Data Science Capstone Project

Enhanced Manhattan Property Search

Introduction

This project builds on the "Segmenting and Clustering Neighborhoods in New York City" exercise done previously. When thinking about what to do for the capstone project, I thought about how useful the neighborhood venue information would be to a client looking for a house or apartment. For this prototype, the client is looking for an apartment in Manhattan. They are totally unfamiliar with Manhattan, but do know the kinds of venues that they want available in their new neighborhood. The client wants to require that the neighborhood have a "Grocery Store", "Coffee Shop" and "Pharmacy" available. My prototype will provide this search method and provide both an ordered list of matching properties based on their "venue score" and a map of the matching properties. The persons most interested in this information would be the clients searching for property containing specific venues. The persons most interested in adding this feature would be property search providers.

Data

- The fundamental neighborhood data was provided in the earlier exercise in the file newyork_data.json. Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.
- As in the earlier exercise, venue information for the neighborhoods was acquired from the Foursquare api at "api.foursquare.com".
- Candidate property data was acquired from New York City sales data at https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page. I used the specific Manhattan data from this resource. I was hoping to find some actual active property search data, but was unable to find a free resource. This would not be an issue with a real-world application as the controller of the property data would be the same entity interested in using the venue data.

Methodology

There are three major sections included in my notebook. First is the acquisition and processing of New York neighborhood venue data. Second is the acquisition and processing of available property data. Third is utilizing the above data to provide a client with properties meeting their criteria.

Acquisition and processing of New York neighborhood venue data

This section was just a duplication of part of the exercise "Segmenting and Clustering Neighborhoods in New York City".

- After loading the provided New York City neighborhood data, it is filtered down to only neighborhoods in the Manhattan borough. Following is a snapshot of some of the neighborhoods and a map of the Manhattan neighborhoods.

```
The Manhattan dataframe has 40 neighborhoods.
```

	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

- Based on the latitude and longitude of each neighborhood Foursquare is used access nearby venue information.
- The venue data is then processed to establish a neighborhood/venue-type dataframe, where each entry represents the frequency of that venue-type in that neighborhood. Following is a snapshot of part of this dataframe.

	Neighborhood	Accessories Store	Acupuncturist	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant
0	Battery Park City	0.0	0.0	0.0	0.0	0.000000	0.010753
1	Carnegie Hill	0.0	0.0	0.0	0.0	0.000000	0.011111
2	Central Harlem	0.0	0.0	0.0	0.0	0.068182	0.045455
3	Chelsea	0.0	0.0	0.0	0.0	0.000000	0.040000
4	Chinatown	0.0	0.0	0.0	0.0	0.000000	0.040000

Acquisition and processing of available property data

- I used Manhattan property sales data as my starting point since I was not able to find a free resource for active property search information.
- The data did not include latitude and longitude information, so I used geocode to add this information. There were about 115 addresses that did not successfully return a location, so I dropped them from the analysis. Following a sample of this data:

	NEIGHBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT PRESENT	ADDRESS	ZIP CODE	YEAR BUILT	SALE PRICE	latitude	longitude
0	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2B	274 EAST 3RD	10009	1900	0	39.180193	-96.560083
1	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2B	301-303 EAST 4TH STREET	10009	1900	3,672,530	40.722295	-73.979974

- I next examined the differences in neighborhood names that were present in the property data and the venue data. There was not a clean conversion, so I assigned a “Foursquare Neighborhood” to each row in the property data based on the closest neighborhood latitude and longitude.
- Looking at the original property data neighborhood names vs. the “Foursquare Neighborhood”, I saw there were a number of property data entries that had incorrectly associated neighborhoods. The following shows a sample of the neighborhood name changes. Some minor, some completely different.

	NEIGHBORHOOD	Foursquare Neighborhood
0	ALPHABET CITY	Battery Park City
1	ALPHABET CITY	Lower East Side
15	CHELSEA	Chelsea
16	CHELSEA	Chelsea

Utilize the venue and property data to provide a client with properties meeting their criteria.

- Started with the assumption the client wants to require a neighborhood to include a Grocery Store, Pharmacy, and Coffee Shop.
- I then created a "scored" neighborhood dataframe that includes the required venue scores, the total of those scores, and a Boolean to indicate if all the required venues were present.
- I reduced that dataframe to only those neighborhoods that included all the require venues. Following is that dataframe.

	Neighborhood	Coffee Shop	Grocery Store	Pharmacy	Score	All Present
10	Flatiron	0.030000	0.010000	0.010000	0.050000	True
11	Gramercy	0.042553	0.021277	0.010638	0.074468	True
15	Inwood	0.018868	0.018868	0.018868	0.056604	True
17	Lincoln Square	0.030612	0.010204	0.010204	0.051020	True
25	Morningside Heights	0.073171	0.024390	0.024390	0.121951	True
26	Murray Hill	0.040000	0.020000	0.010000	0.070000	True
34	Turtle Bay	0.050000	0.010000	0.010000	0.070000	True
37	Washington Heights	0.022727	0.022727	0.011364	0.056818	True

- I then reduced the property dataframe to only include properties whose neighborhood included all the venues. Following is a sample of that data.

ADDRESS	ZIP CODE	YEAR BUILT	SALE PRICE	latitude	longitude	Foursquare Neighborhood
233 EAST 3 STREET	10009	1910	7,500,000	43.150612	-77.597139	Inwood
446 WEST 55TH STREET	10019	1901	4,550,000	40.767257	-73.988725	Lincoln Square
532 9 AVENUE	10018	1901	4,000,000	40.808930	-73.953850	Morningside Heights
129 EAST 17TH STREET	10003	1900	0	40.735814	-73.986666	Gramercy
30 EAST 14TH STREET	10003	1910	23,500,000	40.735139	-73.992286	Flatiron

- I then joined the reduced property dataframe with "scored" neighborhood dataframe to have the scores directly available with each eligible property. This sorted list now provided one of the desired outputs of a venue-scored ordered list of properties.
- Finally I provided the second desired output of a map of the eligible properties.

Results

Shown below is part of the ordered list of properties that meet the clients criteria:

ADDRESS	ZIP CODE	YEAR BUILT	SALE PRICE	latitude	longitude	Foursquare Neighborhood	Neighborhood	Coffee Shop	Grocery Store	Pharmacy	Score	All Present
353 WEST 115 STREET	10026	1900	11,000,000	40.805649	-73.960339	Morningside Heights	Morningside Heights	0.073171	0.024390	0.024390	0.121951	True
344 MANHATTAN AVENUE	10026	1900	1,623,468	40.803910	-73.957328	Morningside Heights	Morningside Heights	0.073171	0.024390	0.024390	0.121951	True
204 WEST 121ST STREET	10027	1910	4	40.806486	-73.950625	Morningside Heights	Morningside Heights	0.073171	0.024390	0.024390	0.121951	True
213 WEST 115 STREET	10026	1900	810,630	40.805649	-73.960339	Morningside Heights	Morningside Heights	0.073171	0.024390	0.024390	0.121951	True

Shown below is a map of all the properties meeting the clients criteria:

Discussion

It seems this functionality would be very helpful to clients desiring to specify some venue criteria in their property search. It could be integrated fairly easily into a web application. The neighborhood venue information could be updated periodically. When a client is searching in an area where venue information was available, the venue criteria option could be provided on the search form

It would be desirable to make the Foursquare neighborhood venue information more accurate. Only venues within a 500 meter radius of neighborhood center latitude/longitude are provided.

Saw some inconsistent results from geocode. Sometimes returned incorrect latitude/longitude for valid Manhattan address. Would want to have a reliable service for this.

An option that could be provided to the client would be to not require all venues to be present and provide a score-ordered list of all eligible properties.

Conclusion

This project demonstrated the utility of using data (Foursquare venue data) to solve a problem (clients wanting to specify venue information in their property search criteria).