Project Proposal

The topic I have chosen is New York City Schools.

Background and Motivation

I am very interested in educational technology and how technology can enhance the educational experience. When I was searching the Internet for data that would be openly available, I encountered the New York City Open Data project. Within that project, I found granular data that I thought could be put together to make a complete picture of New York City Schools. The large amount of open education data available from the New York City Open Data project solidified this choice.

A big influencing factor in selecting the topic for the project was what education data was openly available at a low enough granularity to create a meaningful visualization. I originally wanted to compare granular education data across cities, yet I found locating this data challenging, so I have narrowed the scope down to New York City only.

Project Objectives

The goal of this project is to provide a holistic picture of all New York City schools. This will include several dimensions of data collected across four different types of stakeholders: school, student, teacher, and parent. By using this visualization, the end user should be able to determine how a given school compares against all of the rest of the New York City schools and how some factors of data influence other factors of data. The benefits of this visualization include helping school stakeholders understand where the schools are excelling and can be improved, helping parent and student stakeholders select an ideal school to meet key criteria that are important to them, and informing the overall public of how the schools rate against each other.

Some of the sample types of questions that could be answered with this visualization include:

- Which school districts and schools within New York City have the best/worst:
    - Overall Rating
    - Average Class Size
    - Safety Rating
    - Quality Rating
    - Environment Rating
    - Graduation Rate
    - Drop Out Rate
    - Attendance Rate
    - English Test Scores
    - Math Test Scores
    - Teacher Scores

- How do some of the factors listed in the previous bullet point influence some of the other factors in the previous bullet point? For instance, do schools with the highest attendance have the highest graduation outcomes and does the overall rating influence the graduation outcome.
- What is the teacher, student, and parent perception of the school district?

Data

I am collecting data from:

- New York City Open Data
  - https://nycopendata.socrata.com/data?cat=education
  - School Attendance by District
  - Demographics by District
  - Class size by school
  - School Progress Report
  - Graduation Outcomes
  - NYC School Survey
  - English Test Scores
  - Math Test Scores
  - School Safety Report
  - Quality Review
  - Location [Potential]
- New York City Bytes of the Big Apple
  - http://www.nyc.gov/html/dcp/html/bytes/districts_download_metadata.shtml
  - School District Shape File [Potential]
- Newsday – New York City Department of Education Freedom of Information Act Request – Teacher Performance Data [Potential]
  - http://data.newsday.com/long-island/database/?pid=412&pid=412

All of the data collected from New York City Open Data can be downloaded to flat file or JSON. The New York City shape file data can be downloaded and converted to TopoJSON. All of the data from NYC DOE would need to be screen-scraped and copied into a flat file. The NYC DOE data is only available 100 records at a time for 17,000+ records worth of data, so it may not be practical to obtain this data.

Data Processing

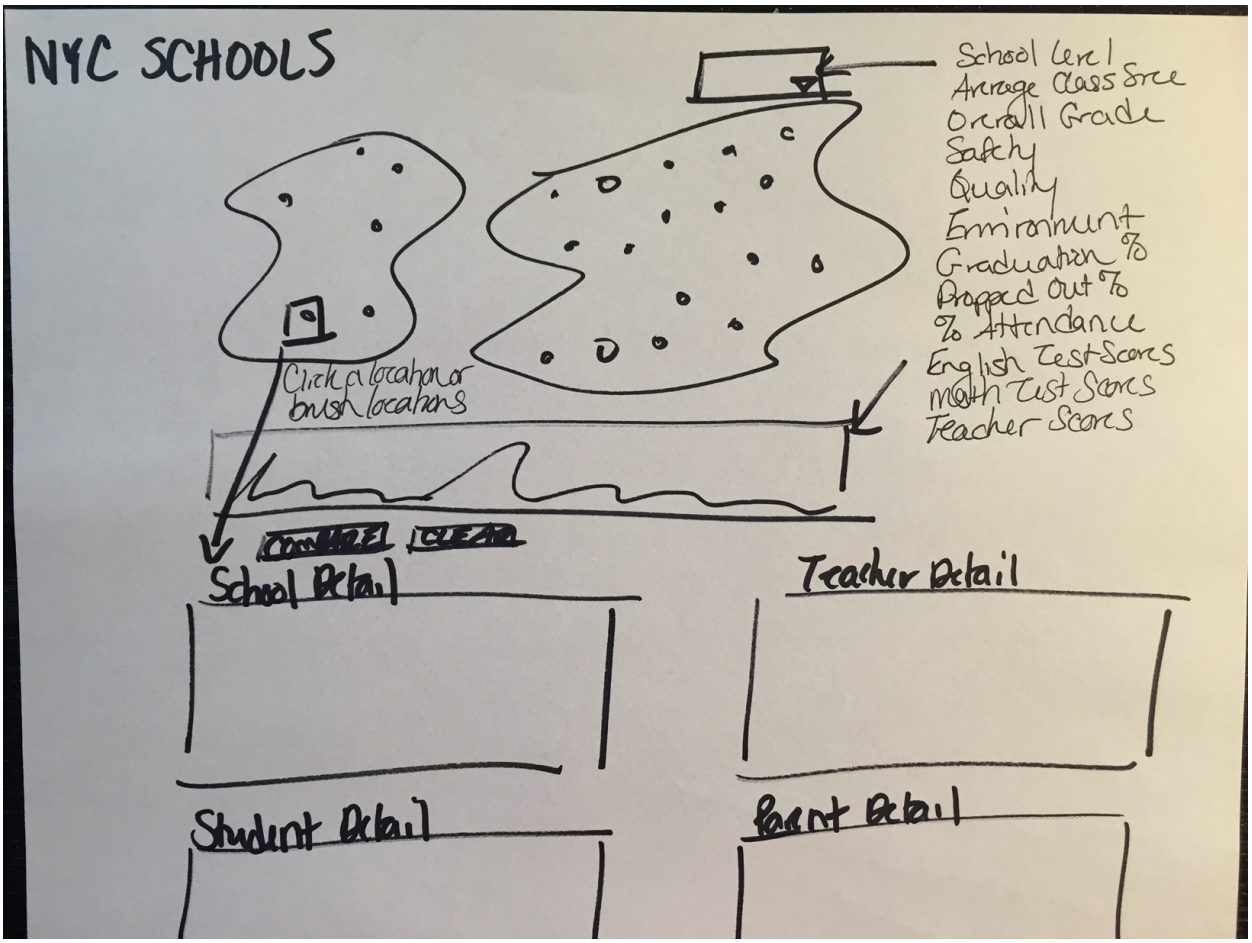
Due to the large number of data files that need to be cleansed and joined, my preference will be to do all of the tabular data processing in a database, then join it and output the result to a JSON file. If this preprocessing is allowed, I will perform all of the preprocessing and only use the preprocessed result in my final project. This will save me from writing a significant amount of JavaScript code. I can write SQL to resolve these data issues much faster than I can write

JavaScript, and I think trying to parse all this JavaScript at runtime would significantly negatively impact performance.

From a cleanup perspective, most of the data has inconsistent keys that need to be parsed to a common format. For instance, there are many different key formats across the files including "DISTRICT 01", "CSD 01 Manhattan", "M015", "01M015", etc., that need to be parsed to a common format. All of the data files will be rolled up to a Borough → School District → School level. The two datasets that are in a lower grain than that are NYC DOE teacher performance data and NYC Open Data class size by school, so these will need to be rolled up to the School level.

The shape file data needs to be converted to TopoJSON in order to use it with DataMaps. I plan to follow these instructions to attempt implementing a custom map of the NYC boroughs, school districts, and schools.

Visualization



I intend to have a map of the boroughs and schools districts in New York City with the schools represented as dots. The user will have the choice to change the metric, for instance average

class size, and the data will be encoded in the dots. I expect to only make three or so different types of dots to keep the visual complexity low, for instance, high – medium – low, or good – average – bad. The dots will update when a different metric is selected.

When a user clicks on a borough shape, school district shape, or school dot, the detail panels will present data about the school, the school's teachers, the school's students, and the school's parents. If the compare button is clicked, that school will remain selected and the user can select an additional school for comparison. There will also be an option to compare two metrics for a school against each other, such as attendance rate and graduation rate.

If I cannot get the map to function, I will present the boroughs, school districts, and schools as nodes and enable the same functionality described. I find having the option of a fallback is important because users have described some difficulties in getting the shape files to convert properly and display with the correct projection.

Must-Have Features

- High level view of all schools, detail level view of a specific school
- Data presented with granularity of Borough → School District → School
- Data from school, student, teacher, and parent stakeholders
- The ability to compare two schools
- The ability to compare two metrics for a school
- At least 6 of the following metrics:
    - School Attendance by District
    - Demographics by District
    - Class size by school
    - School Progress Report
    - Graduation Outcomes
    - NYC School Survey
    - English Test Scores
    - Math Test Scores
    - School Safety Report
    - Quality Review

Optional Features

- Map of boroughs, school districts, and schools
    - Fallback will be nodes
- Brush of several schools to compare
    - This will likely not be practical. Fallback will be clicking one school node, clicking compare, and clicking a second school node.
- Additional metrics

Project Schedule

All dates listed are end dates.

4/10 – Merge and cleanse all data. Test getting DataMaps to work and make a go/no go decision on the map versus the nodes.
4/17 – Milestone 1 – Full data structure, working prototype with a metric
4/26 – TF Project Review – Build out the rest of the metrics
5/5 – Final Project Due – Create the screencasts and wrap up the documentation