Project team 1 – RAM project approach and story

The project team’s original idea was to tell a story about schools in New York City. We asked ourselves: “what are some of the most likely major impacts on school performance.” The group quickly arrived at three major categories:

* Crime (within the schools) – more incidents in the school would typically mean that there are more distractions and less time for learning
* Median Household Income – those that are born into a higher standard of living would be more likely to have the means to afford tutors and the free time to complete class work at home.
* Educational attainment in the family – It is assumed that if the parents have a higher level of education, they would place a greater emphasis on their children’s performance.

The team then began gathering data from NYC’s open data portal. This included the following:

* Crime data within the schools where the students are learning
* Education performance for NYC High Schools, which includes graduation rates and admittance into college.
* Census data, which would cover the median household income and educational attainment variables.

Each member of the team took a dataset to research and better understand. A few key data development decisions occurred early in the process:

* The granularity of the geographic area to be analyzed was critical to establish. The only way that the data could be understood was if all of the datasets had a common link and aggregation level.
  + Census data was typically in the census tract level, which is too detailed for the NYC datasets.
  + School crime data was only available at the School building ID level, which was different than the other data sets.
* New York City did have census data at the Neighborhood Tabulation Area (NTA) level, which meant that the direct US Census download would not be needed for this project.
* The group found that the NTA geography was the common denominator between all the datasets. This became the primary key for our datasets.
* The census data provided by NYC Open data was not formatted well, with custom headers and merged cells in an Excel file. At the same time, it was discovered that median household income was not included in this dataset. We found another dataset that cleaned up the NTA data provided by the NYC Open data and included the median household income. This became the primary source of census data for our project.
* The school crime data was not sufficient to fit into our story. There were not enough years of data (only one intersecting year with school performance), not enough data records, the overall numbers were too low. Once this was discovered our data was expanded to include crime data statistics from NYPD.

The next step for the project team was to ensure that the different datasets talked to each other. We were able to connect the data sets in the following ways:

* School Crime Data 🡪 School Performance.
  + Primary Key: DBN (District Borough Number)
  + Years of overlap: 1
  + Merge these two datasets for further analysis
* School Performance 🡪 Median Household Income & Educational Attainment
  + Primary Key: Neighborhood Tabulation Area (NTA) name (not code).
* School Crime Data 🡪 Median Household income & Educational Attainment
  + Primary Key: NTA. Connection only possible passing through the School Performance dataset.
* School Performance 🡪 Crime Data (NYPD)
  + Data connection / adjustment:
  + Primary Key: Precinct

While connecting and merging data, we discovered several key data elements that needed to be resolved / overcome:

* The use of the NTA name was concerning. We were initially unable to connect the census dataset to the others, until it was found that the Census dataset contained trailing spaces in the NTA name. Once this was discovered and the spaces were stripped, the datasets were connected.
* The school crime data numbers were based on the building that held the schools. If there were multiple schools in a single building, the crime data was aggregated into a consolidated number. This resulted in a smaller than anticipated number of data connections between the school crime dataset and the others.
* The census data for education attainment was based on the population over 25 years of age. Unfortunately, the census data acquired did not include the total population over 25 years of age, which is required to compare educational attainment across the different NTAs.
  + The team compared the individual education attainment values to the values in the originally-downloaded US Census dataset.
  + It was found that adding two of the fields in the newer dataset would result in the total population as needed.
* We needed to coordinate the years of analysis so they were meaningful.

Conclusions

The team found the following as a result of their project work:

* There was a correlation between the schools with the highest and lowest crime rates and their associated graduation rates.
* Educational Attainment was also a strong indicator of school performance.
* High schools with a low (0-30%) college rate had a significantly higher average felony rate. High schools with a high (85-100%) college rate had a significantly lower average felony rate.