Battle of Neighborhoods for setting up

Diagnostic and Treatment center in New York State

Gauri Priya Saran
February 07, 2020

Introduction:

❖ Problem Background:

In New York State the health care landscape is evolving. It is marked in particular by an overall shift toward the provision of primary care and medical care services in neighborhood settings. This movement has been supported by federal and state initiatives aimed at improving overall population health and outcomes of care, and creating more efficient delivery by promoting value throughout the health care delivery system.
In New York State, someone is looking to open a Diagnostic and Treatment Center. Which location should we recommend that they open it?

❖ Problem Description:

This project is an initiative to help provide valuable information to entrepreneurs who want to start a venture in the health care industry by setting up a **Diagnostic and Treatment center** in the borough of Brooklyn, county-level administrative division coterminous with Kings County, New York State. (https://en.wikipedia.org/wiki/Boroughs_of_New_York_City) It is the New York City's most populous borough, with an estimated 2,504,700 residents in 2010.

What is a Diagnostic and Treatment center?

Under the statutory authority of Article 28, Section 3401 of the Public Health Law (PHL), and Title 10 of the New York Codes of Rules and Regulations (NYCRR), Section 405, providers who do not qualify or choose not to operate as private practices may be licensed by the State to operate free-standing clinics are also known as **Diagnostic and Treatment Centers** (Clinics). These free-standing clinics are separately-owned and are not operated by a hospital. By contrast, clinics that are owned and operated by a hospital are known as Hospital Extension Clinics. (https://profiles.health.ny.gov/clinic/)

To find a great location for a Diagnostic and Treatment Center requires intensive research, knowledge gathering of health care sector, and appropriate dataset. A great location in the business plan helps in preparing more accurate earnings estimate for the business and also in fundraising.

The health care industry is a highly controlled sector, with regulation from the administration, hence business plan should address particular issues and help in resolving any queries of prospective investors including **location accessibility, facilities nearby**, and permissibility. A number of businesses are linked with hospitals and health care, from treatment providers and

pharmacy services to eateries, residential areas for health care staff and so on and hence it is critical that the health care business plan deals with the locational issues appropriately.

Data:

❖ Data Sources:

To figure out the best location to set up a Diagnostic and Treatment Center, logically, we need 2 things:

1. Its geographical coordinates (latitude and longitude) to find out where exactly it is located.
2. Population and facilities of the neighborhood where the facility is located.

The data from different health care facilities with their geographical coordinates in the borough of Brooklyn, New York will be used for analyzing different neighborhoods that can be conducive for establishing a new Diagnostic and Treatment Center.

The hospital **data** for New York State **with their geographical coordinates** is available on the website (https://health.data.ny.gov/Health/Health-Facility-General-Information/vn5v-hh5r/data).

The data related to existing diagnostic and treatment centers to is also verifiable at https://health.data.ny.gov/Health/HFIS-Diagnostic-and-Treatment-Centers-General-Outp/tx6m-mpjb

Data Format:

Data is available in the form of **'.csv' files**, there are 36 columns in the dataset of interest, and some of the column names are following:

- Facility ID
- Facility Name
- Short Description
- Description
- Facility Open Date
- Facility Address 1
- .....
- Ownership Type
- Facility Latitude
- Facility Longitude
- Facility Location

Data Cleaning:

Data source was found to include errors and missing values – hence data cleaning was used to address these anomalies. Not cleaning data can lead to problems such as linking errors, model

misspecification, errors in parameter estimation and incorrect analysis leading users to draw false conclusions.

After the data was imported into a Pandas data frame, rows having **Nan values** in columns of interest, specifically **Facility Latitude** and **Longitude**, were dropped.

How Data will be used?

The data is imported into a Pandas dataframe and has been cleaned to perform analysis using KMeans Algorithm to form clusters of different health care facilities that are Diagnostic and treatment centers. Foursquare API was used to gain information about the surrounding neighborhoods. The information gained from surrounding neighborhoods in these clusters has been used to make important conclusions regarding the set up of a new diagnostic and treatment center.

Methodology and Exploratory Data Analysis:

This section describes exploratory data analysis, inferential statistical testing, and machine learning parts of the project. Briefly, the data is imported into a Pandas dataframe and has been cleaned to perform analysis using KMeans Algorithm to form clusters of different health care facilities that are Diagnostic and treatment centers. **Foursquare API** was used to gain information about the surrounding neighborhoods. The information gained from surrounding neighborhoods in these clusters has been used to make important conclusions regarding the set up of a new diagnostic and treatment center.

Based on data, **is there a business opportunity** for setting up a health facility in the area of interest? A **pivot table** using Pandas crosstab method (‘**Hospital_Year’** and ‘**Ownership_Type’**) was constructed **to ascertain the age and ownership of hospitals**. From the below pivot table, it can be observed that maximum health care facilities were built before 1990 and they were mainly Not for Profit Organizations, with second most made by LLC, and then by county, municipality and state. The health care facilities built after 1979 are majorly set up by Not for Profit Organizations, LLC, and business corporations. This indicates **opportunity for setting up Diagnostic and Treatment Center**.

Ownership Type	(unknown)	Business Corporation	County	Individual	LLC	LLP	Municipality	Not for Profit Corporation	Partnership	Public Benefit Corporation	Stat
Hospital_Year											
1901	0	10	24	1	69	1	14	245	0	4	
1960	0	0	0	0	0	0	0	0	1	0	
1970	0	0	0	1	1	0	0	1	0	0	
1977	0	0	0	0	0	0	0	1	0	0	
1978	0	0	0	0	0	0	0	2	0	0	
1979	0	12	44	0	39	0	3	66	0	0	
1980	0	15	0	0	48	0	2	26	1	0	
1981	0	0	0	0	4	0	0	7	0	0	
1982	0	1	1	0	2	0	1	6	0	1	
1983	0	1	0	0	3	0	0	3	0	0	
1984	0	1	0	0	1	0	0	8	0	0	
1985	0	4	0	0	4	0	0	6	0	0	

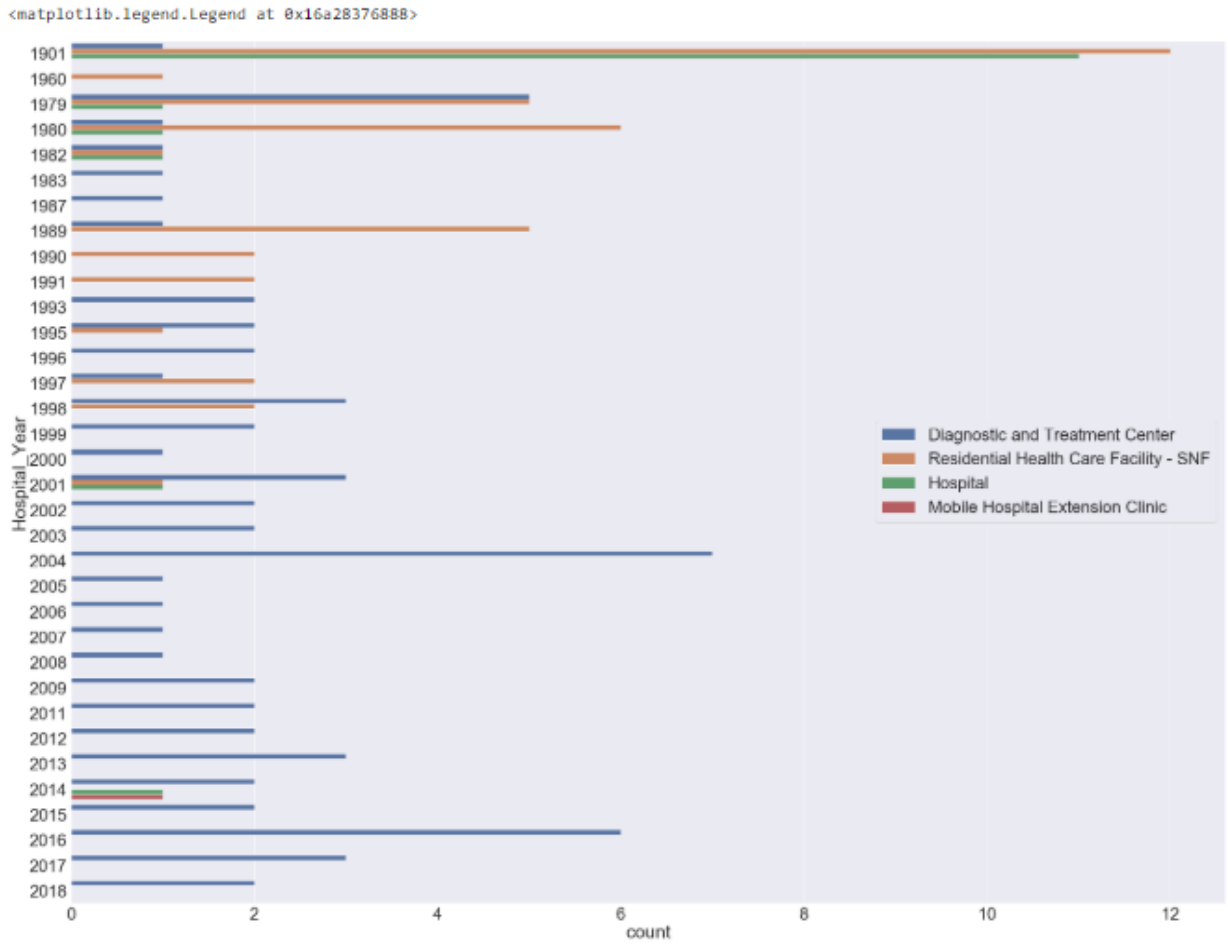
Based on location data, **which borough is most popular for setting up a health facility**? For business purpose it is important to know **which borough in New York has maximum number of health care facilities**. It is normal to assume that such an area will be a natural choice for the patients to visit for diagnostic and treatment. It was an easy task to figure this out **by grouping the data** by **'Facility City'** and **'Facility Name'** and then calling **idxmax** over the **DataFrameGroupBy** object, which returns the index of first occurrence of maximum over requested axis.

From the data set available, the maximum number of health care facilities in New York State are found in Brooklyn city.

Analyzing hospital neighborhoods in Brooklyn

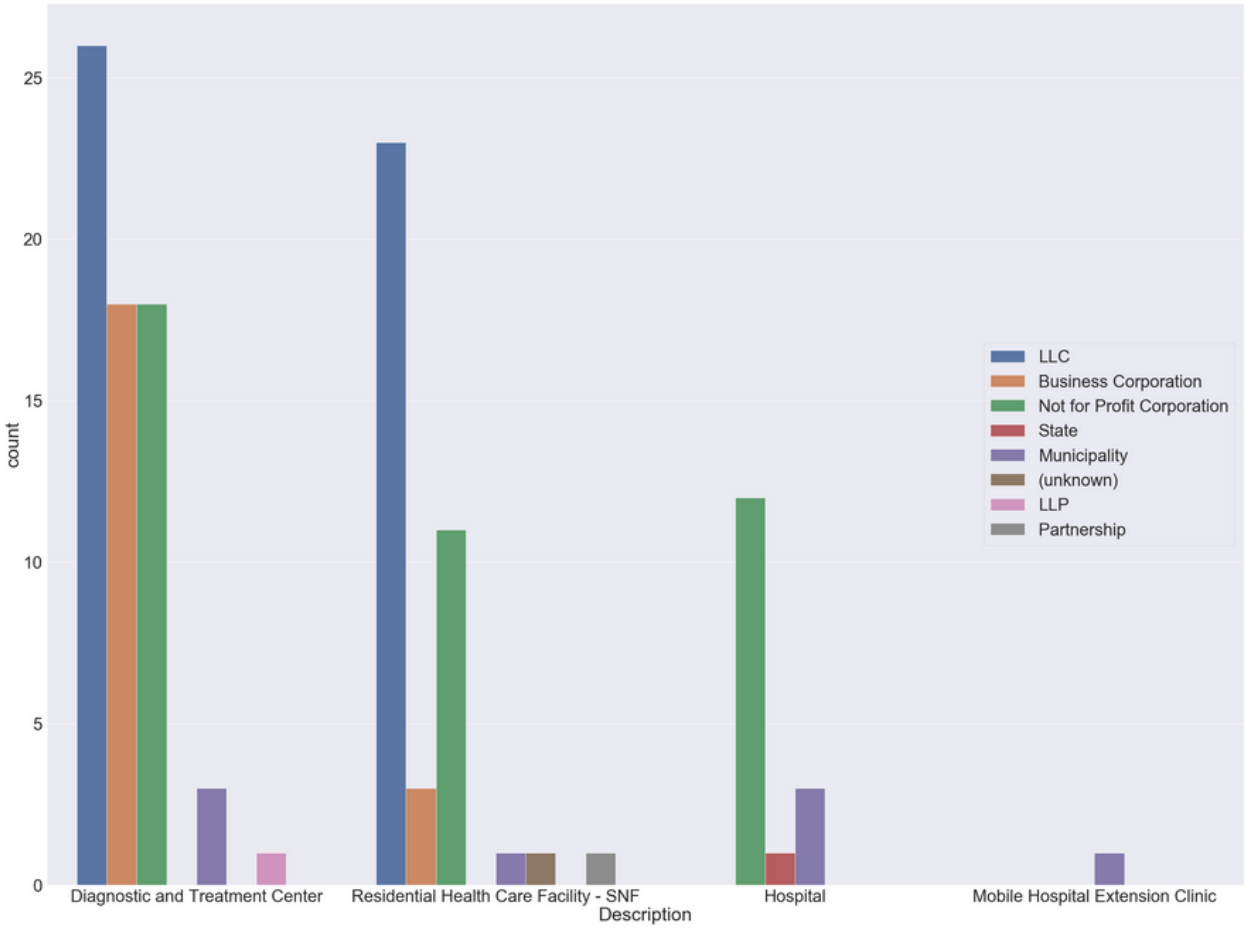
As mentioned above, as per the dataset the maximum number of health care facilities in New York State are in Brooklyn city.

We have found the area, now we must find **which is the most popular type of health facility** from business point of view? For this we create a count plot of different kinds of health care facilities across different years from 1901 to 2019.



Result of Analysis: It can be observed from the above plot that in the recent years, among all the health care facilities, a lot of diagnostic and treatment centers have been established since diagnostic and treatment centers are highly profitable source of revenue generation.

It is also important to know about the **type of ownerships** setting up Diagnostic and Treatment Centers in the area. Are the Diagnostic and Treatment Centers being set up primarily as State run enterprises or are private businesses also setting up? For this purpose we create a histogram of 'Ownership Type'.

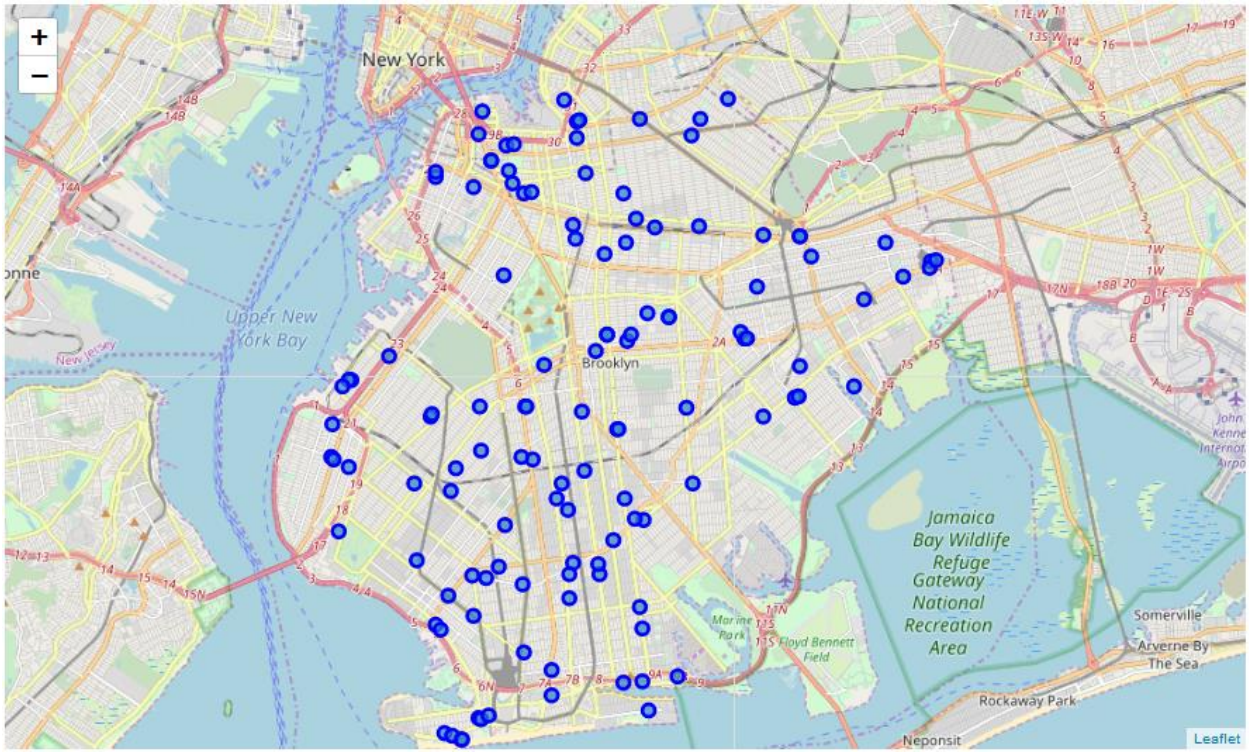


Result of Analysis: It is observed in the above plot that most of the business corporations and the LLC have focused their efforts in developing more diagnostic and treatment centers along with some development of Residential health care facility -SNF as well.

Next we need to explore the neighborhoods around the existing diagnostic centers and segment them by using Foursquare API. For this we first need a map of Brooklyn and need to have a visual understanding of the location.

The latitude and longitude coordinates of Brooklyn are fetched using Nominatim geocoder for OpenStreetMap data (https://wiki.openstreetmap.org/wiki/Nominatim). The fetched latitude and longitude coordinates data is in object format, this is converted by us for our use into float64, and map

of Brooklyn is created using latitude and longitude values and folium https://python-visualization.github.io/folium/).



Using Foursquare API to explore the neighborhoods:

Foursquare API is used to explore the neighborhoods around the diagnostic centers and segment them.

To analyze the surrounding venues in the neighborhood of different diagnostic centers, we fetch the top 100 venues that are within a radius of 500 meters of the Neighborhoods of selected hospital. For this project we limit the number of venues returned by Foursquare API to 100.

We check how many venues were returned for each Hospital in the Neighborhood and group the venues by the hospital. Based on this information we get the hospital name that has maximum number of venue categories in radius of 500m, this is 'Cobble Hill Health Center, Inc.'.

Then we find out how many unique categories can be curated from all the returned venues and analyze each neighborhood.

The venue categories are one-hot encoded. The data frame with one-hot encoded venue categories is in following format.

	Facility Zip Code	Hospital	Neighborhoods	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	...	Winery	Wings Joint	Women's Store	Yoga Studio
0	11209	Bay Ridge Surgi-Center	370 Bay Ridge Parkway	0	0	0	0	...	0	0	0	0
1	11209	Bay Ridge Surgi-Center	370 Bay Ridge Parkway	0	0	0	0	...	0	0	0	0
2	11209	Bay Ridge Surgi-Center	370 Bay Ridge Parkway	0	0	0	0	...	0	0	0	0

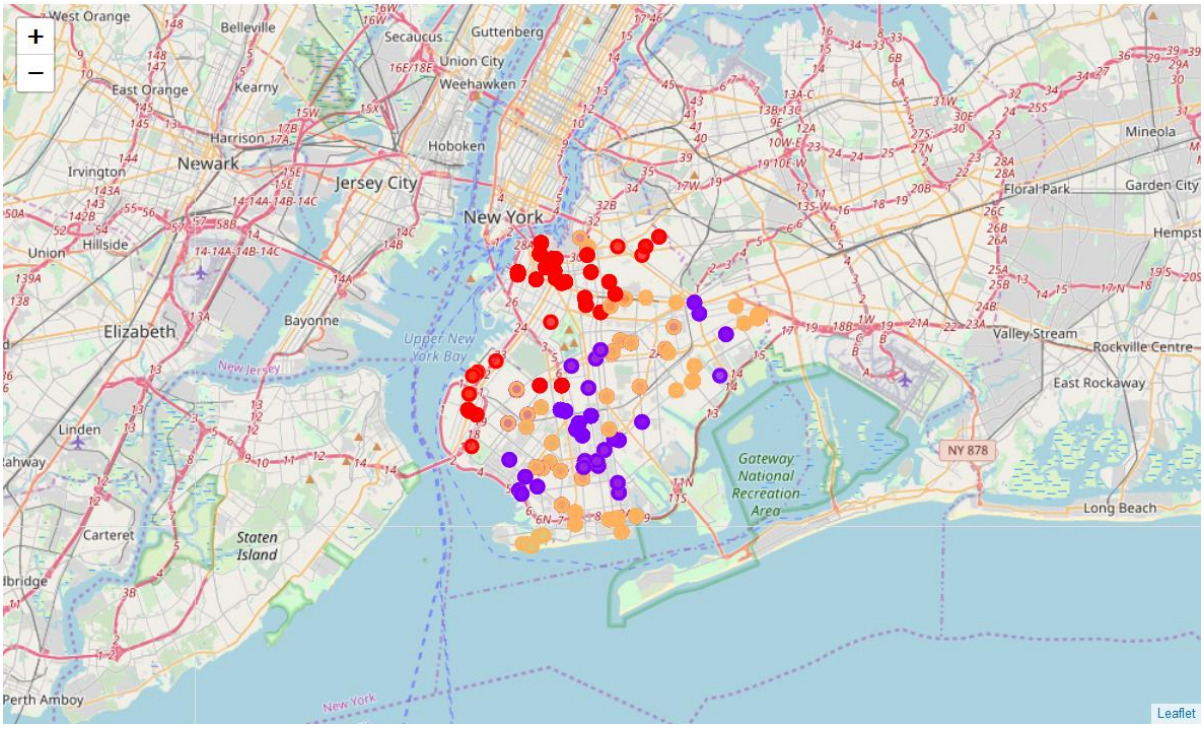
Next, we group the rows by the facility zip code, hospital and neighborhoods and take the mean of the frequency of occurrence of each category so that they can be sorted to give us the top ten most frequently occurring venues in the neighborhoods of the hospital.

Thereafter we fetch the top ten most frequently occurring venues in the neighborhoods of the hospitals and cluster the neighborhoods.

K-Means Algorithm is used to cluster the health care facilities into different clusters. Based on the surrounding venues in each cluster, a plan will be proposed for the development of a new diagnostic and treatment center.

For this we set number of clusters to five, run K-Means clustering, add clustering labels, and merge grouped data frame with hospital data frame and add latitude/longitude, city, state and description for each neighborhood.

A Cluster Map of Brooklyn with the hospitals and their clusters superimposed on it is created.



Results:

1.1 Cluster 1 consist of surrounding venues such as Pizza place, restaurants, parks, grocery stores. There is also residential building and apartment / condos observed in the neighborhoods in this cluster. This might prove convenient for the staff who will work to be able to live close to work or those living in the area to visit the diagnostic center. The cluster 1 is one of the biggest clusters and consist of all the different types of health care facilities such hospitals, residential health care facility and diagnostic centers. There is not much information on the transport connectivity. Therefore, we might need more information to further explore the neighborhoods in this cluster for establishing a new diagnostic and treatment center.

1.2 Cluster 2 neighborhoods have surrounding venues such as pharmacy, metro stations, bus stops, grocery stores, parks, banks, pizza places and other restaurants. This cluster has a high number of already established diagnostic centers. This might pose significant competition for a new business to start a new diagnostic center. However, this cluster also has connectivity for transit and have good surrounding venues that may prove beneficial for facilitating a new business in this cluster.

1.3 Cluster 3 has the lowest number of health facilities in it. Though there are restaurants and pharmacy in the surrounding neighborhoods, there is not much information on the transit connectivity. Therefore, this cluster might not be a good choice based on Foursquare API information alone.

1.4 Cluster 4 seems to be closer to the beach with pharmacy and super-market and Deli in surrounding areas. The cluster mainly consists of residential health care facilities with few diagnostic and treatment centers. The cluster seems to be more isolated on map as well. It might be convenient for those who live in the area to visit the diagnostic centers established in this cluster. But based on Foursquare API information alone, this cluster will not be the wisest choice for establishing a new diagnostic center as there is no information on transit connectivity in this cluster and other residential areas surrounding it.

1.5 Cluster 5 neighborhoods has the highest number of health care facilities in it. The surrounding venues in the neighborhoods mainly consist of cafe, restaurants, banks, pharmacy, bus stations, and some residential buildings too. There could be significant competition posed by other diagnostic centers in this cluster but this cluster has the most promising surroundings to establish a new diagnostic and treatment center.

Discussion:

Overall we can see that cluster 5 seems to be a good choice because of available facilities nearby such as pharmacy, banks, restaurants, bus stations and residential buildings. Cluster 2 being well connected with metro and bus stops seems to be next best choice followed by cluster 1 which has many residential accommodation buildings in the neighborhood seems to be the next promising avenues for further exploration of the neighborhoods based on the Foursquare API information. However, we can not get complete information by using Foursquare API alone for making further conclusions on the best neighborhoods for establishing a new business. If we can further combine the data with data from transit companies and residential data for the city in different neighborhoods, then we might be able to make fair conclusions.

Conclusion:

Based on the data available from the foursquare API, we conclude that cluster 5 is the most promising choice for setting up a new Diagnostic and treatment center followed by cluster 2 and cluster 1. However, to make fair conclusions we will need more transit and residential data in the neighborhoods.