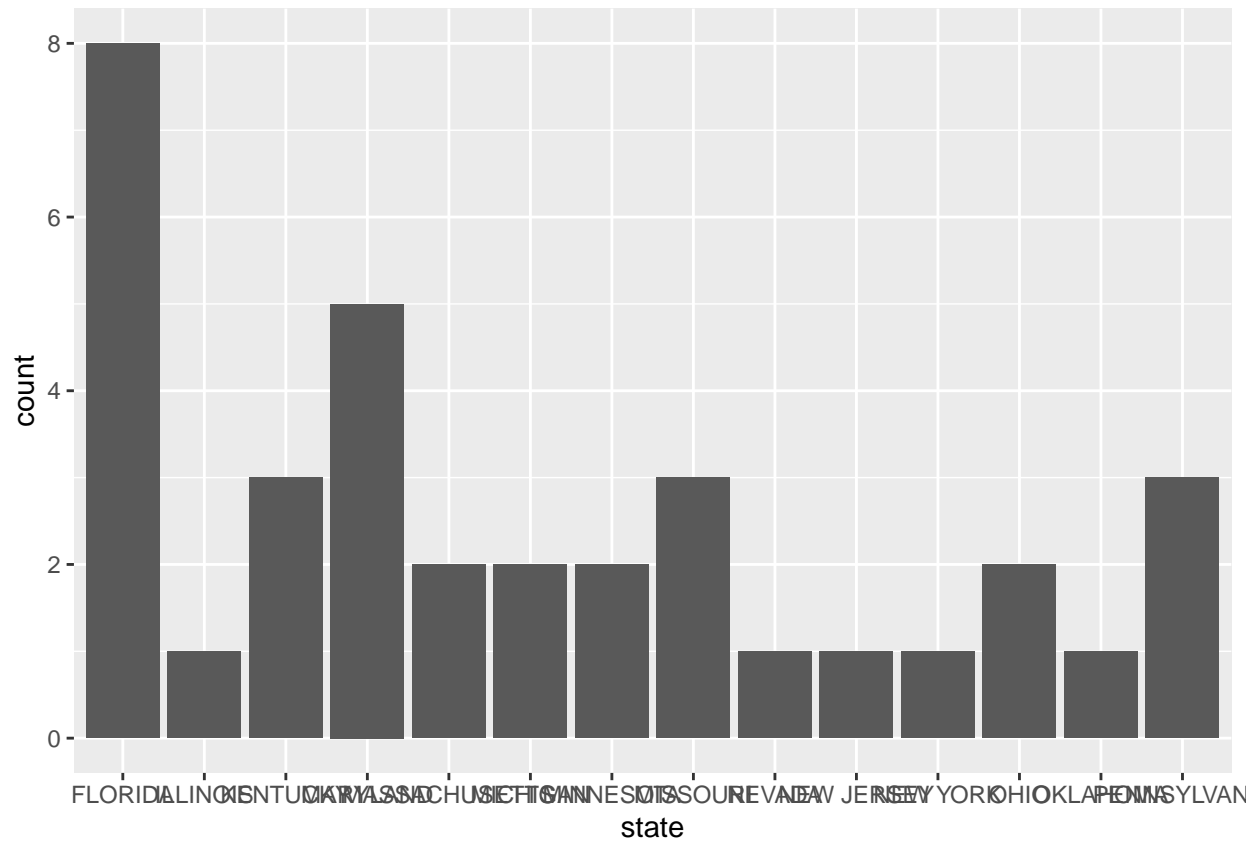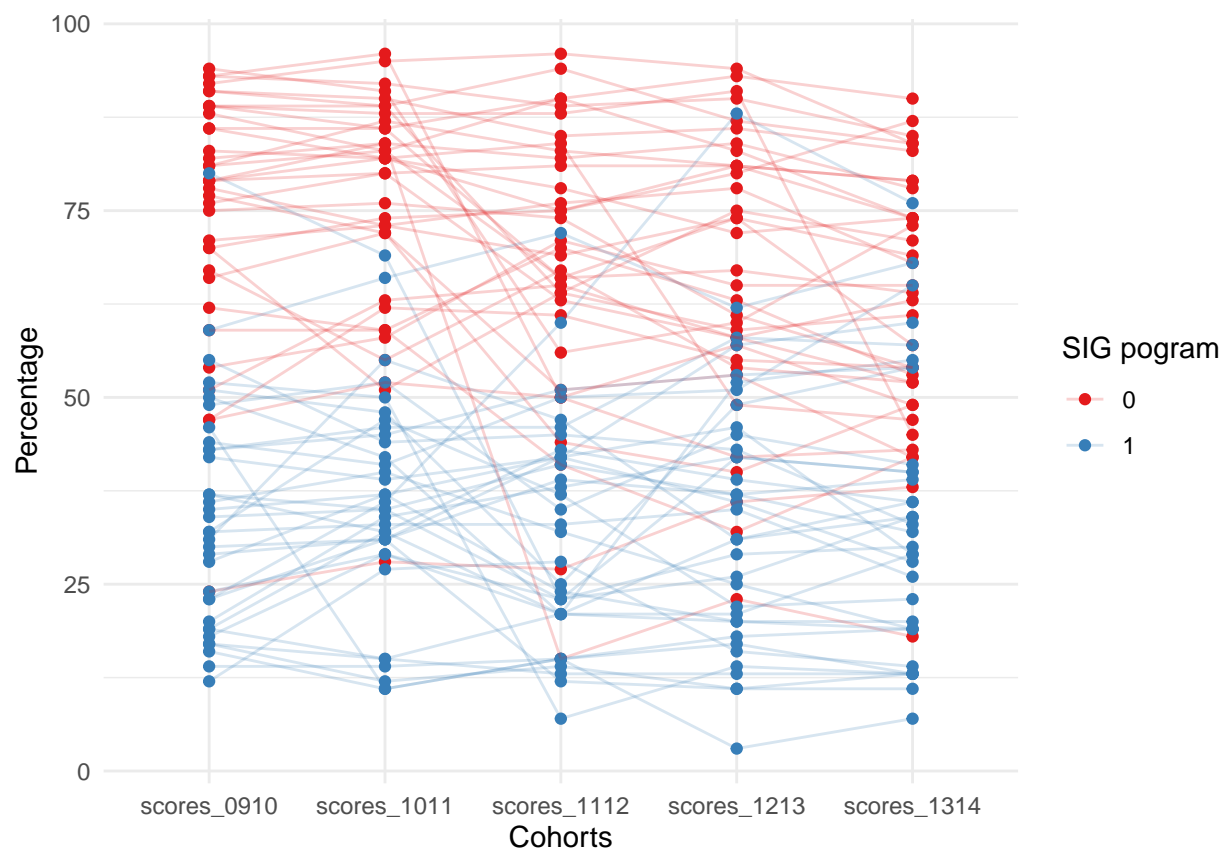
CONTROL GROUP BY COHORTS (Created randomly, keeping treated group distribution)

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

GRAPH OF TREATED AND CONTROLLED GROUPS BY CO-HORTS

APPLYING DIFF IN DIFF TO BY COHORTS

COHORT 2009-10 AND 2010-11

```
##
## Call:
## lm(formula = percentage ~ sig_program * cohort, data = cohort1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.400  -8.907   1.429  10.600  45.800
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          75.4000     2.5736  29.297   <2e-16 ***
## sig_program1        -41.2000     3.6397 -11.320   <2e-16 ***
## cohort1               0.4857     3.6397   0.133    0.894
## sig_program1:cohort1  1.0571     5.1473   0.205    0.838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.23 on 136 degrees of freedom
## Multiple R-squared:  0.6476, Adjusted R-squared:  0.6398
## F-statistic: 83.31 on 3 and 136 DF,  p-value: < 2.2e-16
```

COHORT 2010-11 AND 2011-12

```
## 
## Call:
## lm(formula = percentage ~ sig_program * cohort, data = cohort2)
## 
## Residuals:
##     Min     1Q  Median     3Q     Max 
## -53.571  -8.886   1.186  10.471  40.114 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            75.886      2.680  28.318   <2e-16 ***
## sig_program1          -40.143      3.790 -10.592   <2e-16 ***
## cohort1                -7.314      3.790  -1.930   0.0557 .  
## sig_program1:cohort1    3.457      5.360   0.645   0.5200    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 15.85 on 136 degrees of freedom
## Multiple R-squared:  0.6072, Adjusted R-squared:  0.5986 
## F-statistic: 70.08 on 3 and 136 DF,  p-value: < 2.2e-16
```

COHORT 2011-12 AND 2012-13

```
## 
## Call:
## lm(formula = percentage ~ sig_program * cohort, data = cohort3)
## 
## Residuals:
##     Min     1Q  Median     3Q     Max 
## -53.571 -11.879   0.571  13.300  53.714 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            68.571      2.977  23.035  < 2e-16 ***
## sig_program1          -36.686      4.210  -8.714 8.92e-15 ***
## cohort1                -1.829      4.210  -0.434    0.665    
## sig_program1:cohort1    4.229      5.954   0.710    0.479    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 17.61 on 136 degrees of freedom
## Multiple R-squared:  0.4989, Adjusted R-squared:  0.4878 
## F-statistic: 45.13 on 3 and 136 DF,  p-value: < 2.2e-16
```

COHORT 2012-13 AND 2013-14

```
## 
## Call:
## lm(formula = percentage ~ sig_program * cohort, data = cohort4)
## 
## Residuals:
```

```
##     Min      1Q Median     3Q     Max
## -44.86 -13.77   0.20  14.26  53.71
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           66.743      3.029  22.034  < 2e-16 ***
## sig_program1         -32.457      4.284  -7.577 4.92e-12 ***
## cohort1               -3.886      4.284  -0.907    0.366
## sig_program1:cohort1   3.314      6.058   0.547    0.585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.92 on 136 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4214
## F-statistic: 34.74 on 3 and 136 DF,  p-value: < 2.2e-16
```

Appendix A

```r
knitr::opts_chunk$set(echo = FALSE, include = FALSE)## R Markdown
#Set up the environment
library(readxl)
library(dplyr)
library(tidyr)
library(ggplot2)
library(reshape2)
library(stringr)
#library(groom)

setwd("~/Experimental Design for Data Science/ProblemSet4")

###Upload datasets


#https://www2.ed.gov/programs/sif/data/sy1011-1314.xlsx
treated_schools <- read_excel("sy1011-1314.xlsx")
treated_schools <- janitor::clean_names(treated_schools)


#https://www2.ed.gov/about/inits/ed/edfacts/data-files/math-achievement-sch-sy2009-10.csv
math_assessment_09_10 <- read.csv("math-achievement-sch-sy2009-10.csv")
math_assessment_09_10 <- janitor::clean_names(math_assessment_09_10)

#https://www2.ed.gov/about/inits/ed/edfacts/data-files/math-achievement-sch-sy2010-11.csv
math_assessment_10_11 <- read.csv("math-achievement-sch-sy2010-11.csv")
math_assessment_10_11 <- janitor::clean_names(math_assessment_10_11)

#https://www2.ed.gov/about/inits/ed/edfacts/data-files/math-achievement-sch-sy2011-12.csv
math_assessment_11_12 <- read.csv("math-achievement-sch-sy2011-12.csv")
math_assessment_11_12 <- janitor::clean_names(math_assessment_11_12)

#https://www2.ed.gov/about/inits/ed/edfacts/data-files/math-achievement-sch-sy2012-13.csv
math_assessment_12_13 <- read.csv("math-achievement-sch-sy2012-13.csv")
math_assessment_12_13 <- janitor::clean_names(math_assessment_12_13)

#https://www2.ed.gov/about/inits/ed/edfacts/data-files/math-achievement-sch-sy2013-14.csv
math_assessment_13_14 <- read.csv("math-achievement-sch-sy2013-14.csv")
math_assessment_13_14 <- janitor::clean_names(math_assessment_13_14)


head(treated_schools)

#

treated_all_cohorts <- filter(treated_schools,
                            sy201011sig_model != is.na(sy201011sig_model) &
                            sy201112sig_model != is.na(sy201112sig_model) &
                            sy201213sig_model != is.na(sy201213sig_model) &
                            sy201314sig_model != is.na(sy201314sig_model) &
                            ncessch_1011 == ncessch_1112 &
```

```r
                              ncessch_1011 == ncessch_1213 &
                              ncessch_1011 == ncessch_1314 &
                              ncessch_1112 == ncessch_1213 &
                              ncessch_1112 == ncessch_1314 &
                              ncessch_1213 == ncessch_1314 &
                              sy201011sig_model == "Turnaround" &
                              sy201112sig_model == "Turnaround" &
                              sy201213sig_model == "Turnaround" &
                              sy201314sig_model == "Turnaround"
                              )
treated_all_cohorts <- select(treated_all_cohorts, state, leaid_10_11,
                         leanm_1011, ncessch_1011, schnam_1011)

treated_all_cohorts$ncessch_1011 <- as.numeric(treated_all_cohorts$ncessch_1011)


colnames(treated_all_cohorts) <- c("state", "lea_id","lea_name","ncessch","school_name")


head(treated_all_cohorts)
treated_math_09_10 <- select(math_assessment_09_10, stnam, leaid,
                              leanm, ncessch, schnam09, all_mth00pctprof_0910)

treated_math_10_11 <- select(math_assessment_10_11, stnam, leaid,
                              leanm10, ncessch, schnam10, all_mth00pctprof_1011)

treated_math_11_12 <- select(math_assessment_11_12, stnam, leaid,
                              leanm, ncessch, schnam11, all_mth00pctprof_1112)

treated_math_12_13 <- select(math_assessment_12_13, stnam, leaid,
                              leanm, ncessch, schnam, all_mth00pctprof_1213)

treated_math_13_14 <- select(math_assessment_13_14, stnam, leaid,
                              leanm, ncessch, schnam, all_mth00pctprof_1314)

merged_math <- merge(treated_math_09_10, treated_math_10_11, by= "ncessch")
merged_math <- select(merged_math, ncessch, schnam09, stnam.x,
                      leaid.x, leanm, all_mth00pctprof_0910, all_mth00pctprof_1011)

merged_math <- merge(merged_math, treated_math_11_12, by= "ncessch")
merged_math <- select(merged_math, ncessch, schnam09, stnam.x,
                      leaid.x, leanm.x, all_mth00pctprof_0910, all_mth00pctprof_1011,
                      all_mth00pctprof_1112)

merged_math <- merge(merged_math, treated_math_12_13, by= "ncessch")
merged_math <- select(merged_math, ncessch, schnam09, stnam.x,
                      leaid.x, leanm.x, all_mth00pctprof_0910, all_mth00pctprof_1011,
                      all_mth00pctprof_1112, all_mth00pctprof_1213)

merged_math <- merge(merged_math, treated_math_13_14, by= "ncessch")
merged_math <- select(merged_math, ncessch, schnam09, stnam.x,
                      leaid.x, leanm.x, all_mth00pctprof_0910, all_mth00pctprof_1011,
                      all_mth00pctprof_1112, all_mth00pctprof_1213, all_mth00pctprof_1314)
```

```r
merged_math
### Create Treatment DataSet ###

results_treated <- merge(treated_all_cohorts, merged_math, by= "ncessch")

results_treated <- filter(results_treated,
                          !grepl("(([0-9]+[-][0-9]+)|([a-zA-Z]+[0-9]+)|([a-zA-z]+))",
                                 all_mth00pctprof_0910) )

results_treated <- filter(results_treated,
                          !grepl("(([0-9]+[-][0-9]+)|([a-zA-Z]+[0-9]+)|([a-zA-z]+))",
                                 all_mth00pctprof_1011) )

results_treated <- filter(results_treated,
                          !grepl("(([0-9]+[-][0-9]+)|([a-zA-Z]+[0-9]+)|([a-zA-z]+))",
                                 all_mth00pctprof_1112) )

results_treated <- filter(results_treated,
                          !grepl("(([0-9]+[-][0-9]+)|([a-zA-Z]+[0-9]+)|([a-zA-z]+))",
                                 all_mth00pctprof_1213) )

results_treated <- filter(results_treated,
                          !grepl("(([0-9]+[-][0-9]+)|([a-zA-Z]+[0-9]+)|([a-zA-z]+))",
                                 all_mth00pctprof_1314) )

results_treated <- select(results_treated, ncessch, schnam09, stnam.x,
                          leaid.x, leanm.x, all_mth00pctprof_0910,
                          all_mth00pctprof_1011, all_mth00pctprof_1112,
                          all_mth00pctprof_1213, all_mth00pctprof_1314)

colnames(results_treated) <- c("ncessch", "school_name", "state", "lea_id",
                               "lea_name", "scores_0910", "scores_1011",
                               "scores_1112", "scores_1213", "scores_1314")

results_treated

diff_in_diff_treated <- melt(results_treated, measure = 6:10)
diff_in_diff_treated[,7] <-
sapply(diff_in_diff_treated[,7], function(x) as.numeric(x))

colnames(diff_in_diff_treated) <- c("ncessch", "school_name", "state", "lea_id",
                                    "lea_name", "cohort","percentage")

diff_in_diff_treated %>%
  ggplot(aes(x = cohort,
             y = percentage
             #color = state
             )) +
  geom_point() +
  geom_line(aes(group = ncessch), alpha = 0.2) +
  theme_minimal() +
```

```r
  labs(x = "Periods",
       y = "Percentage",
       color = "Groups") +
  scale_color_brewer(palette = "Set1")

ggplot(results_treated, aes(x=state)) +
  geom_bar()

### Create Control Group ###

merged_math_cleaned <- merged_math

unique_treated_schools <- melt(treated_schools, measure = 10:13)

unique_treated_schools <- unique(unique_treated_schools$value)

merged_math_cleaned <- filter(merged_math, !(ncessch %in%
                              as.numeric(unique_treated_schools)))


merged_math_cleaned <- filter(merged_math_cleaned,
                        !grepl("(([0-9]+[-][0-9]+)|([a-zA-Z]+[0-9]+)|([a-zA-z]+))",
                            all_mth00pctprof_0910) )

merged_math_cleaned <- filter(merged_math_cleaned,
                        !grepl("(([0-9]+[-][0-9]+)|([a-zA-Z]+[0-9]+)|([a-zA-z]+))",
                            all_mth00pctprof_1011) )

merged_math_cleaned <- filter(merged_math_cleaned,
                        !grepl("(([0-9]+[-][0-9]+)|([a-zA-Z]+[0-9]+)|([a-zA-z]+))",
                            all_mth00pctprof_1112) )

merged_math_cleaned <- filter(merged_math_cleaned,
                        !grepl("(([0-9]+[-][0-9]+)|([a-zA-Z]+[0-9]+)|([a-zA-z]+))",
                            all_mth00pctprof_1213) )

merged_math_cleaned <- filter(merged_math_cleaned,
                        !grepl("(([0-9]+[-][0-9]+)|([a-zA-Z]+[0-9]+)|([a-zA-z]+))",
                            all_mth00pctprof_1314) )

results_control <- data.frame()

for (i in unique(results_treated$state)) {

    count_state <- count(filter(results_treated, state == i))

    filter_by_state <- which(merged_math_cleaned$stnam.x == i)

    set.seed(123)
    sample_by_state <- merged_math_cleaned[sample(filter_by_state,
                                          as.numeric(count_state)),]

    results_control <- rbind(results_control, sample_by_state)
```

```
    }

colnames(results_control) <- c("ncessch", "school_name", "state", "lea_id",
                               "lea_name", "scores_0910", "scores_1011",
                               "scores_1112", "scores_1213", "scores_1314")


diff_in_diff_controled <- melt(results_control, measure = 6:10)
diff_in_diff_controled[,7] <-
sapply(diff_in_diff_controled[,7], function(x) as.numeric(x))

colnames(diff_in_diff_controled) <- c("ncessch", "school_name", "state", "lea_id",
                                      "lea_name", "cohort","percentage")

diff_in_diff_controled %>%
  ggplot(aes(x = cohort,
             y = percentage
             #color = state
             )) +
  geom_point() +
  geom_line(aes(group = ncessch), alpha = 0.2) +
  theme_minimal() +
  labs(x = "Periods",
       y = "Percentage",
       color = "Groups") +
  scale_color_brewer(palette = "Set1")

ggplot(results_control, aes(x=state)) +
  geom_bar()

### Create diff in diff dataset with both treated and controlled groups ###

diff_in_diff_treated <- mutate(diff_in_diff_treated, sig_program = "1")
diff_in_diff_controled <- mutate(diff_in_diff_controled, sig_program = "0")

diff_in_diff_both <- rbind(diff_in_diff_controled, diff_in_diff_treated)

diff_in_diff_both %>%
  ggplot(aes(x = cohort,
             y = percentage,
             color = sig_program
             )) +
  geom_point() +
  geom_line(aes(group = ncessch), alpha = 0.2) +
  theme_minimal() +
  labs(x = "Cohorts",
       y = "Percentage",
       color = "SIG pogram") +
  scale_color_brewer(palette = "Set1")


cohort1 <- filter(diff_in_diff_both, cohort %in% c("scores_0910","scores_1011"))
```

```r
cohort1$cohort <- if_else(cohort1$cohort == "scores_0910", "0","1")

cohort1$sig_program <- as.factor(cohort1$sig_program)
cohort1$cohort <- as.factor(cohort1$cohort)

diff_in_diff_cohort1_model <- lm(percentage ~ sig_program*cohort,
                      data = cohort1)

summary(diff_in_diff_cohort1_model)


cohort2 <- filter(diff_in_diff_both, cohort %in% c("scores_1011","scores_1112"))
cohort2$cohort <- if_else(cohort2$cohort == "scores_1011", "0","1")

cohort2$sig_program <- as.factor(cohort2$sig_program)
cohort2$cohort <- as.factor(cohort2$cohort)

diff_in_diff_cohort2_model <- lm(percentage ~ sig_program*cohort,
                      data = cohort2)

summary(diff_in_diff_cohort2_model)


cohort3 <- filter(diff_in_diff_both, cohort %in% c("scores_1112","scores_1213"))
cohort3$cohort <- if_else(cohort3$cohort == "scores_1112", "0","1")

cohort3$sig_program <- as.factor(cohort3$sig_program)
cohort3$cohort <- as.factor(cohort3$cohort)

diff_in_diff_cohort3_model <- lm(percentage ~ sig_program*cohort,
                      data = cohort3)

summary(diff_in_diff_cohort3_model)


cohort4 <- filter(diff_in_diff_both, cohort %in% c("scores_1213","scores_1314"))
cohort4$cohort <- if_else(cohort4$cohort == "scores_1213", "0","1")

cohort4$sig_program <- as.factor(cohort4$sig_program)
cohort4$cohort <- as.factor(cohort4$cohort)

diff_in_diff_cohort4_model <- lm(percentage ~ sig_program*cohort,
                      data = cohort4)

summary(diff_in_diff_cohort4_model)
```