The Best Neighborhood in New Jersey for setting up a outpatient physical therapy facility

1. Introduction

1.1 Business Problem

A friend of mine is thinking of moving to New Jersey to open her physical therapy facility. Both of us met for a drink and she discussed her ideas with me. She has every other thing figured out except the best location (optimal neighborhood) to set up the facility. Her criteria for the optimal neighborhood is as follows:

- Less crime (safe areas)
- Cost of rent
- Close to long-term health care facilities e.g. hospitals, nursing homes, home-care centers, rehabs etc.

This report outlines some basic assumptions, data sets, and analysis that can inform our decision when selecting the optimal neighborhood in New Jersey for setting up a physical therapy facility.

1.2 Target audience

The target audience of this report would be anyone who wants to buy or set up a physical therapy business in New Jersey. This report will also be useful for government, hospitals or other health care facilities that are interested in setting up outpatient physical therapy facility.

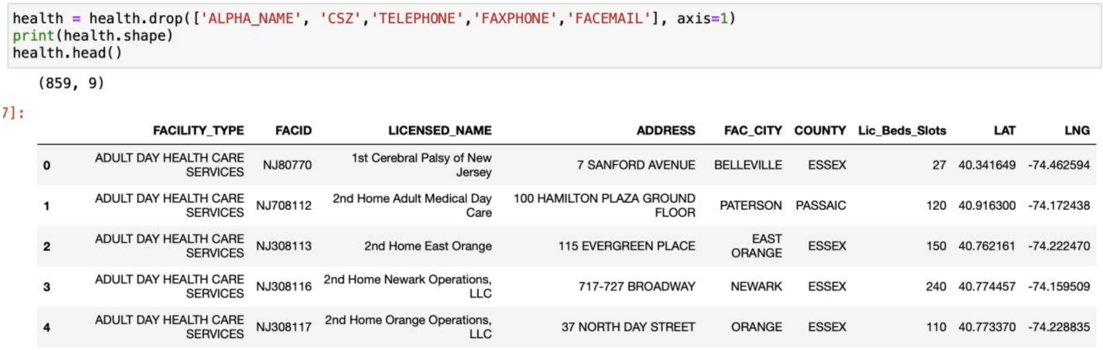
2.0 Data

Making the best decision for the best neighborhood for a physical therapy facility is not a trivial process because several factors have to be considered such as the closeness to long-term health care facilities, areas with low crime rate and cost of rent. To be able to make the best decision, data is needed. Fortunately, New Jersey has several public databases that describes various aspects of the state and Foursquare API allows free access to some of its venue and location data. We will use four sets of data for our analysis. They are:

1. New Jersey Health care facilities data
2. New Jersey Crime Data
3. New Jersey average rent data
4. Foursquare data

2.1 New Jersey Health care facilities data

We have a list of all registered health care facilities in New Jersey from New Jersey Department of Health website - https://healthapps.state.nj.us/Facilities/fsSearch.aspx. We downloaded this data and place it into a pandas dataframe using python. This data will help us to know where most of the health facilities are located. Our physical therapy facility will be situated close to one of these, since most of the people in this health care facilities are likely to need physical therapy.



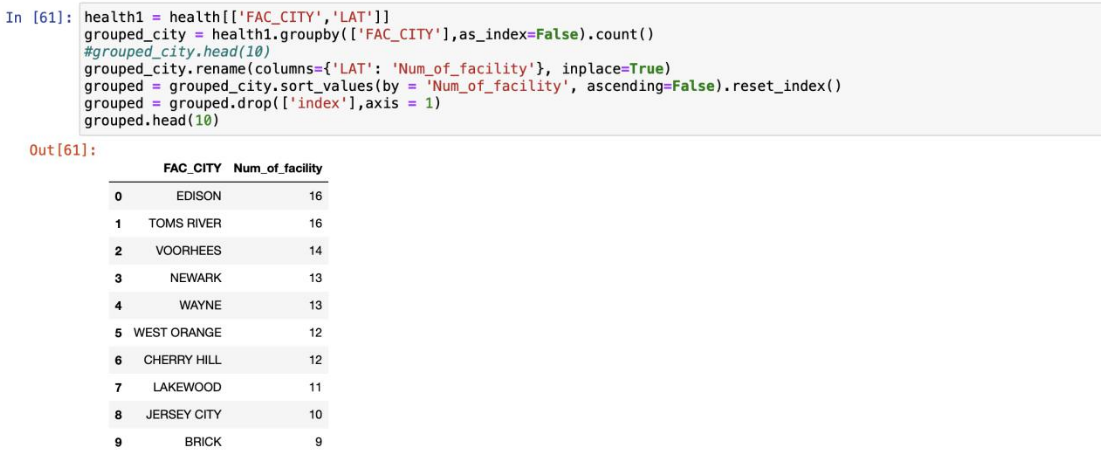
```
health = health.drop(['ALPHA_NAME', 'CSZ','TELEPHONE','FAXPHONE','FACEMAIL'], axis=1)
print(health.shape)
health.head()
```

```
(859, 9)
```

7]:

	FACILITY_TYPE	FACID	LICENSED_NAME	ADDRESS	FAC_CITY	COUNTY	Lic_Beds_Slots	LAT	LNG
0	ADULT DAY HEALTH CARE SERVICES	NJ80770	1st Cerebral Palsy of New Jersey	7 SANFORD AVENUE	BELLEVILLE	ESSEX	27	40.341649	-74.462594
1	ADULT DAY HEALTH CARE SERVICES	NJ708112	2nd Home Adult Medical Day Care	100 HAMILTON PLAZA GROUND FLOOR	PATERSON	PASSAIC	120	40.916300	-74.172438
2	ADULT DAY HEALTH CARE SERVICES	NJ308113	2nd Home East Orange	115 EVERGREEN PLACE	EAST ORANGE	ESSEX	150	40.762161	-74.222470
3	ADULT DAY HEALTH CARE SERVICES	NJ308116	2nd Home Newark Operations, LLC	717-727 BROADWAY	NEWARK	ESSEX	240	40.774457	-74.159509
4	ADULT DAY HEALTH CARE SERVICES	NJ308117	2nd Home Orange Operations, LLC	37 NORTH DAY STREET	ORANGE	ESSEX	110	40.773370	-74.228835

From the figure above, there is a total of 859 entries in the health care facility data frame. It has the name of each heath care facility, their type, ID, address, number of bed slots, latitude and longitude. This data frame contains other features but they have been dropped during the process of cleaning the data. We want to see the number of heath care facilities registered in New Jersey, grouped by city (neighborhood). After coding in python to get the neighborhood with the most heath care facilities, we get the data frame below:



```
In [61]: health1 = health[['FAC_CITY','LAT']]
         grouped_city = health1.groupby(['FAC_CITY'],as_index=False).count()
         #grouped_city.head(10)
         grouped_city.rename(columns={'LAT': 'Num_of_facility'}, inplace=True)
         grouped = grouped_city.sort_values(by = 'Num_of_facility', ascending=False).reset_index()
         grouped = grouped.drop(['index'],axis = 1)
         grouped.head(10)
```

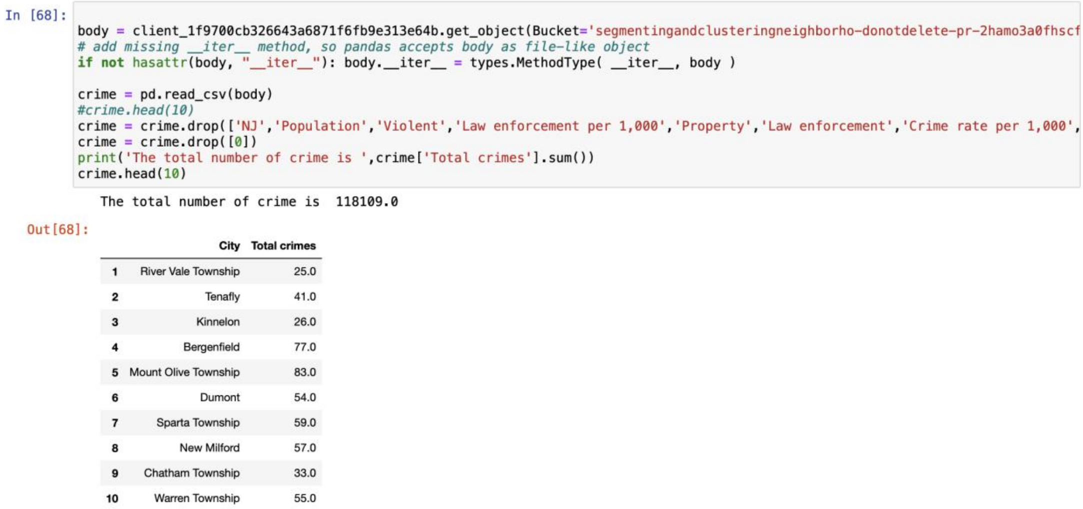
Out[61]:

	FAC_CITY	Num_of_facility
0	EDISON	16
1	TOMS RIVER	16
2	VOORHEES	14
3	NEWARK	13
4	WAYNE	13
5	WEST ORANGE	12
6	CHERRY HILL	12
7	LAKEWOOD	11
8	JERSEY CITY	10
9	BRICK	9

Edison and Toms river have the most heath care facilities in the state but the top 10 are all looking like they have a good amount of heath care facilities.

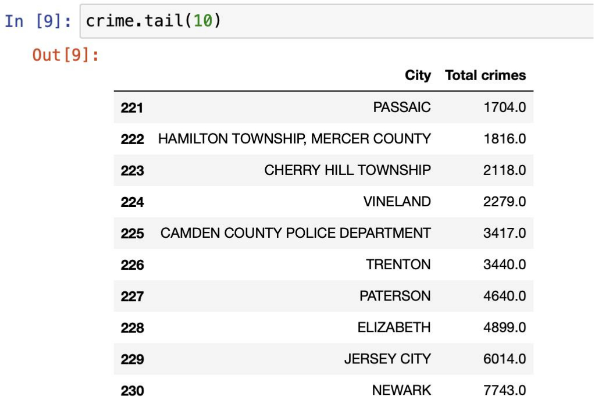
2.2 New Jersey Crime Data

We got the New Jersey crime data from the New Jersey state police website - https://www.njsp.org/ucr/uniform-crime-reports.shtml. This will help us to select the safest place for our physical therapy facility.



```
In [68]:
body = client_1f9700cb326643a6871f6fb9e313e64b.get_object(Bucket='segmentingandclusteringneighborho-donotdelete-pr-2hamo3a0fhscf
# add missing __iter__ method, so pandas accepts body as file-like object
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )

crime = pd.read_csv(body)
#crime.head(10)
crime = crime.drop(['NJ','Population','Violent','Law enforcement per 1,000','Property','Law enforcement','Crime rate per 1,000',
crime = crime.drop([0])
print('The total number of crime is ',crime['Total crimes'].sum())
crime.head(10)

The total number of crime is  118109.0

Out[68]:
```

	City	Total crimes
1	River Vale Township	25.0
2	Tenafly	41.0
3	Kinnelon	26.0
4	Bergenfield	77.0
5	Mount Olive Township	83.0
6	Dumont	54.0
7	Sparta Township	59.0
8	New Milford	57.0
9	Chatham Township	33.0
10	Warren Township	55.0



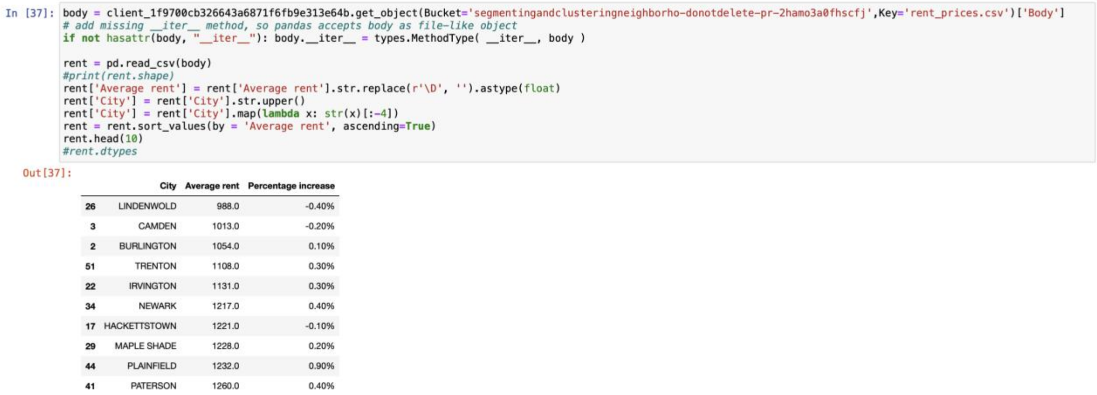
```
In [9]: crime.tail(10)

Out[9]:
```

	City	Total crimes
221	PASSAIC	1704.0
222	HAMILTON TOWNSHIP, MERCER COUNTY	1816.0
223	CHERRY HILL TOWNSHIP	2118.0
224	VINELAND	2279.0
225	CAMDEN COUNTY POLICE DEPARTMENT	3417.0
226	TRENTON	3440.0
227	PATERSON	4640.0
228	ELIZABETH	4899.0
229	JERSEY CITY	6014.0
230	NEWARK	7743.0

We have a total of 118,109 crime reports in this data frame. We grouped this data into different neighborhood using python code. This data frame gives us a good indication of how safe or dangerous each neighborhood is today. River Vale Township is the safest city in New Jersey while Newark is the most dangerous place to set up a business.

2.3 New Jersey average rent report

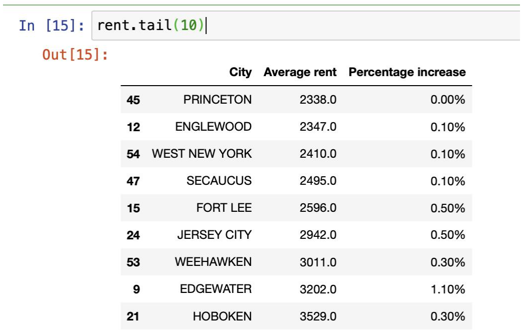
We will use the New Jersey average rent report to select a neighborhood with affordable rent. The data was downloaded from the website of RentCafe - https://infogram.com/rentcafe_new-jersey-rentreport-january2020-1h7k23n3m8dv6xr.



```
In [37]: body = client_1f9700cb326643a6871f6fb9e313e64b.get_object(Bucket='segmentingandclusteringneighborho-donotdelete-pr-2hamo3a0fhscfj',Key='rent_prices.csv')['Body']
         # add missing __iter__ method, so pandas accepts body as file-like object
         if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )

         rent = pd.read_csv(body)
         #print(rent.shape)
         rent['Average rent'] = rent['Average rent'].str.replace(r'\D', '').astype(float)
         rent['City'] = rent['City'].str.upper()
         rent['City'] = rent['City'].map(lambda x: str(x)[:-4])
         rent = rent.sort_values(by = 'Average rent', ascending=True)
         rent.head(10)
         #rent.dtypes
```

Out[37]:

	City	Average rent	Percentage increase
26	LINDENWOLD	988.0	-0.40%
3	CAMDEN	1013.0	-0.20%
2	BURLINGTON	1054.0	0.10%
51	TRENTON	1108.0	0.30%
22	IRVINGTON	1131.0	0.30%
34	NEWARK	1217.0	0.40%
17	HACKETTSTOWN	1221.0	-0.10%
29	MAPLE SHADE	1228.0	0.20%
44	PLAINFIELD	1232.0	0.90%
41	PATERSON	1260.0	0.40%



```
In [15]: rent.tail(10)
```

Out[15]:

	City	Average rent	Percentage increase
45	PRINCETON	2338.0	0.00%
12	ENGLEWOOD	2347.0	0.10%
54	WEST NEW YORK	2410.0	0.10%
47	SECAUCUS	2495.0	0.10%
15	FORT LEE	2596.0	0.50%
24	JERSEY CITY	2942.0	0.50%
53	WEEHAWKEN	3011.0	0.30%
9	EDGEWATER	3202.0	1.10%
21	HOBOKEN	3529.0	0.30%

From the figure, the LINDENWOLD neighborhood has the lowest rent price while Hoboken is the most expensive rent price in the state.

2.4 New Jersey Neighborhood map

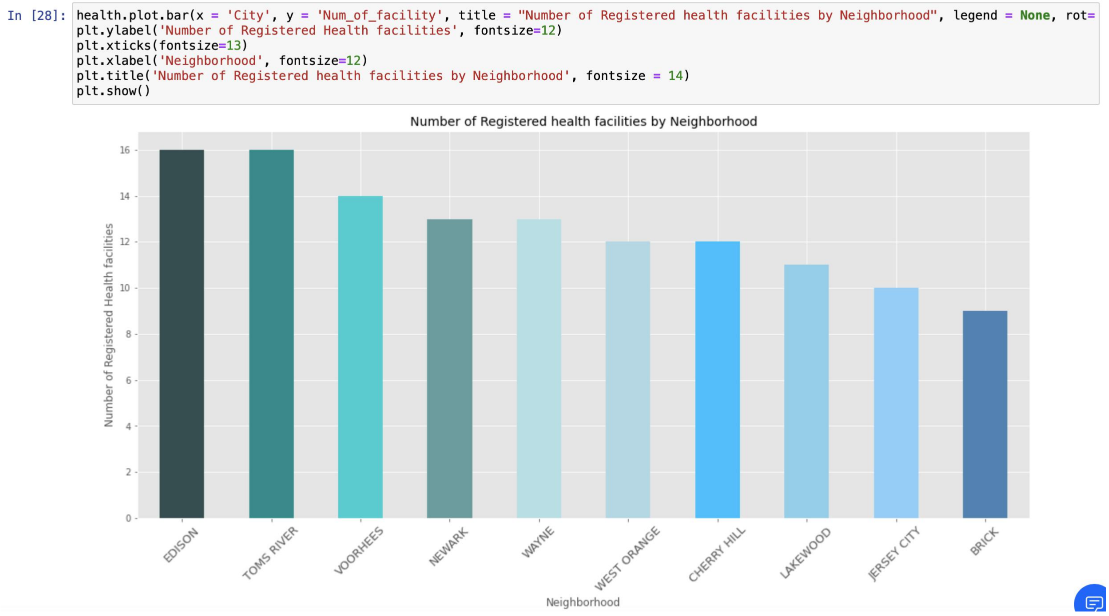
According to ontheworldmap.com - http://ontheworldmap.com/usa/state/new-jersey/ , there are 565 neighborhoods in New Jersey. The largest by population is Newark with 277,140 residents while the smallest is Tavistock.

New Jersey
State capital
County seat
City or Town
kilometers
0 10 20 30
0 10 20
miles
Copyright © Ontheworldmap.com
New York
Sussex
Newton
Passaic
Ringwood
Wanaque
Ramsey
Bergen
Ridgewood
Paramus
Fair Lawn
Paterson
Teaneck
Hackensack
Fort Lee
Hopatcong
Warren
Hackettstown
Belvidere
Dover
Morris
Morristown
Clifton
Essex
North Bergen
Union City
East Orange
Hudson
New York City
Madison
Summit
Newark
Jersey City
Bayonne
Elizabeth
Union
Westfield
Phillipsburg
North Plainfield
Plainfield
South Plainfield
Linden
Hunterdon
Somerville
Carteret
Middlesex
Metuchen
Manville
Somerset
Perth Amboy
Flemington
New Brunswick
South River
Keansburg
Middlesex
Pennsylvania
Red Bank
Princeton
Eatontown
Long Branch
Mercer
Freehold
Trenton
Tinton Falls
Monmouth
Asbury Park
Point Pleasant
Burlington
Moorestown
Mount Holly
Toms River
Seaside Heights
Camden
Ocean
Burlington
Woodbury
Barnegat Light
Penns Grove
Gloucester
Camden
Glassboro
Pennsville
Salem
Ship Bottom
Salem
Hammonton
Beach Haven
Vineland
Atlantic
Mays Landing
Atlantic Ocean
Pleasantville
Brigantine
Bridgeton
Millville
Atlantic City
Cumberland
Somers Point
Ventnor City
Ocean City
Cape May
Sea Isle City
Delaware
Delaware Bay
Cape May Court House
North Wildwood
Wildwood
Cape May
OnTheWorldMap.com

3.0 Methodology: Data Visualization and exploration

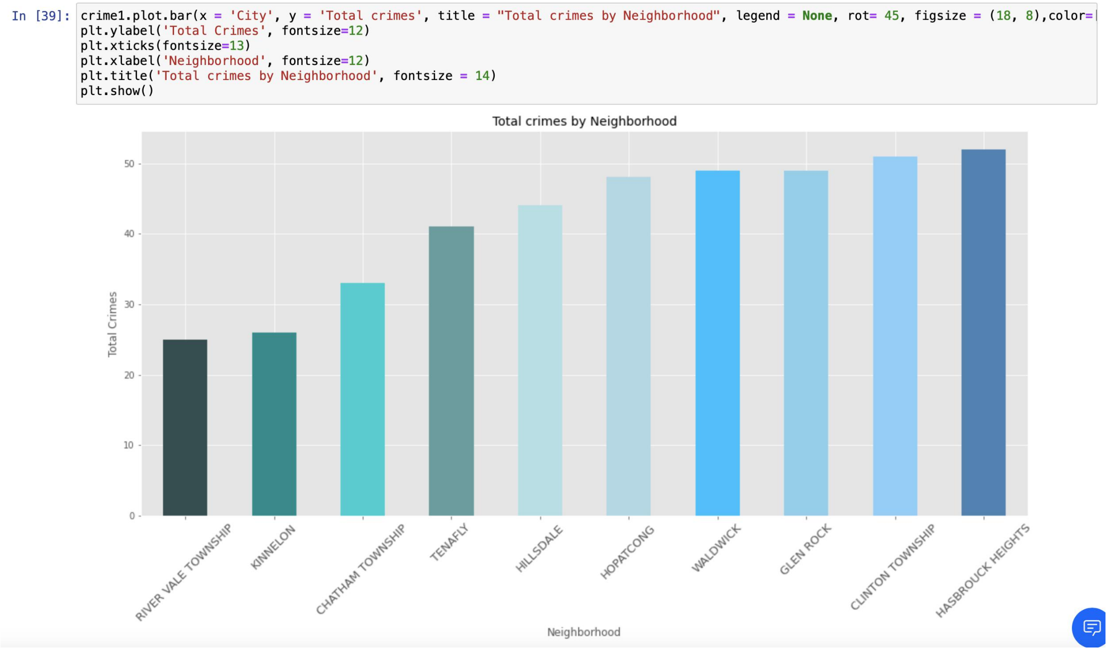
3.1 Narrowing down neighborhoods

We can use simple visualizations to examine our datasets and narrow down our options for which neighborhood we will like to set up our physical therapy facility. First, we start with the first 10 neighborhood with the most heath care facilities.



```
In [28]: health.plot.bar(x = 'City', y = 'Num_of_facility', title = "Number of Registered health facilities by Neighborhood", legend = None, rot=
         plt.ylabel('Number of Registered Health facilities', fontsize=12)
         plt.xticks(fontsize=13)
         plt.xlabel('Neighborhood', fontsize=12)
         plt.title('Number of Registered health facilities by Neighborhood', fontsize = 14)
         plt.show()
```

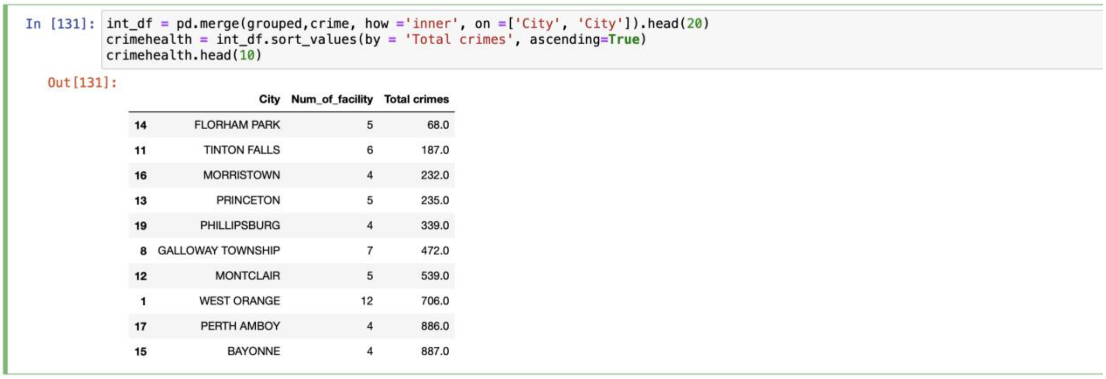
Let us take a look at the neighborhood with the least crime.



```
In [39]: crime1.plot.bar(x = 'City', y = 'Total crimes', title = "Total crimes by Neighborhood", legend = None, rot= 45, figsize = (18, 8),color=
         plt.ylabel('Total Crimes', fontsize=12)
         plt.xticks(fontsize=13)
         plt.xlabel('Neighborhood', fontsize=12)
         plt.title('Total crimes by Neighborhood', fontsize = 14)
         plt.show()
```

Unfortunately, none of the neighborhood with the most heath care facilities are not in the first 10 safest cities in New Jersey. We will only be able to avoid the 10 most dangerous cities in New Jersey.

Hence, we merge the two data frames together and select 20 cities with the most heath care facilities. This data frame is then sorted based on the total number of crimes in ascending order.

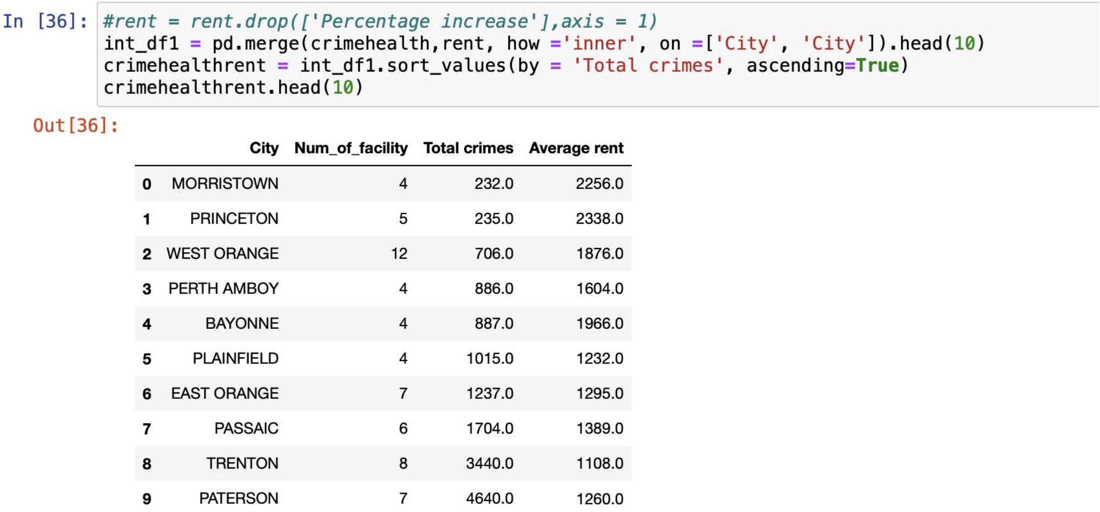


```
In [131]: int_df = pd.merge(grouped,crime, how ='inner', on =['City', 'City']).head(20)
          crimehealth = int_df.sort_values(by = 'Total crimes', ascending=True)
          crimehealth.head(10)
```

Out[131]:

	City	Num_of_facility	Total crimes
14	FLORHAM PARK	5	68.0
11	TINTON FALLS	6	187.0
16	MORRISTOWN	4	232.0
13	PRINCETON	5	235.0
19	PHILLIPSBURG	4	339.0
8	GALLOWAY TOWNSHIP	7	472.0
12	MONTCLAIR	5	539.0
1	WEST ORANGE	12	706.0
17	PERTH AMBOY	4	886.0
15	BAYONNE	4	887.0

From the data frame above, West Orange has the most heath care facilities with 706 crime cases while Florham Park is the safest with 68 crimes but has 5 heath care facilities. We will work with these 10 cities as we consider other criteria.
Since we have to consider other criteria, we merge the heath care facilities, crime rate and the average rent data to form one data frame. For the purpose of this analysis, we will be flexible with this criteria. My client requested for a neighborhood with rent with ranging from $1000 to $2000. The optimum neighborhood based on these 3 criteria is chosen.

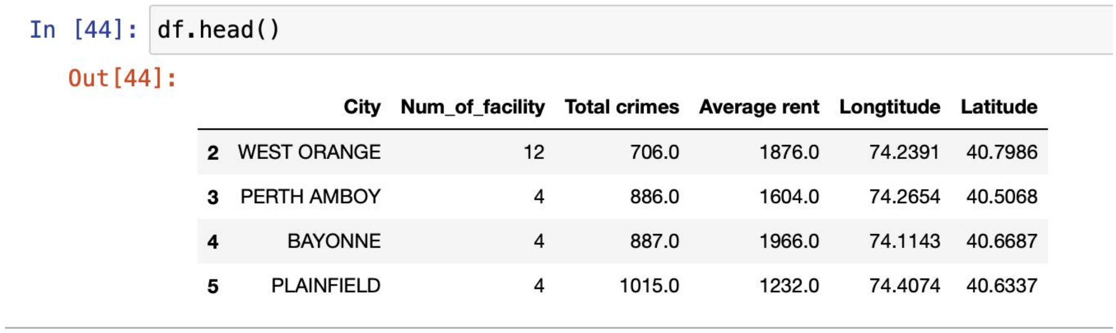


```
In [36]: #rent = rent.drop(['Percentage increase'],axis = 1)
         int_df1 = pd.merge(crimehealth,rent, how ='inner', on =['City', 'City']).head(10)
         crimehealthrent = int_df1.sort_values(by = 'Total crimes', ascending=True)
         crimehealthrent.head(10)
```

Out[36]:

	City	Num_of_facility	Total crimes	Average rent
0	MORRISTOWN	4	232.0	2256.0
1	PRINCETON	5	235.0	2338.0
2	WEST ORANGE	12	706.0	1876.0
3	PERTH AMBOY	4	886.0	1604.0
4	BAYONNE	4	887.0	1966.0
5	PLAINFIELD	4	1015.0	1232.0
6	EAST ORANGE	7	1237.0	1295.0
7	PASSAIC	6	1704.0	1389.0
8	TRENTON	8	3440.0	1108.0
9	PATERSON	7	4640.0	1260.0

Based on these criteria, we have narrowed down our optimum neighborhood to four neighborhoods: West Orange, Perth Amboy, Bayonne and Plainfield. We checked with Google and pull in coordinates. Then, we can split the Coordinates column into Latitude and Longitude and create a new data frame to get a complete picture of our neighborhoods.



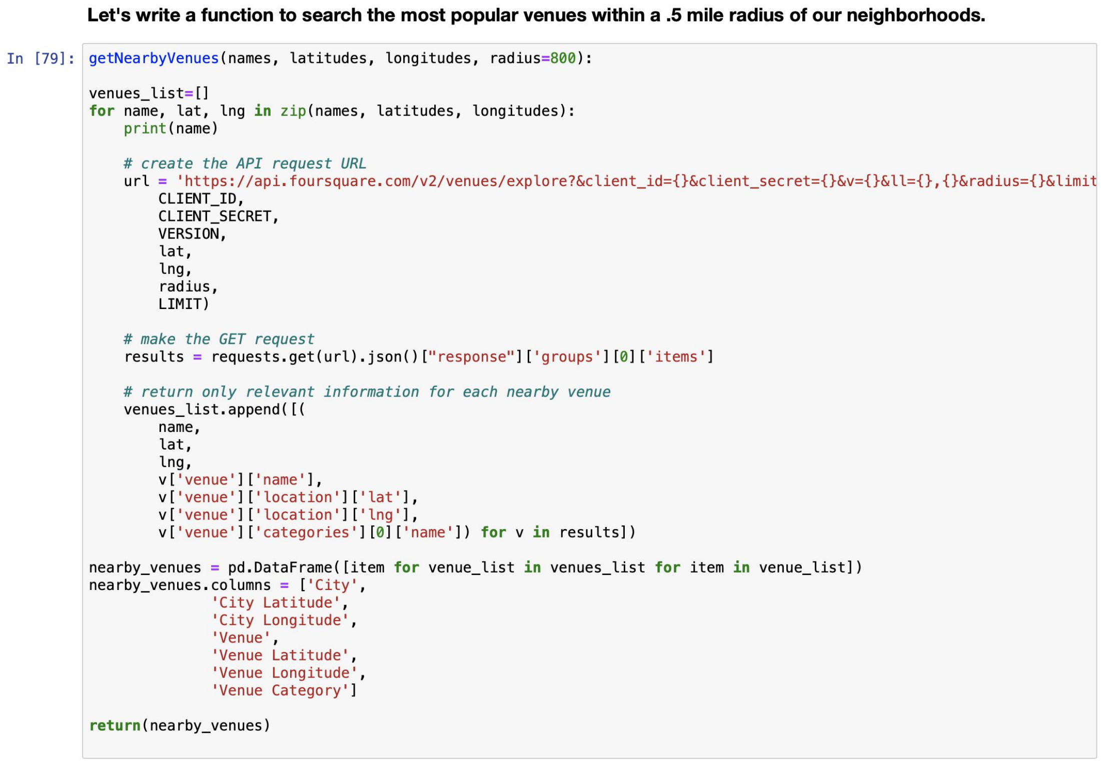
In [44]:
```
df.head()
```

Out[44]:

	City	Num_of_facility	Total crimes	Average rent	Longtitude	Latitude
2	WEST ORANGE	12	706.0	1876.0	74.2391	40.7986
3	PERTH AMBOY	4	886.0	1604.0	74.2654	40.5068
4	BAYONNE	4	887.0	1966.0	74.1143	40.6687
5	PLAINFIELD	4	1015.0	1232.0	74.4074	40.6337

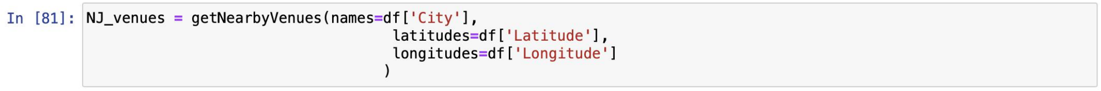
3.2 Foursquare data

Foursquare data is robust location platform that provides data for companies such as Apple and Uber. We can retrieve information about the most popular spots in each neighborhood in New Jersey using the Foursquare API. This will be an insightful indication of foot traffic for different venue types. Calling the Foursquare API returns a JSON file, which can be turned into a data frame for analysis in a python notebook.
We can start by writing a function that will search for the most popular venues within a half mile radius of our neighborhoods



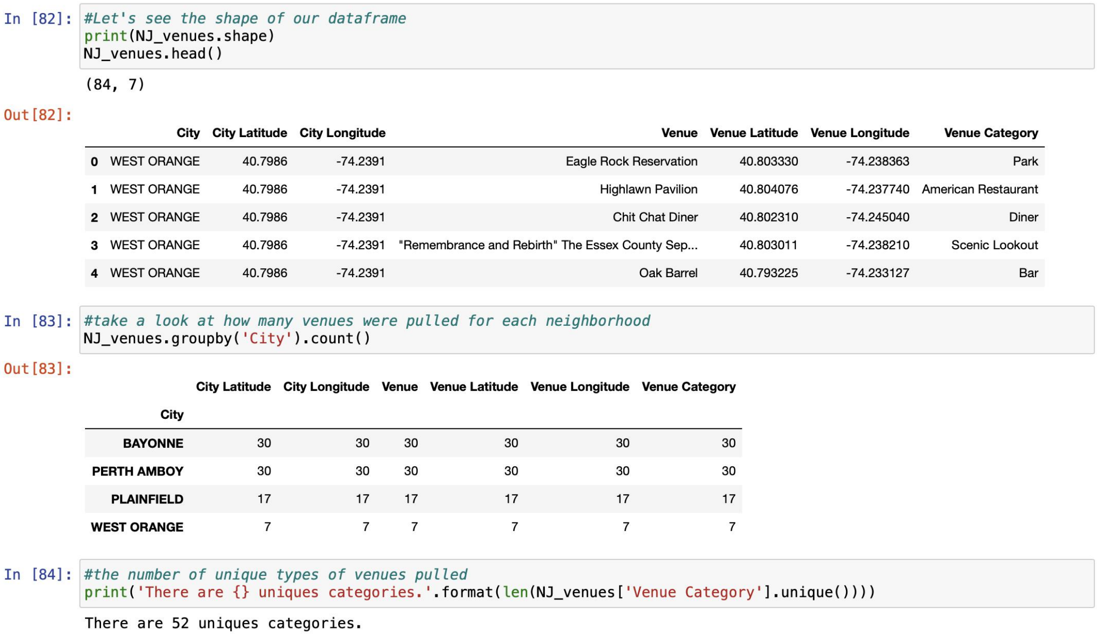
Let's write a function to search the most popular venues within a .5 mile radius of our neighborhoods.

In [79]:
```
getNearbyVenues(names, latitudes, longitudes, radius=800):

venues_list=[]
for name, lat, lng in zip(names, latitudes, longitudes):
    print(name)

    # create the API request URL
    url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit
        CLIENT_ID,
        CLIENT_SECRET,
        VERSION,
        lat,
        lng,
        radius,
        LIMIT)

    # make the GET request
    results = requests.get(url).json()["response"]['groups'][0]['items']

    # return only relevant information for each nearby venue
    venues_list.append([(
        name,
        lat,
        lng,
        v['venue']['name'],
        v['venue']['location']['lat'],
        v['venue']['location']['lng'],
        v['venue']['categories'][0]['name']) for v in results])

nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
nearby_venues.columns = ['City',
              'City Latitude',
              'City Longitude',
              'Venue',
              'Venue Latitude',
              'Venue Longitude',
              'Venue Category']

return(nearby_venues)
```



In [81]:
```
NJ_venues = getNearbyVenues(names=df['City'],
                                   latitudes=df['Latitude'],
                                   longitudes=df['Longitude']
                                  )
```

We get a data frame of 84 entries, having 30 venues for 2 Bayonne and Perth Amboy cities and 17 in Plainfield while 7 in west orange and 52 unique venue categories:



In [82]:
```
#Let's see the shape of our dataframe
print(NJ_venues.shape)
NJ_venues.head()
```

```
(84, 7)
```

Out[82]:

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	WEST ORANGE	40.7986	-74.2391	Eagle Rock Reservation	40.803330	-74.238363	Park
1	WEST ORANGE	40.7986	-74.2391	Highlawn Pavilion	40.804076	-74.237740	American Restaurant
2	WEST ORANGE	40.7986	-74.2391	Chit Chat Diner	40.802310	-74.245040	Diner
3	WEST ORANGE	40.7986	-74.2391	"Remembrance and Rebirth" The Essex County Sep...	40.803011	-74.238210	Scenic Lookout
4	WEST ORANGE	40.7986	-74.2391	Oak Barrel	40.793225	-74.233127	Bar

In [83]:
```
#take a look at how many venues were pulled for each neighborhood
NJ_venues.groupby('City').count()
```

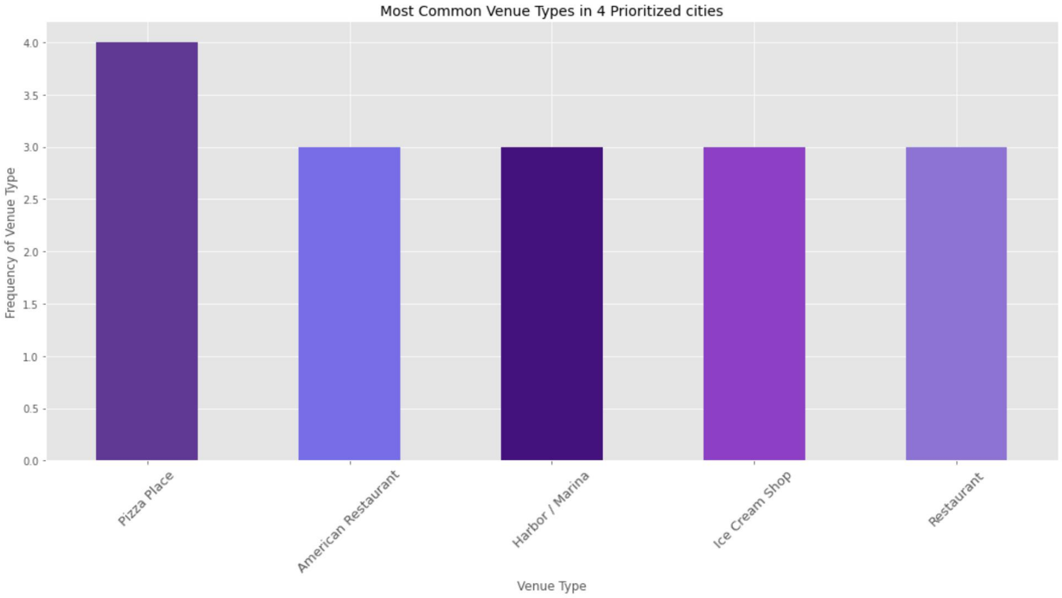
Out[83]:

	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
City						
BAYONNE	30	30	30	30	30	30
PERTH AMBOY	30	30	30	30	30	30
PLAINFIELD	17	17	17	17	17	17
WEST ORANGE	7	7	7	7	7	7

In [84]:
```
#the number of unique types of venues pulled
print('There are {} uniques categories.'.format(len(NJ_venues['Venue Category'].unique())))
```

```
There are 52 uniques categories.
```

We have a graphic representation of the most popular venue categories across all 4 cities:

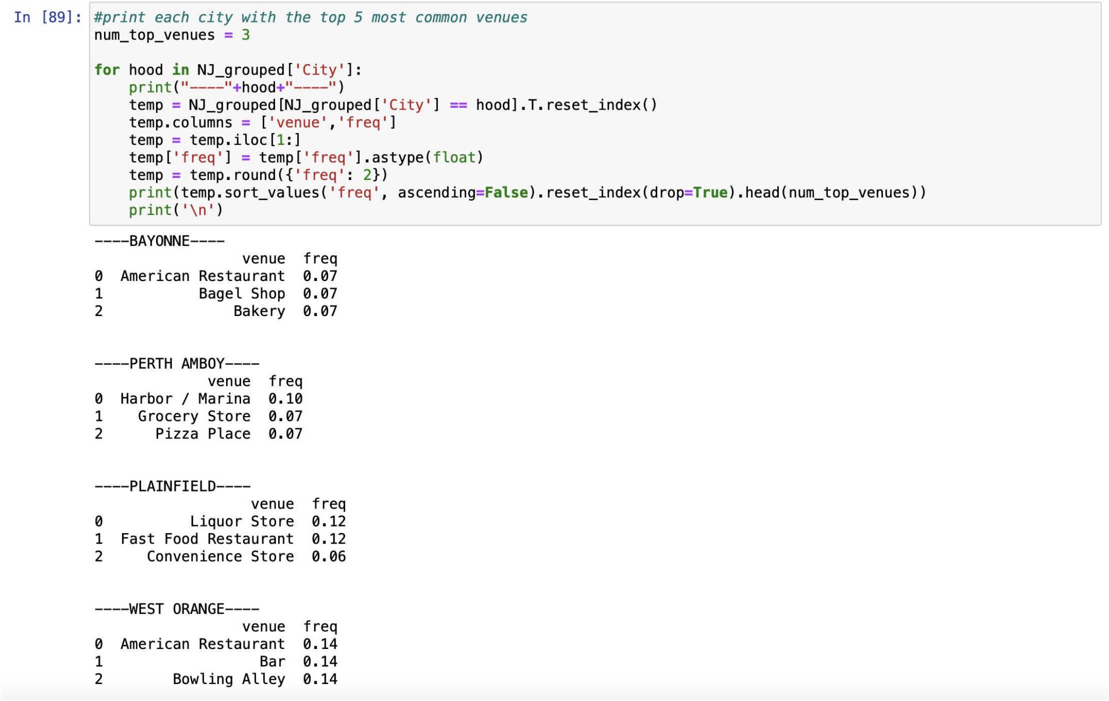


It looks like Pizza place, restaurants, Harbor/Marina and ice cream shops are the most common popular venue type. We can see that physical therapy or Gym are not the most popular type of venue in these cities over all, but they may be more popular in some cities than others.

We dig further into each of the cities to see the most popular types of venues for each city. To do this, we will take the following steps:

- Create a data frame of venue categories with pandas one hot encoding

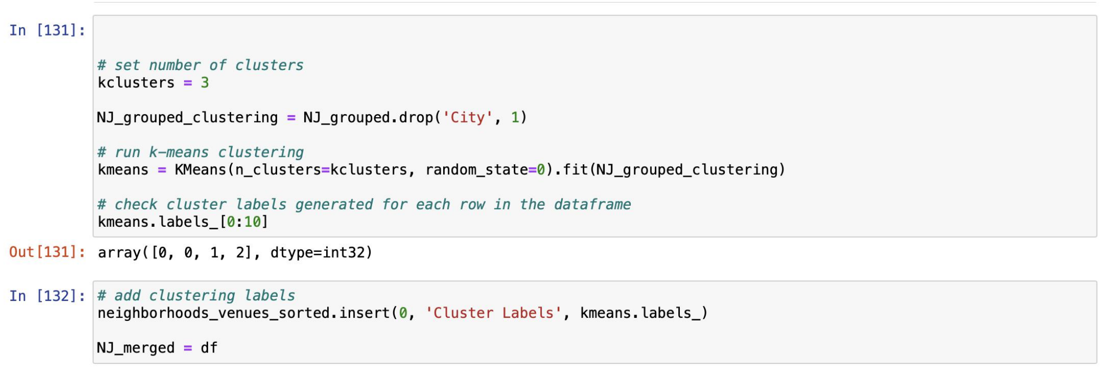
- Use pandas groupby to get the mean of the one-hot encoded venue categories
- Transpose the data frame and arrange in descending order



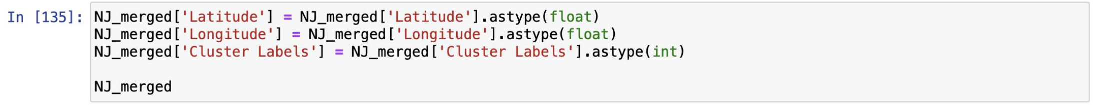
```
In [89]: #print each city with the top 5 most common venues
         num_top_venues = 3

         for hood in NJ_grouped['City']:
             print("----"+hood+"----")
             temp = NJ_grouped[NJ_grouped['City'] == hood].T.reset_index()
             temp.columns = ['venue','freq']
             temp = temp.iloc[1:]
             temp['freq'] = temp['freq'].astype(float)
             temp = temp.round({'freq': 2})
             print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
             print('\n')

         ----BAYONNE----
                         venue  freq
         0  American Restaurant  0.07
         1           Bagel Shop  0.07
         2               Bakery  0.07


         ----PERTH AMBOY----
                     venue  freq
         0  Harbor / Marina  0.10
         1    Grocery Store  0.07
         2      Pizza Place  0.07


         ----PLAINFIELD----
                          venue  freq
         0         Liquor Store  0.12
         1  Fast Food Restaurant  0.12
         2     Convenience Store  0.06


         ----WEST ORANGE----
                         venue  freq
         0  American Restaurant  0.14
         1                  Bar  0.14
         2        Bowling Alley  0.14
```

3.3 Neighborhood Clustering

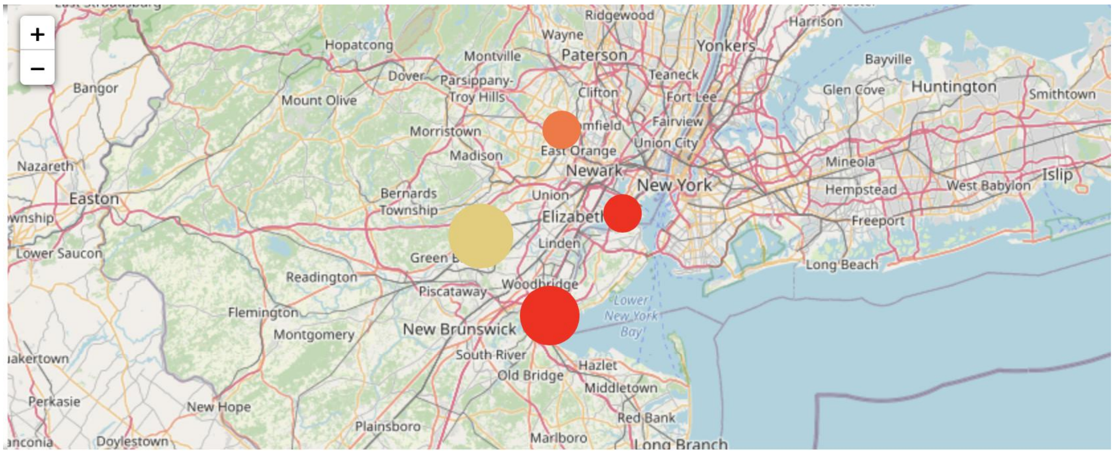
Finally, we can cluster our 4 cities based on their popular venue categories. This will help us get a feel for which city are like each other based on the venues people like to visit in each one. We use K-Means clustering, detailed in the code below, to group our cities into 3 clusters.



```
In [131]:


           # set number of clusters
           kclusters = 3

           NJ_grouped_clustering = NJ_grouped.drop('City', 1)

           # run k-means clustering
           kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(NJ_grouped_clustering)

           # check cluster labels generated for each row in the dataframe
           kmeans.labels_[0:10]

Out[131]:  array([0, 0, 1, 2], dtype=int32)

In [132]:  # add clustering labels
           neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

           NJ_merged = df
```



```
In [135]:  NJ_merged['Latitude'] = NJ_merged['Latitude'].astype(float)
           NJ_merged['Longitude'] = NJ_merged['Longitude'].astype(float)
           NJ_merged['Cluster Labels'] = NJ_merged['Cluster Labels'].astype(int)

           NJ_merged
```



It looks like Bayonne and Perth Amboy fit into one cluster, plainfield fit into the second cluster while west orange stands alone as a third.

4.0 Results and Discussion

We have pulled data on crime rate, rent prices and the number of health care facilities for every city in New Jersey and use this information to narrow down our city options to 4 cities. Our analysis has informed us that:

- Pizza place, restaurants, Harbor/Marina and ice cream shops are the most common popular venue type in our 4 preferred cities.

Clustering cities based on their most popular venues grouped Bayonne and Perth Amboy into one cluster, plainfield and west orange stand as an independent cluster.

- From the rent dataset, plainfield has the cheapest rent while from the crime dataset, west orange has the least number of crimes.

Based on this analysis, the optimal location is west orange since it has the least number of crimes and the highest number of health care facilities around.
A major drawback of this analysis is that the clustering was based on Foursquare' s data for popular venues. There are plenty other ways to assess popularity of cities and the spots inside them, venue popularity is just one of them.

5.0 Conclusion

In this report, we did an end to end data science project using python libraries to manipulate datasets. We use the Foursquare API to explore the cities in New Jersey and Folium map to cluster and segment cities. This analytical tool will help my friend to make the right decision on where to open her physical therapy facility in New Jersey.