Recommend a house for New York customers

Coursera Capstone – The Battle of the Neighborhoods

Changshao Liang

Nov 6, 2020

1.Introduction

1.1 Background

Henda investment&Property consulting is a company located in Brooklyn, New York, dedicated to providing customized consulting and agency services for clients who are interested in buying or renting real estate. With the development of information technology, Henda found that data science technology including machine learning can bring better business opportunities. The recommendation system based on data science technology is gradually becoming a good helper for experienced real estate agents. These techniques are more convincing than personal experience.

1.2 Interest & Problem

Moving is inevitable, but we don't like moving to a very strange environment. Every customer has lived in a Neighborhood for a long time and has developed their unique living habits. Therefore, if you move to a similar neighborhood as before, you can try to avoid the unsuitability caused by environmental changes. Therefore Henda is committed to providing these customers with services to find suitable communities.

Recently, an elderly couple asked Henda to provide them with counseling services. They used to live in the Hillcrest community in Queens, New York. Now they hope to move to Staten Island, New York. They filled out Henda's questionnaire (see link). According to the questionnaire, their main requirements are as follows:

1. They want us to recommend communities similar to Hillcrest for them because they think similar communities can prevent them from changing their living habits.
2. They hope their neighbourhood has various bakeries because they really like donuts.
3. They hope that their neighbourhood has more Bus Stops because they like to travel. And they go to Manhattan to visit their son once a month.

2.Data

2.1Data souces

We will use the map data of New York City provided by NYU spatial data repository.
link ： https://geo.nyu.edu/catalog/nyu_2451_34572
We obtained housing price data in New York City from Kaggle.
Link:https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/data
Foursquare API for developers can provide us with information about venue categories, top tips, locaton data, ratings, etc. to complete this task.

In addition, python libraries such as geopy, forlium, etc. also provide us with necessary help.

2.2Data Processing

Cleaning

The New York City Neighborhood data obtained in this article is a json file. We use some python libraries to get the Brough, Neighborhood, Latitude, and Longitude we want.

Before performing K-Means clustering, we used the One-hot method to encode the Top 10 Commen Venues of each Neighborhood.

When obtaining data on housing prices in New York City, we analyzed the excess data and deleted it. For other data, if there is a null value, we use the average or common value instead.

Feature selection

When exploring the designated Neighborhood, there is no doubt that the four features of Brough, Neighborhood, Latitude, and Longitude are the most important. With this geographic information, we can use the Foursquare API to explore the surrounding Venues to find the most suitable Neighborhood for our customers.

When processing house price data, we pay attention to price, room_type, number_of_reviews, avalilability, etc., because these are all related to whether the customer can really buy a house in the area.

3.Methodology

3.1 Neighbourhood data in New York City

When obtaining the Neighbourhood data of New York City, we mainly use the urlopen function of python's urllib.request to download the json file from the Internet. Then use the json library to read the json file. Obtained Neighborhood data for New York City.

```
In [5]: neighborhoods_data[0]

Out[5]: {'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```

By using pandas, we build a Dataframe and read the Json data into it. In this way we have a usable pd.dataframe. This dataframe mainly contains fields such as Brough, Neighborhood, Latitude, and Longitude.

```
In [9]: neighborhoods.head()
```

Out[9]:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

By using geopy or other methods, we can get the geographic coordinates of New York City. According to the geographic coordinates and the neighborhoods above, we can use folium to map to get the distribution map of the neighborhoods in New York City. Then we screened the customer's original Queens and his destination Staten Island Neighborhood. We obtain these common Venues of Neighbourhood through Foursquare. Then we use One-hot processing to make the data available for cluster analysis. We will use the K-Means method.

```
In [23]: client_area_onehot.head()
```

Out[23]:

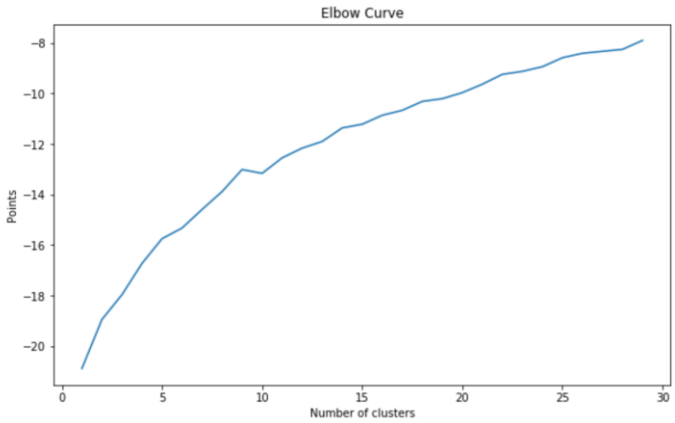
	Yoga Studio	Accessories Store	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Au
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Out[51]:

	Neighborhood	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	Annadale	American Restaurant	Pizza Place	Park	Restaurant	Diner	Train Station	Food	Deli / Bodega	Dance Studio	Flower Shop
1	Arden Heights	Coffee Shop	Pharmacy	Deli / Bodega	Bus Stop	Pizza Place	Eye Doctor	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
2	Arlington	Coffee Shop	Home Service	Grocery Store	Deli / Bodega	Boat or Ferry	Bus Stop	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
3	Arrochar	Bus Stop	Deli / Bodega	Italian Restaurant	Outdoors & Recreation	Mediterranean Restaurant	Bagel Shop	Athletics & Sports	Sandwich Place	Middle Eastern Restaurant	Polish Restaurant
4	Arverne	Surf Spot	Sandwich Place	Metro Station	Bus Stop	Beach	Thai Restaurant	Café	Restaurant	Donut Shop	Board Shop

3.2 K-Means Method

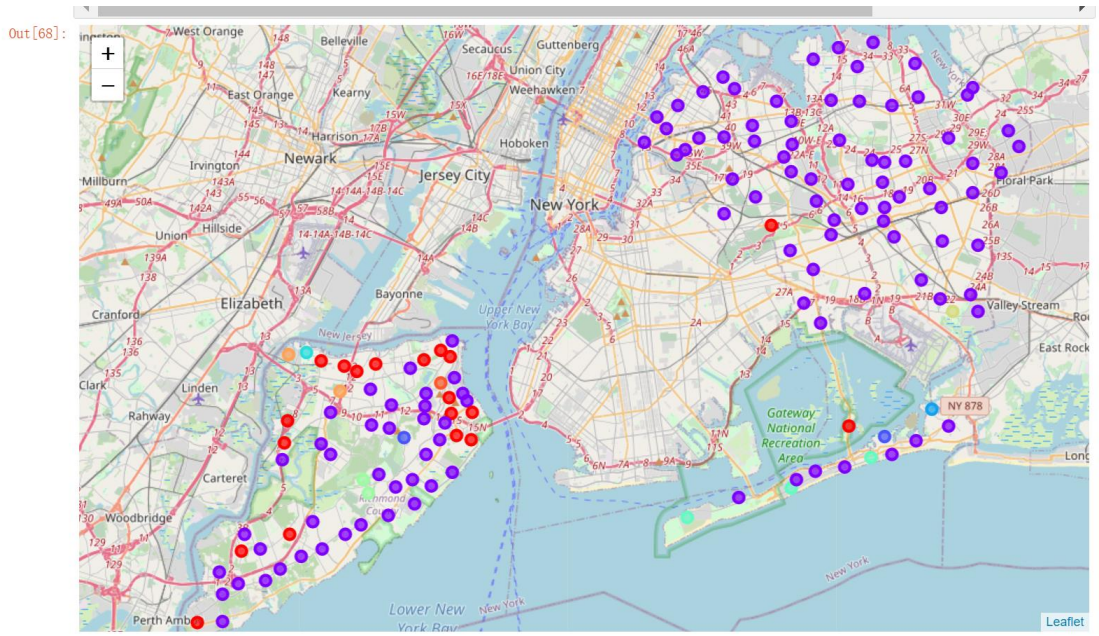
In order to use k-means, a value of k must be determined to determine the number of clusters. Here we use Elbow's method, let k=range(1,30), and determine the best K value according to the elbow point (the point where the curvature of the evaluation curve changes drastically). See https://www.zhihu.com/question/29208148。



We write a piece of code to study the scores of the K-Means method under different K values. We can observe that at the point K=10, the curvature of the curve changes significantly. k>10, the curve becomes flat. Therefore we set K=10.

	Neighborhood	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	Annadale	American Restaurant	Pizza Place	Park	Restaurant	Diner	Train Station	Food	Deli / Bodega	Dance Studio	Flower Shop
1	Arden Heights	Coffee Shop	Pharmacy	Deli / Bodega	Bus Stop	Pizza Place	Eye Doctor	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
2	Arlington	Coffee Shop	Home Service	Grocery Store	Deli / Bodega	Boat or Ferry	Bus Stop	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
3	Arrochar	Bus Stop	Deli / Bodega	Italian Restaurant	Outdoors & Recreation	Mediterranean Restaurant	Bagel Shop	Athletics & Sports	Sandwich Place	Middle Eastern Restaurant	Polish Restaurant
4	Arverne	Surf Spot	Sandwich Place	Metro Station	Bus Stop	Beach	Thai Restaurant	Café	Restaurant	Donut Shop	Board Shop

3.3 Exploring Neighbourhoods



By drawing a map by Folium, we can get the above result. The origin of the same color means that these Neighborhoods belong to the same category, and the surrounding Venues are

very similar.

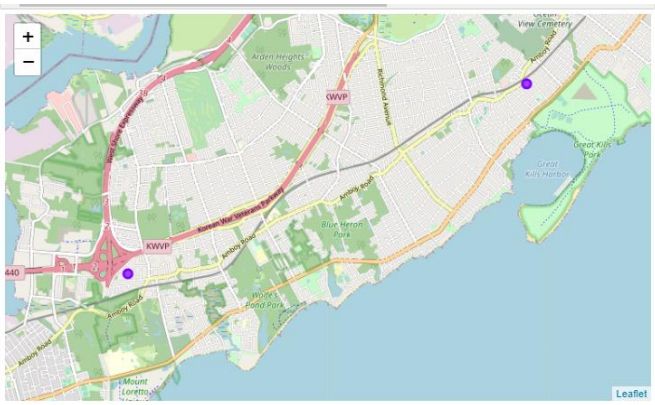
Below we select the same Neighborhood of the Cluster from Staten Island. Then we explore the potential target Neighborhood based on Top comment venues. We use 1st Most Common Venue = "Donut Shop" to be our filter. So we selected two potential neighborhoods, Bay Terrace and Pleasant Plains.

```
In [72]: potential_choice_1 = potential_choice.loc[potential_choice["1st Most Common Venue"] == "Donut Shop"]
         potential_choice_1
```

Out[72]:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th M Comm Ve
22	Staten Island	Bay Terrace	40.553988	-74.139166	1	Donut Shop	Clothing Store	Supermarket	Women's Store	Mobile Phone Shop	Cosme S
24	Staten Island	Pleasant Plains	40.524699	-74.219831	1	Donut Shop	Yoga Studio	Discount Store	Bus Stop	Fast Food Restaurant	Lic S

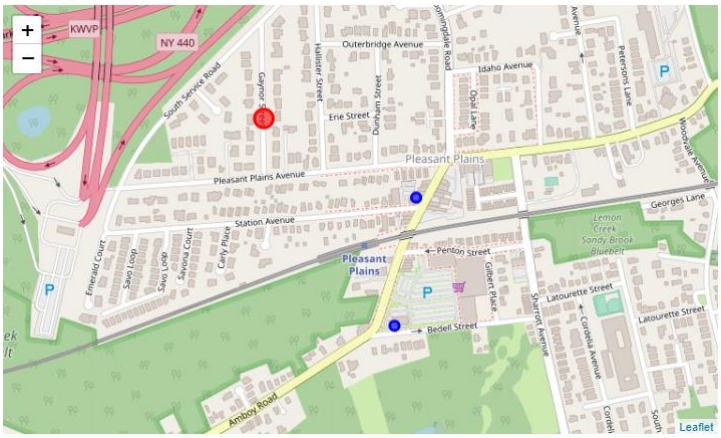
Out[73]:



3.4 The first condition - Bakery

Next, we call the Foursquare API to explore these two Neighborhoods. We issue a call to the Foursquare API with search_query ='Bakery'. It is worth noting that in our questionnaire, the couple stated that they own a car but drive less frequently. They like to walk or walk around the community. So we narrowed the scope of exploration to 1,000 meters so that they can reach their favorite destinations on foot.

	name	categories	address	lat	lng	labeledLatLngs	distance	postalCode	cc	city	state	country
0	La Dolce Bakery & Pastry Shoppe	Bakery	6321 Amboy Rd	40.523319	-74.216365	[{'label': 'display', 'lat': 40.523319, 'lng':...	331	10309	US	Staten Island	NY	United States
1	La Dolce Bakery	Bakery	Amboy Road	40.521113	-74.216860	[{'label': 'display', 'lat': 40.521113, 'lng':...	471	NaN	US	Staten Island	NY	United States



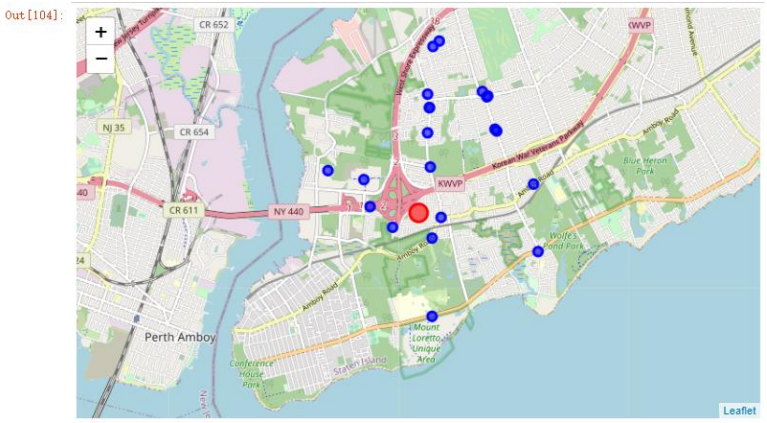
This is the Bakery we found near Pleasant Plains. We continue to use the Foursquare API to explore these stores, but unfortunately, the stores near here are not rated, and there are only the following tips.

	text	agreeCount	disagreeCount	id	user.firstName	user.lastName	user.id
0	I love eating bugs :)	0	0	51801ed2e4b09dd206022131	Nikkilyn	R	11038272

We also did the same exploration on Bay Terrace. The Bakery near Bay Terrace is less, and there are no ratings and tips.

3.5 The second condition – Bus Stops

In order to meet customers' needs to travel by Bus, we have studied Bus Stops between these two communities. Similarly, we called the Foursquare API to obtain information about the surrounding Bus Stops.



It can be seen that the bus stops in Pleasant Plains are more evenly distributed, and there are more bus stops on the line Bloomingdale Road-Amboy Road. In addition, in the middle of Pleasant Plains is the overpass of Veterans Road. This road can quickly help residents of this neighborhood quickly reach other areas of New York.

As you can see from the map, the bus stops on Bay Terrace near Lower Ney York Bay are mainly concentrated on Hylan Boulevard and Giffords Lane. For those who want to travel far by bus, the transportation here is not the most convenient.

3.6 The third condition - House Price

I found some New York City housing price data from Kaggle for reference and to help us make decisions.

```
In [3]: #observe the data
        df.head()
```

Out[3]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	pri
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	1
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	2
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	1
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	

Next, we have to make preliminary observations, clean up and sort out the housing price data. Let's first observe the situation of null values.

```
In [4]: #check the null
        df.isnull().sum()
```

Out[4]:

```
id                                  0
name                               16
host_id                             0
host_name                          21
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                     10052
reviews_per_month               10052
calculated_host_listings_count      0
availability_365                    0
dtype: int64
```

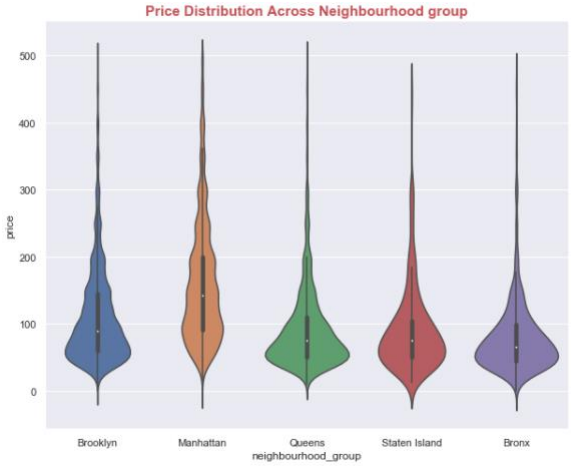
Name: We use 0 to fill, in fact, this feature is not very critical. We already have host_id.
Host_name: It is private information and does not contribute much to our analysis. delete.
last_review & reviews_per_month: There are more empty values. The existence of a null value or 0 means that no one has evaluated, and both of these features indicate that no one has

evaluated, so just keep one. lr is the date, which does not directly help our analysis. We delete last_review and keep reviews_per_month, and then fill the rpm to 0.
Id: Delete. In order to explore the price, just leave the host_id.

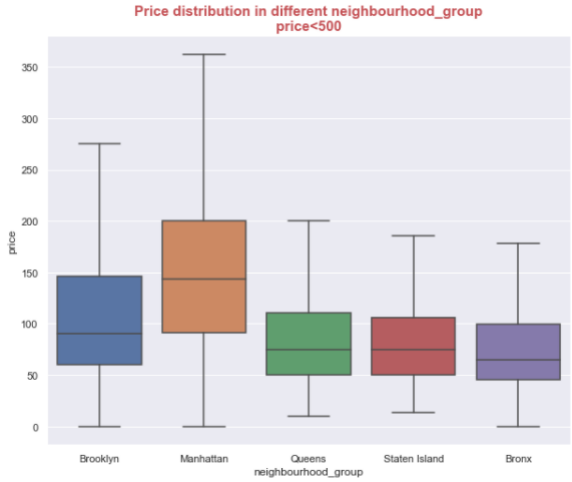
Next, we observe housing price data based on the five regions of New York.

	Brooklyn	Manhattan	Queens	Staten Island	Bronx
count	20104.000000	21661.000000	5666.000000	373.000000	1091.000000
mean	124.383207	196.875814	99.517649	114.812332	87.496792
std	186.873538	291.383183	167.102155	277.620403	106.709349
min	0.000000	0.000000	10.000000	13.000000	0.000000
25%	60.000000	95.000000	50.000000	50.000000	45.000000
50%	90.000000	150.000000	75.000000	75.000000	65.000000
75%	150.000000	220.000000	110.000000	110.000000	99.000000
max	10000.000000	10000.000000	10000.000000	5000.000000	2500.000000

We use the violin chart to observe the mean and variance of the five regions.



We can also use box plots to observe the variance of the mean of each area, which is one of the knowledge points of the IBM data science course.

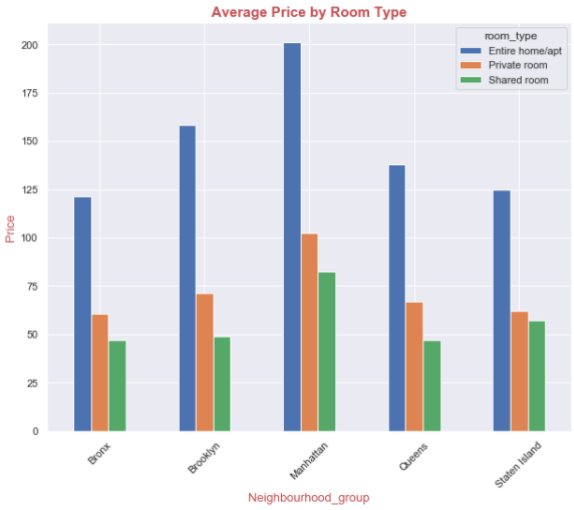


Through the above two figures, we can know:
①Manhattan's housing prices are the highest. The price range is also very large. The average house price in Manhattan is around $1.45 million.
②The housing prices of the client's original Queens and Staten Island that they will move into are very similar. The price range of Queens house prices is relatively large. Therefore, moving from Queens to Staten Island is a reasonable choice from a financial point of view.

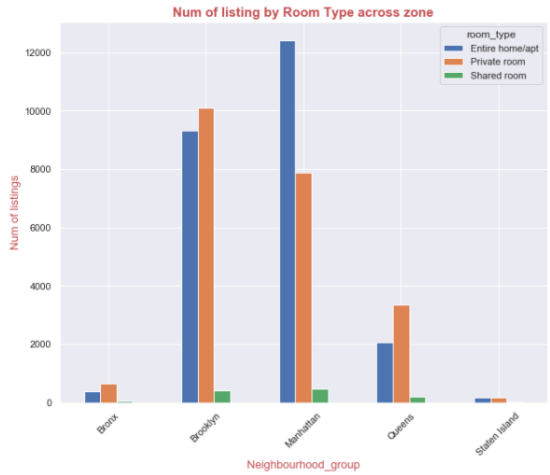
We use geographic information maps to observe the relationship between New York City's geographic location and housing prices.

There is no doubt that in terms of housing prices, the central area of New York City is Manhattan, followed by Brooklyn. There are some high-priced housing areas in the east of Queens, which may be luxurious residential areas.

Next we have to consider the relationship between housing types and housing prices.



We can see that the price of Entire home/apt is significantly higher.
Next, we observe the housing situation, we can see that the housing in Manhattan is much higher than other areas. The second is Brooklyn. And Staten Island has very few listings.



According to the above chart, we can recommend different types of houses to customers according to their purchase budget.

Next, we observe the popular listings based on number_of_review and price, and calculate the average house price in the area.

```
In [54]: avg_price=review_price.price.mean()
         avg_price
Out[54]: 130.0539026437264
```

We conclude that the average house price in New York City is $1.3 million.

Then we want to obtain the average house price of Staten Island and compare it with the house price of the neighborhood where the residents were originally located.
Since the statistical year may be different, we did not find Pleasant Plains in the Kaggle data. Therefore, we replaced the price of Pleasant Plains with the price of Castleton Corners near Pleasant Plains. The distance between the two Neighborhoods is very small, and the difference in housing prices should not be very large.

	name	host_id	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_n
13682	Newly renovated house 4 bedroom. Minutes from NYC	10721093	Staten Island	Castleton Corners	40.61042	-74.12277	Entire home/apt	299	
41361	3 Private rooms,shared space	17473098	Staten Island	Castleton Corners	40.62373	-74.11773	Private room	150	
19778	Delightful studio apartment.	102526590	Staten Island	Castleton Corners	40.62063	-74.13001	Entire home/apt	65	
9345	Spacious, Clean , Close to Local Transportation	37360127	Staten Island	Castleton Corners	40.61363	-74.12152	Private room	45	

```
In [71]: avg_price_pp=review_pp.price.mean()
         avg_price_pp
Out[71]: 139.75
```

It can be seen that the average house price near Pleasant Plains is 139.75. But the price range is from 45 to 299, customers have many choices.

	name	host_id	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_
25146	BIG-3 BDRM house, 1hr to Manhattan, near beach	7927832	Staten Island	Bay Terrace, Staten Island	40.55182	-74.14439	Entire home/apt	150	
42862	Modern studio/private entrance/superb location	250580779	Staten Island	Bay Terrace, Staten Island	40.55105	-74.13660	Entire home/apt	55	

From the above, the average house price near Bay Terrace is 102.5. The price range is 55 to

150. But the choices are not many.

Let us look at Hillcrest's housing prices.
Since the statistical year may be different, we did not find Hillcrest in the Kaggle data. Therefore, we used the housing prices in Jamaica Hills near Hillcrest to replace the housing prices in Hillcrest. The distance between the two Neighborhoods is very small, and the difference in housing prices should not be very large.

	name	host_id	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nig
41559	RM#2-Bright room w/balcony 30 mins NYC & airports	241963980	Queens	Jamaica Hills	40.71686	-73.79742	Private room	67	
19555	Room near NYC airports + trains	87393824	Queens	Jamaica Hills	40.71245	-73.79806	Private room	65	
41048	ALL NEW LUXURY APT- 1ST FL 4BR/2 BTH,10MIN-JFK...	215385823	Queens	Jamaica Hills	40.71378	-73.80292	Entire home/apt	275	
43134	Apartment minutes from manhattan	252168489	Queens	Jamaica Hills	40.71022	-73.79665	Entire home/apt	110	
46061	LUXURY 2ND FL,4BR/2 FUL BTH,SLEEP 8+,15MIN-JFK...	215385823	Queens	Jamaica Hills	40.71209	-73.80151	Entire home/apt	325	

```
In [80]: avg_price_h=top_review_h.price.mean()
         avg_price_h
Out[80]: 132.125
```

We can see that the average price of the customer's original residence was US$1.321 million, which is close to the average price of New York City.

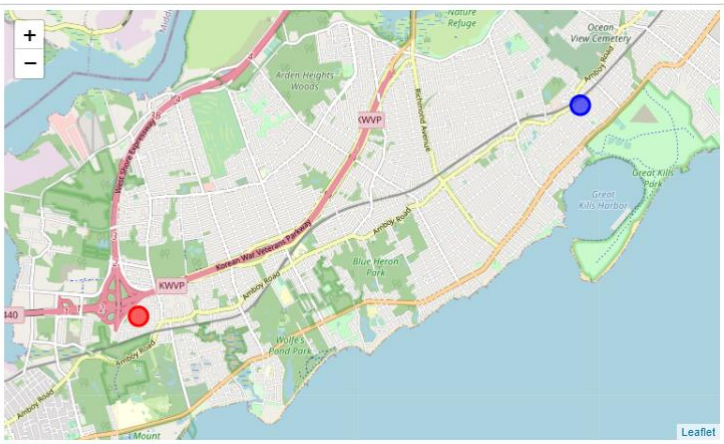
4.Rusults

By exploring customer requirements for Donut Shop and Bus Stops, the two communities: "Pleasant Plains" and "Bay Terrace" can basically meet customer needs.
In addition, they are all in the first cluster just like "Hillcrest". But from the Bakery data returned by Foursquare, it seems that Bakery around Pleasant Plains is more popular.
Compared to Bay Terrace, Pleasant Plains is closer to the main road, which is more convenient for customers to travel. House prices around Bay Terrace and Pleasant Plains are similar. The price of Hillcrest's house is very close to these two Neighborhoods. But there are more houses for sale near Pleasant Plains, and there are many different prices and many choices.
Therefore, our company believes that Pleasant Plains is the best choice for customers on Staten Island.

5.Discussion

In fact, these two blocks have their own characteristics. Let's take a look at their geographic information.

Out [108]:



It can be observed from the map:

Pleasant Plains is very close to New Jersey, it is far from the center of New York City. However, because it is on the main traffic road, the traffic in this Neighborhood is more convenient. Pleasant Plains is far from the coastline, surrounded by Bloomingdale Park and Long Poind Park for people to move around.

There are only general roads near Bay Terrace, so traveling far away may not be convenient. But its location is closer to the city center of New York City. There are Great Kills Park and GreatKills Harbor near Bay Terrace, which have a different beach style from the client's original residence, Hillcrest.

It is likely that those venues of requirements are just between their geographically. Therefore, they are very suitable for our customers. If customers want to try something new, Bay Terrace is also a good choice.

6.Conclusion

This result has limitations. The data used in this project is from the top 10 Venues in each community, which may omit those who have a lot of total venues, and there are also many Bakery and Bus stops, but the proportion of Bakery and Bus stops is low in Neighborhood.

Secondly, we use 1km as the radius to explore Neighborhood, because customers don't drive often. But customers may change their way of travel because of the difference in Venues. So there is room for improvement.

In short, Pleasant Plains and Bay Terrace can meet the basic requirements of our customers. Henda strongly recommends Pleasant Plains. But there may be better options. Our ultimate choice remains in the hands of customers.