Capstone Project

IBM Applied Data Science Capstone

Opening a New Vegan Restaurant in Chicago, IL

By : Shruti Shah

June 20, 2020

1. Introduction:

1.1.Business Problem:

Being a vegetarian, I know how difficult it is to find a right place that serves a really good vegetarian or vegan meal. The objective of this capstone project is to analyze and choose which location would be a best in Chicago to open a new vegan restaurant. If you are or someone who you know are planning to open a vegan restaurant and stressing with the following questions, then this report should help you out :

1) **Should I open a restaurant in a place where there are already lot of restaurants? Or**
2) **Should I open it where there are no vegan restaurants? Or**
3) **Should I try to find some kind of balance between the other two options?**

This report outlines some basic assumptions using data science methodology, machine learning techniques like clustering, and it will give you information based on the data provided by Foursquare API.

1.2. Target Audience:

As I mention earlier, the targeted audience would be anyone who wants to buy or build a property for a vegan restaurant in Chicago.

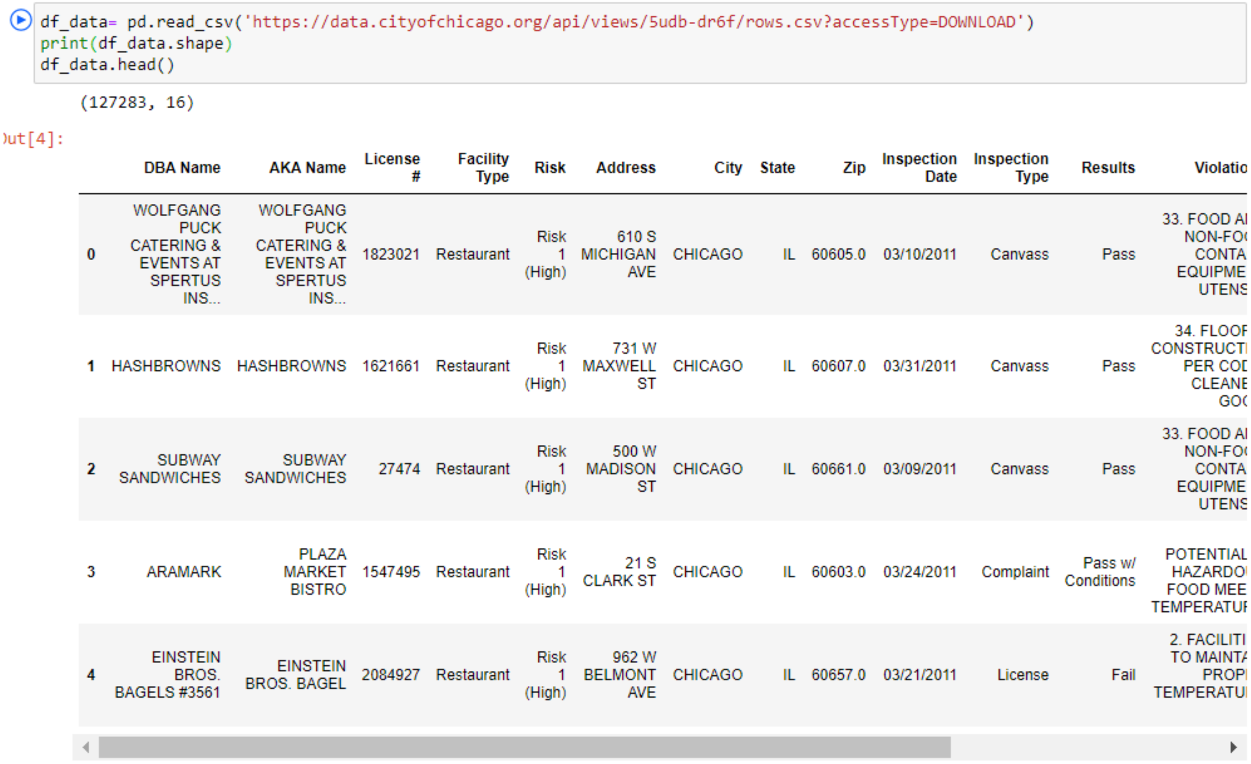
2. Data:

In order to make the best decision, we're going to need some data. Fortunately, there are hundreds of data sets of Chicago and its neighborhoods that describes various aspects of the city like zip codes, crime data, registered business data. Also, the Foursquare API allows free access to some of its location and venue data, so that would be a good option to look at the neighborhood venues. Altogether, we will be looking at these sets of data for our analysis:

1) Chicago and its neighborhoods Registered Restaurant Business Data
2) Chicago and its neighborhoods Reported Crimes Data
3) Clustering Vegan Restaurants in Chicago neighborhoods
4) Data Visualization and Exploration
5) Foursquare Data Analysis

2.1. Chicago Registered Business Data of Restaurants:

We can pull a list of every business registered in Chicago and its neighborhoods from the last ten years from the Chicago Restaurant Data Website. we can extract the data from the website into Pandas data frame using python. This will show us the data of all the businesses located within each neighborhood and help us understand the foot traffic in each neighborhood of Chicago.



```
df_data= pd.read_csv('https://data.cityofchicago.org/api/views/5udb-dr6f/rows.csv?accessType=DOWNLOAD')
print(df_data.shape)
df_data.head()
```

```
(127283, 16)
```

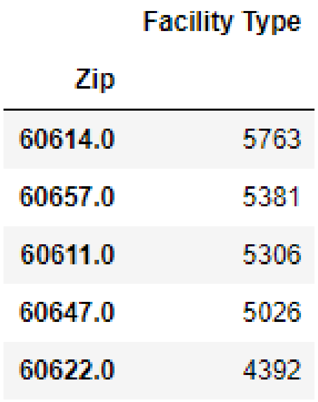
Out[4]:

	DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type	Results	Violatio
0	WOLFGANG PUCK CATERING & EVENTS AT SPERTUS INS...	WOLFGANG PUCK CATERING & EVENTS AT SPERTUS INS...	1823021	Restaurant	Risk 1 (High)	610 S MICHIGAN AVE	CHICAGO	IL	60605.0	03/10/2011	Canvass	Pass	33. FOOD A NON-FO CONTA EQUIPME UTENS
1	HASHBROWNS	HASHBROWNS	1621661	Restaurant	Risk 1 (High)	731 W MAXWELL ST	CHICAGO	IL	60607.0	03/31/2011	Canvass	Pass	34. FLOO CONSTRUCT PER CO CLEAN GO
2	SUBWAY SANDWICHES	SUBWAY SANDWICHES	27474	Restaurant	Risk 1 (High)	500 W MADISON ST	CHICAGO	IL	60661.0	03/09/2011	Canvass	Pass	33. FOOD A NON-FO CONTA EQUIPME UTENS
3	ARAMARK	PLAZA MARKET BISTRO	1547495	Restaurant	Risk 1 (High)	21 S CLARK ST	CHICAGO	IL	60603.0	03/24/2011	Complaint	Pass w/ Conditions	POTENTIAL HAZARDO FOOD MEE TEMPERATU
4	EINSTEIN BROS. BAGELS #3561	EINSTEIN BROS. BAGEL	2084927	Restaurant	Risk 1 (High)	962 W BELMONT AVE	CHICAGO	IL	60657.0	03/21/2011	License	Fail	2. FACILITI TO MAINT PROP TEMPERATU

we can see that there are 127,283 entries in the data frame which includes restaurant names, their license number, address, zip codes, latitude, longitude and more.

2.2. Data Cleaning and Preprocessing:

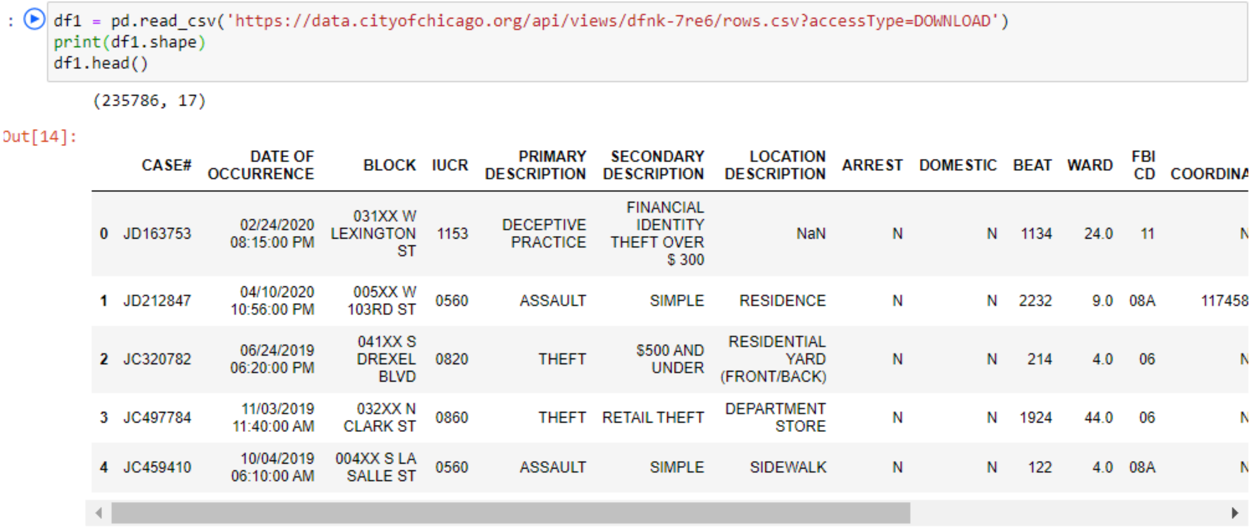
We will user these methods to transform the raw data into an understandable format and make it more efficient. We want to count the number of businesses registered in Chicago in the last 10 years to get better idea for the neighborhoods. First, we will drop the rows where latitude and longitude values are none(NAN), then we will group the data by Zip codes and count the number of businesses registered in those areas. This will give us a rough indication of how much foot traffic each area of the city gets today. After coding, it looks like the zip code – 60614 which is the Northeastern side of Illinois has significantly more business registration than everywhere else:



	Facility Type
Zip	
60614.0	5763
60657.0	5381
60611.0	5306
60647.0	5026
60622.0	4392

2.3. Chicago Crime Data :

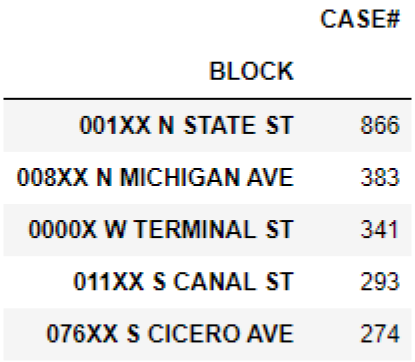
Now, let’s get the Chicago’s crime data for the last few year from Chicago Crime Data. This way we can see which areas are safe to open a restaurant.



```
: df1 = pd.read_csv('https://data.cityofchicago.org/api/views/dfnk-7re6/rows.csv?accessType=DOWNLOAD')
  print(df1.shape)
  df1.head()

    (235786, 17)

Out[14]:
```

	CASE#	DATE OF OCCURRENCE	BLOCK	IUCR	PRIMARY DESCRIPTION	SECONDARY DESCRIPTION	LOCATION DESCRIPTION	ARREST	DOMESTIC	BEAT	WARD	FBI CD	COORDINA
0	JD163753	02/24/2020 08:15:00 PM	031XX W LEXINGTON ST	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	NaN	N	N	1134	24.0	11	N
1	JD212847	04/10/2020 10:56:00 PM	005XX W 103RD ST	0560	ASSAULT	SIMPLE	RESIDENCE	N	N	2232	9.0	08A	117458
2	JC320782	06/24/2019 06:20:00 PM	041XX S DREXEL BLVD	0820	THEFT	$500 AND UNDER	RESIDENTIAL YARD (FRONT/BACK)	N	N	214	4.0	06	N
3	JC497784	11/03/2019 11:40:00 AM	032XX N CLARK ST	0860	THEFT	RETAIL THEFT	DEPARTMENT STORE	N	N	1924	44.0	06	N
4	JC459410	10/04/2019 06:10:00 AM	004XX S LA SALLE ST	0560	ASSAULT	SIMPLE	SIDEWALK	N	N	122	4.0	08A	N

we can see that there are 235,786 entries in the data frame which includes case number, data of incident, block/neighborhood, description of the crime, location, latitude, longitude and more. let us use data cleaning and preprocessing again to clean up the data and make it look more efficient. We will start by dropping the latitude-longitude rows with NAN values, then we’ll group the data by Block and count the number of cases reported in those blocks.



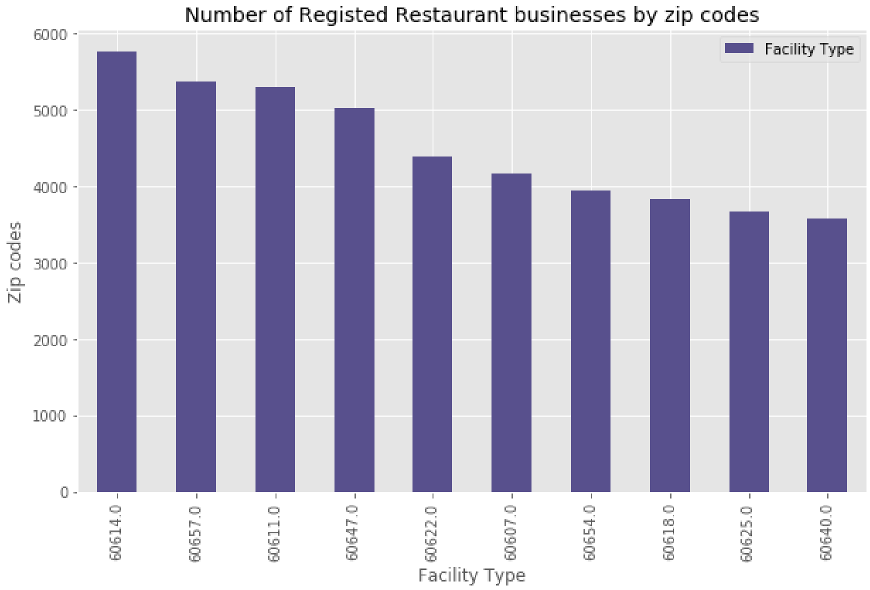
	CASE#
BLOCK	
001XX N STATE ST	866
008XX N MICHIGAN AVE	383
0000X W TERMINAL ST	341
011XX S CANAL ST	293
076XX S CICERO AVE	274

It looks like that more crimes were reported near the North side of Chicago than everywhere else. Now let’s move on to visualizing these data frames.

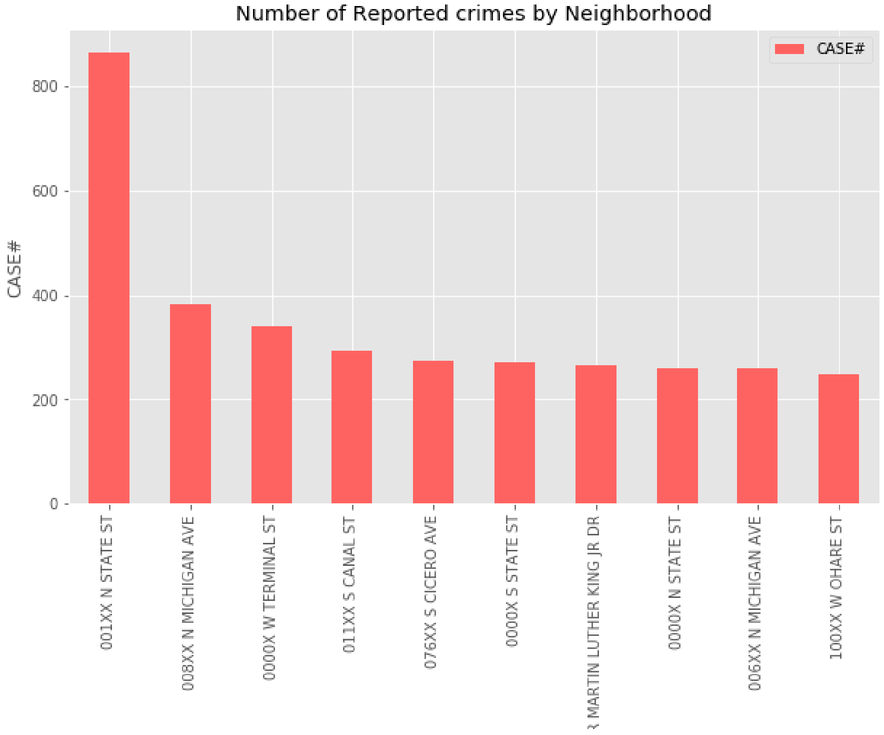
3. Methodology : Data Visualization and Exploration

3.1. Narrowing Down and Plotting the Blocks/Neighborhoods

We can start with simple visualization tool like bar plots to visualize our data sets and narrow it down by only focusing on 10 most registered restaurant businesses, we'll sort the data from most to least.

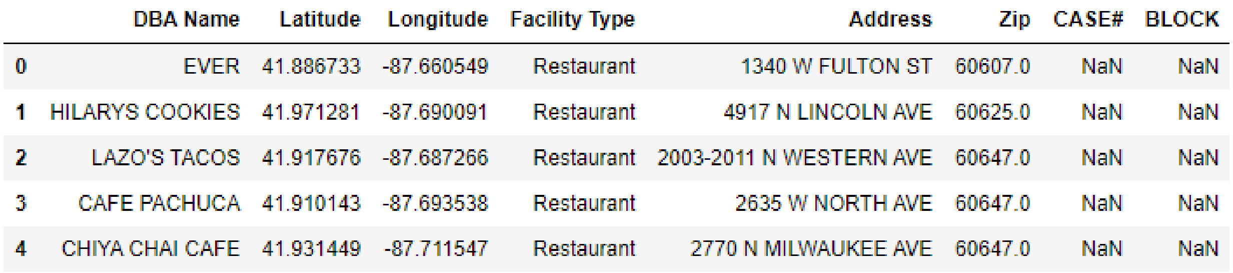


We will do the same process to look at the visualization of neighborhoods with most crimes. We will sort the data in descending order to get the 10 most reported crimes.



The above data shows the most significant crimes in first 10 neighborhoods. Our plan is to open a restaurant, we want customers to fill safe doesn’t matter what time of the day they come to the restaurant. This bar plots indicates that we should try to avoid these areas since there are many crimes reported and we should try not to open a restaurant in an area where there’s already many restaurants.

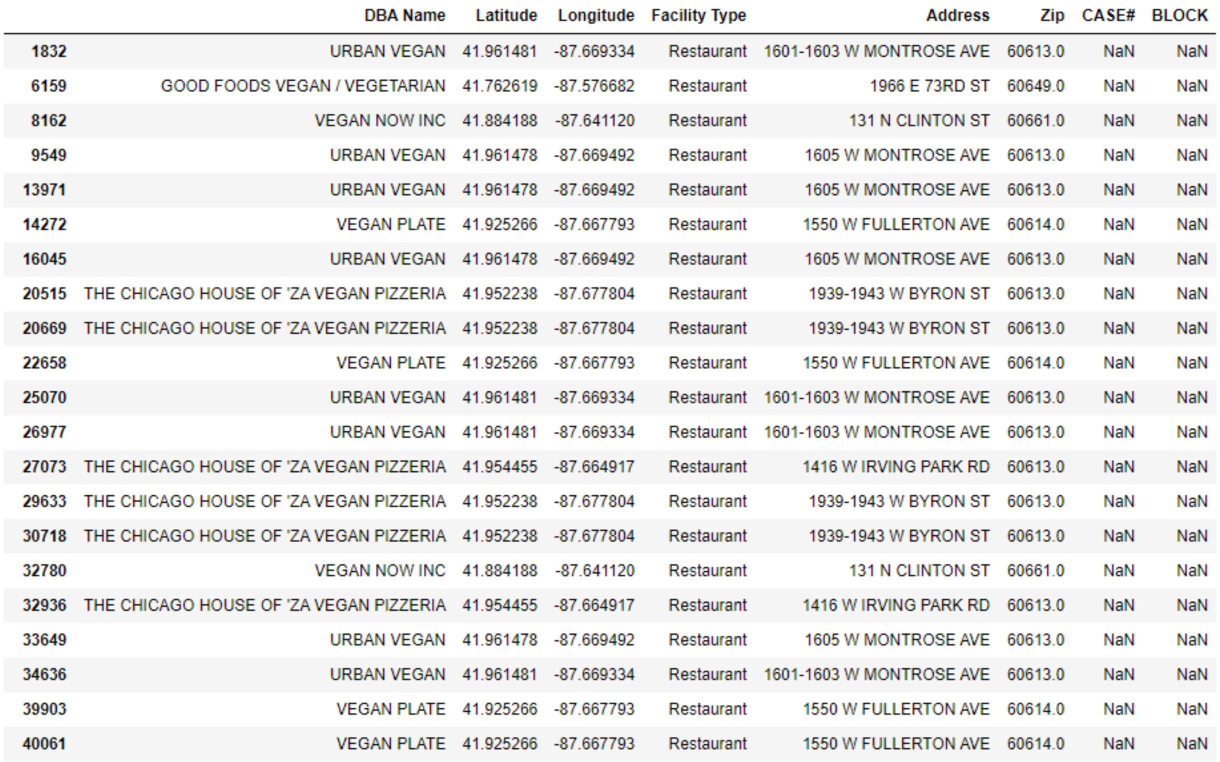
Let’s move on to the next step and merge the two data frames : registered business data and crimes data frames and ger a refined list of neighborhoods.



	DBA Name	Latitude	Longitude	Facility Type	Address	Zip	CASE#	BLOCK
0	EVER	41.886733	-87.660549	Restaurant	1340 W FULTON ST	60607.0	NaN	NaN
1	HILARYS COOKIES	41.971281	-87.690091	Restaurant	4917 N LINCOLN AVE	60625.0	NaN	NaN
2	LAZO'S TACOS	41.917676	-87.687266	Restaurant	2003-2011 N WESTERN AVE	60647.0	NaN	NaN
3	CAFE PACHUCA	41.910143	-87.693538	Restaurant	2635 W NORTH AVE	60647.0	NaN	NaN
4	CHIYA CHAI CAFE	41.931449	-87.711547	Restaurant	2770 N MILWAUKEE AVE	60647.0	NaN	NaN

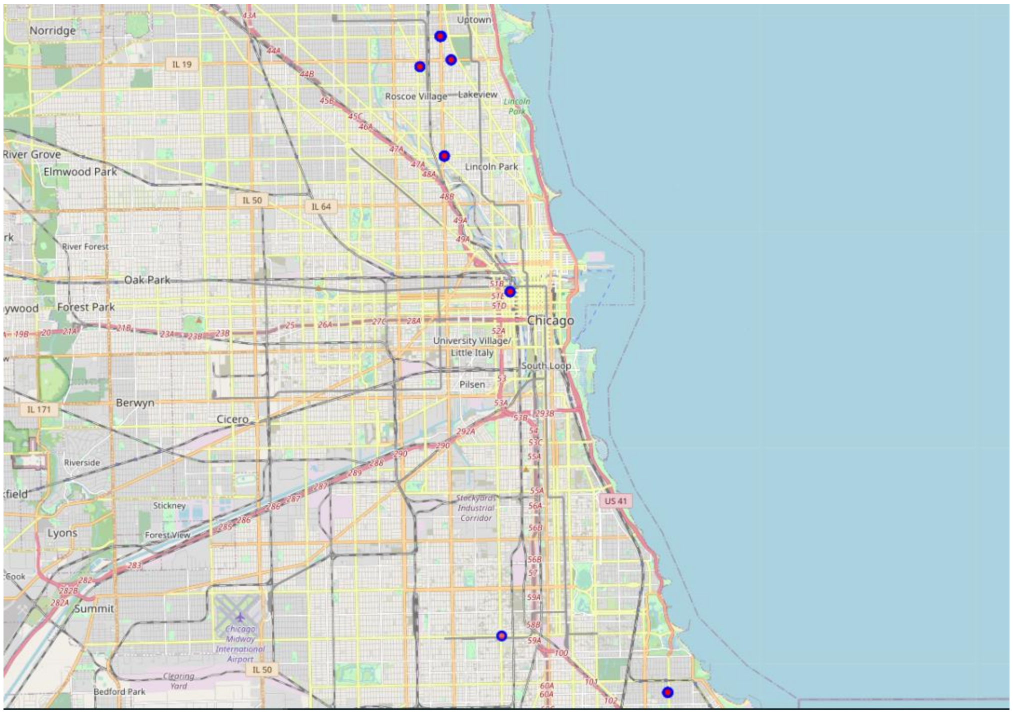
4. Clustering Vegan Restaurants in Chicago Neighborhoods:

We will be using k-means clustering. K-means is vastly used for clustering in many data science application. in real-world, k-means includes customer segmentation, pattern recognition and data comparison. Let’s begin with getting all the rows from the data frame which contains the word ‘VEGAN’ in their DBA Name. we should be getting something like this:

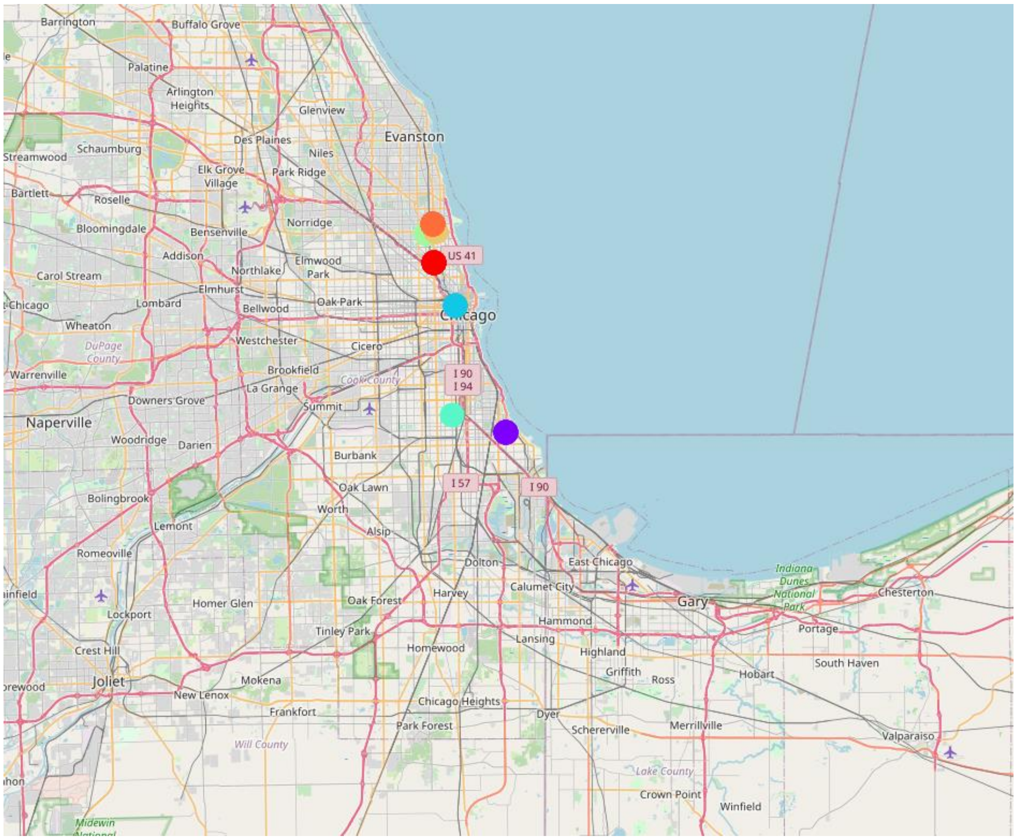


	DBA Name	Latitude	Longitude	Facility Type	Address	Zip	CASE#	BLOCK
1832	URBAN VEGAN	41.961481	-87.669334	Restaurant	1601-1603 W MONTROSE AVE	60613.0	NaN	NaN
6159	GOOD FOODS VEGAN / VEGETARIAN	41.762619	-87.576682	Restaurant	1966 E 73RD ST	60649.0	NaN	NaN
8162	VEGAN NOW INC	41.884188	-87.641120	Restaurant	131 N CLINTON ST	60661.0	NaN	NaN
9549	URBAN VEGAN	41.961478	-87.669492	Restaurant	1605 W MONTROSE AVE	60613.0	NaN	NaN
13971	URBAN VEGAN	41.961478	-87.669492	Restaurant	1605 W MONTROSE AVE	60613.0	NaN	NaN
14272	VEGAN PLATE	41.925266	-87.667793	Restaurant	1550 W FULLERTON AVE	60614.0	NaN	NaN
16045	URBAN VEGAN	41.961478	-87.669492	Restaurant	1605 W MONTROSE AVE	60613.0	NaN	NaN
20515	THE CHICAGO HOUSE OF 'ZA VEGAN PIZZERIA	41.952238	-87.677804	Restaurant	1939-1943 W BYRON ST	60613.0	NaN	NaN
20669	THE CHICAGO HOUSE OF 'ZA VEGAN PIZZERIA	41.952238	-87.677804	Restaurant	1939-1943 W BYRON ST	60613.0	NaN	NaN
22658	VEGAN PLATE	41.925266	-87.667793	Restaurant	1550 W FULLERTON AVE	60614.0	NaN	NaN
25070	URBAN VEGAN	41.961481	-87.669334	Restaurant	1601-1603 W MONTROSE AVE	60613.0	NaN	NaN
26977	URBAN VEGAN	41.961481	-87.669334	Restaurant	1601-1603 W MONTROSE AVE	60613.0	NaN	NaN
27073	THE CHICAGO HOUSE OF 'ZA VEGAN PIZZERIA	41.954455	-87.664917	Restaurant	1416 W IRVING PARK RD	60613.0	NaN	NaN
29633	THE CHICAGO HOUSE OF 'ZA VEGAN PIZZERIA	41.952238	-87.677804	Restaurant	1939-1943 W BYRON ST	60613.0	NaN	NaN
30718	THE CHICAGO HOUSE OF 'ZA VEGAN PIZZERIA	41.952238	-87.677804	Restaurant	1939-1943 W BYRON ST	60613.0	NaN	NaN
32780	VEGAN NOW INC	41.884188	-87.641120	Restaurant	131 N CLINTON ST	60661.0	NaN	NaN
32936	THE CHICAGO HOUSE OF 'ZA VEGAN PIZZERIA	41.954455	-87.664917	Restaurant	1416 W IRVING PARK RD	60613.0	NaN	NaN
33649	URBAN VEGAN	41.961478	-87.669492	Restaurant	1605 W MONTROSE AVE	60613.0	NaN	NaN
34636	URBAN VEGAN	41.961481	-87.669334	Restaurant	1601-1603 W MONTROSE AVE	60613.0	NaN	NaN
39903	VEGAN PLATE	41.925266	-87.667793	Restaurant	1550 W FULLERTON AVE	60614.0	NaN	NaN
40061	VEGAN PLATE	41.925266	-87.667793	Restaurant	1550 W FULLERTON AVE	60614.0	NaN	NaN

Now we can visualize the data on map to get better idea of where all the vegan restaurants are located.



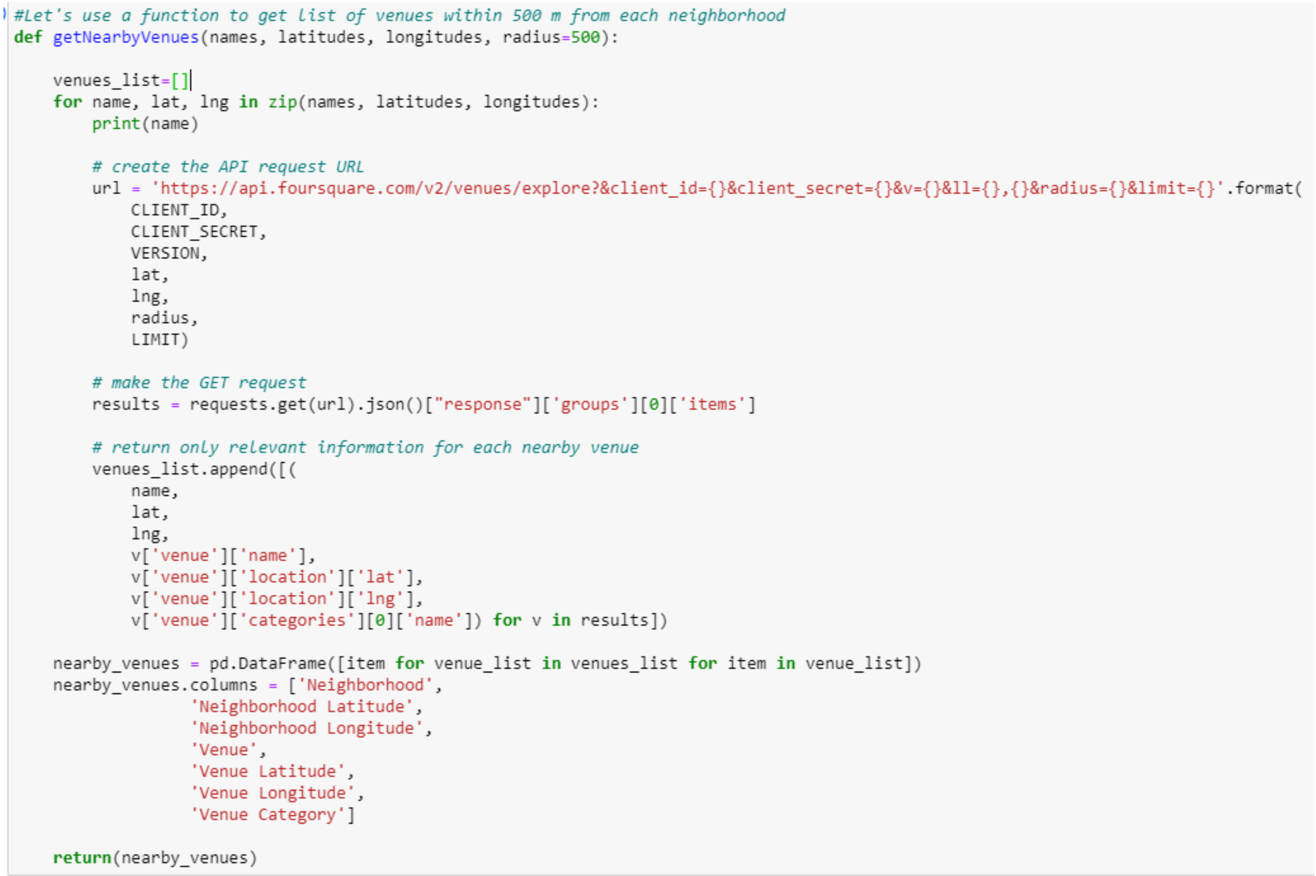
Next, we will use k-means clustering and set 8 clusters on our new vegan data frame.



5. Foursquare Data Analysis:

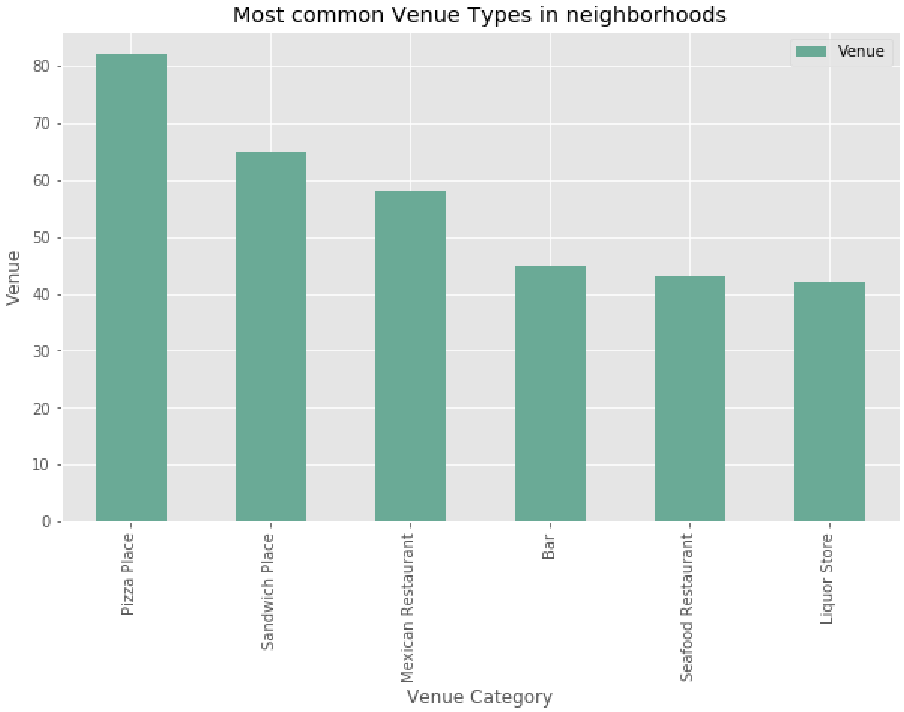
Foursquare is useful to send a request to API to search for a specific type of venue, to explore a geographical location, and to get trending venues around a location. We will use Foursquare API to retrieve information about the most popular sports/venues in each neighborhood in Chicago.

We can start by defining your Foursquare credentials and version. after that, we will write a function to get list of venues within 500 m from each neighborhood. It looks like there are 1520 entries for each neighborhood. The code should look like this:



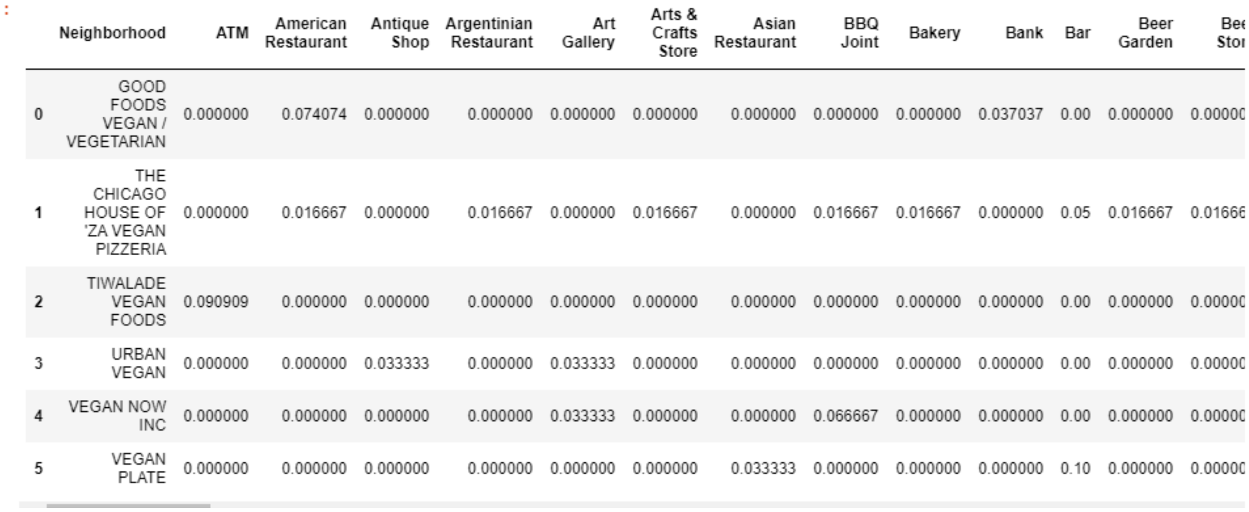
```
#Let's use a function to get list of venues within 500 m from each neighborhood
def getNearbyVenues(names, latitudes, longitudes, radius=500):

    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={}'.format(
            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 = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

We can see a graphic representation of the most popular venue categories across the neighborhoods using bar plot just the way we did earlier. The graph below counts the most frequently occurring popular venue types in the prioritized neighborhoods, sorted from most frequent to least. It seems like pizza places are most popular venue type. So clearly vegan restaurants are not the most popular venue in the neighborhoods, but maybe they are more popular in some other neighborhoods.



Once we look at the graphical representation, we will go in depth of each neighborhood to see the most popular types of venues for each neighborhood.

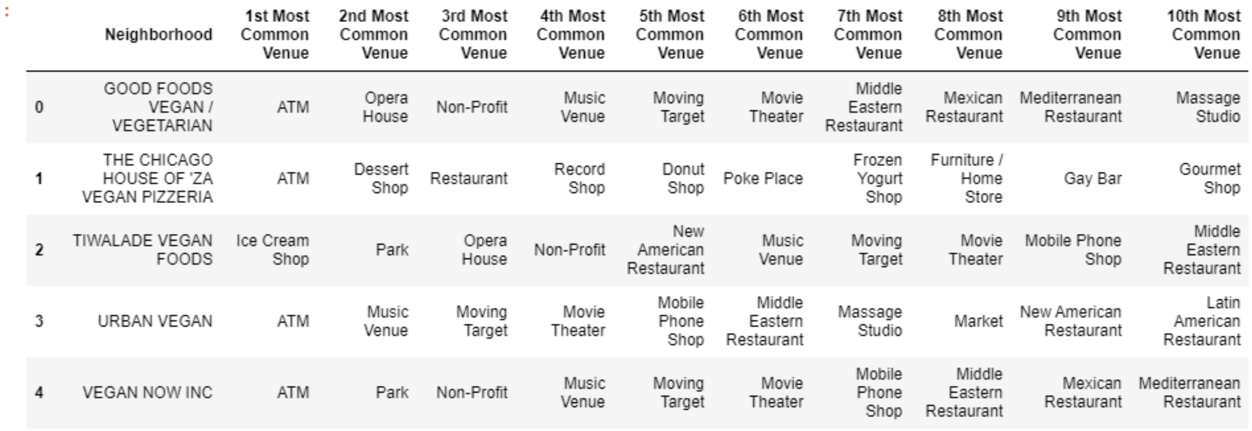
Now, we will create a data frame of venue categories with pandas one hot encoding and using Pandas group by to ger the mean of the one-hot encoded venue categories.



	Neighborhood	ATM	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bakery	Bank	Bar	Beer Garden	Bee Stor
0	GOOD FOODS VEGAN / VEGETARIAN	0.000000	0.074074	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.037037	0.00	0.000000	0.00000
1	THE CHICAGO HOUSE OF 'ZA VEGAN PIZZERIA	0.000000	0.016667	0.000000	0.016667	0.000000	0.016667	0.000000	0.016667	0.016667	0.000000	0.05	0.016667	0.01666
2	TIWALADE VEGAN FOODS	0.090909	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.00000
3	URBAN VEGAN	0.000000	0.000000	0.033333	0.000000	0.033333	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.00000
4	VEGAN NOW INC	0.000000	0.000000	0.000000	0.000000	0.033333	0.000000	0.000000	0.066667	0.000000	0.000000	0.00	0.000000	0.00000
5	VEGAN PLATE	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.033333	0.000000	0.000000	0.000000	0.10	0.000000	0.00000

We will transport the data frame and arrange it in descending order to return most common venues.

Now let’s create the new data frame and display the top 10 venues for each neighborhood to get better idea.



	Neighborhood	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	GOOD FOODS VEGAN / VEGETARIAN	ATM	Opera House	Non-Profit	Music Venue	Moving Target	Movie Theater	Middle Eastern Restaurant	Mexican Restaurant	Mediterranean Restaurant	Massage Studio
1	THE CHICAGO HOUSE OF 'ZA VEGAN PIZZERIA	ATM	Dessert Shop	Restaurant	Record Shop	Donut Shop	Poke Place	Frozen Yogurt Shop	Furniture / Home Store	Gay Bar	Gourmet Shop
2	TIWALADE VEGAN FOODS	Ice Cream Shop	Park	Opera House	Non-Profit	New American Restaurant	Music Venue	Moving Target	Movie Theater	Mobile Phone Shop	Middle Eastern Restaurant
3	URBAN VEGAN	ATM	Music Venue	Moving Target	Movie Theater	Mobile Phone Shop	Middle Eastern Restaurant	Massage Studio	Market	New American Restaurant	Latin American Restaurant
4	VEGAN NOW INC	ATM	Park	Non-Profit	Music Venue	Moving Target	Movie Theater	Mobile Phone Shop	Middle Eastern Restaurant	Mexican Restaurant	Mediterranean Restaurant

The above code provides us with the top 10 venues for each neighborhood.

6. Results and Discussion :

This data is important because it gives us an idea of the atmosphere of each neighborhood. As someone trying to open a vegan restaurant, I might want to know whether my location is already a hot spot for other bars and restaurants. It looks like most common venue for these vegan restaurants are ATMs, Opera House, Music Venue. So, if someone wants to open a vegan restaurant in Chicago or its neighborhood, they should consider opening it by Movie Theater or Market. Having restaurant by movie theater and marker would be a good deal because you know that people always go to these places and would love to hangout or have lunch or dinner after movie or after going to market.

7. Conclusion:

Finally, we have executed an end-to-end data science project using popular python libraries to manipulate data sets, clustering to get vegan restaurants, Foursquare API to explore the neighborhoods of Chicago, and Folium map to cluster and segment neighborhoods. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence. These analytical tools open a world of possibilities for strategic decision making across various business platforms.