Where to Live-College Edition

Let the data help you choose the neighborhood

August 31, 2020

1.1 Introduction

Second to choosing a college or university, where to live "off campus" can be the most stressful decision students and parents must make. There are numerous factors to consider about a city the student may have zero knowledge of: safety, transportation, shopping, restaurants, fitness, social scene, etc. While there may be a plethora of apps now to close the gap on knowledge about restaurants, shopping, and social connections, other information continues to be more elusive. Understanding how to compare neighborhoods based on safety is easily one of the more difficult. While that information should be readily available to the public, law enforcement may not make it easy to obtain and review for the average person. To make an unbiased decision based on multi-factor considerations: safety, distance, quality of life, becomes nearly impossible without a system to chart, visualize and map the information. My proposed business challenge is the first step in providing a solution to aid parents and students in making that decision using a data-driven process.

1.2 Problem

Data may be readily available, but it is of disparate sources, types, and relationships. In order for students to weigh them together, we must bring them together for consideration. For example, a choropleth map of crime rates labeled with the neighborhoods recommended by the university. When you can start to tie individual pieces of information together for the customer, it becomes easier to make a decision.

1.3 Interest

In a world of international students, helicopter parents, and Varsity Blues parents buying their children's way into college, the demand to ensure students have a safe and enjoyable place to live is without question. However, not all parents can afford to hire fancy real estate agents to do the heavy lifting for them to determine the best neighborhoods for low crime, best food, bars, yoga. That is where an app could come in and provide similar service at a lower price point to those with more modest budgets.

2. Data acquisition and cleaning

2.1 Data sources

The first step was identifying a university to demo the capability. Columbia University in New York City, NY, was chosen because there was a plethora of data readily available for this project. The geolocational data required included boroughs, neighborhoods, latitudes, longitudes, which was available in the

NYC (JSON): https://cocl.us/new_york_dataset

The crime data was comprehensive in both coverage of Major Crime Indicators (MCI) and of Precinct areas covered in all of New York City, beyond Manhattan – the area than what was needed for our project. The crime data was obtained from the following sources:

Crimes: https://data.cityofnewyork.us/Public-Safety/NYC-crime-qb7u-rbmr

GeoJSON: https://raw.githubusercontent.com/dwillis/nyc-maps/master/police_precincts.geojson

2.2 Cleaning

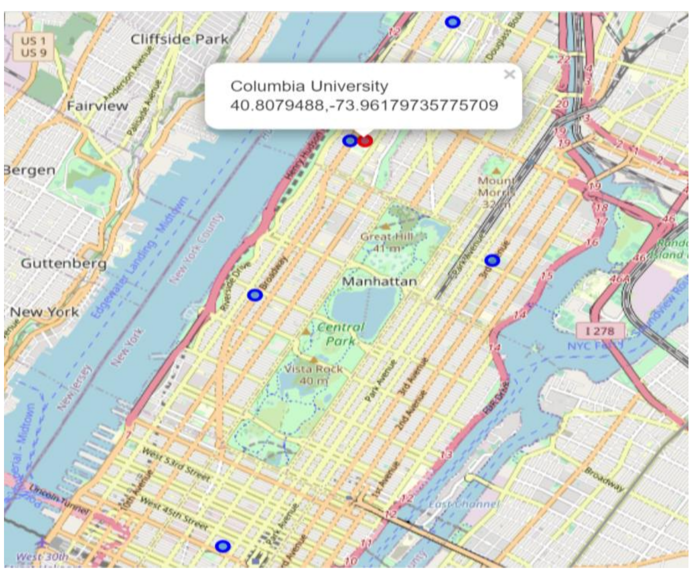
Where possible, data was restricted to only the neighborhoods of interest, or only Manhattan. GeoJSONs did not allow for easy editing and were not narrowed to Manhattan.

3. Methodology

Folium maps were created to show the location of Columbia University and the neighborhoods that were recommended by the School of International Affairs and Public Administration student blog. This was the starting point. After crime rates were calculated, they were mapped on a choropleth map with pop-up labels to easily identify the neighborhoods with highest/lowest crime rates. Finally, when the FourSquare survey was complete and the venues/services had been assessed in each of the neighborhoods, the K-Means clustering analysis was performed on the results. This allowed analyzing which neighborhoods share similarities or clusters and which stood out from the rest. Tables were also created to display the distance from the centroid of the neighborhood to Columbia University, to compare relative crime rates, and to compare top venues.

Results

As stated above, the first step was to map the locations of Columbia University relative to the neighborhoods. Each circle is a neighborhood with a pop-up label.



Next, charts were prepared to convey latitude, longitude and distance from centroid of the neighborhood to the center of Columbia University's campus.

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688
5	Manhattan	Manhattanville	40.816934	-73.957385
6	Manhattan	Central Harlem	40.815976	-73.943211
7	Manhattan	East Harlem	40.792249	-73.944182
8	Manhattan	Upper East Side	40.775639	-73.960508
9	Manhattan	Yorkville	40.775930	-73.947118

```
Distance in km between Columbia and Hamilton Heights is:
2.0165162360278153  kms
Distance in km between Columbia and Manhattanville is:
1.0650302019280105  kms
Distance in km between Columbia and Central Harlem is:
1.8038495539117967  kms
Distance in km between Columbia and East Harlem is:
2.2911192832517915  kms
Distance in km between Columbia and Upper West Side is:
2.595422209139266  kms
Distance in km between Columbia and Midtown is:
6.147523013104539  kms
Distance in km between Columbia and Morningside Heights is:
0.17719496689619196  kms
Distance in km between Columbia and Columbia University is:
0.0  kms
```

Next, the crime rates were calculated in several steps. First, an overall sum of Major Crime Indicators was calculated, as was a percentage.

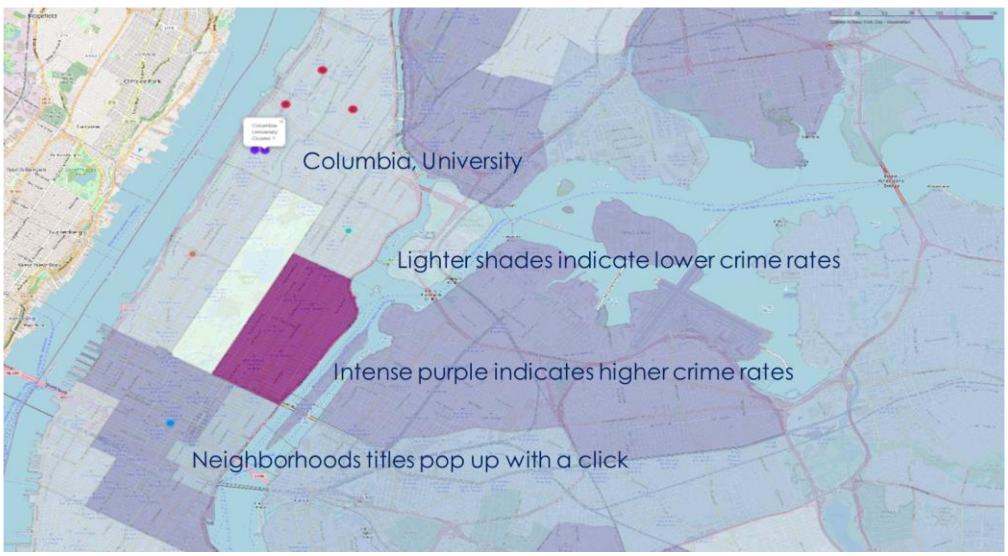
	MCI	Number of Occurrence in 2019	Percentage of Occurrence in 2019
0	PETIT LARCENY	1171.0	22.813170
1	GRAND LARCENY	1104.0	21.507890
2	HARRASSMENT 2	512.0	9.974674
3	CRIMINAL MISCHIEF & RELATED OF	466.0	9.078512
4	OFF. AGNST PUB ORD SENSBLTY &	339.0	6.604325
5	THEFT-FRAUD	280.0	5.454900
6	SEX CRIMES	190.0	3.701539
7	ASSAULT 3 & RELATED OFFENSES	153.0	2.980713
8	MISCELLANEOUS PENAL LAW	135.0	2.630041
9	RAPE	96.0	1.870251
10	BURGLARY	80.0	1.558543
11	FRAUDS	80.0	1.558543
12	UNAUTHORIZED USE OF A VEHICLE	77.0	1.500097
13	DANGEROUS DRUGS	71.0	1.383207
14	FELONY ASSAULT	65.0	1.266316
15	GRAND LARCENY OF MOTOR VEHICLE	58.0	1.129944
16	OFFENSES AGAINST PUBLIC ADMINI	45.0	0.876680
17	FORGERY	43.0	0.837717
18	NYS LAWS-UNCLASSIFIED FELONY	31.0	0.603935
19	ADMINISTRATIVE CODE	20.0	0.389636
20	ARSON	15.0	0.292227
21	ROBBERY	15.0	0.292227
22	CRIMINAL TRESPASS	13.0	0.253263
23	VEHICLE AND TRAFFIC LAWS	11.0	0.214300
24	OFFENSES AGAINST THE PERSON	9.0	0.175336
25	POSSESSION OF STOLEN PROPERTY	9.0	0.175336
26	OTHER OFFENSES RELATED TO THEF	9.0	0.175336
27	THEFT OF SERVICES	7.0	0.136372
28	DANGEROUS WEAPONS	7.0	0.136372
29	OTHER STATE LAWS (NON PENAL LA	5.0	0.097409
30	MURDER & NON-NEGL. MANSLAUGHTER	5.0	0.097409
31	ANTICIPATORY OFFENSES	3.0	0.058445
32	INTOXICATED & IMPAIRED DRIVING	3.0	0.058445
33	FRAUDULENT ACCOSTING	2.0	0.038964
34	OFFENSES INVOLVING FRAUD	2.0	0.038964
35	PROSTITUTION & RELATED OFFENSES	2.0	0.038964

Next, the individual precincts were considered in order to construct a choropleth map to augment our neighborhood mapping.

	Precinct Number	Number of Incidents
0	75	156.0
1	19	155.0
2	113	136.0
3	70	111.0
4	105	110.0
5	67	107.0
6	18	100.0
7	44	100.0
8	14	100.0
9	114	98.0

	Precinct Number	Borough	Occurrence Year	MCI	Lat	Long	Coordinates
0	5	MANHATTAN	2019	SEX CRIMES	40.716196	-73.997491	(40.716195914000025, -73.99749074599998)
1	23	MANHATTAN	2019	OFF. AGNST PUB ORD SENSBLTY &	40.799665	-73.947200	(40.799665264000055, -73.94719977999995)
2	5	MANHATTAN	2019	RAPE	40.716196	-73.997491	(40.716195914000025, -73.99749074599998)
3	5	MANHATTAN	2019	PETIT LARCENY	40.714431	-74.006101	(40.714430898000046, -74.00610127799997)
4	9	MANHATTAN	2019	DANGEROUS DRUGS	40.722397	-73.978536	(40.72239709900003, -73.97853584199999)
5	18	MANHATTAN	2019	THEFT-FRAUD	40.770827	-73.992611	(40.770827222000044, -73.99261118899994)
6	14	MANHATTAN	2019	GRAND LARCENY	40.747881	-73.991040	(40.74788104700008, -73.99104019099997)
7	19	MANHATTAN	2019	GRAND LARCENY	40.775773	-73.954750	(40.77577282900006, -73.95475034499998)
8	30	MANHATTAN	2019	PETIT LARCENY	40.824602	-73.950114	(40.82460236600008, -73.95011391699995)
9	19	MANHATTAN	2019	OFF. AGNST PUB ORD SENSBLTY &	40.767336	-73.954875	(40.76733554700007, -73.95487521099994)

Here is the choropleth map that resulted from this analysis:



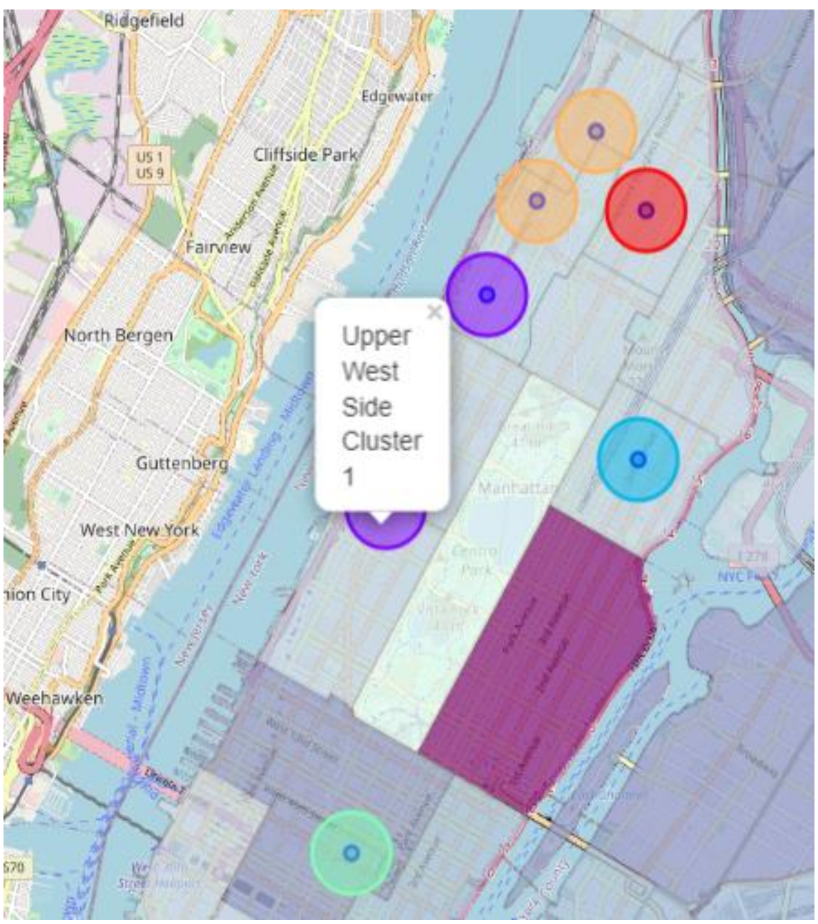
A survey of FourSquare data was conducted for “nearby venues” using the neighborhoods above and their latitudes, and longitudes. This provided 164 unique venue categories and a great starting point for analyzing the top 10 most common venues in each neighborhood. It also provided a basic grouping of data to conduct a K-means analysis for clustering.

	Location	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	Central Harlem	Southern / Soul Food Restaurant	African Restaurant	Café	Pizza Place	Seafood Restaurant	Sushi Restaurant	French Restaurant	Bar	Lounge	American Restaurant
1	East Harlem	Mexican Restaurant	Café	Bakery	Pizza Place	Deli / Bodega	Plaza	Thai Restaurant	Italian Restaurant	Coffee Shop	Gym
2	Hamilton Heights	Coffee Shop	Park	Mexican Restaurant	Bar	Café	Yoga Studio	Ethiopian Restaurant	Sushi Restaurant	Deli / Bodega	Chinese Restaurant
3	Manhattanville	Park	Italian Restaurant	American Restaurant	Seafood Restaurant	Mexican Restaurant	Café	Coffee Shop	Cocktail Bar	Indian Restaurant	Tennis Court
4	Midtown	Theater	Plaza	Steakhouse	Coffee Shop	American Restaurant	Gourmet Shop	Hotel	Bookstore	Concert Hall	Cuban Restaurant
5	Morningside Heights	Coffee Shop	Park	Italian Restaurant	American Restaurant	Chinese Restaurant	Grocery Store	Playground	Bookstore	Bakery	Mexican Restaurant
6	Upper West Side	Italian Restaurant	Coffee Shop	Bakery	Café	Gym	Park	American Restaurant	Wine Bar	Bar	Ice Cream Shop

The K-means clustering tells us that Hamilton Heights and Manhattanville are similar in venue options, just as are Morningside Heights and Upper West Side are similar to each other.

	Borough	Location	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Manhattan	Central Harlem	40.815976	-73.943211	0	Southern / Soul Food Restaurant	African Restaurant	Café
1	Manhattan	East Harlem	40.792249	-73.944182	2	Mexican Restaurant	Café	Bakery
2	Manhattan	Hamilton Heights	40.823604	-73.949688	4	Coffee Shop	Park	Mexican Restaurant
3	Manhattan	Manhattanville	40.816934	-73.957385	4	Park	Italian Restaurant	American Restaurant
4	Manhattan	Midtown	40.754691	-73.981669	3	Theater	Plaza	Steakhouse
5	Manhattan	Morningside Heights	40.808000	-73.963896	1	Coffee Shop	Park	Italian Restaurant
6	Manhattan	Upper West Side	40.787658	-73.977059	1	Italian Restaurant	Coffee Shop	Bakery

This becomes easier to visualize when the clusters are brought together on the choropleth map where each colored circle marker represents a different cluster.



5. Example

If we consider the demo student from the presentation, the results might look something like this:

1. Our student chooses the Upper West Side based on the access to ice cream and wine bars close to home.
2. In addition to the SIPA blog, our data could provide the following:
3. The location is actually the shown on the choropleth map on the previous page. If you notice the background color is tinted quite light. That suggests that enjoyed a relatively low rate of crime in 2019.

```
Distance in km between Columbia and Upper West Side is:
2.595422209139266  kms
```

Distance to Columbia University from the Centroid of the Neighborhood

	Location	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	Central Harlem	Southern / Soul Food Restaurant	African Restaurant	Café	Pizza Place	Seafood Restaurant	Sushi Restaurant	French Restaurant	Bar	Lounge	American Restaurant
1	East Harlem	Mexican Restaurant	Café	Bakery	Pizza Place	Deli / Bodega	Plaza	Thai Restaurant	Italian Restaurant	Coffee Shop	Gym
2	Hamilton Heights	Coffee Shop	Park	Mexican Restaurant	Bar	Café	Yoga Studio	Ethiopian Restaurant	Sushi Restaurant	Deli / Bodega	Chinese Restaurant
3	Manhattanville	Park	Italian Restaurant	American Restaurant	Seafood Restaurant	Mexican Restaurant	Café	Coffee Shop	Cocktail Bar	Indian Restaurant	Tennis Court
4	Midtown	Theater	Plaza	Steakhouse	Coffee Shop	American Restaurant	Gourmet Shop	Hotel	Bookstore	Concert Hall	Cuban Restaurant
5	Morningside Heights	Coffee Shop	Park	Italian Restaurant	American Restaurant	Chinese Restaurant	Grocery Store	Playground	Bookstore	Bakery	Mexican Restaurant
6	Upper West Side	Italian Restaurant	Coffee Shop	Bakery	Café	Gym	Park	American Restaurant	Wine Bar	Bar	Ice Cream Shop

Ten Most Common Venues in the Neighborhood (notice the commonalities with cluster partner Morning Side Heights).

5. Future Efforts

While not possible in this iteration, future efforts would seek to add simple transportation/pedestrian paths with distance from the neighborhood to the university. By adding all of this data to a single map with ability toggle layers, students and parents would truly be empowered to make data driven decisions.