Capstone Project - The Battle of the Neighborhoods (Week 4)

Applied Data Science Capstone by IBM/Coursera

Choosing the best neighborhood to locate a Zimbabwean cuisine restaurant in New York City

Table of contents

1. Introduction and problem definition.

During a recent trip to New York, a Zimbabwean delegation noticed several African cuisine restaurants but none serving Zimbabwean dishes. One of the business delegation members seeks to open a restaurant specifically to serve Zimbabwean cuisine to ride on the growth in demand of African cuisines in a cosmopolitan city like New York.

No matter where in the world one considers starting a new business, they are bound to be both nervous and anxious. For an entrepreneur, it can be nerve-wracking attempting to manoeuvre through a big city looking for a suitable location to establish company operations. Location is one of the pivotal points of consideration and can affect almost every aspect of the company. Taking the time to investigate multiple locations wastes both time and money. It is judicious to ask for the services of a data scientist so that they recommend the most suitable location based on agreed criteria.

For this project, I seek to answer the question; '*Where would we recommend that a Zimbabwean cuisine restaurant be located in the city of New York*?'

Where to locate the restaurant business depends in part on the location of the target market (expatriate areas with a limited number of African restaurants), business partners (other expat restaurants to assist with visibility), and personal preferences (easy access from subway and airports).

We aim to come up with clusters of recommended neighborhoods and offer our client options to consider as they select the most suitable location. Results of this work can be of interest to anyone looking to establish a restaurant business in a big city like New York.

2. Data

The data we need to narrow down the site choices of the restaurant as contemplated in this project is as follows: expatriate areas already having African restaurants, near a financial district, and a transport hub.

We will show skills of moving data around, using databases, using APIs, performing data transformation, and creating visualizations using Python.

Data will be gathered from https://geo.nyu.edu/catalog/nyu_2451_34572 and https://www.restohub.org/

I shall use the Foursquare API to explore neighborhoods in New York City and apply the explore function to get the most common venue categories in each of the neighborhoods. I will then use this feature to group the neighborhoods into clusters and use the k-means clustering algorithm to refine the groupings. Finally, I will use the Folium library to visualize the neighborhoods in New York City and their emerging clusters as well as push the right neighborhood for setting up the restaurant business.

3. Methodology and Analysis

1. Download and Explore Dataset

2. Explore Neighborhoods in New York City

3. Analyse Each Neighborhood

4. Cluster Neighborhoods

5. Examine Clusters

I shall use Github as the repository to share my notebook for this project, Medium for my blog post and k-means for clustering neighborhoods.

3.1 New York City Dataset

New York City, in the state of New York, is by far the largest city in the United States, with an estimated 2016 population of 8.55 million. The city features **5** separate boroughs: *Staten Island, The Bronx, Brooklyn, Queens, and Manhattan* and **306** neighborhoods. The latest research shows that people who live in New York City have a higher life expectancy than the rest of the country. In 2010, the life expectancy of a person living in New York City was 80.9 years of age which is 2.2 years longer than the life expectancy of the entire country.

The income disparity between the citizens of New York City is vast. According to the latest census, the median household income for a wealthy citizen was $188,697 per year, and the poorest median income was reported at $9,320. New York City's population is expected to reach 9 million by 2040, based on recent projections created by the city. Among the five boroughs, the Bronx's growth is projected to be the highest at 14% between 2010 and 2040. On the flip side, Manhattan is expected to grow by 6.7% by 2040.

If these projections are accurate, Brooklyn will extend its lead over Queens as the largest borough in New York City, growing to nearly 3 million by 2040.
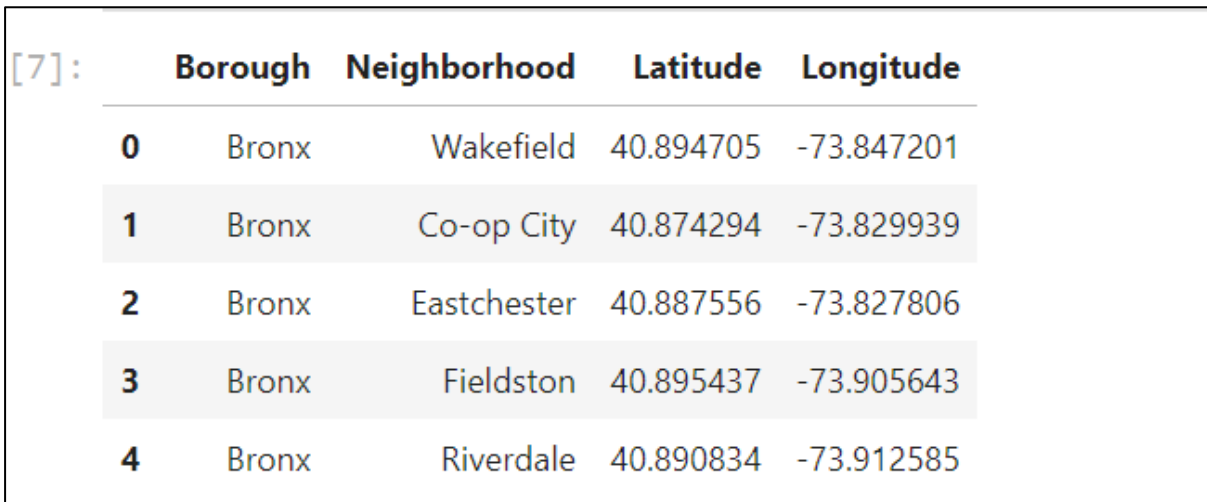
Download and exploratory analysis

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.

Luckily, this dataset exists for free on the web through the link: https://geo.nyu.edu/catalog/nyu_2451_34572

For our convenience, the downloaded files are available on the IBM server, so we can simply run a '*wget*' command and access the data.

The json file is read as a tuple for New York City neighbourhood data and we realise that all the information required is under the '*feature*' key. Our intention is to end up with a dataframe like Figure 1 that contains all the information of interest to enable exploratory analysis.

	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

Figure 1: New York City nyneighborhoods dataframe top 5 rows only

Using Folium to plot a map of New York City

I use geopy library to get the latitude and longitude values of New York City and create a map of New York City with neighborhoods superimposed on top as shown in Figure 2.
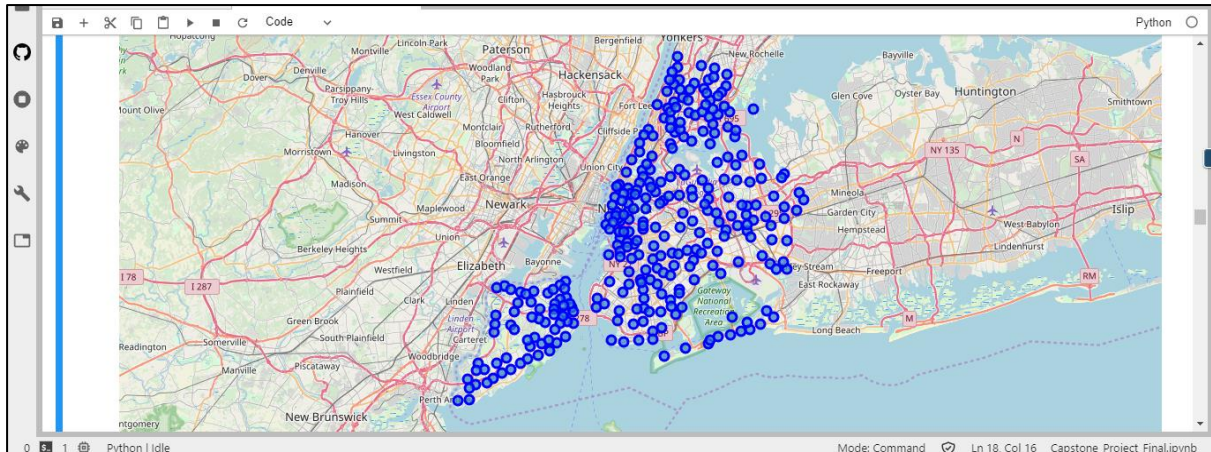
Figure 2: Folium map for New York City showing all neighborhoods

3.2 Exploring neighborhoods in New York City

We are going to utilize the Foursquare API to explore the neighborhoods in New York City and segment them in order to come up with a short list of possible locations for the Zimbabwean restaurant guided by the mentioned preconditions of other African restaurants, near a financial district and also near a major transportation hub.

To begin with we look at one neighborhood using the **GET request URL** to obtain a json file, clean it up and create a dataframe. We repeat this process for all the neighborhoods in New York in order to get top 100 venues within 1km of the chosen center as provided by the coordinates from Foursquare.

We apply the **get_category_type function** from the Foursquare lab, clean the json file and create a pandas dataframe as shown in Figure 3.

	name	categories	lat	lng
0	Lollipops Gelato	Dessert Shop	40.894123	-73.845892
1	Ripe Kitchen & Bar	Caribbean Restaurant	40.898152	-73.838875
2	Ali's Roti Shop	Caribbean Restaurant	40.894036	-73.856935
3	Rite Aid	Pharmacy	40.896649	-73.844846
4	Jackie's West Indian Bakery	Caribbean Restaurant	40.889283	-73.843310

Figure 3: Top 5 rows for nearby_venues dataframe from nearby venues to Wakefield

In total 46 venues were returned by Foursquare for the Wakefield neighborhood.

Repeating this approach for the entire city of New York and consolidating the information in one dataframe for ease of manipulation and analysis yields a huge dataframe of 10 382 rows with the top 5 rows as shown in Figure 4.

[22]:	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue_Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898276	-73.850381	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Figure 4: Top 5 rows for nyneighborhoods_venues dataframe for the venues returned for all New York City neighborhoods

The number of unique categories to be curated from all returned venues is 427. Note that this number changes depending on time of the day you run the Foursquare API. Fordham has 91 venues, Central Harlem has 42 and University Heights has 28. We will realise later that these are our own 3 options to locate the restaurant.

3.3 Analysis of Neighborhoods

When extracting features, from a dataset, it is often useful to transform categorical features into vectors so that you can do vector operations. Pandas provides the very useful get_dummies method on Dataframe, which does what we want. By default, the get_dummies() does not do dummy encoding, but one-hot encoding and that will be adequate for our needs. The column values will return 1 where the given venue exists for that row or 0 on the contrary.

Figure 5 shows typical one hot encoding results.

[25]:		Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	Auditoriu
0		0	0	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	0	0	0	
2		0	0	0	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	0	0	0	
4		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figure 5: One hot encoded results dataframe head

We need to group rows by neighborhood and take the mean of the frequency of occurrence of each category in order to see the top venues around each neighborhood. We can create a dataframe from a sorted list of these results and get an overview of the top 10 types of venues around neighborhoods in New York City as shown in Figure 6.

[30]:	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	Allerton	Pizza Place	Deli / Bodega	Supermarket	Pharmacy	Chinese Restaurant	Spa	Spanish Restaurant	Martial Arts Dojo	Gas Station	Breakfast Spot
1	Annadale	Pizza Place	Cosmetics Shop	Dance Studio	Park	Pub	Train Station	American Restaurant	Restaurant	Sports Bar	Diner
2	Arden Heights	Pharmacy	Coffee Shop	Business Service	Pizza Place	Women's Store	Financial or Legal Service	Factory	Falafel Restaurant	Farm	Farmers Market
3	Arlington	Intersection	Coffee Shop	American Restaurant	Bus Stop	Women's Store	Fish Market	Factory	Falafel Restaurant	Farm	Farmers Market
4	Arrochar	Deli / Bodega	Bus Stop	Italian Restaurant	Pizza Place	Outdoors & Recreation	Sandwich Place	Athletics & Sports	Mediterranean Restaurant	Bagel Shop	Hotel

5

Figure 6: Neighborhoods and their top 10 venues dataframe head

3.4 Clustering neighborhoods

In clustering, we try to find homogeneous subgroups within our data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

In this case we are going to achieve clustering of neighborhoods based on types of nearby venues. We seek to segment neighborhoods in such a way that we can locate those that best meet the criteria of our client who wants to locate a Zimbabwean cuisine restaurant. Our objective is to find the most suitable location based on proximity to expatriates (African restaurants), financial district (banks) and a transport hub (along a major road or subway).

Clustering is considered an unsupervised learning method since we don't have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. We only want to try to investigate the structure of the data by grouping the data points into distinct subgroups.

In this project, we will use K-means which is considered as one of the most used clustering algorithms due to its simplicity. At a glance the K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.

The way the k-means algorithm works is as follows:

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

We assume that our data is suitable for k-means clustering and that 5 clusters give optimum results. The resulting dataframe is shown in Figure 7 and Folium map with the clusters can be visualised in Figure 8.

	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 Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bronx	Wakefield	40.894705	-73.847201	0.0	Pharmacy	Gas Station	Sandwich Place	Laundromat	Ice Cream Shop	Caribbean Restaurant	Donut Shop	Dessert Shop	Food Truck	Farmers Market
1	Bronx	Co-op City	40.874294	-73.829939	0.0	Bus Station	Baseball Field	Ice Cream Shop	Gift Shop	Fast Food Restaurant	Grocery Store	Park	Mattress Store	Pharmacy	Pizza Place
2	Bronx	Eastchester	40.887556	-73.827806	3.0	Bus Station	Caribbean Restaurant	Deli / Bodega	Diner	Metro Station	Bus Stop	Intersection	Business Service	Chinese Restaurant	Seafood Restaurant
3	Bronx	Fieldston	40.895437	-73.905643	1.0	River	Bus Station	Plaza	Women's Store	Fish & Chips Shop	Eye Doctor	Factory	Falafel Restaurant	Farm	Farmers Market
4	Bronx	Riverdale	40.890834	-73.912585	1.0	Bus Station	Park	Plaza	Playground	Bank	Food Truck	Gym	Home Service	Factory	Falafel Restaurant

Figure 7: Clusters and the top 10 venues for each neighborhood.
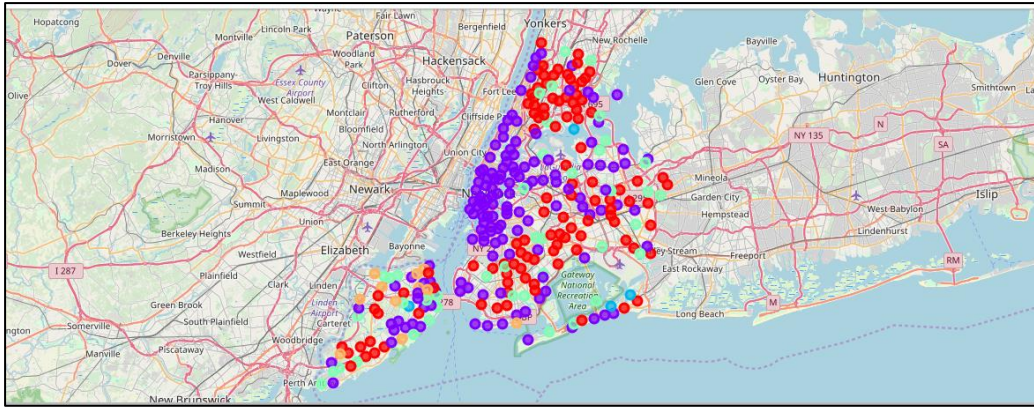
6

Figure 8: Visualisation of the clusters using Folium

3.5 Examining clusters

Now, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, we can then assign a name to each cluster.

On examining the first cluster we notice that there is no occurrence of our first condition which is having other African restaurants in the vicinity. There is no need to reinvent the wheel on locating areas where expat from Africa are most likely to settle or visit. From the foregoing it seems that our initial condition of at least 5 African restaurants is rather stringent and won't be met easily. We will contend with the location that has a maximum number possible instead.

We revisit the *nyneighborhoods_venues* dataframe and zoom in to locate neighborhoods with African restaurants (see Figure 9). On locating these neighborhoods we will then assess them to find a location that is closest to banks and a transport hub. That place is likely to be the most suitable location for the Zimbabwean restaurant for our client and a similar approach can be employed by any entrepreneur looking at locating a business in a big city.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue_Category
325	University Heights	40.855727	-73.910416	Accra Resturant	40.853871	-73.908421	African Restaurant
375	Fordham	40.860997	-73.896427	Papaye Restaurant	40.857407	-73.899738	African Restaurant
3815	Central Harlem	40.815976	-73.943211	Ponty Bistro Harlem	40.817886	-73.941522	African Restaurant
3826	Central Harlem	40.815976	-73.943211	Keur Sokhna	40.813556	-73.945001	African Restaurant
3837	Central Harlem	40.815976	-73.943211	Africa Kine Restaurant	40.813728	-73.944426	African Restaurant

Figure 9: Dataframe showing neighborhoods with African Restaurant venues

Since we only have five venues satisfying our first criteria of which they are found in 3 neighborhoods we can simplify Figure 8, and examine only the venues in **University Heights, Fordham and Central Harlem.** So let's slice the original dataframe and create a new dataframe of **that meets our highlighted criteria if any** to start with. The resulting venues are shown in Figure 10.

Figure 10: Visualisation of the 3 neighborhoods and venues

Let's visualise some quick statistics on the frequency of different restaurant types in our chosen neighborhoods (see Figure 11).
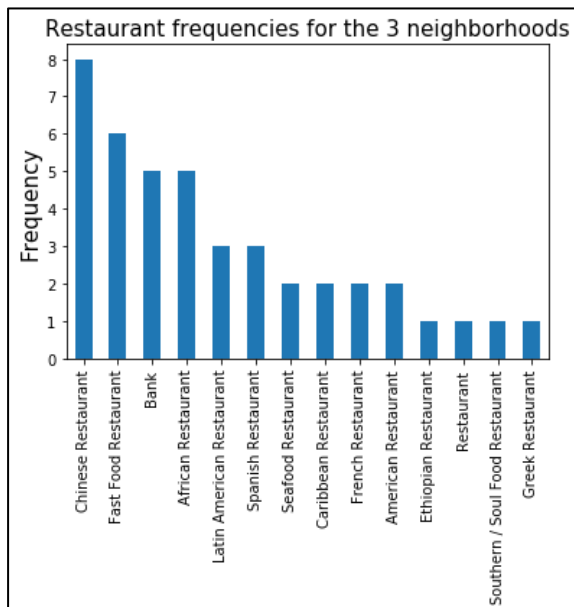

Figure 11: Types of restaurants in the 3 neighborhoods

A Wordcloud (or Tag cloud) is a visual representation of text data as shown in Figure 12. It displays a list of words, the importance of each being shown with font size or color. This format is useful for quickly perceiving the most prominent terms. We seek to employ it in this case to give our client a quick proof that restaurants dominate the venues across the three neighborhoods. This indicates that we are indeed using the appropriate data for our analysis.

Figure 12: WordCloud for all Venues Categories for the 3 neighborhoods

4. Discussion

Option 1 (Best)

Harlem is a neighborhood in the northern section of the New York City borough of Manhattan. It is an Uptown area with a population of 200,754. The employment numbers show that there are 89.47% white collar employees and 10.53% blue collar employees in Harlem.

Central Harlem is closest to the initial condition of our client which is to locate near other African restaurants. There are 4 African restaurants in this particular neighborhood and therefore it seems the location is already popular with African expatriates or visitors.

Option 2

Fordham is geographically located in the west Bronx, New York City. Fordham has a population of 77,840 people. Fordham Road's successful mixture of small independent shops and national and regional chain stores has kept it as the largest and most prominent shopping district in the Bronx and one of the largest in New York City. Fordham is a bustling, vibrant community with a lot of business potential to an aspirational entrepreneur.

Option 3

University Heights is a low income residential neighborhood geographically located in the west Bronx, New York City. The neighborhood is part of Bronx. University Heights has a population of over 40,000. For decades University Heights has been one of the poorest communities in America. Over half the population lives below the poverty line and receives public assistance (AFDC, Home Relief, Supplemental Security Income, and Medicaid). The neighborhood is now predominantly Dominican with a significant longstanding Puerto Rican and African American population. The vast majority of households are renter occupied.

University Heights has the least of venues perhaps due to it being a college environment. This may be the least suitable place to locate since already they are few such facilities. There could be a barrier of entry which we need to look closely at.

5. **Conclusion**

As the most populated city in the United States, New York City (NYC) is a great place to open a restaurant. Research shows that the city's numbers have reached a record high of more than 8.6 million people, an increase of more than 5% in the last few years – and everyone needs to eat! Harlem has a huge population of employed people who already have expatriate tastes and is thus the best location.

As the population continues to grow, the number of restaurants in NYC keeps growing as well. According to the National Restaurant Association, there were nearly 45,000 eating and drinking establishments, nearly 26,700 open restaurants in the city alone. Though the competition is high among New York City restaurants, there is great potential for success with creative differentiation like the proposed offer for a Zimbabwean cuisine.

Choosing the right location for a restaurant is extremely important. Not only will it affect how much rentals are paid, but it also affects the target audience, competition, and the spaces available to choose from.

It is wise to think upfront about location and take some time to explore different areas of a city to see what could be a good fit. This project performed location analysis for a prospective restaurateur and the same approach can be considered by any entrepreneur looking at starting a business or expanding an existing operation.

Data science is a crucial skill that enables informed decision making.

References

1. https://geo.nyu.edu/catalog/nyu_2451_34572
2. http://worldpopulationreview.com/us-cities/new-york-city-population/
3. http://www.citypopulation.de/en/usa/newyorkcity/
4. https://www.restohub.org/operations/planning/new-york-restaurant/
5. https://www.restohub.org/operations/planning/choosing-a-restaurant-location/
6. https://urbanareas.net/info/resources/neighborhoods
7. https://www.point2homes.com/US/Neighborhood/NY/
8. https://furmancenter.org/neighborhoods/view/