The Battle of Neighborhoods: Capstone Project

Opening a bakery in a convenient neighborhood in

New York City

Introduction/Business Problem

Because New York City has been a major point of entry for immigrants in United States, is a metropolis with a large and ethnical diversity. With over 8.3 million people in 2019, and with a fast-growing population, it looks like a good place to open a business or buying a business already installed. Considering that "bread" is part of many cultures, bakery is a good option to open a business. There are two approach that will be considered to open or buying a bakery: In a business area, or close to a residential area. The main reason in a business area is giving more emphasis to provide other businesses like restaurants, cafes, sandwich places, etc., and the main reason close to a residential area is to provide services directly to people. In this case, it will be considered to open a bakery close to a residential area, because it would be a small family business.

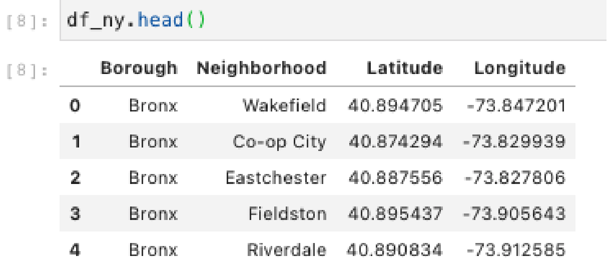
Many questions come out about this business opening. For example: What is the best borough and neighborhood? What indicators determine a good borough and a good neighborhood? In this case, population, density, and the best Gross Domestic Product (GDP) will be considered as good indicator. A place with more people and more concentration of people implies more customers, and when GDP is growing, especially if inflation is not a problem, is better for workers and businesses.

Data Section

I believe that the depth of a study, always depends on data available for analysis.

1.- Dataset

A dataset that contains Borough, Neighborhoods of New York City with their latitudes and longitudes will be use. The source of this dataset is https://cocl.us/new_york_dataset. This dataset will help to explore the neighborhoods and start to answer the question: What is the best neighborhood should be considered for a bakery.
The output of the exploring is:
First, latitude and longitude of every neighborhood.

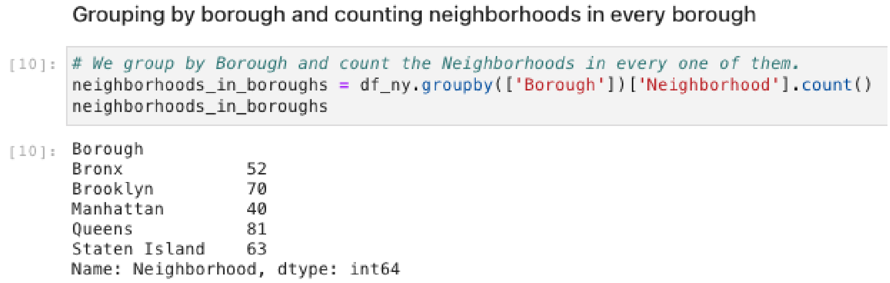


```
[8]: df_ny.head()
```

[8]:

	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

Second, the borough with more neighborhood.

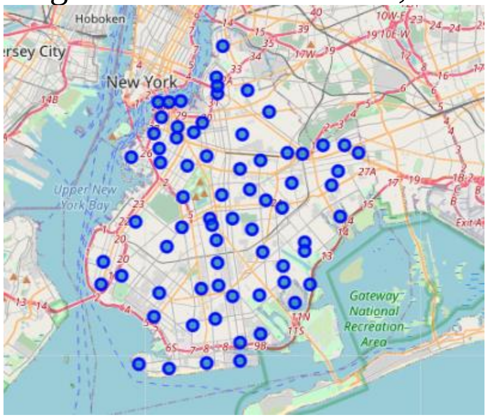


Grouping by borough and counting neighborhoods in every borough

```
[10]: # We group by Borough and count the Neighborhoods in every one of them.
      neighborhoods_in_boroughs = df_ny.groupby(['Borough'])['Neighborhood'].count()
      neighborhoods_in_boroughs
```

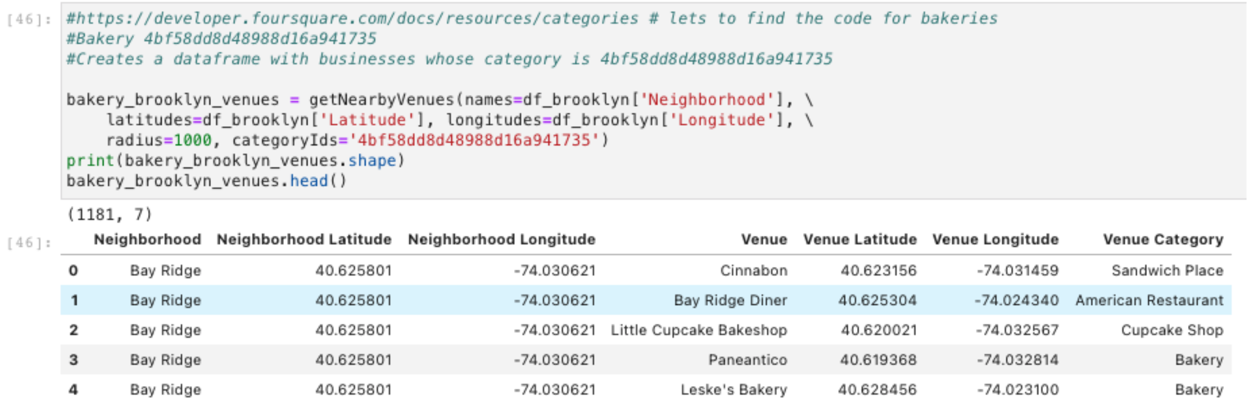
```
[10]: Borough
      Bronx            52
      Brooklyn         70
      Manhattan        40
      Queens           81
      Staten Island    63
      Name: Neighborhood, dtype: int64
```

Third, The map-visualization of neighborhoods in the borough with more neighborhood. In this case, Brooklyn.



2.- Link

The link https://developer.foursquare.com/docs/resources/categories is a resource for developers with codes that helps to identify the type of business in venues when Foursquare API is used. So, it lets to filter the type of business. The code for bakeries is 4bf58dd8d48988d16a941735. There is a little issue with this code because it also includes a couple other type of business. Even though this additional businesses are related with bakeries, an additional filter is necessary to select only bakeries. The first filter gives 1180 business with code (4bf58dd8d48988d16a941735), but the real number of bakeries in Brooklyn is 945. The map shows 1180 business.



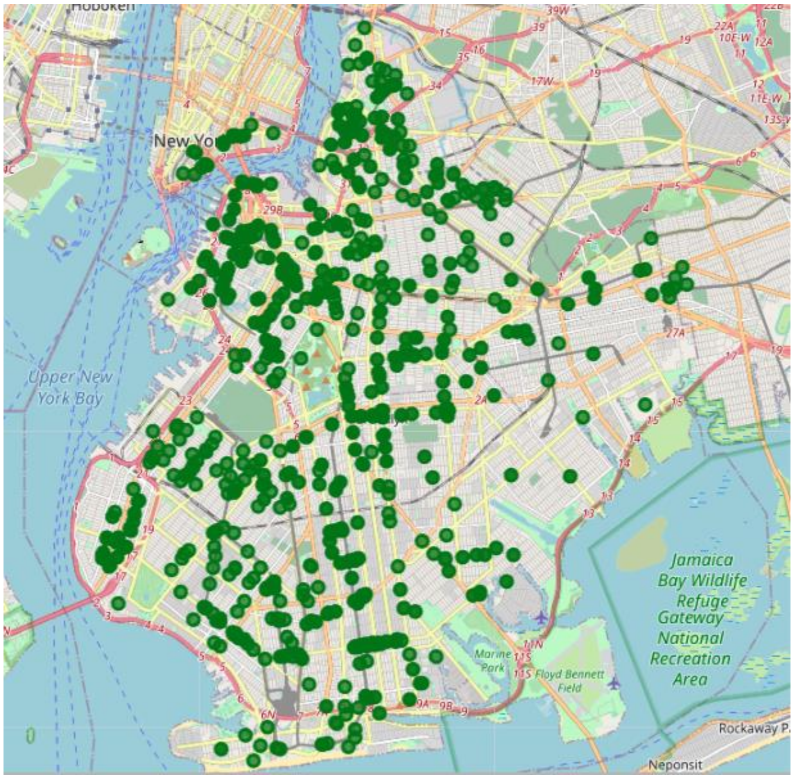
```
[46]: #https://developer.foursquare.com/docs/resources/categories # lets to find the code for bakeries
      #Bakery 4bf58dd8d48988d16a941735
      #Creates a dataframe with businesses whose category is 4bf58dd8d48988d16a941735

      bakery_brooklyn_venues = getNearbyVenues(names=df_brooklyn['Neighborhood'], \
          latitudes=df_brooklyn['Latitude'], longitudes=df_brooklyn['Longitude'], \
          radius=1000, categoryIds='4bf58dd8d48988d16a941735')
      print(bakery_brooklyn_venues.shape)
      bakery_brooklyn_venues.head()
```

```
(1181, 7)
```

[46]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bay Ridge	40.625801	-74.030621	Cinnabon	40.623156	-74.031459	Sandwich Place
1	Bay Ridge	40.625801	-74.030621	Bay Ridge Diner	40.625304	-74.024340	American Restaurant
2	Bay Ridge	40.625801	-74.030621	Little Cupcake Bakeshop	40.620021	-74.032567	Cupcake Shop
3	Bay Ridge	40.625801	-74.030621	Paneantico	40.619368	-74.032814	Bakery
4	Bay Ridge	40.625801	-74.030621	Leske's Bakery	40.628456	-74.023100	Bakery



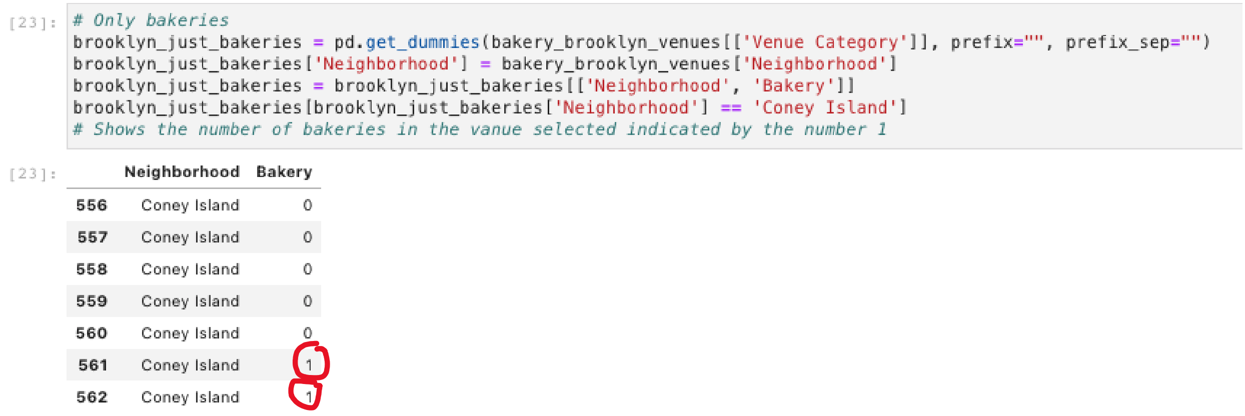
Analysis

We already can make an analysis with this information. Brooklyn has the highest number of neighborhoods, so is a good borough to analyze

Grouping the businesses by venue, we can get the number of related businesses in every venue in Brooklyn.

Identifying the number of bakeries in every venue we can get the average of bakeries. The venue with the smallest average could be a good choice for a bakery. Or a venue with small average with venues around that have small average.

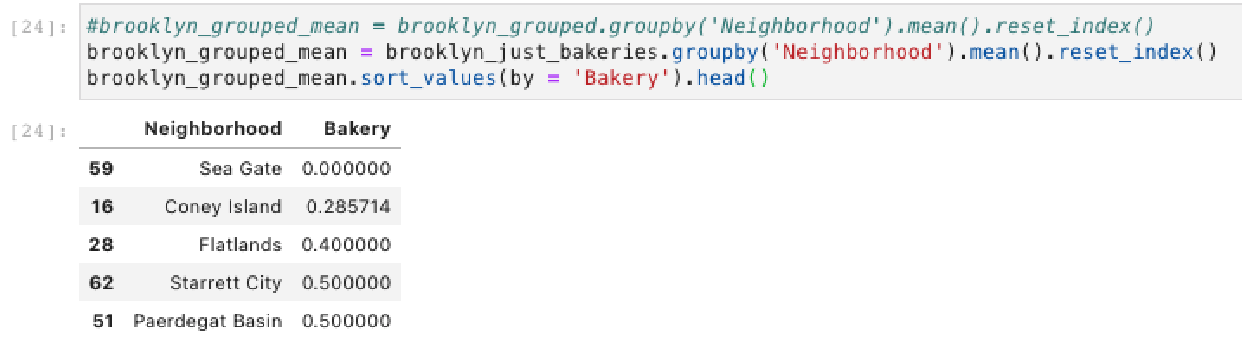
Under this criteria Coney Island could be a good choice.



```
[23]: # Only bakeries
brooklyn_just_bakeries = pd.get_dummies(bakery_brooklyn_venues[['Venue Category']], prefix="", prefix_sep="")
brooklyn_just_bakeries['Neighborhood'] = bakery_brooklyn_venues['Neighborhood']
brooklyn_just_bakeries = brooklyn_just_bakeries[['Neighborhood', 'Bakery']]
brooklyn_just_bakeries[brooklyn_just_bakeries['Neighborhood'] == 'Coney Island']
# Shows the number of bakeries in the vanue selected indicated by the number 1
```

[23]:

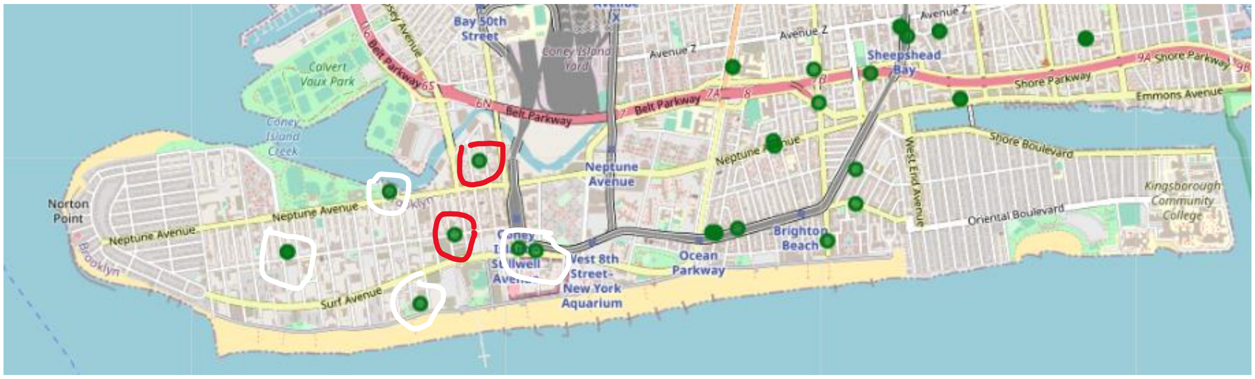
	Neighborhood	Bakery
556	Coney Island	0
557	Coney Island	0
558	Coney Island	0
559	Coney Island	0
560	Coney Island	0
561	Coney Island	1
562	Coney Island	1



```
[24]: #brooklyn_grouped_mean = brooklyn_grouped.groupby('Neighborhood').mean().reset_index()
brooklyn_grouped_mean = brooklyn_just_bakeries.groupby('Neighborhood').mean().reset_index()
brooklyn_grouped_mean.sort_values(by = 'Bakery').head()
```

[24]:

	Neighborhood	Bakery
59	Sea Gate	0.000000
16	Coney Island	0.285714
28	Flatlands	0.400000
62	Starrett City	0.500000
51	Paerdegat Basin	0.500000

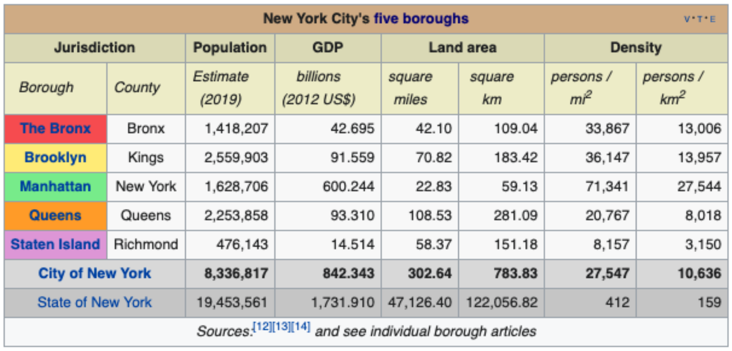


There are few numbers of bakeries in a broad area.

Additional data for and additional analysis.

Other dataset that could be explored is one that helps to analyze the population, the Gross Domestic Product, and the density of people in every borough. It is supposed that a borough with more population, a higher GDP, and a higher density is a good option to open the business.

The source of this dataset is https://en.wikipedia.org/wiki/Demographics_of_New_York_City



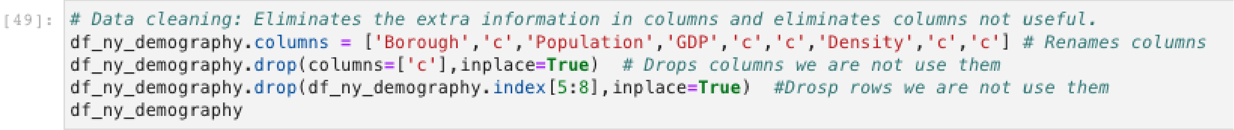
New York City's five boroughs

v • t • e

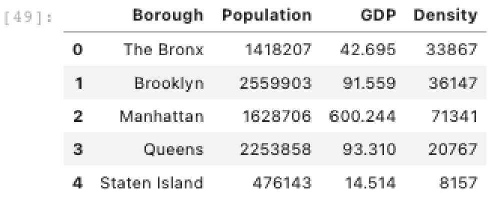
Jurisdiction		Population	GDP	Land area		Density	
Borough	*County*	*Estimate (2019)*	*billions (2012 US$)*	*square miles*	*square km*	*persons / mi²*	*persons / km²*
The Bronx	Bronx	1,418,207	42.695	42.10	109.04	33,867	13,006
Brooklyn	Kings	2,559,903	91.559	70.82	183.42	36,147	13,957
Manhattan	New York	1,628,706	600.244	22.83	59.13	71,341	27,544
Queens	Queens	2,253,858	93.310	108.53	281.09	20,767	8,018
Staten Island	Richmond	476,143	14.514	58.37	151.18	8,157	3,150
City of New York		**8,336,817**	**842.343**	**302.64**	**783.83**	**27,547**	**10,636**
State of New York		19,453,561	1,731.910	47,126.40	122,056.82	412	159

Sources:[12][13][14] *and see individual borough articles*

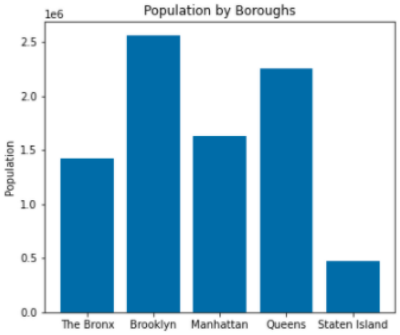
The output of the exploring is: Population, GDP, and density of every borough in New York City.

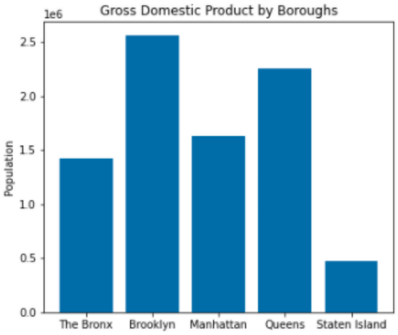


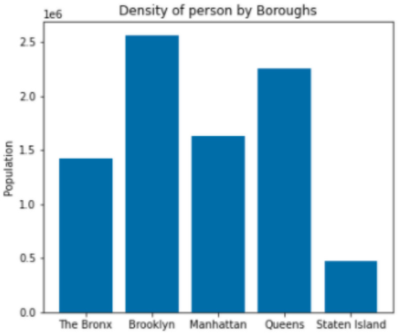
```
[49]: # Data cleaning: Eliminates the extra information in columns and eliminates columns not useful.
      df_ny_demography.columns = ['Borough','c','Population','GDP','c','c','Density','c','c'] # Renames columns
      df_ny_demography.drop(columns=['c'],inplace=True)  # Drops columns we are not use them
      df_ny_demography.drop(df_ny_demography.index[5:8],inplace=True)  #Drosp rows we are not use them
      df_ny_demography
```



```
[49]:        Borough  Population      GDP  Density
      0    The Bronx     1418207   42.695    33867
      1     Brooklyn     2559903   91.559    36147
      2    Manhattan     1628706  600.244    71341
      3       Queens     2253858   93.310    20767
      4  Staten Island    476143   14.514     8157
```







Analysis

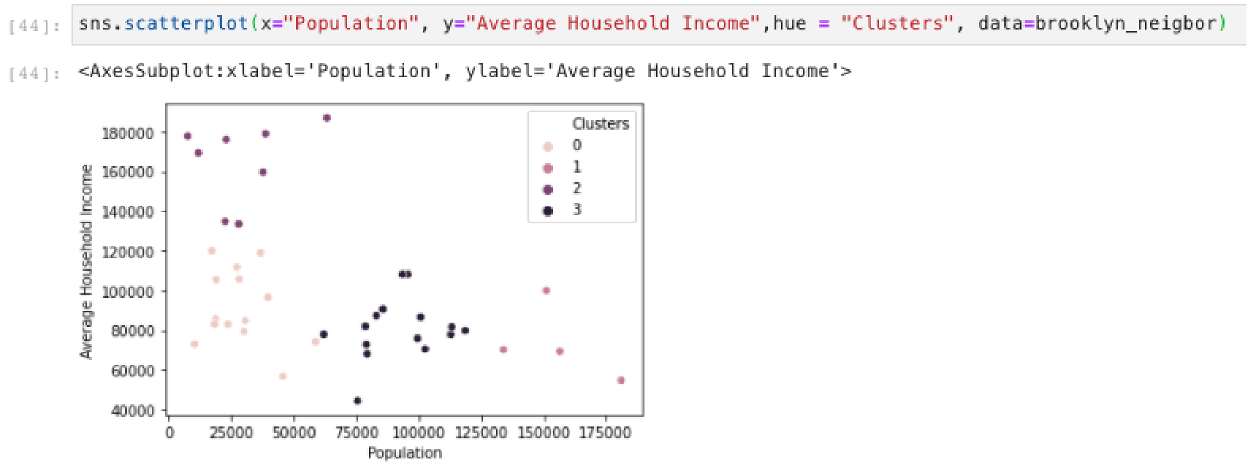
Population, density and GDP also show that Brooklyn is a good choice to be explored as the analysis before showed.

Dataset

A final dataset used for analysis is an excel worksheet built using information extracted from the web site www.point2homes.com.
Example: www.point2homes.com/US/Neighborhood/NY/Brooklyn/Brooklyn-Height-Demographics.html.
This dataset contains the "Population" and "Average Household Income" (AHI) of venues in Brooklyn that also are good indicators for analysis. Populated neighborhood with a high household income is a good choice too.

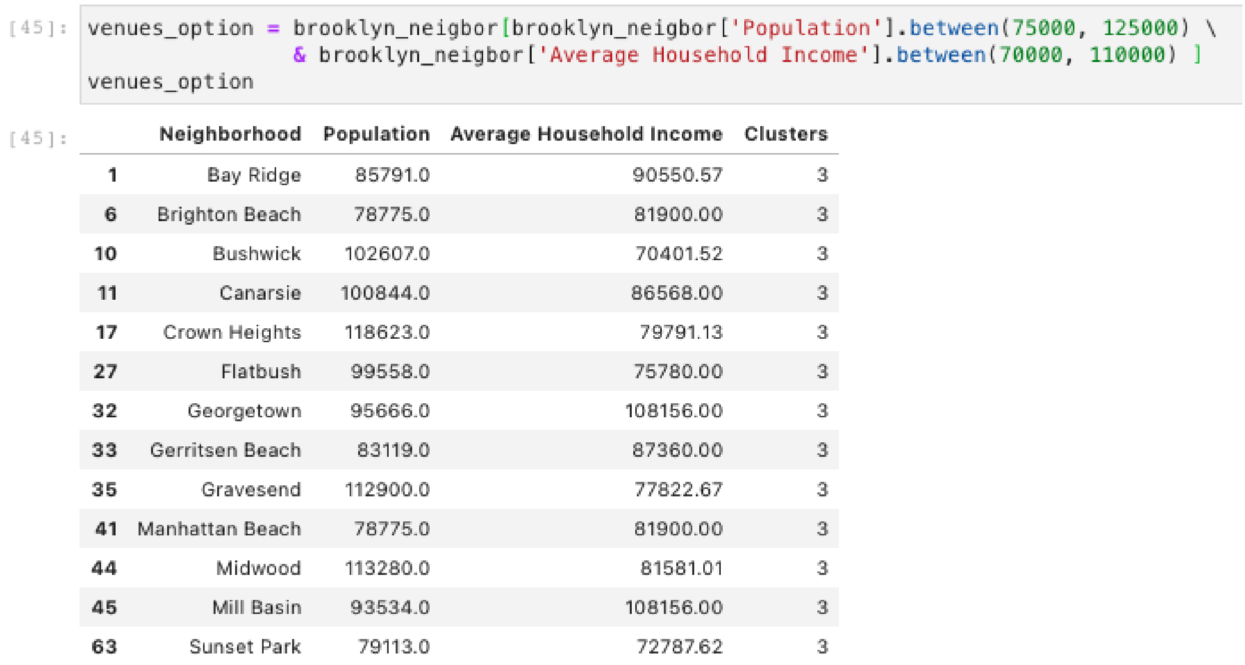
Analysis

Clustering shows us venues with high AHI and low population, venues with high population and low AHI, and those that are more balanced.
Because bread is not an expensive product and could be a daily necessity, population should have more weight than income. So, the venues under these criteria can be found in cluster 2, with population between 75,000 and 125,000, and AHI between 70,000 and 110,000.



```
[44]: sns.scatterplot(x="Population", y="Average Household Income",hue = "Clusters", data=brooklyn_neigbor)

[44]: <AxesSubplot:xlabel='Population', ylabel='Average Household Income'>
```

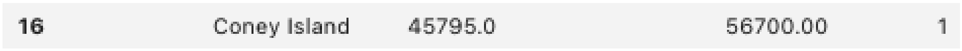


```
[45]: venues_option = brooklyn_neigbor[brooklyn_neigbor['Population'].between(75000, 125000) \
                      & brooklyn_neigbor['Average Household Income'].between(70000, 110000) ]
      venues_option
```

[45]:

	Neighborhood	Population	Average Household Income	Clusters
1	Bay Ridge	85791.0	90550.57	3
6	Brighton Beach	78775.0	81900.00	3
10	Bushwick	102607.0	70401.52	3
11	Canarsie	100844.0	86568.00	3
17	Crown Heights	118623.0	79791.13	3
27	Flatbush	99558.0	75780.00	3
32	Georgetown	95666.0	108156.00	3
33	Gerritsen Beach	83119.0	87360.00	3
35	Gravesend	112900.0	77822.67	3
41	Manhattan Beach	78775.0	81900.00	3
44	Midwood	113280.0	81581.01	3
45	Mill Basin	93534.0	108156.00	3
63	Sunset Park	79113.0	72787.62	3

Although, Coney Island is not in the range of this dataframe, still is an acceptable option under these last criteria.



16	Coney Island	45795.0	56700.00	1

Conclusion

Here are the analysis criteria and the different options resulting.
Even though data used in this case help to make good decisions, more data could be necessary to make a deeper analysis and answer questions like; what is the income of bakeries around the selected areas? or what is the rate of crimes in that areas? etc. that could affect the business.
How many variables should be considered for an analysis? Depend on how deep is necessary going is to satisfy the stakeholders requirements.