IBM Data Science Certificate Program

Capstone project

Analysis of business opportunities for new childcare facilities in the city of Chicago, Illinois

Author: Elina Vitol

July 2020

TABLE OF CONTENTS

1. Introduction/ Business problem

Childcare facilities for children aged 0 – 5 years old in the United States are private businesses. Exceptions include part-time facilities which offer morning or afternoon care, typically in a pre-school setting provided by school districts.

Some of the daycares offer franchise opportunities, for example Goddard School and Kiddie Academy [1-2]. At the same time, there are a lot of independent daycare facilities which cater to specific childcare needs in selected neighborhoods. Opening a new daycare business requires a thorough analysis of the existing competition as well as the analysis of socioeconomic situation in the given neighborhood. Running a daycare is considered an attractive opportunity for people looking to own a business, providing the much-needed options for the childcare needs for families with small children [3].

The city of Chicago, being the largest city in the Midwest United States, with the estimated population of 2.7 million [4], serves as a home to many different companies within the city limits. As a result, there is a large number of working parents looking for childcare arrangements in the Chicago area. According to the census data, there are approximately 175,000 children under the age of 5 in Chicago [4].

The goal of this project is to create a recommendation on which of the neighborhoods in the city of Chicago would be attractive for starting a new daycare business. The target audience for this project is people looking to start a new daycare business in Chicago.

2. Data Sources

Foursquare API [5] will be used for analyzing the neighborhood data in Chicago. There is a specific venue category for daycares (4f4532974b9074f6e4fb0104). Specifying the venue type will help to narrow down the results.

We will explore the number of daycare facilities in the city of Chicago per neighborhood as well as the type of the facility. In addition to the Foursquare data, we will analyze the data provided by the city of Chicago about the existing early education programs [6].

Information about specific neighborhoods in the city, including neighborhood names and corresponding zip codes is provided by the city of Chicago [7]. This data will be used for creating a map of existing childcare facilities and linking it to the socioeconomic data. The latter will be used as an indicator for the daycare price point opportunities in a given area, i.e. low income vs high income neighborhoods. The socioeconomic information, including per capita income per neighborhood is available from the Chicago census data [8].

3. References

1. https://kiddieacademy.com/franchising/
2. https://www.goddardschoolfranchise.com/franchise-cost.html
3. https://smallbiztrends.com/2017/11/child-care-franchise.html
4. https://www.census.gov/quickfacts/fact/table/chicagocityillinois,US/PST045219
5. https://developer.foursquare.com/
6. https://data.cityofchicago.org/Education/Chicago-Early-Learning-Programs-Map/2kih-a5ex
7. https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Chicago-Zip-Code-and-Neighborhood-Map/mapn-ahfc
8. https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

4. Methodology

4.1 Extracting data about daycares in Chicago from Foursquare

I used QGIS open source software in order to create the dataframe which contains the neighborhood names along with the corresponding coordinates (https://qgis.org/en/site/). The

original file provided by the city of Chicago had the polygon coordinates rather than the coordinates of neighborhoods centers. The latter is needed in order to enable venue search on the Foursquare API. Below is the example of the first two rows of the resulting data frame for Chicago neighborhoods.

	Neighborhoods	Longitude	Latitude
0	Grand Boulevard	-87.617860	41.812949
1	Printers Row	-87.629035	41.870981

To start the venue search using Foursquare API, it is required to define access credentials (client ID, client secret), the API version (today's date) and the search parameters. The parameters include the radius of search with respect to the given coordinates, the search limit (maximum number of venues to be returned by Foursquare for a given set of geographic coordinates) and the coordinates. First, I ran the unrestricted search for venues in all Chicago neighborhoods, without specifying the venue category. An excerpt from the resulting dataframe is shown below.

```
chicago_venues.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Grand Boulevard	41.812949	-87.61786	Parkway Ballroom	41.813142	-87.616064	Food
1	Grand Boulevard	41.812949	-87.61786	Peach's Restaurant	41.809481	-87.617009	Breakfast Spot
2	Grand Boulevard	41.812949	-87.61786	Blues Brothers Mural / Shelly's Loan & Jewelry...	41.809391	-87.619517	Plaza
3	Grand Boulevard	41.812949	-87.61786	Chicago Blues District	41.810071	-87.614105	Jazz Club
4	Grand Boulevard	41.812949	-87.61786	Harold Washington Cultural Center	41.809395	-87.616302	Performing Arts Venue

Next, I searched for the daycare venues within these results.

```
df_daycare = chicago_venues[chicago_venues['Venue Category'].str.contains('Daycare')]
```

```
df_daycare.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category

Searching for daycares within 2687 venues found by the Foursquare returns zero results. This is most likely due to the fact that daycares are not the most popular types of venues on the Foursquare.

Next, I restricted the category ID specifically to the daycare. The specific category number for daycares is '4f4532974b9074f6e4fb0104', according to the Foursquare website.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Printers Row	41.870981	-87.629035	It Takes A Village	41.872199	-87.633867	School
1	Sheffield & DePaul	41.927188	-87.653670	Tiny Tots Incorporated	41.929380	-87.648550	Daycare
2	Hermosa	41.924348	-87.734740	Kidslife Daycare Center	41.922340	-87.737204	School
3	Avondale	41.938666	-87.711211	A-Karrasel Child Care	41.936617	-87.707797	Daycare
4	Avondale	41.938666	-87.711211	Kat Slawson	41.935631	-87.715263	Daycare

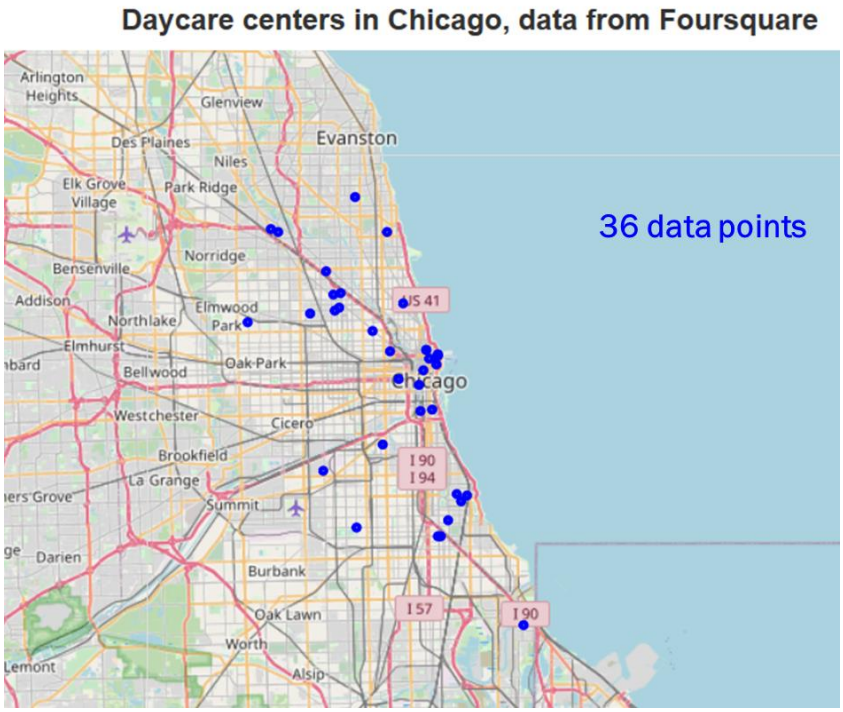
```
chicago_venues2.shape
```

```
(39, 7)
```

Venue category had to specified to find daycares on Foursquare. 39 results were returned, 3 had to be excluded since they were erroneously categorized as daycares. Excerpt from resulting dataframe is shown below.

	COMMUNITY AREA NAME	Neighborhood Latitude	Neighborhood Longitude	Site Name	Latitude	Longitude	Venue Category
0	Printers Row	41.870981	-87.629035	It Takes A Village	41.872199	-87.633867	School
1	Sheffield & DePaul	41.927188	-87.653670	Tiny Tots Incorporated	41.929380	-87.648550	Daycare
2	Hermosa	41.924348	-87.734740	Kidslife Daycare Center	41.922340	-87.737204	School
3	Avondale	41.938666	-87.711211	A-Karrasel Child Care	41.936617	-87.707797	Daycare
4	Avondale	41.938666	-87.711211	Kat Slawson	41.935631	-87.715263	Daycare
5	Logan Square	41.923193	-87.707389	Christopher House	41.926334	-87.709773	Daycare
6	Logan Square	41.923193	-87.707389	A-Karrasel Child Care	41.924895	-87.712903	Daycare
7	East Side	41.707314	-87.534902	Shinning Star	41.703346	-87.535548	Daycare
8	Grand Crossing	41.763247	-87.616134	Jellybean Learning Center II	41.765854	-87.615811	Daycare
9	Grand Crossing	41.763247	-87.616134	Allison's Infant &Toddler Center	41.765966	-87.613692	Daycare
10	Loop	41.880052	-87.626993	Bright Horizons at Cook County/City of Chicago...	41.882722	-87.629767	Daycare
11	Magnificent Mile	41.894784	-87.624188	Butler Children's Prep	41.897270	-87.627411	Daycare
12	Magnificent Mile	41.894784	-87.624188	catherineschildren	41.891188	-87.624625	Daycare
13	West Loop	41.877690	-87.648721	Lily Pad Nursery + Preschool	41.876259	-87.652870	Daycare
14	Andersonville	41.979854	-87.667865	Mi Casita Chicago	41.979643	-87.664005	Daycare
15	Woodlawn	41.778787	-87.601686	Busy bumble bee daycare Academy	41.777096	-87.606196	Nursery School

Next, I plotted the Foursquare data on the Chicago map using Folium library.



By looking at the map, one can easily see that the information about daycares, obtained from the Foursquare, is very limited. In order to perform a more detailed analysis of the daycares in Chicago, we will supplement this dataset with the data from the city of Chicago.

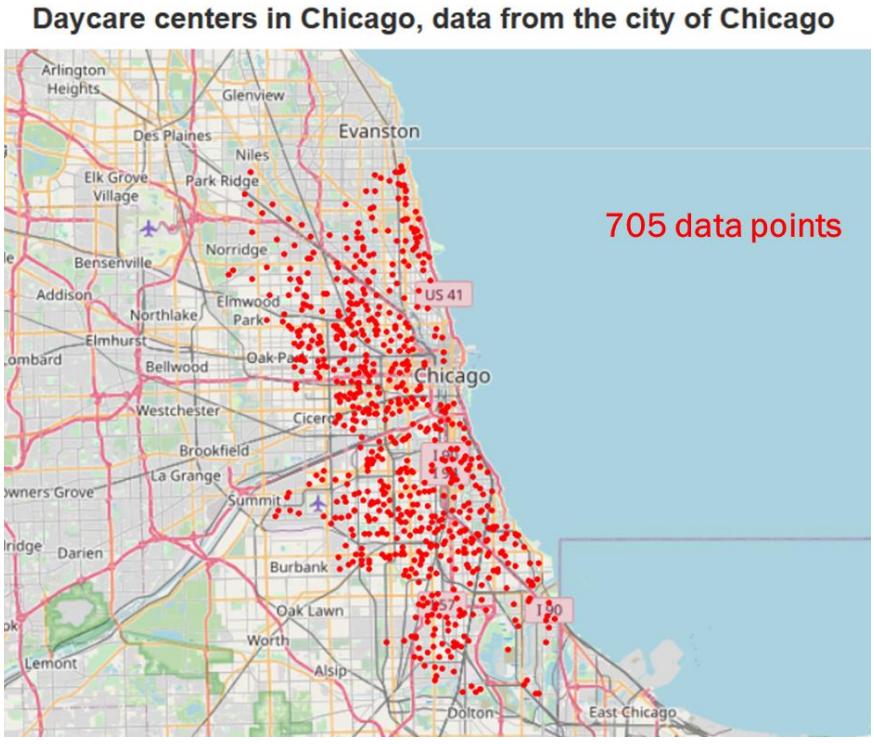
4.2 Explore the daycare data provided by the city of Chicago

Information about early childhood education centers was downloaded from [8]. There are 705 entries, 23 features in the dataframe with early childhood education programs provided by the city of Chicago. An excerpt from this dataframe is shown below.

```
df_daycare.head()
```

	Key	Site Name	Description	Address	City	State	ZIP	Phone Number	URL	Ages Served	Weekday Availability	Duration/Hours	Program Information	Languages other than English	Other Features	Accreditation	Quality	Quality Rating	Latitude	Longitude
0	463	Orr Family Development Center YMCA	<p>The Orr Family Development Center's purpose...	730 N Pulaski Rd	Chicago	IL	60624	(773) 565-0130	http://www.ymcachicago.org/Orr	6 weeks - 3 years old	Full Week, Full Day	NaN	Community Based, Early Head Start	NaN	Accepts CCAP	NAEYC	gold	Gold	41.894071	-87.726046
1	213	All About Kids Learning Academy	<p>At All About Kids Learning Academy we focus...	514 E 75th St	Chicago	IL	60619	(773) 892-2800	http://www.allaboutkidslearningacademy.com	6 weeks to 5 years	Full Day	NaN	Community Based, Head Start	NaN	Accepts CCAP	QRS level 2	silver	Silver	41.758683	-87.611946
2	368	One Hope United - Wings	<p>One Hope United Wings services help new par...	707 E 47th St	Chicago	IL	60653	(312) 949-5590	http://www.onehopeunited.org	0-3	Full Week, Full Day	NaN	Offers Home Visiting, Community Based	NaN	NaN	COA	licensed	Licensed	41.809476	-87.608710
3	375	Catholic Charities - Our Lady of Lourdes Child...	<p>Our Lady of Lourdes provides support for lo...	1449 S Keeler Ave	Chicago	IL	60623	773.521.3126	http://www.catholiccharities.net/services/chil...	6 weeks to 5 years	Full Week, Full Day	NaN	Community Based, Head Start	NaN	Accepts CCAP	None	licensed	Licensed	41.860922	-87.729880
4	1000	Ada S. McKinley - Dream Child Development	NaN	1836 W 87th St	Chicago	IL	60620	(773) 445-5100	NaN	6 weeks to 5 years	None	NaN	Community Based	NaN	NaN	None	licensed	Licensed	41.735887	-87.669375

To visualize this dataset, I used Folium library. The resulting map is shown below. Looking at the map, one can easily see that this dataset covers significantly larger number of daycare centers compared to what is available on the Foursquare. It is therefore important to continue using both datasets for the purposes of this project. In the next section, we merge both daycare datasets to carry out exploratory analysis and machine learning.



4.1 Merge the daycare data from the Foursquare and the city of Chicago

First, we prepare the city of Chicago daycare data for merge with the Foursquare data. Second, merge the daycare data obtained from the Foursquare and the city of Chicago on the common column "COMMUNITY AREA NAME". The resulting dataset contains the neighborhood (Community Area) name, neighborhood coordinates, daycare name and daycare coordinates.

```
merged_city_foursq.head(10)
```

	COMMUNITY AREA NAME	Neighborhood Latitude	Neighborhood Longitude	Site Name	Latitude	Longitude	Venue Category
0	Printers Row	41.870981	-87.629035	It Takes A Village	41.872199	-87.633867	School
1	Sheffield & DePaul	41.927188	-87.653670	Tiny Tots Incorporated	41.929380	-87.648550	Daycare
2	Hermosa	41.924348	-87.734740	Kidslife Daycare Center	41.922340	-87.737204	School
3	Avondale	41.938666	-87.711211	A-Karrasel Child Care	41.936617	-87.707797	Daycare
4	Avondale	41.938666	-87.711211	Kat Slawson	41.935631	-87.715263	Daycare
5	Logan Square	41.923193	-87.707389	Christopher House	41.926334	-87.709773	Daycare
6	Logan Square	41.923193	-87.707389	A-Karrasel Child Care	41.924895	-87.712903	Daycare
7	East Side	41.707314	-87.534902	Shinning Star	41.703346	-87.535548	Daycare
8	Grand Crossing	41.763247	-87.616134	Jellybean Learning Center II	41.765854	-87.615811	Daycare

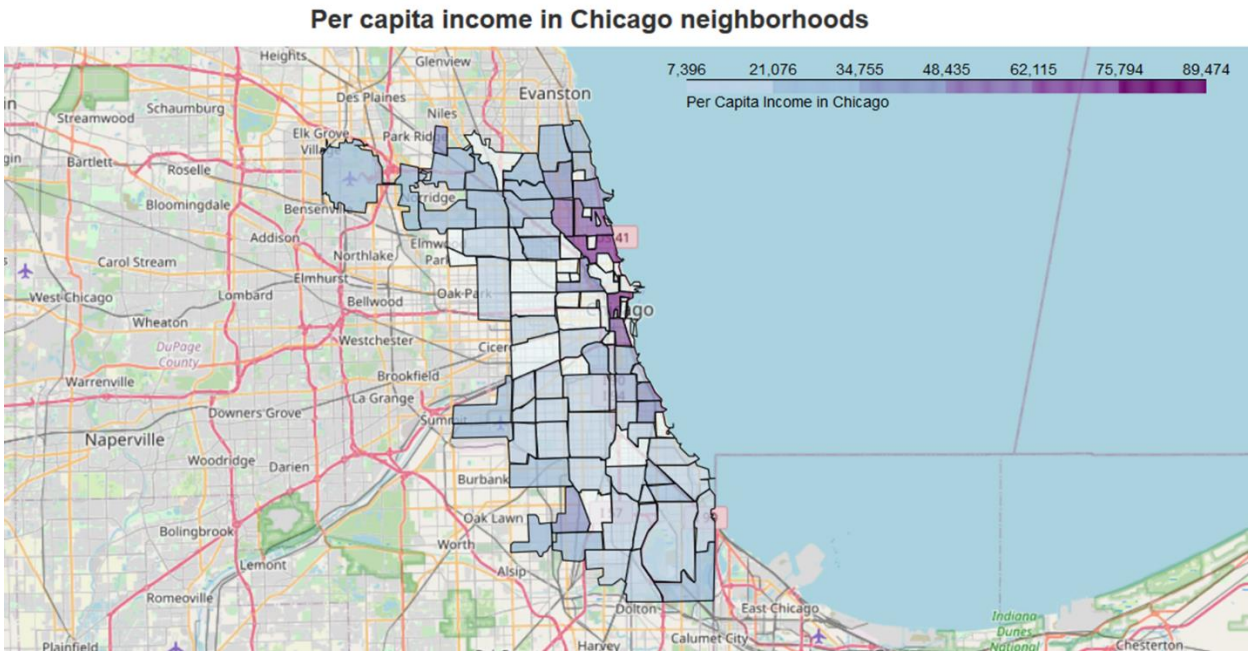
4.2.1 Analysis of the merged dataset in combination with the socioeconomic data on Chicago neighborhoods

In this section, we introduce the dataframe with the socioeconomic data on the city of Chicago. A snapshot of the dataframe is shown below.

df_income

	Community Area Number	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
0	1.0	Rogers Park	7.7	23.6	8.7	18.2	27.5	23939	39.0
1	2.0	West Ridge	7.8	17.2	8.8	20.8	38.5	23040	46.0
2	3.0	Uptown	3.8	24.0	8.9	11.8	22.2	35787	20.0
3	4.0	Lincoln Square	3.4	10.9	8.2	13.4	25.5	37524	17.0
4	5.0	North Center	0.3	7.5	5.2	4.5	26.2	57123	6.0
5	6.0	Lake View	1.1	11.4	4.7	2.6	17.0	60058	5.0
6	7.0	Lincoln Park	0.8	12.3	5.1	3.6	21.5	71551	2.0
7	8.0	Near North Side	1.9	12.9	7.0	2.5	22.6	88669	1.0
8	9.0	Edison Park	1.1	3.3	6.5	7.4	35.3	40959	8.0
9	10.0	Norwood Park	2.0	5.4	9.0	11.5	39.5	32875	21.0
10	11.0	Jefferson Park	2.7	8.6	12.4	13.4	35.5	27751	25.0
11	12.0	Forest Glen	1.1	7.5	6.8	4.9	40.5	44164	11.0
12	13.0	North Park	3.9	13.2	9.9	14.4	39.0	26576	33.0
13	14.0	Albany Park	11.3	19.2	10.0	32.9	32.0	21323	53.0
14	15.0	Portage Park	4.1	11.6	12.6	19.3	34.0	24336	35.0

Using Folium library, I plotted the choropleth map with the color scheme reflecting the distribution of per capita income in Chicago neighborhoods, as shown below.



The choropleth map clearly shows that there are only several neighborhoods with high income. These neighborhoods may be attractive to people, looking to start a daycare with high price point. Let's print out the names of those neighborhoods (communities).

Top 5 Chicago neighborhoods with highest per capita income

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
Near North Side	1.9	12.9	7.0	2.5	22.6	88669	1.0
Lincoln Park	0.8	12.3	5.1	3.6	21.5	71551	2.0
Loop	1.5	14.7	5.7	3.1	13.5	65526	3.0
Lake View	1.1	11.4	4.7	2.6	17.0	60058	5.0
Near South Side	1.3	13.8	4.9	7.4	21.8	59077	7.0

Top 5 Chicago neighborhoods with lowest per capita income

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
West Englewood	4.8	34.4	35.9	26.3	40.7	11317	99.0
West Garfield Park	9.4	41.7	25.8	24.5	43.6	10934	92.0
Fuller Park	3.2	51.2	33.9	26.6	44.9	10432	97.0
South Lawndale	15.2	30.7	15.8	54.8	33.8	10402	96.0
Riverdale	5.8	56.5	34.6	27.5	51.5	8201	98.0

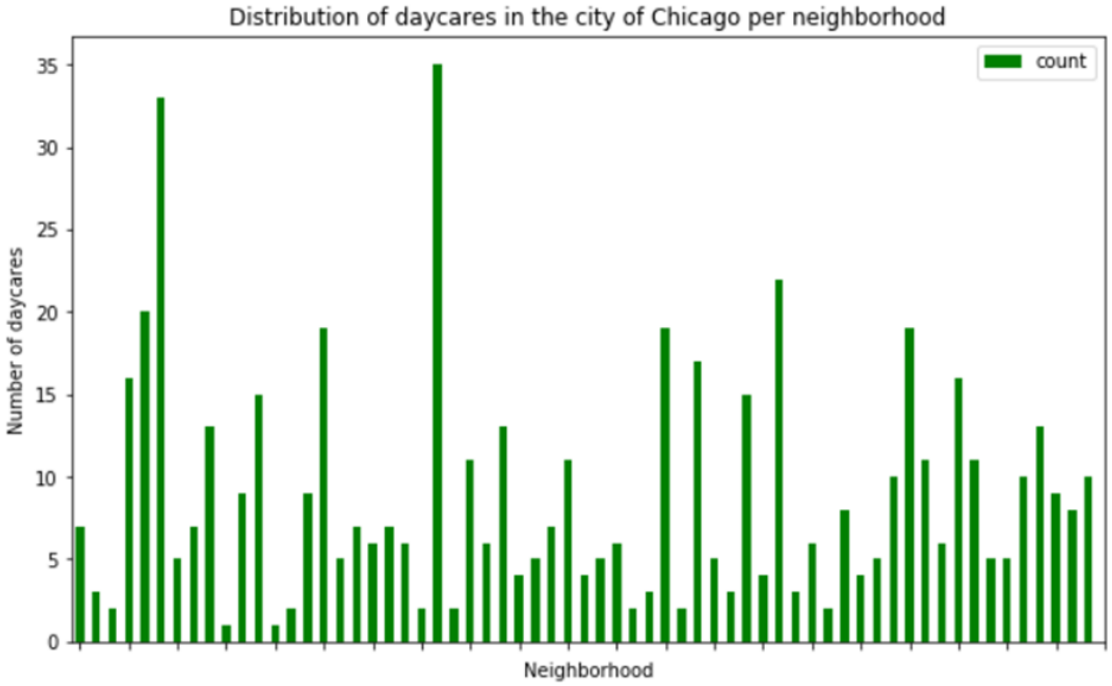
Now, let's add the data about per capita income to the daycare dataframe. Let's remove the duplicate values in the merged dataframe. In order to keep the first row out of several duplicates, we set the parameter keep to 'first' in the drop_duplicates method. There are 63 unique neighborhoods and 557 unique daycare centers in those neighborhoods.

Next, we explore the merged dataset further. First, we calculate the number of existing daycares per neighborhood.

```
daycare_count.head()
```

	COMMUNITY AREA NAME	count
0	Albany Park	7
1	Archer Heights	3
2	Armour Square	2
3	Ashburn	16
4	Auburn Gresham	20

Second, we plot the number of daycares per neighborhood. Neighborhood names are not displayed on the graph since they are not important at this point of analysis.



Now let's merge the socioeconomic data per neighborhood, including per capita income, together with the number of daycares per neighborhood to see if there is any correlation.

10 Neighborhoods with the largest number of daycares

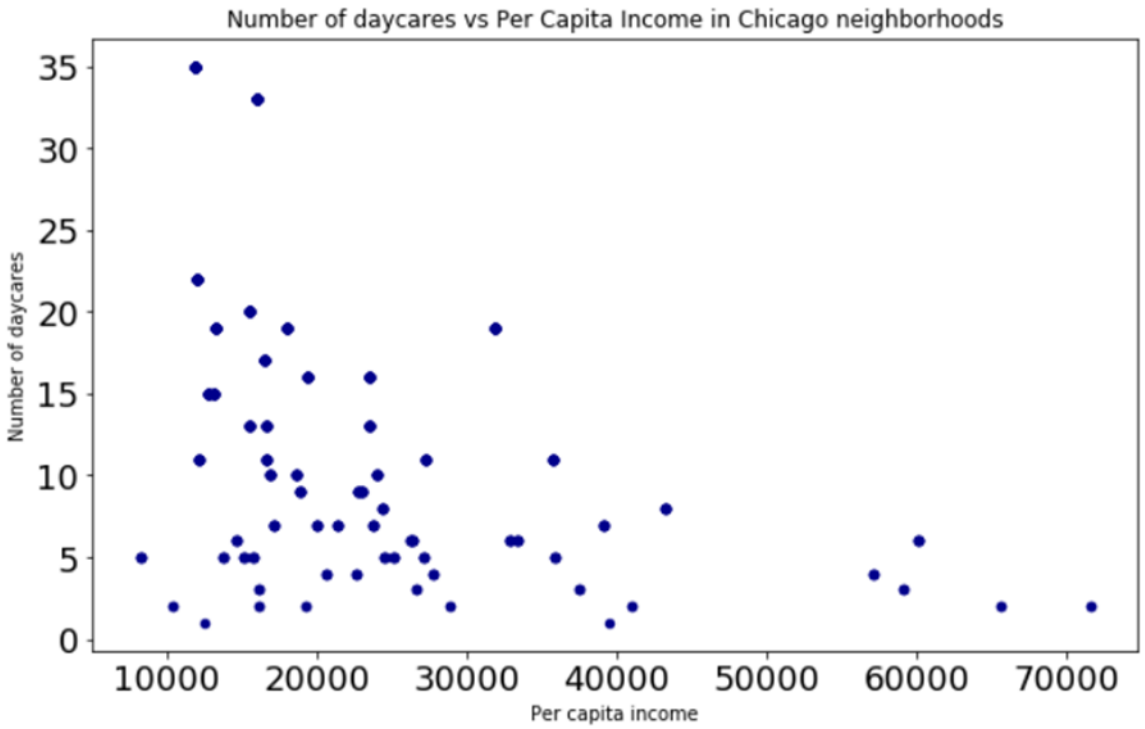
	COMMUNITY AREA NAME	count
22	Englewood	35
5	Austin	33
43	North Lawndale	22
4	Auburn Gresham	20
36	Logan Square	19
51	Roseland	19
15	Chicago Lawn	19
38	Lower West Side	17
3	Ashburn	16
54	South Shore	16

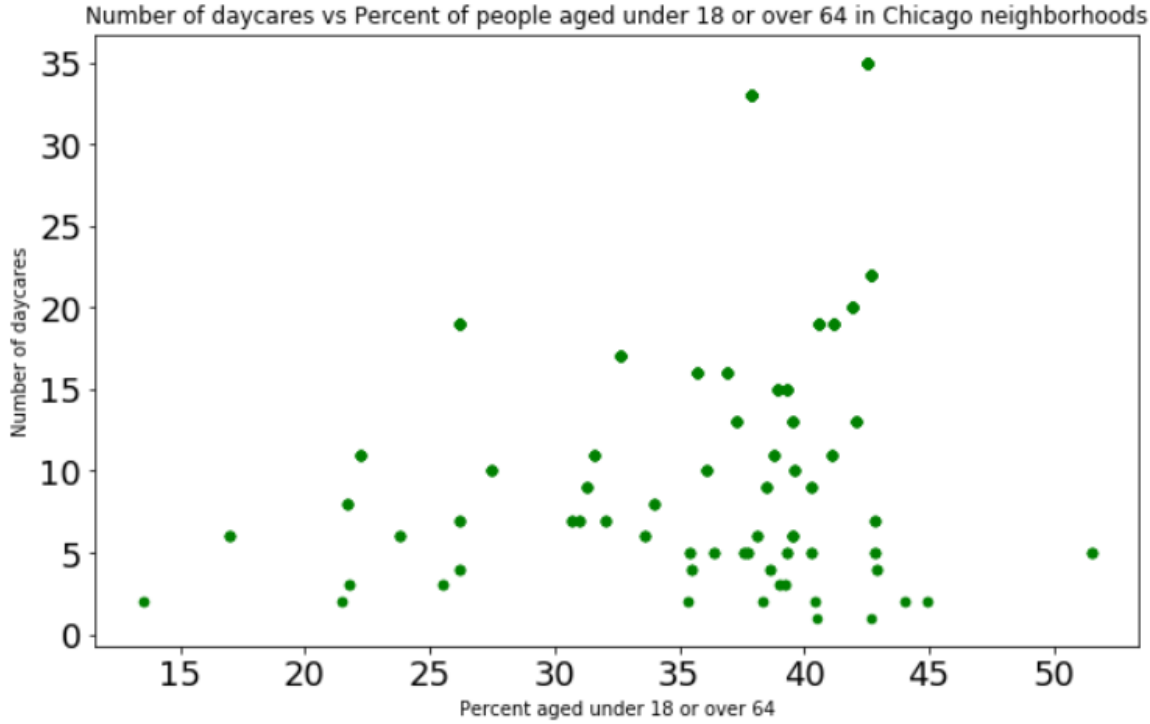
10 Neighborhoods with the smallest number of daycares

	COMMUNITY AREA NAME	count
1	Archer Heights	3
21	Edison Park	2
23	Fuller Park	2
13	Calumet Heights	2
37	Loop	2
34	Lincoln Park	2
2	Armour Square	2
46	Oakland	2
12	Burnside	1
9	Beverly	1

4.2.2 Analysis of socioeconomic data and daycare density per neighborhoods in Chicago

In this section, we carry out the exploratory analysis of the daycare density (number of neighborhood) and the socioeconomic factors, such as per capita income and percent of people aged under 18 or 64. Please see the scatter plots below.


Number of daycares vs Per Capita Income in Chicago neighborhoods
Number of daycares
Per capita income
0
5
10
15
20
25
30
35
10000
20000
30000
40000
50000
60000
70000


Number of daycares vs Percent of people aged under 18 or over 64 in Chicago neighborhoods
Number of daycares
Percent aged under 18 or over 64
0
5
10
15
20
25
30
35
15
20
25
30
35
40
45
50

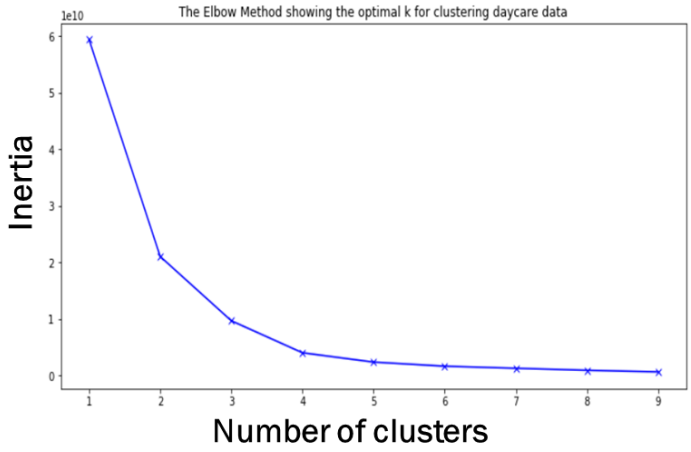
Based on the plots above, it appears that there is inverse correlation between the number of daycares in a given area and the per capita income. Let's explore this further. We will use the K-means clustering model to cluster the daycare data.

4.2.3 K-means clustering machine learning method

In order to perform the K-means clustering, we define the optimal number of clusters k. Calculate inertia for different number of clusters and analyze the Elbow plot. Cluster is performed on the daycare data containing information on per capita income and the number of daycares per neighboorhood.

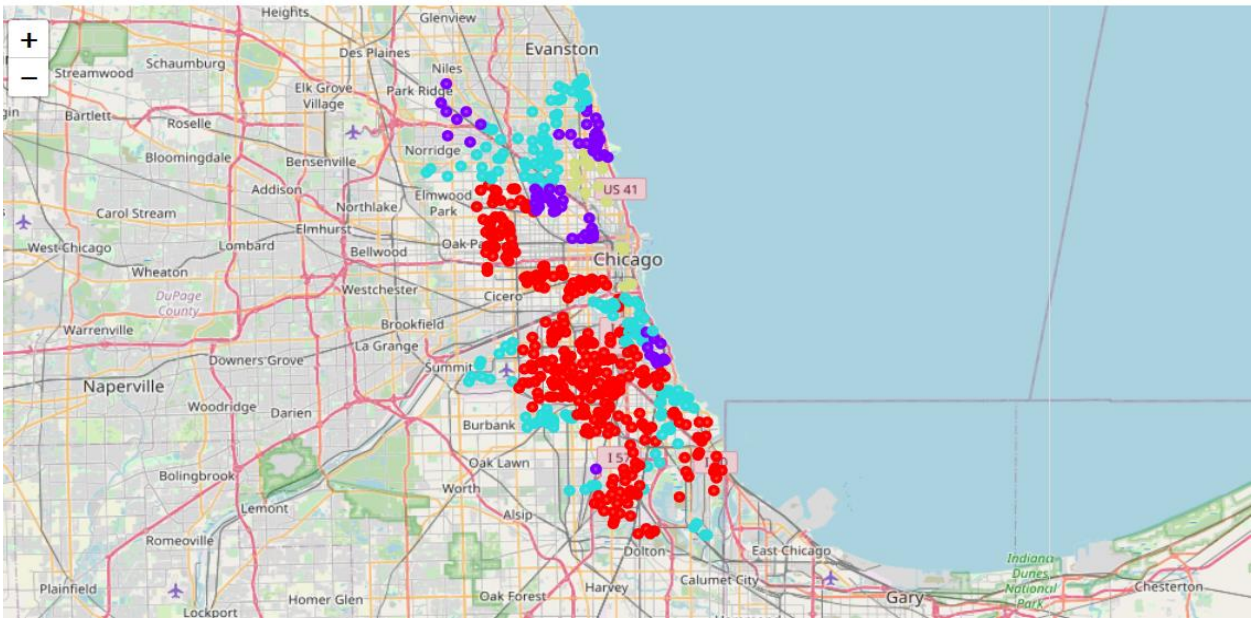
```
inertia1 = []
K = range(1, 10)
for k in K:
    kmeanModel = KMeans(n_clusters=k, random_state = 0)
    kmeanModel.fit(daycare_clustering)
    inertia1.append(kmeanModel.inertia_)
```

```
plt.figure(figsize=(12,6))
plt.plot(K, inertia1, 'bx-')
plt.title('The Elbow Method showing the optimal k for clustering daycare data')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.show()
```



According to the elbow plot, the optimal number of clusters = 4. The map of daycare centers with the resulting clusters is shown below.

Daycare data clustering, based on per capita income and number of daycares per neighborhood



Next, we examine each individual cluster, so we can assign meaningful cluster labels. After looking at each cluster, it turns out that they correspond to different average per capita incomes. The results are shown in the table below.

cluster 1 = 35735 average income, 68 data points

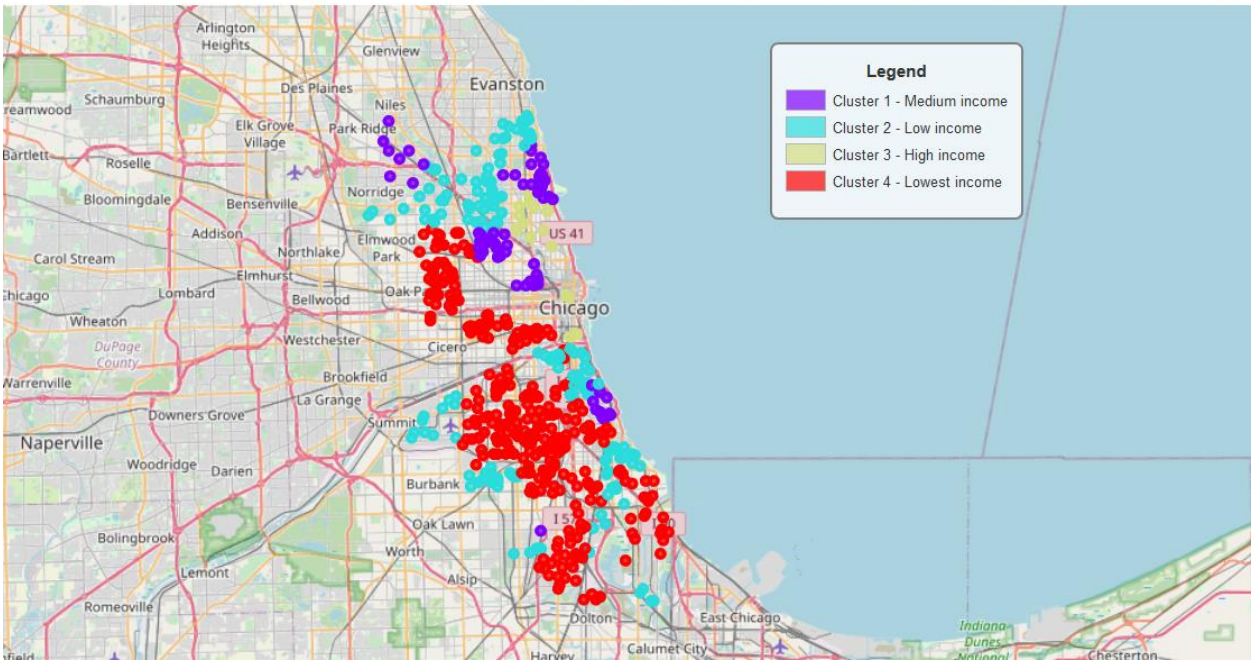
cluster 2 = 23598 average income, 159 data points

cluster 3 = 61189 average income, 17 data points

cluster 4 = 14767 average income, 313 data points

Based on the cluster analysis, we can label daycare datapoints accordingly. The map below includes the legend for each cluster.

Daycare data clustering, based on per capita income and number of daycares per neighborhood



5. Results

In this project, I extracted the data about daycares for children in Chicago from the Foursquare data platform. This data was very limited (36 datapoints, after removing the mislabeled venues), so I supplemented it with the data provided by the city of Chicago. After merging the datasets, the total number of daycare datapoints was 741.

Exploratory data analysis for performed to look for the correlation between the number of daycares per neighborhood and the socioeconomic factors, such as per capita income per neighborhood. The number of daycares per neighborhood was calculated.

I applied a machine learning k-means algorithm to cluster the daycare data combined with per capita income. The optimal number of clusters was defined using the elbow method: the inertia for a range of the number of clusters was calculated and plotted. From the plot it was determined that the optimal number of clusters is 4.

From examining the individual clusters, one can conclude that the clusters correspond to per capita income. The income increases from cluster to cluster. Cluster 4, with the largest number of rows, corresponds to the areas with the lowest average income of 14767. Cluster 3 corresponds to the area with the highest average income: 61189. Cluster 2 contains the daycare data corresponding to the neighborhoods with low average per capita income of 23598. Finally, cluster 1 contains the data for the medium range per capita income of 35735.

6. Discussion

First of all, I would like to point out that using the Foursquare data platform, only 36 daycares were found for all the neighborhoods in Chicago. This number is very low and clearly shows the limitations of the Foursquare platform for searching to this type of venue. In fact, when I ran the unrestricted search for venues per Chicago neighborhood, without specifying a venue category, the results included almost 2700 different venues. Out of those venues, none were daycares. This makes sense, assuming that daycares are not the most popular venue type on Foursquare. It was therefore necessary to specify venue category in order to find daycares.

The data set about daycare centers from the city of Chicago was much more informative compared to the Foursquare data. I merged both datasets in order to perform more detailed analysis. In the final dataset, there were 63 unique neighborhoods and 557 unique daycare centers in those neighborhoods.

Using the Chicago census data, I analyzed the correlation between per capita income and the number of daycares per neighborhood. Interestingly, there was a negative correlation, with very few exceptions: there are less daycares in high income neighborhoods, compared to the low income areas. Low income neighborhoods is the largest number of daycares included Englewood (35 daycares), Austin (33 daycares), North Lawndale (22 daycares), Auburn Gresham (20 daycares). Areas with high income and low daycare count included the Loop (2 daycares), Archer Heights (3 daycares), Edison Park (2 daycares).

For a person interested in starting a new daycare business, the results of this analysis will help estimate the acceptable price point for a given neighborhood. Low income neighborhoods will not be receptive to a daycare with high weekly payments. By looking at the daycares in each of the four data clusters, one can make a decision whether to pursue a new daycare business venue in a given neighborhood.

Notably, the daycare centers listed on Fousquare were mostly private businesses, judging by their names, whereas the Chicago data included daycares funded by Chicago Public Schools. Further analysis can be performed by specifying a type of daycares, such as full time or part-time, or the age of children that a daycare serves.

It is important to note that according to this data, there are much less daycares available in the neighborhoods with high per capita income. This may be attributed to the fact that in those areas people may be able to afford hiring nannies instead of using daycares.

7. Conclusions

The goal of this project was to provide insight on daycares in Chicago, Illinois. The results of analysis showed that there are a lot of existing daycare businesses. The data about daycares on the Foursquare platform, which was required to be used in this project, is very limited.

Supplementing this dataset with the data from the city of Chicago allowed to perform a more detailed analysis. Impact of socioeconomic factors on the density of daycares per neighborhood was evaluated using k-means clustering machine learning algorithm.

The results demonstrate that the daycare data can be divided into 4 individual clusters, based on per capita income in given neighborhoods. The results of this project can be used to evaluate new daycare business opportunities in Chicago area.