Group 1: Pooja Patel, Zehra Amir, Eric Navarro, Nelson Corral
STAT 382
Professor Embers
December 4, 2020

Final Project Report: Academic Performance Across CPS High Schools

Introduction

The Chicago Public School (CPS) system consists of more than 600 hundred schools of different levels that altogether serve over 350,000 students. These students come from diverse socioeconomic and cultural backgrounds and cover a wide range in terms of academic performance. The academic performance of students in the CPS school system varies tremendously across the many different schools present all over the large city of Chicago. In this project, we opted to focus specifically on examining the academic performance of high school students in the CPS school system. The goal of this project was to examine the academic performance of students across CPS high schools, using each school's average composite ACT scores as a primary indicator of academic performance. We wanted to see how factors such as the student attendance rates and suspension rates of a school and the school type impacted the school's average composite ACT score. Additionally, we also wanted to look at the four year graduation rates of CPS high schools and how they compare to the CPS reported average and the national graduation rate. In this project, we had five prominent questions that we hoped to answer through our analyses:

1. Is there a linear association between ACT scores and attendance rates?
2. Is there a linear association between ACT scores and suspension rates?
3. Does the average ACT score differ between different school types?
4. Does the average CPS high schools graduation rate of our data set differ from the average graduation rate reported by CPS?
5. Does CPS's graduation rate differ from the national graduation rate?

After carrying out various statistical tests, we found that there was a strong positive linear association between the average composite ACT scores of schools and the attendance rates of each of the schools. Looking at the suspension rates and the average composite ACT scores of schools, we found that there was a negative linear association between the two variables. We also determined that the average ACT score does differ between the seven different represented school types. As for whether the average CPS high school graduation rate of our data set differs from the average graduation rate reported by CPS, we found that there was no difference between the two graduation rates in question. On the other hand, we did determine that the four year graduation rate of CPS schools did differ from the national graduation rate reported for 2016-2017 which was 84.5% (Education Week 2017).

Data

The data set used for the various statistical analyses in this report originates from two separate larger datasets. One of these datasets was a complete account of all the data included in CPS's School Progress Reports for the academic year, 2016-2017, for schools of all primary levels, while the other dataset was a complete account of the average ACT scores of CPS high schools. Both of these datasets were merged in creating the final smaller, subsetted data set. The smaller, subsetted data set used throughout the analyses in this report focused more specifically on data regarding high schools and their students. It consists of 81 observations of 39 variables. The 81 observations represented 81 separate high schools. The 39 variables consisted of a mix of both categorical and quantitative variables that contained information regarding many aspects of the high schools from average ACT math, reading, english, science, and composite scores to attendance and suspension rates along with the results of student and parent surveys. Out of these 39 variables, we focused upon and analyzed 7 particular variables: School Type, Student Attainment Ratings, School Survey Results Regarding the Supportive Environment of the School, Suspensions Per 100 Students in 2016, Student Attendance in 2016, 4 Year Graduation Rate in 2016, and Average Composite ACT Score.

Analysis

To answer the five questions aforementioned in the introduction of this report, we used a myriad of different analytical methods. Before starting any exploratory analysis or statistical analyses, we used the summary functions and is.na functions in R to clean our data set and remove any high schools with missing values. We then performed exploratory analyses on seven variables for which we began by using the summary function on all the variables. For the categorical variables, we then used the table function and barplot function to find and plot the number of occurrences of each possible value for each variable. For the quantitative variables, we used the histogram function, the qqnorm and qqline functions, and the function for the Shapiro-Wilk test (shapiro.test) to assess the normality of such variables. We used this set of three variables to also assess the normality of residuals during our linear regression analyses. For significant testing, we used the function for t-tests (t.test), the function for the ANOVA test (anova), the TukeyHSD() function for running a post-hoc Tukey test. We used the t tests to determine if there was a difference between the average graduation rate of CPS high schools in our data set and the reported CPS average as well as to look at if there was a difference between the average graduation rate of CPS high schools and the national average. The ANOVA test was used to determine if the average composite ACT score differs between different school types, and the Tukey test was used after the ANOVA test to determine what the specific differences were. To look at whether there was a linear association between the average composite ACT score and suspension rates as well as student attendance rates, we conducted a linear regression and did follow up tests to address the appropriateness of linear regression models. For the linear regressions, we used the plot and abline functions, the correlation function (cor()) to determine the correlation coefficient, the correlation test function (cor.test()), the function to generate a linear regression model (lm(y~x)), and the function to calculate residuals (resid()),

Issues

The first set of issues we faced in attempting to examine student performance across CPS high schools and answer the questions we posed in the introduction of this report included finding a data set that contained all the relevant variables and information that we would need to answer these questions. The first data set we found was one that summarized and contained information regarding the CPS school reports for all CPS schools for the academic year 2016-2017. While this first data set had a wide range of variables including results of surveys and average performance on standardized tests, it lacked data for the average ACT scores of the students at each high school. We found an additional data set containing the average ACT scores of students at each CPS high school and merged it with our original data set.

Additionally, the first data set that was found had several high schools that were missing values for some of the variables we wanted to look at, so we decided to omit those high schools from our data set simply because we wanted to examine the same set of high schools in all of our analyses opposed to looking at different sets for each variable. Because a number of high schools were removed from the first data set, when we merged the second data set containing the average ACT scores of each high school, we had to take caution in ensuring that both data sets contained the same set of high schools and did so partially through manual methods opposed through the use of R.

Furthermore, the data set for Suspensions Per 100 Students was a cause for concern as well because some of the percentages seem surprisingly high and felt that some of this data might've been misinterpreted. As a group, we determined that some of higher percentages accounted for multiple suspensions from the same students at the school. However, as a group, we proceeded with caution and were open to any determinations and conclusions that were associated with this dataset.

Results

Exploratory Analysis

Prior to carrying out any statistical tests, we performed some basic exploratory analysis on 7 different variables in our data set, three of which were categorical variables and four were quantitative variables.

School Type	Career Academy	Contract	Magnet	Military Academy	Neighborhood	Selective Enrollment	Small
Number of Schools	4	2	4	5	43	10	13

Table 1. Summary Statistics of School Type. As shown in Table 1, the summary statistics of the categorical variable, school type, demonstrated that the majority of the schools in our data set were neighborhood schools. Following neighborhood schools, small and selective enrollment schools were the next most represented school types. It is important to note here that for some of the school types there were very small numbers of observations.

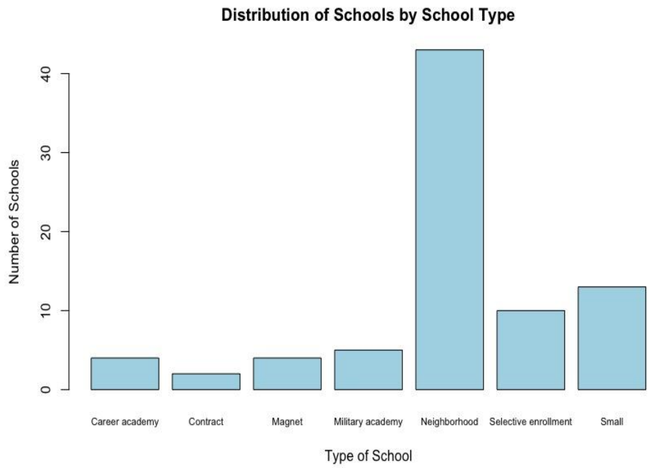

Figure 1. Distribution of Schools by School Type. Figure 1 is a bar plot depicting the breakdown of the number of schools per each school type. Once again, it can be seen that neighborhood schools are the best represented in this data set.

Student Attainment Rating	Above Average	Average	Below Average	Far Above Average	Far Below Average
Number of Schools	3	12	40	5	21

Table 2. Summary Statistics of Student Attainment Ratings. As shown in Table 2, the summary statistics of student attainment ratings revealed the rating below average had the most number of schools and far below average had the second highest number of schools. Additionally, only 8 schools had a rating of either above average or far above average.

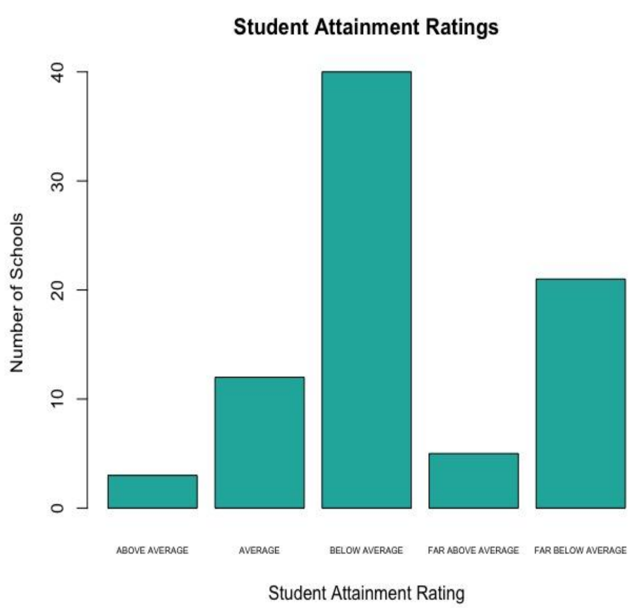

Figure 2. Bar Plot of Student Attainment Ratings. Figure 2 shows the number of schools for each possible rating of student attainment ratings. It clearly shows how the number of schools rated as being below average and far below average predominate over the other ratings.

School Survey - Supportive Environment	Neutral	Not Enough Data	Strong	Very Strong	Weak
Number of Schools	38	1	29	6	7

Table 3. Summary Statistics of School Survey - Supportive Environment Ratings. Table 3 depicts the number of schools for each rating of the school in terms of providing a supportive environment. The statistics reveal that a large number of the schools were rated as being neutral for this metric with the next highest number of schools being rated as strong. It is also important to note that there was one school that did not have enough data for this metric.

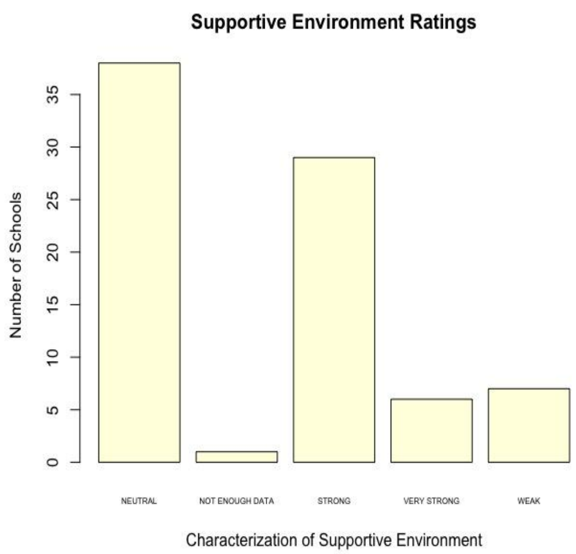

Figure 3. Bar Plot of Supportive Environment Ratings. Figure 3 contains a bar plot that depicts the statistics reported in Table 3. Once again, it can be seen that the majority of the schools are rated as being either neutral or strong in regards to the metric of being a supportive environment.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0.50	6.70	14.20	27.74	34.90	193.40

Table 4. Summary Statistics of Student Suspensions Per 100 Students in 2016. The summary statistics of student suspensions per 100 students of the school reveal that the data may be skewed to the right as the mean, 27.74, is larger than the median, 14.20. Also, it is important to

note that these numbers are presented as percentages which explains the large value of the maximum.

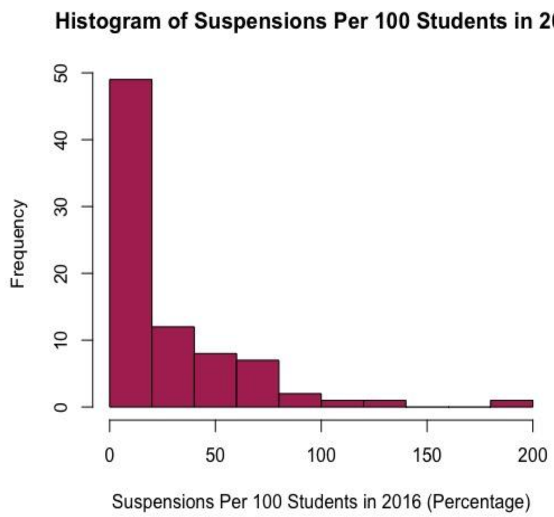

Figure 4. Histogram of Suspensions Per 100 Students in 2016. Figure 4 is a histogram of the suspensions per 100 students in 2016. The histogram reveals that the distribution of the data for this variable is unimodal and right-skewed, and thus, it is likely that the data does not follow a normal distribution.

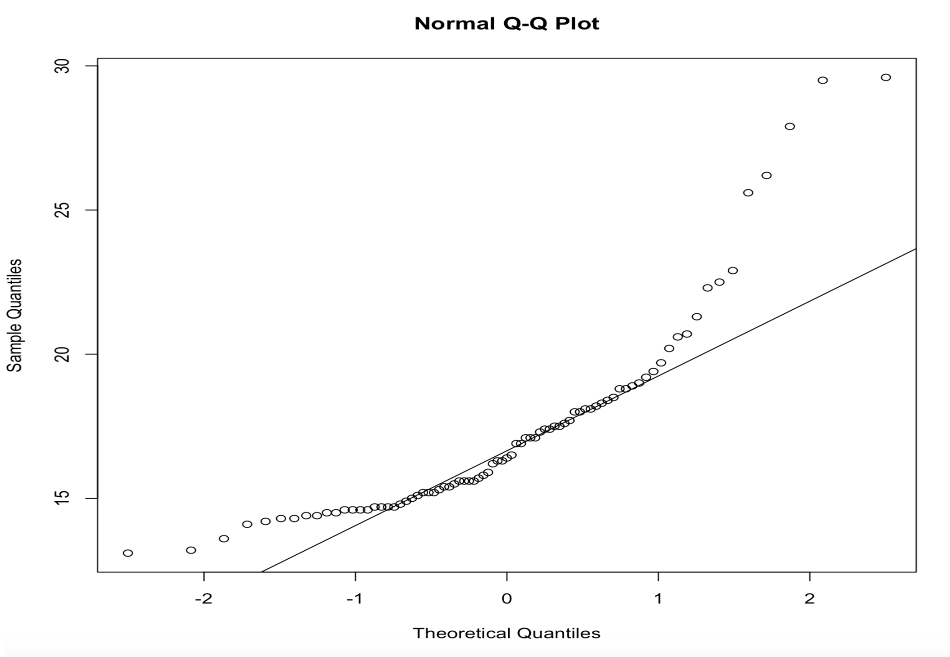

Figure 5. Q-Q Plot for Assessing Normality of Suspensions Per 100 Students. To assess the normality of the distribution for suspensions per 100 students, we used both a Q-Q plot and Shapiro-Wilk test. The data points on the Q-Q plot do not follow the Q-Q line. Additionally, the results of the Shapiro-Wilk test indicate the data does not come from or follow a normal

distribution ($P < 0.05$). Thus, with the results of Q-Q plot along with the Shapiro-Wilk test, we can conclude that it is very likely that our data does not come from a normal distribution.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
74.30	85.90	89.30	88.38	92.60	97.00

Table 5. Summary Statistics of Student Attendance Rates in 2016. It is important to note that the data for this variable is reported as percentages. Since the median and mean are relatively close to one another, looking solely at the summary statistics, there does not seem to be a skew in either direction. The values for student attendance rates range from 74.30% to 97.00%.

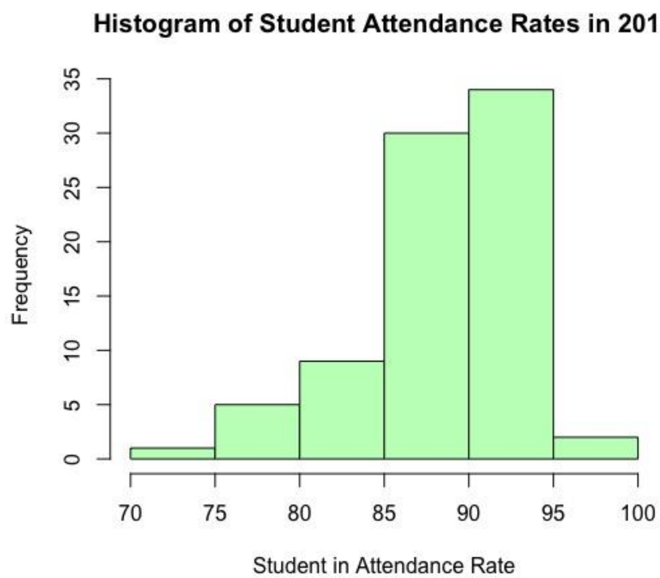

Figure 6. Histogram of Student Attendance Rates in 2016. The histogram reveals that the distribution of data for student attendance rates in 2016 is unimodal and slightly left skewed. The frequencies of attendance rates that are between 85% and 90% as well as 90% and 95% are the two highest frequencies.

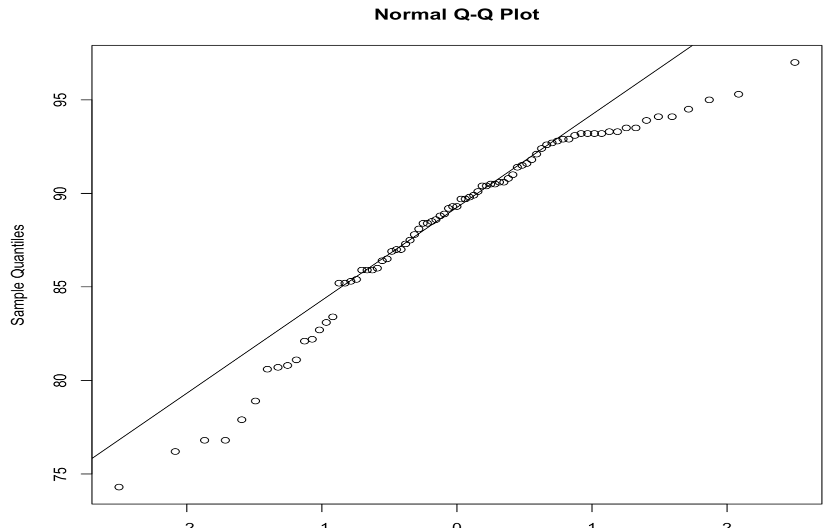

Figure 7. Q-Q Plot for Assessing Normality of Student Attendance Rates in 2016. To assess the normality of the distribution for student attendance rates, we used both a Q-Q plot and Shapiro-Wilk test. The data points on the Q-Q plot stray away from the Q-Q line. Additionally, the results of the Shapiro-Wilk test indicate the data does not come from or follow a normal distribution ($P < 0.05$). From the results of the Q-Q plot along with the Shapiro-Wilk test, we can conclude that it is very likely that our data does not come from a normal distribution.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
31.40	58.90	73.00	72.14	83.80	97.60

Table 6. Summary Statistics of 4 Year Graduation Rates in 2016. It is important to note that the data for this variable is reported as percentages. The median and mean are relatively close to one another; thus, looking solely at the summary statistics, there does not seem to be a skew in the data in either direction. The values for the 4 year graduation rate range from 31.40% to 97.60%.

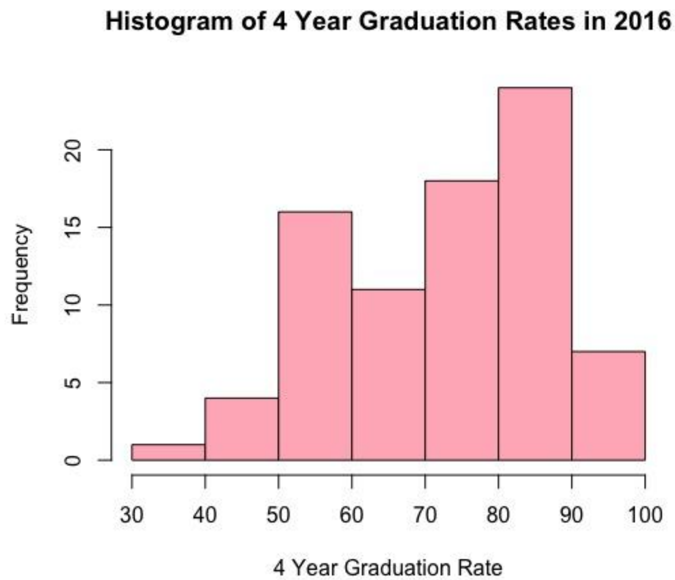

Figure 8. Histogram of 4 Year Graduation Rates in 2016. The histogram of the data for 4 Year Graduation Rates reveals a unimodal, almost symmetrical distribution. The histogram also reveals that graduation rates between 80 and 90% occur with the highest frequency.

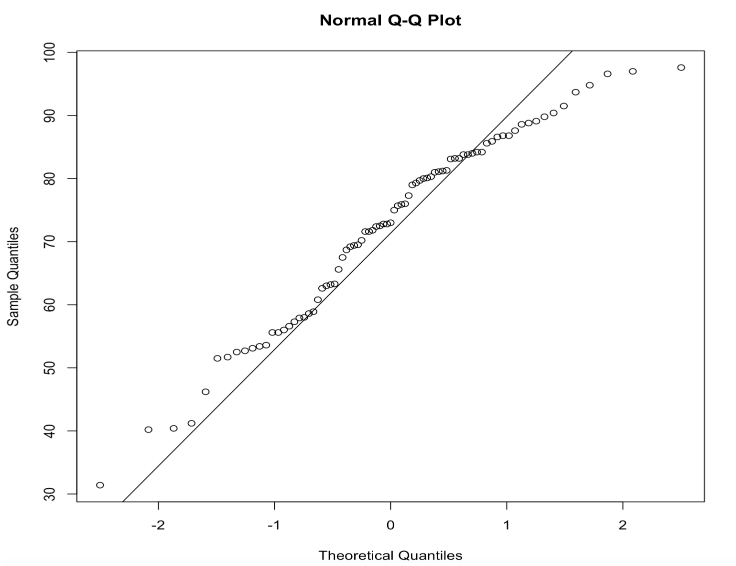

Figure 9. Q-Q Plot for 4 Year Graduation Rates in 2016. To assess the normality of the distribution for 4 year graduation rates, we used a Q-Q plot along with the Shapiro-Wilk test. The data points on the Q-Q plot do not follow the Q-Q line and instead stray away from it towards the maximum and minimum values. Additionally, the results of the Shapiro-Wilk test indicate the data does not come from or follow a normal distribution ($P < 0.05$). Due to our Q-Q plot and the results of the Shapiro-Wilk test, we can conclude that it is very likely that our data does not follow a normal distribution.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
13.10	14.90	16.40	17.36	18.40	29.60

Table 7. Summary Statistics of Average Composite ACT Scores. Once again, the summary statistics reveal that the median and mean are relatively close to one another in value. There is a wide range of values for the average composite ACT scores from 13.10 to 29.60.

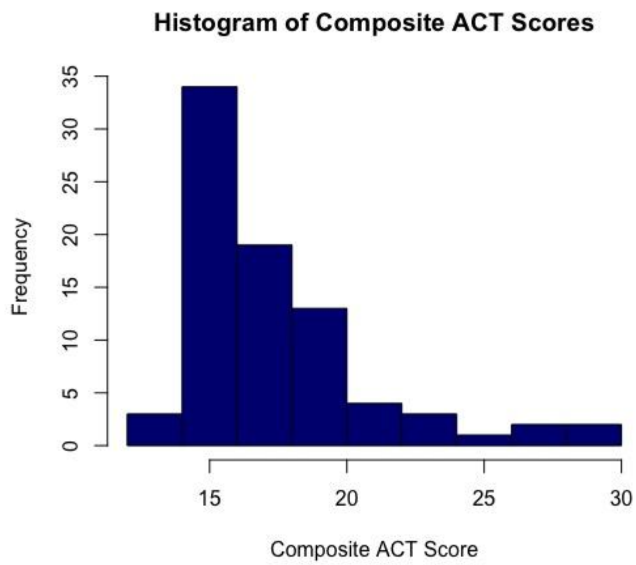

Figure 10. Histogram of Composite ACT Scores. The histogram reveals that the data for the average composite ACT scores is unimodal and left-skewed. There is a wide range of the values for the average composite ACT score, with the majority of the data points being around an average score of 15.

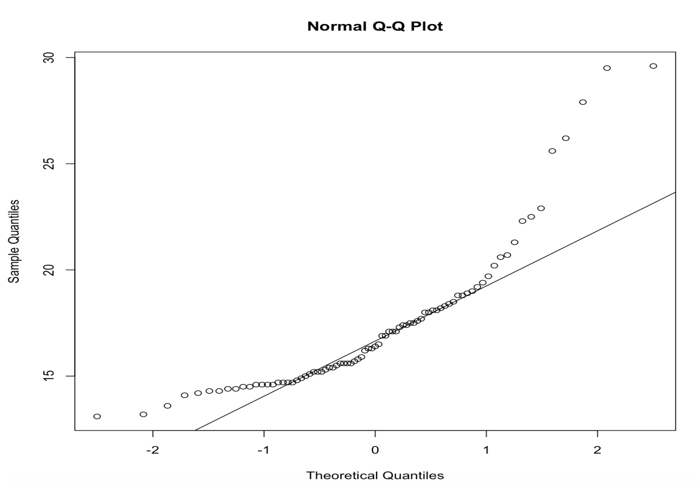

Figure 11. Q-Q Plot for Average Composite ACT Scores. To assess the normality of the distribution for average composite ACT scores, we used a Q-Q plot in addition to a Shapiro-Wilk test. The data points on the Q-Q plot do not follow the Q-Q line and instead stray away from it especially towards the maximum value. Additionally, the results of the Shapiro-Wilk test indicate the data does not come from or follow a normal distribution ($P < 0.05$). Due to the generated Q-Q plot and the results of the Shapiro-Wilk test, we can conclude that it is very likely that our data does not follow a normal distribution.

```
	One Sample t-test

data:  projectdata$Graduation_4_Year_School_Pct_Year_2
t = -0.80714, df = 80, p-value = 0.422
alternative hypothesis: true mean is not equal to 73.5
95 percent confidence interval:
 68.79365 75.49030
sample estimates:
mean of x
 72.14198
```

```
	One Sample t-test

data:  projectdata$Graduation_4_Year_School_Pct_Year_2
t = -7.4044, df = 80, p-value = 1.174e-10
alternative hypothesis: true mean is not equal to 84.6
95 percent confidence interval:
 68.79365 75.49030
sample estimates:
mean of x
 72.14198
```

Figure 12. Statistical Analysis between CPS and National Graduation Rate.
Null Hypothesis (H_0): The population mean graduation rate is 73.5.
Alternate Hypothesis (H_1): The population mean graduation rate is not equal to 73.5.

As shown in Figure 12, we did two different significance tests to determine whether the average graduation rate of the schools nationally would differ from the reported average graduation rate of Chicago Public High Schools (CPS). The reported average graduation rate is 73.5% for CPS and 84.5% for other schools nationally. The t.test provided on the left in the figure above was our test done on CPS schools and because it generated a P-Value of greater than 0.05 ($0.422 > 0.05$), then there is insufficient evidence that the mean graduation rate of the schools in our dataset differs from the reported average 4 year graduation rate of 73.5. We can then conclude that there is no difference in mean graduation rates from the CPS reported average graduation rate. However, the t.test done on the right in the figure above was our test done on national graduation rates. To mention again, the national average graduation rate is 84.5% and the test shown is that because our p-value is significantly lower than 0.05 ($1.174e\text{-}10 < 0.05$), we reject the null hypothesis that the average graduation rate of our sample of CPS high schools is the same as the national average graduation rate. From the tests and results provided, we can conclude that the average graduation rate of CPS high schools differs from the national graduation rate.

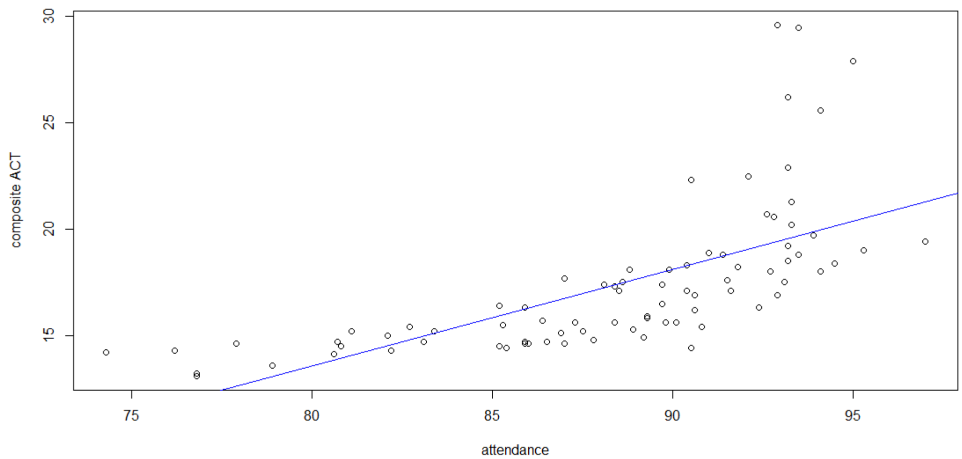

Figure 13. Linear association between ACT Scores and Attendance Rates
Here is the scatter plot and the regression line where the plot seems to show that there is a positive linear association between both composite ACT scores and percentage of students who attended the whole school year.
We found that the coefficient r = 0.65, which implies that there might be a strong positive linear association between the two variables. In order to be confident that r was nonzero, we performed the cor.test() and found that $p < 0.05$ so we accepted the alternative hypothesis that rho was non-zero. We also found the regression line to be the equation $y = 0.454x - 22.79$. We wanted to check if the regression line was appropriate and if R^2 is close 1 to see if the model fits the data well. We plotted the residuals and found that they were somewhat randomly scattered and somewhat equally spread out.

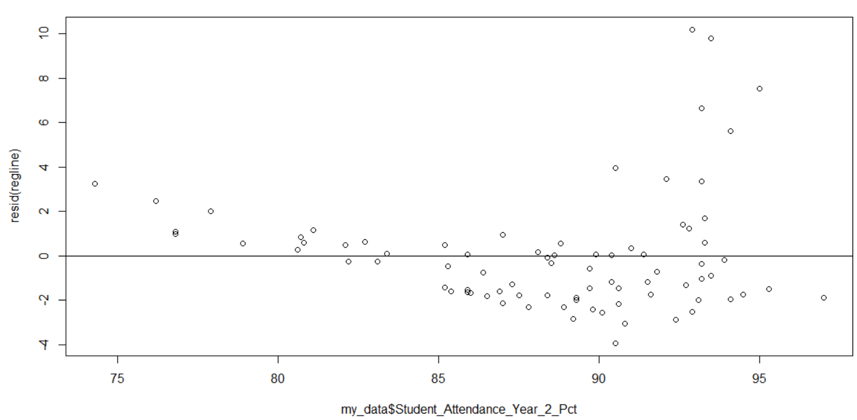

Figure 14. Regression Line Data between ACT Scores and Attendance Rates

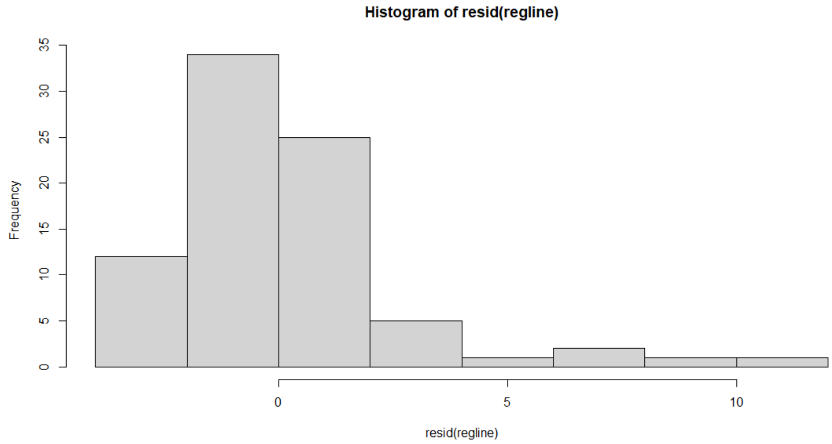

Figure 15. Histogram of Regression Line Data

However, we found that the histogram was not normal.

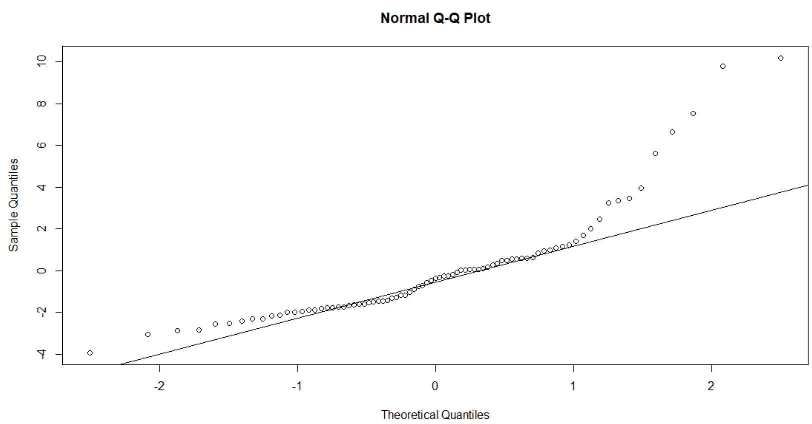

Figure 16. QQ Plot of Regression Line Data

And a good chunk of the data was not on the qqline.

We also tested the residuals for normality using the shapiro test and found that the p-value <0.05 and we also computed R^2 and found that it was 0.43.

The results of trying to find a linear association between both composite ACT scores and percentage of students who attended the whole school year ended showing that there was strong positive linear association since r =0.65 which was close to 1. So it did validate our hypothesis (there was a positive linear association between the two variables). Since the residuals were not normally distributed and R^2 was also close to zero, we concluded that the regression line may not be appropriate and the model did not fit the data well.

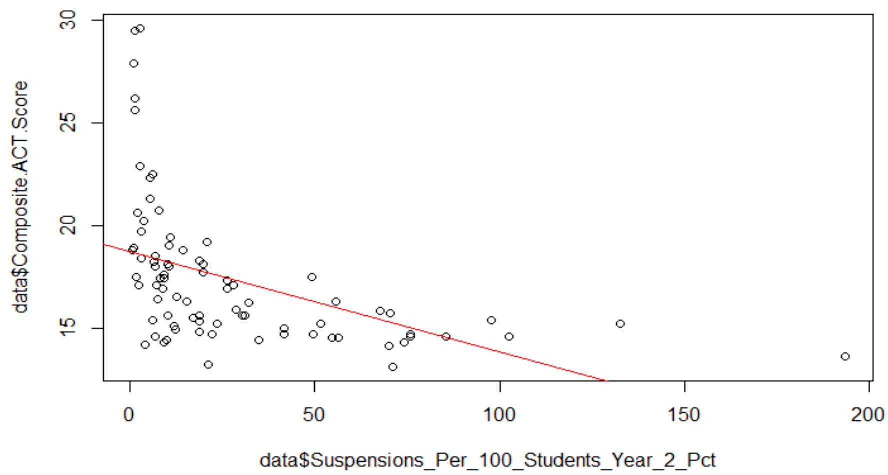

Figure 17. Linear association between ACT scores and Suspension Rates
The scatter plot of residuals shows that residuals are randomly scattered, however, they are not equally spread out.

A scatter plot showing the relation between the ACT Scores and Student Suspension. ACT Scores is the dependent variable and student suspensions is the independent variable. Computing the r, gets us r = -0.47249. This explains that r is negative and is closer to 0 so we know that the linear association is negative and is probably weak. Using cor.test then gives us a confidence interval and checks whether rho is non zero which we find out when we get our $p < 0.05$. We then find the equation of the regression line by lm(y~x). The regression line is $y = -0.04887x + 18.71965$.

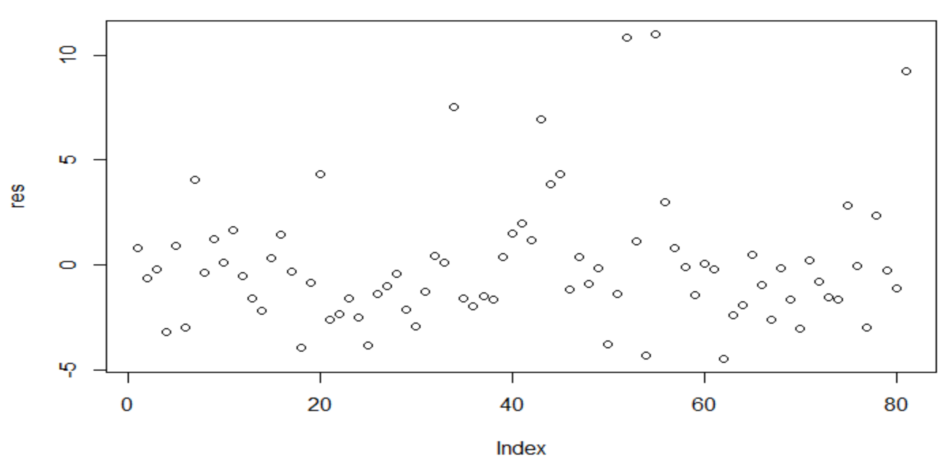

Figure 18. Regression Line Data between ACT Scores and Suspension Rates

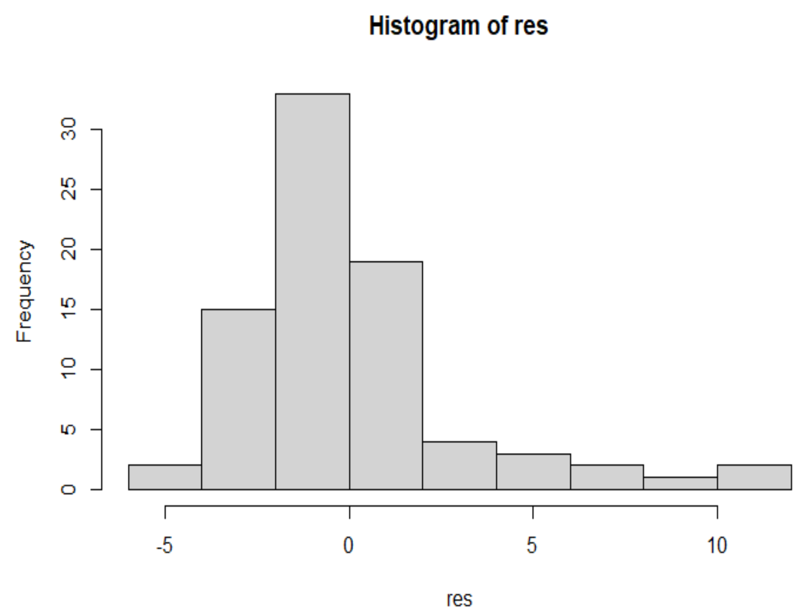

Figure 19. Histogram of Regression Line Data
We then draw a histogram to check the normality of the residuals and we can tell that it is not normally distributed.

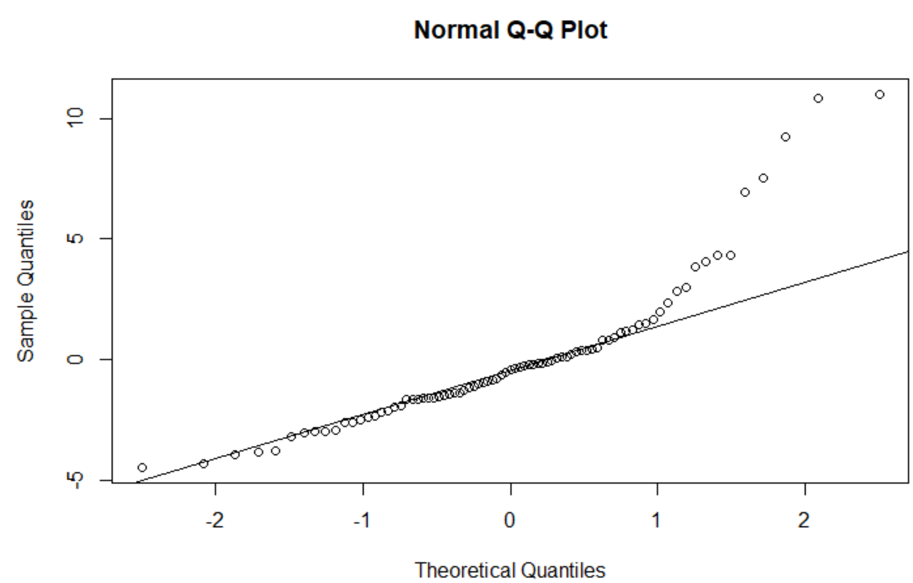

Figure 20. Histogram of Regression Line Data
Next, we check the normality using qqline and qqnorm and looking at it can tell that it is not normal.

Lastly, we test the residuals for normality using the Shapiro test and we get the p-value = 1.521e-07 < 0.05. So we conclude that the residuals are not normal and that the regression line may not be appropriate.

Conclusion: Finding the linear regression with ACT Scores and Student Suspension helped us find the relationship between the two variables. We have a negative weak linear association between the two. We found the equation of the regression line and also determined that the residuals are not spread out normally. Finally, we also computed the R^2 and found that it was 0.22 which is closer to 0, so we know that the model did not fit the data well.

Examining Differences in Mean Composite ACT Scores Between School Types

One of our aims was to examine whether there is a difference in the average composite ACT scores between different school types. In order to determine whether there was a difference in the composite ACT scores across the seven school types (neighborhood, small, selective enrollment, military academy, magnet, career academy, and contract schools), we used a one-way ANOVA test. The results of the ANOVA test indicated that there was a significant difference in the mean composite ACT scores amongst the groups of schools by school type ($F(6,74) = 18.95$, $p < 0.001$). It was found that at both a 0.05 and 0.001 significance level, there was a difference between the mean composite ACT scores between the different types of schools. A post-hoc Tukey test revealed that at a 0.05 significance level there were significant differences in the mean composite ACT scores between enrollment schools and each of the six other types of schools. In particular, the Tukey test revealed significant differences in the mean ACT scores between the following pairs of school types: selective enrollment and career academy, selective enrollment and contract, selective enrollment and magnet, selective enrollment and military academy, selective enrollment and neighborhood, and small and selective enrollment. For each pair, the 95% confidence intervals generated indicate that the mean composite ACT score of the selective enrollment schools is larger in comparison to the mean composite ACT scores of the other 6 types of schools.

```
            Df Sum Sq Mean Sq F value  Pr(>F)    
School_Type  6  587.2   97.86   18.95 3.1e-13 ***
Residuals   74  382.1    5.16                    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 21. Results of the ANOVA Test for Differences in Average ACT Score Between Different School Types. This figure summarizes the results of the ANOVA test that was performed to determine whether there is a difference in the average composite ACT score between different school types. There was a significant difference between the average composite ACT score between the school types ($F(6,74) = 18.95$, $P < 0.05$).

```
$School_Type
                                           diff        lwr        upr     p adj
Contract-Career academy              1.42500000  -4.539415  7.3894151 0.9906685
Magnet-Career academy                2.72500000  -2.144925  7.5949245 0.6208524
Military academy-Career academy      2.06500000  -2.555016  6.6850161 0.8234906
Neighborhood-Career academy         -0.01220930  -3.612370  3.5879518 1.0000000
Selective enrollment-Career academy  8.16500000   4.090529 12.2394712 0.0000010
Small-Career academy                -0.05961538  -3.997473  3.8782423 1.0000000
Magnet-Contract                      1.30000000  -4.664415  7.2644151 0.9942967
Military academy-Contract            0.64000000  -5.122172  6.4021724 0.9998757
Neighborhood-Contract               -1.43720930  -6.419101  3.5446822 0.9752888
Selective enrollment-Contract        6.74000000   1.405265 12.0747350 0.0047766
Small-Contract                      -1.48461538  -6.715753  3.7465227 0.9772395
Military academy-Magnet             -0.66000000  -5.280016  3.9600161 0.9994651
Neighborhood-Magnet                 -2.73720930  -6.337370  0.8629518 0.2558573
Selective enrollment-Magnet          5.44000000   1.365529  9.5144712 0.0023303
Small-Magnet                        -2.78461538  -6.722473  1.1532423 0.3389684
Neighborhood-Military academy       -2.07720930  -5.331367  1.1769486 0.4645449
Selective enrollment-Military academy 6.10000000  2.327773  9.8722273 0.0001083
Small-Military academy              -2.12461538  -5.748854  1.4996234 0.5676379
Selective enrollment-Neighborhood    8.17720930   5.759295 10.5951238 0.0000000
Small-Neighborhood                  -0.04740608  -2.227250  2.1324375 1.0000000
Small-Selective enrollment          -8.22461538 -11.121488 -5.3277423 0.0000000
```

Figure 22. Results of the Post-Hoc Tukey Test. This figure contains the results of the post-hoc Tukey test that was conducted after the ANOVA test. It consists of 95% confidence intervals calculated for each pair of school types that were compared. Of the confidence intervals, the following confidence intervals indicated there was a significant non-zero difference in the average composite ACT scores between each pair: selective enrollment - career academy, selective enrollment - contract, selective enrollment - magnet, selective enrollment - military academy, selective enrollment - neighborhood, and small - selective enrollment ($P < 0.05$). For each of these pairs, the corresponding 95% confidence intervals revealed that the average composite ACT scores of selective enrollment schools is greater than the average scores of each of the other school types.

Discussion

Through our statistical analysis and the results of our various statistical tests, we were able to answer the five prominent questions we discussed in the introduction of this report. The results of our t tests indicated that there is no significant difference between the average four year graduation rate of the CPS high schools represented in our data set and the average graduation rate reported by CPS. Additionally, the results of our t tests also show that there is a significant difference between the average four year graduation rate of CPS high schools and the national graduation rate. Examining the linear associations between variables, the results of our linear regression models and running correlation tests indicated there was a moderate positive linear association between attendance rates and average composite ACT scores. However, the results of our linear regression also indicated that the suspension rate does not correlate with the average ACT scores of the schools. The results of the ANOVA test we performed indicated that the mean composite ACT scores of schools do differ significantly between the seven different school types; the Post-Hoc Tukey test revealed that selective enrollment schools outperformed the other types of schools with a higher average composite ACT score.

To conclude, our group would like to summarize our initial thoughts before we analyzed the data and had our final results. In our linear regression analysis, we expected to see a strong positive correlation between ACT scores and attendance rates and a strong negative correlation between ACT scores and suspensions. Our initial thoughts were accurate for the positive correlation between ACT scores and attendance rates, however the result between ACT Scores and suspensions showed no correlation. In terms of the different types of schools analyzed, our prediction constructed was that there was going to be a difference in ACT scores. Our group expected to see differences between many of the different types of schools; however, the results primarily showed differences between school types solely due to selective enrollment schools. This may have been because some schools such as magnet schools or military academy schools were not well represented. Finally, our group assumed that there would be a difference among graduation rates between the CPS and national average. We expected the average graduation rate to not be different when analyzing CPS data and in comparison to the national average, there would be a significant difference. Overall, our analysis and results were to prove our initial hypothesis that there was a difference in average graduation rates between CPS and national data. For further analysis and to expand on our findings, it would have been critical to consider the average ACT Scores in comparison to Graduation Rate or College Enrollment. By researching this aspect, it would have helped us understand the influence in graduation rates and the impact of college enrollment. Overall, our group was able to correctly validate most of our initial claims and even more importantly answer all the questions we aimed to learn more about.

References

Chicago Public Schools. Average ACT Scores [Data file]. Retrieved from https://www.cps.edu/about/district-data/metrics/assessment-reports/.

Chicago Public Schools. (2017). Chicago Public Schools - School Progress Reports SY1617 [Data file]. Retrieved from https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Progress-Reports-SY1/cp7s-7gxg.

Data: U.S. Graduation Rates by State and Student Demographics. (2017, December 7). Retrieved December 04, 2020, from https://www-edweek-org.proxy.cc.uic.edu/ew/section/multimedia/data-us-graduation-rates-by-state-and.html.