Analyzing COVID-19 Vaccination Access in the City of Chicago: A Machine Learning Approach

Clinton McClure

IBM Data Science Professional Certificate

Coursera

April 3, 2021

Introduction

Like many U.S. cities, Chicago is working to provide as many COVID-19 vaccines as possible to its residents. As of March 24, 2021, only 24% of city residents have received at least one dose of a COVID-19 vaccine[1], and 12% of residents have been fully vaccinated. As eligibility requirements are relaxed, and as supply struggles to meet demand, city health officials will be interested in finding ways to make vaccinations as easy and as widely available as possible, especially those that they have identified as being uniquely vulnerable to infection, such as poor and minority communities[2]. Besides state-run sites, venues such as grocery stores, pharmacies, and medical centers will serve as local distribution sites that will administer doses to city residents[3]. This project will seek to compare neighborhoods across Chicago based on the number of grocery stores, pharmacies, and medical centers and identify any similarities or differences between neighborhoods using machine learning. Based on this comparison, this project will remark on any areas of the city that have relatively few venues where vaccines are likely to be obtained.

Data

The data that will be used in the project will be the following:

Names and Lat/Long coordinates of Chicago neighborhoods. These will be obtained from Wikipedia and the ArcGIS Geocoder package. There are 77 community areas in Chicago and over 200 neighborhoods. Community areas may contain multiple neighborhoods and may sometimes share the same name as a neighborhood. Since other data in the Chicago Data Portal that will be used in this project includes community areas, we will use them as the basis for comparison instead of neighborhoods to maintain consistency and to make it easier to join different sets of data. Using the Pandas library, we extract a table of community area names from the Wikipedia page, "Community Areas in Chicago" (https://en.wikipedia.org/wiki/Community_areas_in_Chicago #cite_note-City_basics-9). We run the community names through the ArcGIS Geocoder in order to obtain the coordinates of each community. After plotting a map of the neighborhoods using Folium and checking each location for accuracy, we found that the coordinates for the North Lawndale neighborhood were not accurate because the geocoder mistook North Lawndale for the name of a street in Chicago instead of a neighborhood. We performed a Google search for the coordinates of the neighborhood and replaced the coordinates in the dataframe. Plotting another map of the neighborhoods, we confirmed that the coordinates were more accurate.

Location data for grocery stores, pharmacies, and medical centers in Chicago. These will be obtained from HERE using its Geocoding & Search API (https://developer.here.com/documentation/geocoding-search-api/dev_guide/index.html). HERE contains many kinds of venue categories as well as location coordinates for those venues. This will enable us to easily use machine learning to find clusters of potential vaccine sites as well as areas with little to no access to vaccines.

Methodology

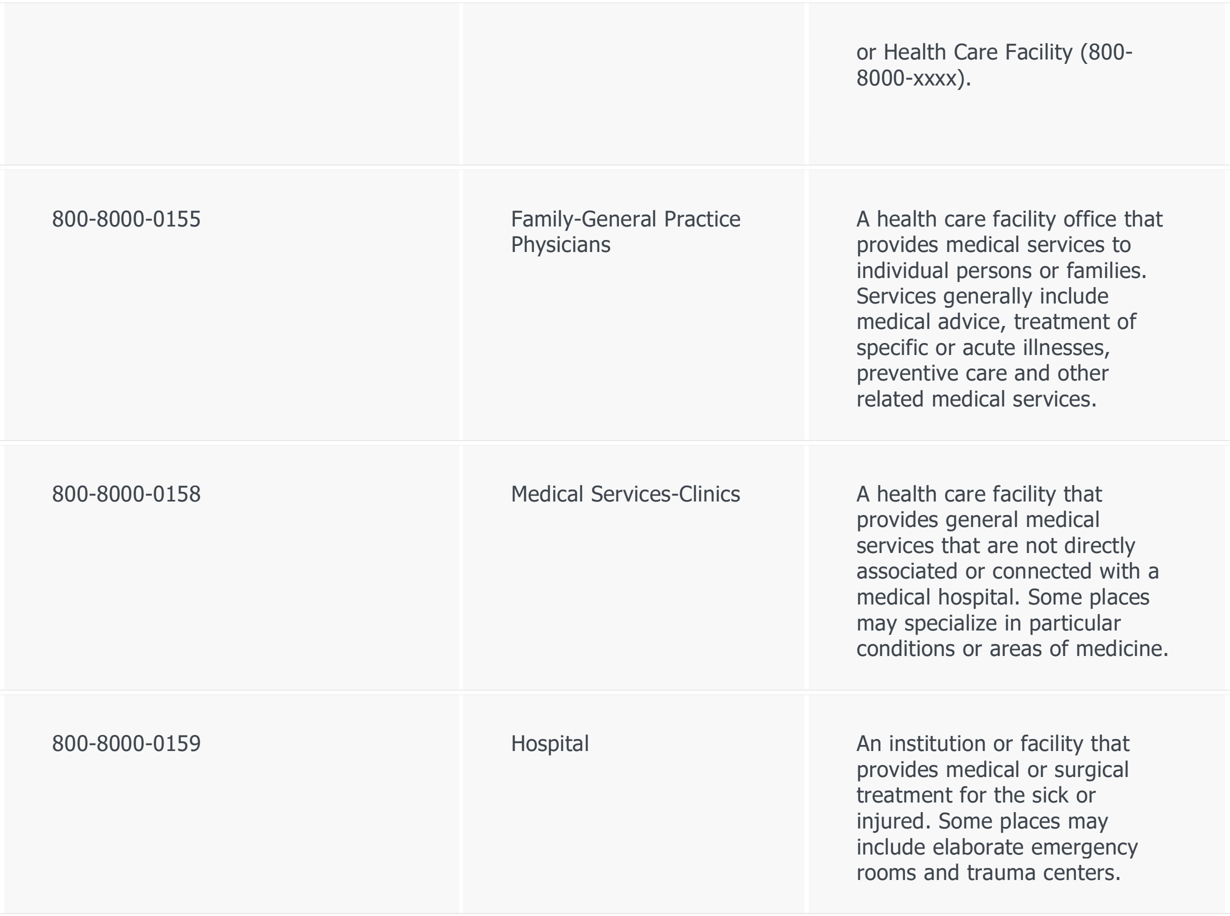
Data Wrangling and Cleaning

Since we are scraping a website and geocoding latitude/longitude coordinates, there will inevitably be some cleaning, wrangling, and verification before we can proceed to conducting our analysis. One issue that arose from using the ArcGIS Geocoder had to do with incorrectly reading two neighborhood names – North Lawndale and South Lawndale - as though they were street names. We found the errors after an initial plot of the neighborhoods in Chicago using Folium. Using Google search, we found the correct coordinates and replaced the erroneous ones in our dataframe.

Accessing the HERE Geocoding & Search API

HERE is a location services company that provides an API that can be used for requesting information on venues, geographic and administrative boundaries, and transportation networks. One feature of the API is that we can search for venues within a specific radius of a set of coordinates that match a particular category type (i.e. restaurants, schools, movie theaters, etc.). Since we are interested in only a few types of venues, we narrowed our GET request by including the following Category ID numbers in the URL that we pass to the API:

Category ID	Category (Level 3)	Description
600-6400-0000	Drugstore or Pharmacy	A business that sells medications, toiletry items and other retail cosmetics. Drugstores and Pharmacies generally sell both prescription and non-prescription medications. This is a base-level category that should be used for all places that do not fit other categories defined for Pharmacy (600-6400-xxxx).
600-6300-0066	Grocery	A business that sells a large variety of food including fresh produce, frozen foods, packaged goods, bakery items and meat products.
800-8000-0000	Hospital or Health Care Facility	An institution or facility that provides medical or surgical treatment for the sick or injured. Places range from small clinics and doctor's offices to urgent care centers and large hospitals containing elaborate emergency rooms and trauma centers. This is a base-level category that should be used for all places that do not fit other categories defined for Hospital

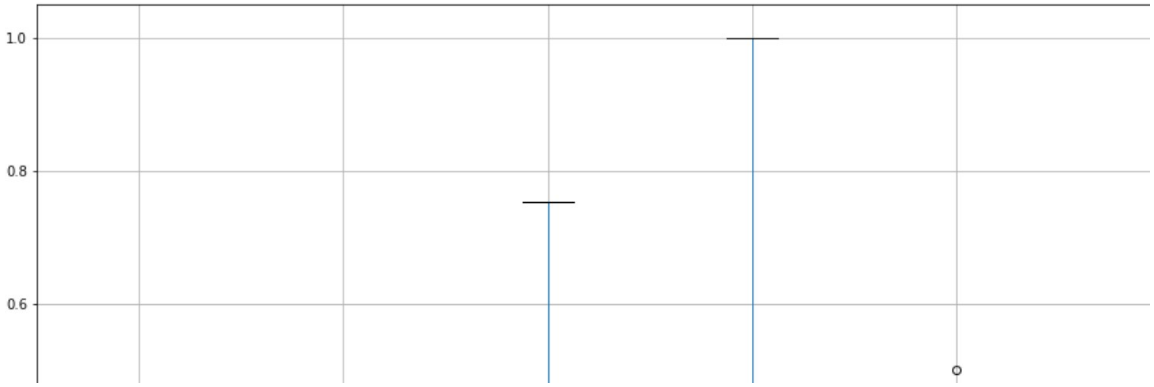


		or Health Care Facility (800-8000-xxxx).
800-8000-0155	Family-General Practice Physicians	A health care facility office that provides medical services to individual persons or families. Services generally include medical advice, treatment of specific or acute illnesses, preventive care and other related medical services.
800-8000-0158	Medical Services-Clinics	A health care facility that provides general medical services that are not directly associated or connected with a medical hospital. Some places may specialize in particular conditions or areas of medicine.
800-8000-0159	Hospital	An institution or facility that provides medical or surgical treatment for the sick or injured. Some places may include elaborate emergency rooms and trauma centers.

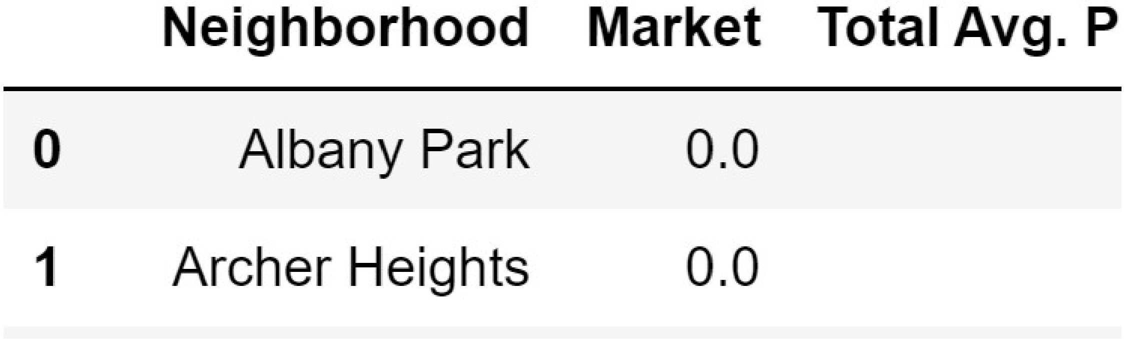
HERE Place Category System for grocery and healthcare-related venues (https://developer.here.com/documentation/geocoding-search-api/dev_guide/topics-places/places-category-system-full.html)

Exploratory Data Analysis

Before building our k-means clustering model, it is prudent to explore our data in order to identify any patterns and/or outliers that may skew our analysis. First, we draw a boxplot of the venues dataframe:



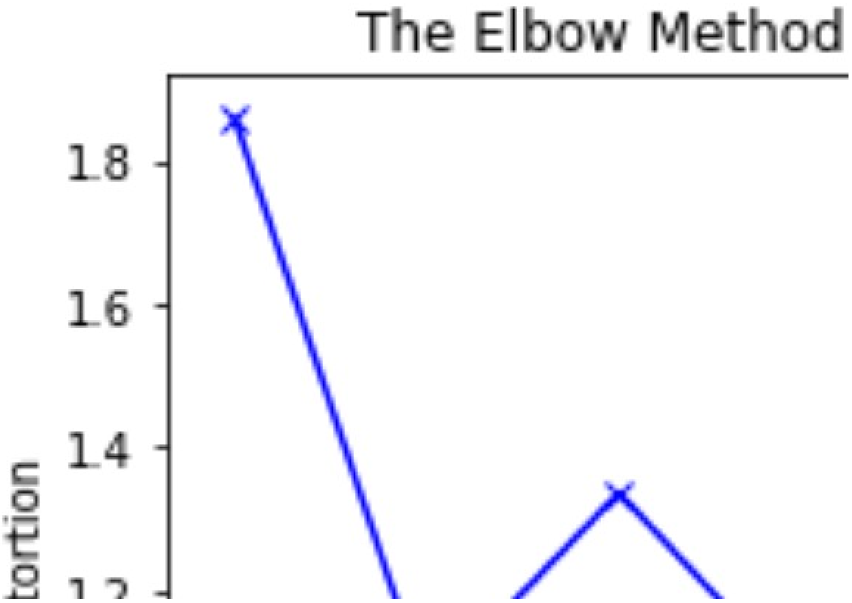
This boxplot shows the spread of the average frequency of occurrence of each venue type. The values for Family/General Practice Physicians, Grocery, and Medical Services/Clinics appear to be much greater than the rest of our venue data. Although we are interested in these venue categories, we don't want to skew our clustering analysis, so we will drop these columns from our dataframe. After dropping these columns, we combined columns with similar venues in order to make it easier to conduct our analysis and to interpret our results. Below is the head of our new dataframe:



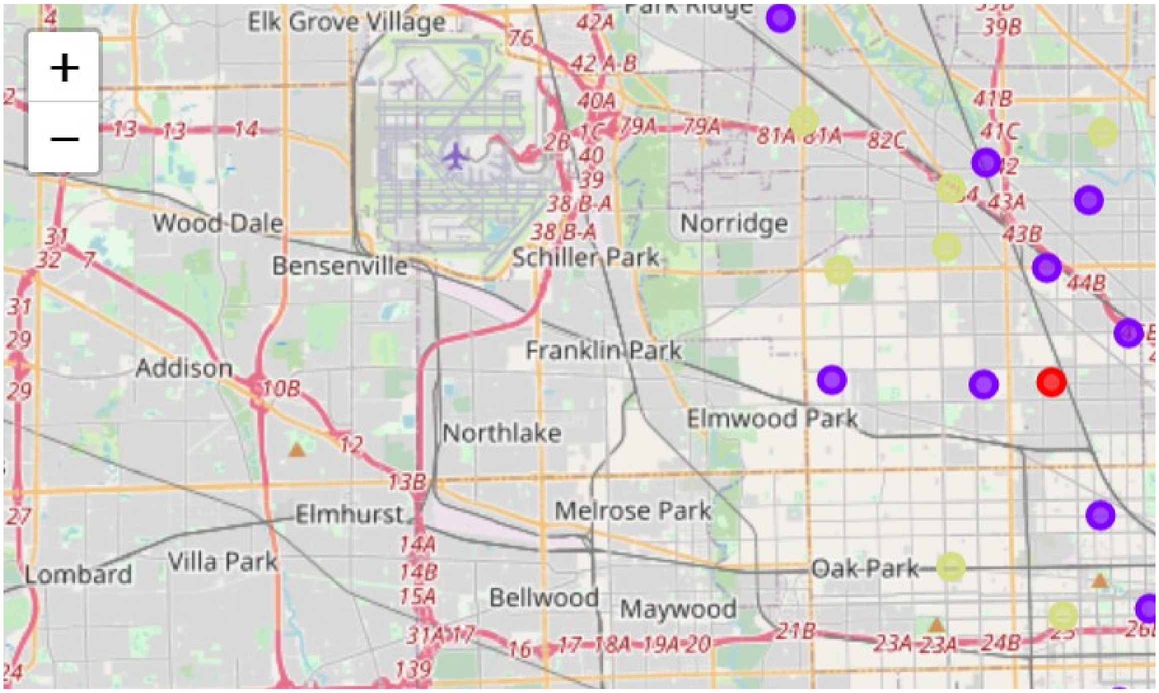
	Neighborhood	Market	Total Avg. P
0	Albany Park	0.0	
1	Archer Heights	0.0	

K-Means Clustering

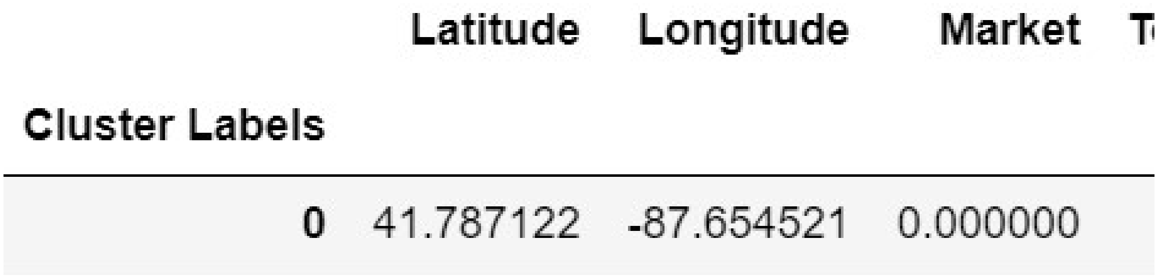
K-Means clustering is the most common unsupervised machine learning technique in data science. It is an iterative process that finds the centroid(s) of a cluster or clusters such that all points within a cluster are as close as possible and all points between clusters are as far apart as possible. In this analysis, we used k-means clustering in order to identify all unique clusters among neighborhoods in Chicago based on the average number of pharmacies, supermarkets, and healthcare centers found in each neighborhood. We can choose any number of clusters in order to create our model, but there are techniques to help us select the right number of clusters. We use the elbow method in order to find the optimal number of clusters. The elbow method relies on calculating and visualizing the average sum of the square of the distance between the cluster and the cluster centroids. This is known as distortion, and the optimal number of clusters can be found where the distortion is found to be declining at a linear rate. Using the graph below, we find that the optimal number is k = 4:



Now that we have the number of clusters that we need to model, we build our k-means cluster model and plot the clusters on our map of the city:



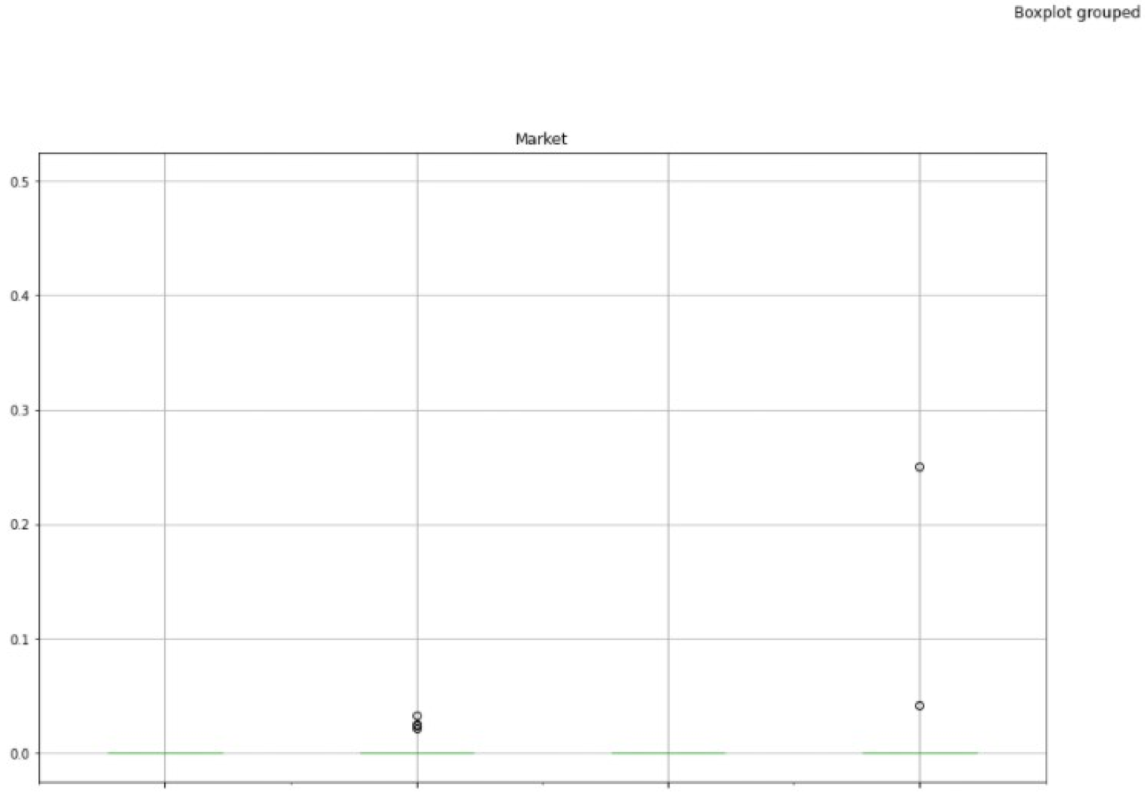
The red dots are part of Cluster 1 (where Cluster Label = 0), the purple dots are part of Cluster 2 (where Cluster Label = 1), the teal dots are part of Cluster 3 (where Cluster Label = 2), and the green dots are part of Cluster 4 (where Cluster Label = 3). We aggregate the average number of vaccine sites based on the cluster label:



	Latitude	Longitude	Market	T
Cluster Labels				
0	41.787122	-87.654521	0.000000	

Results

Let’s explore the data again using our new cluster labels with a boxplot:



Based on the above, we can describe our clusters in the following ways:

Cluster 1 (label 0) has, on average, more health care centers than the others and few markets.

Cluster 2 (label 1) has the most number of neighborhoods of all clusters. It has the second most number of pharmacies on average, and the second highest average number of markets.

Cluster 3 (label 2) has significantly more pharmacies than the other clusters. It has the fewest number of markets and healthcare centers.

Cluster 4 (label 3) has a few neighborhoods with access to markets, but it has, on average, very few pharmacies and healthcare centers.

Discussion

We have found four distinct clusters across Chicago based on neighborhood access to pharmacies, markets, and health care centers, which will be heavily used by residents to receive vaccines for COVID-19 for the coming months. Supermarkets, on the whole, are not the best sites for wide distribution of vaccines, as they are not as plentiful as drugstores and clinics. Public health officials would be wise to make pharmacies a primary distribution channel for vaccines, as they are plentiful across many neighborhoods. Two neighborhoods in particular, Chatham and West Pullman, would rely solely on pharmacies as a means to vaccinate residents. Health care centers, including hospitals and clinics, can certainly supplement the distribution of vaccines since they cover a measurable portion of neighborhoods.

Conclusion

Machine learning has a wide array of applications, including in the area of public health. By using k-means clustering, an unsupervised learning technique, we can group neighborhoods across the City of Chicago into clusters and compare them based on the number of venues that city officials and local media have identified as vaccination sites. We analyzed each cluster and remarked on the defining characteristics of each one and made recommendations for distribution strategies based on our findings. During our initial data wrangling, we chose to exclude data that appeared to be highly skewed. Should this analysis be refined, we can find a way to incorporate and normalize this data so as not to skew the results of our clustering. Public health officials are certainly interested in finding ways to vaccinate as many residents as possible. Data science can be a powerful tool in order to help them do just that.

[1] "COVID-19 Daily Vaccinations - Chicago Residents - Cumulative Doses by Day." Chicago Data Portal, City of Chicago. https://data.cityofchicago.org/Health-Human-Services/COVID-19-Daily-Vaccinations-Chicago-Residents-Cumu/rna5-2pgy.

[2] "Chicago COVID-19 Community Vulnerability Index Chicago CCVI." City of Chicago. https://www.chicago.gov/content/dam/city/sites/covid/reports/012521/Community_Vulnerability_Index_012521.pdf

[3] “How to Sign Up for COVID Vaccine Appointments in Illinois” NBC Chicago. March 2, 2021. https://www.nbcchicago.com/news/local/how-to-sign-up-for-covid-vaccine-appointments-in-illinois/2451346/