Capstone Project-A Study on small business in Buffalo

06/24/2020

1 Introduction

1.1 Background and business problem

Buffalo is the second largest city in New York State, as well as the largest city in Western New York [1]. Say you live in Buffalo and plan to do some investment or start your own business. It is necessary to do some research on what kind of business is more likely to survive in this area. There are so many factors to be considered such as funding, accessibility (traffic), labor cost, *etc.*, however, it is impractical to run an analysis for each factor here. Herein, we only study which district in Buffalo is a better choice for starting or investing a business, and the types of business that are popular in this district.

1.2 Objective

In this study, we will figure out:

2 which district is likely to have more potential customers;

3 how many different types of venues there are in this district;

4 what the most popular venues in this area are.

Our target audience are entrepreneurs or business owners who want to start a new business or invest in the existing businesses in Buffalo.

2 Data preparation

Data sources used in the analysis are listed below.

The population data in each district of Buffalo: https://www.zip-codes.com/city/ny-buffalo.asp.

The zip code latitude and longitude data of Buffalo:

https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/?q=buffalo.

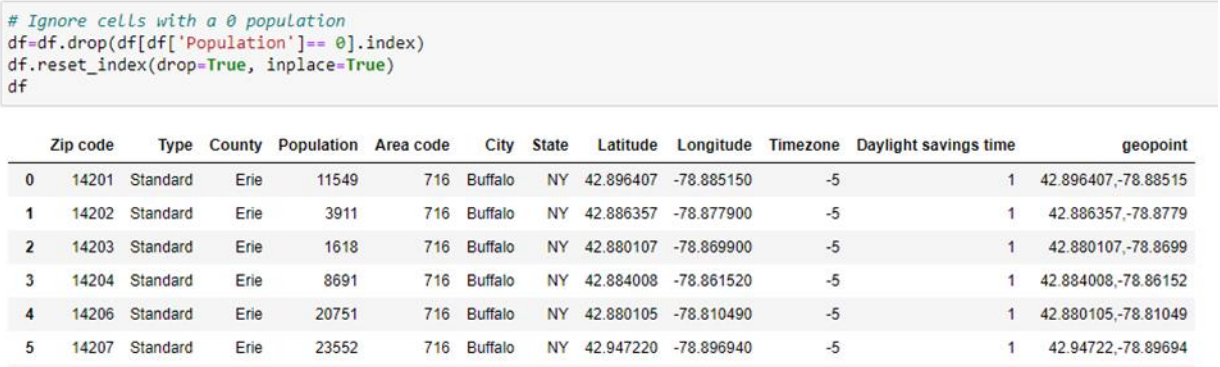
Both forms were exported as csv files.

The information on venues in Buffalo were obtained by utilizing Foursquare API.

3 Methodology

First of all, all the population data and the geopoint data of Buffalo were pulled from csv files to create dataframes. They were all grouped by zip code.

Then, we merged the population data and geopoint data into one dataframe based on zip code. There are 44 zip codes in Buffalo. It is obvious that those districts with a zero population are useless in this study, so these cells were deleted.



```
# Ignore cells with a 0 population
df=df.drop(df[df['Population']== 0].index)
df.reset_index(drop=True, inplace=True)
df
```

	Zip code	Type	County	Population	Area code	City	State	Latitude	Longitude	Timezone	Daylight savings time	geopoint
0	14201	Standard	Erie	11549	716	Buffalo	NY	42.896407	-78.885150	-5	1	42.896407,-78.88515
1	14202	Standard	Erie	3911	716	Buffalo	NY	42.886357	-78.877900	-5	1	42.886357,-78.8779
2	14203	Standard	Erie	1618	716	Buffalo	NY	42.880107	-78.869900	-5	1	42.880107,-78.8699
3	14204	Standard	Erie	8691	716	Buffalo	NY	42.884008	-78.861520	-5	1	42.884008,-78.86152
4	14206	Standard	Erie	20751	716	Buffalo	NY	42.880105	-78.810490	-5	1	42.880105,-78.81049
5	14207	Standard	Erie	23552	716	Buffalo	NY	42.947220	-78.896940	-5	1	42.94722,-78.89694

With the population data in each district of Buffalo obtained, it is our aim to find a place where there are more potential customers, so the larger population, the better. Herein, we chose only the districts with a population larger than 25,000.

Next, we utilized the Foursquare API to explore the venues in selected districts. I designed the limit as 100 venues and the radius 2000 meters for each district from their given latitude and longitude information.

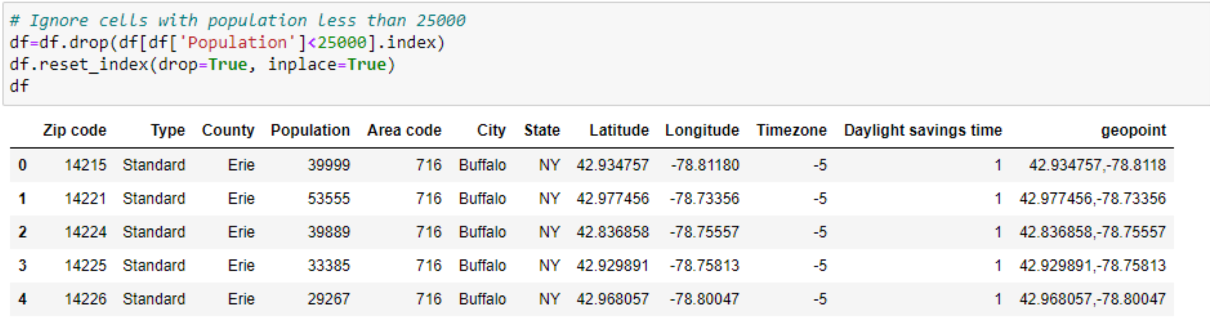
To analyze the data, we used one hot encoding (it helps to transform categorical data into numerical data) to group data, and find out the top ten venues present in each district.

Finally, K-Means Clustering was applied to cluster these selected districts based on the venue categories.

4 Results and discussion

4.1 Select the districts suitable for small businesses in Buffalo

As Figure 1 shows, there are only five districts in which the population is larger than 25,000. Among them, population in 14221 (zip code) district is the largest. Throughout this report, zip codes are used to represent the districts they are corresponding to.



```
# Ignore cells with population less than 25000
df=df.drop(df[df['Population']<25000].index)
df.reset_index(drop=True, inplace=True)
df
```

	Zip code	Type	County	Population	Area code	City	State	Latitude	Longitude	Timezone	Daylight savings time	geopoint
0	14215	Standard	Erie	39999	716	Buffalo	NY	42.934757	-78.81180	-5	1	42.934757,-78.8118
1	14221	Standard	Erie	53555	716	Buffalo	NY	42.977456	-78.73356	-5	1	42.977456,-78.73356
2	14224	Standard	Erie	39889	716	Buffalo	NY	42.836858	-78.75557	-5	1	42.836858,-78.75557
3	14225	Standard	Erie	33385	716	Buffalo	NY	42.929891	-78.75813	-5	1	42.929891,-78.75813
4	14226	Standard	Erie	29267	716	Buffalo	NY	42.968057	-78.80047	-5	1	42.968057,-78.80047

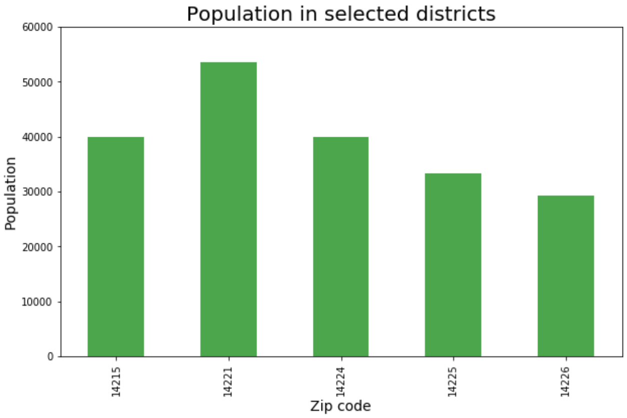


Figure 1. The population data in the five selected districts in Buffalo

We also generated a map of Buffalo and located all the five selected districts on it (blue dots) as shown in Figure 2.

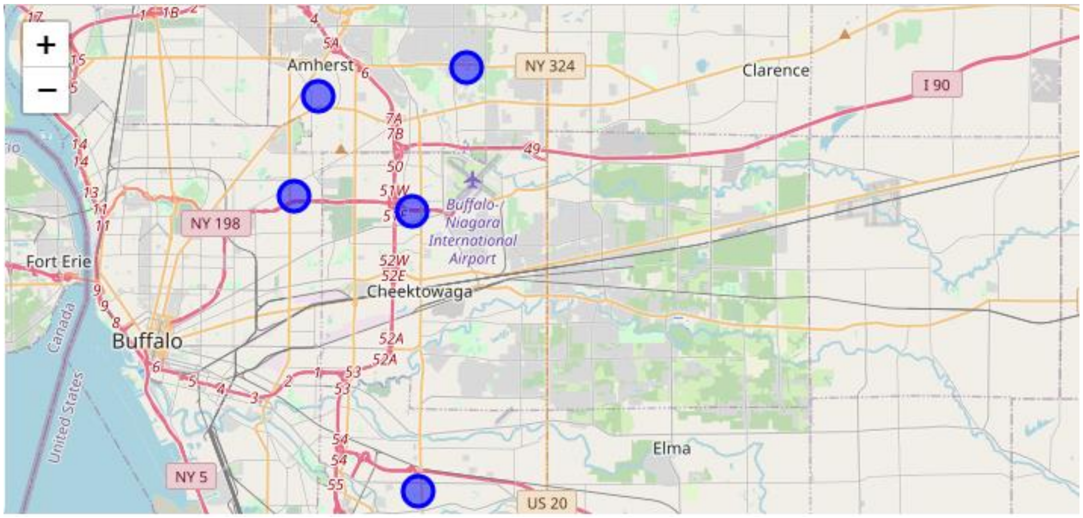
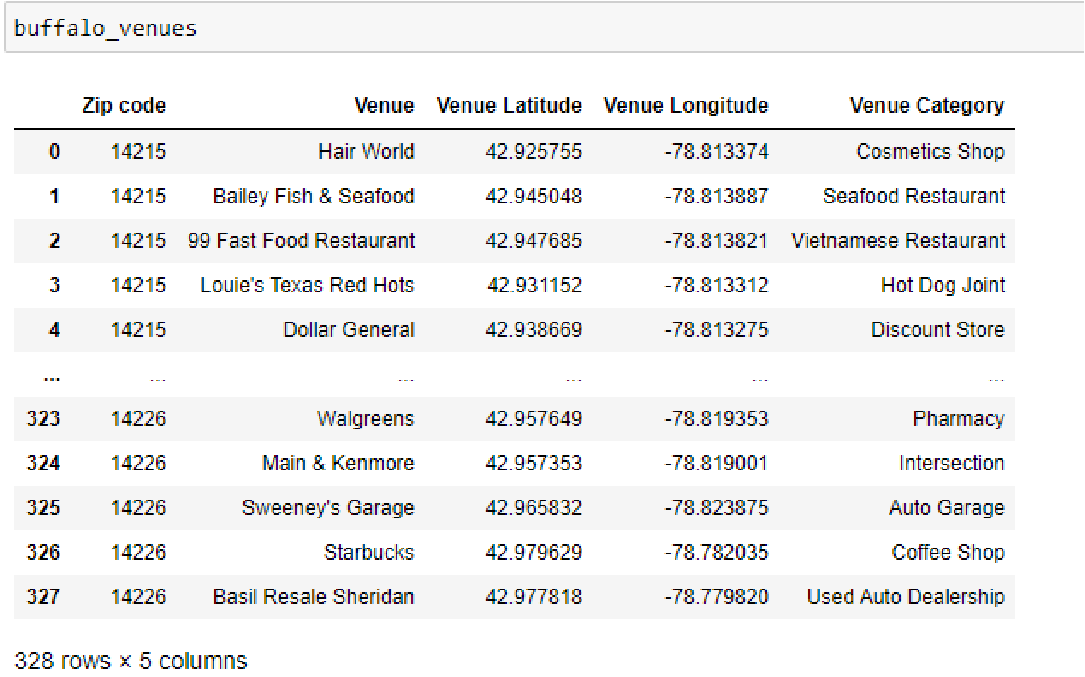


Figure 2. Buffalo map

4.2 Explore the venues in each district

By utilizing the Foursquare API, a function to look up the venues was defined and the venues in each district were obtained (includes the venue name, latitude, longitude and category data). The information of 328 venues were extracted, and the total number of unique categories in these selected areas is 130.



```
buffalo_venues
```

	Zip code	Venue	Venue Latitude	Venue Longitude	Venue Category
0	14215	Hair World	42.925755	-78.813374	Cosmetics Shop
1	14215	Bailey Fish & Seafood	42.945048	-78.813887	Seafood Restaurant
2	14215	99 Fast Food Restaurant	42.947685	-78.813821	Vietnamese Restaurant
3	14215	Louie's Texas Red Hots	42.931152	-78.813312	Hot Dog Joint
4	14215	Dollar General	42.938669	-78.813275	Discount Store
...	...	...	...	...	...
323	14226	Walgreens	42.957649	-78.819353	Pharmacy
324	14226	Main & Kenmore	42.957353	-78.819001	Intersection
325	14226	Sweeney's Garage	42.965832	-78.823875	Auto Garage
326	14226	Starbucks	42.979629	-78.782035	Coffee Shop
327	14226	Basil Resale Sheridan	42.977818	-78.779820	Used Auto Dealership

328 rows × 5 columns

The number of each unique category was counted and assorted. The ten most frequently occurring venues in these districts are shown in Figure 3. The top three venue categories are coffee shops, sandwich places and pizza places.

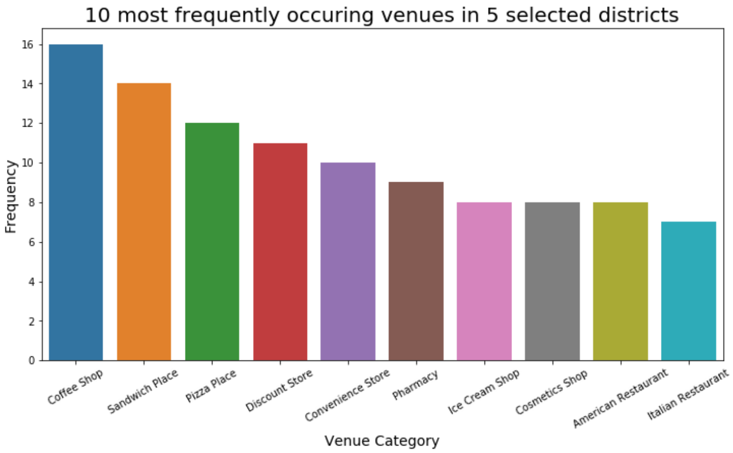


Figure 3. The 10 most frequently occurring venues in selected districts in Buffalo

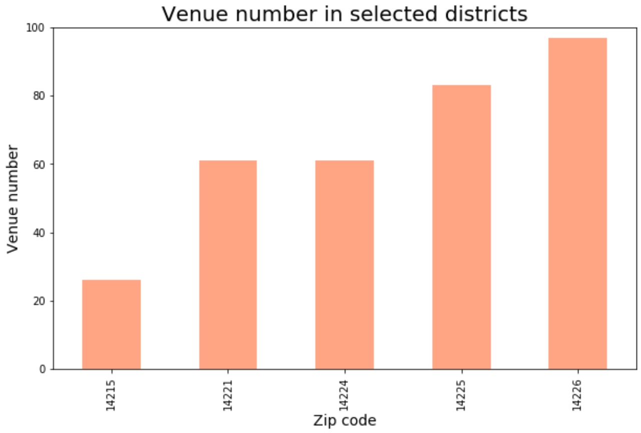


Figure 4. The venue numbers in each selected districts in Buffalo

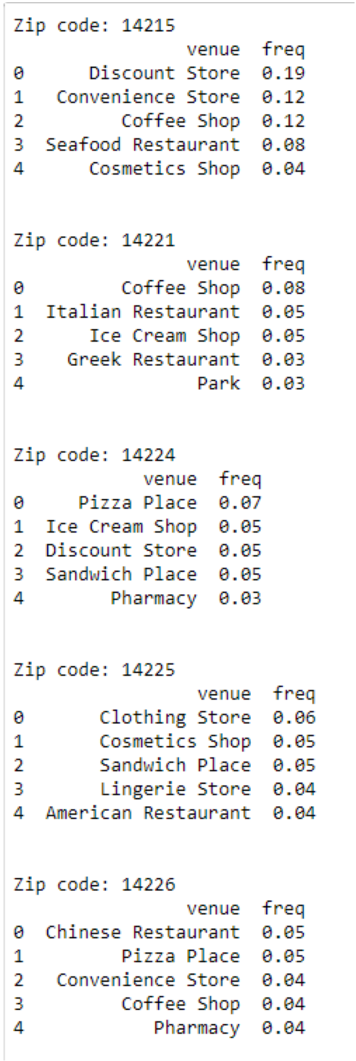
The number of venues in each district was counted. Although 14225 and 14226 districts were ranked as the fourth and fifth in population (Figure 1), the venue number in these two districts were higher than the other three districts. It is important to figure out what types of venues are popular in each district. Are they also the coffee shops, sandwich places, which are the most frequently occurring venues under the condition that all the five districts were treated as a whole?

To answer this question, we created a dataframe with pandas one hot encoding for the venue categories, which helps to transform categorical data into numerical data.

We found out the top five venues of each district.

In district 14221, which is more likely to have more potential customers, coffee shops, Italian restaurants and ice cream shops are more popular.

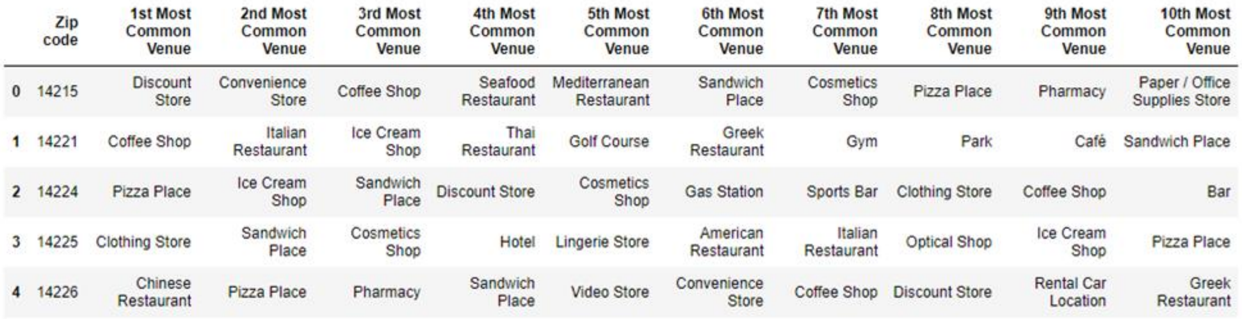
In district 14225, clothing stores, sandwich places and cosmetics shops are more popular; in district 14226, Chinese restaurants and pizza places are more popular.



```
Zip code: 14215
                venue  freq
0      Discount Store  0.19
1   Convenience Store  0.12
2         Coffee Shop  0.12
3  Seafood Restaurant  0.08
4      Cosmetics Shop  0.04


Zip code: 14221
                venue  freq
0         Coffee Shop  0.08
1  Italian Restaurant  0.05
2      Ice Cream Shop  0.05
3    Greek Restaurant  0.03
4                Park  0.03


Zip code: 14224
            venue  freq
0     Pizza Place  0.07
1  Ice Cream Shop  0.05
2  Discount Store  0.05
3  Sandwich Place  0.05
4        Pharmacy  0.03


Zip code: 14225
                 venue  freq
0       Clothing Store  0.06
1       Cosmetics Shop  0.05
2       Sandwich Place  0.05
3       Lingerie Store  0.04
4  American Restaurant  0.04


Zip code: 14226
                venue  freq
0  Chinese Restaurant  0.05
1         Pizza Place  0.05
2   Convenience Store  0.04
3         Coffee Shop  0.04
4            Pharmacy  0.04
```

4.3 K-Means Clustering

We have some common venue categories in these districts. We use the K-Means clustering technique to cluster the districts.



	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	14215	Discount Store	Convenience Store	Coffee Shop	Seafood Restaurant	Mediterranean Restaurant	Sandwich Place	Cosmetics Shop	Pizza Place	Pharmacy	Paper / Office Supplies Store
1	14221	Coffee Shop	Italian Restaurant	Ice Cream Shop	Thai Restaurant	Golf Course	Greek Restaurant	Gym	Park	Café	Sandwich Place
2	14224	Pizza Place	Ice Cream Shop	Sandwich Place	Discount Store	Cosmetics Shop	Gas Station	Sports Bar	Clothing Store	Coffee Shop	Bar
3	14225	Clothing Store	Sandwich Place	Cosmetics Shop	Hotel	Lingerie Store	American Restaurant	Italian Restaurant	Optical Shop	Ice Cream Shop	Pizza Place
4	14226	Chinese Restaurant	Pizza Place	Pharmacy	Sandwich Place	Video Store	Convenience Store	Coffee Shop	Discount Store	Rental Car Location	Greek Restaurant

Finally, these five districts were clustered into three clusters based on the similarities of venue categories. Districts 14224, 14225 and 14226 are in the same cluster (cluster 0), 14215 is cluster 1, 14221 is cluster 2. The matplotlib and folium packages were used to visualize the three clusters (green, red and purple markers) on a map of Buffalo.

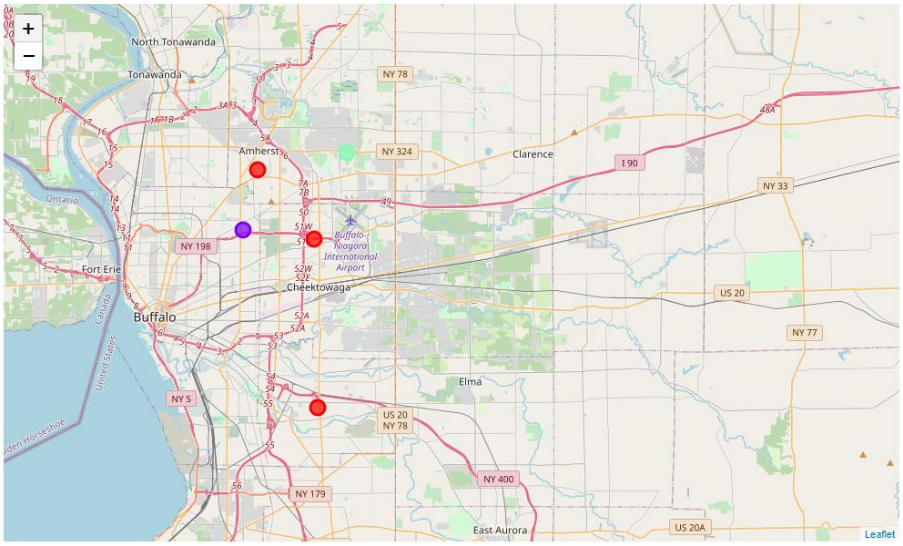


Figure 5. The clusters on the map of Buffalo

5 Conclusion

Deciding the type and location of a business to run (or invest) in a big city is a complex problem and many factors need considering. In this report, some frequently used python libraries and Foursquare API were used to scrap data from the Internet, to explore the venues in a few major districts of Buffalo. Here a few conclusions were drawn to solve the question that which district in Buffalo is a better choice for starting or investing a business, and the types of business that are popular in this district:

- Overall, coffee shops, sandwich places and pizza places are the three most popular venues in five major districts (14215, 14221, 14224, 14225 and 14226) in Buffalo.
- District 14221 is more likely to have more potential customers, and coffee shops, Italian restaurants and ice cream shops are more popular in this area.
- District 14225 and 14226 has more venues survived. In district 14225, clothing stores, sandwich places and cosmetics shops are more popular; in district 14226, Chinese restaurants and pizza places are more popular.

References

[1] https://en.wikipedia.org/wiki/Buffalo,_New_York