Report: FINAL PROJECT PAPER RACHEL JONES
Introduction
Because of the pandemic, many schools have reevaluated their relationship to standardized testing a thing that was once a tool to understanding student success. Many who still believe in these scores appreciate them because they see them as the only tool for standardization. In another class, I did my project on creating a predictive model that uses non-test score factors to estimate whether an NYC high school had a goal-level graduation rate (NOTE: I DID MAKE A NEW MODEL FOR THIS CLASS).

Collecting the data for this project was arduous, and I could not find other examples of this data being used in a dynamic and predictive way. Thus I created my project to people to conceptualize how non-academic factors relate to academic success or graduation rate, and better appreciate their situation, and better grasp what graduation rates are really a reflection of. Moreover, I wanted to explore ways of presenting the inner workings of an ML model in an accessible way.

Related Work
Functionally
Whether it be Buzzfeed or, the new york times article that inspired me to take this class.(https://www.nytimes.com/interactive/2021/09/08/opinion/republicans-democrats-parties.html)
Quiz-based data narratives seem to have a way of connecting people with the data that they are learning about and make different outcomes seem personal and worth investing in.

With this storytelling assignment, I wanted to piggyback on that concept to create something compelling to viewers in this way.

Topic
Other tools like great schools, or this (https://tools.nycenet.edu/snapshot/2020/01M034/PK/#CT) snapshot visualizer that NYC and other schools use. This allows interested parties to understand how their school is doing relatively on these factors. These are very useful in understanding the school as a whole, however, they do not show overarching data trends on these factors, nor do they show importance.

This visualization has the opportunity to push farther beyond these methods by creating a more aggregate understanding.

Methods
Developing Visualization 1
To create this visualization I first crated a dataset of State Names and their stated goal graduation rates. Each state does requirements differently, so I initially thought I would need different datasets for 4, 5, & 6-year graduation data sets, however upon closer inspection, every state which had multiple graduation rates listed, used the same % for each year target, telling me that perhaps there was not too much difference, so I used the join function to make a graduation rate goal column that was agnostic of year.

Developing the Quiz
To create a more active and interactive narrative I re-ran my models using harsher feature selection and a tree-based algorithm so that they would be more suitable for a quiz. I did this using Lightside and Weka, using my development set to come up with ideas, training on my training set, testing on my testing set to develop the right model, and using the data from my holdout set as a useful evaluation of that data to impart into my article. I want to re-iterate, that I made sure that this model worked well with the data, and did not build from it backward, this was very useful, and is not statistically less helpful than the model I created earlier.

Model Statistics From Holdout Set:
Model – J48 with Attribute Selection
Kappa = .68
% Correct = 87

With this model created. I knew I wanted to represent it through interactions and questions. To ensure interactivity I used Observables input functions and added an if, then model that would calculate the direction that one results indicated that they should be in.

The Algorithm as it exists in Observable:
Where:
ELLP = % English Language Learners
ELP = Economic Need Index
RigScore = Academic Rigor Perception Score

```
{
  if(ELLP <= 5 && RigScore > .66 || ELLP > 5 && RigScore > .9 && RigScore<=.92 ||
ELLP > 5 && RigScore > .9 && RigScore>.92 && EIP <= .84 && ELLP>10){
    return "Your school IS predicted to have reached the 90% goal"
  }else {
    return "Your school IS NOT predicted to have reached the 90% goal"
  }
}
```

Ensuring the visualizations could respond to the quiz
To make the most of these inputs, I wanted to integrate their results into my other visualizations. I do this by creating a function that creates ranges for each of the data inputs. Before results could be directly compared, I had to transform the data from several inputs.

Creating Accademic Rigor Score
One of the factors that my model determined was important to this model was the academic rigor score. This comes from the annual DOE NYC Schools Survey, which is given to teachers, parents, and students. This score has been out of several different numbers in the past, so the data set reflects a norm percentage score. For the purpose of this activity, I chose to use the most recent 1–4 question format, because I do think that forced choice is an effective technique.

Keeping in mind this normed scoring, I averaged the scores from this section to genre that the Academic Rigor Score shown on the page.

Creating Paramaters
I wanted the information from the quizzes to be shown on the graphs below. To enable this to happen I had to change the scale of my inputs at the input level to work.

Results
Visualization 1 – 90%
This bar chart is intended to teach the audience about the most common 90% goal so that they can see the value in this choice as there is no national standard. Initially, I had planned to make the visualization interactive so that one could see if this applied in their state. But I realized that making this data available upfront may make an audience member not care about the findings of this research, because it does not apply directly to their state. I understand the dangers of hiding information, but I do think that this restriction in transparency can serve to help create a more maintainable goal.

The use of a bar chart was a useful choice in building this case as it makes the chosen class value very clear to the reader.

Quiz Interaction
I wanted to keep the mood of this article one narrow of discovery, so I opted to make the reveal of one's outcome a two-step process.

For slider selectors, I chose to make the initial settings at the initial split marks for those categories, to ensure that those going through the quiz, may have a memory of saying that the initial split value was wrong and they should be on a different side of it.

Visualizations of Factors
These were are all made to show one how the two factors are related, and in the grand scheme of things. I believe that using the interaction from the quiz earlier enables this to demonstrate a higher level of congruity, making th amount of questions more worth it. Additionally, because it uses that data from above it can give users a better base to evaluate where they are.

When looking at the results of one's quiz, the visualizations of each factor are intended to create a sense of grounding and understanding in the following relations:
1. Factor & Graduation Rate
2. Input & Data
3. Input & Model
The first is achieved by creating a standard X Y dost plot of the data from which the reader can sense the clustering of data and draw conclusions of its importance.

In that realm, one can see where their value for that factor, their input is in relation to other data points. With Opacity representing # of inputs, it is easy to see at a glance where the majority of schools sit on a given input.

The relationship between the input and the model is strengthened in three key ways on each graph. First, in hovering over each data point one is confronted with the other values of factors used in the model. This comparison at every stage gives the user the ability to find a peer and explore in that way. Secondly, each chart very visibly has grey bars which lay at the values for each split, serving as a visual reminder of how the tree algorithm split data. Finally, the 90% and over or not color-coding serves as an aid in a visual reminder for the user to more easily understand why a model may have made a split at that point, as they can easily see where the green squares are.

Additionally, each model is accompanied by text, which describes the factor in more detail and its importance.

I think that the connection between invested action, explanation, and visualization should be associated with more readers understanding why such factors work as measurement tools at the moment, and what we can do to make it less of an issue.

Discussion
I hope that the audience can gain a greater appreciation for how non-standardized test factors can help predict graduation rates, and where their school lies on that equation. In doing so I hope that in understanding why these non-test factors are predictive and learning remedies to address them, they feel compelled to aid in areas that are creating low support.

In doing so I hope that they learn about groups that need support if they haven't already.

I know that in a perfect world a test on academics should be neutral, but if these outside factors are predictive then they cannot actually be, no matter who writes the questions, or the graduation requirements. I hope that my audience takes away an understanding of the complexities of educational success.

Future Work
One thing that I think that my article misses out on conceptually is the ability to look at individual NYC schools through that lens. The data has the full capability of creating such specialized visualizations and as this data did come from 28 different datasets it is likely that this would be a useful tool. I had chosen to write for a more general audience and talk about predictive measures so I did miss that opportunity.

If I were to spend more time with the dataset, I know that some cool, new york-based data visualizations are possible.