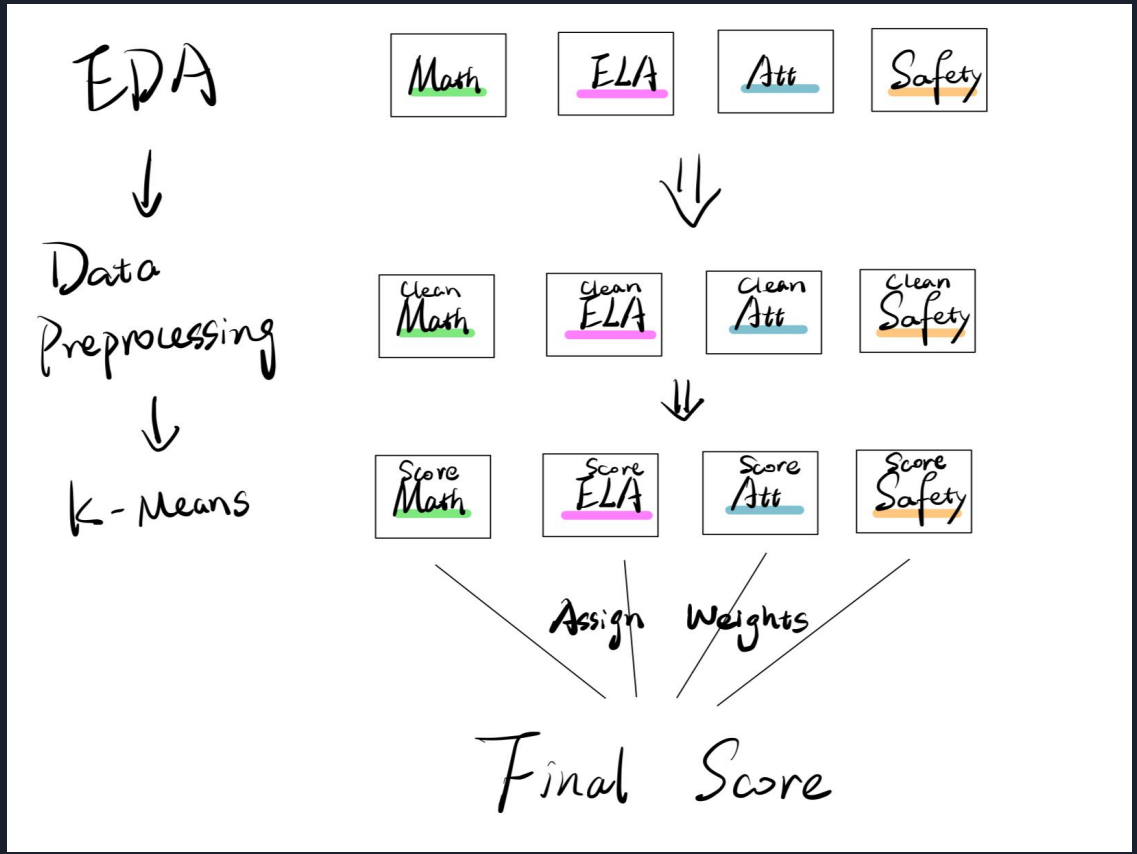
Education Dataset Analysis

By Yalong Zhang, Tianqi Liu, Wenjie Bi

Motivation

- Elementary education has a profound effect on a child’s future success. Our city's elementary schools face significant challenges, from resource constraints to varying safety standards. It's clear that improvements are needed.
- Ranking data empowers parents, policymakers, and schools. It guides decisions on school choice, resource allocation, and policy reforms.
- In this analysis, we aim to rank NYC's elementary schools based on a comprehensive set of factors, including students' test scores, attendance, safety, and more. We also focus on the relationship between school registration and safety scores. In the end, we analyze some potential education bias.

Modeling Flowchart

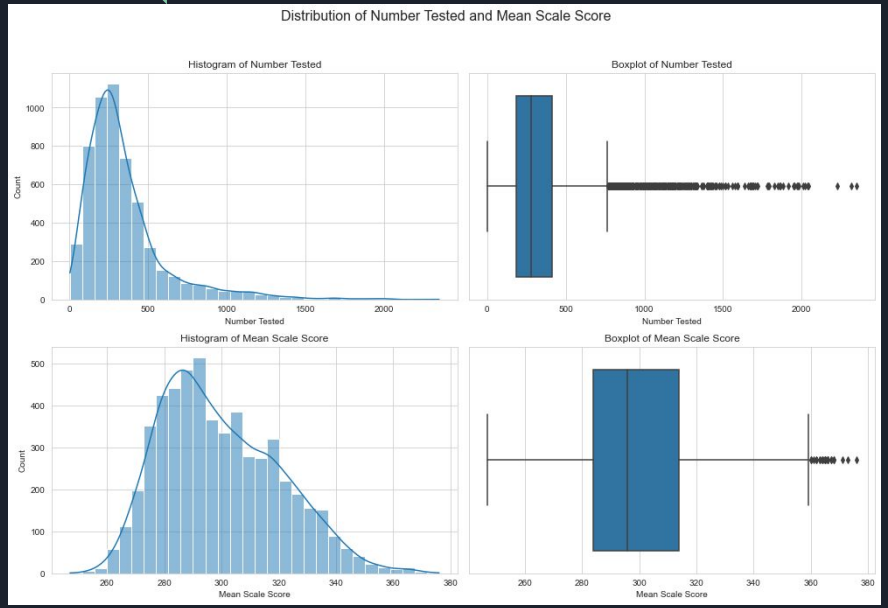


Data Preprocessing

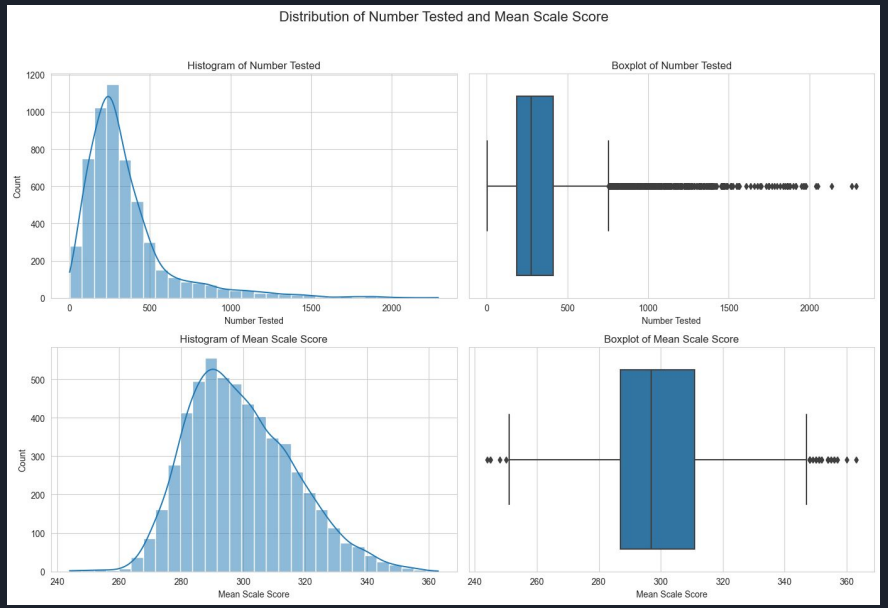
- We have accessed several different datasets which covers a wide range of variables, including math score, ELA(English Language and Arts) score, attendance, and safety datasets
- Data can be messy. We've addressed **missing values**, **outliers**, and **inconsistencies** to ensure the data's accuracy.
- For missing data, we employed techniques such as **imputation** or **exclusion** based on the context of the missing values.
- We've scaled or normalized numerical features to ensure they're on a consistent scale for analysis(May be done after EDA due to visibility).

EDA in Math and ELA

- We've done EDA on each of the datasets we obtain.

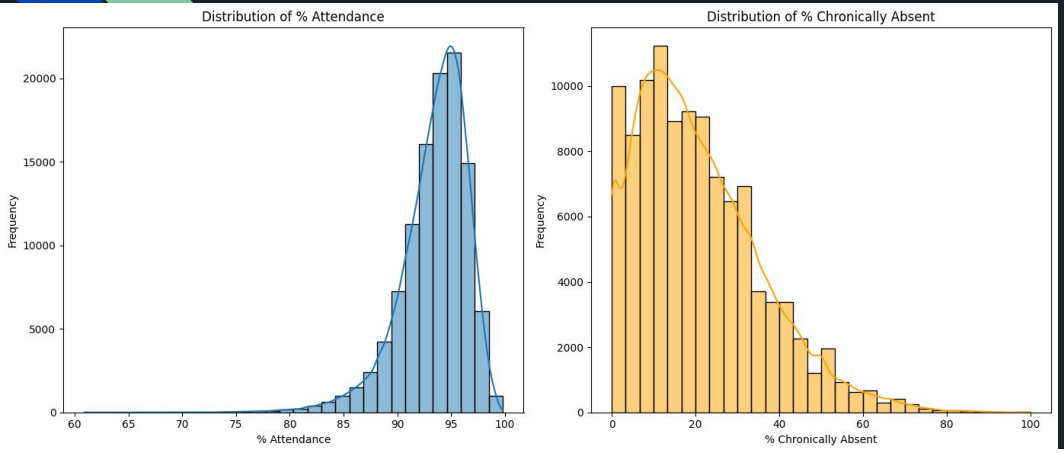


Math



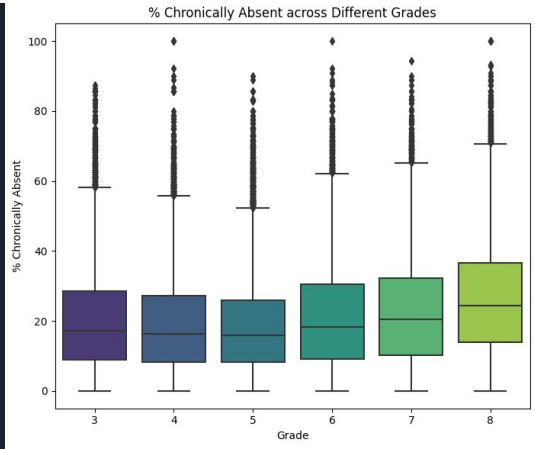
ELA

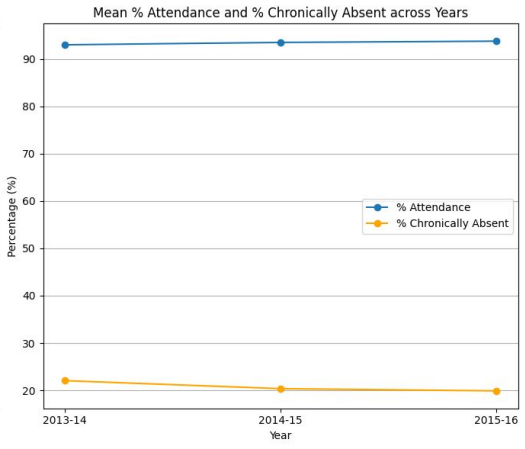
EDA in attendance



We saw that the histogram of % attendance is left skewed while the % chronically absent is right skewed. From these two graphs, we acknowledge that for most school, the attendance rate is around 95 percent. Also, we need to pay attention this skewness during the modeling session.

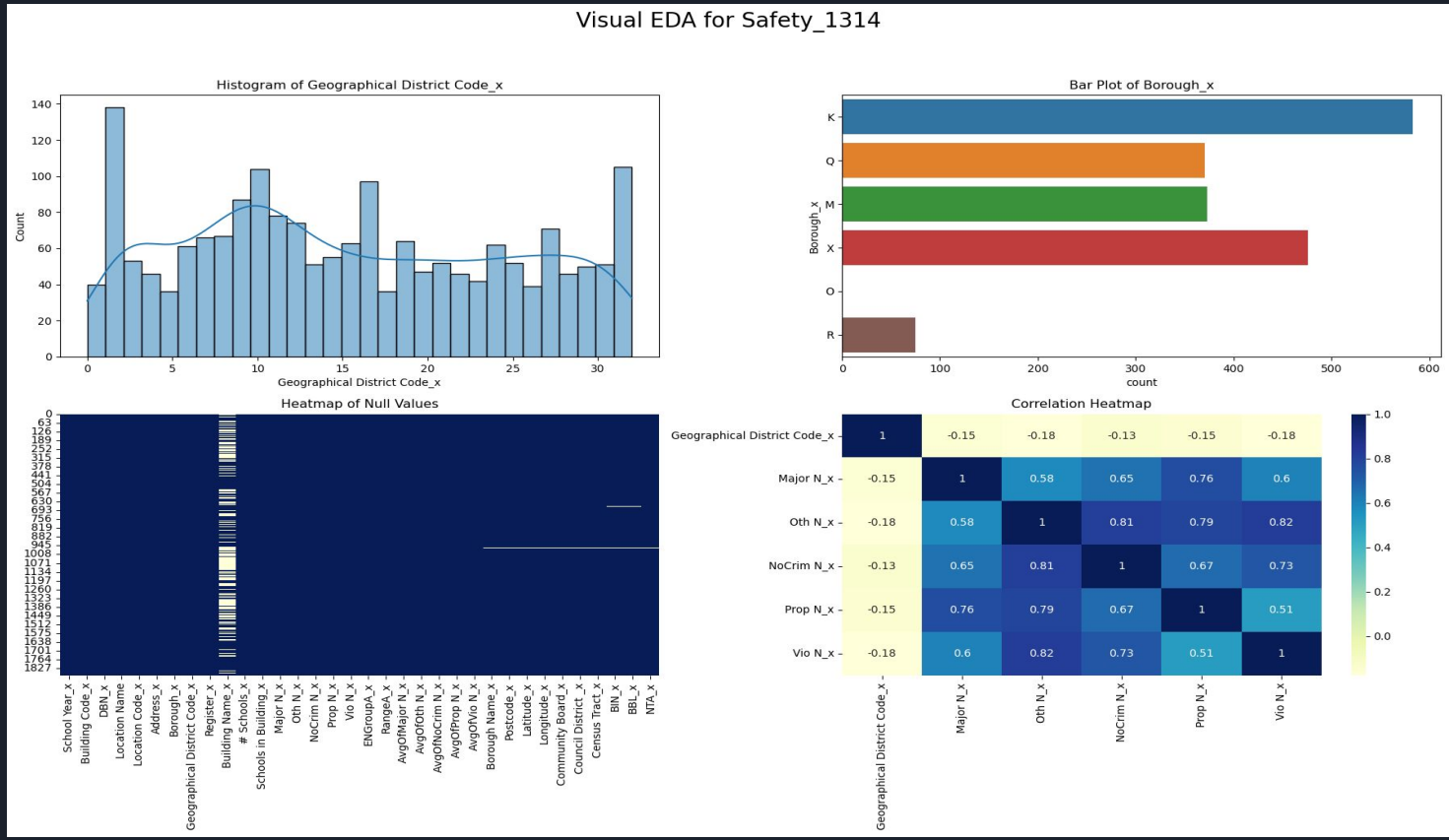
The attendance rate generally increases from low to higher grades, while the chronically absent rate decreases.





EDA in safety

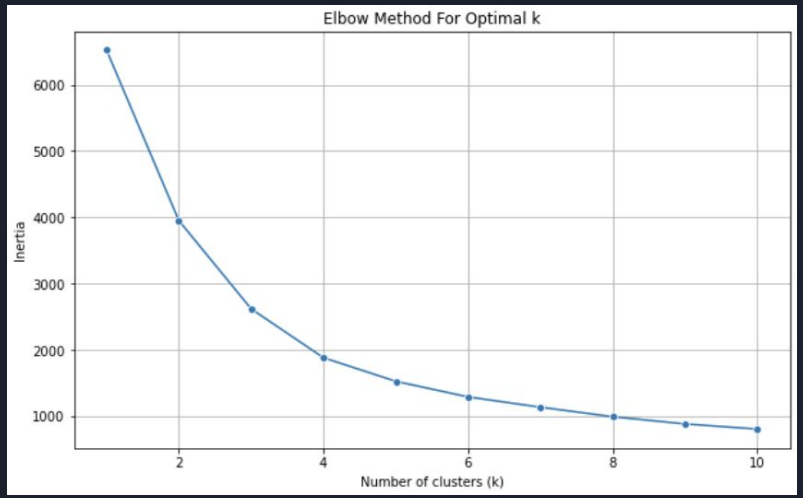
From these graphs on the safety dataset, we obtain that the data is about evenly distributed in case of geographical distinct code, indicating no bias on the locations. Also, we can see that there are some columns that are highly correlated, which we should pay attention to in future analysis.



Data Modeling

- We've created separate rankings for four key aspects: Math performance, ELA performance, Attendance rates, and Safety, since separate rankings allow us to capture the nuances of each aspect and address specific concerns in these areas.
- We've employed a K-means algorithm to calculate rankings for attendance and safety aspect.
- We also used the elbow method to better understand what is the optimal cluster number.

	DBN	ELA_Rank	MATH_Rank	ATTENDENCE_rank
0	01M015	5	5	2
1	01M019	5	5	5
2	01M020	4	4	2
3	01M034	3	3	5
4	01M063	5	5	2
...	...	...	...	...
1132	32K377	4	4	5
1133	32K383	1	1	2
1134	32K384	2	2	5
1135	32K554	3	3	2
1136	32K562	4	4	2

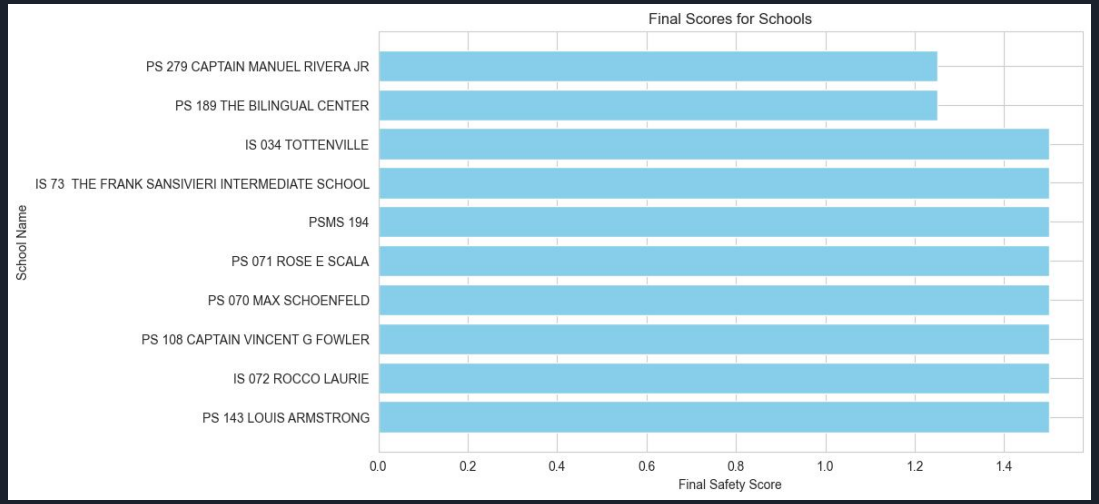


Results

	DBN	School Name	ELA_Rank	MATH_Rank	ATTENDENCE_rank	Composite_Score	Final_Rank	final_safety_score
0	10X279	PS 279 CAPTAIN MANUEL RIVERA JR	1	1	1	1.0	1	1.25
17	17K189	PS 189 THE BILINGUAL CENTER	1	1	1	1.0	1	1.25
37	31R034	IS 034 TOTTENVILLE	1	1	1	1.0	1	1.50
36	24Q073	IS 73 THE FRANK SANSIVIERI INTERMEDIATE SCHOOL	1	1	1	1.0	1	1.50
33	11X194	PSMS 194	1	1	1	1.0	1	1.50
28	08X071	PS 071 ROSE E SCALA	1	1	1	1.0	1	1.50
27	09X070	PS 070 MAX SCHOENFELD	1	1	1	1.0	1	1.50
45	27Q108	PS 108 CAPTAIN VINCENT G FOWLER	1	1	1	1.0	1	1.50
22	31R072	IS 072 ROCCO LAURIE	1	1	1	1.0	1	1.50
40	24Q143	PS 143 LOUIS ARMSTRONG	1	1	1	1.0	1	1.50

After adding safety measures into account, we arrived with our final model. We rank those schools by math score, ELA score, attendance and safety, and selected the top 10 school under this model.

	DBN	safety_rank
0	19K292	3.75
1	19K292	3.75
2	02M347	2.00
3	27Q475	3.25
4	27Q475	3.25
5	06M540	3.50
6	10X459	1.50
7	19K661	3.50
8	13K499	2.75
9	75Q277	2.00



More analysis on SAFETY

- Since safety is one of the most important factors to consider when choosing a school, we want to see whether there is a significant relationship between school registration numbers and various safety scores.
- We use linear regression, decision trees, and random forests to fit our models.
- We follow the regular data analysis steps to get our final models, which include data exploration, data preprocessing, modeling, and model evaluation.

Linear regression, decision tree and random forest

- By using decision trees, we can see the top 3 important features, and we fit models by using these features.
- After fitting the models, we can see the R^2 for all the models are around 0.8.
- Therefore, we can use these three features to predict schools' registration numbers.

Feature	Importance
AvgOfNoCrim N	0.611302
AvgOfProp N	0.122951
AvgOfOth N	0.116307

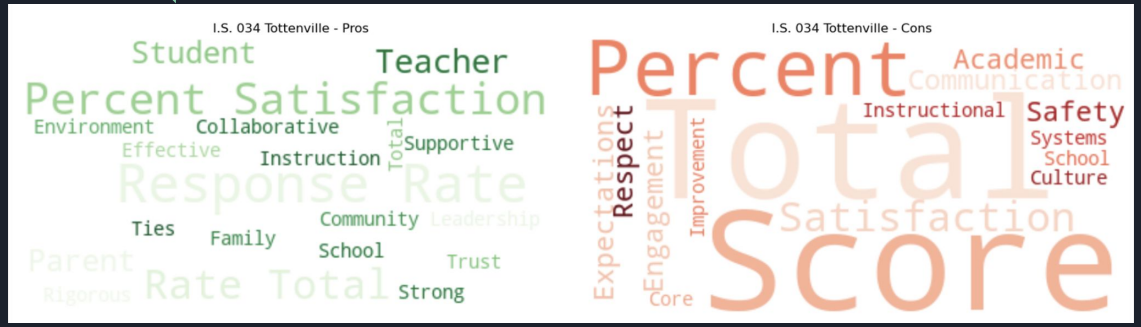
Top 10 Schools in NYC

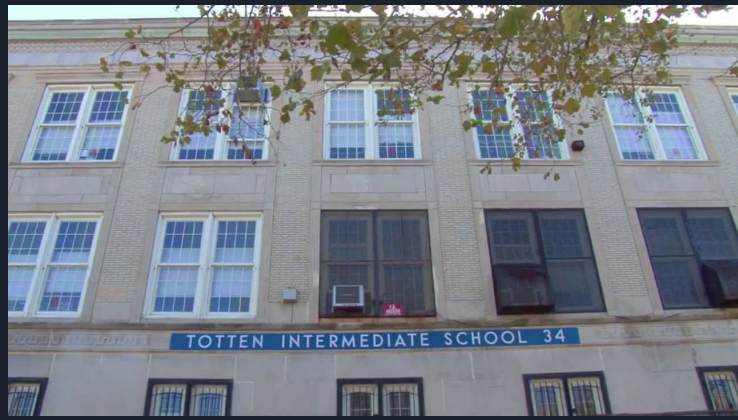
10. PS 279 CAPTAIN MANUEL RIVERA JR 🏆🏆🏆🏆

9. PS 189 THE BILINGUAL CENTER 🏆🏆🏆🏆

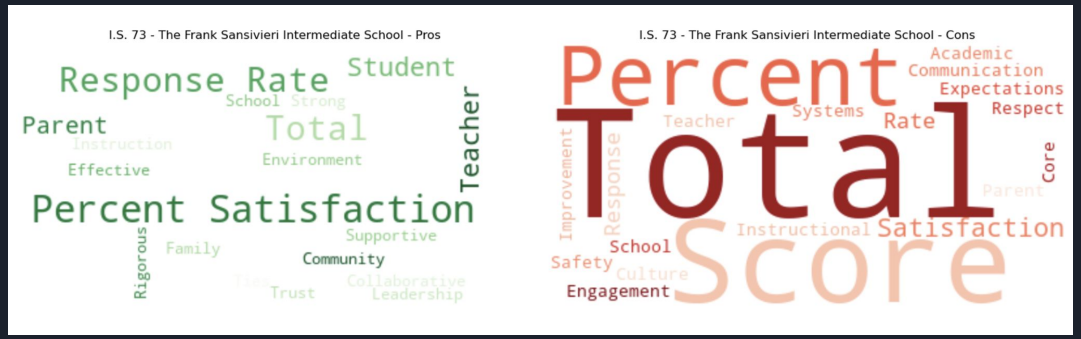
Top 10 Schools in NYC

8. IS 034 TOTTENVILLE 🏆🏆🏆🏆🏆





7. IS 73 THE FRANK SANSIVIERI INTERMEDIATE SCHOOL 🏆🏆🏆🏆🏆



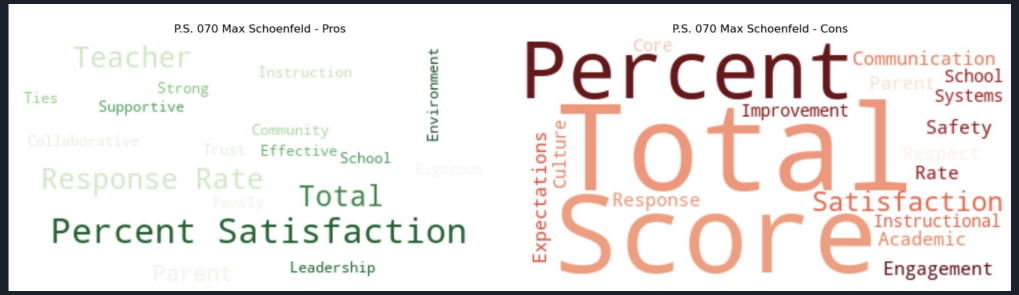
Top 10 Schools in NYC

6. PSMS 194 🏆🏆🏆🏆🏆

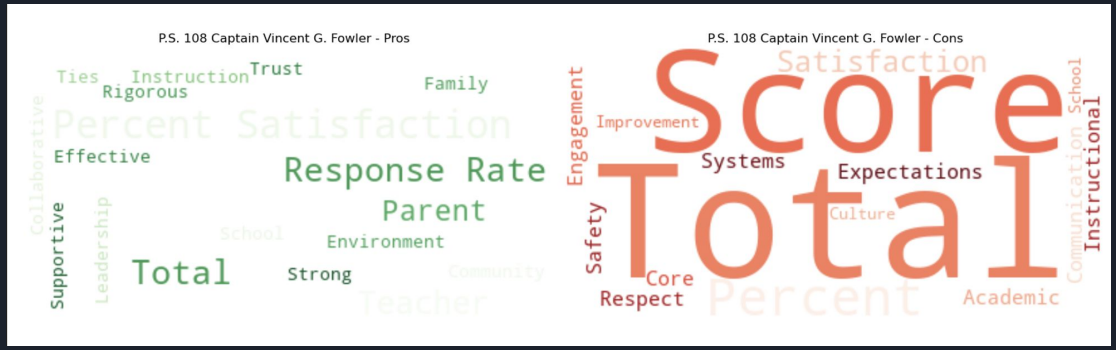
5. PS 071 ROSE E SCALA 🏆🏆🏆🏆🏆

Top 10 Schools in NYC

4. PS 070 MAX SCHOENFELD 🏆🏆🏆🏆🏆

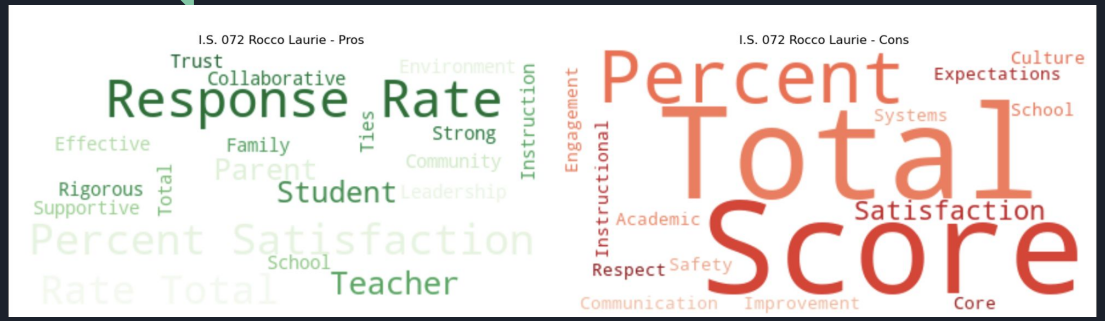


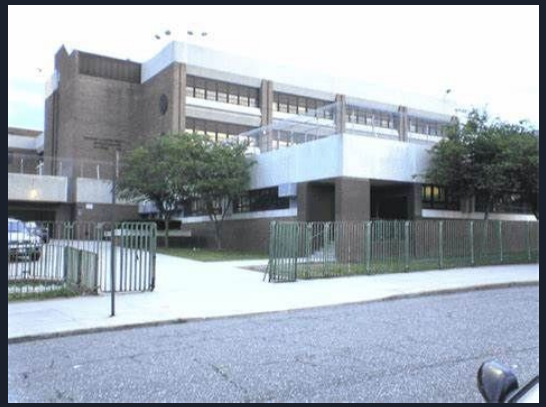
3. PS 108 CAPTAIN VINCENT G FOWLER 🏆🏆🏆🏆🏆



Top 10 Schools in NYC

2. IS 072 ROCCO LAURIE 🏆🏆🏆🏆🏆





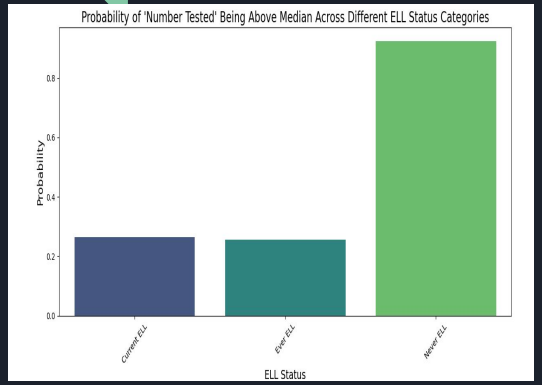
1. PS 143 LOUIS ARMSTRONG 🏆🏆🏆🏆🏆

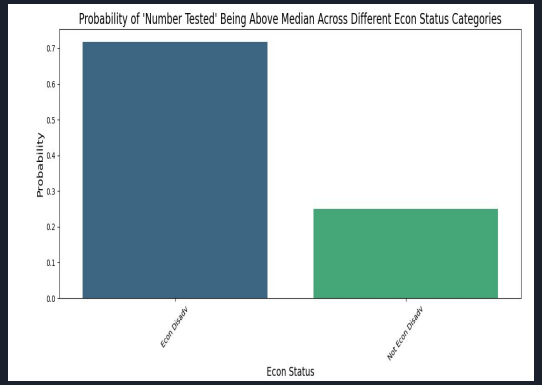
Insights

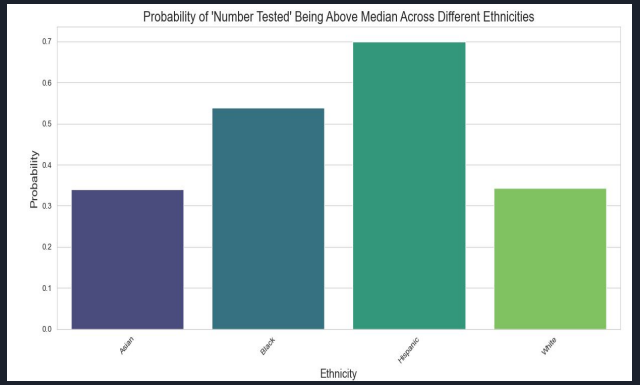
- Our analysis has identified the top-performing schools in NYC, considering various aspects. These schools consistently excel in percent satisfaction.
- We've also uncovered disparities among schools, indicating areas that need improvement. These disparities can be attributed to factors like scores and expectations.
- We emphasize transparency in our ranking model, allowing stakeholders to adjust weights based on their priorities as well as adding new features into the model.
- Our analysis has the potential to influence policy decisions related to education funding, school improvement initiatives, and resource allocation.

Bias in the education system ?

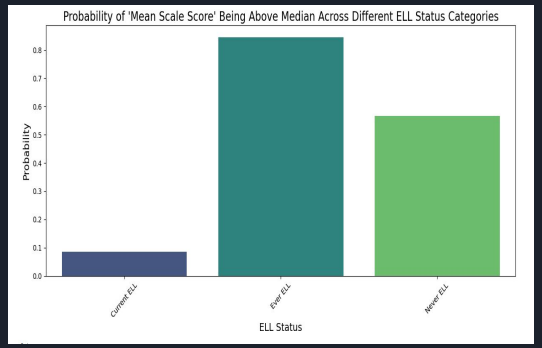
Start with EDA on original math dataset: Number Tested

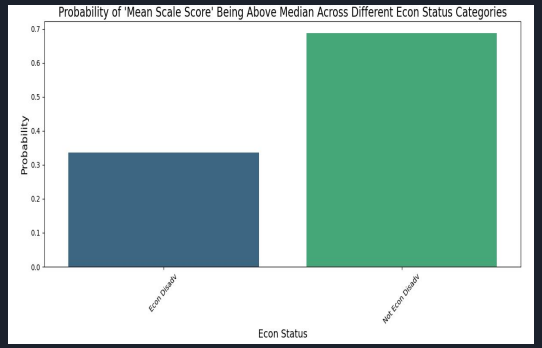


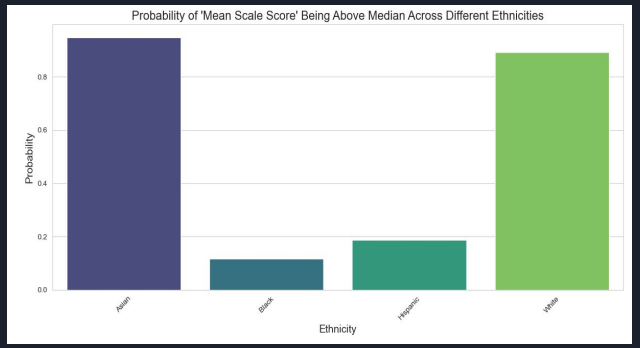




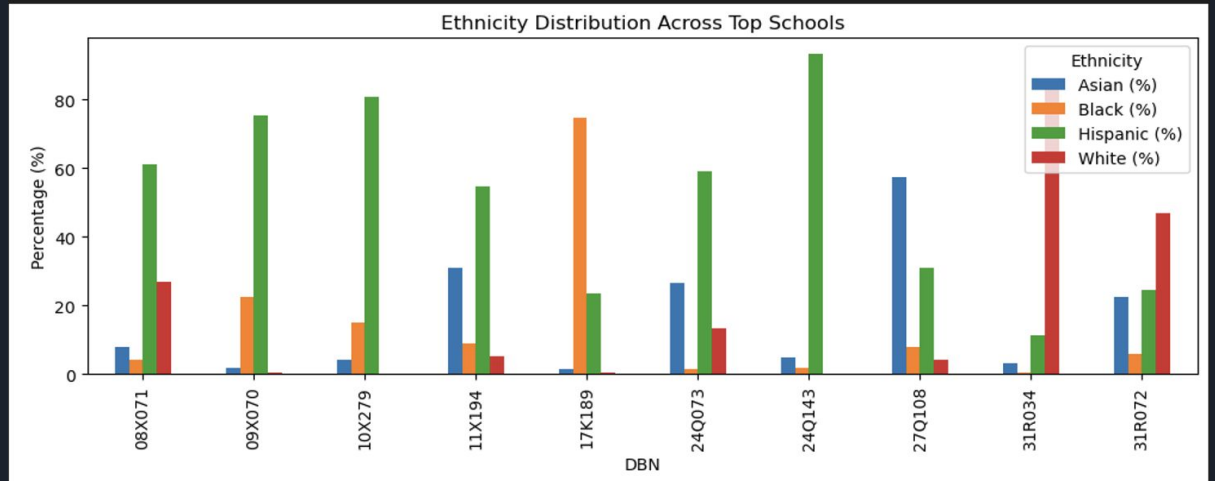
Then with EDA on original math dataset: Mean scale score with fill NaN

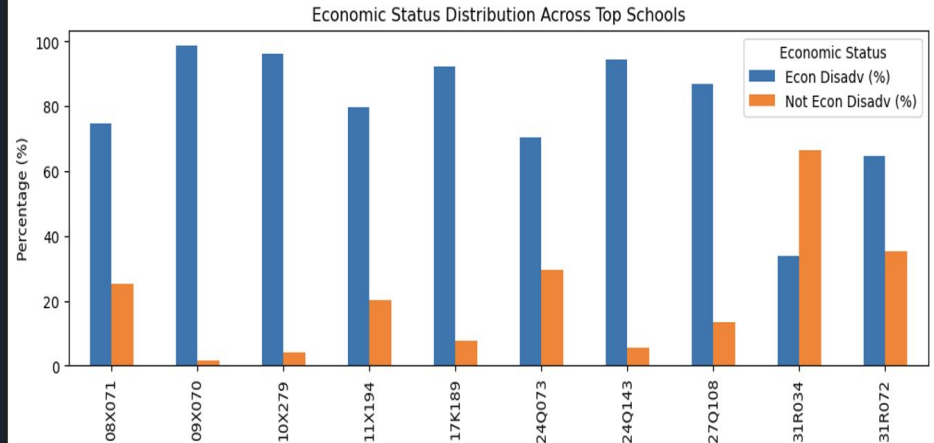


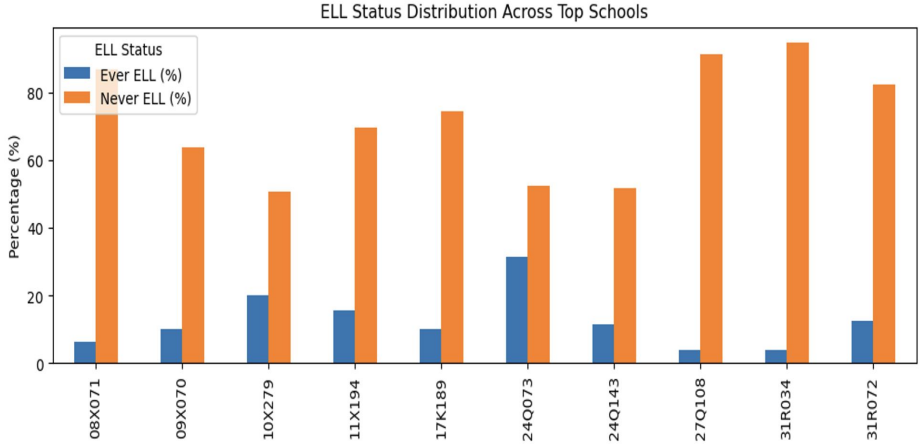




Demographic Parity (DP) in Top 10







Thank you!