Final report STA207

Haojian Li

3/16/2025

Introduction of the dataset

The Tennessee Student/Teacher Achievement Ratio study (known as Project STAR) was conducted in the late 1980s to assess the impact of class size on test scores. This dataset has served as a classic example in numerous textbooks and research papers. The Project STAR public access data set includes test scores, treatment groups, and student and teacher characteristics from the experiment's four years, from 1985-1986 to 1988-1989. The test scores analyzed in this chapter are the sum of the Stanford Achievement Test's math and reading portions.

In this study, we focused on whether the experiment of class size effect the math scaled scores of the first grade. The research was consist of three parts: Descriptive analysis, inferential anslysis and sensitivity analysis.

Background

Project STAR (Student/Teacher Achievement Ratio) conducted in 1985 as a educational experiment tracking students from kindergarten to third grade in 79 Tennessee schools. The study random assigned to place both students and teachers in three different classroom environments: small classes (13-17 students), regular classes (22-25 students), and regular classes with a teaching assistant (22-25 students). Researchers aimed to evaluate the long-term effects of class size on educational outcomes by consistently implementing the class size categories for four years.

This study tested first-year math scores from the Project STAR dataset to figure out how class size predict academic achievement. The data was avaliable to the public from Harvard Dataverse. The primary hypothesis suggesting that different classroom settings significantly affect first-grade math grade. Further analysis will identify which classroom size have continuous effect on students grade in the following years

In this study, the classroom serves as the central unit of analysis, with various factors were used to assess their collective impact on math grade. This study aims to establish direct Project STAR (Student/Teacher Achievement Ratio) conducted in 1985 as a educational experiment tracking students from kindergarten to third grade in 79 Tennessee schools. The study random assigned to place both students and teachers in three different classroom environments: small classes (13-17 students), regular classes (22-25 students), and regular classes with a teaching assistant (22-25 students). Researchers aimed to evaluate the long-term effects of class size on educational outcomes by consistently implementing the class size categories for four years.

This study tested first-year math scores from the Project STAR dataset to figure out how class size predict academic achievement. The data was avaliable to the public from Harvard Dataverse. The primary hypothesis suggesting that different classroom settings significantly affect first-grade math grade. Further analysis will identify which classroom size have continuous effect on students grade in the following years

In this study, the classroom serves as the central unit of analysis, with various factors were used to assess their collective impact on math grade. This study aims to establish direct causal relationships by capturing demographic variables and other confoundng variables that may influence student performance. The final purpose is to provide evidence-based insights for educational policymakers and practitioners about the influence of class size, classroom composition, and school environment on academic performance of different students, supporting related law and descision-making process.

Experimental design

The STAR Project studied whether class size significantly impacts educational outcomes. Students were randomly assigned to three conditions: small classes (13-17 students), regular classes (22-25 students), or regular classes with a teaching aide (22-25 students). This randomization aimed to reduce selection bias by building direct relationship between experimental outcome and class size. The study also kept similar class size from kindergarten through third grade to ensure the continuity of the study. STAR collected comprehensive data including academic performance outcome and non-academic factors like attendance and engagement, while also considering demographic variables, teacher patterns, and school contexts for potential confounding factors.

Some major flows appears in this experimental design. One is Project STAR used discrete class size categories rather than considering treating class size as continuous variable. This categorical approach creates strange artificial boundaries and fails to capture potential effects in the middle of excluded range (17-22 students). And the designer of this experiment should explain more about the justification for why 13-17 students can consider as small class but not setting the class size even more smaller or higher, is there evidence proving that? Another flow is the study's longitudinal design, while very interesting, suffered from student mobility issues. The mobility issue of participants might introduced systematic biases since students who remained in the study for its duration may differ fundamentally from those who left, potentially skewing results.

The study's randomization also raised questions. Complete randomization of students didn't consider the demographic patterns of the area, which reduced the external validity and generalizability.

In conclusion, while Project STAR project builds direct causal relationship between class size and study outocmes, its experimental design have limitations and its results should be interpreted in caution.

Descriptive analysis

Because we are using the full version of STAR dataset this time, we need to redo the descriptive analysis

This descriptive statistics include the categorical variables of interest, mainly gender, race, class type, whether have free lunch and the location of school.

Also with the dependent varibale, I chose the total grade of first year math and first year reading because I think this two grades can well-reflect the study ability of the student.

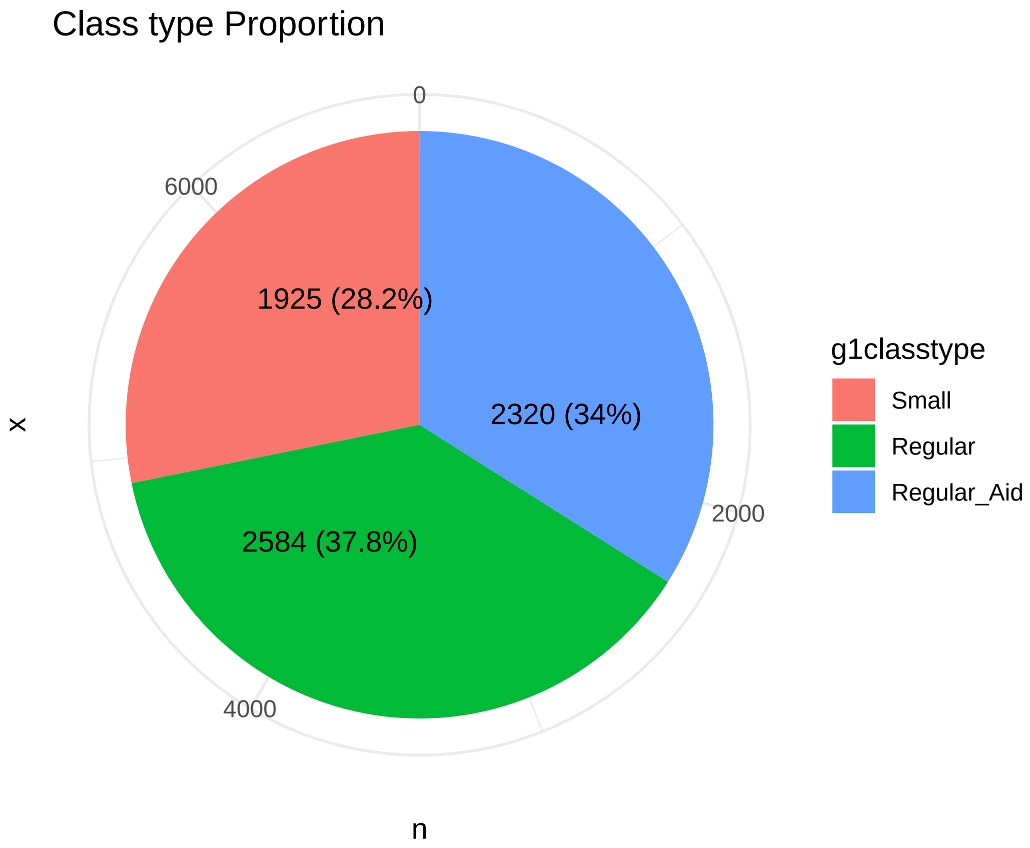
```
##     gender                  race          g1schid    
##  Male  :3541   White          :4528   169229 : 238  
##  Female:3275   Black          :2221   201449 : 149  
##  NA's  :  13   Asian          :  22   244806 : 142  
##                Hispanic       :   9   257905 : 134  
##                Native American:   9   244708 : 127  
##                Other          :  11   244755 : 127  
##                NA's           :  29   (Other):5912  
```

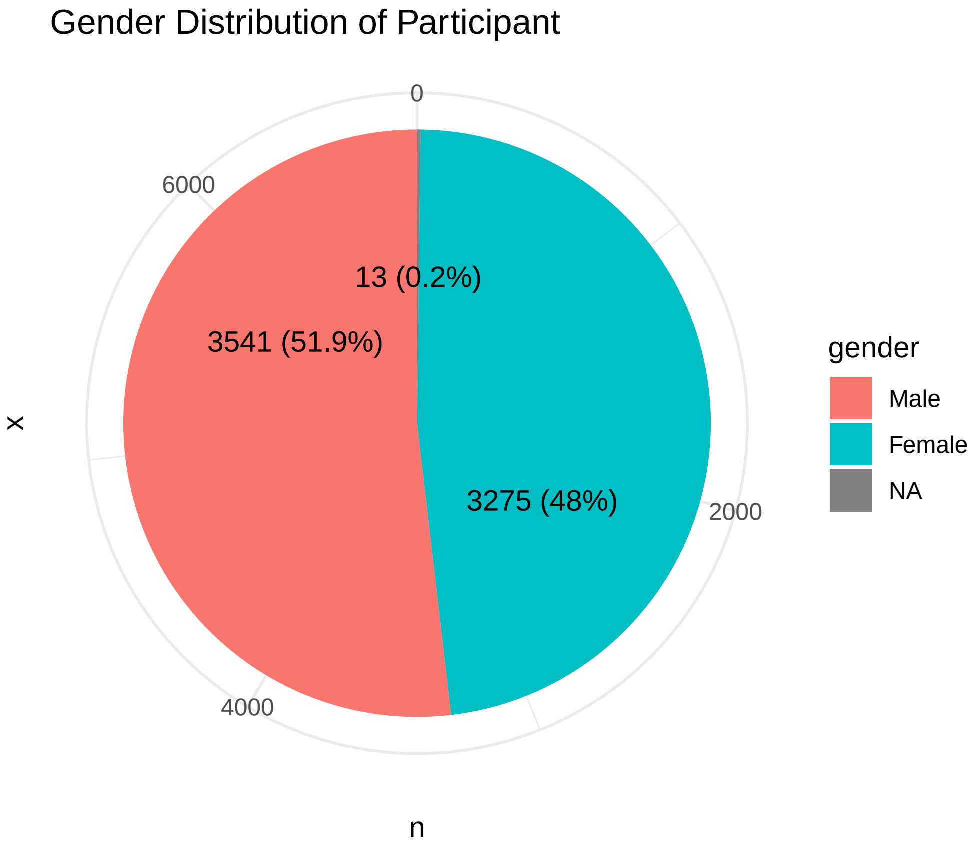
```
##       g1classtype           g1freelunch         g1surban
##  Small      :1925   Free-lunch    :3429   Inner City:1380
##  Regular    :2584   Non-free lunch:3221   Suburban  :1586
##  Regular_Aid:2320   NA's          : 179   Rural     :3237
##                                           Urban     : 626
```

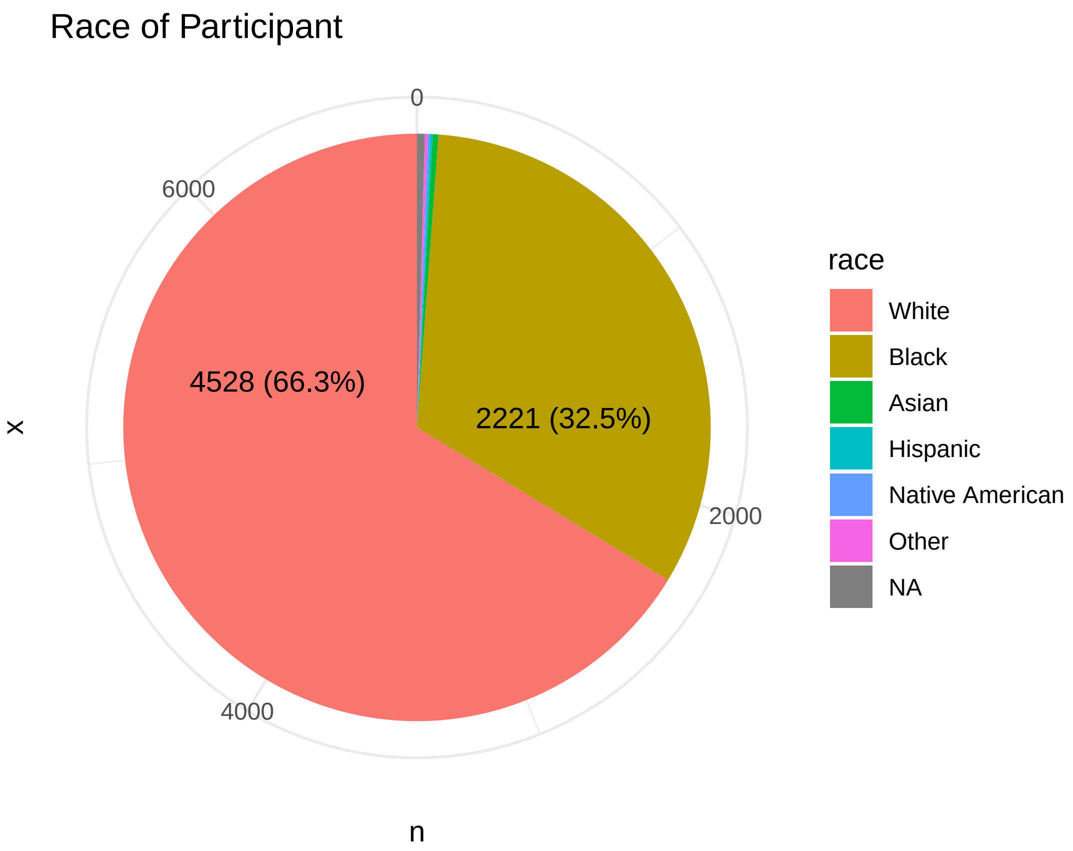
```
## # A tibble: 1 x 4
##   meang1treadss sdg1treadss meang1tmathss sdg1tmathss
##           <dbl>       <dbl>         <dbl>       <dbl>
## 1          521.        55.2          531.        43.0
```

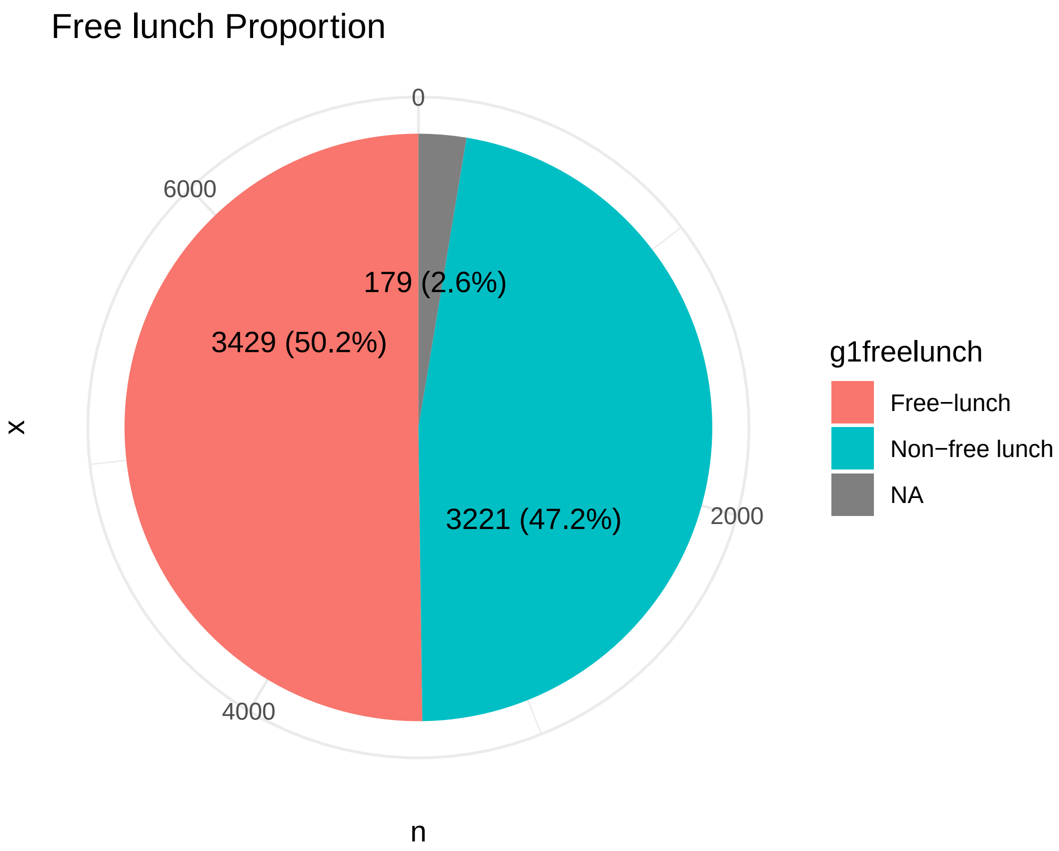
As the primary question of interest in this study is whether there are any differences in math scaled scores in 1st grade across class types (class size and school differences), our variables of interest would be g1classtype, which represented the class type of 1st year students; and our dependent variable would be g1tmathss, which represented the math grade of first year students.

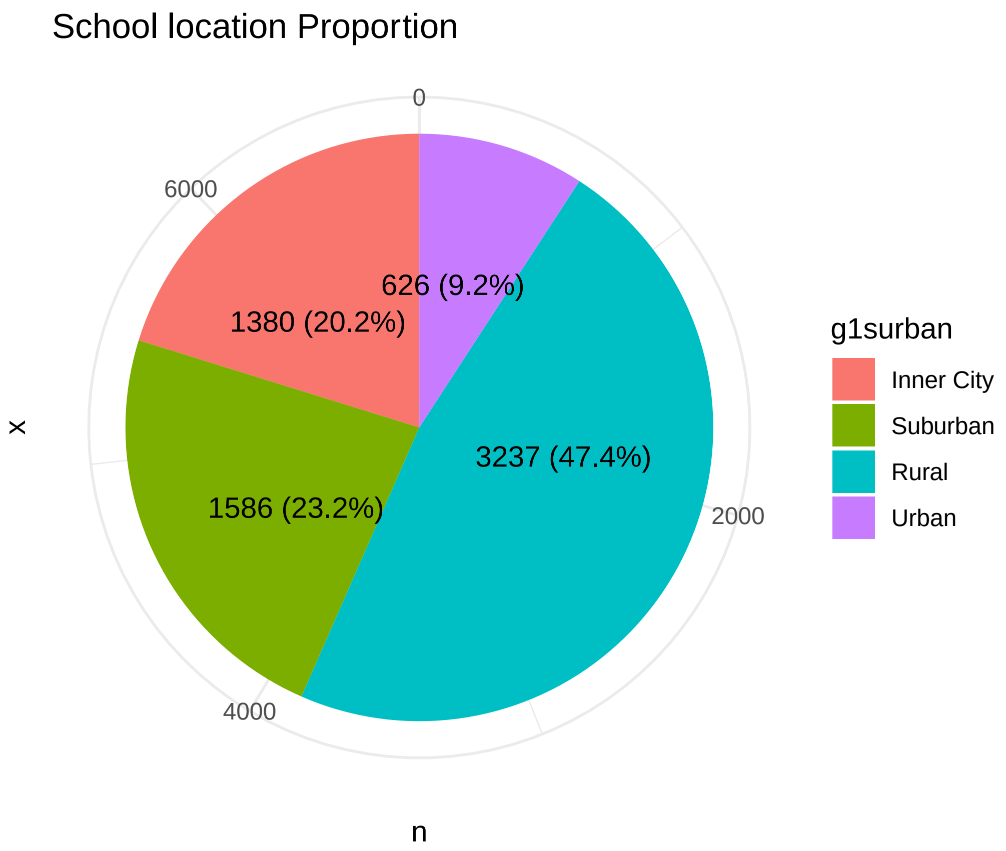
Here's graphical representation of the categorical variables of interest.



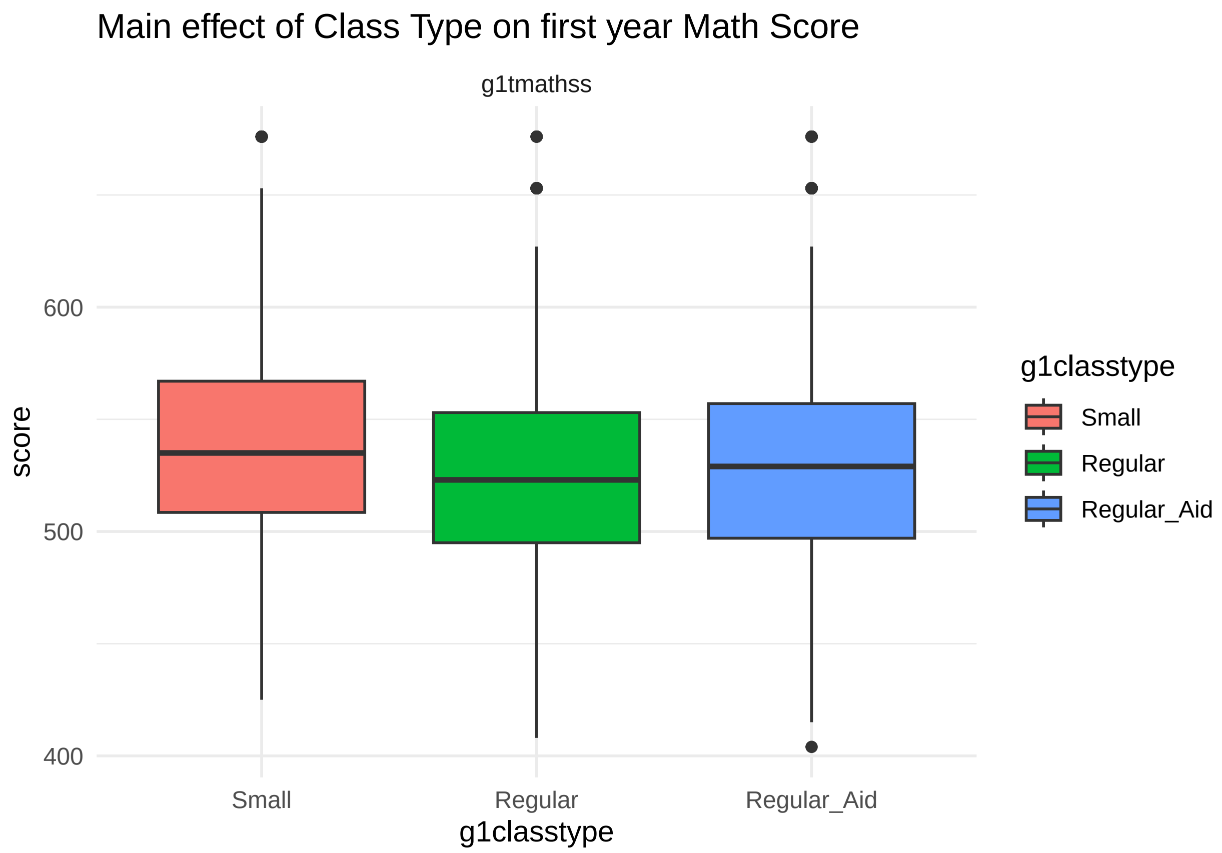

Gender Distribution of Participant
0
6000
2000
4000
13 (0.2%)
3541 (51.9%)
3275 (48%)
x
n
gender
Male
Female
NA

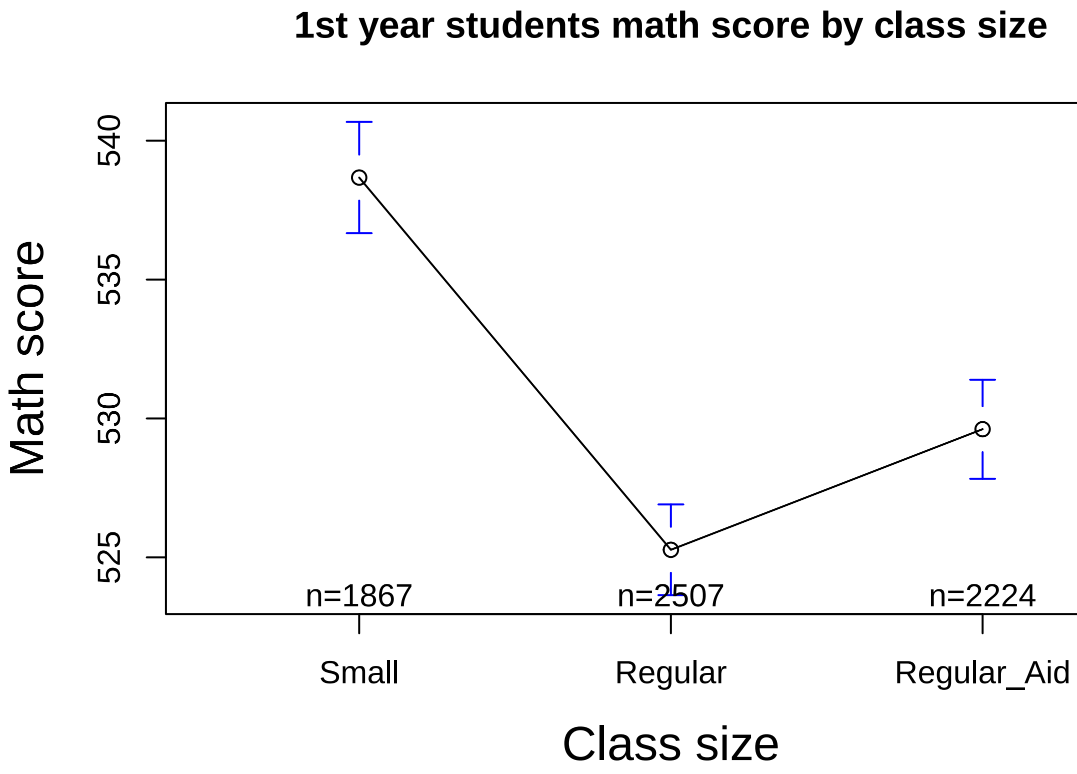

Race of Participant
0
6000
2000
4000
4528 (66.3%)
2221 (32.5%)
x
n
race
White
Black
Asian
Hispanic
Native American
Other
NA


Free lunch Proportion
0
6000
2000
4000
x
n
179 (2.6%)
3429 (50.2%)
3221 (47.2%)
g1freelunch
Free−lunch
Non−free lunch
NA



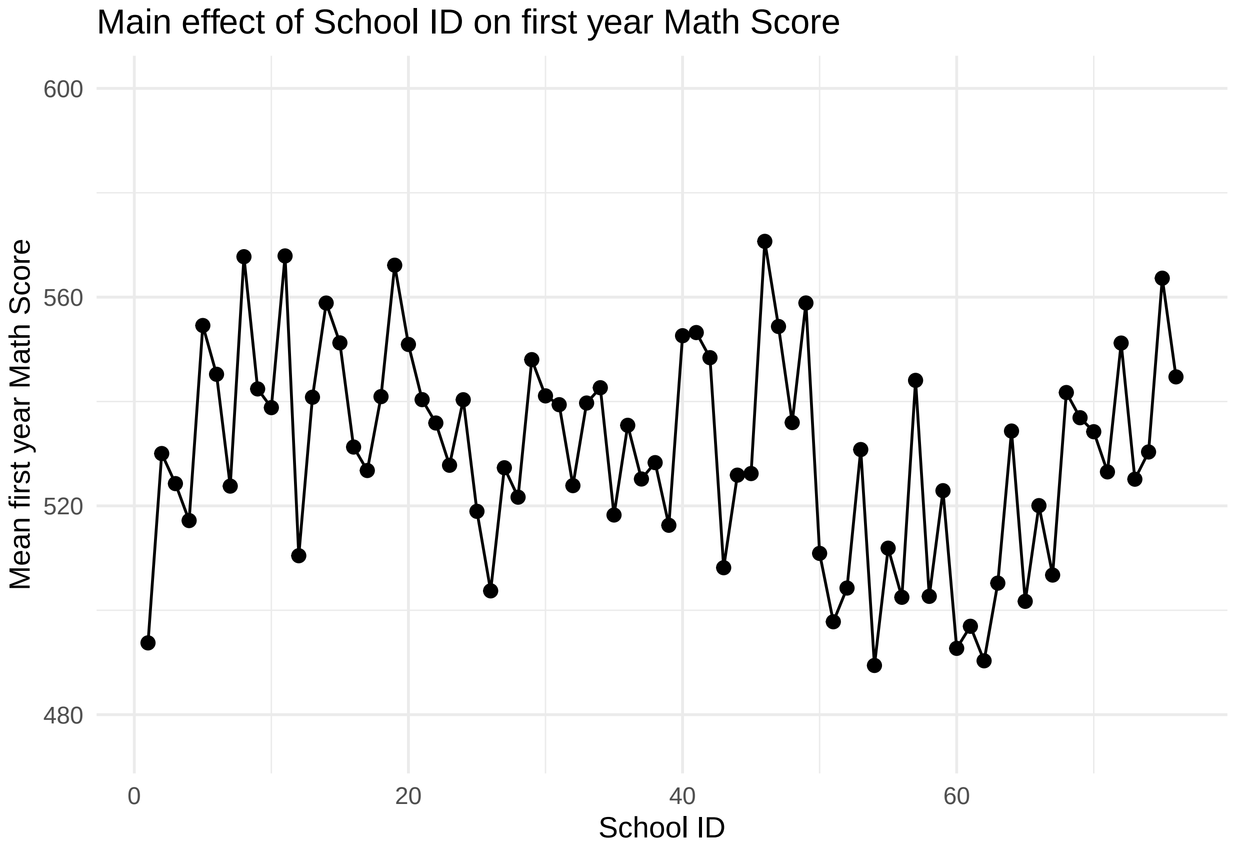
As we are interested in how the class type influencing the first year math grade, we conducted a main effect test to discover how first year math grade changed by class size


Main effect of Class Type on first year Math Score
g1tmathss
score
600
500
400
Small
Regular
Regular_Aid
g1classtype
g1classtype
Small
Regular
Regular_Aid



As school id is also another concerning variables in our preliminary analysis report, we also need to take a look of how first year math grade changed by school id.

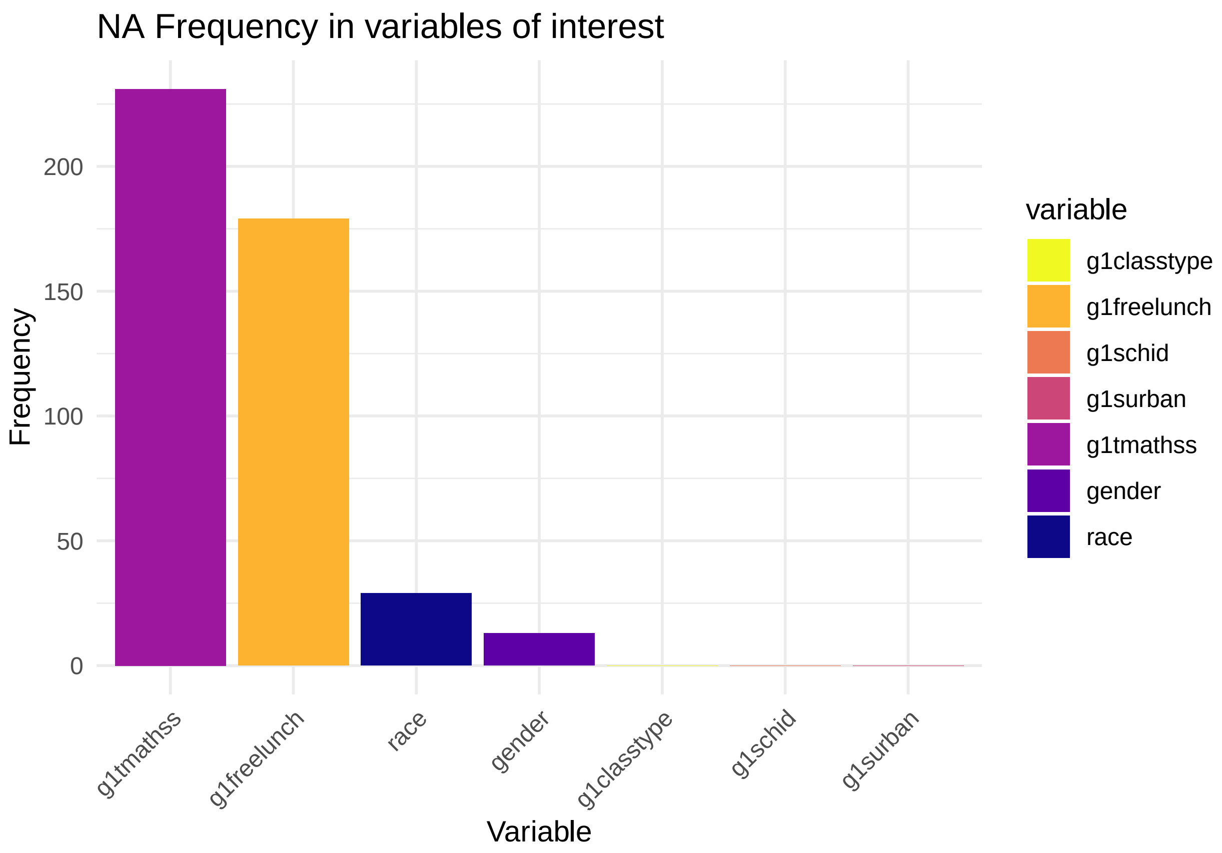
Here's the graph of math grade vs school_id. From the graph we can see that the differences on school have huge impact on the STA math score.



Missing value handeling

One important caveats we haven't take into through consideration is about handling the missing values. Previously, we simply getting rid of all the missing value, which is problematic because we didn't take into consideration of the patterns within the missing value. Thus, we need to take more careful look of how to handle the missing value.

To look into missing value, we first need to know where are our missing values coming from. Currently we have several categorical variable of interest, which includes gender, race, school location, class type, free lunch and one dependent continuous variable of interest total first year math score. Here's the proportion of NA for these several values:



We can see that for all the categorical variables except free lunch, the number of NA is extremely small (13 for gender and 29 for race) comparing the total sample size of 6829. So it would be convenient to just delete the rows with missing value for race and gender. For the free lunch, NA numbers are a little higher (around 179), but still can consider as extremely small comparing with the total sample size of 6289. More importantly, as it is extremely hard to predict the missing value statistically for categorical variables, it would be better to just cutting off the rows with missing values for the categorical variables.

After cleaning the NA values of categorical variables, we figure out that we still have 189 NA values for the math first year scores, which counts for around 3% of the total sample. To handle these NA values, a common practice to replace NA with the median, called median imputation.

```
## # A tibble: 1 x 1
##   g1tmathss
##       <int>
## 1         0
```

Inferential analysis

The primary question of interest in this study is whether there is any differences in math scaled scores in 1st grade across class types (class size and school id), and if so, a secondary question of interest is which class type is associated with the highest math scaled scores in 1st grade. To address these two problems, a two-way ANOVA model in additive form were fitted as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

Parameters:

- i: levels of class type, fixed effect, $i = 1, 2, 3$
- j: levels of school ID, random effect, $j = 1, 2, \ldots, 76$
- k: representing number of subjects for each class type within each school, $k = 1, 2, \ldots, n_{ij}$
- Y_{ijk}: first year math score of the kth student in the ith class type within the jth school
- μ: Overall average math score across all class types and schools

Constraints:

- α_i: The fixed effect of the ith class type on the math score,

$$\sum_{i=1}^{I} \alpha_i = 0$$

- β_j: The random effect of the jth school on the math score, capturing variations due to schools.

$$\beta_j \sim N(0, \sigma_\beta^2)$$

- ϵ_{ijk}: Random error term,

$$\epsilon_{ijk} \sim N(0, \sigma^2)$$

Assumptions and model hypothesis

To fit a two-way anova model, several general assumptions must be made:

Independent of observations

$$\epsilon_{ijk} \text{ are independent for all } i, j, k.$$

Homoscedasticity

$$\text{Var}(\epsilon_{ijk}) = \sigma^2 \quad \forall \quad i, j, k.$$

Normality of residuals

$$\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2).$$

No interaction term, but we are going to test whether we really have a interaction term or not.

$$H_0 : (\alpha\beta)_{ij} = 0 \quad \forall i, j.$$

No multicollinearity

$$\text{Cor}(\alpha_i, \beta_j) \approx 0.$$

The hypothesis of the F-test listed here. Alpha represented the main effect from class size, beta represented the main effect from school indicator, alpha-beta represented the interaction effect between class size and school indicator.

For the main effect of class type:

$$H_0 : \alpha_i = 0 \ \forall i \qquad H_1 : \text{not all } \alpha_i \text{ are zero.}$$

For the main effect of school indicator:

$$H_0 : \beta_j = 0 \; \forall j \qquad H_1 : \text{not all } \beta_j \text{ are zero.}$$

For now we haven't discuss the possibility of having an interaction term, so in case the interaction term really exist, we need to write down the hypothesis here:

$$H_0 : (\alpha\beta)_{ij} = 0 \; \forall i, j \quad v.s. \quad H_1 : \text{not all } (\alpha\beta)_{ij} \text{ are zero.}$$

To figure out whether the interaction effect exist, this study compared between reduced model and full model:

Reduced model: $Y_{ijk} = \mu. + \alpha_i + \beta_j + \epsilon_{ijk}$ Full model: $Y_{ijk} = \mu. + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$

The significant level of all below tests is

$$p < .05$$

.

ANOVA Test results:

```
## Analysis of Variance Table
##
## Response: median
##              Df Sum Sq Mean Sq F value    Pr(>F)
## g1schid      75 135952  1812.7  6.0858 < 2.2e-16 ***
## g1classtype   2  11071  5535.6 18.5848  2.87e-08 ***
## Residuals   259  77145   297.9
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## Analysis of Variance Table
##
## Response: median
##                      Df Sum Sq Mean Sq F value    Pr(>F)
## g1schid              75 135952  1812.7  6.1173 < 2.2e-16 ***
## g1classtype           2  11071  5535.6 18.6808 9.609e-08 ***
## g1schid:g1classtype 145  43364   299.1  1.0092    0.4822
## Residuals           114  33781   296.3
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## Analysis of Variance Table
##
## Model 1: median ~ g1schid + g1classtype
## Model 2: median ~ g1schid + g1classtype + g1classtype * g1schid
##   Res.Df   RSS  Df Sum of Sq      F Pr(>F)
## 1    259 77145
## 2    114 33781 145     43364 1.0092 0.4822
```

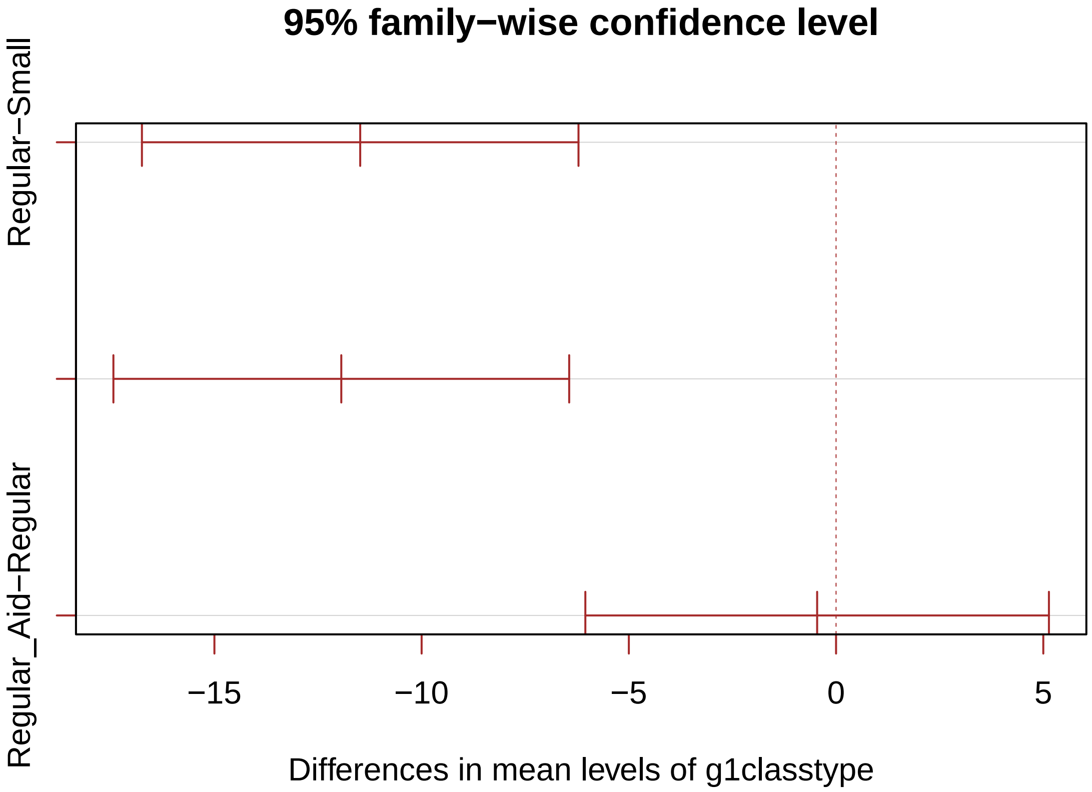
The reduced model, which included only the main effects, showed a significant main effect of class size, with an F(2, 259)= 18.58, p<.001. Similarly, the main effect of school id was significant, with an F(75, 259)=6.09, p< .001. These results suggested that both class size and school id independently influence math1 scores.

The full model included an interaction term (class size × school id) in addition to the main effects. The results indicated that the main effects remained significant. Class size had an F(2,114)= 18.68, p<.001. School id had an F(75,114)=6.12, p< .001. Importantly, the interaction between class size and school id was not significant, F(145,114)=1.01,p >.05. This indicates that the effect of class size on 1st year math grade scores was not depends on school id, which means the interaction term was not exist.

To verify whether the interaction term really not significant, a F-test between the full model and reduced model was conducted. The F test results showed that adding interaction term not significantly improved the model, F(145, 114)= 1.01,p>.05.H0 for the interaction effect was not rejected.

In conclusion, class size and school id have significant influence to the math score of the first year students.

To test which class type was associated with the highest math scaled scores in 1st grade, Tukey's HSD test was applied here.

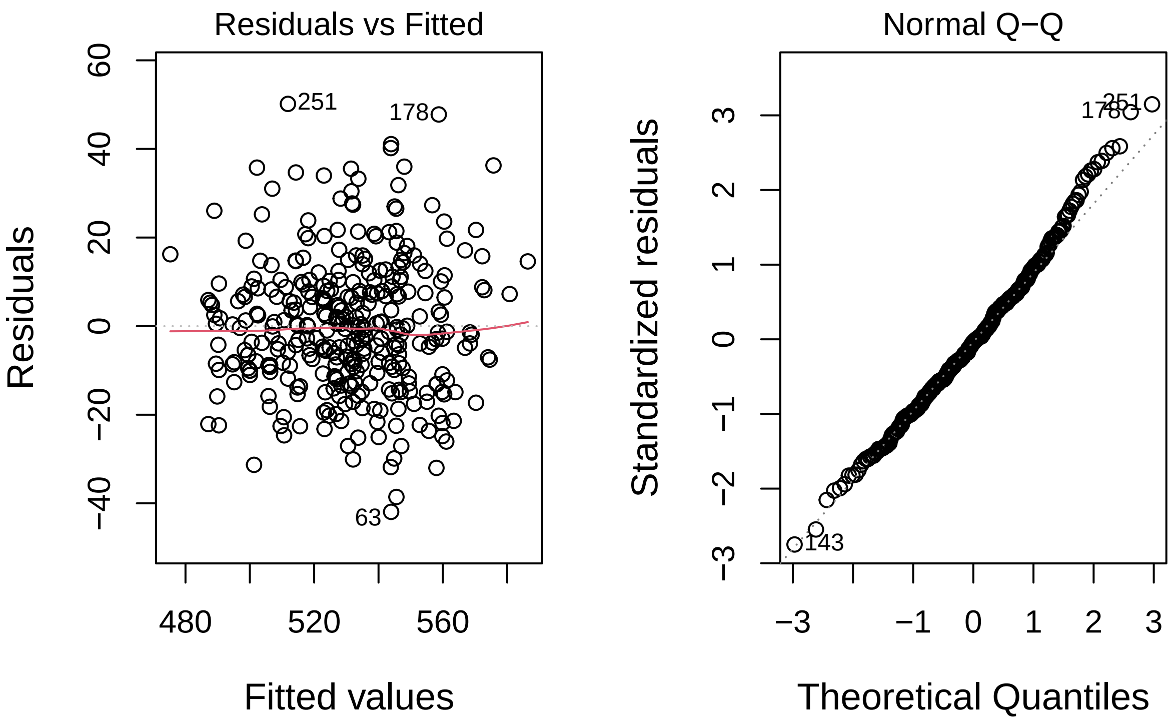


```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = median ~ g1schid + g1classtype, data = summary_data)
##
## $g1classtype
##                              diff        lwr       upr     p adj
## Regular-Small       -11.4818478 -16.748715 -6.214981 0.0000016
## Regular_Aid-Small   -11.9385447 -17.437317 -6.439772 0.0000018
## Regular_Aid-Regular  -0.4566969  -6.049648  5.136254 0.9797835
```

A Tukey's HSD test was conducted to examine pairwise differences between levels of class size on 1st year math scores of students following a significant main effect in the ANOVA. The results indicated a significant difference between the grade of small and regular class size, with regular class significantly lower than small class. mean difference of 11.48, a 95% confidence interval ranging from -16.75 to -6.21, and an adjusted p-value of < .001. Also, there're a significant differences between the regular+aide with small class size, which regular aid class significantly lower than small class, with a mean difference of -11.94, a 95% confidence interval ranging from -17.43 to -6.44, and an adjusted p-value of < .001.The regular aid group did not have significant differences with the regular group.

Since the confidence interval did not include zero and the p-value was significant, this result suggested that students in the small size class had significantly higher math scores compared to those in the regular group or regular group with aide. In conclusion, smaller classes yielding better math outcomes for 1st year students.

To assess whether the model violates the fundamental assumptions of ANOVA, this study examined the residuals vs. fitted plot and the Q-Q plot. The residuals vs. fitted plot was used to evaluate homoscedasticity. If residuals were randomly scattered without a discernible pattern, homoscedasticity assumption was considered satisfied. From the plot, we observed the residuals does not have structured pattern. The Q-Q plot was used to assess the normality of residuals. In the analysis, the plot indicated a slight skewness, but it remained within an acceptable range, suggesting that the normality assumption was approximately satisfied.

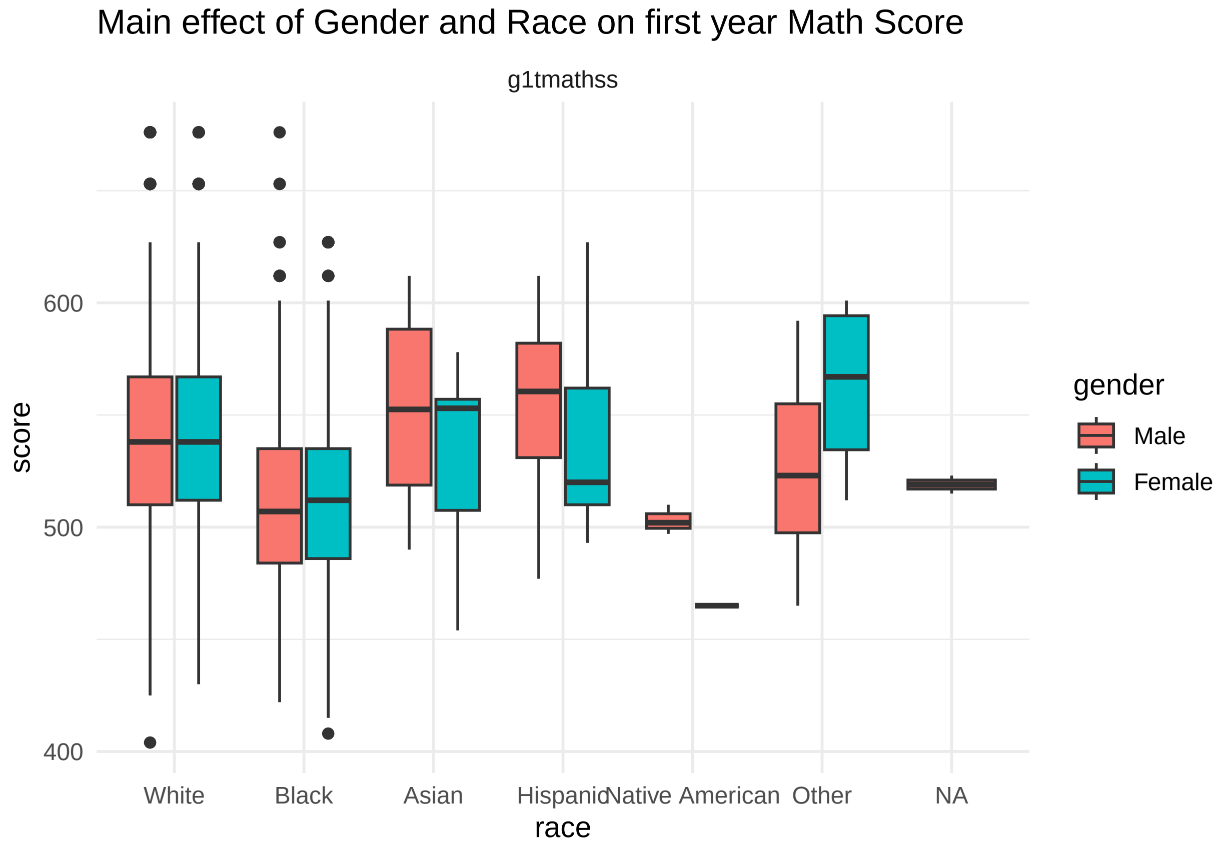


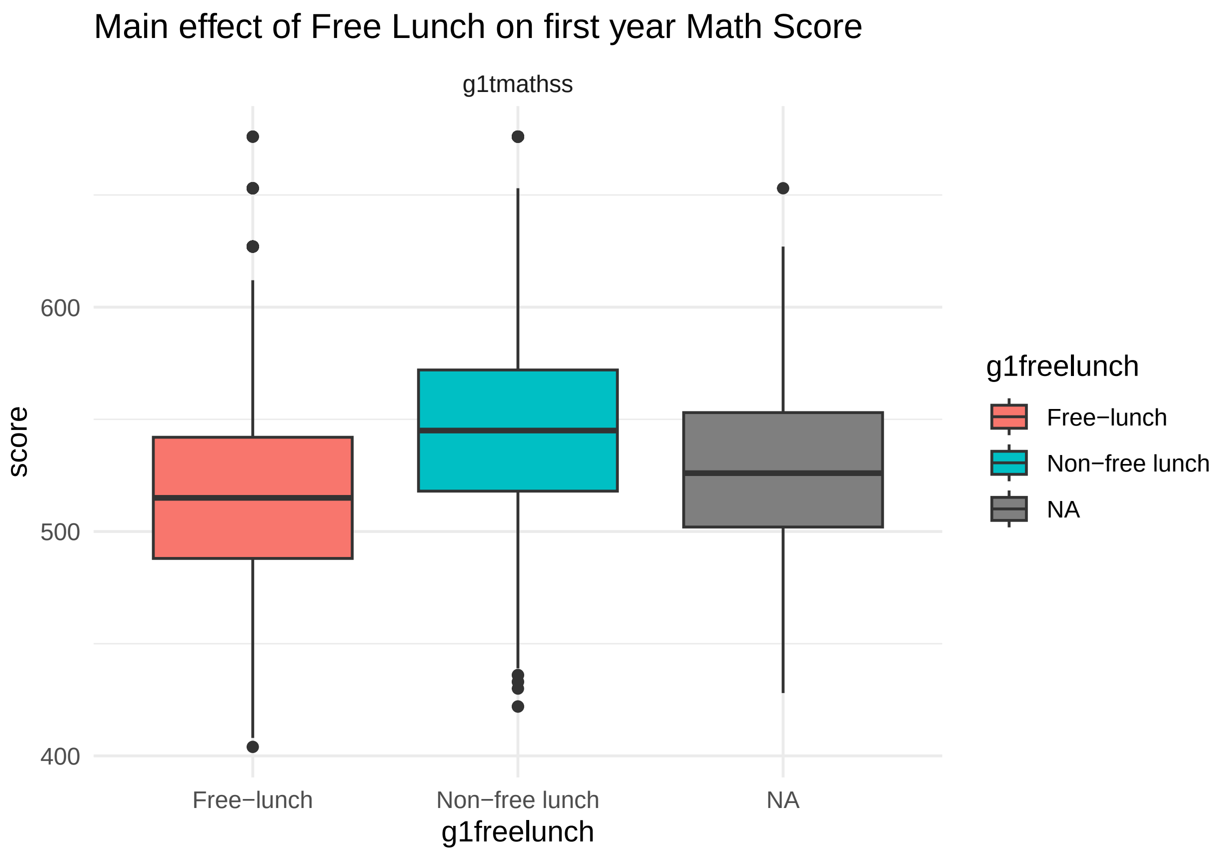
Potential Caveats of Two-way ANOVA model

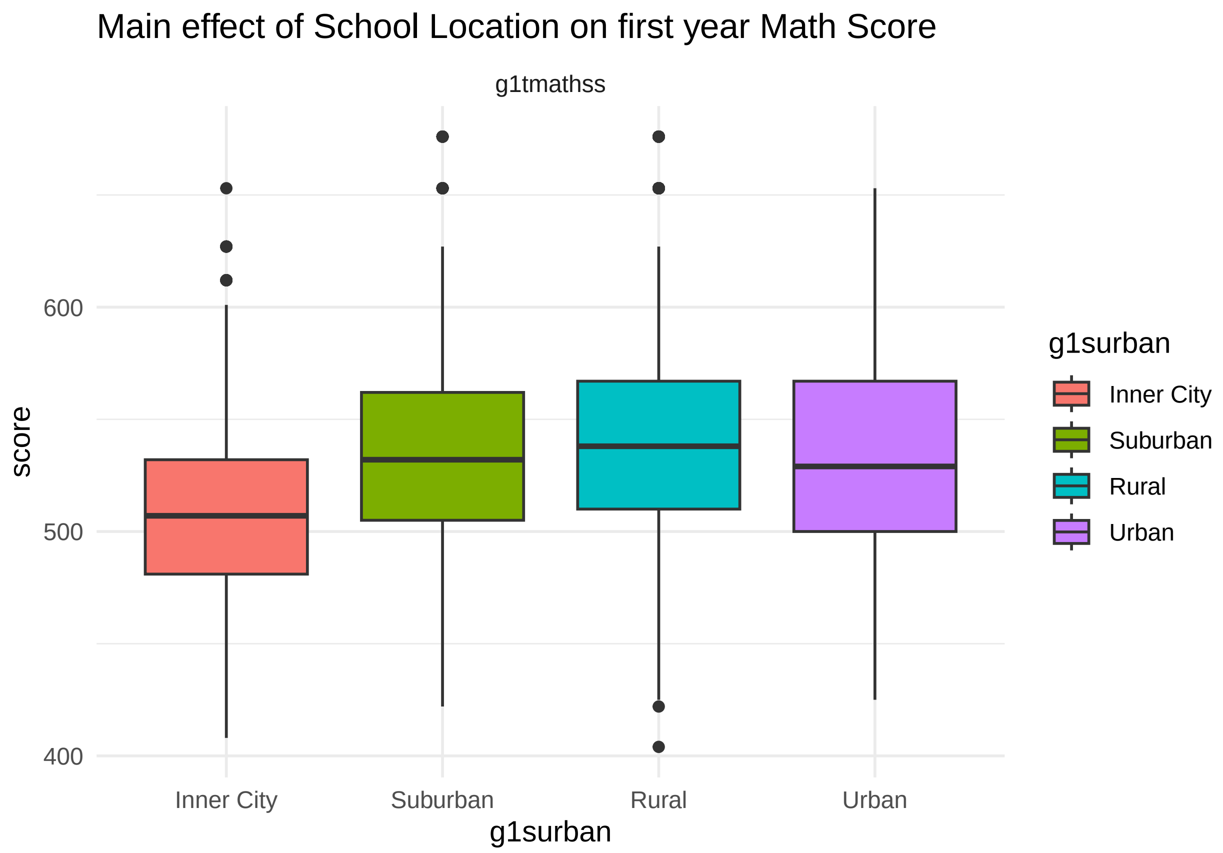
Caveat 1: failed to consider the demographic variables

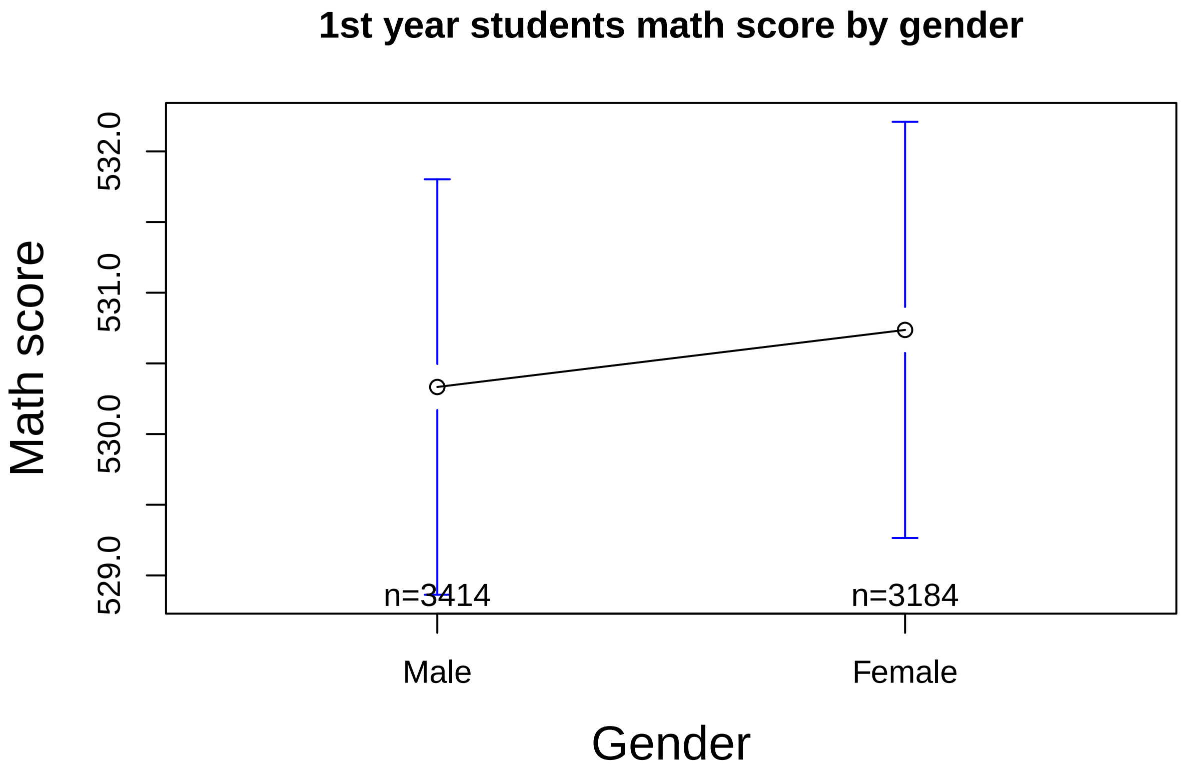
Apart from the null values, another caveat of the initial analysis report is the failure to consider demographic variables. Since we only fitted a two-way ANOVA model, we did not account for the effect of demographic variables. However, it is reasonable to expect that demographic factors could influence students' grades, especially variables like free lunch, which implies family income. For example, students from wealthier families tend to receive more educational investment and social support from their families, leading to better math grades. However, we lack evidence to confirm whether demographic variables significantly affect outcomes in the STAR dataset. If they do not, excluding them would not be a major issue.

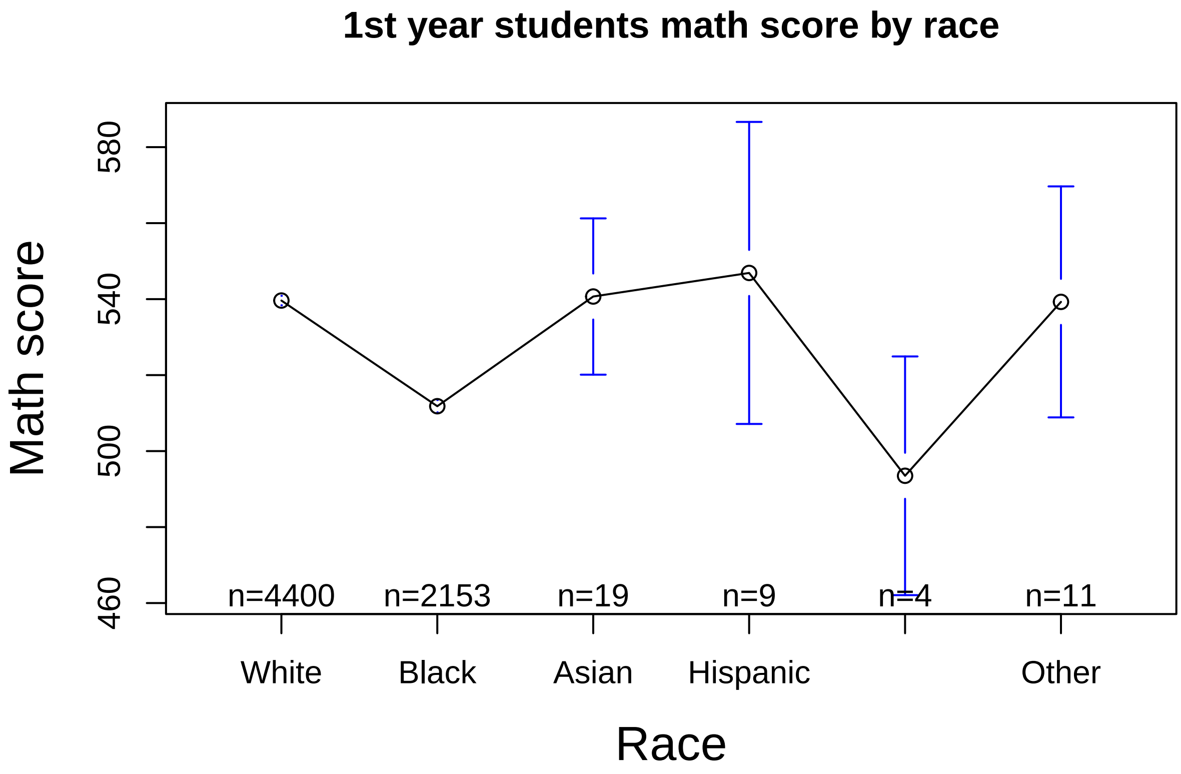
To assess whether demographic variables matter, we conducted a main effect test of demographic variables on first-year math grades. The boxplots and main effect plot for class type and first-year math scores are presented below

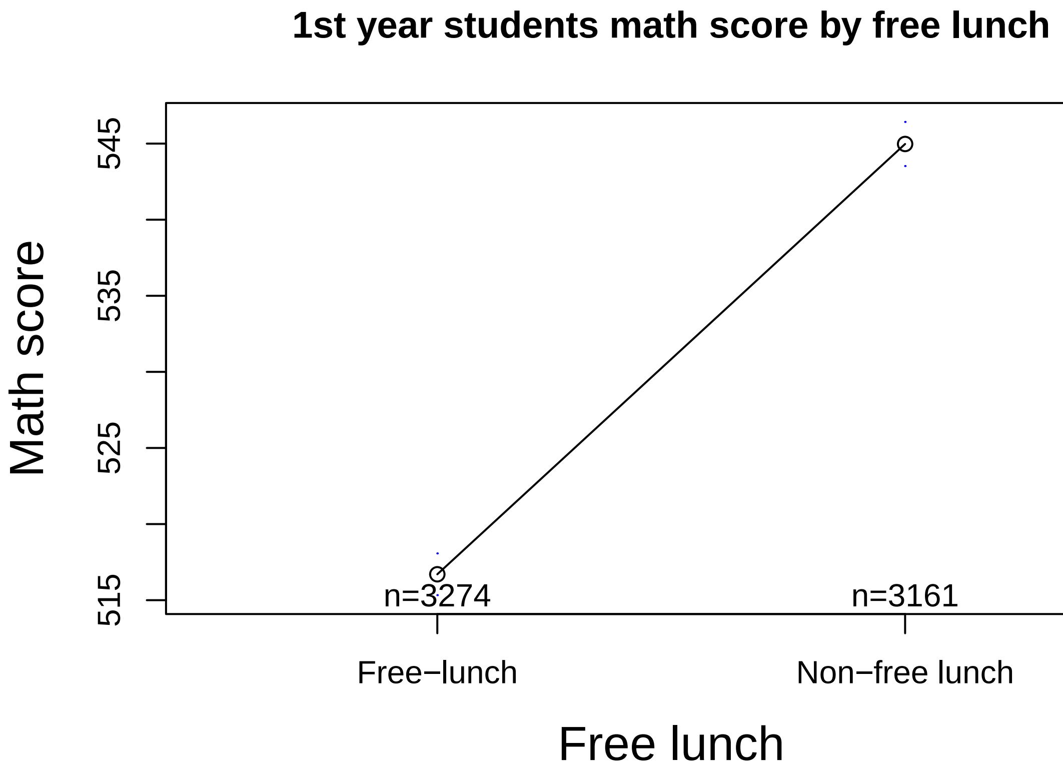


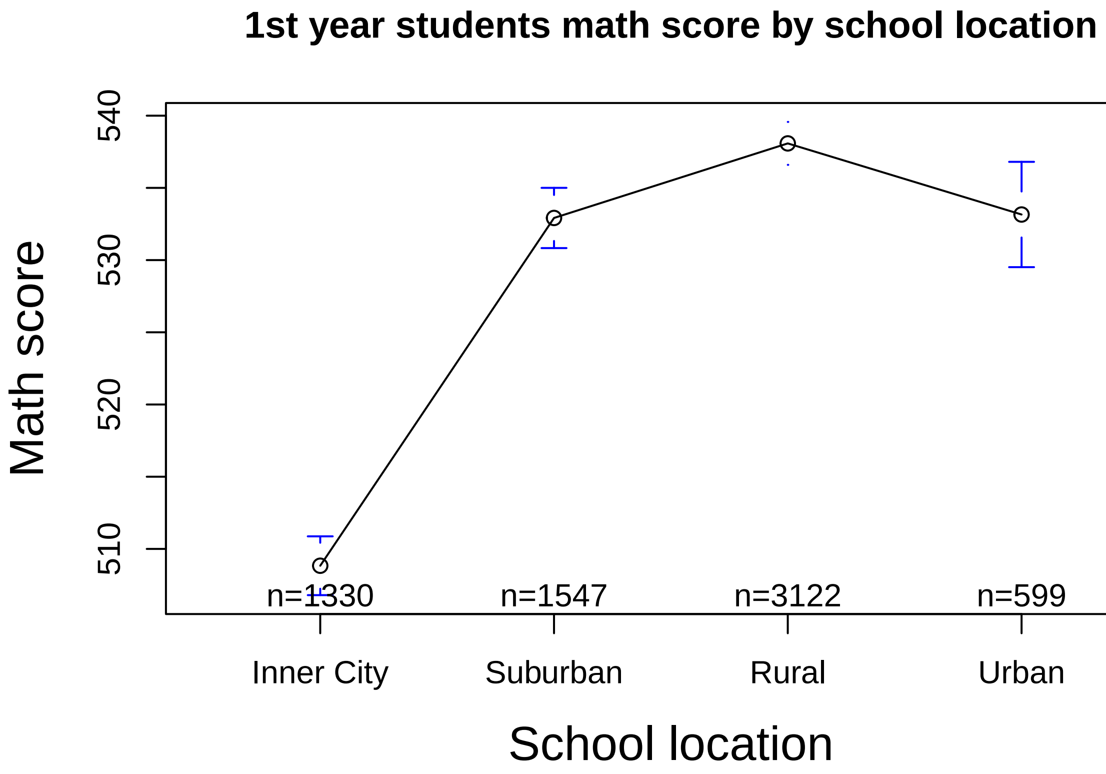

Main effect of Free Lunch on first year Math Score
g1tmathss
score
600
500
400
Free−lunch
Non−free lunch
NA
g1freelunch
g1freelunch
Free−lunch
Non−free lunch
NA


Main effect of School Location on first year Math Score
g1tmathss
score
600
500
400
Inner City
Suburban
Rural
Urban
g1surban
g1surban
Inner City
Suburban
Rural
Urban


1st year students math score by gender
Math score
532.0
531.0
530.0
529.0
n=3414
n=3184
Male
Female
Gender


1st year students math score by race
Math score
580
540
500
460
n=4400
n=2153
n=19
n=9
n=4
n=11
White
Black
Asian
Hispanic
Other
Race


1st year students math score by free lunch
Math score
545
535
525
515
n=3274
n=3161
Free−lunch
Non−free lunch
Free lunch

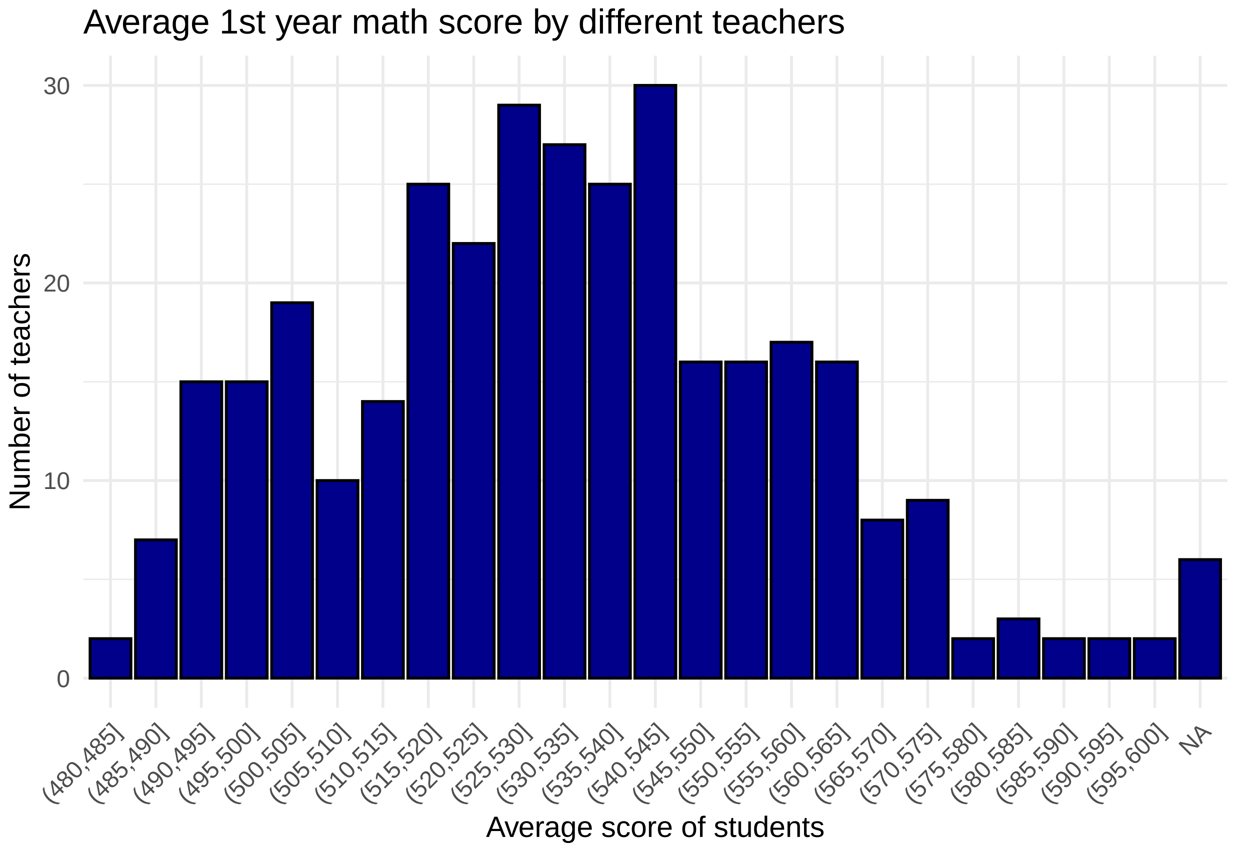


From the graph, we can roughly speaking that demographic variables seems played an important part in predicting the 1st year math grade. So it is necessary to consider the effect from demographic variables.

Caveat 2: haven't consider the effect from teachers

One important factor to consider is the ability of teachers. Teachers may vary significantly in their teaching style, professional attitude, and level of responsibility, which can have a huge influence on student outcomes. These individual differences among teachers may also overlap with other factors. For example, schools in inner cities may have less experienced or less capable teachers because they face financial shortage and cannot afford the salary of experiences teachers. Due to significant influence of teachers, we need to highlight this as a caveat and control for teacher effects to avoid compromising the validity of our findings. After all, our main research question focuses on the effect of class size.

Currently, we lack evidence on how teachers may affect students' math grades. Therefore, we need to determine whether there is significant variation in teaching outcomes among teachers in our dataset.

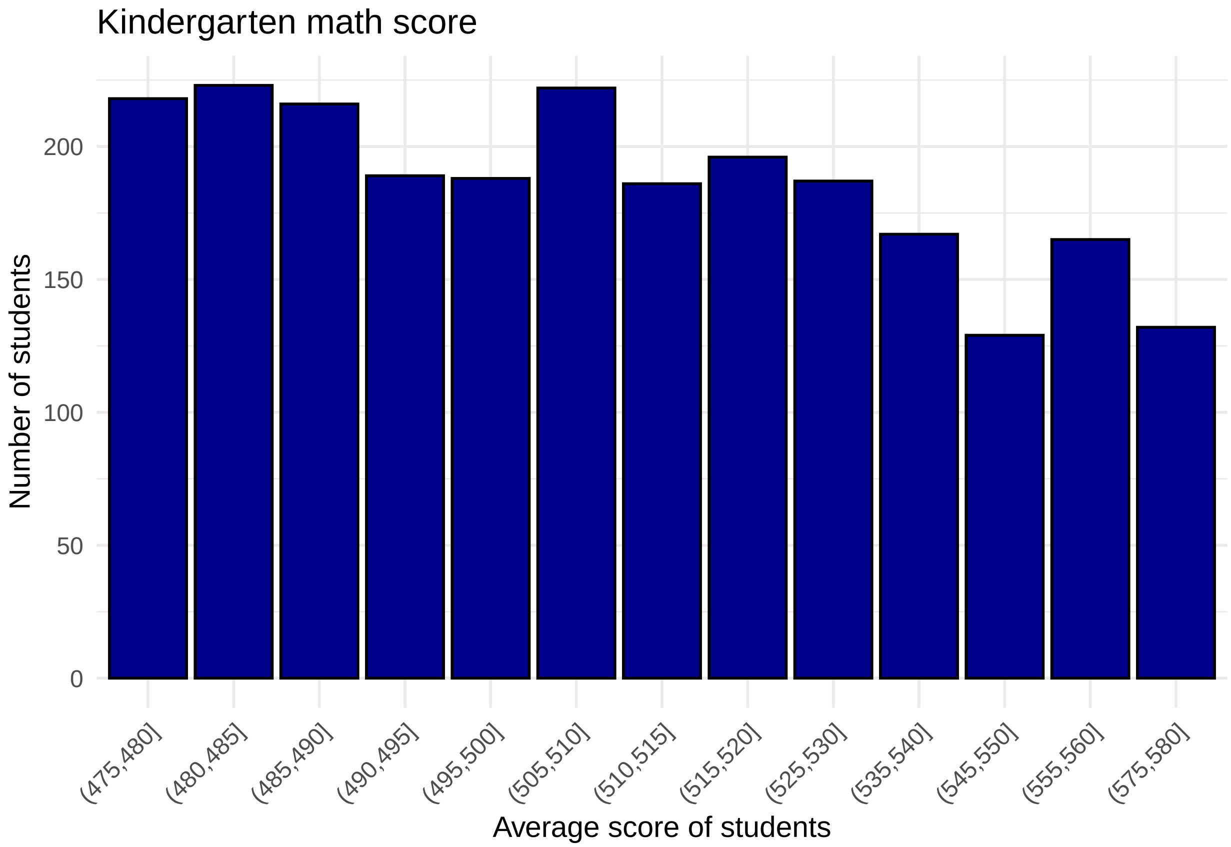


From the results above, we calculated the average score of each teacher's students and measured the range of these mean scores. The mean scores ranged from 480 to 600—a substantial variation of nearly 120 points in average math scores. This suggests that individual teacher factors play a highly significant role in students' math outcomes and should be addressed in our model.

Caveat 3: should consider the factor of students' intelligence

Individual differences also need to be considered when building the model. We have already handled several demographic variables, such as gender and race, to address individual differences. However, a more fundamental factor is intelligence, which is a key determinant of a student's ability to absorb the knowledge taught by teachers. Although the knowledge covered in first-year primary school is very simple, many students may still struggle to keep up with the pace of instruction. Therefore, it is still important to consider the intelligence factors in our study.

Unfortunately, we do not have IQ data in our dataset. However, we could use another variable as a proxy for intelligence: math grades in kindergarten. Since most kindergarten students do not prepare for exams, their kindergarten math grades are likely to reflect their initial intelligence level.



The results show that kindergarten math scores ranged from 475 to 580, which is a substantial variation. This suggests that intelligence exhibits significant individual differences and should be addressed in our model.

Fixing the caveats: linear mixed model with teachers' id

We can address the teacher's issue by applying a linear mixed model, which is presented below. Since we already tested the school ID in the ANOVA model, we will exclude it from the linear mixed model and instead focus on the demographic variables. And the reason why we need to treat teachers as random variable not fixed factors is because three classes types were selected for the study but the number of teacher was far more than class types and the teachers were randomly assigned to the classes. This teacher selection process indicates that the specific levels of teachers are unknown, so we need to treat teachers as random factor.

Applying the linear mixed model, our model should be written as the following:

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \eta_m + \epsilon_{ijklm}$$

Terms:
- i: class type, fixed effect $i = 1, 2, 3$
- j: teacher's id, random effect $j = 1, 2, \ldots, 337$
- k: level of race, fixed effect
- l: level of school location, fixed effect
- m: level of free-lunch, fixed effect
- n_{ijklm}: number of observations for each combination of class type, teacher's id, race, school location, and free-lunch

- Y_{ijklm}: The math score for the mth student in the ith class type within the jth teacher, kth race, lth school location, and mth level of free-lunch.
- μ: Overall average math score across all class types, teacher, races, school location levels, and free-lunch levels.

Effects:
- α_i: The fixed effect of the ith class type on the math score, representing the differential effect of class types on math scores.
- β_j: The random effect of the jth teacher on the math score, capturing variations due to schools.

$$\beta_j \sim N(0, \sigma_\beta^2)$$

- γ_k: The fixed effect of the kth race (ratio of black students) on the math score
- η_m: The fixed effect of the mth level of free-lunch on the math score
- δ_l: The fixed effect of the lth level of school location on the math score
- ϵ_{ijklm}: Random error term,

$$\epsilon_{ijklm} \sim N(0, \sigma^2)$$

Also, β_j and ϵ_{ijklm} are assumed to be mutually independent.

Let's run the model

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: median ~ g1classtype + race + gender + g1freelunch + g1surban +
##     (1 | g1tchid)
##    Data: summary_data_new
##
## REML criterion at convergence: 14635.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1689 -0.5608 -0.0236  0.5872  4.9792
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  g1tchid  (Intercept) 328.8    18.13
##  Residual             555.9    23.58
## Number of obs: 1558, groups:  g1tchid, 337
##
## Fixed effects:
##                            Estimate Std. Error t value
## (Intercept)                529.2564     3.5818 147.764
## g1classtypeRegular          -8.2560     2.7854  -2.964
## g1classtypeRegular_Aid      -7.6044     2.9062  -2.617
## raceBlack                  -17.8642     1.6244 -10.998
## raceAsian                    1.0168     5.6279   0.181
## raceHispanic                11.3472     8.4456   1.344
## raceNative American         -8.3475     8.9976  -0.928
## raceOther                    5.2424     7.6157   0.688
## genderFemale                -0.2366     1.2102  -0.195
## g1freelunchNon-free lunch  17.5368     1.2548  13.975
## g1surbanSuburban             3.7518     3.7387   1.004
## g1surbanRural                5.7055     3.4645   1.647
## g1surbanUrban                3.2846     4.7916   0.685
```

From the results, we can see that we have a huge variance of 328.8 among teachers, which means considering teachers as a random effect is quite reasonable. Interpreting the fixed effect, we can see that after controlling the variance from the teachers, class type still had a huge effect on the math outcome (t<-2). Still, the regular class had the lowest math score compared with the small class, and the regular aid class also had a much smaller math score. Race played a role in the math score, but only for African Americans (t <-10) significantly lower than white Americans. For other races, the effect is very small. Free lunch had a huge effect on the math grade, which free-lunch group was significantly lower than the non-free lunch group. But the results of school location became really weak (t<2), we can consider that the effect brought by school location was mainly the differences among teachers.

To test whether our model fit with the assumption of multicolinarity, we conduct a VIF test:

```
##                 GVIF Df GVIF^(1/(2*Df))
## g1classtype 1.011610  2        1.002890
## g1tchid     1.100571  1        1.049081
## race        1.422617  5        1.035878
## gender      1.002127  1        1.001063
## g1freelunch 1.048559  1        1.023992
## g1surban    1.460822  3        1.065204
```

From the data we can see there's no clear sign of multicollinearity.

Fixing the caveats: considering the intelligence factor

As discussed above, since there are no direct measurements of intelligence, we used kindergarten math grades as an indicator. However, we encountered several issues: (1) There are too many missing values in the kindergarten math grades—specifically, 2,582 NA values. Since nearly half of the kindergarten math grades are missing, simply using them as a predictor in our previous model could introduce bias. (2) Kindergarten grades are continuous variables, and small differences in scores may not reliably indicate meaningful differences in intelligence.

Therefore, one necessary step is to categorize kindergarten math grades by ranking: students in the top 33% should be classified as "high intelligence," those in the middle 33% as "medium intelligence," and those in the bottom 33% as "low intelligence."Another necessary step is to build a new model that includes only students with recorded kindergarten math grades. Since half of the sample has missing values, using any form of imputation could significantly impact the results.

Our model can be written as following. To avoid conflict, we used the κ_p to represent the intelligence factor in the model:

$$Y_{ijklm} = \mu + \alpha_i + \kappa_p + \gamma_k + \delta_l + \eta_m + \epsilon_{ijklm}$$

Terms:
- i: class type, fixed effect $i = 1, 2, 3$
- p: intelligence, fixed effect $j = 1, 2, 3$
- k: level of race, fixed effect
- l: level of school location, fixed effect
- m: level of free-lunch, fixed effect

Effects:
- α_i: The fixed effect of the ith class type on the math score, representing the differential effect of class types on math scores.
- κ_p: The fixed effect of the pth intelligence, capture the individual differences.
- γ_k: The fixed effect of the kth race (ratio of black students) on the math score.

- η_m: The fixed effect of the mth level of free-lunch on the math score.
- δ_l: The fixed effect of the lth level of school location on the math score.
- ϵ_{ijklm}: Random error term,

$$\epsilon_{ijklm} \sim N(0, \sigma^2)$$

```
##
## Call:
## lm(formula = median ~ g1classtype + race + gender + g1freelunch +
##     g1surban + gktmathss_factor, data = summary_data_kind)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.936  -9.878   1.890  11.402  68.120
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                500.222      4.667 107.187  < 2e-16 ***
## g1classtypeRegular         -10.034      2.995  -3.350 0.000937 ***
## g1classtypeRegular_Aid      -8.392      3.020  -2.779 0.005883 **
## raceBlack                  -10.826      2.602  -4.160 4.41e-05 ***
## raceAsian                   18.968      7.326   2.589 0.010199 *
## raceHispanic                41.663     20.043   2.079 0.038692 *
## raceOther                   12.053     11.756   1.025 0.306292
## genderFemale                 3.780      2.467   1.532 0.126821
## g1freelunchNon-free lunch   13.367      2.488   5.372 1.82e-07 ***
## g1surbanSuburban             8.743      3.754   2.329 0.020699 *
## g1surbanRural               10.255      3.803   2.696 0.007501 **
## g1surbanUrban                8.331      3.938   2.115 0.035406 *
## gktmathss_factormedium      26.493      2.955   8.964  < 2e-16 ***
## gktmathss_factorhigh        48.841      3.042  16.058  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 19.64 on 243 degrees of freedom
## Multiple R-squared:  0.6152, Adjusted R-squared:  0.5946
## F-statistic: 29.88 on 13 and 243 DF,  p-value: < 2.2e-16
```

Interpreting the fixed effect, we can see that after controlling the variance from the intelligence and other demographic variables, class type still had a huge effect on the math outcome, p<.001. Still, the regular class had the lowest math score compared with the small class, and the regular aid class also had a much smaller math score. Race played a role in the math score, and for African Americans (b=-10.83, t=-4.16, p<.001) significantly lower than white Americans, while Hispanic American (b=41.66, t=2.08, p<.05) and Asian American (b=18.97, t=2.69, p<.05) have slightly higher grade comparing with white american. Free lunch had a huge effect on the math grade, which free-lunch group was significantly lower than the non-free lunch group(b=13.37, t=5.37, p<.001). Suburban school(b=8.74, t=2.32, p<.05), rural schools(b=10.26, t=2.70, p<.01) and urban schools(b=8.31, t=2.12, p<.05) have slightly higher grade comparing with inner city schools. Intelligence factor have huge impact on the math grade, intelligence medium group have significantly higher grade comparing with intelligence low group (b=26.49, t=8.96, p<.001), as well as intelligence high group comparing with intelligence low group(b=48.84, t=16.06, p<.001).

To test whether our model fit with the assumption of multicolinarity, we conduct a VIF test:

```
##                        GVIF Df GVIF^(1/(2*Df))
## g1classtype      1.032208  2        1.007957
## race             1.218776  4        1.025039
## gender           1.013792  1        1.006872
## g1freelunch      1.030066  1        1.014921
## g1surban         1.145542  3        1.022905
## gktmathss_factor 1.033194  2        1.008197
```

From the data we can see there's no clear sign of multicollinearity.

Discusison and conclusion

The anova model in inferential analysis revealed a significant positive effect of smaller class sizes on student performance within the STAR project. Several factors contribute to the benefits of reduced-size classrooms. Smaller classes may foster greater student engagement, giving the more opportunities to asking and answering questions. The classroom environment will also benefited from reduced amount of people, and fewer students per class enables closer student-teacher interactions, allowing for identification of learning difficulties and timely intervention.

Our linear mixed model and regression model adding students' intelligence strengthened our findings of class size. Result suggest that the positive effect of reduced class size remains robust across variations in teacher ability, student intelligence levels, and demographic variables. This indicates that smaller classes benefit all students regardless of their individual capabilities, teacher quality, or demographic characteristics, which is a compelling argument for implementing class size reduction as a universal educational improvement strategy.

Our analysis also showed a negative impact of free-lunch status and student performance. This factor probably serves as an indicator of the boarder social economic challenge, as low income family more likely to require free lunch welfare. Students from such environments often face severe economic constraints, leading to limited educational resources and opportunities. These conditions create barriers to quality education and hinder focus on academic pursuits, contributing to the observed decrease in performance.

Interestingly, adding teachers into our model, the effects brought by school location and racial factors reduced a lot. This suggests that teacher quality is a critical determinant of student academic outcomes. The disadvantages of math grade observed in inner-city schools and Afircan american students appear largely due to limited access to highly effective teachers. This inequality stems from fiscal constraints, neighborhood safety concerns, and various socioeconomic factors that affect teacher recruitment and retention in these communities.

Based on these findings, we proposal several policies recommendations. First, reducing class size can enhances student participation and improves learning outcomes. Second, increased support for economically disadvantaged students is essential. Policies should implement targeted programs and resources for low-income students and their schools. Third, facilitating cross-regional teacher exchanges by giving subsidies to experienced teachers who willing to teach in disadvantaged inner-city schools. Additionally, organizing structured opportunities for teachers to share pedagogical experiences across different school districts, which would help the spread effective teaching practices.

Our report has several limitations. First, we haven't compare alternative imputation methods to justify our use of median imputation when dealing with missing values. Second, other important variables such as teacher qualifications, experience, and student attendance rates were not fully explored. Third, our additional model was limited to first-order effects, without examining interactions between variables. We will address these limitations in the future research.

Reference

Imbens, G., & Rubin, D. (2015). Stratified Randomized Experiments. In Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction (pp. 187-218). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139025751.010

Session info

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 26100)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] car_3.1-2       carData_3.0-5   lme4_1.1-32     Matrix_1.5-4
##  [5] gplots_3.2.0    broom_1.0.5     haven_2.5.2     psych_2.1.9
##  [9] lubridate_1.9.2 forcats_1.0.0   stringr_1.5.0   dplyr_1.1.4
## [13] purrr_1.0.1     readr_2.1.2     tidyr_1.3.0     tibble_3.2.1
## [17] ggplot2_3.4.4   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.10        lattice_0.21-8     gtools_3.9.5       digest_0.6.31
##  [5] utf8_1.2.3         R6_2.5.1           backports_1.4.1    evaluate_0.23
##  [9] highr_0.10         pillar_1.9.0       rlang_1.1.1        rstudioapi_0.15.0
## [13] minqa_1.2.5        nloptr_2.0.3       rmarkdown_2.25     labeling_0.4.3
## [17] splines_4.1.2      munsell_0.5.0      compiler_4.1.2     xfun_0.37
## [21] pkgconfig_2.0.3    mnormt_2.1.1       htmltools_0.5.6    tidyselect_1.2.1
## [25] viridisLite_0.4.2  fansi_1.0.4        tzdb_0.3.0         withr_3.0.0
## [29] MASS_7.3-58.3      bitops_1.0-7       grid_4.1.2         nlme_3.1-162
## [33] gtable_0.3.4       lifecycle_1.0.4    magrittr_2.0.3     scales_1.2.1
## [37] KernSmooth_2.23-20 cli_3.6.0          stringi_1.7.12     farver_2.1.1
## [41] generics_0.1.3     vctrs_0.6.5        boot_1.3-29        tools_4.1.2
## [45] glue_1.6.2         hms_1.1.3          abind_1.4-5        parallel_4.1.2
## [49] fastmap_1.1.1      yaml_2.3.7         timechange_0.2.0   colorspace_2.1-0
## [53] caTools_1.18.2     knitr_1.42
```