Solution 2

```
library(tidyverse)
library(miceadds)

df <- read.csv('https://raw.githubusercontent.com/rjblake34/waenroll/main/waenrolldf.csv', header=TRUE)
```

Run regressions

```
mod1 <- miceadds::glm.cluster(data=df,
                              formula = TwoYear ~ LowIncomePct + WhitePct + ELLPct + HighlyCapablePct +
                                TotalStudents,
                              cluster = "District",
                              family='gaussian')

mod2 <- miceadds::glm.cluster(data=df,
                              formula = FourYear ~ LowIncomePct + WhitePct + ELLPct + HighlyCapablePct
                                TotalStudents,
                              cluster = "District",
                              family='gaussian')

mod3 <- miceadds::glm.cluster(data=df,
                              formula = TwoYear+FourYear ~ LowIncomePct + WhitePct + ELLPct + HighlyCapa
                                TotalStudents,
                              cluster = "District",
                              family='gaussian')
```

Print results of model 1

```
summary(mod1$glm_res)
```

```
##
## Call:
## stats::glm(formula = formula, family = family, data = data, weights = wgt__)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.212050  -0.044995   0.000021   0.046856   0.264098
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.918e-01  9.011e-02   5.458 6.19e-08 ***
## LowIncomePct      5.421e-02  1.816e-02   2.984  0.00292 **
## WhitePct          9.693e-03  2.439e-02   0.397  0.69116
## ELLPct            6.579e-02  4.838e-02   1.360  0.17415
## HighlyCapablePct -2.516e-01  6.610e-02  -3.807  0.00015 ***
```

```
## FemalePct        -5.190e-01  1.764e-01  -2.942  0.00334 **
## TotalStudents     1.033e-07  4.048e-07   0.255  0.79856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.005852072)
##
##     Null deviance: 5.7568  on 926  degrees of freedom
## Residual deviance: 5.3839  on 920  degrees of freedom
## AIC: -2126
##
## Number of Fisher Scoring iterations: 2
```

```
with(summary(mod1$glm_res), 1 - deviance/null.deviance) #print r squared
```

```
## [1] 0.06477535
```

Print results of model 2

```
summary(mod2$glm_res)
```

```
##
## Call:
## stats::glm(formula = formula, family = family, data = data, weights = wgt__)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.38773  -0.06305  -0.00905   0.05898   0.31297
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.639e-01  1.114e-01   2.369   0.0181 *
## LowIncomePct     -4.787e-01  2.246e-02 -21.313  < 2e-16 ***
## WhitePct         -1.353e-01  3.016e-02  -4.486 8.19e-06 ***
## ELLPct            8.815e-02  5.982e-02   1.474   0.1409
## HighlyCapablePct  4.136e-01  8.173e-02   5.061 5.05e-07 ***
## FemalePct         7.037e-01  2.181e-01   3.226   0.0013 **
## TotalStudents     7.651e-07  5.006e-07   1.528   0.1268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.008947925)
##
##     Null deviance: 14.7365  on 926  degrees of freedom
## Residual deviance:  8.2321  on 920  degrees of freedom
## AIC: -1732.4
##
## Number of Fisher Scoring iterations: 2
```

```
with(summary(mod2$glm_res), 1 - deviance/null.deviance) #print r squared
```

```
## [1] 0.4413794
```

Print results of model 3

```
summary(mod3$glm_res)
```

```
##
## Call:
## stats::glm(formula = formula, family = family, data = data, weights = wgt__)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.41618  -0.04952  -0.00161   0.05434   0.33263
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.557e-01  9.923e-02   7.616 6.49e-14 ***
## LowIncomePct     -4.245e-01  2.000e-02 -21.222  < 2e-16 ***
## WhitePct         -1.256e-01  2.686e-02  -4.676 3.37e-06 ***
## ELLPct            1.539e-01  5.327e-02   2.890  0.00395 **
## HighlyCapablePct  1.620e-01  7.279e-02   2.225  0.02629 *
## FemalePct         1.847e-01  1.942e-01   0.951  0.34188
## TotalStudents     8.684e-07  4.458e-07   1.948  0.05172 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.007097036)
##
##     Null deviance: 11.0121  on 926  degrees of freedom
## Residual deviance:  6.5293  on 920  degrees of freedom
## AIC: -1947.2
##
## Number of Fisher Scoring iterations: 2
```

```
with(summary(mod3$glm_res), 1 - deviance/null.deviance) #print r squared
```

```
## [1] 0.4070797
```

This analysis relies on data from the Washington State Office of the Superintendent of Public Instruction (OSPI). The panel data is created by combining datasets on college enrollment and demographics at the school district level for each year from 2014 to 2019. The purpose of this analysis is estimate the effect of school district demographic characteristics on college enrollment–i.e., does the makeup of a school district predict how many students choose college after high school?

Three models are estimated in this report. Enrollment at two year, enrollment at four year, and enrollment at either are used as the dependent variables in each model, respectively. The independent variables are rate statistics for low income, whiteness, English language learners, highly capable, female, and district size. All variables are calculated at the district level.

Overall, the analysis suggests districts that send more students to college tend to be higher income, less white, and have more highly capable students. However, there are significant differences between enrollment level. Districts with high levels of two year enrollment tend to be much whiter with much higher populations of low income students than districts with high levels of four year enrollment. Districts with high levels of four year enrollment tend to have many more highly capable students.

A better use for this panel data is to serve as a control for regression analysis on individual-level microdata from ACS.

The next step for this analysis is to perform a cluster analysis to learn if there are any commonalities among school districts with high college go-on rates, and then to incorporate ACS microdata.