Applied Data Science Capstone Project

Introduction:

New York City is one of the most diverse cities in the world, home to millions of immigrants from all around the world. It is one of the world's most populous cities, and is America's largest city, home to about 8.4 million people. The surrounding NYC Metropolitan Area is home to over 23 million people, and continues to grow to this day.

With such a diverse and populous citizen base, New York City is home to hundreds of high schools for the thousands of students who study in this metropolitan area. Schools are incredibly important to advance the education of its students, to prepare them for college education, the work force, and basic life skills. All in all, good schools produce good students, who will likely go on to be successful in their endeavors. Nowadays, one of the key features that highlight a school's level is the average SAT score of its students. The SAT is a standardized test which is required in college admissions. Higher scores on the exam scored out of 2400 represent students who likely study more than students with lower scores. Therefore, average SAT scores show a good metric of how to rank NYC high schools. The higher the average SAT score of the school, the better it likely is due to its smarter student body.

Problem:

Suppose a family is moving to New York City and they have children they need to put through school. Families with children likely care about the child's education and want to do everything they can to ensure that their child goes to a good school that will put them in a position to succeed later in life. Therefore, we will be trying to answer the following questions:

Which borough has the highest mean average SAT score?
Which borough(s) has the most schools that have average SAT scores higher than the national average?
What venues surrounding these high performing schools seem to be recurring?

Data:

To effectively carry out this project, we need to obtain information about the different high schools in New York City, their location, borough, and average SAT scores. We also need to obtain the neighboring venues to each of these schools.

To satisfy the requirements of the first dataset, I used a dataset found on kaggle.com, a renowned data science website that contains a multitude of free data sets. The exact URL to this dataset is: https://www.kaggle.com/new-york-city/new-york-city-sat-results. This dataset contains information of over 450 NYC high schools, their names, borough, latitude/longitude values, average SAT scores by section (math, reading, writing), etc. With this, we can explore each of the schools and where the better schools are located in NYC by exploring their locations and their average SAT scores.

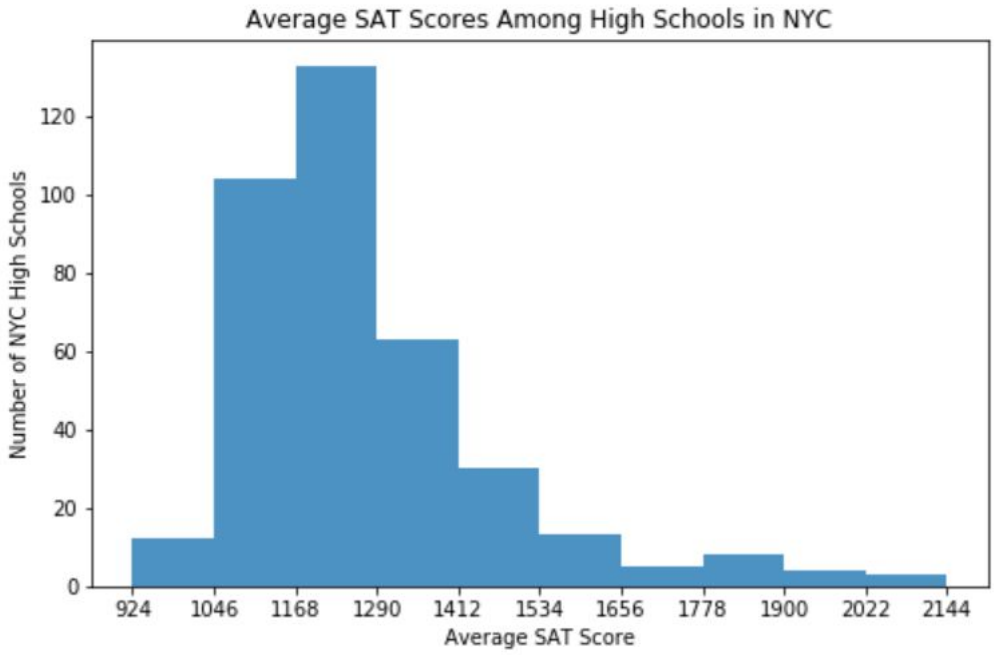
To satisfy the second requirement of obtaining venues, I used the Foursquare API. By calling the Foursquare API, I was able to retrieve a number of venues surrounding each school and from there explore venues around certain schools.

Methodology:

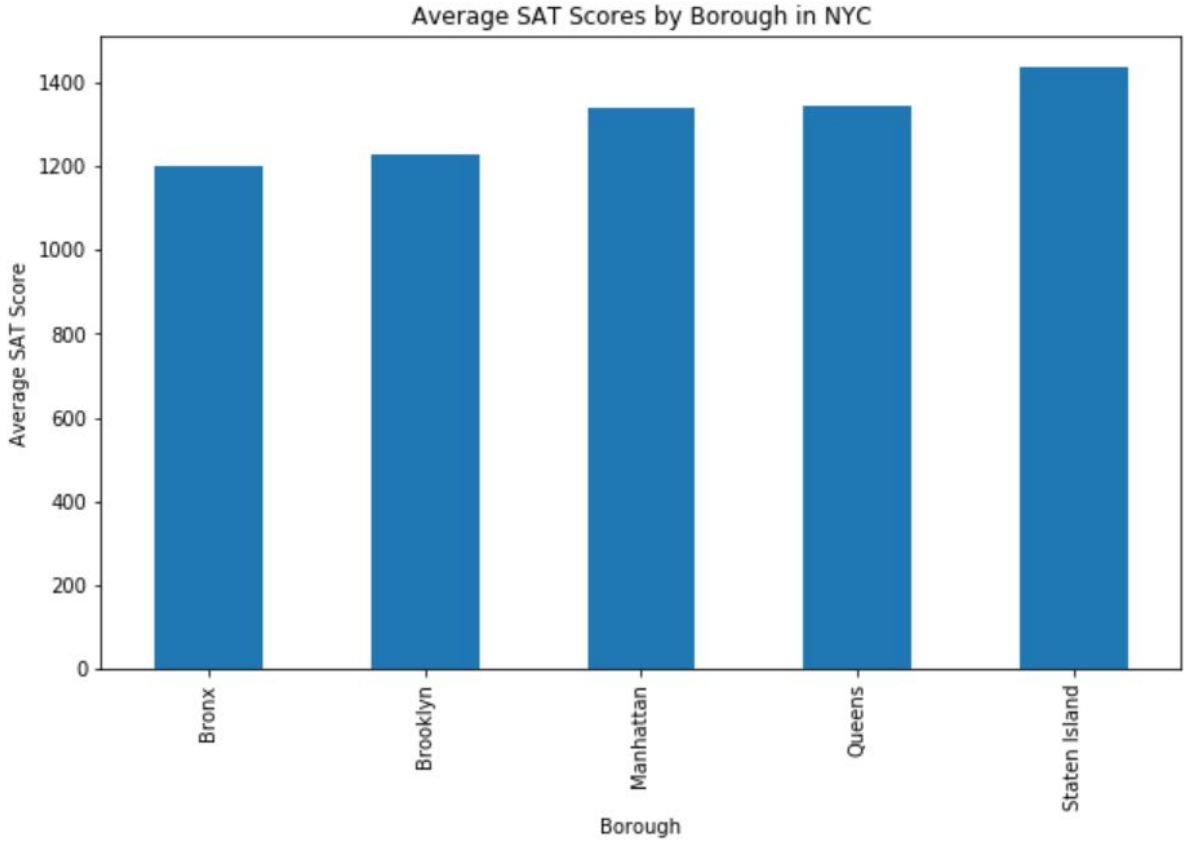
First, I downloaded the data from https://www.kaggle.com/nycopendata/high-schools and turned it into a pandas dataframe

	School ID	School Name	Borough	Building Code	Street Address	City	State	Zip Code	Latitude	Longitude	...	End Time	Student Enrollment	Percent White	Percent Black	Percer Hispani
0	02M260	Clinton School Writers and Artists	Manhattan	M933	425 West 33rd Street	Manhattan	NY	10001	40.75321	-73.99786	...	NaN	NaN	NaN	NaN	Na
1	06M211	Inwood Early College for Health and Informatio...	Manhattan	M052	650 Academy Street	Manhattan	NY	10002	40.86605	-73.92486	...	3:00 PM	87.0	3.4%	21.8%	67.8
2	01M539	New Explorations into Science, Technology and ...	Manhattan	M022	111 Columbia Street	Manhattan	NY	10002	40.71873	-73.97943	...	4:00 PM	1735.0	28.6%	13.3%	18.0
3	02M294	Essex Street Academy	Manhattan	M445	350 Grand Street	Manhattan	NY	10002	40.71687	-73.98953	...	2:45 PM	358.0	11.7%	38.5%	41.3
4	02M308	Lower Manhattan Arts Academy	Manhattan	M445	350 Grand Street	Manhattan	NY	10002	40.71687	-73.98953	...	3:00 PM	383.0	3.1%	28.2%	56.9
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
430	27Q302	Queens High School for Information, Research, ...	Queens	Q465	8-21 Bay 25th Street	Far Rockaway	NY	11691	40.60199	-73.76283	...	4:10 PM	381.0	2.1%	49.1%	43.6

Then I did some initial exploratory analysis after cleaning up the dataframe. I wanted to see what the distribution of average SAT scores was.

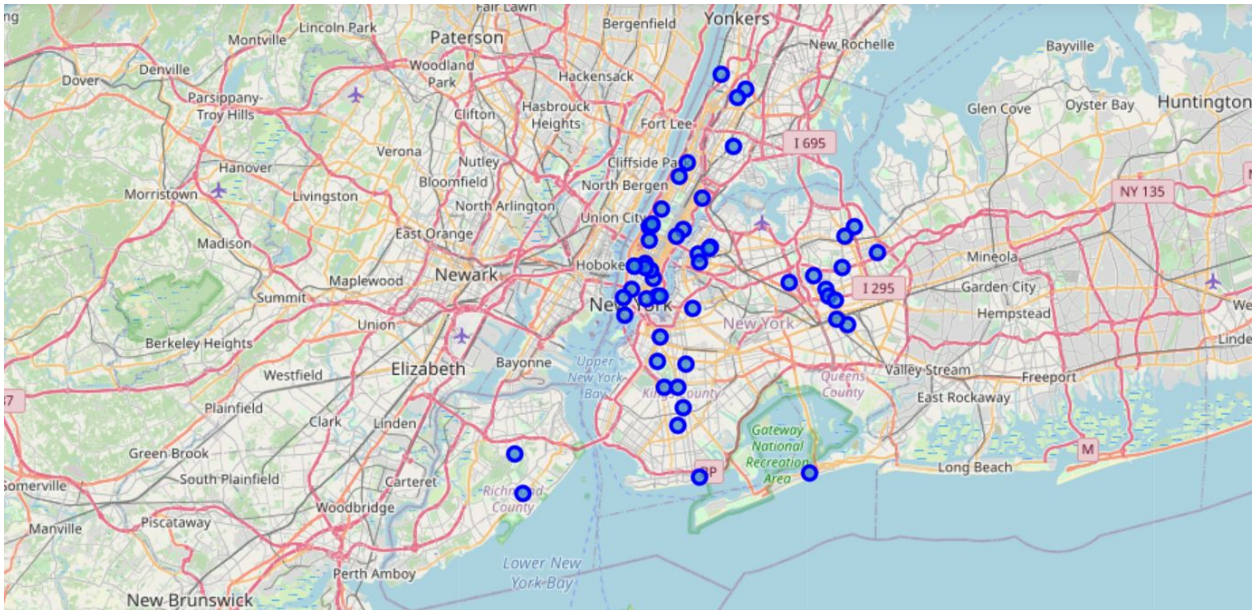


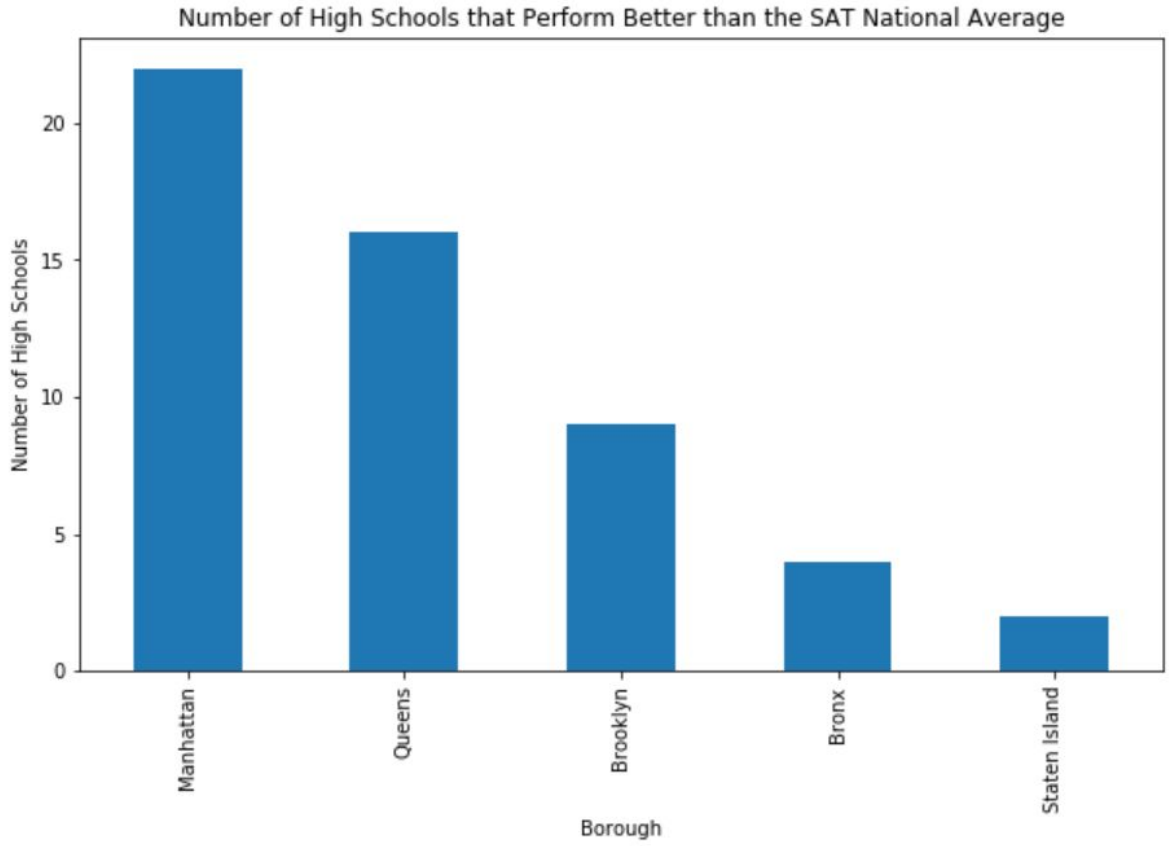
I also wanted to see what borough had the highest average SAT score



I then wanted to focus only on the schools with average SAT scores above 1450, which is around

the national average, specifically where they were located and which boroughs had the most of such schools.



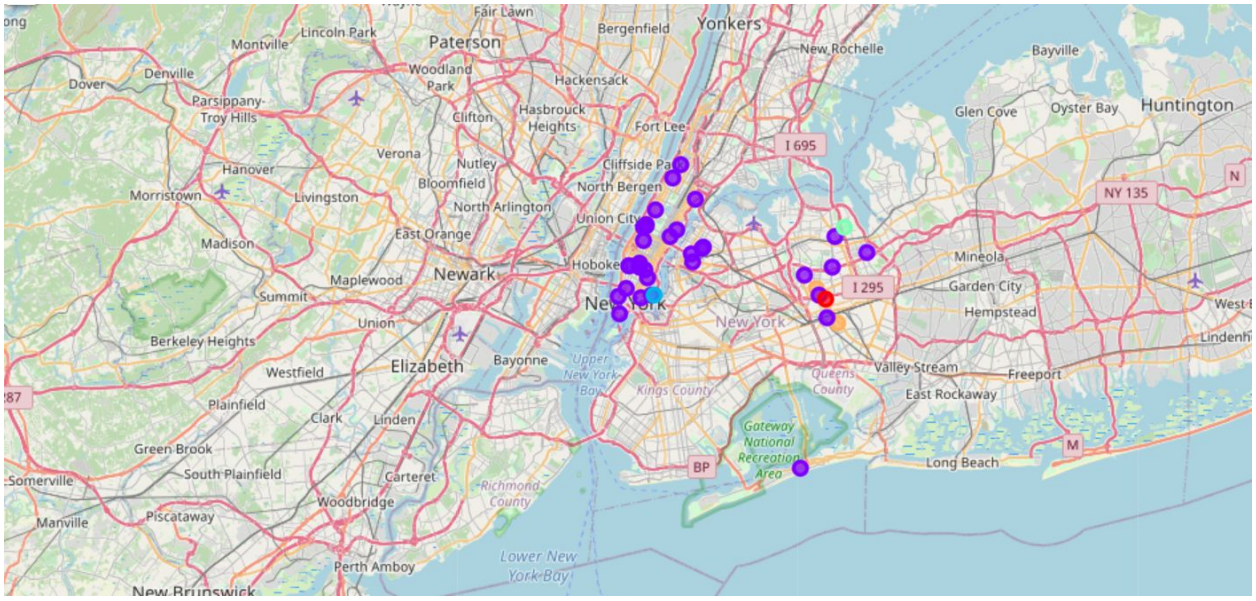


From this, it's evident to see most of the high performing high schools are located within the Manhattan/Queens boroughs. I then made a new dataframe containing only the above average

schools from these two boroughs and generated the nearby venues for each school to explore if there were any patterns through K Means Clustering.

	School ID	School Name	Borough	Building Code	Street Address	City	State	Zip Code	Latitude	Longitude	...	Student Enrollment	Percent White	Percent Black	Pe His
0	01M539	New Explorations into Science, Technology and ...	Manhattan	M022	111 Columbia Street	Manhattan	NY	10002	40.71873	-73.97943	...	1735.0	28.6%	13.3%	1
1	02M545	High School for Dual Language and Asian Studies	Manhattan	M445	350 Grand Street	Manhattan	NY	10002	40.71687	-73.98953	...	416.0	1.7%	3.1%	
2	01M696	Bard High School Early College	Manhattan	M097	525 East Houston Street	Manhattan	NY	10002	40.71896	-73.97607	...	545.0	45.3%	17.2%	1
3	02M407	Institute for Collaborative Education	Manhattan	M475	345 East 15th Street	Manhattan	NY	10003	40.73249	-73.98305	...	482.0	56.5%	14.1%	1
4	02M418	Millennium High School	Manhattan	M824	75 Broad Street	Manhattan	NY	10004	40.70492	-74.01151	...	659.0	32.8%	7.6%	1

	School Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Academy of American Studies	Hotel	Coffee Shop	Bar	Pizza Place	Deli / Bodega	Café	Donut Shop	Cocktail Bar	Italian Restaurant	Mexican Restaurant
1	Baccalaureate School for Global Education	Mexican Restaurant	Bar	Gym	Discount Store	Athletics & Sports	Italian Restaurant	Seafood Restaurant	Bakery	Restaurant	Café
2	Bard High School Early College	Track	Park	Tennis Court	Baseball Field	Nightclub	Yoga Studio	Electronics Store	Dumpling Restaurant	Eastern European Restaurant	Empanada Restaurant
3	Bard High School Early College Queens	Café	Coffee Shop	Bakery	Halal Restaurant	Deli / Bodega	Donut Shop	Juice Bar	Tennis Court	Cafeteria	Cocktail Bar
4	Baruch College Campus High School	Coffee Shop	Hotel	Yoga Studio	Gym / Fitness Center	Vegetarian / Vegan Restaurant	Café	Hotel Bar	Japanese Restaurant	Grocery Store	American Restaurant
	Bayside High	Farmers	Gym /	Empanada				Dumpling	Eastern	Electronics	English



Results:

We can answer the questions from the Problem section.

From the analysis, Staten Island had the highest average SAT scores from its high schools, but likely due to its smaller number of high schools. The boroughs that had the most schools that had an average SAT score higher than the national average were Manhattan, followed by Queens. Lastly, from the clusters and the venue information generated by the Foursquare API, some recurring venues tend to be coffee shops and various types of restaurants, including Indian, Chinese, American, and Italian.

Discussions:

From this preliminary analysis made from one single dataset of SAT scores of various NYC high schools, I'd recommend parents who want to enroll their children into good schools to stay in either Manhattan or Queens. Though Staten Island had the highest average SAT scores amongst the boroughs, it was due to the fact that it had the fewest number of high schools of all the boroughs, and there were far more schools that performed above the national average in the other two boroughs. However, it is important to be wary of these results since it was all based off of one dataset primarily based on average SAT scores. Though the SAT is a good metric to measure a school's rating, there are a number of methods that can be done to improve the analysis of what is a good school and what isn't for new parents to NYC.

Conclusion:

The above project can be refined with more data, maybe from better sources and in higher quantity. Another way to improve the results would be to include other metrics that demonstrate high performing schools. SAT scores are only one metric to provide a basic analysis of a school's level of education, but there are others. The solution provided is by no means the correct solution or a completely accurate portrayal and analysis of the data and can easily be improved depending on the availability of more data.