An Analysis of the Influence of Various Factors on High School Graduation Rates in Chicago Public Schools

Abstract: The purpose of this paper is to ascertain which factors have the greatest influence over graduation rate in high schools in Chicago Public Schools (CPS). Using the 2018-2019 CPS report, with identification information, student surveys, and teacher surveys about schools within CPS, we found that mobility rate, school type, involvement of parents, attendance average, school survey safety, ambitious instruction, and culture climate rating have the most significant impact on graduation rate. Mobility rate and attendance were found to be particularly useful predictors in our model. We hypothesize that attendance average is highly correlated with graduation rate because students are more likely to learn and retain information when they are present in the classroom as opposed to when they are missing lectures. Mobility rate may be a significant factor due to the difficulty and culture shock that transferring schools can inflict upon students, hindering their ability to learn comfortably. Because Chicago public schools have some of the lowest graduation rates in the US, research is necessary to determine the causes of low graduation rates and the changes required to accommodate the needs of students in the district.

I. Background and Introduction

The United States is known for having the third-highest population, the most powerful military, and the largest economy in the world (WENR). Despite all of these accomplishments, there are underperforming schools across the American landscape. The urban setting, in particular, accounts for the majority of low-performing high schools and the weakest promoting powers (promotion of 50% or fewer freshmen to senior status on time) of all American high schools (Balfanz & Legters). Characteristics of urban schools include high poverty rates, lower student performance, engagement in risk taking behaviors, higher safety and health risks, and a large minority population (NCES). With 30% of students attending urban public schools and low high school graduation rates in the urban setting, the US is likely to face an uncertain future (EdWeek). American industries and government agencies may soon encounter a skills gap in STEM fields and inadequate numbers of students academically and physically able to serve (TNScore). Thus, low graduation rates would pose significant risks to American economic prosperity and national security.

II. Data and Exploratory Analysis

Data & Variables. The Chicago Public Schools (CPS) system, the third largest in the US, serves 104,733 students, with more than 80% economically-disadvantaged students of color (CPS). The data from the 2018-2019 CPS Report contains a comprehensive yearly report with statistics and metrics derived from identification information and surveys taken by students, parents, and teachers within the CPS system. This information is published and updated by the City of Chicago's Dataworld portal (ibid).

The original data includes 661 observations of 182 variables for elementary, middle, and high schools. This study focuses on CPS high schools with reported graduation rates in the 2018-2019 school year, with 122 observations of 32 variables. 56 schools were omitted due to failure to report information, fundamental differences in function, or small sample size.

The predictors are explained below. Categories within variables were combined due to either a lack of data within a category or a similarity in graduation rate across categories.

Mobility rate is the change of student enrollment from the first school day to the last school day of the given year, calculated by taking the sum of students who transferred out and the students who transferred in and dividing it by the average daily enrollment, then multiplying by 100. As students are counted each time they transfer in or out of the school, students may be counted more than once.

The variable school type reflects the type of each school. The variable initially contained 10 categories: career academy, contract, magnet, military academy, selective enrollment, neighborhood, small, charter, special education, and city-wide option. Due to either similarities in graduation rate or lack of data, the first five were collapsed into a new category "specialized," and the subsequent two became "traditional." Charter remained its own category. Special education and city-wide option were excluded.

Involvement of parents is a measure of the level of involvement from the parents of students in each school based on survey responses. It originally contained the categories very weak, weak, neutral, strong, and very strong, but we collapsed very weak and weak under the category "weak."

School survey safety shows how safe students feel at school based on their responses to the survey. The variable was originally broken down into the categories very weak, weak, neutral, strong, and very strong. We combined very weak and weak into "weak" and very strong and strong into "strong." Neutral remained its own category.

Ambitious instruction is a variable that measures the degree to which classes are academically demanding and engage students by emphasizing the application of knowledge. It originally contained the responses weak, neutral, strong, and very strong. Neutral and weak were combined under the label "neutral."

Culture climate rating reports the results of the "My Voice, My School 5Essentials Survey," which is intended to gauge the level of organization within each school. It originally contained the

categories not yet organized, partially organized, moderately organized, organized, and well organized. The first four were joined into the category "partially organized," and well organized is its own category.

Graduation rate is a reflection of the percentage of students that graduate from each four year school. It is the response variable for our study. The raw data includes two variables for graduation rate: the graduation rate for year one, and the graduation rate for year two. We averaged these two variables in order to obtain the average graduation rate for each school.

Exploratory Data Analysis

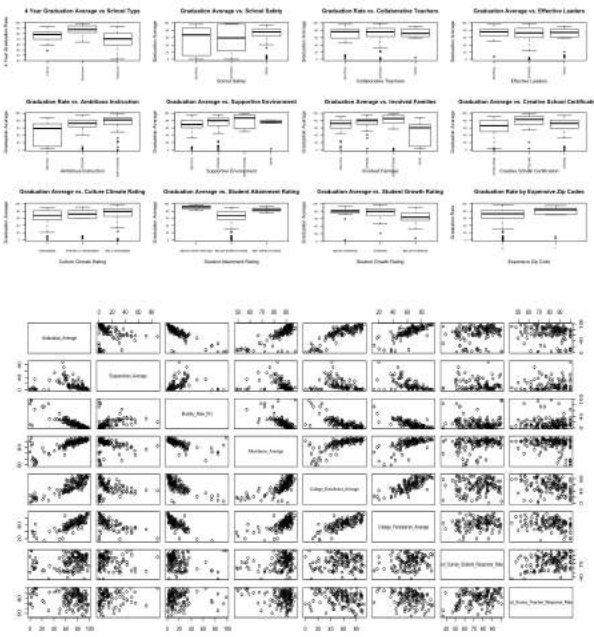


Fig 1. Graduation rates showed large variations based on the school type. Weak and very weak school safety levels have smaller spread but relatively similar graduation rates. Collaborative teachers, effective leaders, supportive environment, and culture climate rating show roughly the same medians but different spreads among the categories in each model. Weak ambitious instruction is associated with an overall lower graduation average. More students from the lower 50% of weak involved families have graduation rates lower than the median of other categories. Creative school certification shows roughly equal medians but emerging creative schools show the largest spread. Below and far below student attainment ratings show the greatest spread in graduation average. Higher student growth rating is associated with smaller spread and higher graduation average. High schools in Chicago's top 25 most expensive zip codes have higher graduation rates and smaller spread (Chicago Tribune).

Fig 2. Among all quantitative variables, mobility rate, attendance average, college enrollment average, and college persistence display moderate to high linearity with graduation rate average.

III. Model and Results

Analytical Methods. A multiple linear regression model was used with High School Graduation Rate as the response variable. We decided that it didn't make sense to use college enrollment rate or college persistence rate as predictors, as these variables are measured after our variable of interest. We also determined that we would not include the number of suspensions per 100 students because 44 of the high schools in Chicago did not have data for this variable, and of those that did, mobility rate and attendance average accounted for much of the variability in graduation rate that could be explained by suspensions average (Appendix A). Additionally, Collaborative Teachers and Effective Leaders did not show a significant difference in graduation rate between the categories (Appendix B and C).

We fit a linear regression model with the remaining variables, and performed selection procedures. Two possible models were calculated, one with 7 predictors, and another with 4 predictors (which were a subset of the 7 predictors). These 7 predictors were then isolated from the data set, allowing 7 more high schools to enter the model as we now had less variables of interest. See Appendix D for more details on selection procedures once these rows were added back into the data set. Partial F-tests were performed on all of the 7 variables in the larger model, and all of these predictors were found to be significant (at a $\alpha = 0.05$ significance level).

Final Model and Results

$$\text{Graduation Average} = -7.16078 + 2.16588\,(I_{AmbitiousInstruction-Strong}) + 5.25548(I_{AmbitiousInstruction-VeryStrong}) + 5.31106(I_{SchoolType-Specialized}) + 3.75162(I_{SchoolType-Traditional}) - 0.85049\,(I_{SchoolSurveySafety-Strong}) - 3.85947\,(I_{SchoolSurveySafety-Weak}) + 4.38409(I_{SchoolSurveyInvolvedFamilies-Strong}) + 4.51926(I_{SchoolSurveyInvolvedFamilies-VeryStrong}) + 1.82953(I_{SchoolSurveyInvolvedFamilies-Weak}) + 2.68795(I_{CultureClimateRating-PartiallyOrganized}) - 3.77494(I_{CultureClimateRating-WellOrganized}) - 0.66966\,(MobilityRatePct) + 0.99164\,(AttendanceAverage).$$

	2.5 %	97.5 %
(Intercept)	-39.5468025	25.2252483
Mobility_Rate_Pct	-0.8432235	-0.4961045
Attendance_Average	0.6558617	1.3274239
factor(School_Survey_Ambitious_Instruction)STRONG	-1.5044291	5.8361813
factor(School_Survey_Ambitious_Instruction)VERY STRONG	1.1640849	9.3468801
factor(School_Type)Specialized	2.3997184	8.2223993
factor(School_Type)Traditional	1.1050008	6.3982419
factor(School_Survey_Safety)STRONG	-7.7640358	6.0630548
factor(School_Survey_Safety)WEAK	-6.9945856	-0.7243582
factor(School_Survey_Involved_Families)STRONG	1.3692188	7.3989628
factor(School_Survey_Involved_Families)VERY STRONG	0.6073546	8.4311601
factor(School_Survey_Involved_Families)WEAK	-3.4521450	7.1112074
factor(Culture_Climate_Rating)PARTIALLY ORGANIZED	-1.0929898	6.4688857
factor(Culture_Climate_Rating)WELL ORGANIZED	-6.8742689	-0.6756205

This is with "Neutral" being the baseline group for Ambitious Instruction, "Charter" being the baseline group for School Type, "Neutral" being the baseline group for School Survey Safety, "Neutral" being the baseline group for School Survey Involved Families, and "Organized" being the baseline group for Culture Climate Rating. See Appendix E for model assumptions.

Fig 3. The table above contains the 95% confidence intervals for the slopes of all the variables in the model. It is important to note that some of the indicator variables contain 0 in their 95% confidence intervals. However, their corresponding categorical variables have significant results from partial F-test (Appendix G).

The model has an F-value of 45.18, with p-value of less than 2.2e-16. Therefore, our full model is very effective as compared to the intercept model. We have a residual standard error (the standard deviation of the errors) of 5.398. R^2 is 0.8447, which means that 84.47% of the variation in high school graduation rates is accounted for by the multiple linear regression model with attendance average, mobility rate, ambitious instruction rating, school type, safety rating, involved families rating, and culture climate rating. R^2_{adj} is 0.826, which reflects the number of predictors and the amount of variability in high school graduation rate explained by the predictors. See Appendix F and H for summary and ANOVA output.

IV. Conclusion and Discussion

Mobility rate, attendance average, school type, school safety, ambitious instruction, culture climate, and involved families were important in predicting average graduation rate in CPS high schools. Of these predictors, mobility rate and attendance average were the most significant in predicting the graduation rate (Appendix G).

Attendance average is highly correlated with graduation rate because students are more likely to learn and retain information when they are physically present than absent. Mobility rate could be explained by the difficulty and culture shock of the transfer process, hindering their ability to learn comfortably. More selective school types, such as specialized schools, may have higher graduation rates due to the type of students they attract. Involved families and ambitious instruction are both positively correlated with graduation rate since teachers and relatives create both a student's learning environment and support system.

Given the other predictors, culture climate and school safety were negatively correlated with graduation rate. A well-organized culture climate rating, for example, may be explained by ambitious instruction. School safety ratings may be explained by school type, since selective schools are more likely to be located in safer areas and have better security protocols.

The limitations of the study include response rate, type of data gathering, and omitted schools. With a median of only 78.10% of students and 83.85% of teachers responding to the survey, the data may not accurately reflect the opinions of all CPS personnel (Appendix I). Most of our predictor variables in the model were categorical that were measured in a survey with imprecise options, such as "strong" and "weak". A way to improve our study would be finding more precise ways to measure these variables. Mobility rate may also mildly violate the independence assumption as a school's transfer could affect another school's transfers. We were unable to include 47 schools due to blanks in the data. 9 of these were omitted due to fundamental differences in function, high variance in graduation, and small sample size. Therefore, we were unable to model graduation rates for special education and citywide option schools.

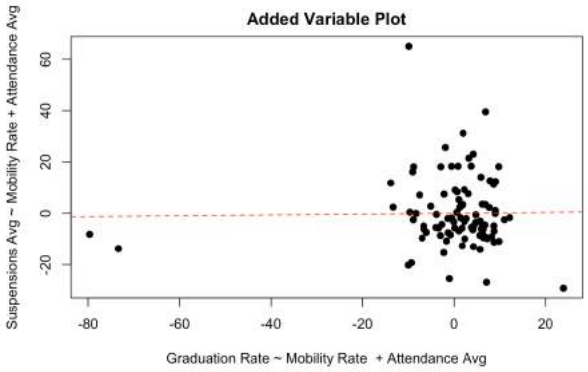
Chicago has one of the largest populations of students that are unable to graduate high school. It is extremely important for research to be done on changes that governments, school systems, faculty, teachers, and parents can make.

References

1. "A Matter of National Security: How K-12 Education Impacts America's Military." Education Impact Series Military, TNSCORE, 2017, tnscore.org/wp-content/uploads/2018/09/A-Matter-Of-National-Security_Education-MilitaryImpact2017.pdf.
2. Balfanz, Robert, and Legters, Nettie. "LOCATING THE DROPOUT CRISIS." Johns Hopkins University Center for Social Organization of Schools, June 2004.
3. "Chicago Public Schools - School Progress Reports SY1819 - Dataset by Cityofchicago." Data.world, 13 Nov. 2019, data.world/cityofchicago/dw27-rash.
4. "CPS Stats and Facts." Chicago Public Schools, cps.edu/About_CPS/At-a-glance/Pages/Stats_and_facts.aspx.
5. Loo, Bryce. "Education in the United States of America." WENR, 16 Apr. 2019, wenr.wes.org/2018/06/education-in-the-united-states-of-america.
6. Riser-Kositsky, Maya. "Education Statistics: Facts About American Schools." Education Week, 2 Apr. 2020, www.edweek.org/ew/issues/education-statistics/index.html.
7. Rockett, Darcel. "These Are the Most Expensive ZIP Codes in the Chicago Area." Chicagotribune.com, Chicago Tribune, 13 Jan. 2020, www.chicagotribune.com/real-estate/ct-re-listicle-most-expensive-chicago-area-zip-codes-tt-20200109-20200113-hvaxizr6pvbnzpluibmruhlgg4-list.html.
8. "SCHOOL QUALITY RATING POLICY (EFFECTIVE FOR THE 2020-2021 SCHOOL YEAR)." Chicago Public Schools Policy Manual, 26 June 2019, https://policy.cps.edu/download.aspx?ID=267.
9. "Urban Schools: The Challenge of Location and Poverty." Urban Schools: Executive Summary, National Center for Educational Statistics, nces.ed.gov/pubs/web/96184ex.asp.

Appendix

Appendix A



Above is an added variable plot for adding Average Suspensions to a model predicting Graduation Rate using Mobility Rate and Average Attendance. Since there is no linear pattern on the plot, it probably won't be worth adding Average Suspensions to the model. It also had a F value of 0.0219, with a p-value of 0.8827556, for it's partial F-test. Average Suspensions also had 44 high schools without data, so we would have to omit 44 different high schools from our model if we were to include Average Suspensions. Therefore, due to the random pattern on the added variable plot and the large p-value in our partial F-test, we decided to not include Average Suspensions to our model.

Appendix B

```
  Posthoc multiple comparisons of means : Bonferroni
    85% family-wise confidence level

$School_Survey_Effective_Leaders
                          diff     lwr.ci   upr.ci   pval
STRONG-NEUTRAL      -151.87222  -505.6494 201.9050 1.0000
VERY STRONG-NEUTRAL -189.86129 -1006.9244 627.2018 1.0000
WEAK-NEUTRAL        -251.19020  -642.0639 139.6835 0.8851
VERY STRONG-STRONG   -37.98907  -870.5212 794.5431 1.0000
WEAK-STRONG          -99.31797  -521.5738 322.9378 1.0000
WEAK-VERY STRONG     -61.32890  -910.2893 787.6315 1.0000

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                             Df   Sum Sq Mean Sq F value Pr(>F)
School_Survey_Effective_Leaders  3  1450425  483475   0.804  0.494
Residuals                      134 80547206  601099
```

Above is a Bonferroni 85% family-wise confidence level for an ANOVA model with graduation rate as the response variable and School Survey Effective Leaders. Model assumptions were met by transforming Graduation Rate to $Graduation\ Rate^{1.75}$. It is clear that even at an 85% confidence level, all of the confidence intervals contain 0, so there is not a significant difference in graduation rate between the different categories, so we did not include Effective Leaders in our model.

Appendix C

```
  Posthoc multiple comparisons of means : Bonferroni
    85% family-wise confidence level

$School_Survey_Collaborative_Teachers
                          diff     lwr.ci   upr.ci   pval
STRONG-NEUTRAL       143.28101 -203.6856 490.2476 1.0000
VERY STRONG-NEUTRAL  184.05293 -308.3648 676.4707 1.0000
WEAK-NEUTRAL          66.21535 -474.1179 606.5486 1.0000
VERY STRONG-STRONG    40.77192 -432.8353 514.3791 1.0000
WEAK-STRONG          -77.06566 -600.3137 446.1824 1.0000
WEAK-VERY STRONG    -117.83758 -747.0259 511.3507 1.0000

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                 Df   Sum Sq Mean Sq F value Pr(>F)
School_Survey_Collaborative_Teachers  3   702545  234182   0.386  0.763
Residuals                           134 81295087  606680
```

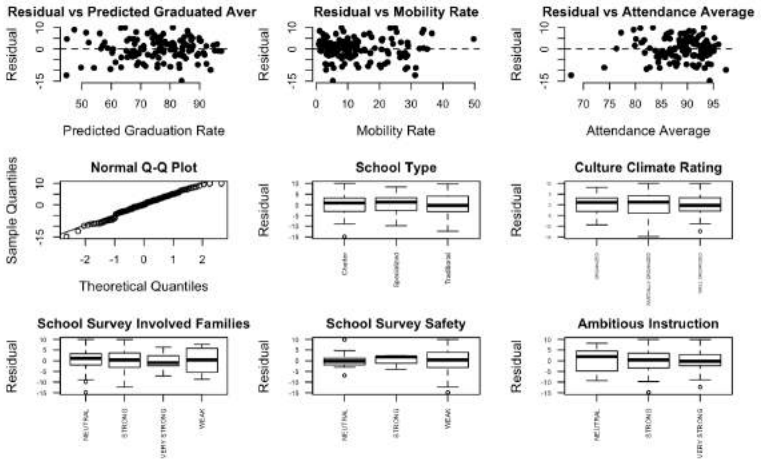
Above is a Bonferroni 85% family-wise confidence level for an ANOVA model with graduation rate as the response variable and School Survey Collaborative Teachers. Model assumptions were met by transforming Graduation Rate to $Graduation\ Rate^{1.75}$. It is clear that even at an 85% confidence level, all of the confidence intervals contain 0, so there is not a significant difference in graduation rate between the different categories, so we did not include Collaborative Teachers in our model.

Appendix D

```
Start:  AIC=424.5
Graduation_Average ~ School_Type + Culture_Climate_Rating + School_Survey_Involved_Families +
    School_Survey_Ambitious_Instruction + School_Survey_Safety +
    Mobility_Rate_Pct + Attendance_Average

                                    Df Sum of Sq    RSS    AIC
<none>                                          3146.5 424.50
- School_Survey_Safety               2    182.57 3329.1 427.38
- School_Survey_Involved_Families    3    269.53 3416.1 428.53
- School_Survey_Ambitious_Instruction 2   264.37 3410.9 430.35
- Culture_Climate_Rating             2    280.35 3426.9 430.92
- School_Type                        2    442.39 3588.9 436.55
- Attendance_Average                 1    998.36 4144.9 456.12
- Mobility_Rate_Pct                  1   1704.15 4850.7 475.31

Call:
lm(formula = Graduation_Average ~ School_Type + Culture_Climate_Rating +
    School_Survey_Involved_Families + School_Survey_Ambitious_Instruction +
    School_Survey_Safety + Mobility_Rate_Pct + Attendance_Average,
    data = chicagoModel2)

Coefficients:
                                (Intercept)                          School_TypeSpecialized
                                    -7.1608                                          5.3111
                        School_TypeTraditional       Culture_Climate_RatingPARTIALLY ORGANIZED
                                     3.7516                                          2.6879
           Culture_Climate_RatingWELL ORGANIZED         School_Survey_Involved_FamiliesSTRONG
                                    -3.7749                                          4.3841
    School_Survey_Involved_FamiliesVERY STRONG           School_Survey_Involved_FamiliesWEAK
                                     4.5193                                          1.8295
     School_Survey_Ambitious_InstructionSTRONG  School_Survey_Ambitious_InstructionVERY STRONG
                                     2.1659                                          5.2555
                   School_Survey_SafetySTRONG                   School_Survey_SafetyWEAK
                                    -0.8505                                         -3.8595
                          Mobility_Rate_Pct                            Attendance_Average
                                    -0.6697                                         0.9916
```

Above is a backward selection procedure, showing that all of the variables were deemed significant enough to be left in the final model. The same model was found by forward and stepwise selection.

Appendix E



It appears that all of the model assumptions are met. There is a random scatter of the residuals around 0 on the residual vs predicted graduation rate plot, and random scatter of the residuals around 0 on the residual vs quantitative variables plots. Generally, it appears that the mean of all of the categories' residuals are around 0. There is a slight deviation in Ambitious Instruction's Neutral Category, but that is a very small difference and I don't think it has much of an effect on the model. Thus, linearity and

constant variance are met. The observations are independent of each other. The number of kids that graduate from one school does not affect the number of kids that graduate from another. The only variable that might be a violation of independence is mobility rate, which is the sum of the students who transferred out and the students who transferred in, divided by the average daily enrollment, multiplied by 100. We are aware that students could transfer to other schools in Chicago, making their mobility rates dependent on each other. However, we will move forward with caution, and assume that the independence assumption is met. Finally, the normal qqplot of residuals follows a relatively straight line; there is slight deviation on the left tail, but it is a small violation. Thus, normality is also satisfied.

Appendix F

```
Analysis of Variance Table

Response: Graduation_Average
                                          Df  Sum Sq Mean Sq  F value    Pr(>F)
Mobility_Rate_Pct                          1 14742.4 14742.4 506.0133 < 2.2e-16 ***
Attendance_Average                         1   896.2   896.2  30.7614 2.093e-07 ***
factor(School_Survey_Ambitious_Instruction)  2   277.8   138.9   4.7680  0.010367 *
factor(School_Type)                        2   617.4   308.7  10.5965 6.280e-05 ***
factor(School_Survey_Safety)               2   182.9    91.5   3.1389  0.047309 *
factor(School_Survey_Involved_Families)    3   112.9    37.6   1.2919  0.280964
factor(Culture_Climate_Rating)             2   280.4   140.2   4.8113  0.009962 **
Residuals                                108  3146.5    29.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table for our final regression model.

Appendix G

Predictor Variable	**F value**	**p-value**
Mobility_Rate_Pct	58.4925	8.993e-12
Attendance_Average	34.2674	5.234e-08
factor(School_Survey_Ambitious_Instruction)	4.5370	0.012824
factor(School_Type)	7.5922	0.000822
factor(School_Survey_Safety)	3.1331	0.04757
factor(School_Survey_Involved_Families)	3.0837	0.03042
factor(Culture_Climate_Rating)	4.8113	0.009962

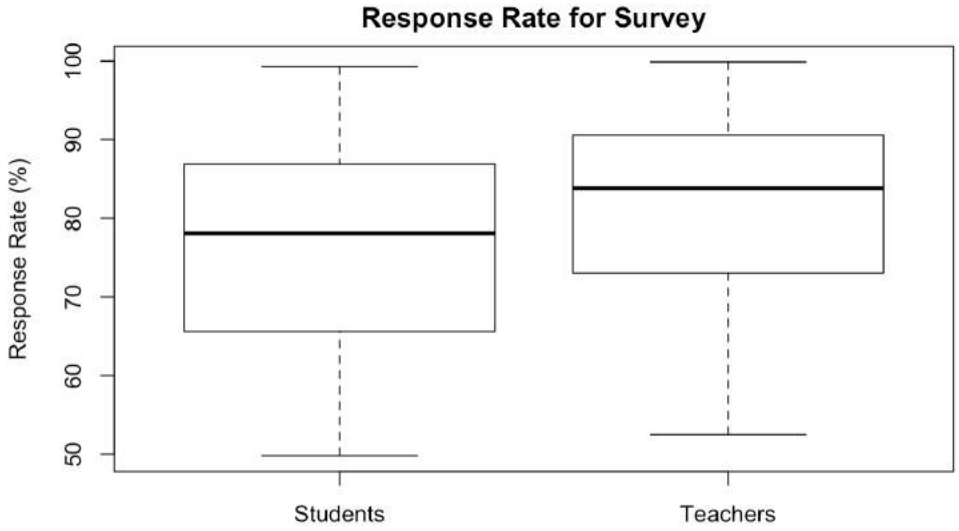
The partial F-tests for all of the predictor variables. All of the predictor variables are significant, with p-values less than 0.05. It is clear that Mobility_Rate_Pct and Attendance_Average are the most important variables to the model. In fact, removing Mobility_Rate_Pct brings R^2_{adj} from 0.826 to 0.7342, and removing Attendance_Average brings R^2_{adj} from 0.826 to 0.7729.

Appendix H

```
Call:
lm(formula = Graduation_Average ~ Mobility_Rate_Pct + Attendance_Average +
    factor(School_Survey_Ambitious_Instruction) + factor(School_Type) +
    factor(School_Survey_Safety) + factor(School_Survey_Involved_Families) +
    factor(Culture_Climate_Rating), data = chicagoModel2)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8132  -3.0127   0.2215   3.5708   9.9526

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      -7.16078   16.33864  -0.438 0.662064
Mobility_Rate_Pct                                -0.66966    0.08756  -7.648 8.99e-12 ***
Attendance_Average                                0.99164    0.16940   5.854 5.23e-08 ***
factor(School_Survey_Ambitious_Instruction)STRONG 2.16588    1.85166   1.170 0.244698
factor(School_Survey_Ambitious_Instruction)VERY STRONG 5.25548 2.06410  2.546 0.012302 *
factor(School_Type)Specialized                    5.31106    1.46876   3.616 0.000456 ***
factor(School_Type)Traditional                    3.75162    1.33521   2.810 0.005887 **
factor(School_Survey_Safety)STRONG               -0.85049    3.48786  -0.244 0.807815
factor(School_Survey_Safety)WEAK                 -3.85947    1.58165  -2.440 0.016308 *
factor(School_Survey_Involved_Families)STRONG     4.38409    1.52099   2.882 0.004763 **
factor(School_Survey_Involved_Families)VERY STRONG 4.51926   1.97354   2.290 0.023969 *
factor(School_Survey_Involved_Families)WEAK       1.82953    2.66459   0.687 0.493801
factor(Culture_Climate_Rating)PARTIALLY ORGANIZED 2.68795    1.90747   1.409 0.161658
factor(Culture_Climate_Rating)WELL ORGANIZED     -3.77494    1.56360  -2.414 0.017448 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.398 on 108 degrees of freedom
Multiple R-squared:  0.8447,    Adjusted R-squared:  0.826
F-statistic: 45.18 on 13 and 108 DF,  p-value: < 2.2e-16
```

This is the summary output from our final model. It contains all of the slopes for the predictor variables, along with their standard errors, t-values, and p-values. It also contains residual standard error, F-statistic, R^2 and R^2_{adj}.

Appendix I



The above boxplot shows the response rate for the schools involved in the survey. Students had a median response rate of 78.10% and a mean of 76.68%, with an IQR of 21.18% (86.90 - 65.72). Teachers had a median response rate of 83.85% and a mean of 81.28% with an IQR of 17.30% (90.38 - 73.08). This is a limitation of our study. With only a median of 78.10% of students responding to the survey, and a median of 83.85% of teachers responding to the survey, we can't be sure that our data accurately reflects the opinions of all of the students and teachers in Chicago high schools. Perhaps those that responded were more likely to have positive opinions about their school, and thus our model was built with data that overestimates the true opinions of Chicago high schools, or vice versa.