Capstone Technical Report

Juliann Groglio, Azure Eller, Widnie Dorilas, Seungmin Kim

Introduction

After being assigned our capstone project and knowing we had to explore finance, we decided to explore how the allocation of funds across various aspects of schools affected students. We originally wanted to look at it across the entire country between different states. However, the more we explored, the more we realized that it would be better to pick a specific state or region and solely focus on that. After a group discussion we decided to focus on the state of New York since we all had a connection to it.

We started getting curious about the effects, if there were any, in the funding allocation of different school districts and the outcome in their graduation rates. We decided to look at the demographic breakdown of the school districts, the graduation rates, and the amount of funding they were receiving and how it was being dispersed. After initial data exploration, we decided that we wanted to find the answers to the following questions:

1. Top 5 districts with highest/lowest female to male ratios?
2. Do districts with higher female to male ratios have higher graduation rates?
3. How do homelessness rates affect graduation rates?
4. Does the amount of pupil support expenditures school districts spend influence Advanced Regents diplomas acquired?
5. What is the racial distribution among all school districts?
6. Is there a correlation between property taxes and enrollment?
7. What are top 5 districts with the highest dropout rate?
8. What is the racial makeup of the top 5 districts with the highest dropout rate?

9. Does an increase in total staff wages increase the number of students graduating?
10. Can we predict school dropout rates based on financial allocations and demographic factors?
11. What is the correlation between the number of economically disadvantaged students and graduation rate?

Data Sources

The data sets that we used on our project came from two separate sources:

- 2019 Public Elementary-Secondary Education Finance Data obtained from the U.S. Census Bureau
    - 2019 Public Elementary-Secondary Education Finance Data (census.gov)
- 2019-20 Enrollment Database from the New York State Education Department
    - Downloads | NYSED Data Site
- 2019-20 Graduation Rate Database from the New York State Education Department
    - Downloads | NYSED Data Site

Data Overview

Once the ETL of the datasets was complete, the data we have is:

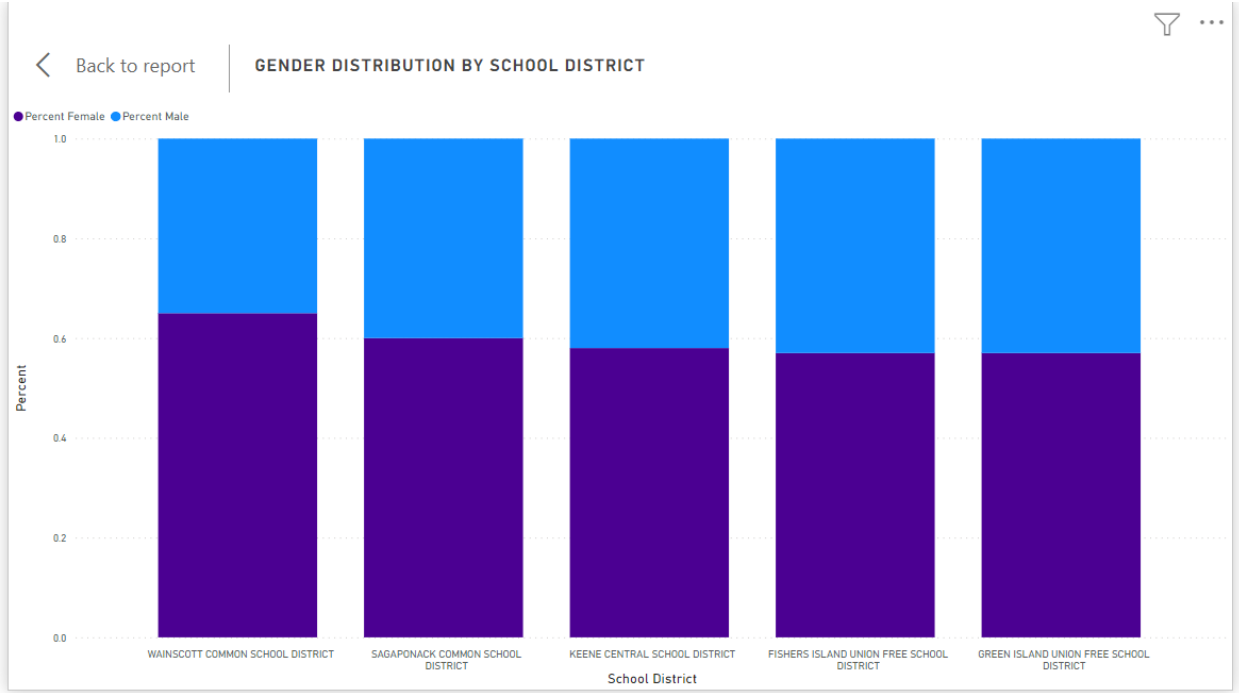
- The financial distribution of funds in 2019 for school districts
- The 2019 graduation rates and degrees earned for school districts
- The 2019 demographic factors for school districts

Before we started exploring our questions, we did some basic data exploration. During the exploratory data analysis, we saw that the column for COE Pupil Support had only zero values, so we made the decision to use the Total Expenditure Support Services and revised question four.

While looking at graphs dealing with dropout count and graduation count, we saw strong relationships, however, we realized that dealing with “counts” instead of “rates” lead to skewed data, since graduation and dropout counts are bound to increase as enrollment count increases. To fix this issue, we created new columns to function as rates by dividing the count values by enrollment counts.

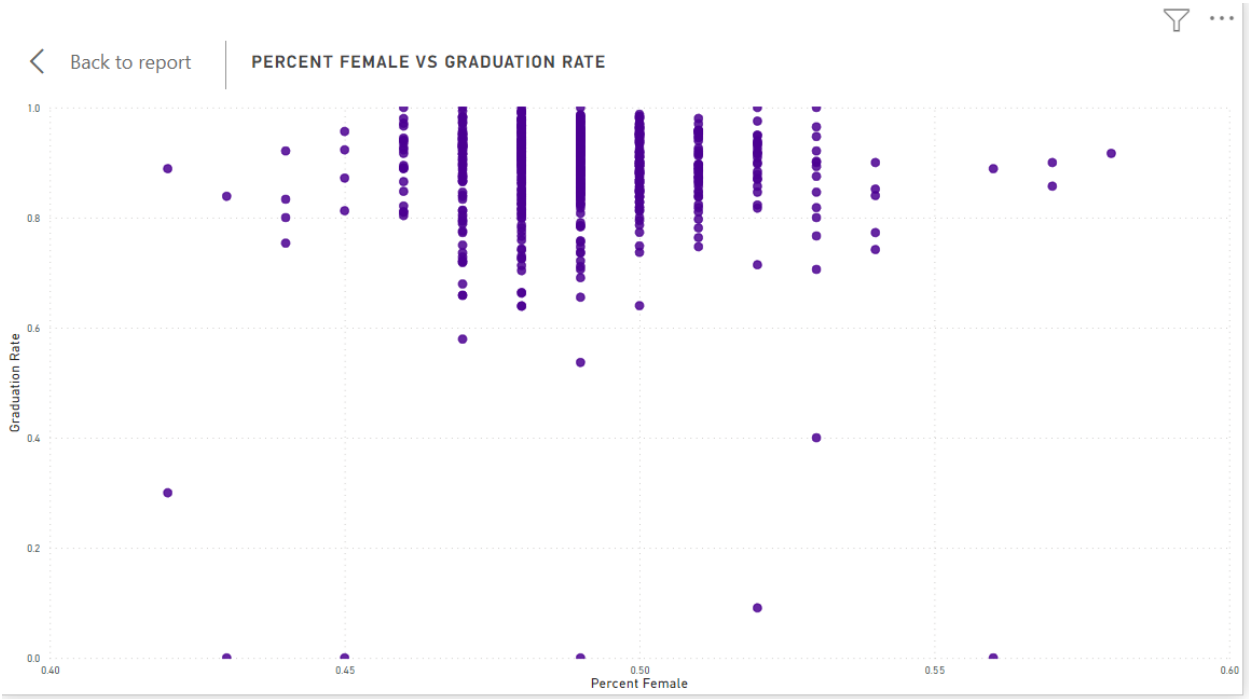
Question Exploration

Q1: Top 5 districts with highest/lowest female to male ratios?



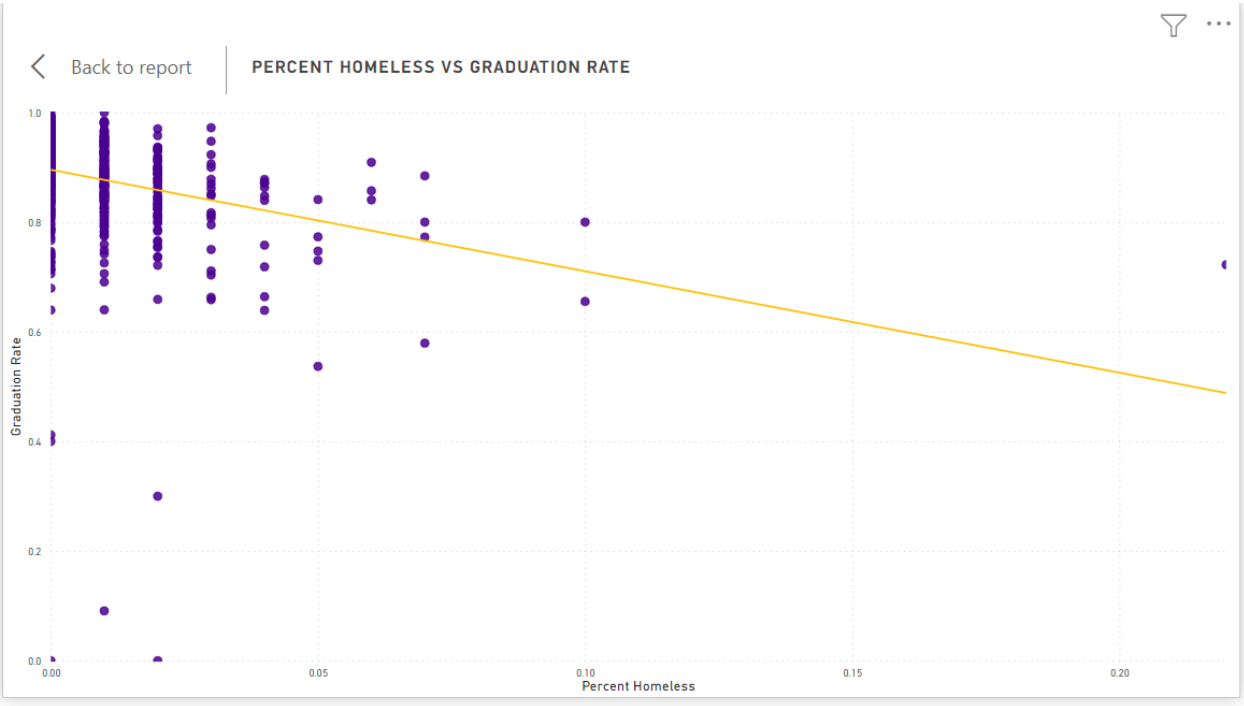
We wanted to see what the school districts were that had the highest female to male ratio and see if that had any effect on the graduation rates.

Q2: Do districts with higher female to male ratios have higher graduation rates?



Once we knew what districts had the highest female to male ratios, we then decided to look at the graduations rates comparatively. There are a few outliers on the graph but for the majority, it looks like districts with an even ratio have a standard graduation rate starting in the low 70s while districts with a higher female ratio have graduation rates starting in the low 80s.

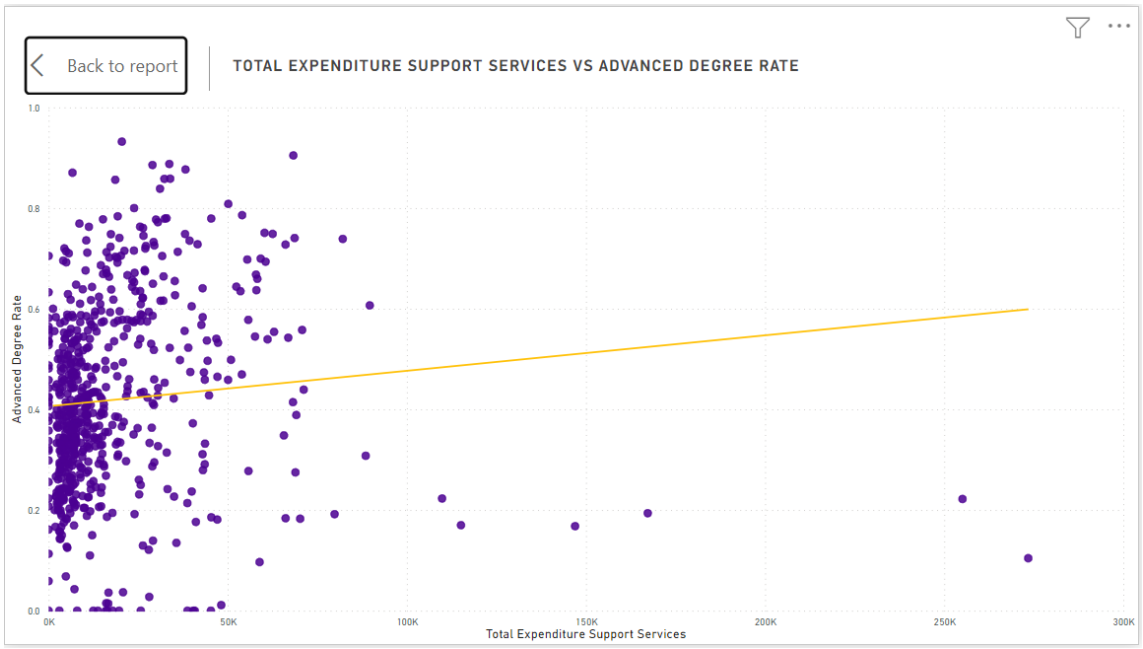
Q3: How do homelessness rates affect graduation rates?



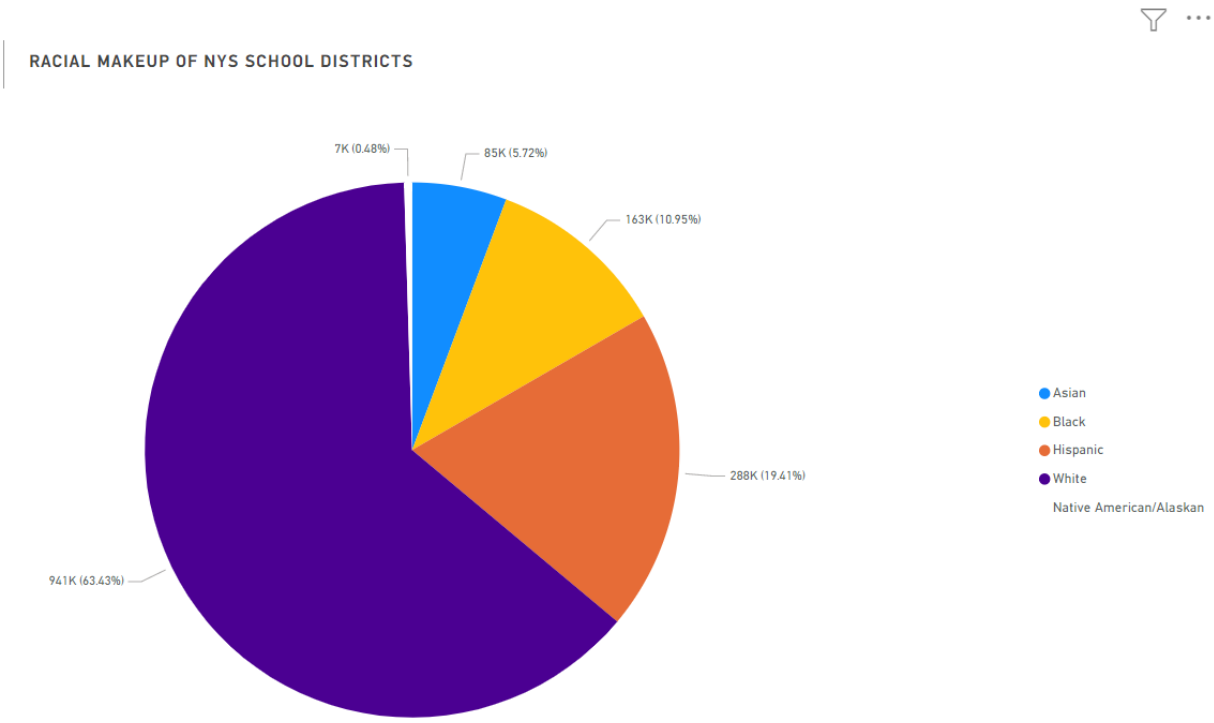
Another question we wanted to explore was see the effects of schools with a homeless student population and how it affected the graduation rates. As the graph shows, districts with a lower homelessness percentage had higher graduation rates with few outliers.

Q4: Does the amount of pupil support expenditures school districts spend influence Advanced Regents diplomas required?

We asked this question because we wanted to know if spending more money on accelerating student studies and making sure they had the tools and support they needed would lead to them excelling more. According to this graph, the data shows that when districts are spending between $50-100K on expenditures, more students are getting advanced degrees. However, when districts spend over $100K, the number of advanced degrees stagnates around 20%.

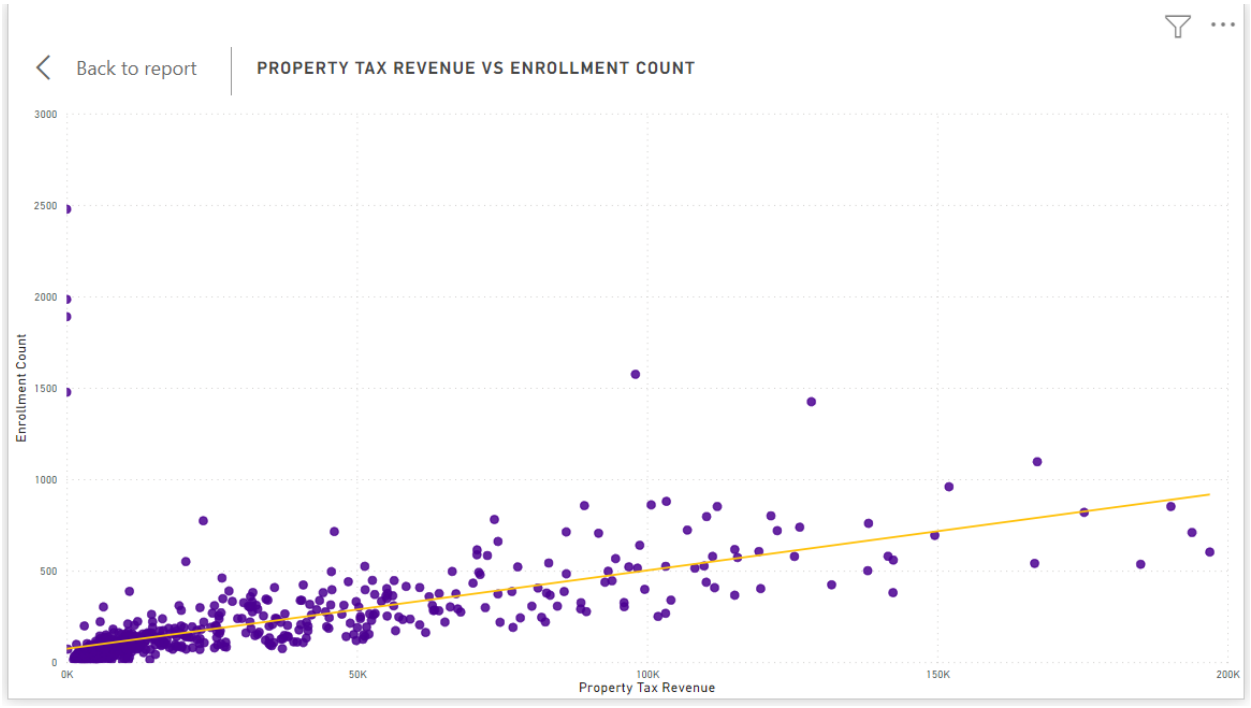


Q5: What is the racial distribution among all school districts?



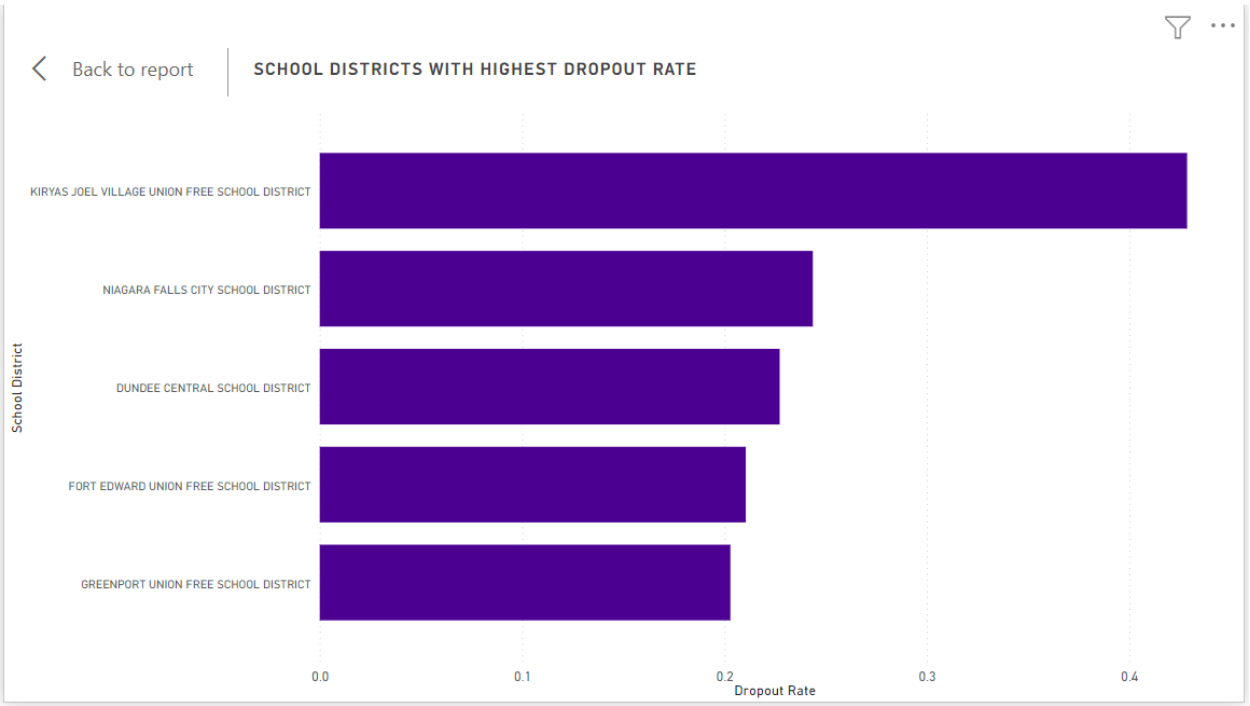
Considering New York is a state known for being diverse and where a lot of immigrants go, we were curious about the racial makeup of the school districts. This chart shows that almost 64% of the students are White with Hispanic and Black students following behind.

Q6: Is there a correlation between property taxes and enrollment?



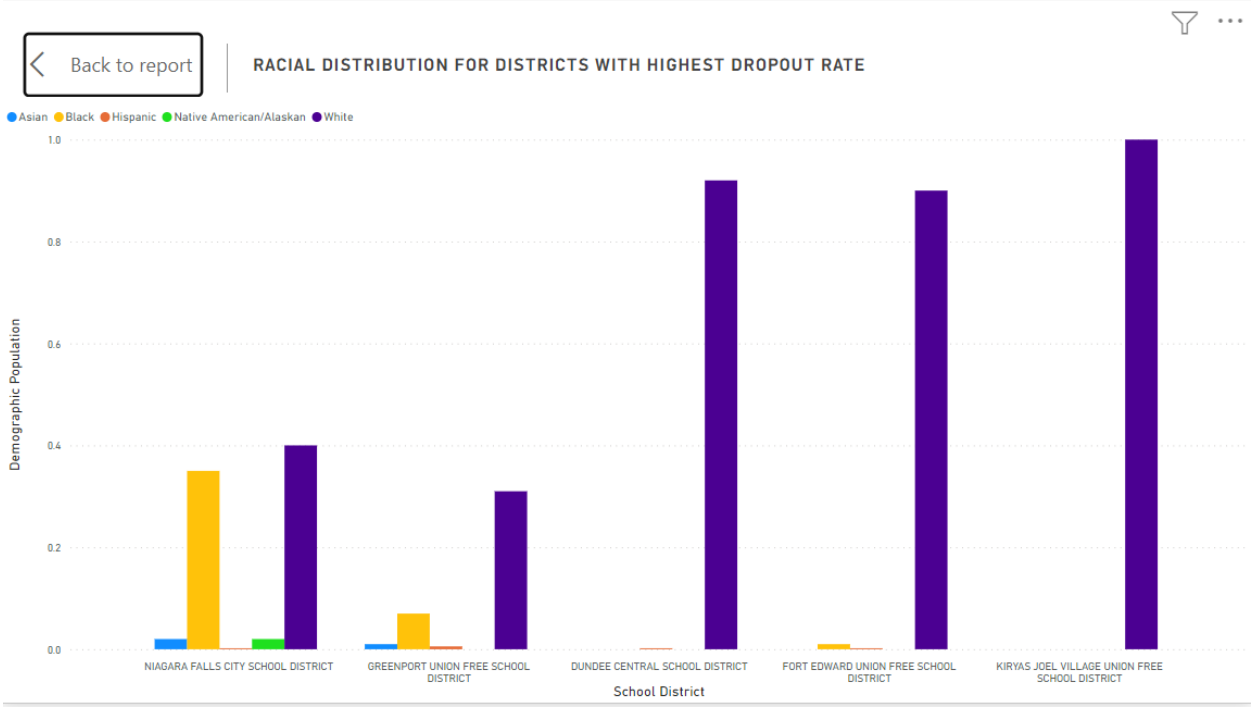
Considering that districts are mostly funded through property taxes we wanted to see if having more property tax which equates to more district funding would also equate to higher student enrollment. In this graph, the data shows that as property tax revenue increases so does the school enrollment until the revenue hits $150K which is when the enrollment tapers off.

Q7: What are the top 5 districts with the highest dropout rate?



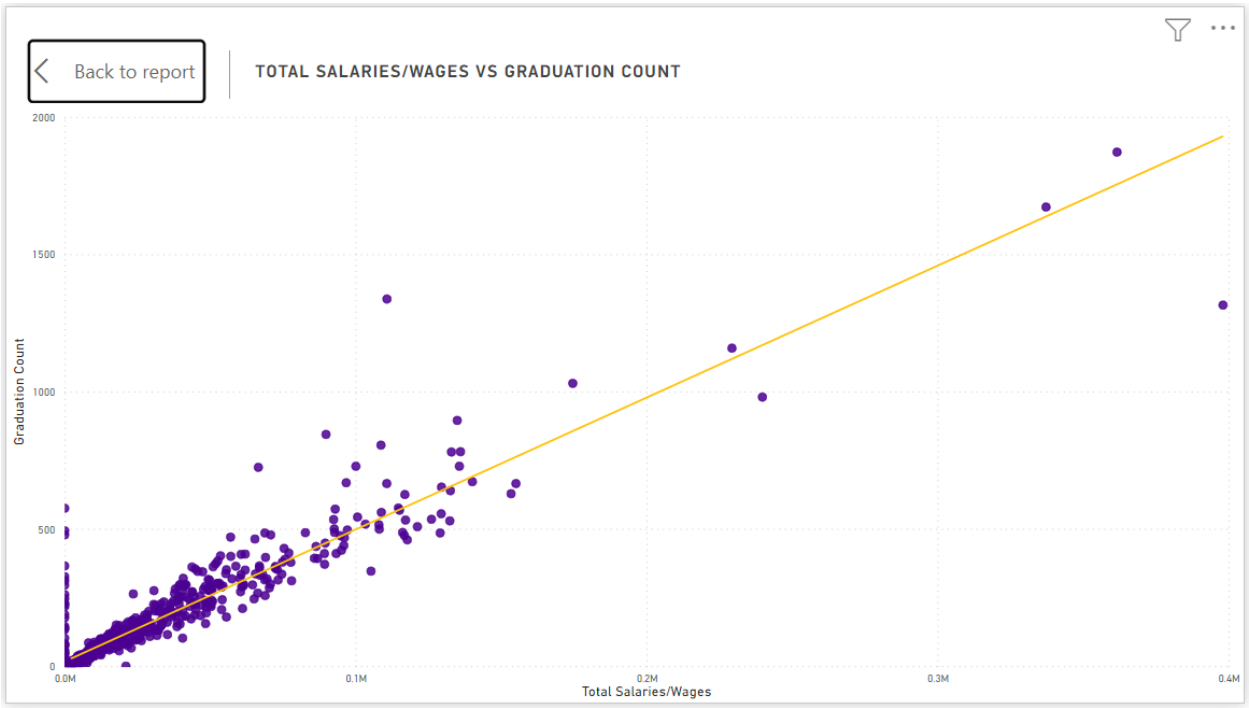
This question was asked in the same realm as figuring out which districts had the highest female to male ratios along with seeing the racial distribution of all the districts. Having this information showed us that the number one district with a dropout rate of over 40% was 20% more than the number five district of highest dropout rates. This information then prompted us to try and figure out why the dropout rate was so high.

Q8: What is the racial makeup of the top 5 districts with the highest dropout rate?



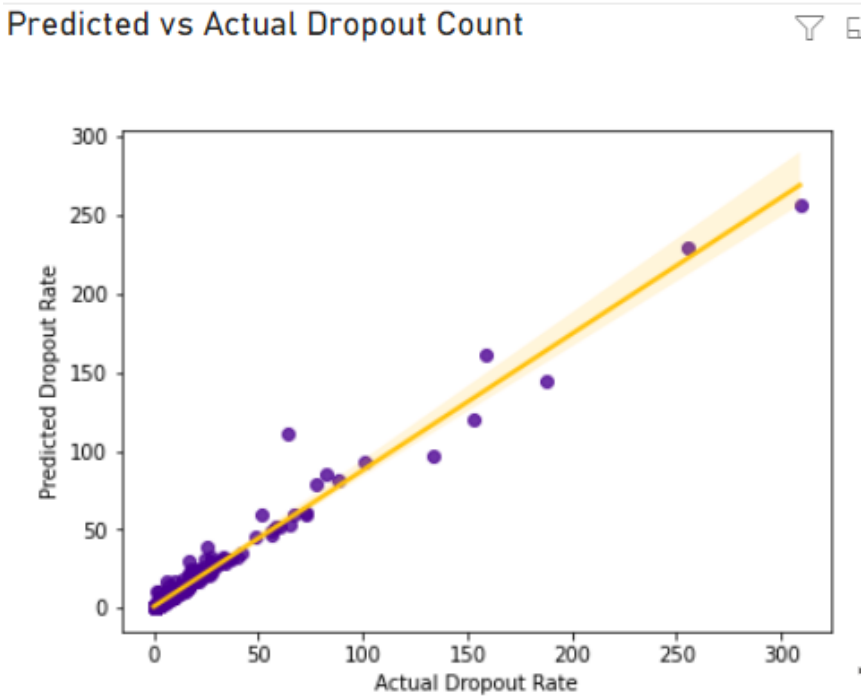
Once we knew who the districts with the highest dropout rates were, we decided to see what the racial distribution was out of curiosity. The data is ordered from the fifth to the first. The racial makeup of all the districts is White followed by a smaller percentage of Black, Native American, and Asian students.

Q9: Does an increase in total staff wages increase the number of students graduating?



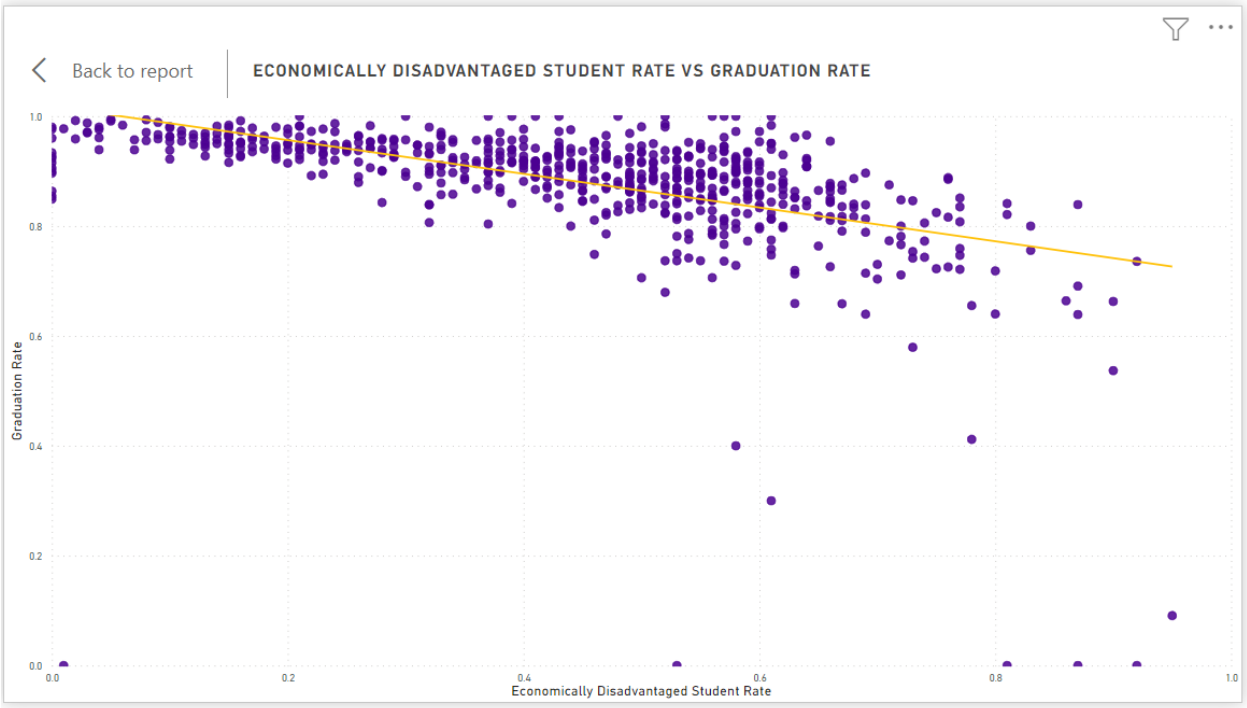
As students who have been through the education system, we hypothesized that staff who got paid more could put more time and effort into making sure students were successful which could lead to higher graduation rates. The data in this chart proves our hypothesis because as the number of total wages increases, so does the number of students graduating.

Q10: Can we predict school dropout rates based on financial allocations and demographic factors?



This graph was made based on our machine learning model to see if we could take the financial allocations along with the demographic factors to predict the dropout rates of school districts. When sampled against the actual dropout rates, the machine learning algorithm was very acutely predicting the rates.

Q11: What is the correlation between the number of economically disadvantaged students and graduation rate?

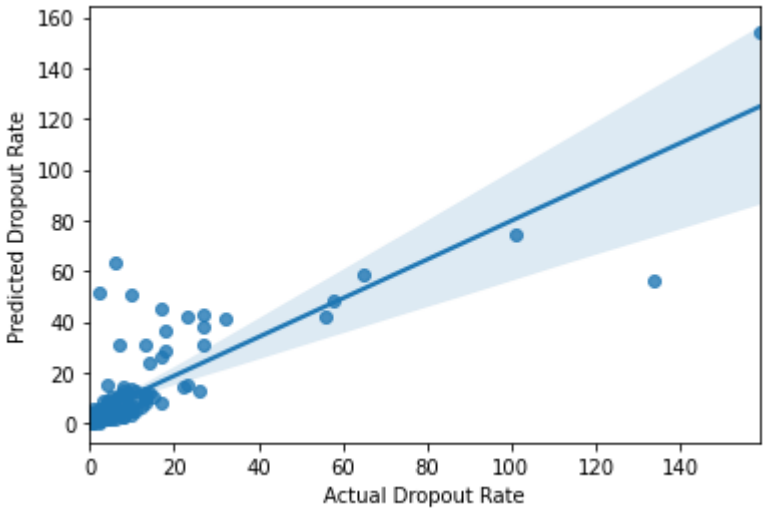


Our last and final question was to see how the rate of economically disadvantaged students affected the graduation rates. When a student is economically disadvantaged, they do not have the access to tools that will help them graduate such as extra tutoring and special attention if the school does not provide it. As can be seen by the data, as the rate of economically disadvantaged students rises, the graduation rates of the districts go down.

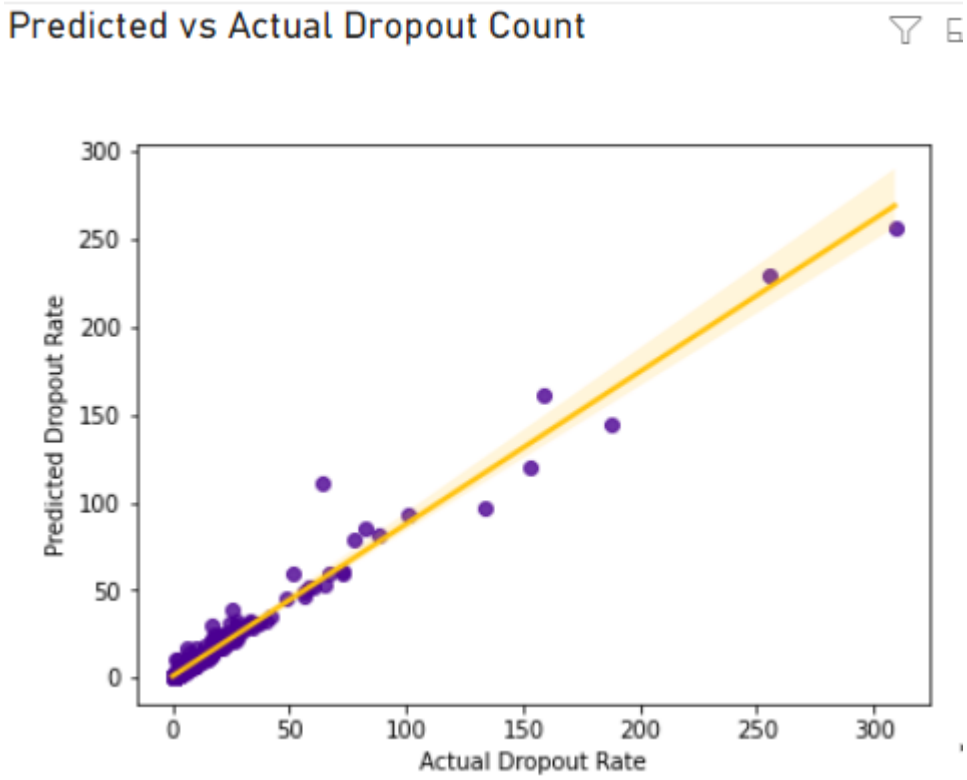
Machine Learning Process

The breakdown of a school districts' success can be seen in graduation count and dropout count. For our Machine Learning Model, we decided to dive deeper into the dropout count to see if we could predict it. Using all three data sets from the US Census Bureau and NYS School District data, we built a model with the prediction variable as the dropout count and the rest of the variables as dependent variables, excluding all categorical variables.

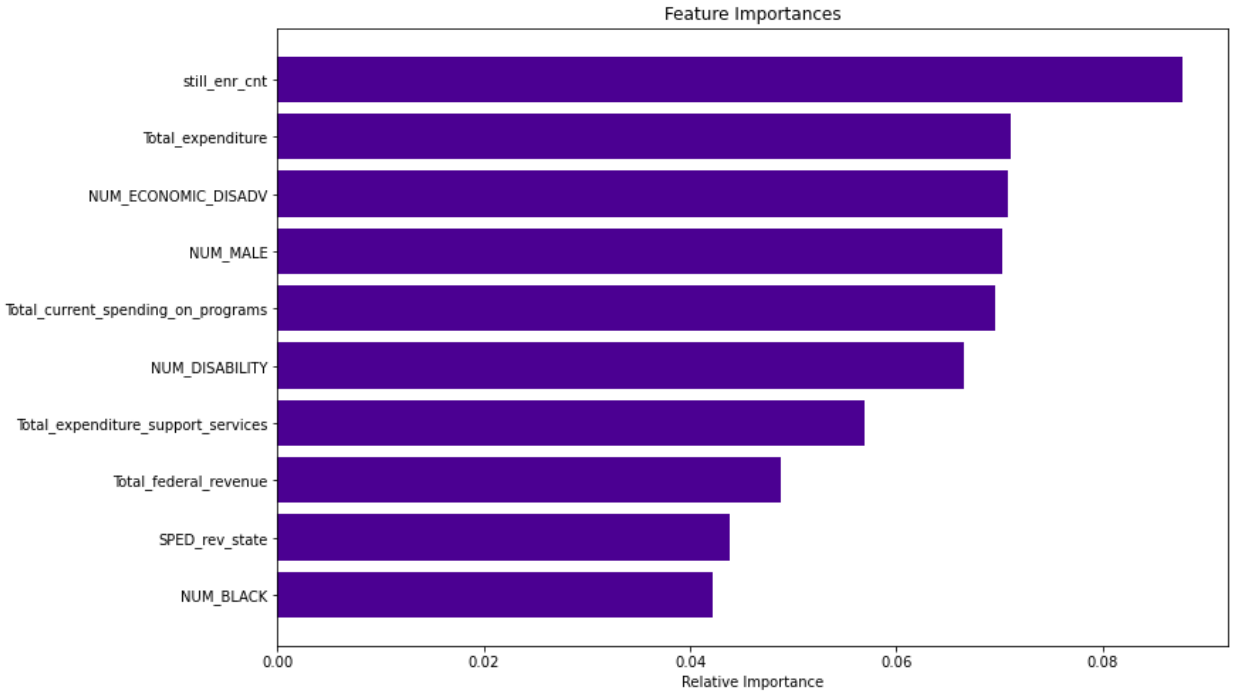
Using the Random Forest Algorithm, we first created a regression model without any hyperparameters with 100 estimators. Our accuracy score for our testing data was 0.7247, while our training score was 0.9616. This showed that our model was overfit, so we decided to introduce some hyperparameters to try and reduce the overfitting.



In our base model, we added max_depth, bootstrap, min_samples_split, min_samples, max_features, and n_estimators and input safe values in to see how our model performs. We found that the results were like our previous model and decided to introduce a GridSearchCV algorithm to help tune our hyperparameters. Using the GridSearchCV algorithm, we found the most optimal hyperparameters to be max_depth=10, max_features='sqrt', n_estimators=90, min_samples_split = 2, min_samples_leaf = 1, and bootstrap = True. For the Gridsearch we used a 3-fold cross validation with n_jobs = 1 and verbose = 0. We were able to improve our testing accuracy score to 0.9652, which is a huge increase from our previous untuned model.



We then created a horizontal bar chart to see the feature importance of the model. This helped give more solid recommendations because we can see which features had the largest impact in predicting the testing data set.



Conclusion and Recommendations

In closing, this project has given us more of an insight into how school districts are spending their money along with how the allocations of these funds in conjunction with their demographics is affecting the students. We believe that further research would need to be done to give more unique solution recommendations for each district, but we do have some general recommendations that can be applied across the board. Our recommendations include helping students and families find affordable housing, more research into how students are economically disadvantaged and how to best help those students, have an equal male to female ratio of students, spend a greater percentage of budget on staff wages, and increase pupil support expenditures.