USING FOURSQUARE’s DATA IN TIMES OF COVID 19

Roberto Arriaga Omacell

August, 2020

1. Introduction

1.1 Background

During 2020, COVID 19 epidemic has become a global concern, killing thousands, evidencing deficiencies in several healthcare systems, lack of political leadership, and radically changing our way of life.

Many countries have implemented different measures to try to stop contagion, mainly quarantine and social distancing. Isolation had its effects on supply chain management, and communities are beginning to depend more on local resources.

As this can be considered a war against COVID 19, and because the life of people is in the middle, access to healthcare venues, such as hospitals and pharmacies, is critical for any population.

During the first stages of the pandemic, in the US, the city of New York was one of the most affected, and have been working hard to get things under control.

1.2 Problem

Datascience has been about using information to answer specific questions, so for this project: Can we use data from FOURSUARE (FS), a platform mostly focused on categories related to entertainment, leisure and shopping, to evaluate if a neighborhood has low or high access to healthcare venues?

1.3 Interest

Besides the people of New York City, especially those living in Manhattan, and under quarantine, the possibility of using non-conventional sources to evaluate the accessibility to resources can be of high value in a near future.

2. Data

2.1 Data sources

This project will use information provided by FOURSQUARE regarding venues with tags like '**hospital**' (both, public and private) and ‘**pharmacy**’. After an initial exploration of the site, tags like 'clinic' although related to health, were mostly associated with cosmetic and even veterinary (animals), so will be out of this scope, so just the 2 tags defined earlier will be used.

The focus will be New York City, USA; and using Manhattan for the first sample of neighborhoods (40).
Location information for the neighborhoods will be obtained from the NYU Spatial Data Repository (2014).

2.2 Data cleaning

From the downloaded json file, the following values were scraped: Borough, Neighborhood, Latitude and Longitude. These were assigned to a dataframe called "*neighborhoods*". From this, a slice was extracted with: "Borough" = "Manhattan", and the new dataframe was called "***manhattan_data***".

3. Methodology

3.1 Calling the FOURSQUARE API

With the location information in ***manhattan_data***, FOURSQUARE was queried for venues using categories 'hospital' and 'pharmacy', radius = 500, limit = 100.
Each category was queried separated, so two json files were retrieved from FS: "results" for 'hospital' and "results1" for 'pharmacy'.
Data regarding venue, location and distance were scrapped from both json files and added to a dataframe "***manhattan_venues***" by way of a defined function, along with the name and location of neighborhood.

3.2 Exploration

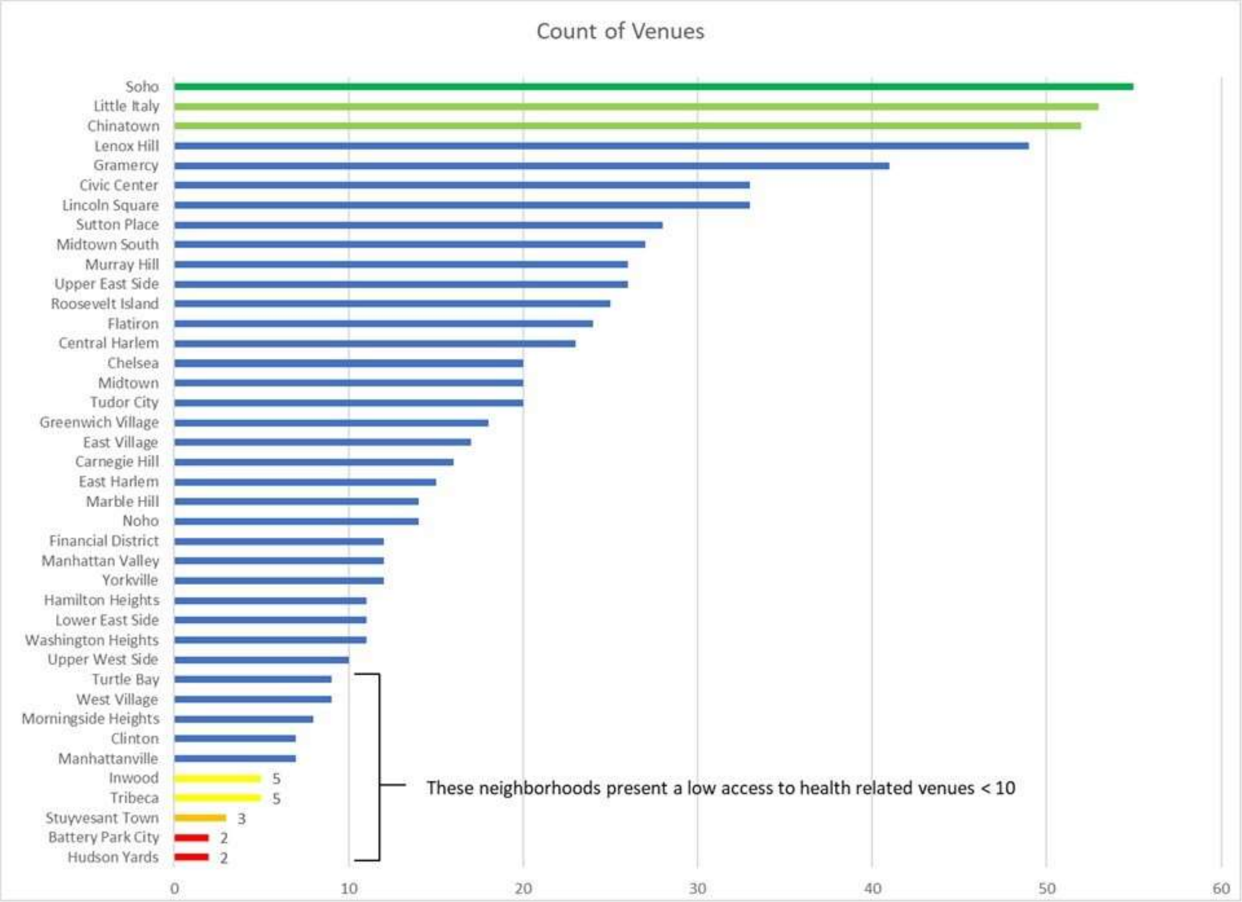
The venue data was grouped by their respective neighborhoods, counted and estimated frequency.
The top 5 venues by neighborhood by frequency were displayed.

3.3 Clustering

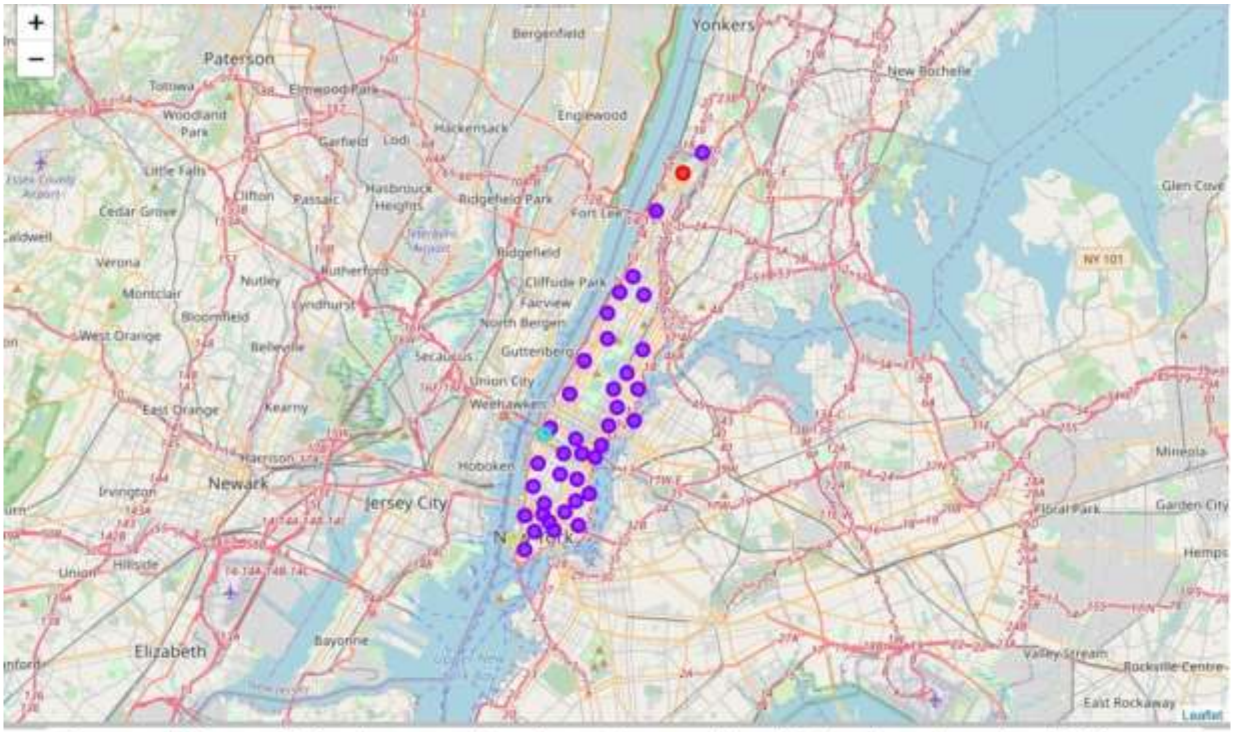
Initially the chosen algorithm for clustering was to be DBSCAN, as this had more opportunities of identifying outliers, but as the scope was reduced to just the Manhattan area and its 40 neighborhoods, the possibility of outliers is not significant, so K-means algorithm was selected instead (not only it's calculation is fastest, but also have been used successfully in previous laboratories).
A value of k = 4 was used (no optimal 4 was used as we are just interested in a general idea of how "accessible" each neighborhood is).
After the clustering, the cluster labels were added to the original dataframe and mapped.
Each of the 4 clusters were later analyzed.

4. Results

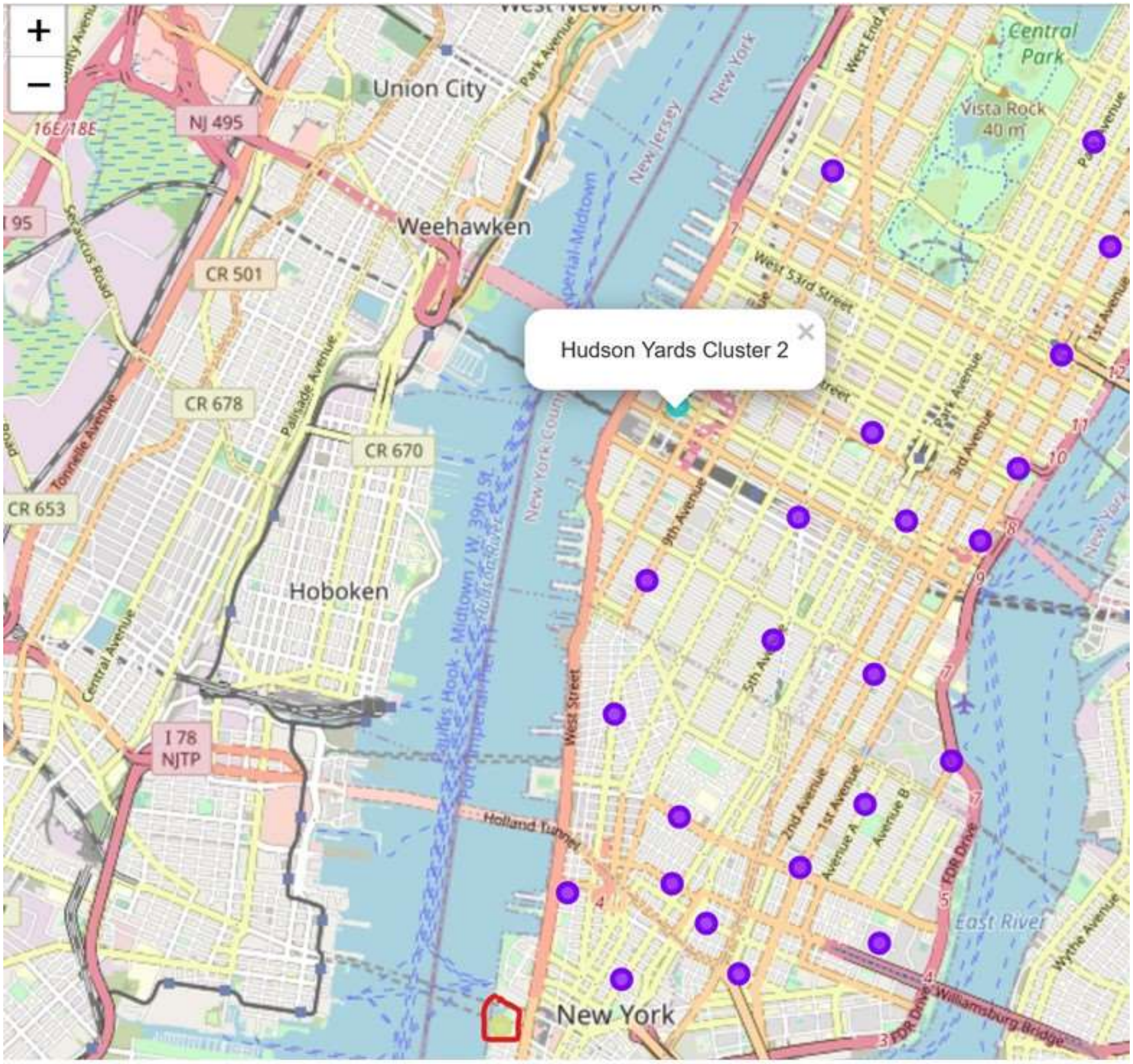
From the analysis resulting of counting the number of venues per neighborhood, we get an idea that Soho, Little Italy and Chinatown are the neighborhoods with the highest level of ‘contact’ with healthcare venues, with more than 50.



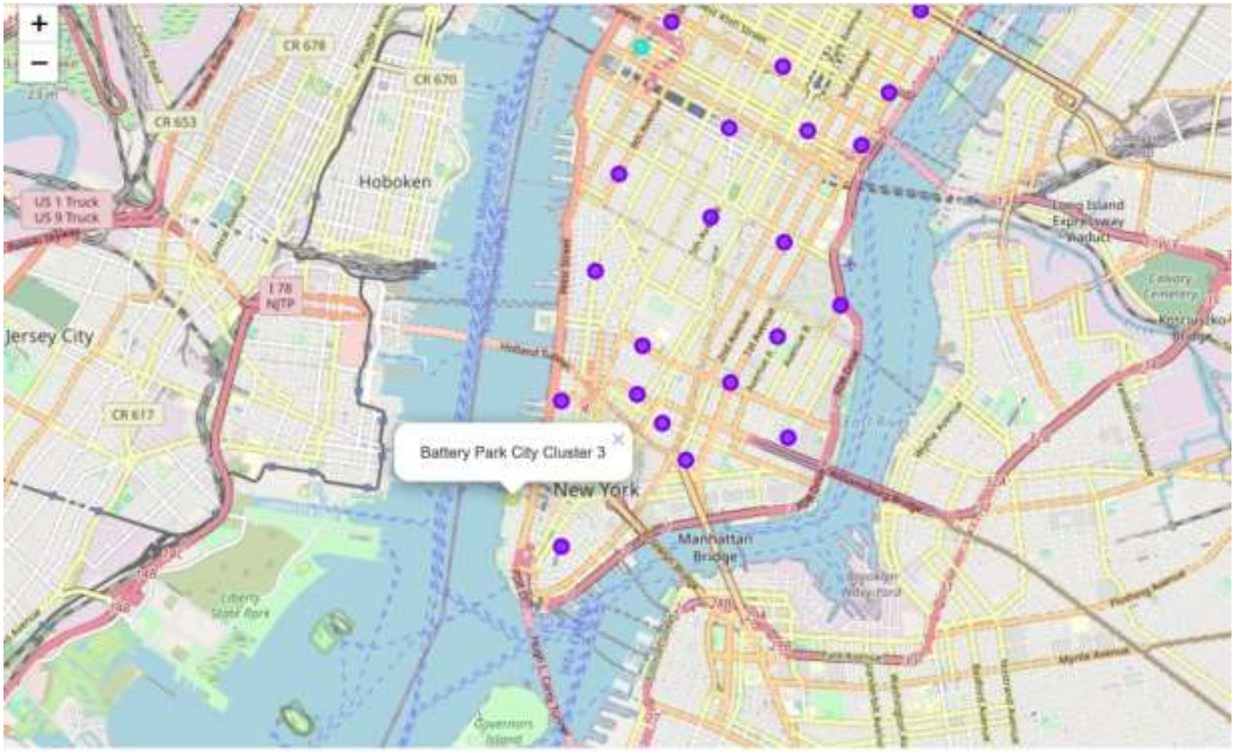
On the other side, there are 2 neighborhoods with the lowest amount: Battery Park City, and Hudson Yards. When we find the frequency tables for these two, the picture is even grimmer: in Battery Park City one of the venues is a pharmacy and the other is for animal health (veterinary hospital). For Hudson Yards, one is a pharmacy, the other a “*hospitality*” venue. Stuyvesant Town doesn’t do better, but at least is 1 public hospital and 2 veterinary hospitals.



General Map of Manhattan showing clustered neighborhoods using K-Means



Detail of cluster 2 Hudson Yards,
with Battery Park highlighted in the south



Detail of cluster 3: Battery Park City

As seen in the maps, K-Means did an interesting job by clustering both Battery Park City and Hudson Yards in their own clusters, so a more in-depth analysis can be performed. Visual information tell us that these are neighborhoods on "coastal" areas, and towards South-West. Stuyvesant Town discussed early is also coastal and in the south part.

Lastly Cluster 0 is Inwood, with 5 venues, mostly pharmacies. The interesting part of this cluster is that is tied with Tribeca with 5 venues, but one of the venues in Tribeca is an animal hospital. For future exploration, other values of k should be tested.

5. Discussion

It can be seen that some of the venues analyzed are related to animal hospitals (veterinaries), there are some 'hospitalities', and a few 'other' categories such as cafes and offices. This may be product of stating initial query limit of 100 venues [100 venues*40 neighborhoods*2 categories = 8000 possible, and near 519 were found, so the depth of the query was very high]. Instead of eliminating them, each case was analyzed in context. In this way the real situation of our "data points of interest" was better understood.

From the results, Soho, Little Italy and Chinatown are neighborhoods with high access to hospitals and pharmacies (>50 venues) while Battery Park City and Hudson Yards are the neighborhoods with the least access to healthcare facilities (each have 2 venues, but actually just a pharmacy each, as the other venue is either an animal hospital or a “hospitality”). It is important although to remember some limitations about the source of the data and the methodology before thinking of these findings as correct or absolute: first the venue information came from FS, so there may be pharmacies, or even hospitals, that are not included in their database, so would not show up in this project. Second, many of the tags in **FS** are either wrong or ambiguous (it mainly depends on how popular is the venue) so when making deep searches, you can get a lot of noise. Also, a quick review shows that some of the venues info has 2 years or more without any update (so there are venues that probably have already disappeared, but still shows in the database, or again, venues that have not yet being included). Lastly the search radius for the venues in this “run” was just 500 meters; by increasing the radius more venues will be included (although for coastal neighborhoods the effect will be lessen as it will be searching in the body of water’s area).

With regards of K-means, it is of notice that included Hudson Yards as cluster 2; and Battery Park City as cluster 3, and definitely these two present interesting characteristics as being the ones with less access, both are coastal and really just 1 pharmacy. It is possible that an algorithm such as DBSCAN would have clustered these two together.

Inwood, or cluster 0, has also a relatively low access (5 venues), but they are all in the main category of hospitals or pharmacies. By increasing k, eventually a set of clusters with neighborhoods from the lowest to the highest access, could be obtained.

6. Conclusion

So, answering the initial question: yes, FOURSQUARE (FS) data can be used to have a general estimation for the ‘accessibility’ of neighborhoods to healthcare venues, at least, with some considerations:

- Some of the data in FS is tagged ‘ambiguously’ or wrong, and deep searches could bring these values, producing undesired noise.
- FS is mainly devoted to venues not strongly related to healthcare, so there may be various important ones missing from their database, or could be mislabeled (as indicated before). There are other platforms more focused on the Health field with more current data. For future models the information from these sites can be combined with social media data to produce stronger models.
- K-means worked well identifying 2 interesting data points, but because its inherited random element, each run could yield different results.