Final Project

Alan Qiao, Jacqueline Dinh

Read in the data

Introduction

Through our project, we hope to predict grade level proficiency of a school across all EOG subjects based on demographics of the school using predictors such as total number of female, male, American Indian, Asian, Black, Hispanic, two or more races, white, economically disadvantaged, limited English proficient, and academically or intellectually gifted students, along with the total number of students and the number of students with disabilities. We intend to create a linear regression model and will calculate the predictors by manipulating the dataset.

Our EDA will include a plot of percentages of each demographic that is proficient to illustrate disparities and box plots categorized by school size.

Our hypothesis is that schools with higher percentages of white students will, on average, have a greater percentage of students that are grade level proficient. In addition, we hypothesize that a larger school will, on average, have a greater variability in students that are grade level proficient.

```
33.33333% 66.66667%
 29.98547  61.69422
```

no american indian

no AIG

```
Call:
lm(formula = all ~ female_dem + male_dem + asian_dem + black_dem +
    hispanic_dem + white_dem + EDS_dem + LEP_dem + SWD_dem +
    AIG_dem + white_comp, data = nc_school)
```

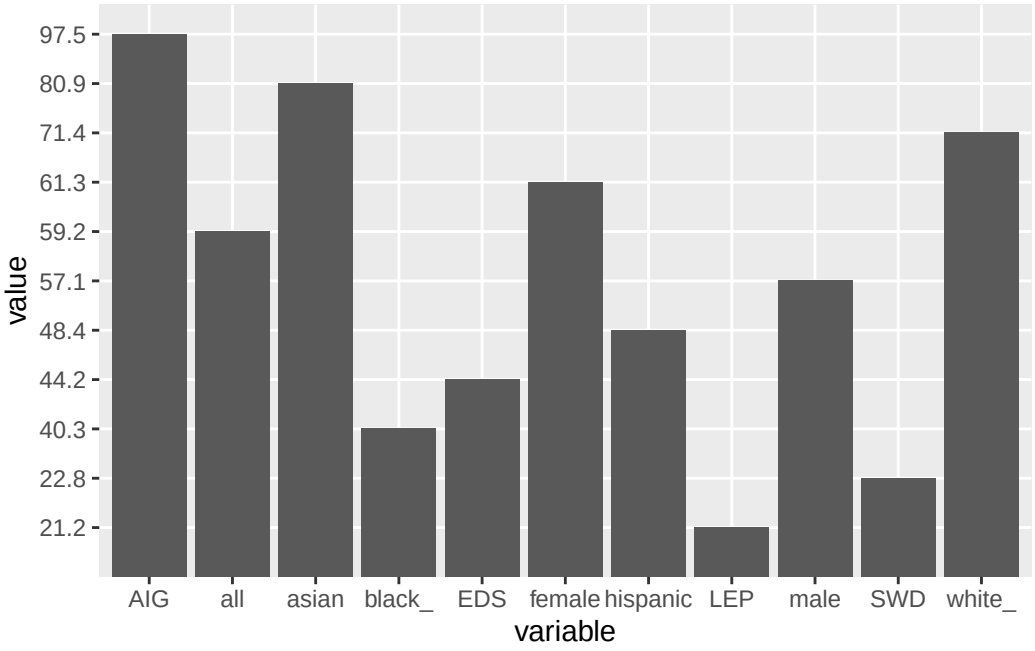
```
Residuals:
     Min       1Q   Median       3Q      Max
-19.9618  -3.8129  -0.0862   3.8170  24.1717

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               251.387096  47.776804   5.262 1.78e-07 ***
female_dem                 -1.797082   0.489232  -3.673 0.000253 ***
male_dem                   -1.963939   0.463676  -4.236 2.51e-05 ***
asian_dem                   0.192906   0.054991   3.508 0.000474 ***
black_dem                  -0.000600   0.033876  -0.018 0.985874
hispanic_dem                0.120169   0.040113   2.996 0.002812 **
white_dem                   0.169690   0.041335   4.105 4.40e-05 ***
EDS_dem                    -0.244715   0.015923 -15.369  < 2e-16 ***
LEP_dem                    -0.053874   0.010088  -5.340 1.17e-07 ***
SWD_dem                    -0.080386   0.008947  -8.985  < 2e-16 ***
AIG_dem                     0.506832   0.033728  15.027  < 2e-16 ***
white_compbmid white        2.186893   0.877580   2.492 0.012881 *
white_compmajority white   -0.185998   1.450784  -0.128 0.898015
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.081 on 912 degrees of freedom
  (1753 observations deleted due to missingness)
Multiple R-squared:  0.8208,    Adjusted R-squared:  0.8184
F-statistic: 348.1 on 12 and 912 DF,  p-value: < 2.2e-16
```

```r
nc_school1 <- nc_school1 |>
  filter(variable %in% c("all", "female", "male", "asian", "black_", "hispanic", "white_",

ggplot(nc_school1, aes(x = variable, y = value)) +
  geom_bar(stat = "identity")
```


97.5
80.9
71.4
61.3
59.2
57.1
48.4
44.2
40.3
22.8
21.2
value
AIG
all
asian
black_
EDS
female
hispanic
LEP
male
SWD
white_
variable