Michelle Weng

NYC Districts Regent Scores Pre vs Post-Covid

Abstract

Covid-19 had a huge impact on children, especially on education. Families in poor communities might not have the resources and technology for their child to succeed in learning while at home. These families also might have had a rough time dealing with finances and school wouldn't have been a priority for them. Using data from the New York State Education Department on all the districts in NYC, compare scores prior to covid-19 to scores after for each district to see if there's a significant difference between poor and rich districts and push for more funding in poor communities.

Hypothesis / Goal

I hypothesize economically worse off schools would have less funding since NYC schools get some of its funding directly from its neighborhood property tax. Therefore, if a neighborhood is a slum and the buildings there aren't worth as much, the school would get less funding as the government would collect less property taxes.

Data Collection

Find data on the regents scores scored and funding each school gets for the 32 NYC School Districts. Find which district is worse/better off economically and then compare the regent scores before and after Covid-19.

3-8 Assessment Database from the 2018-19 and 2020-21 school years
https://data.nysed.gov/downloads.php.
Contained comprehensive 3-8 ELA and Math Researcher Files, with district, public, and charter school level enrolled, tested, and not tested aggregate data results and another files called 3-8_ELA_AND_MATH_NYC_SUMMARY_2019/21 which gives the summary for NYC schools. Dataset info: Regent scores range from levels 1 to 4. Level 3 is passing while level 4 is excellent. Level corresponds to the regent score and there is a column to count the number of students who scored in that level along with the percentage out of the total who took that test in that school. The dataset also includes data for the whole district meaning all the schools in a said district.

2019 NYC school transparency funding data for each individual school (Excel):
https://www.budget.ny.gov/schoolFunding/2021/new-york-city/index.html
The 2020-21NYC school transparency excel file contains 5 more sheets about

1. Total funding for all New York City schools
2. Basic school info such as enrollment and staff numbers, school status, etc.
3. Basic school allocations along with funding per pupil
4. School spendings on Pre-K and after school programs
5. Locally implemented funding formula

Out of these sheets, Sheet 2 and 3 are the most useful.

Data Cleaning

These datasets are provided in .xlsx format so I installed the readxl package and used read_excel(<filename>) and set the working directory to the project folder to import the data into a R data frame.

The 2018-19 school year columns removed:

- SCHOOL YEAR END DATE: Redundant info, all observations in the table have the same date.
- STUDENT SUBGROUP: All observations in the table have the same subgroup of All Students.
- LEVEL 2-4 PCT: This percentage is insignificant to this problem.
- NAME: Name of the school or name of a NYC school district

The 2019-21 school year columns removed:

- SCHOOL YEAR END DATE: Redundant info, all observations in the table have the same date.
- STUDENT SUBGROUP: All observations in the table have the same subgroup of All Students.
- NAME: Name of the school or name of a NYC school district

DATA CLEANING 1: Data transformation, character to numeric

I noticed that all the entries are of the character type. Using the mutate_at function, I did data transformations on TOTAL ENROLLED, TOTAL NOT ENROLLED, the count for the number of students who scored in each level 1-4, and the MEAN SCALE SCORE columns to transform them from character type to numeric type. This function also automatically sets the '-' value in the table to NA.

DATA CLEANING 2: Data transformation, character to decimal

I also wanted to convert all the percentages to decimal such as 72% to 0.72 as this will work well with the summary function. To do this, I wrote a function that checks if a cell character ends with a % character. I used the grepl() function which takes a pattern and returns TRUE if a string contains the pattern, otherwise FALSE. The pattern is a regex and I used %$ which checks if a string ends with %. I paired this function with the mutate_if function to convert all the cells with characters ending with % to a decimal. This also sets the '-' values in the rest of the table to NA.

```
is.percentage <- function(x) any(grepl("%$", x))
nyc_2019 <- nyc_2019 %>% mutate_if(is.percentage, ~as.numeric(sub("%", "", .))/100)
```

DATA CLEANING 3: Remove NA

In the 2018-19 school year, I decided to remove the observation that contained NA. This is because all the observations that had NA had no information on TOTAL TESTED so I will be assuming that nobody in that school took that specific test. Also a majority of these observations have Grade 8 Math as the subject and from my experience I believe that the 8th grade math regent is not a mandatory test and students can opt to take the NYC Algebra 1 Regents instead. Using the *na.omit()* I removed 24 entries and should not significantly impact the dataset.

I have also decided to remove observations containing NA from the 2020-21 dataset as many schools did not require their students to take the regent in 2021. There were many schools that were lenient as the DOE did not require students to take it and allowed them to pass. The dataset in the coming years will definitely contain way more data as currently 2962 observations are NA compared to 24 from 2018-19 and not to mention that observations without NA may contain data on top students as it is more likely that parents who are invested in their child education would push them to take the regents even if it is not mandatory.

DATA CLEANING 4:

Removing the school name as it is inconsistent between documents. BEDS CODE is the better unique identifier for schools so I created a data table that contains Beds Code(primary id), School Name, District #. Use the 2020-21 assessment dataset for this since it includes all the schools that still exist after the pandemic. The steps:

1. Copied the 2020-21 assessment dataset.
2. Keep only the BEDS CODE and NAME columns and remove duplicates.
3. This excel sheet is ordered by the BEDS CODE and starts with the “NEW YORK CITY GEOGRAPHIC DISTRICT # 1” and every school below this row is a part of District 1 until we get to the row where it’s NAME is “NEW YORK CITY GEOGRAPHIC DISTRICT # 2”, then every school below that will be part of district 2 and so on. To do so, I had to use a loop function and using an if else statement, checked if a row contained “NEW YORK CITY GEOGRAPHIC DISTRICT”, and stored it in the variable district. If a row NAME doesn’t match then we update its DISTRICT to the district variable.
4. Rename the new third column to DISTRICT.
5. Then saved into the schools.xlsx file.

DATA CLEANING 5:

For now, I’m not exacly sure what data is the most important or is useful in the budget dataset pages 2 and 3. They both contain alot of detailed information, for example in Sheet 3 it includes where they use their fundings on. Currently, the best thing to do is to remove all the schools from Sheet 2 and 3 that don't serve grades 3-8 since that is what we care about. To do this first, since the data wasn’t imported correctly, I had to extract the column names and rename the table. Then, I used the *merge()* to inner join the SCHOOLS dataframe with sheets 2 and 3 by BEDS CODE. This removed 446 entries from both sheets.

Saving to File

After performing the data cleaning, I saved the files using write.xlsx(), so the data cleaning will be preserved.

EDA

1. Box plot to show the distribution of funding per pupil for each district. Wanted to see which district has the most/least funding utilizing the budget_info data table. From this, we can see district 2 has the smallest range and mean while district 5 has the largest mean.

Code: `` boxplot(`Total School Funding per Pupil` ~`DISTRICT`, data=budget_info) ``

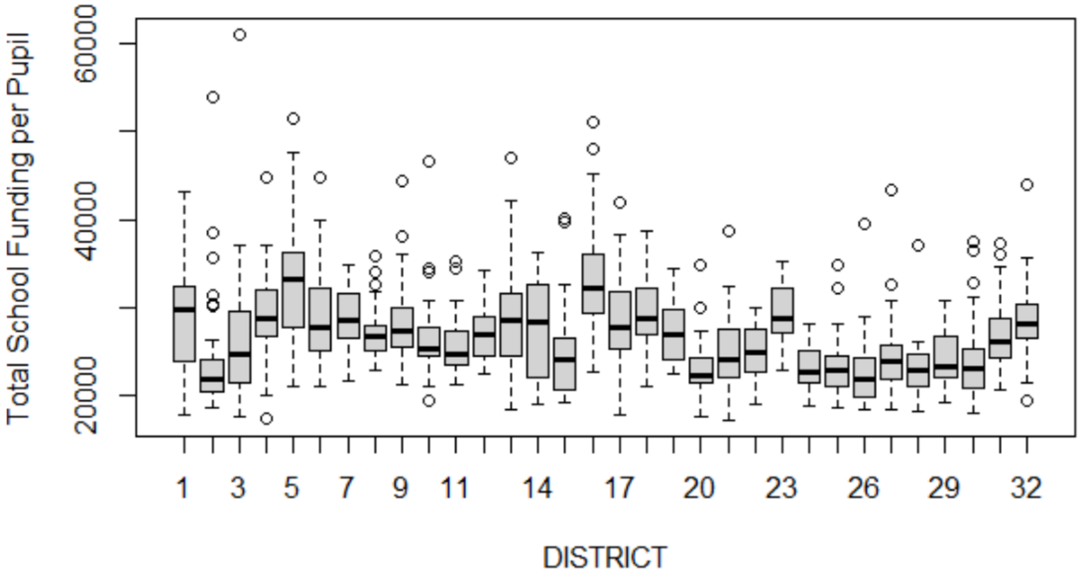


Figure 1. Total School Funding per Pupil in Every District

2. 2019 NYC Regent Score scatter plot based on funding per pupil. Wanted to see if there is any correlation between mean scale score and funding per pupil. Merged 2019 regent score data set and budget_info data set on beds code then display it using ggplot. The scatter plot shows that funding per pupil might not fully correlate to the mean scale score. In the plot below, we can see that the higher mean scale scores almost all have a Total Funding per Pupil under $30,000. Showing that total school funding per pupil may not have any direct impact on the regent score.

Code:
```
ggplot(pupil_mean) + geom_point(mapping = aes(x=`MEAN SCALE SCORE`,
y=`Total School Funding per Pupil`))
```

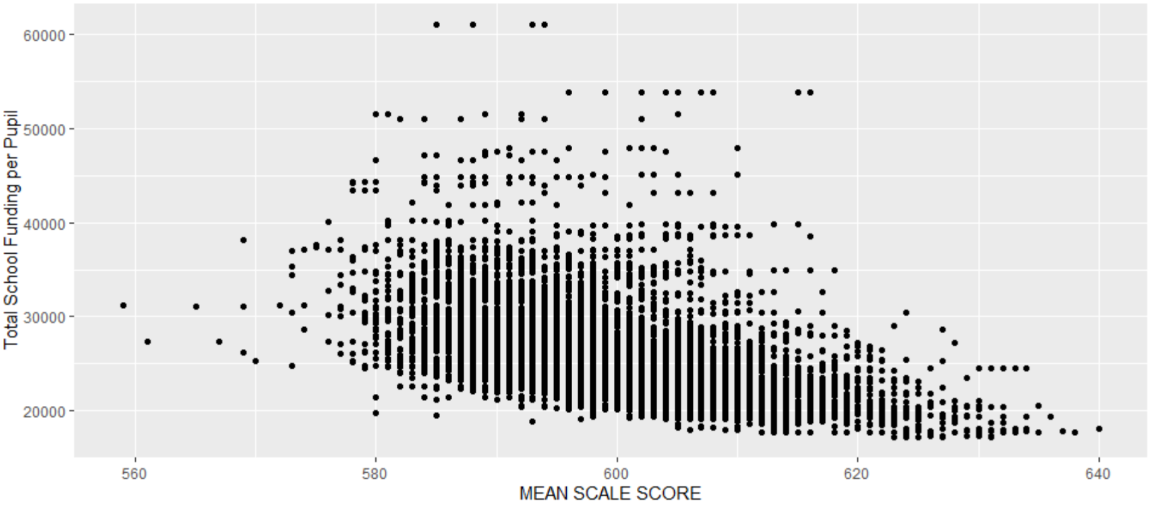


Figure 2. 2019 NYC Regent Score scttaer plot based on funding per pupil

3. 2019 and 2021 NYC Regent Score box plot based on District. In Figure 1, we saw District 2 has the lowest mean funding. However, we see district 2 has one of the highest mean in comparison to all other districts before and after covid-19, followed by district 26. To do this, I had to add a year column and added 2019 as the year for 2019 regent data and 2021 as the year for 2021 regent data. Then I removed TOTAL NOT TESTED, TOTAL ENROLLED from 2021 data set so I could concat 2021 data to 2019 without any issue. In this new combined data set I am able to graph the mean scale score for 2019 and 2021 side by side as a box plot. From this chart we also see a wider range of scores after covid than before. However, we must note that taking the regent was not mandatory in 2021 so it is not the best indicator about pre covid scores versus post covid scores.

Code:

```
bp_2019_2021 <- ggplot(regent_19_21, aes(x=DISTRICT, y=`MEAN SCALE SCORE`,
     fill=YEAR)) + geom_boxplot()
bp_2019_2021
```

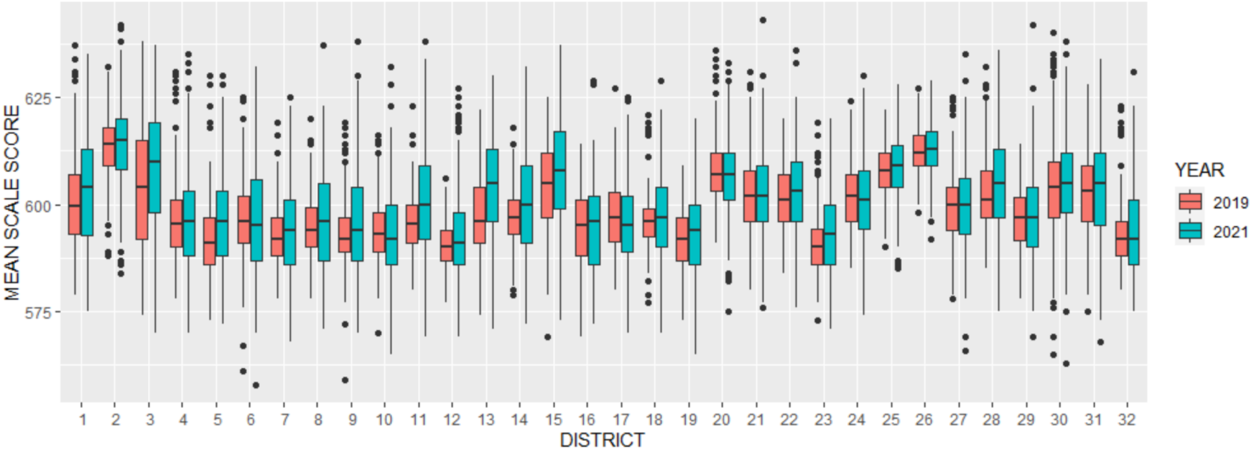
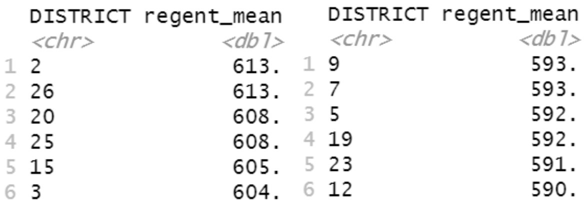


Figure 3. 2019 and 2021 NYC Regent Score box plot based on District

4. Using dplyr library, ordered the district by the highest/lowest mean score for both 2019 and 2021.

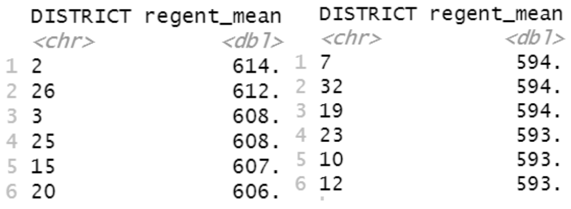
Code:

```
highest_rengent <- regent_2019 %>%
  group_by(DISTRICT) %>% #group by DISTRICT
  summarize(regent_mean = mean(`MEAN SCALE SCORE`, na.rm = TRUE)) %>%
  arrange( desc(regent_mean) )
head(highest_rengent)  ##tail for worst districts
```



```
  DISTRICT regent_mean   DISTRICT regent_mean
  <chr>          <dbl>   <chr>          <dbl>
1 2               613. 1 9               593.
2 26              613. 2 7               593.
3 20              608. 3 5               592.
4 25              608. 4 19              592.
5 15              605. 5 23              591.
6 3               604. 6 12              590.
```

Figure 4. 2019 Top and Bottom Districts



```
  DISTRICT regent_mean   DISTRICT regent_mean
  <chr>          <dbl>   <chr>          <dbl>
1 2               614. 1 7               594.
2 26              612. 2 32              594.
3 3               608. 3 19              594.
4 25              608. 4 23              593.
5 15              607. 5 10              593.
6 20              606. 6 12              593.
```

Figure 5. 2021 Top and Bottom Districts

5. Now the question is why does District 2 have a higher mean test score when they have one of the least funding per pupil in comparison to the other districts? A factor to take into consideration could be the school size. Another factor could be class room size. In the code I found the ratio between number of students and number of classroom teachers and stored it in a new column, "Teacher To Student Ratio". Then I plotted "Teacher To Student Ratio" versus the District. Surprisingly, District 2 and 26 both have a higher student to teacher ratio meaning there are more students than teachers which is the opposite from my hypothesis.

Code:

```
for(i in 1:nrow(staff_info)) {
  teachers <- staff_info[i, "Total Classroom Teachers"]
  students <- staff_info[i, "K-12 Enrollment"]
  staff_info[i, "Teacher To Student Ratio"] <- (students/teachers)
}
boxplot(`Total Classroom Teachers` ~`DISTRICT`, data=staff_info)
```

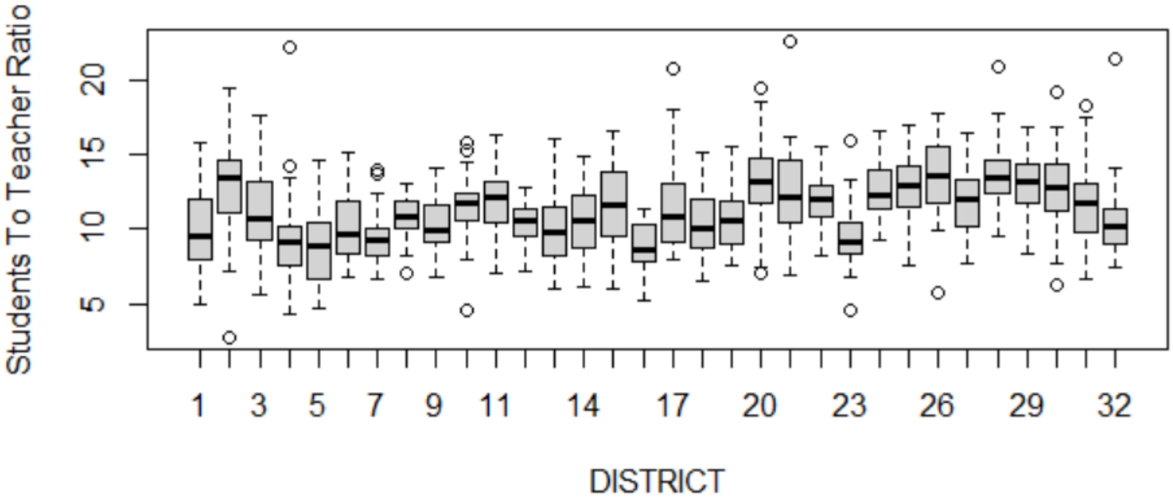


Figure 6. Students to Teacher Ratio

Modeling and Analysis

During EDA I found that the Top School District was District 2 for both 2019 and 2021 even though the Total School Funding per Pupil in both districts was much lower than in any other district. I realized that funding the school funding per pupil doesn't correlate with the student's performance. I decided to use the RESEARCHER_FILE_2019 from the New York State Education Department website. This file had more detailed information regarding students' economic status, gender, ethnicity, etc.

I had to do some data cleaning to retrieve only infomation on NYC schools and districts. I removed some columns that contained redundant information and only changed MEAN_SCALE_SCORE to numeric as that was the only column I needed. Then I saved the files as subgroup_2019 and subgroup_2021.

I performed some EDA by grouping the different sub groups.

1. Ethnicity

Code:

```
ethnicity <- list("04", "05", "06", "07", "08", "09")
subgroups_2021_ethn <- subgroups_2021 %>%
  filter((SUBGROUP_CODE %in% ethnicity))
ggplot(subgroups_2021_ethn, aes(fill=subgroup_name, y=TOTAL_ENROLLED,
x=DISTRICT)) +
  geom_bar(position="fill", stat="identity")
```

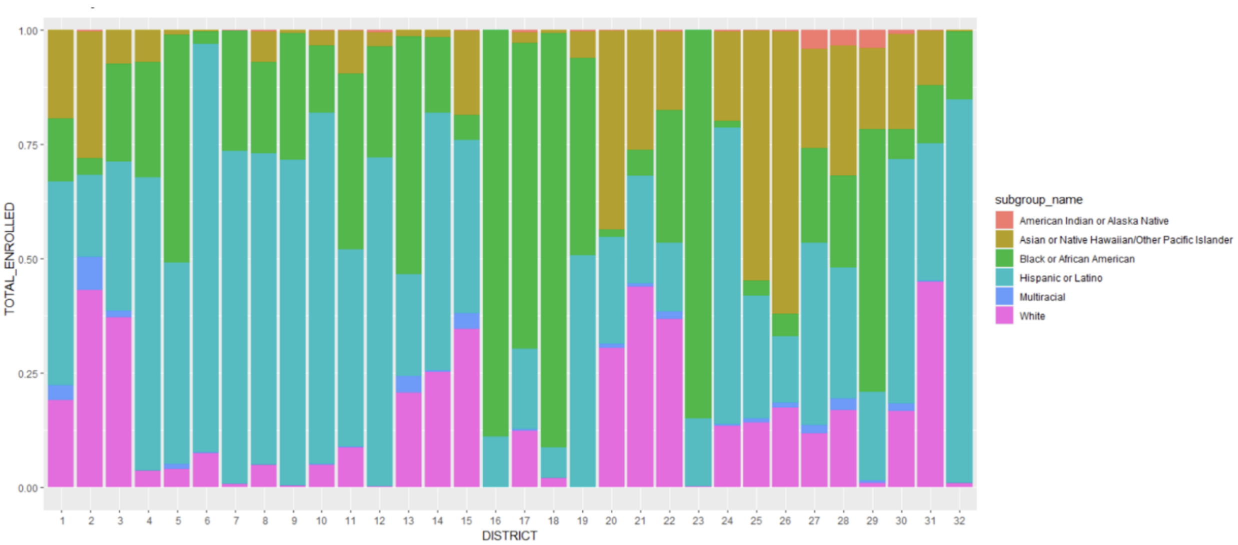


Figure 7. Ethnicity Ratio Across Districts

2. Economically Disadvantaged vs Not Economically Disadvantaged

Code:

```
eco_status <- list("15", "16")
subgroups_2021_status <- subgroups_2021 %>%
  filter((SUBGROUP_CODE %in% ethnicity))
ggplot(subgroups_2021_status, aes(fill= subgroup_name, y =
TOTAL_ENROLLED, x=DISTRICT)) +
  geom_bar(position="fill", stat="identity")
```

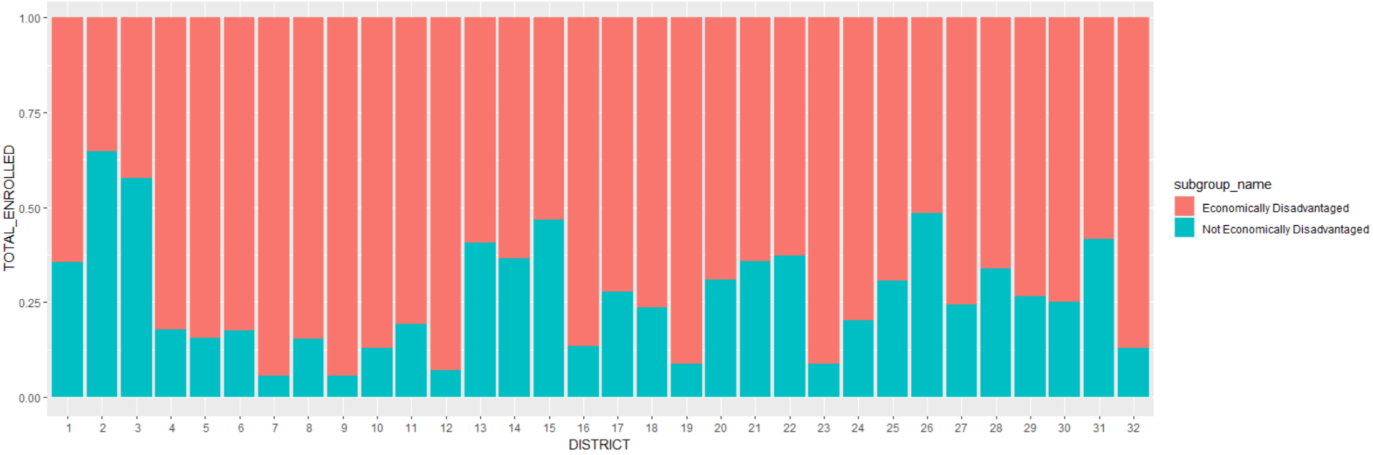


Figure 8. Economically Disadvantaged Ratio Across Districts

3. Economically Disadvantaged vs Not Economically Disadvantaged

Code:

```
homeless <- list("20", "21")
subgroups_2021_homeless <- subgroups_2021 %>%
  filter((SUBGROUP_CODE %in% ethnicity))
```

```
ggplot(subgroups_2021_homeless, aes(fill = subgroup_name, y =
TOTAL_ENROLLED, x=DISTRICT)) +
  geom_bar(position="fill", stat="identity")
```

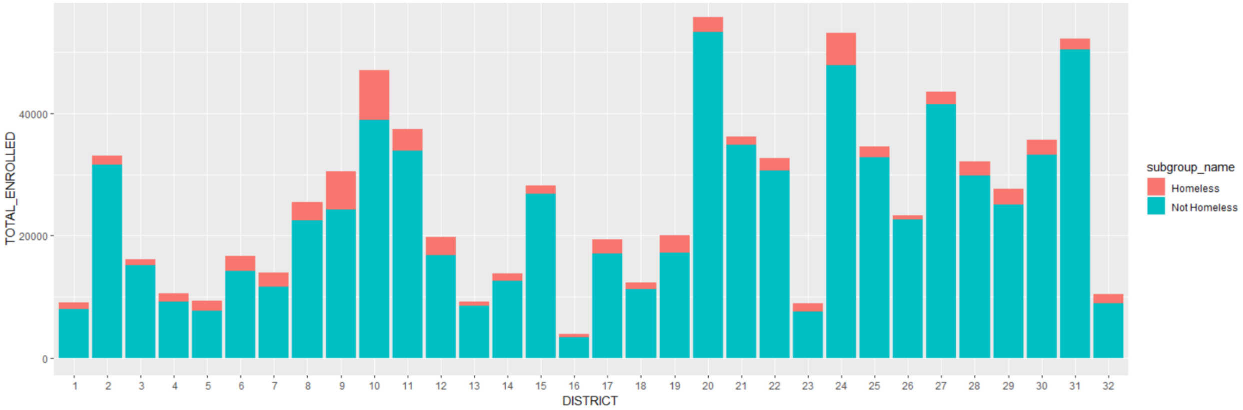


Figure 9. Homeless Ratio Across Districts

After running some EDA, I realized that what made District 2 stand out was that it was the district with the highest % of not economically disadvantaged students to economically disadvantaged students.

I decided to run some logistic regression on the probability the student is of economical disadvtange based on the regent scores. The results were pretty telling.

Logistic Regression: Probability of Being Economically Disadvtange by the Mean Scale Score

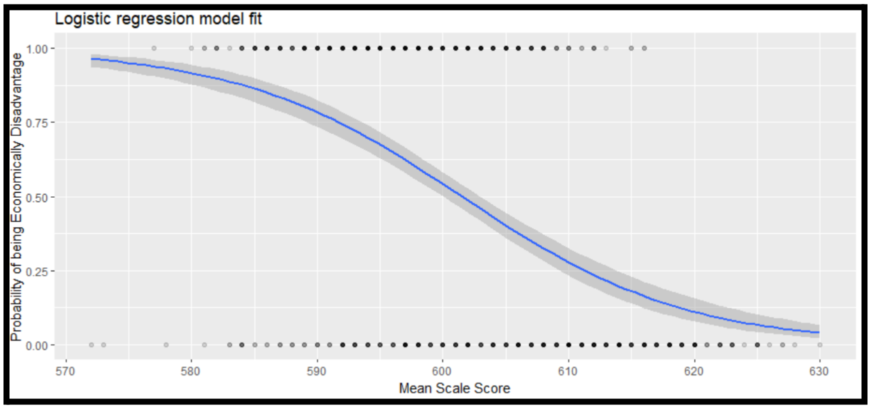


Figure 10. 2019 ED and Regent Score

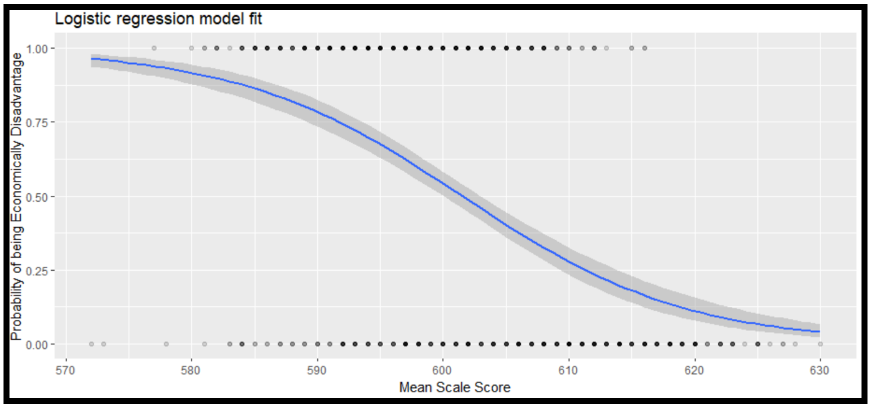
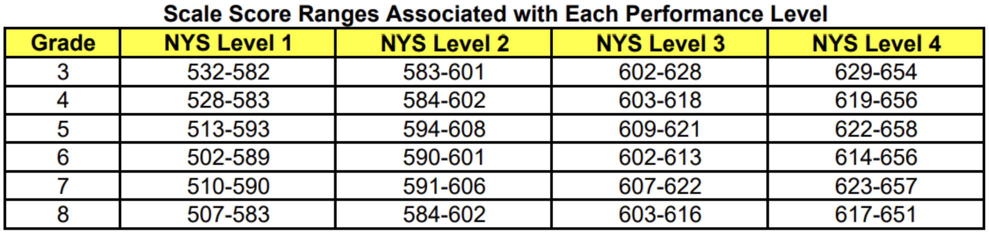


Figure 11. 2021 ED and Regent Score

From the logistic regression, we can see that there is a relationship between a student's score and whether or not they are economically disadvantaged. As the score decreases, there is a higher chance that the students are economically disadvantaged.



Scale Score Ranges Associated with Each Performance Level

Grade	NYS Level 1	NYS Level 2	NYS Level 3	NYS Level 4
3	532-582	583-601	602-628	629-654
4	528-583	584-602	603-618	619-656
5	513-593	594-608	609-621	622-658
6	502-589	590-601	602-613	614-656
7	510-590	591-606	607-622	623-657
8	507-583	584-602	603-616	617-651

Figure 12. NYC Regent Raw Score to Performance Level

Given that a Level 2 is below expectation, I used the calculated logistic regression to get the probability that a student who scored 590 was at an economical disadvantage. The result was pretty bad. There's an 86.18% probability that the students are at an economic disadvantage for 2019 and 77.5% for 2021 if they scored a 590. Given that Level 4 is above standard, I used the calculated logistic regression to get the probability that a student who scored 615 (around 4) and the result were pretty bad again. It is very unlikely for an economically disadvantaged student to do excellent in school. There is a 7.8% chance that a student of economic disadvantage will score 615 for 2019, but was better in 2021 at 19.22%, given the fact that not every school required their students to take the regent in 2021.

Logistic Regression: Probability of Being Homeless by the Mean Scale Score

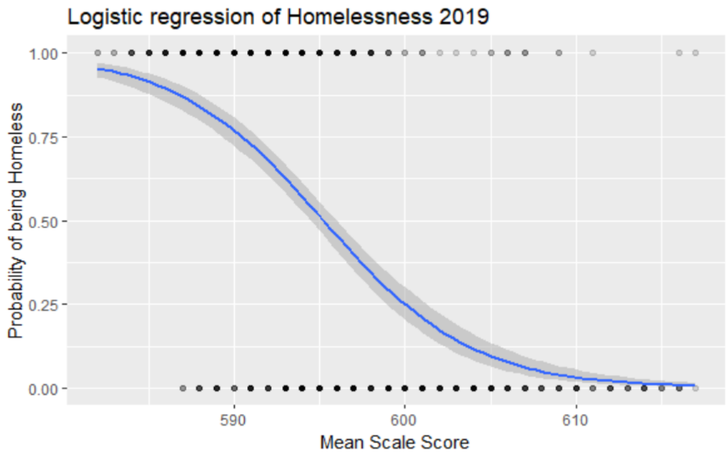


Figure 13. 2019 Homelessness to Score

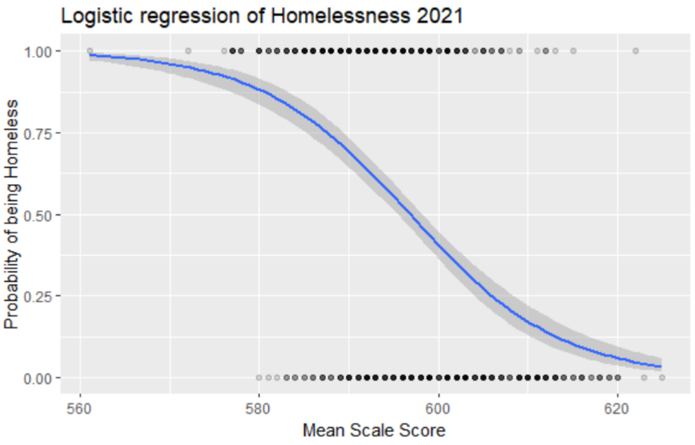


Figure 14. 2021 Homelessness to Score

There is a probability of 68.45% that the student was homeless in 2019 and 78.2% for 2021 if they scored 590 (almost a 2). There is only a probability of 12.13% that the student was homeless in 2019 and .09% for 2021 if they scored 615 (almost a 4). Similarly, being homeless affects the score of the students. They are far more likely to do worse as school is not one of their top priorities.

Linear Regression: School enrollment and Funding Per Pupil

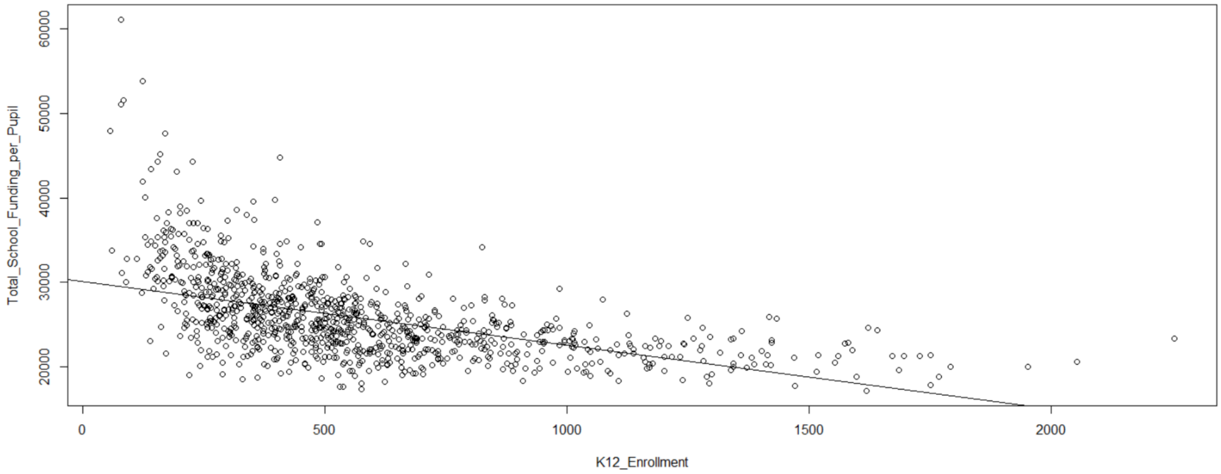
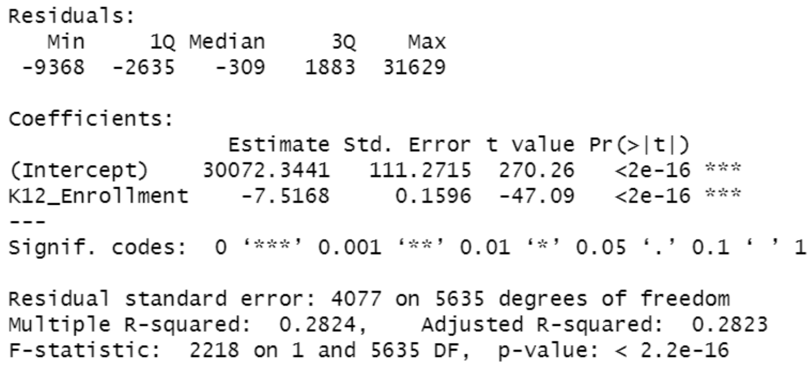


Figure 15. School enrollment and Funding Per Pupil Regression

Linear formula: 30072.344094 - 7.516809x



```
Residuals:
   Min     1Q Median     3Q    Max 
 -9368  -2635   -309   1883  31629 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    30072.3441   111.2715  270.26   <2e-16 ***
K12_Enrollment    -7.5168     0.1596  -47.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4077 on 5635 degrees of freedom
Multiple R-squared:  0.2824,	Adjusted R-squared:  0.2823 
F-statistic:  2218 on 1 and 5635 DF,  p-value: < 2.2e-16
```

Figure 16. Linear Regression Summary

The residuals are the difference between the actual values and the predicted values. It looks like our distribution is not quite symmetrical and is slightly right-skewed as the max is comparatively larger than the min. This tells us that the model is not predicting as well at the larger school sizes compared to smaller schools. The quantile-quantile plot also helps visualize this, you can see there are outliers on both ends of the chart, but those on the upper end look more severe than those on the bottom. Overall the residuals look to have a fairly normal distribution.

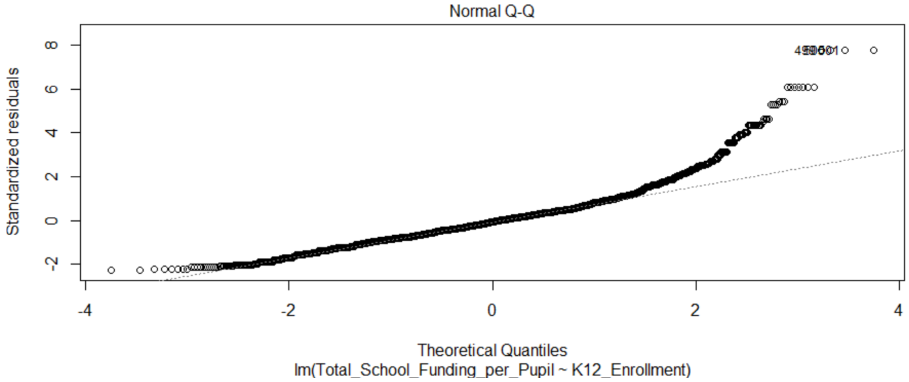


Figure 17. Linear Regression Normal Q-Q

~28.24% of the variation within Total_School_Funding_per_Pupil, our dependent variable. This means that points help to explain some of the variations within K12_Enrollment, but not as much as we would like. Ultimately, our model isn’t fitting the data very well.

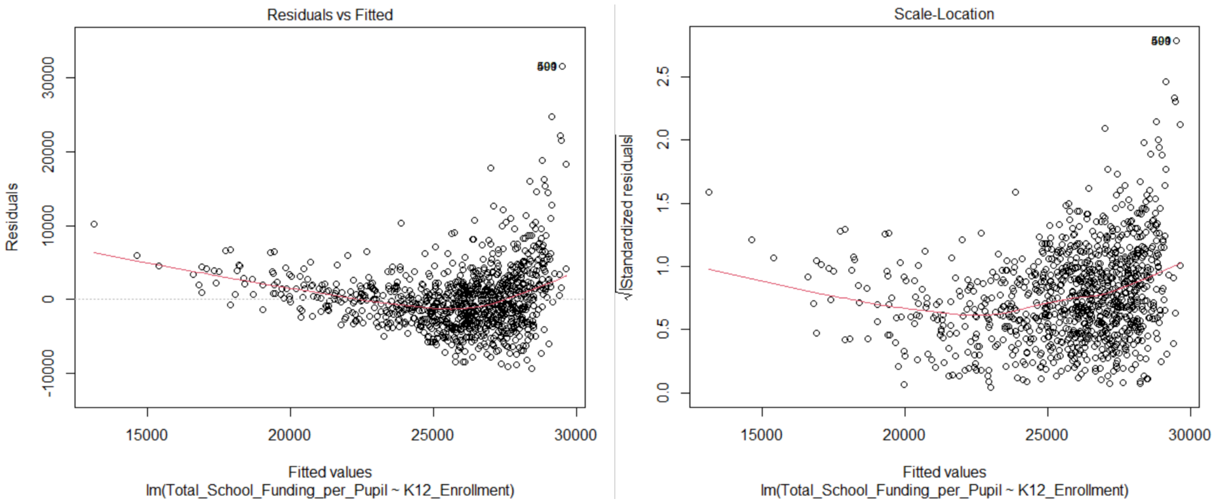


Figure 18. Linear Regression Residuals vs.Fitted and Scale-Location

For the residual vs fitted and scale location plots we can see that there is a concentration of points around the right side meaning the spread is not constant.

K-Means: Mean Score and Funding Per Pupil

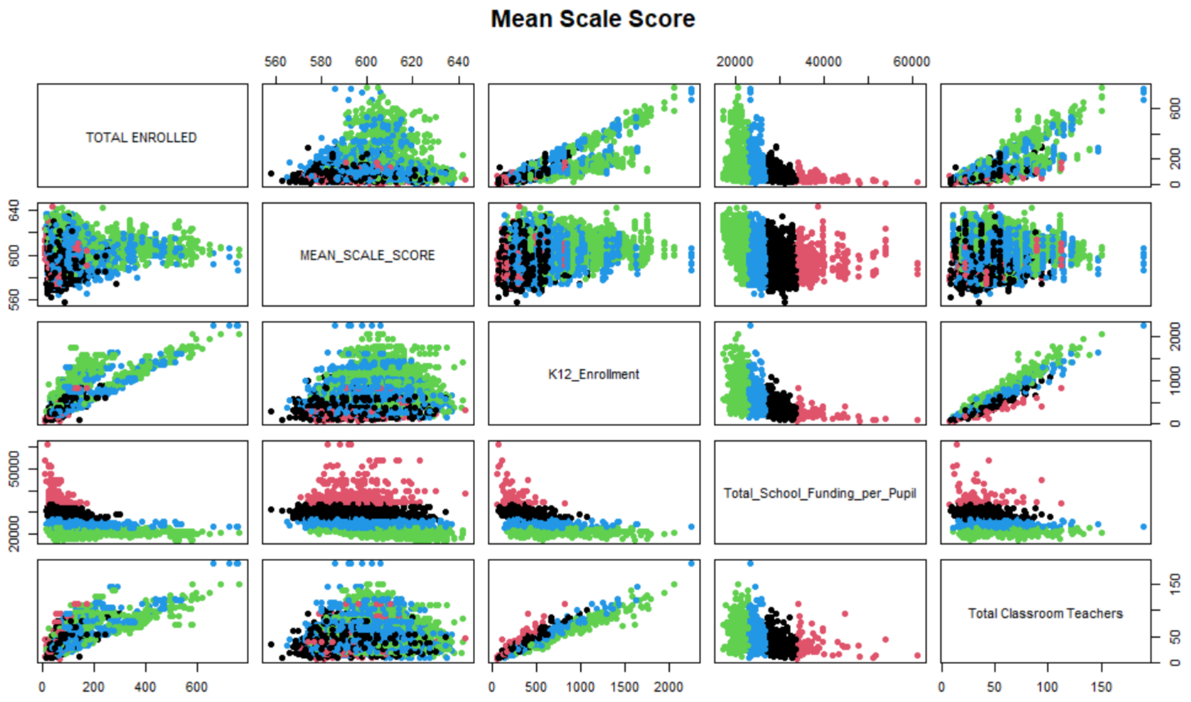
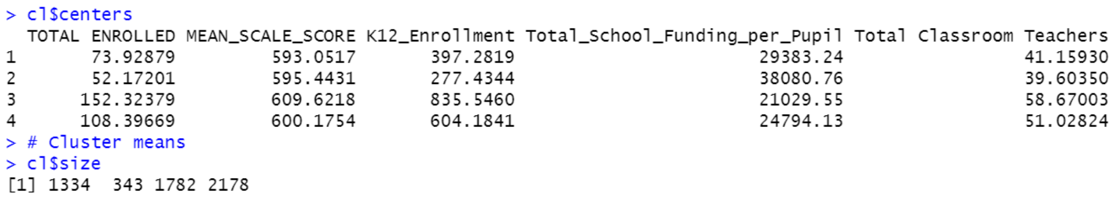


Figure 19. K-Means Between Total_Enrolled, Mean_Score, K-12_Enrollment, Funding_per_Pupil and Total_Classroom_Teachers



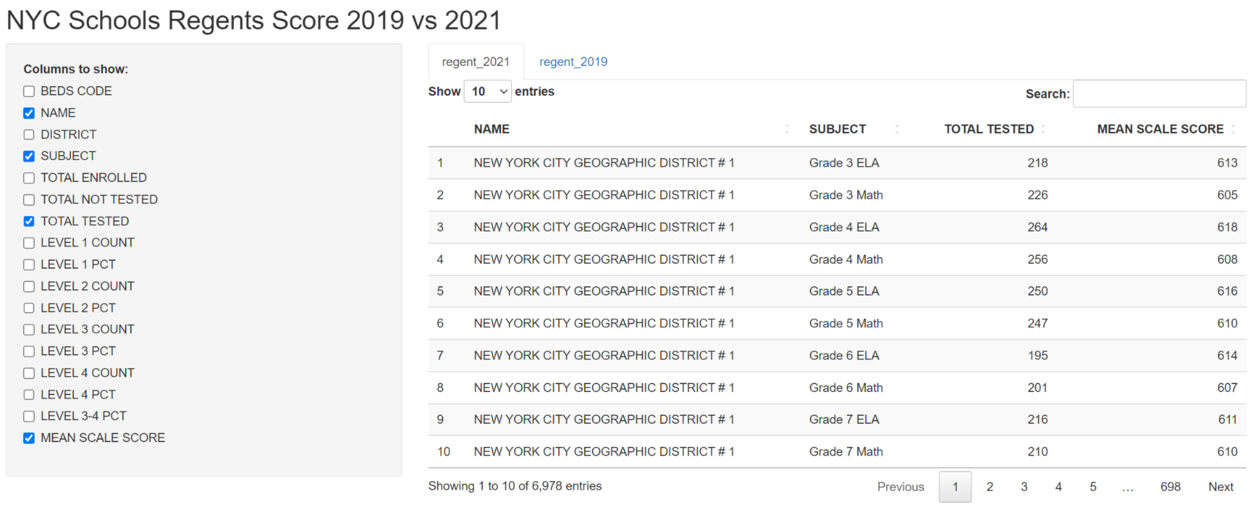
```
> cl$centers
  TOTAL ENROLLED MEAN_SCALE_SCORE K12_Enrollment Total_School_Funding_per_Pupil Total Classroom Teachers
1       73.92879         593.0517       397.2819                       29383.24                 41.15930
2       52.17201         595.4431       277.4344                       38080.76                 39.60350
3      152.32379         609.6218       835.5460                       21029.55                 58.67003
4      108.39669         600.1754       604.1841                       24794.13                 51.02824
> # Cluster means
> cl$size
[1] 1334  343 1782 2178
```

Figure 20. K-Mean Cluster Centers and Means

Interestingly, the Mean Scale Score and Total School Funding Per Pupil graph categorize the relationship based on the amount of funding. We can see the green cluster has the least funding per pupil but it has a bigger concentration of around 600-640 than any other cluster. This once again shows that the amount of school funding per pupil doesn't correlate with how well its student performs. There could also be a more hidden factor. If a school accepts more students with disabilities, they might get more funding in total but it may not be distributed equally.

User Interface

Utilize Shiny, R package for building interactive web apps in R. The UI is pretty rudimentary. It allows users to interact with the 2019 and 2021 regent dataset. The data is loaded using the saved excel files from the data cleaning phase and the fluidPage() helps create the UI. Users have a side bar where they can choose which fields/columns they are interested in. In the body, the user can toggle between viewing the 2019 or 2021 datasets. The users can search for different school names, filter/sort the columns alphabetically or numerically, and navigate through different pages.



NYC Schools Regents Score 2019 vs 2021

Columns to show:

☐ BEDS CODE
☑ NAME
☐ DISTRICT
☑ SUBJECT
☐ TOTAL ENROLLED
☐ TOTAL NOT TESTED
☑ TOTAL TESTED
☐ LEVEL 1 COUNT
☐ LEVEL 1 PCT
☐ LEVEL 2 COUNT
☐ LEVEL 2 PCT
☐ LEVEL 3 COUNT
☐ LEVEL 3 PCT
☐ LEVEL 4 COUNT
☐ LEVEL 4 PCT
☐ LEVEL 3-4 PCT
☑ MEAN SCALE SCORE

Show 10 **entries** **Search:**

	NAME	SUBJECT	TOTAL TESTED	MEAN SCALE SCORE
1	NEW YORK CITY GEOGRAPHIC DISTRICT # 1	Grade 3 ELA	218	613
2	NEW YORK CITY GEOGRAPHIC DISTRICT # 1	Grade 3 Math	226	605
3	NEW YORK CITY GEOGRAPHIC DISTRICT # 1	Grade 4 ELA	264	618
4	NEW YORK CITY GEOGRAPHIC DISTRICT # 1	Grade 4 Math	256	608
5	NEW YORK CITY GEOGRAPHIC DISTRICT # 1	Grade 5 ELA	250	616
6	NEW YORK CITY GEOGRAPHIC DISTRICT # 1	Grade 5 Math	247	610
7	NEW YORK CITY GEOGRAPHIC DISTRICT # 1	Grade 6 ELA	195	614
8	NEW YORK CITY GEOGRAPHIC DISTRICT # 1	Grade 6 Math	201	607
9	NEW YORK CITY GEOGRAPHIC DISTRICT # 1	Grade 7 ELA	216	611
10	NEW YORK CITY GEOGRAPHIC DISTRICT # 1	Grade 7 Math	210	610

Showing 1 to 10 of 6,978 entries

Figure 21. UI Screen

Conclusion

In conclusion the regent scores from the 2021 dataset wasn't conclusive enough to use for measuring the impact of Covid-19 on students. This was mainly because students in the 2020-21 school year were not required to take the NYS regents. Students had the option to opt in to take the regents, therefore the score was more skewed as those who are more confident in their ability would take it. On the other hand, a key take away from this project is that school funding in NYC doesn’t have a big impact on the student’s regent score and that whether the student is from an economically disadvantaged family has more impact. Through exploring with the researcher files from the New York State Education Department website, it is now clear why certain districts score better on the regents than others. Wealth of the child’s parents plays a huge part. If a child comes from a wealthy family that lives comfortably, they are more likely to score higher in comparison to kids from economically disadvantaged families.