Applied Data Science Capstone Final Project Report
Using Venue information to cluster Zip Codes Areas in the city of Chicago

Julio Guerrero

August 2021

I. Introduction

I.1. Background

The city of Chicago, located in the state of Illinois, is one of the most populated cities in the U.S, with an estimated population of 2,693,976 people. It also has one of the biggest GDP in the world, $689 Billion in 2018 (1). The city can be divided in many ways. The two main ways this is done for data collection is by: Zip codes or Community Areas. There is map that shows in detail how these boundaries are drawn (2).

All the data in this project will be from dataset organized by Zip Codes. It is important to note that there are a couple of Zip Codes that encompass areas that are outside city limits (Example: Zip Code 60827).

As many US Cities, Chicago economy suffers because of the pandemic. Trying to control the pandemic, the Government stablished strict lockdowns. This cause many Chicago-Area business closing their doors, increasing unemployment (3). This makes the risk of future lookdowns a very real threat for venues owner.

I.2. Problem

The aim of the project is to be able to cluster different areas of the City of Chicago so to label them based on their nearby venues and contrast these findings with the percentage of the population fully vaccinated.

It´s logical to think that the faster everyone gets vaccinated, the faster the city can go to pre-pandemic movement. This can only be done by vaccinating more people. Having local outbreaks pushes local officials to impose restrictions and regulations that can be detrimental to stores and venues. While many US states are close at becoming fully vaccinated, having the full picture of this reality in number is important so investors

can make inform decisions. This makes areas with higher vaccination more desirable for possible investor.

This project can help features venues owners or investors see which area is better suited for their type of venues, while helping them understand the risk of suffering local outbreaks and lose money because of it.

II. The Data

The city of Chicago has a vast amount of Data on their official web page. The information that will be using will be:

1. A GeoJSON file with the Zip Codes boundaries of the City of Chicago. This will be used for visualization of the data collected.(4)
2. A Dataset of the City of Chicago Data Portal call “COVID-19 Vaccine Doses by ZIP Code - Series Completed”. We are interested in the percentage of the population fully vaccinated by zip code and the coordinates of the area. (5)
3. We will use the Foursquare API to get the most common venues of each Zip Code.

We will use the coordinates of the second Dataset as the center of our search for venues using the Foursquare API. We will organize this information by Zip Code to use it for clustering each Zip Code center. Using the data of the percentage of people fully vaccinated by zip code, we will make a choropleth map making use of the GeoJSON file. After clustering each zip code center, we will plot each point that map.

III. Methodology

The main dataset we will use in this project is the “COVID-19 Vaccine Doses by ZIP Code – Series Completed” from the City of Chicago Data Portal. This dataset has 22 columns, and every row is an entry on the state of each Zip Code Area on a given date.

We are only interested in the “Population”, “Zip Code Location”, “Zip Code” and “Vaccine Series Completed - Percent Population” columns. Also, we want to know the latest values of last one. We will also need to manipulate the Location information to extract the latitude and longitude. The resulting data frame can be view in the next figure, it has a total of 59 different Zip Codes.

	Zip Code	Population	Latitude	Longitude	Vaccine Series Completed-Percent Population
0	60601	15083	41.886262	-87.622844	0.676
1	60602	1145	41.883136	-87.628309	0.958
2	60603	1052	41.880112	-87.625473	1.000
3	60604	823	41.878153	-87.629029	0.883
4	60605	29060	41.867824	-87.623449	0.648

Figure 1. Main Data Frame after cleaning

We used folium library to visualize geographic information in this project. First, we used the information of the previous data frame to mark the Zip Code coordinates in a map with the boundaries of each Zip Code (using the GeoJSON file). We can see the resulting map in the next figure.

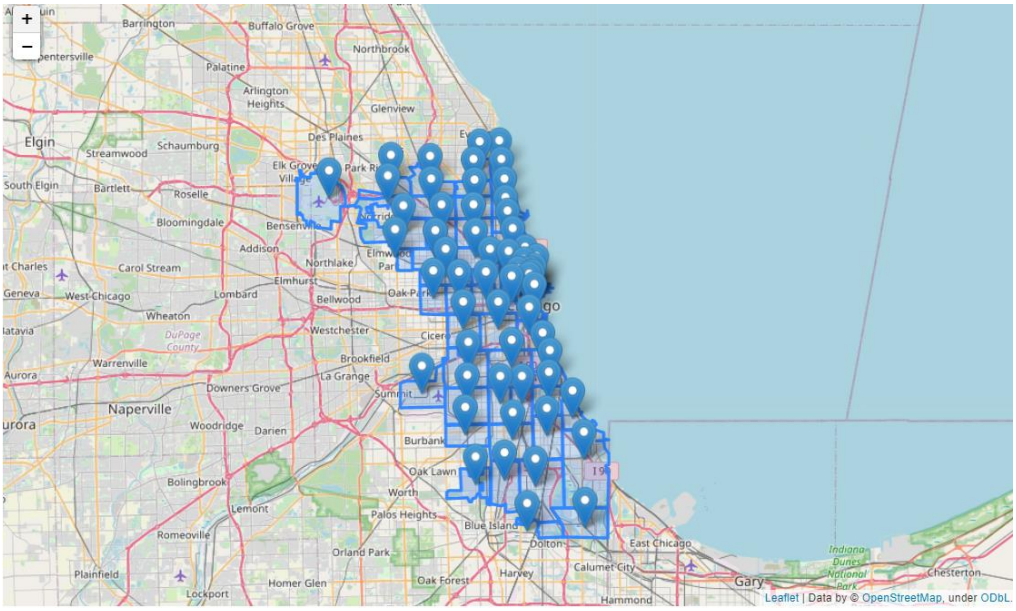

Figure 2. City of Chicago Divided by Zip Codes Areas and their center

In a closer look at the map, we can see that a couple of markers (Zip Codes: 60827 and 60707) seem to be out of their boundaries. This is because the way the city of Chicago is divided. This Zip Codes have parts of their area outside city limits and only the section belonging to the City itself is present in the GeoJSON file.

III.1. Foursquare API

With the coordinates of each Zip Code center, we run multiples queries with the Foursquare API to gather information about the type of venues that are located close to each Zip Code. We establish a limit of **100 venues** and a radius of **950-meters**. With the information collected we make a new data frame with information of each venue, as we can see in the next figure. We had a total of **3467 venues**, with **321 unique categories**

	Zip Code	Zip Code Latitude	Zip Code Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	60601	41.886262	-87.622844	Chicago Architecture Center	41.887720	-87.623650	Tour Provider
1	60601	41.886262	-87.622844	Roti Modern Mediterranean	41.886048	-87.624948	Mediterranean Restaurant
2	60601	41.886262	-87.622844	sweetgreen	41.884964	-87.624728	Salad Place
3	60601	41.886262	-87.622844	Wildberry Pancakes & Cafe	41.884412	-87.623047	Breakfast Spot
4	60601	41.886262	-87.622844	St. Jane Chicago	41.886573	-87.624902	Hotel

Figure 3. Data Frame with each venue collected

Next, we group each venue by Zip Code to see the total amount of venues by Zip Code. We can se that information in the next graph.

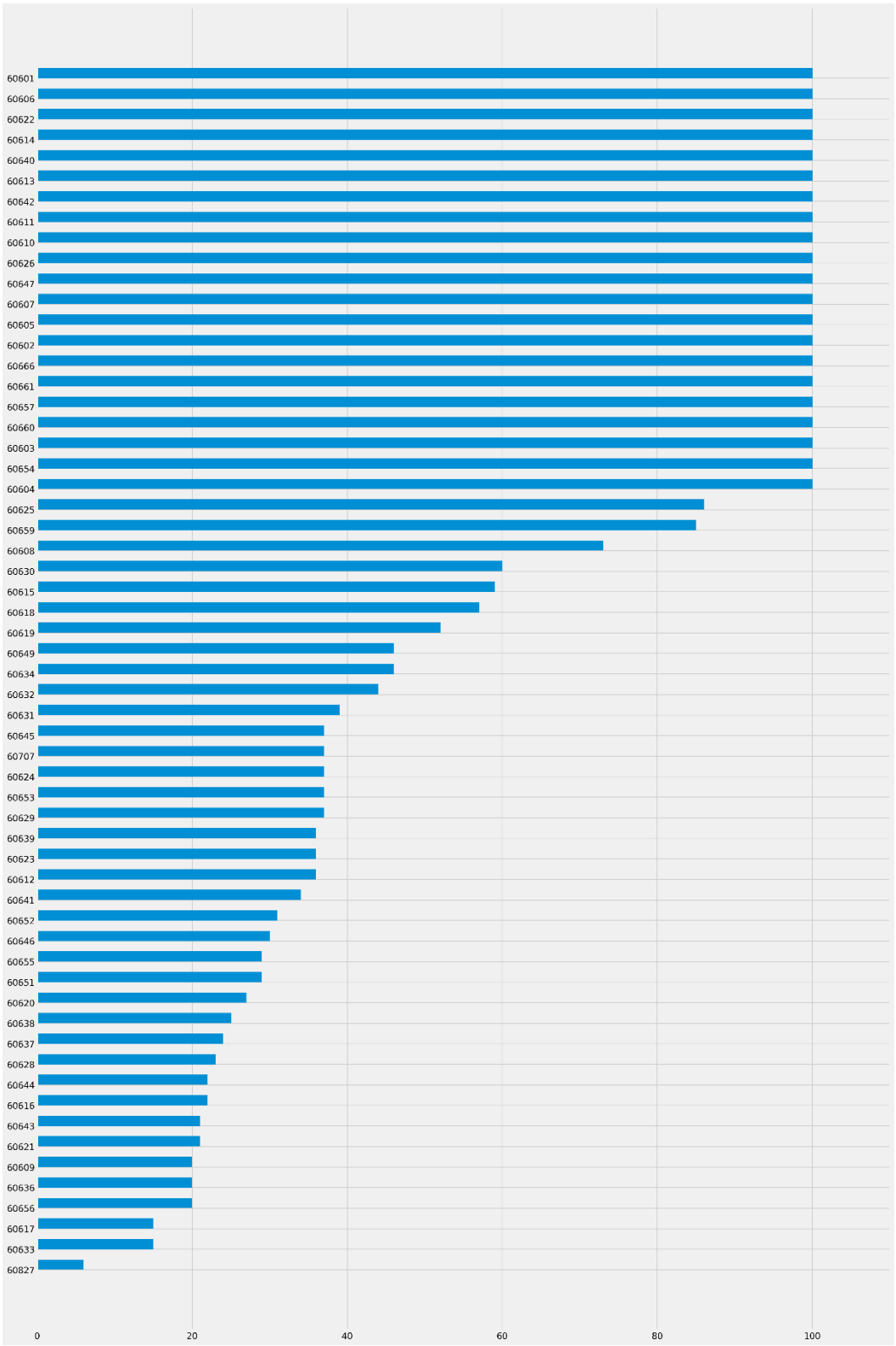

Figure 4. Bar chart of the number of venues in each Zip code

There are venues with less than 100 venues. This does not mean that those are all the venues in that Zip Code area. This is likely the effect of the coordinates that the Zip Code center is located. It is possible that in larger Zip Codes Areas the denser venues sections are not close to the center.

The next thing we did was to one-hot encode each venues base on their category. This data frame will be used to group the venues by Zip Code based on the frequency of each venue category in the area.

We make a Data frame with the top 10 venues of each Zip Code, as we can see in the next figure.

	Zip Code	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	60601	Hotel	Seafood Restaurant	American Restaurant	Theater	Snack Place	Pizza Place	Cosmetics Shop	Park	Steakhouse	Breakfast Spot
1	60602	Hotel	Museum	Steakhouse	Theater	Pizza Place	Coffee Shop	Seafood Restaurant	Salad Place	Bar	Park
2	60603	Coffee Shop	Theater	Museum	Italian Restaurant	Hotel	Salad Place	Asian Restaurant	Sushi Restaurant	Concert Hall	Garden
3	60604	Coffee Shop	Hotel	Italian Restaurant	Museum	Salad Place	Asian Restaurant	Theater	Café	Middle Eastern Restaurant	Pub
4	60605	Pizza Place	Aquarium	Gym / Fitness Center	Yoga Studio	Burger Joint	Hotel	Sandwich Place	History Museum	Football Stadium	American Restaurant

Figure 5. Data Frame of most common venues by Zip Code

Clustering

For this project we want to create labels for each Zip Code Area based on the most common venues. For this we used K means, which is a form of unsupervised clustering algorithm. The main parameter that needs to be tune is the number of clusters (also known as K). For this we train the model with different values of K and use the dissimilarity between the clusters to measure their performance. With those values in function of K, we can plot the result and use the elbow method to determine the best value for K, as we did for the next graph.

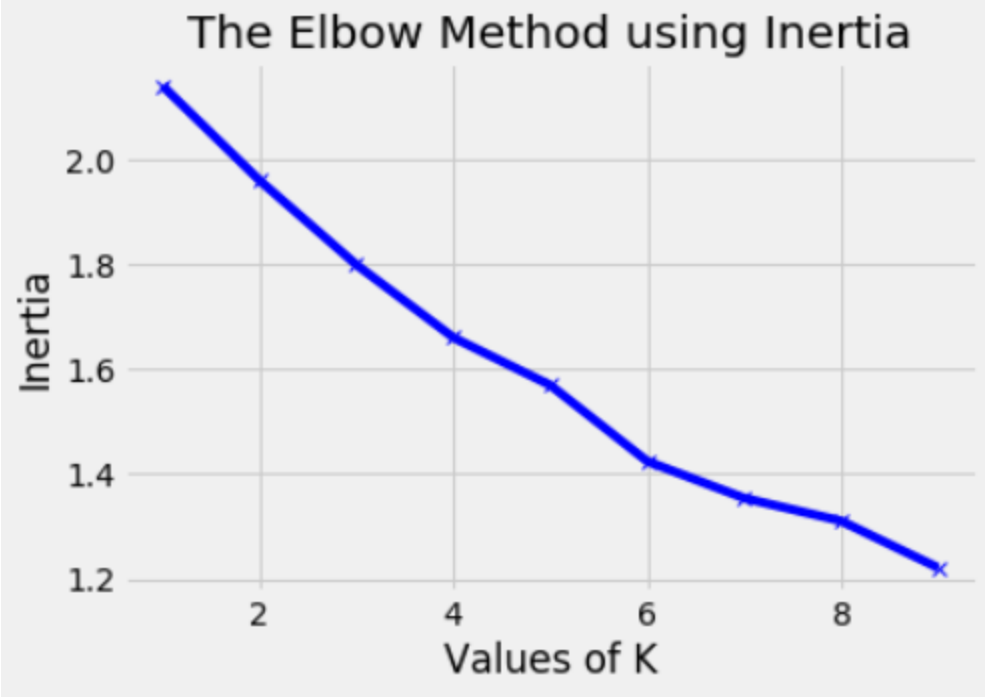

Figure 6. Graph of elbow method

Although there is not a clear elbow, the value of K equals 5 produced clusters with big enough number and sufficient difference between them that satisfies us.

After training the model with K equal 5 we use the new labels to group each venue base on their new label to look at the frequency of label´s top 5 venues. With that, we name each label based on those things, except the cluster with only one Area.

Cluster 0: Park and fast food (Red in the map)

Cluster 1: Coffee and Italian food (Purple in the map)

Cluster 2: 60827 (Blue in the map)

Cluster 3: Mexican Food and Pizza (Green in the map)

Cluster 4: Parks and Grocery (Orange in the map)

After that we use folium to make a choropleth map of the Chicago Zip Code Areas base on the percentage of people fully vaccinated and use markers of different colors based un the Zip Code cluster. For this reason, we try to use DBSCAN, but were unable to tune to produce good clusters.

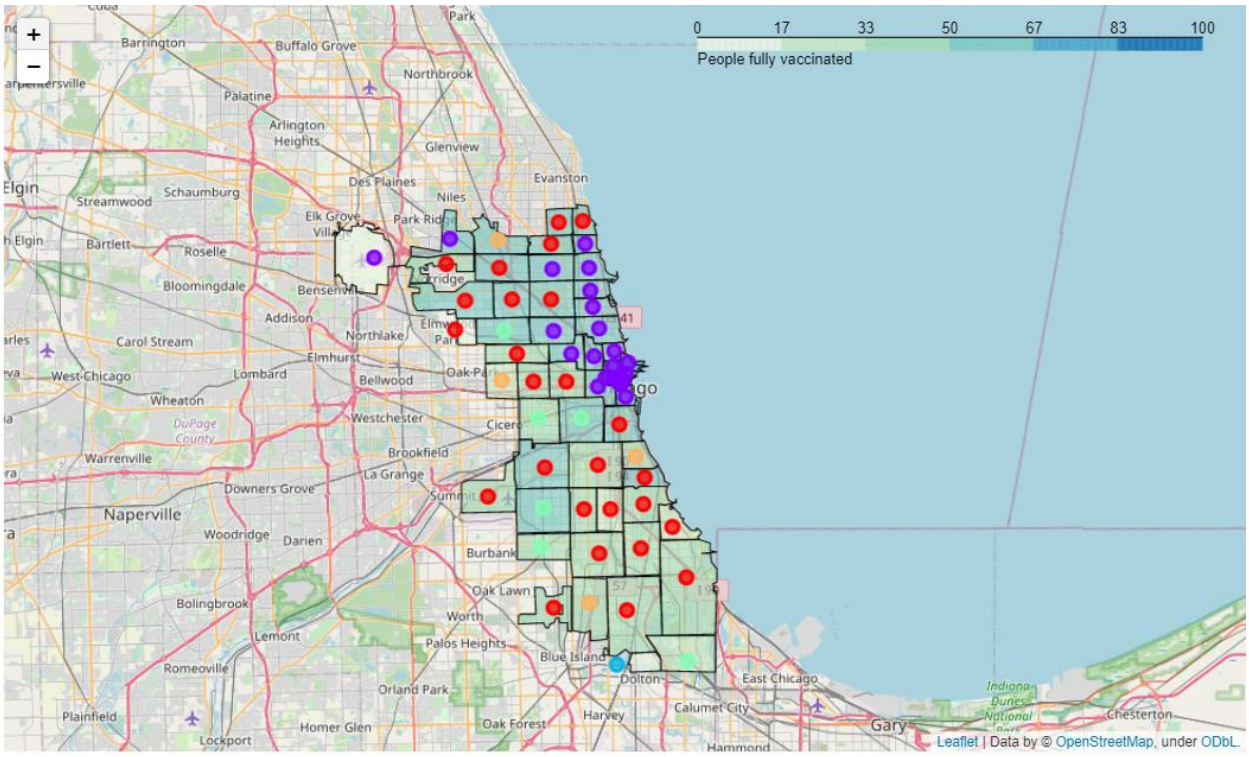

Figure 7. Resulting Choropleth map of the percentage of people vaccinated and markers identifying each cluster

IV. Results

There seems to be that areas of cluster 1 (Coffee and Italian food, the purple dots) are closer to the city center. Also, this points area located in areas that are almost fully vaccinated. Cluster 0 (Park and fast food, red dots) seems to surround cluster 1, which may explain the higher frequency of parks.

Cluster 4 (Parks and grocery) seems to be the one farther away of the city center, which may explain why it has more “convenient” types of foods and locals. The last cluster seems to have higher frequency of Mexican restaurants and pizza venues.

We also see that the closer we get to the city center, the higher the percentage of people fully vaccinated. It does not seem to be a correlation between our clusters and the percentage of people vaccinated.

V. Discussions

This project could be complemented with more demographic information about the city of Chicago (as Per Capita Income and population density) to give the stakeholders more information about what type of business is better suited for each area. This information is available in the Chicago Data portal but is given by Community Areas.

This project can be improved by dividing the city in smaller areas. An idea for this is to create hexagons around the city and give each one the data on the area they are located. That could solve the problem of having data that divided the city in a different way. After that each, hexagon can be cluster, as done in this project. This could improve the way we look at each section of the city.

The Zip Code 60666 belongs to the airport. Given that airports are very different from the normal areas of a city, is important that a stake holder who wants to target the airport looks for more specific information for that.

Because there is no clear elbow in the clustering section of the project, there is argument for using other type of clustering algorithm. We try to use DBSCAN but were unable to tune the parameters to produce good clusters.

VI. Conclusions

We were able to group each Zip Code of the city in 5 clusters based on their venues. With that information, we made a couple of data frames that can tell future investors which are the most common venues and their frequency. We also created a Map to give an idea of the distribution of our clusters and the percentages of people fully vaccinated in the city of Chicago.