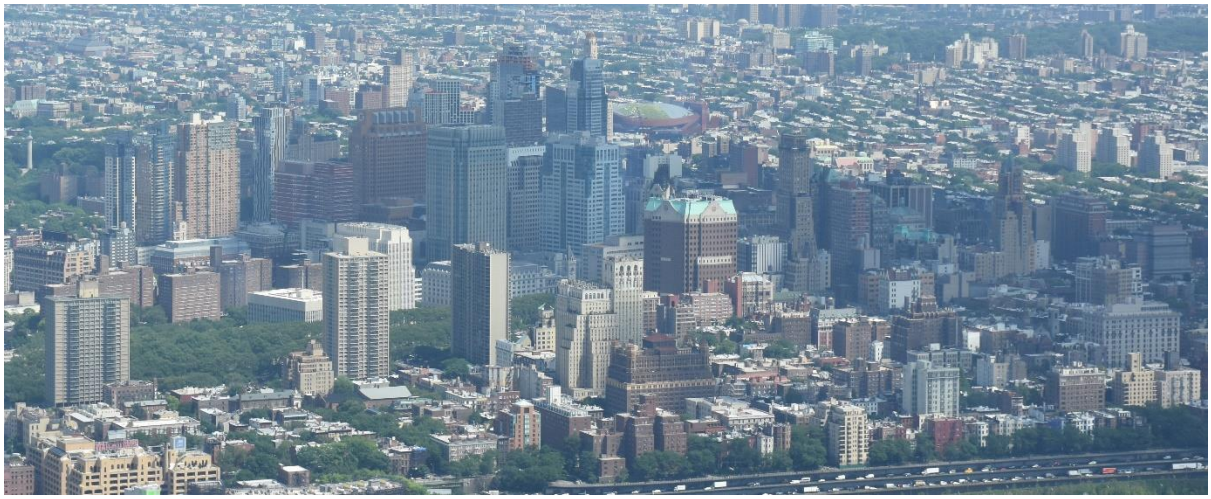


Finding the Best House to Buy in Brooklyn, New York

By: Mona Berdugo

Introduction:

Buying a house is one of the biggest financial investments that people make during their lives and many factors go into that decision. The choice of what city to live in is often determined by the location of one's employment, friends, family and other considerations. But even within certain cities, neighborhoods vary greatly in terms of cost and character. Knowing what you are looking for in a neighborhood as well as how much money they can spend can help a person decide which neighborhoods to look at when buying a house.

In this paper I will use the example of a young couple with small children and a limited budget. They both have friends and family in different parts of Brooklyn and want to buy a house there.

Since they have small children they want to be close to schools. Neighborhoods with a variety of schools can give them more choices when choosing a school for their children. Also, neighborhoods with many schools are likely to have more young families making it more likely they and their children will make friends.

They also want to be near parks and playgrounds and other outdoor recreational venues. Being able to easily enjoy fresh air and exercise outside will benefit both their and their children's quality of life.

Lastly, though they do not go to restaurants very often, they want to be near grocery stores, supermarkets, and other stores where they can buy food. Having these types of stores nearby is a convenience that can save them lots of time in their busy lives.

This couple is would like to buy a home in a neighborhood that fulfils the above needs and want to know what their options are as well as how much houses in those neighborhoods generally cost.

Data

To find the right neighborhood for the couple I will use information taken from the following locations:

- Geographic data on Brooklyn, including the longitude and latitude of each neighborhood was taken from https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json. These are the coordinates used to search for the desired nearby venues.
- A geojson of New York neighborhoods used to create a choropleth map of Brooklyn was downloaded from https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas-NTA-/cpf4-rkhq. This map can help visualize the location of each neighborhood.
- For data on the cost of housing in Brooklyn I used the table provided at: https://www.propertyshark.com/mason/market-trends/residential/nyc/brooklyn/brooklyn-heights. Not all neighborhoods were listed on this table so neighborhoods with no housing data were removed from the neighborhoods dataframe. Information was available for 50 out of 70 neighborhoods. Those were the neighborhoods used in this analysis.
- I got information about schools, parks and shops in Brooklyn from the **Foursquare API**. This data will be used to see the relationship between each of those factors and the price of houses in each neighborhood.

Methodology

Starting Point and Preprocessing:

I started out by downloading and organizing my data. I downloaded the list of neighborhoods with their geographic coordinates and joined it to the list of hosing prices, getting rid of unnecessary columns as well as any neighborhoods for which information was not available. I had to make sure that the names of the neighborhoods were written the same way in both tables so that the join would work and I changed the field type of the "Price" column to float, instead of string so I could do arithmetic operations on that column. The final table looked like this:

	Neighborhood	Latitude	Longitude	Median Sale Price
0	Bath Beach	40.599519	-73.998752	520000.0
1	Bay Ridge	40.625801	-74.030621	430500.0
2	Bedford	40.687232	-73.941785	773360.0
3	Bensonhurst	40.611009	-73.995180	686555.0
4	Bergen Beach	40.615150	-73.898556	740000.0

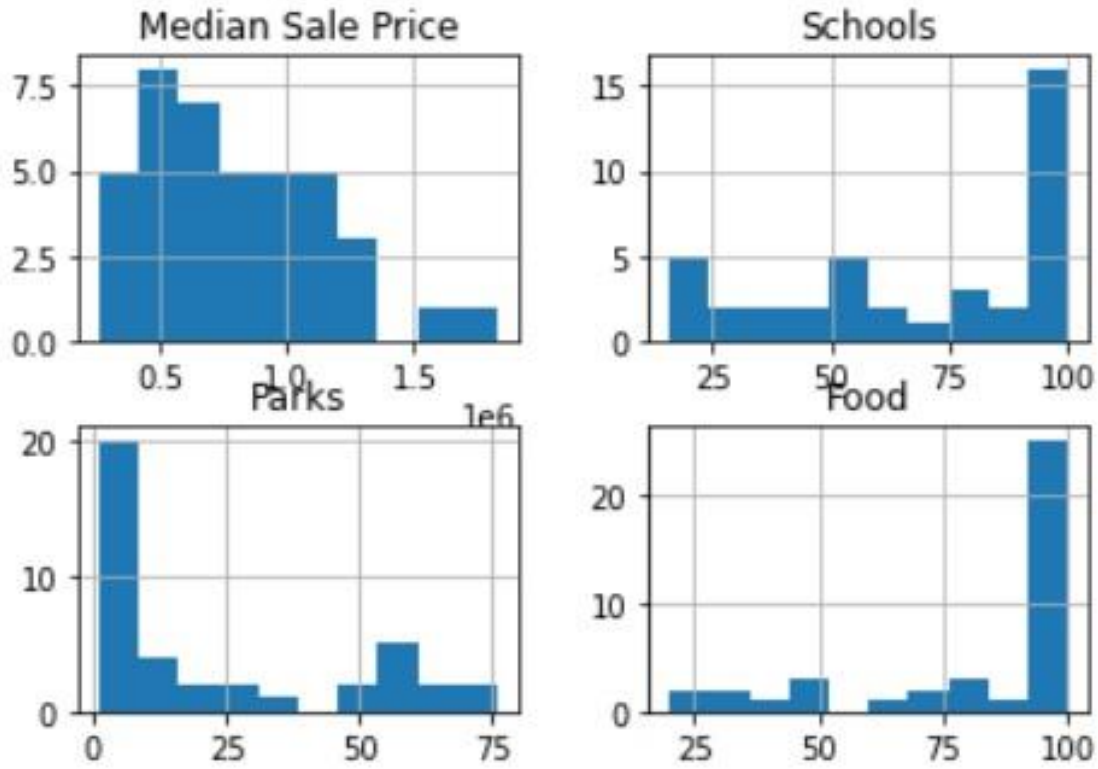
Gathering Information:

I used the Foursquare API to gather information about each neighborhood. In particular, I wanted to see how many schools, parks and other outdoor venues, and food shops were in each neighborhood. After passing the appropriate API request into foursquare and grouping the results by neighborhood, I joined the lists to the original table and got the following:

	Neighborhood	Latitude	Longitude	Median Sale Price	Schools	Parks	Food
0	Bay Ridge	40.625801	-74.030621	430500.0	43	2	98
1	Bedford	40.687232	-73.941785	773360.0	95	18	100
2	Boerum Hill	40.685683	-73.983748	932000.0	100	70	100
3	Borough Park	40.633131	-73.990498	800000.0	50	1	92
4	Brighton Beach	40.576825	-73.965094	440000.0	30	5	50

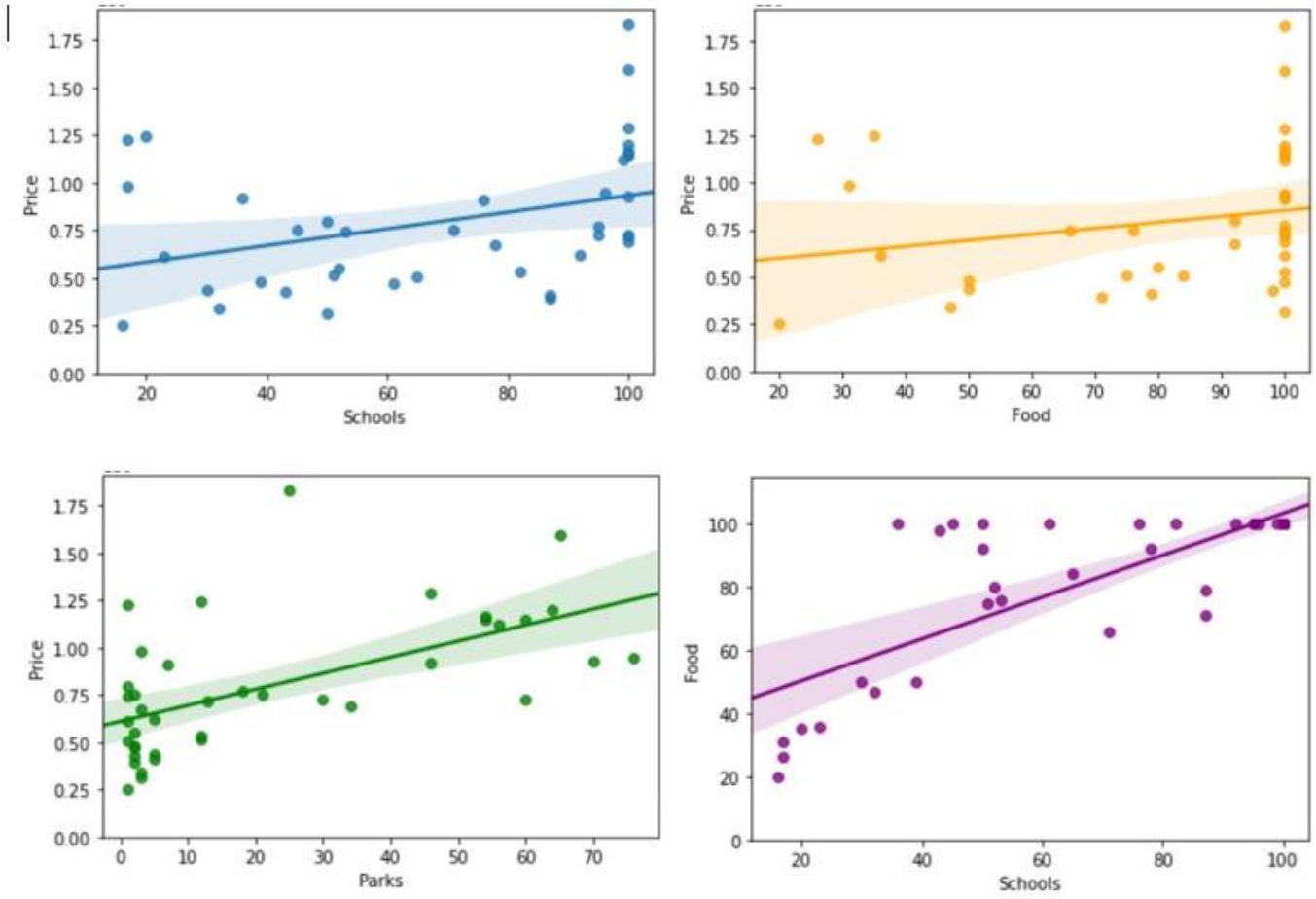
Analysis:

Once I had the information I needed I could explore and analyze the data. First, I made a histogram of all the relevant information to see the range and other statistical information for each parameter.

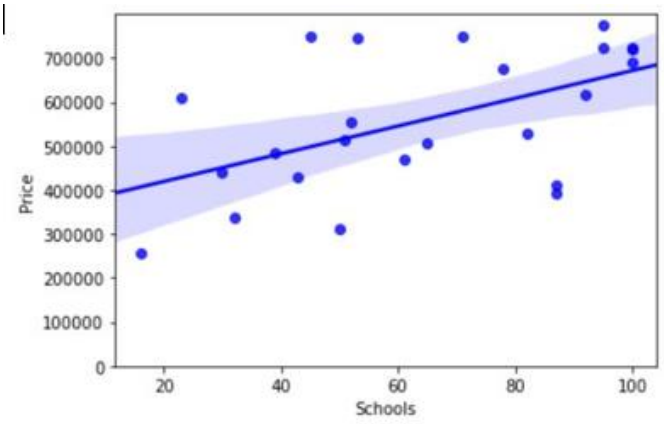


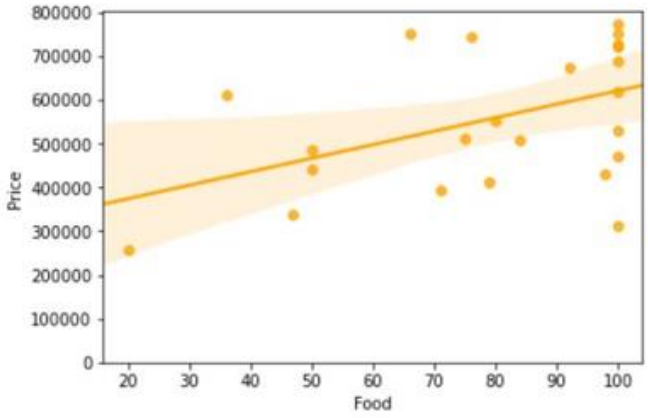
I saw that housing costs range from a couple of hundred thousand dollars to almost two million. A bar chart gave me some more specific information including the names of the neighborhoods. Viewing the neighborhoods from least expensive to most gives the following result:

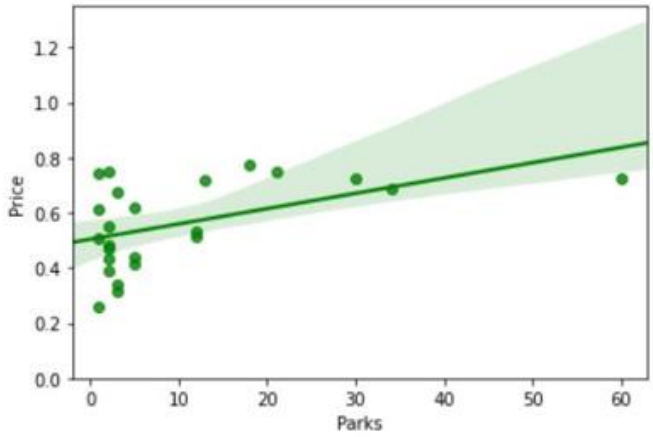
Then I looked to see if there is a correlation between the cost of houses and the factors I was considering. I created regression plots for each of the variables:



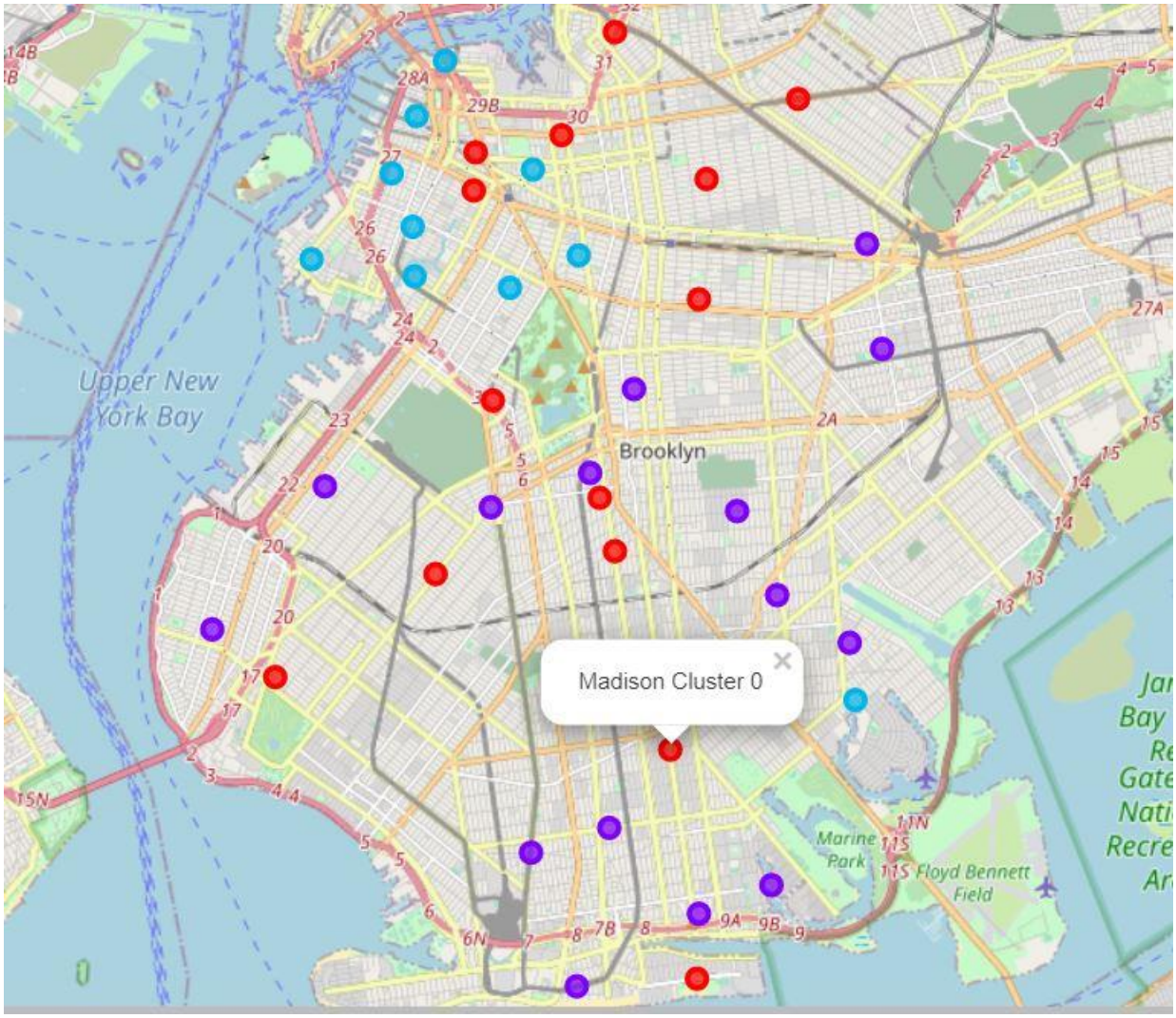
If we create the same regression plots using only neighborhoods with home values of less than $800,000 we get the following:







I then ran a k-means clustering algorithm to cluster the neighborhoods into three clusters based on their attributes:



Finally, I searched for neighborhoods that fill the following requirements:

- Under $800,000
- More than 80 schools
- More than 80 Food establishments
- At least 20 parks or outdoor spaces

This gave me the following three options:

	Neighborhood	Price	Schools	Parks	Food
7	Bushwick	724000.0	95	30	100
9	Clinton Hill	689000.0	100	34	100
13	DUMBO-Vinegar Hill-Downtown Brooklyn-Boerum Hill	725000.0	100	60	100

Results

Histograms:

From the histograms we see that houses in Brooklyn go for anywhere between $250,000 to over $1,800,000. This is a huge range, but the vast majority are below $1,000,000 with the greatest number of neighborhoods selling houses in the $500,000 range.

As for schools and food establishments, Foursquare seemed to limit the number of venues it returned to 100. This is fine since 100 is more than enough for the purpose of this study. A majority of neighborhoods come close to or exceed the 100 venue limit for schools and food. This means that, hopefully, it shouldn't be too difficult to meet those criteria.

Parks, on the other hand, are not as common. There are many neighborhoods with few, if any outdoor spaces. However, there are some neighborhoods with quite a few.

Regression plots:

It is possible that the slope was not reflective of the actual correlation because the upper limit of 100 for food and schools changed the results. In any case, if you only use the "under $800,000" neighborhoods a correlation is slightly more evident, though it still seems to be the case that there are other factors driving the price that we have not considered. The highest correlation seems to be between food and schools, regardless of cost.

While the very low cost houses seem to have the fewest outdoor spaces, schools, and food, there still seems to be many options in the mid-range of neighborhoods that fulfill the requirements.

K-means clusters:

The statistics for the three clusters are shown below.

Cluster	Average Price	Average Number of Schools	Average Number of Parks	Average Number of Food Shops
0	$802,595	74.13	25.67	90.40
1	$458,044	53.93	3.87	72.47
2	$1,296,913	83.60	43.70	86.10

Cluster #2 (the light blue circles) has the most expensive houses. They also have the most schools and parks on average. It comes in a close second for the average number of food stores. While these neighborhoods fit the criteria, they are very expensive and we would prefer something more affordable.

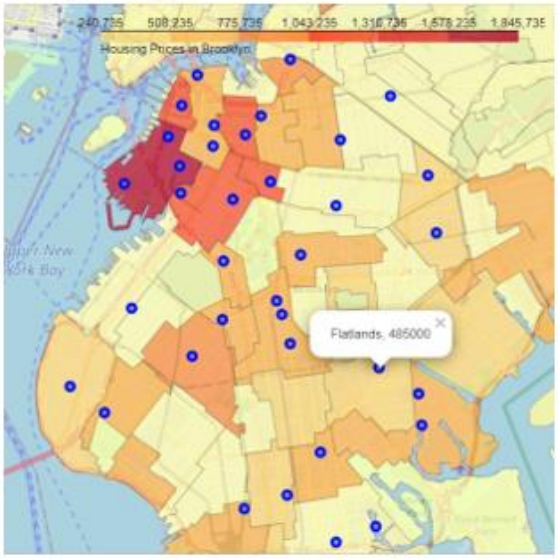
Cluster # 1 (purple) has the cheapest homes, but they also have the fewest schools and food shops and almost no parks. (Less than 4 on average.)

Cluster #0 (red) has the midrange housing prices with a good number of schools and parks and the most food shops. It is in these neighborhoods that the couple should probably look for a home.

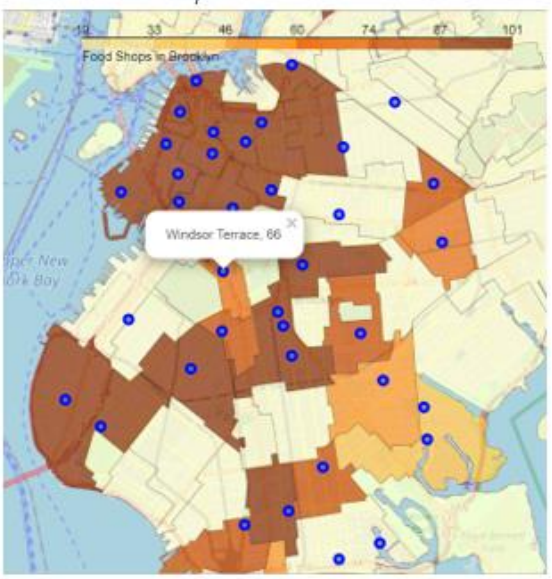
Choropleth map:

Using a Choropleth map we can see the locations of these neighborhoods in relation to each other and other parts of the city.

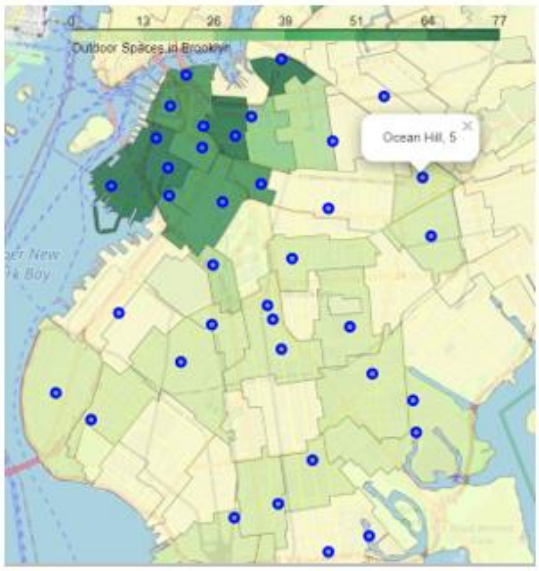
Average Cost of Homes



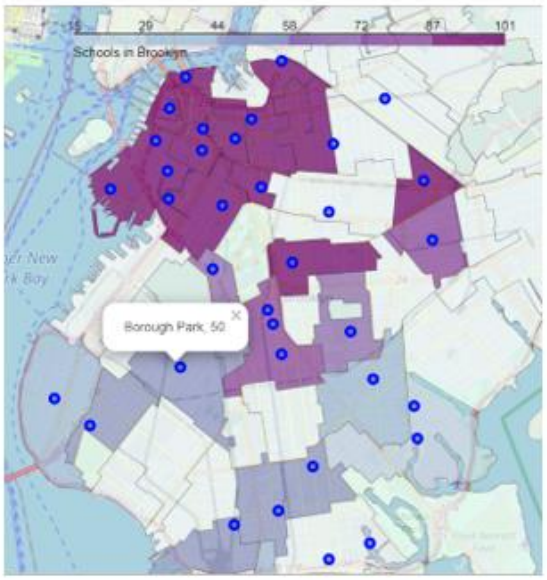
Food & Drink Shops



Parks & Outdoor Spaces



Schools



The northwestern part of the borough is clearly where there is the most outdoor spaces and schools. Food shops are plentiful in that section but also in other areas. Housing prices seem to be highest in that area, but there are also some neighborhoods there that are mid-range in terms of prices.

Final search:

After getting a general idea about the different neighborhoods and what the numbers look like for the various parameters, I did a very specific search for homes under $800,000 with more than 80 schools and food shops and over 20 outdoor places:

```
bk_options = PD[(PD['Price'] <= 800000) &
                (PD['Schools'] > 80) &
               (PD['Food'] > 80) &
               (PD['Parks'] > 20)]
bk_options
```

	Neighborhood	Price	Schools	Parks	Food
7	Bushwick	724000.0	95	30	100
9	Clinton Hill	689000.0	100	34	100
13	DUMBO-Vinegar Hill-Downtown Brooklyn-Boerum Hill	725000.0	100	60	100

This search turned up three very decent options for the couple that meet all their criteria.

Discussion

Using various methods of analysis, we were able to find three neighborhoods which best meet the requirements for this couple. The fact that the correlation between their requirements and the price of housing in different neighborhoods is minimal means that there must be other factors driving the cost of housing. Such factors might have to do with crime rates, size and type of houses for sale, ethnic/religious considerations and more.

I chose three parameters based on the needs of a specific couple. These criteria are clearly not the only factors affecting housing prices but they are what was appropriate for this specific case.

The fact that there is a correlation between food and schools in these areas indicates that these are probably neighborhoods with many families and children. The lack of (or minimal) correlation between prices and the other criteria prove that there are many other factors and considerations determining prices.

While 100 schools seems like a lot, the results included everything from nursery schools through universities. It also included driving schools, adult education and more. The vast majority of schools, however, were elementary and high schools. Taking this into account, limiting the search for over 80 schools is reasonable.

Similarly, 100 food establishments at first seems extreme, but it includes not only supermarkets and grocery stores, but butchers, liquor stores, and all kinds of specialty stores that sell any type of food and drink. Therefore, given the range, searching for over 80 seemed like a reasonable amount to provide a nice selection.

The category of Outdoors and Recreation included both indoor and outdoor sports facilities, studios, as well as playgrounds and gardens. Selecting at least 20 would ensure variety of both indoor and outdoor play areas for any weather throughout the year.

The final search returned three excellent suggestions that perfectly fulfilled the criteria we were looking for. They are midrange in terms of cost, but, as we saw, the cheaper neighborhoods did not have what we were looking for. The couple should narrow their search for home to these three neighborhoods.

Conclusion

“Location, location, location” is a common phrase when buying real estate. But what determines a desirable location? Prices can depend upon man many factors. In our case, we were able to narrow down the housing options for this couple to three neighborhoods based on their specific needs. The process is easily repeatable using different criteria or different cities for anyone who can narrow down what they are looking for to a few specific factors. The geographical data was easy to come by although it needed some preprocessing. Also, real estate is a very volatile market so be sure to use the most recent data available!