Finding the best neighborhood in Chicago, IL to open a new Starbucks Shop

Mustafa Abdullayev

July 5, 2019

1. Introduction

Coffee is one of the most popular drink all over the world. From morning routines to sweet conversations with our friends, it is almost essential to our daily life. The United States imports in excess of 4 Billion worth of coffee per year. Americans consume 400 million cups of coffee per day making tje United States the leading consumer of soffee in the world. Independent coffee shops equal $12 billion in annual sales. At the present time there are approximately 24,000 Coffee Shops across the country. This statistics shows how vital coffee is for our daily routines and how perfect business opportunity. When talking about coffee it is ineviateble to mention one company - Starbucks. Starbucks has almost 30000 stores all over the world. Opening another store is very risky and its location must be chosen wisely. In this analysis, I will find best spots in Chicago, IL to open and operate new coffee shop. This analysis is for two types of business people :

1. Starbuck stakeholders - To find best spots for maximum profitability
2. Coffee businesspeople - To find best spots to escape competition with Starbucks and increase profit.

2. Data

For this problem I need 2 types of data :

1. Chicago Census Data.
2. Location data of Starbucks.

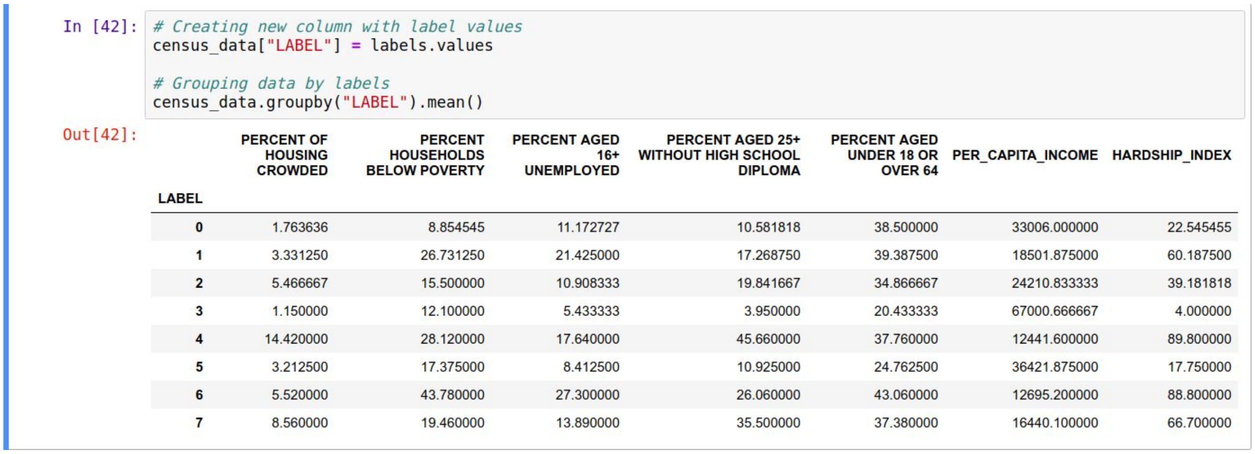
First of them is Chicago Census Data, which provided in this course before and I have this data locally. This data will be used to find and classify best neighborhoods of Chicago according to many factors such as Per Capita Income. Then I will choose these best community areas and will use second data to analyse further.

Secondly, to get location of Starbucks Coffee shops I will leverage Foursquare. I will use premiım calls to get Coffee shops around each neighborhood chosen in previos step, and will further analyse each place.

Finally, with these data I will determine which place is best to open a new Starbucks coffee Shop.

3. Methodology and Analysis

To achieve what I want, some essential libraries will be imported. Also, I will create my Foursquare agent.
Firstly, I will import the Chicago Census Data and use it to cluster Chicago neighborhoods. In order to classify our neighborhoods, we only need numeric data. So let's create a new dataframe with only numeric data of census_data.
Then I will check if there is any nan value. If so, with using Simple Imputer I will fill them. Now, since I will use K means clustering algorithm to cluster neighborhoods, it is absolutely essential to normalize our data. I will use Standart Scaler to do this.Now, lets cluster our neighborhoods with KMeans algorithm.Now, since we found a label of each neighborhoods, let's add these label dataframe to our original dataframe. Then we will find the best cluster in our dataframe and will continue with these neighborhoods.



```
In [42]: # Creating new column with label values
         census_data["LABEL"] = labels.values

         # Grouping data by labels
         census_data.groupby("LABEL").mean()
```

Out[42]:

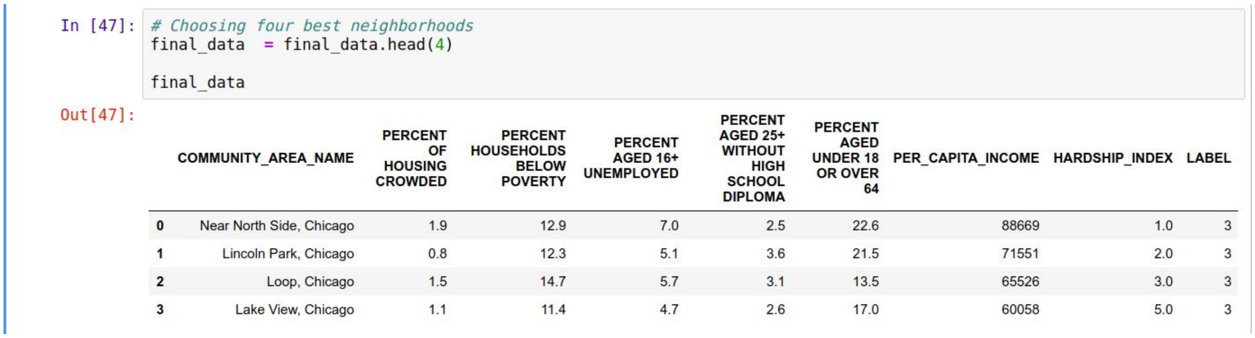
LABEL	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PERCENT AGED UNDER 18 OR OVER 64	PER_CAPITA_INCOME	HARDSHIP_INDEX
0	1.763636	8.854545	11.172727	10.581818	38.500000	33006.000000	22.545455
1	3.331250	26.731250	21.425000	17.268750	39.387500	18501.875000	60.187500
2	5.466667	15.500000	10.908333	19.841667	34.866667	24210.833333	39.181818
3	1.150000	12.100000	5.433333	3.950000	20.433333	67000.666667	4.000000
4	14.420000	28.120000	17.640000	45.660000	37.760000	12441.600000	89.800000
5	3.212500	17.375000	8.412500	10.925000	24.762500	36421.875000	17.750000
6	5.520000	43.780000	27.300000	26.060000	43.060000	12695.200000	88.800000
7	8.560000	19.460000	13.890000	35.500000	37.380000	16440.100000	66.700000

We observe that 3rd cluster is by far best cluster in terms of Per Capita Income, Hardship Index and so on. So I will consider only these neighborhoods in the 3rd cluster to analyze further

Best Cluster = 3rd Cluster

Now let's check how many neighborhoods are there in each cluster.

In order to get the location of each neighborhood without any error, I will add Chicago to each neighborhood. This is because there can be more than 1 place with the same neighborhood name. In order to Narrow down our Analysis, I will choose 4 best neighborhoods to analyze further. Since we found 4 best neighborhoods to analyze, It is time to use Foursquare to analyze further and vizualie them.



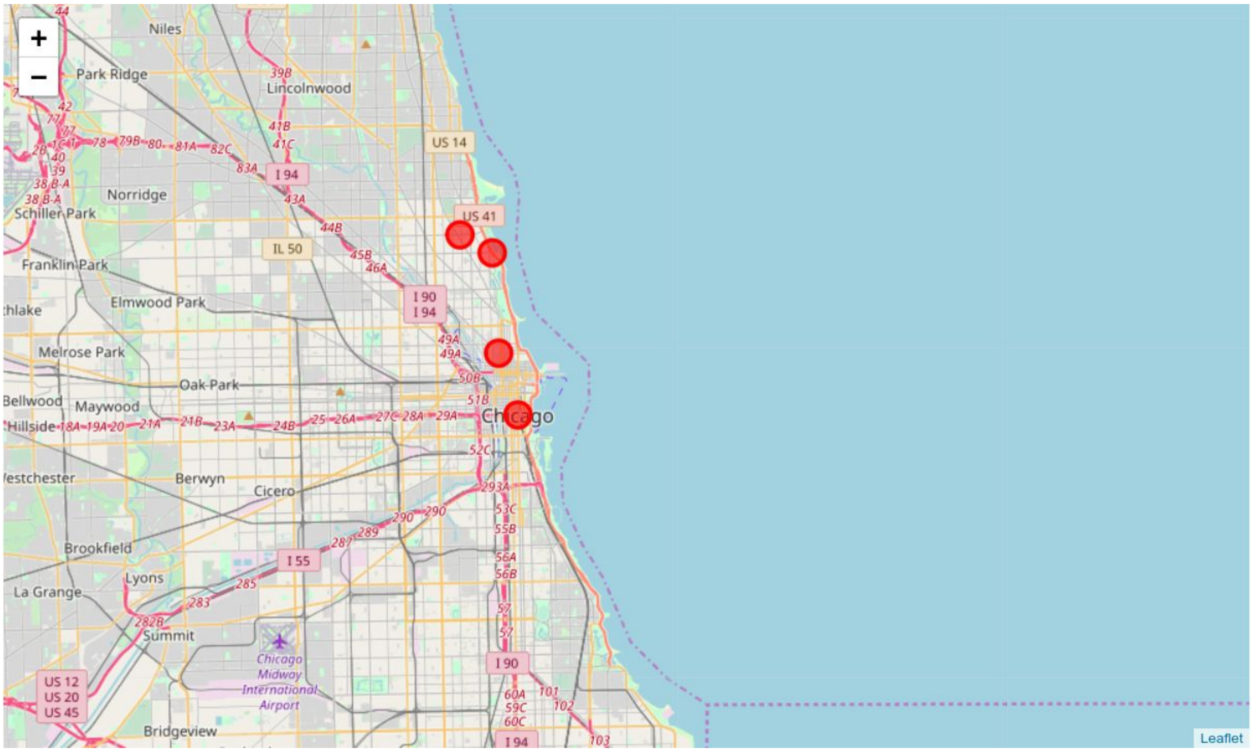
```
In [47]: # Choosing four best neighborhoods
         final_data  = final_data.head(4)

         final_data
```

Out[47]:

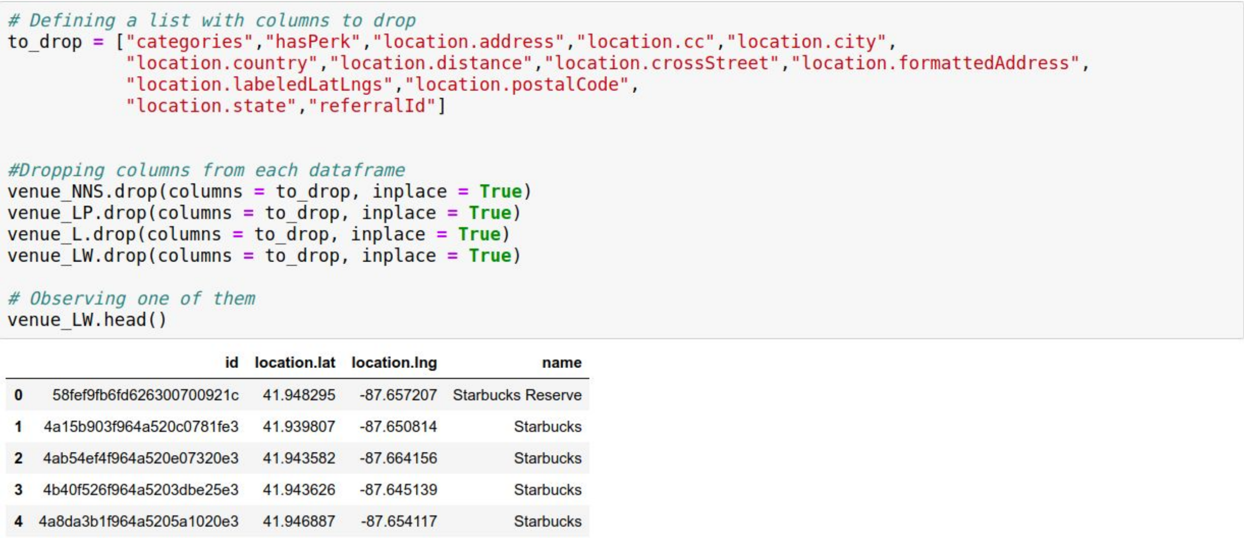
	COMMUNITY_AREA_NAME	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PERCENT AGED UNDER 18 OR OVER 64	PER_CAPITA_INCOME	HARDSHIP_INDEX	LABEL
0	Near North Side, Chicago	1.9	12.9	7.0	2.5	22.6	88669	1.0	3
1	Lincoln Park, Chicago	0.8	12.3	5.1	3.6	21.5	71551	2.0	3
2	Loop, Chicago	1.5	14.7	5.7	3.1	13.5	65526	3.0	3
3	Lake View, Chicago	1.1	11.4	4.7	2.6	17.0	60058	5.0	3

Now, I will use Foursquare to get the location of each neighborhood that I will use.
Then I will visualize them on the map.
Our user agent name will be capstone_agent. Now, Let's extract latitude and longitude of each neighborhood.
I will use abbreviations to represent neighborhood names. Let's visualize each neighborhood on the map.



Now let's search for Starbucks in each neighborhood. For this, I will create an URL for each neighborhood and then use request library to get JSON formatted data. Let's examine one of them to see what is in.We see that what we want is stored in : response - venues. So I will get data from these tags because they are all we need to analyze.

For each neighborhood I got JSON data, I will create new dataframe with the desired information. We see that the majority of columns are useless for my analysis and I will get rid of them



```
# Defining a list with columns to drop
to_drop = ["categories","hasPerk","location.address","location.cc","location.city",
           "location.country","location.distance","location.crossStreet","location.formattedAddress",
           "location.labeledLatLngs","location.postalCode",
           "location.state","referralId"]


#Dropping columns from each dataframe
venue_NNS.drop(columns = to_drop, inplace = True)
venue_LP.drop(columns = to_drop, inplace = True)
venue_L.drop(columns = to_drop, inplace = True)
venue_LW.drop(columns = to_drop, inplace = True)

# Observing one of them
venue_LW.head()
```

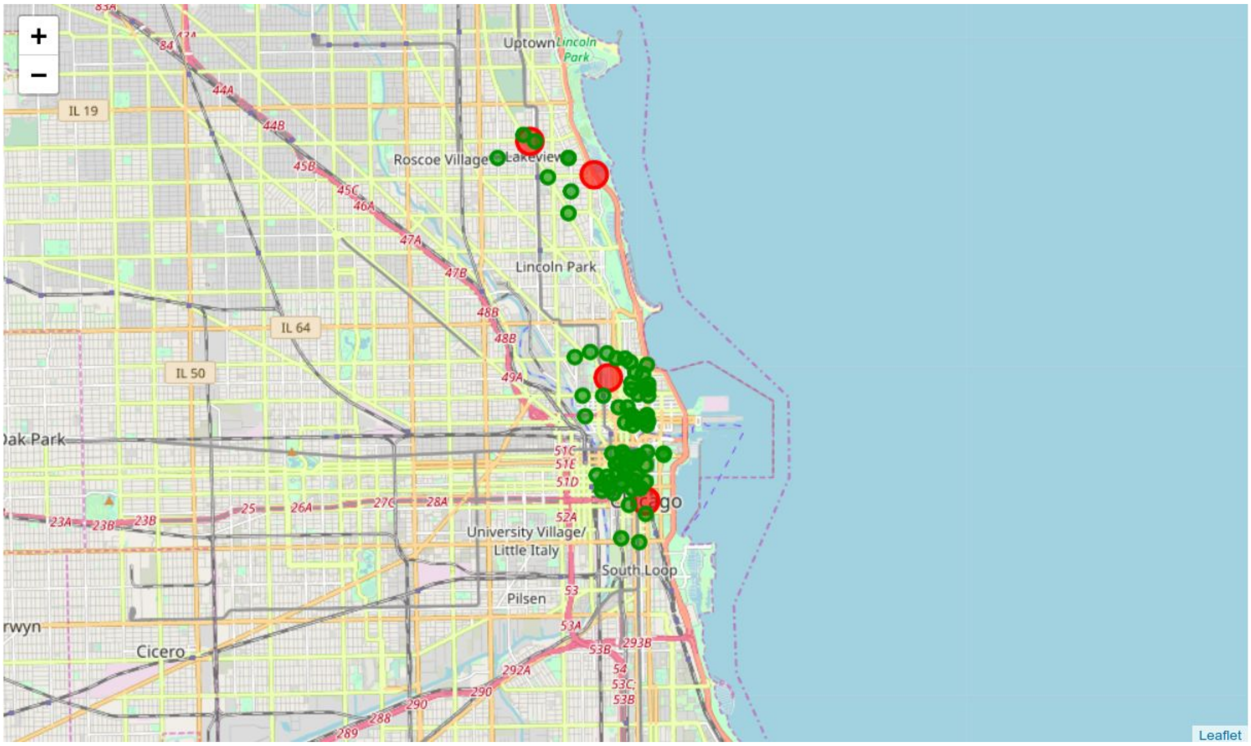
	id	location.lat	location.lng	name
0	58fef9fb6fd626300700921c	41.948295	-87.657207	Starbucks Reserve
1	4a15b903f964a520c0781fe3	41.939807	-87.650814	Starbucks
2	4ab54ef4f964a520e07320e3	41.943582	-87.664156	Starbucks
3	4b40f526f964a5203dbe25e3	41.943626	-87.645139	Starbucks
4	4a8da3b1f964a5205a1020e3	41.946887	-87.654117	Starbucks

Let's create new lists to get lattitude and longitude for each neighborhood in our dataframes. This will make easier to visualize each coffee shop in each neighborhood on the map. Visualize each spot on the map with neighborhood centers.

First, it is observed that Both of Near North Side and Loop community areas are very crowded whereas Lake View and Lincoln Park are good places to open a new Starbucks shop:

1. Crowded: Near North Side and Loop
2. Not Crowded: Lake View and Lincoln Park

But I am going to analyze further and use statistics to improve my analysis



In this part, I am going to get ratings of each Starbucks shop and analyse these ratings to determine which neighborhoods are in desperate need of new and better serving Shop. To do this I will use premium Foursquare calls:

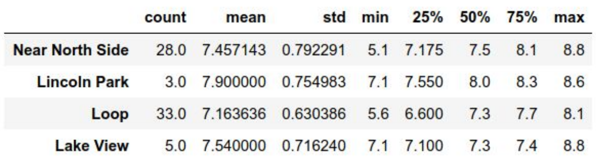
1. Create a list to store ratings
2. Get venue ids from neighborhoods dataframes
3. Use these ids to get information about specific places and store their ratings
4. Analyze ratings

Note: For any place without a rating, I will pass them. Ratings are stored in response-venue-rating

Let's check each dataframe with ratings with describe a method and gain some insights. To ease comparison I will create a new dataframe containing all of the data from individual neighborhoods dataframe describe a method. But First, let's see what describe method gives us

Now, I will use this method and get values. Then I will concatenate them for ease comparison

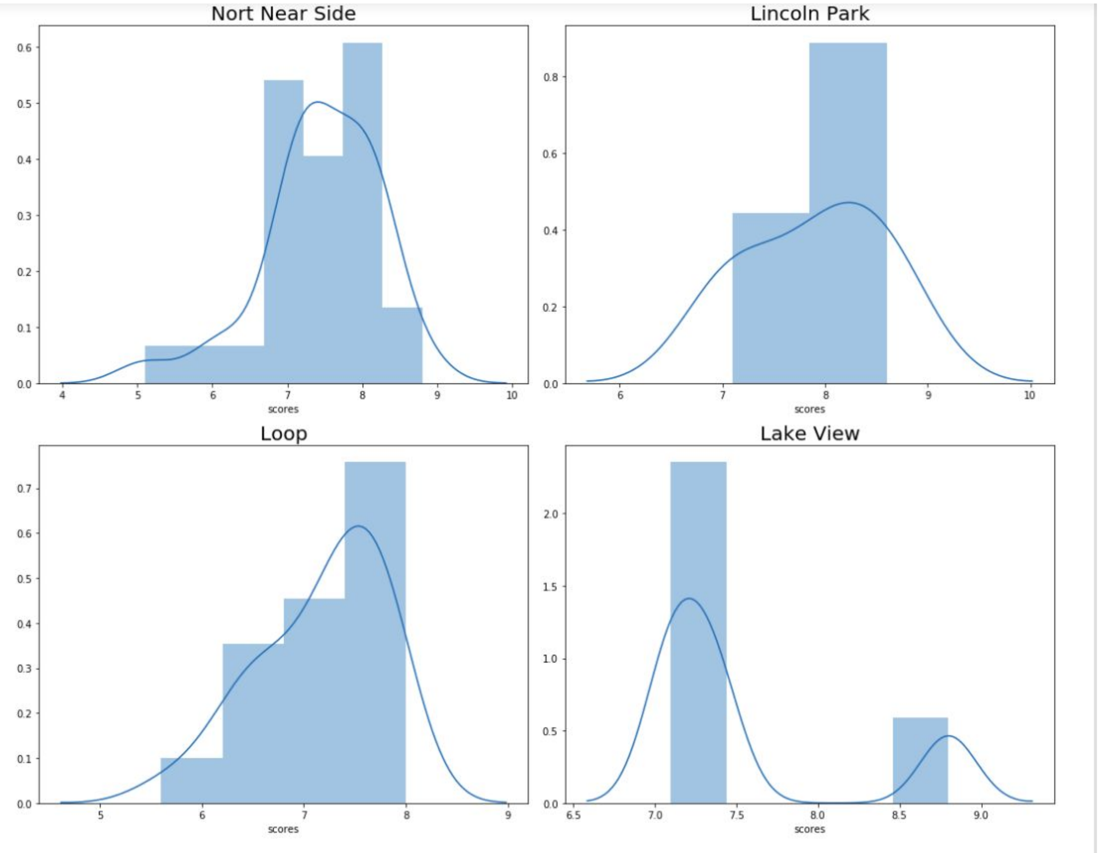
Lets concatanate them to one dataframe



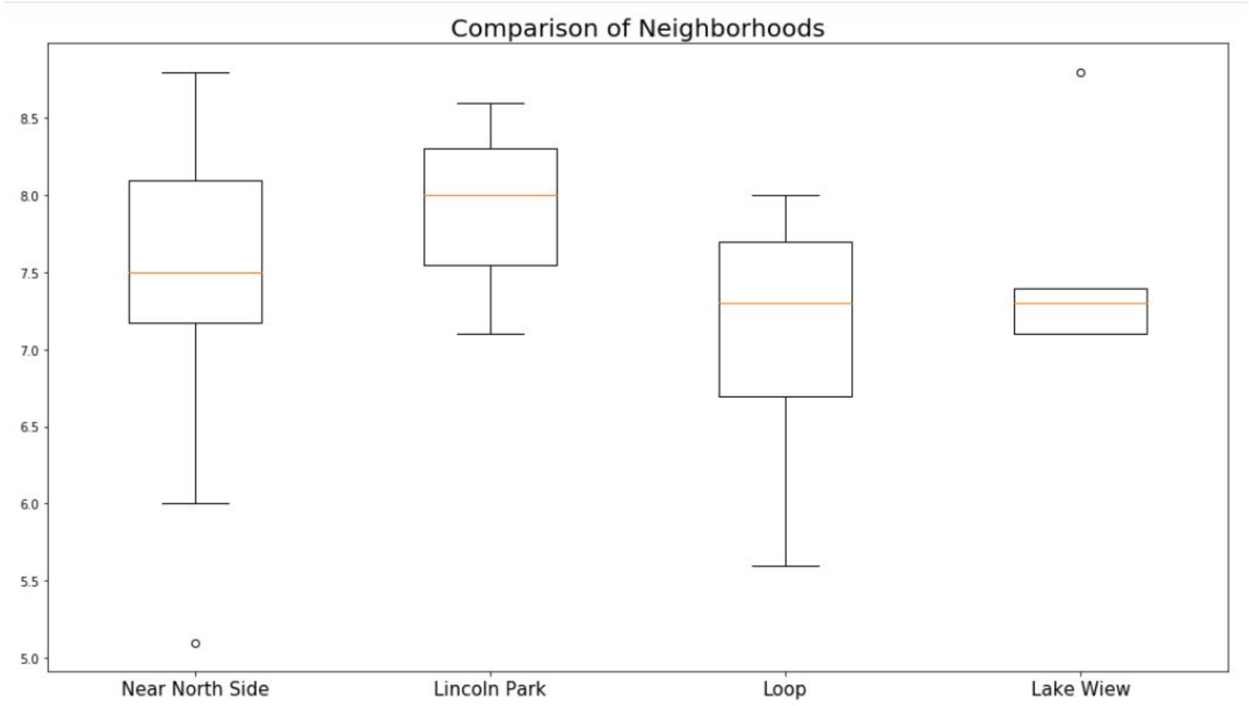
	count	mean	std	min	25%	50%	75%	max
Near North Side	28.0	7.457143	0.792291	5.1	7.175	7.5	8.1	8.8
Lincoln Park	3.0	7.900000	0.754983	7.1	7.550	8.0	8.3	8.6
Loop	33.0	7.163636	0.630386	5.6	6.600	7.3	7.7	8.1
Lake View	5.0	7.540000	0.716240	7.1	7.100	7.3	7.4	8.8

In here we observe again that, Near North is very crowded and ratings are high. On the other hand, Lincoln park is less crowded and the ratings are very high. Let's visualize further.

Lets Use Seaborn library to visualize ratings for each neighborhood



Now let's use boxplots to compare each of them and visualize at the same graph. For that, I will create a dictionary, Store values in appropriate keys and visualize them



4. Results and Discussion

After Analysing and visualizing my data, my conclusion is that **Lake View** is the best neighborhood to open a new Starbucks shop. My reasons are that :

1. It is not crowded
2. Rating scores are not very high.

The Loop on the other side is fairly crowded. When looking to average score it is very low but observing its distribution, median and quartile scores it is not as good as **Lake View**. There is some kind of need for a new better serving shop.

Thirdly, **Lincoln Park** has only 3 shops. It is definitely now crowded but the average rating is 7.9 which is very high. With a small amount of sample, it is hard to find a reason to open a new restaurant there.

Finally, I think the **Near North Side** is the worst place to operate. It is very crowded, has a decent average score, median and quartiles are high. Also, it is normally distributed. I think one must stay away from **Near North Side**

In my analysis , I used census data and Foursquare data to find crowded neighborhoods with Starbucks. But one thing I missed is that crowded is not necesseraly bad. There can be way more people living in one area and thus, Reducing per capita starbucks in final. For instance :

1. X neighbırhood has 50 Starbucks and 1 million population.
2. Y meighborhood has 5 starbucks and 20 k population.

At First it seems that X is overcrowded. But When we observe that per capita starbucks is equal to (to how many people one starbucks shop serve) 20k in X and only 4k in Y, it is reosanable to think X is not overcrowded.

But we did not have population data in our census data and I omittied this part.

Moreover, in general 2 of our neighborhoods has fewer than 6 starbucks and that is very little data to analyse properly. But again since that is all we have, I have done my analysis with that.

5. Conclusion

Purpose of this project was to identify the best spot in Chicago neighborhoods to open a new Starbucks shop. Firstly I used the KMeans clustering algorithm to cluster neighborhoods in order to find the best of them. Then I leveraged Foursquare to get shop detail and their ratings to analyze. Finally Using data visualization methods and Statistical inferences to find optimal spots to open a new Starbuck shop.

Also, I discussed possible flaws in my analysis such as data scarcity and missing of population data.

The final decision on optimal shop location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like shop ratings, neighborhood attractiveness possible impact of missing data and etc.