IBM Data Science Capstone Project
The Battle of Neighborhoods (Week 2)

Table of Contents

Introduction/Business Problem

Description & Discussion of the Background

"The main purpose of the American school is to provide for the fullest possible development of each learner for living morally, creatively, and productively in a democratic society." —The ASCD Committee on Platform of Beliefs, Educational Leadership, January 1957 This statement is powerful and convincing. However, can education be given fairly regardless of race, gender, place of origin, or background?

Education indeed is one of the most decisive elements and parameters for the children's economic future and success. I other words, it has a competitive and capitalistic aspect. I was born and raised in Japan, where elementary and middle school education is compulsory. Before kids attend an elementary school, they have an option of an affordable kindergarten and preschool for small toddlers. Even though the idea of sending small kids to these schools was not as popular as now, it is becoming more common as a society to accept and encourage more women to work in the workplace.

Due to the fact that the population is declining, the cost of education for parents per child is also on the rise. The government's budget allocated to each city/ school is also fairly allocated, so that all children, regardless of family income, can study in the same classroom. From these reasons, I was able to grow up without having to feel the difference between wealthier children and me or be particularly aware of the characteristics of the area in which I lived.

I studied abroad in Los Angeles, California, fifteen years ago. I still remember the experience I had at that time. Los Angeles, with a population of 12 million, is one of the largest cities in the United States. It is a melting pot of ethnic groups, with people from all over the world. It was a fascinating city to live in.

In reality, however, the gap between the rich and the poor is stark, and cities are racially segregated, creating divisiveness among people not only psychologically but also socially, culturally, and politically. This fact came as a shock to me when I first came to America. People with lower incomes cannot attend schools with a good program and high scores because of their school district, which results in limiting the possibilities of those kids who live in these particular areas. However, I only knew those were facts from my perspective and not based on data.

I am a father of 1, and I have been looking for the best preschool for my child in my city. However, it has not been successful because there is not much place. Even if I found someplace, either my child or I did not like the place. Besides, the prices were over our budget. On the other hand, in Santa Monica, CA, one of the most expensive cities to live in Los Angeles, there are many places. It made me wonder if the family wealth and educational opportunity, even for smaller kids, have a direct relationship.

Through IBM's Data Scientist course, I will address these questions with the help of the power of data. I will reveal the facts and provide valuable information through the visual materials that policymakers and government officials could use to take appropriate actions and measures in the social welfare and school system fields.

Data Description

The primary data I will use in this project is the median household income in Los Angeles by zip code. I will use a technique called K-means, which is frequently used in data science to create clusters. Moreover, I will color for each cluster with boundaries so that we can intuitively see which areas are wealthy and which areas are not.

On top of that, I would like to pin the preschool locations that I will obtain from Foursquare API. If my assumptions are correct, then the wealthier neighborhoods should have more educational facilities, and the lower-income neighborhoods should have fewer educational facilities.

I would also like to find out the population (preferably under six years old) and the number of educational facilities in the area where each facility is located and plot a histogram or graph with an index of educational opportunities per person.

Based on these objective data, I will test, evaluate, and discuss my hypothesis and draw conclusions.

```
import numpy as np
import pandas as pd
from pandas import DataFrame
import folium
import json
import requests
from geopy.geocoders import Nominatim
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

pd.set_option('display.max_rows', 600 )

# from pandas.io.json import json_normalize
# import matplotlib.cm as cm
# import matplotlib.colors as colors
# import random

print('Libraries imported.')
```

```
Libraries imported.
```

import an estimated median income by zip code informatin.

```
with open('df_zip.csv') as zipcd:
    df_zipcode = pd.read_csv(zipcd)

df_zipcode["ZIP"] = df_zipcode["ZIP"].map(lambda x: int(x))
df_zipcode["Estimated Median Income"] = df_zipcode["Estimated Median Income"].map(lambda x: int(x))

data_zip = df_zipcode.drop(['ZipCode'],1)
data_zip.head(3)
```

	ZIP	Estimated Median Income
0	90001	38521
1	90002	35410
2	90003	37226

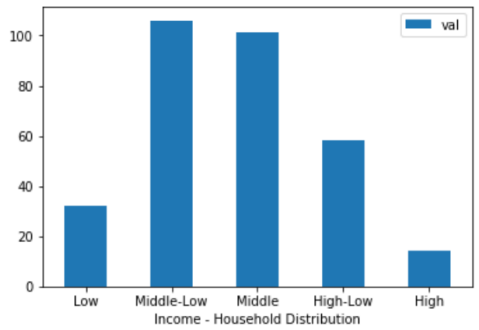
Methodology - K-Means

Split the data to 5 groups based on the median household income, and plot it to the bar chart.

```
k = 5

# The new dataframe only with Latitude & Longtitude.
la_cluster = data_zip.drop(['ZIP'], 1)
kmeans = KMeans(n_clusters = k, random_state = 0).fit(la_cluster)
kmeans.labels_

data_zip.insert(0, 'Cluster Labels', kmeans.labels_)
data_zip["Cluster Labels"] = data_zip["Cluster Labels"].replace({3:0, 0:1, 4:3, 1:4})

data_zip.head(3)
```

	Cluster Labels	ZIP	Estimated Median Income
0	1	90001	38521
1	1	90002	35410
2	1	90003	37226

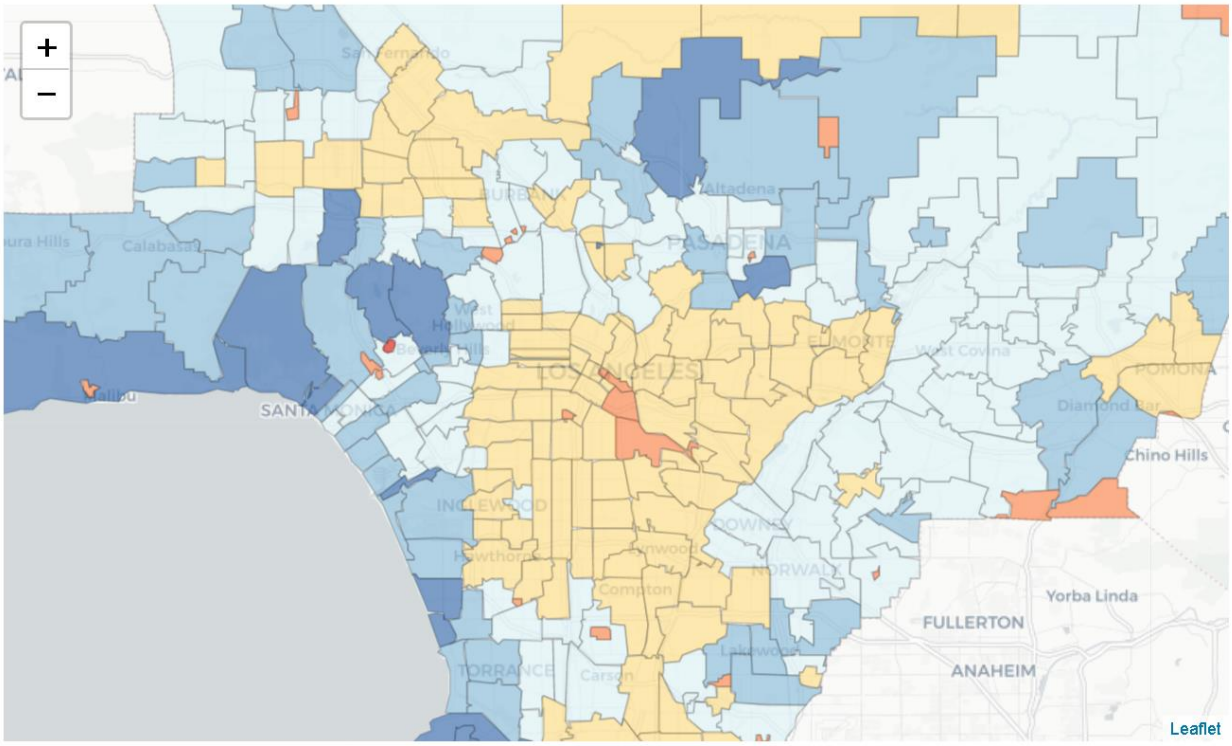
```
cluster0 = len([i for i in data_zip["Cluster Labels"] if i == 0])
cluster1 = len([i for i in data_zip["Cluster Labels"] if i == 1])
cluster2 = len([i for i in data_zip["Cluster Labels"] if i == 2])
cluster3 = len([i for i in data_zip["Cluster Labels"] if i == 3])
cluster4 = len([i for i in data_zip["Cluster Labels"] if i == 4])

df = pd.DataFrame({'Income - Household Distribution':['Low', 'Middle-Low', 'Middle', 'High-Low', 'High'], 'val':[cluste
r0, cluster1, cluster2, cluster3, cluster4]})
ax = df.plot.bar(x='Income - Household Distribution', y='val', rot=0)
```



GeoJson

I used the python folium library to visualize the geographic details of Los Angeles and each city's estimated median income. I used latitude and longitude values to get the map as below.

The city boundaries are shown by zip code on the map, and the different colors show the clusters of estimated median income. Red is lower-income areas, and Blue is higher-income areas.



As I expected, those beach cities are blue, which means high-income areas as well as surrounding suburb cities.
Please note that some of the zip code does not have estimated household income data resulted in being shown with red.

Foursquare

I used the Foursquare API to explore the preschools in Los Angeles. It limits the response of 100 venues for my API call, which may not be sufficient to support my hypothesis accurately, but I believe it still gives us a specific perspective and fact.

```
CLIENT_ID = 'D0AGU1UH3RETNJF5RE3E5P5TAKJN2FIWM5BNUPYDJ2VUJAKQ' # your Foursquare ID
CLIENT_SECRET = 'TJC11CVWBBPGGKK441R04TFEA1ULFO1Y5H1T3ONPNKYXECIR' # your Foursquare Secret
VERSION = '20180604'

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)


address = 'Los Angeles, CA'
geolocator = Nominatim(user_agent="foursquare_agent")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude

print(latitude, longitude)
```

```
search_query = 'Preschool'
radius = 100000
LIMIT = 10000
categoryId = "52e81612bcbc57f1066b7a45"
url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&query={}&radius={}&limi
t={}&categoryId={}'.format(CLIENT_ID, CLIENT_SECRET, latitude, longitude, VERSION, search_query, radius, LIMIT, categor
yId)
url
```

```
results = requests.get(url).json()
venues = results['response']['groups'][0]['items']

def create_dataframe(keyword):
    i=0
    for venues[i] in venues:
        lst.append(venues[i]['venue'][keyword])
        i+=1

def create_dataframe2(list_n, keyword):
    i=0
    for venues[i] in venues:
        list_n.append(venues[i]['venue']['location'][keyword])
        i+=1
```

```
lst=[]
create_dataframe("name")

lst2=[]
create_dataframe2(lst2, 'lat')

lst3=[]
create_dataframe2(lst3, 'lng')

lst4=[]
create_dataframe2(lst4, 'city')

#lst5=[]
#create_dataframe2('postalCode')

df_venues = DataFrame(columns=['Venue Name', 'Latitude','Longitude', 'City'])
df_venues = pd.DataFrame({'Venue Name':lst, 'Latitude':lst2, 'Longitude':lst3, 'City':lst4})
```

Here is a head of the list Venues name, category, latitude and longitude information from Forsquare API.

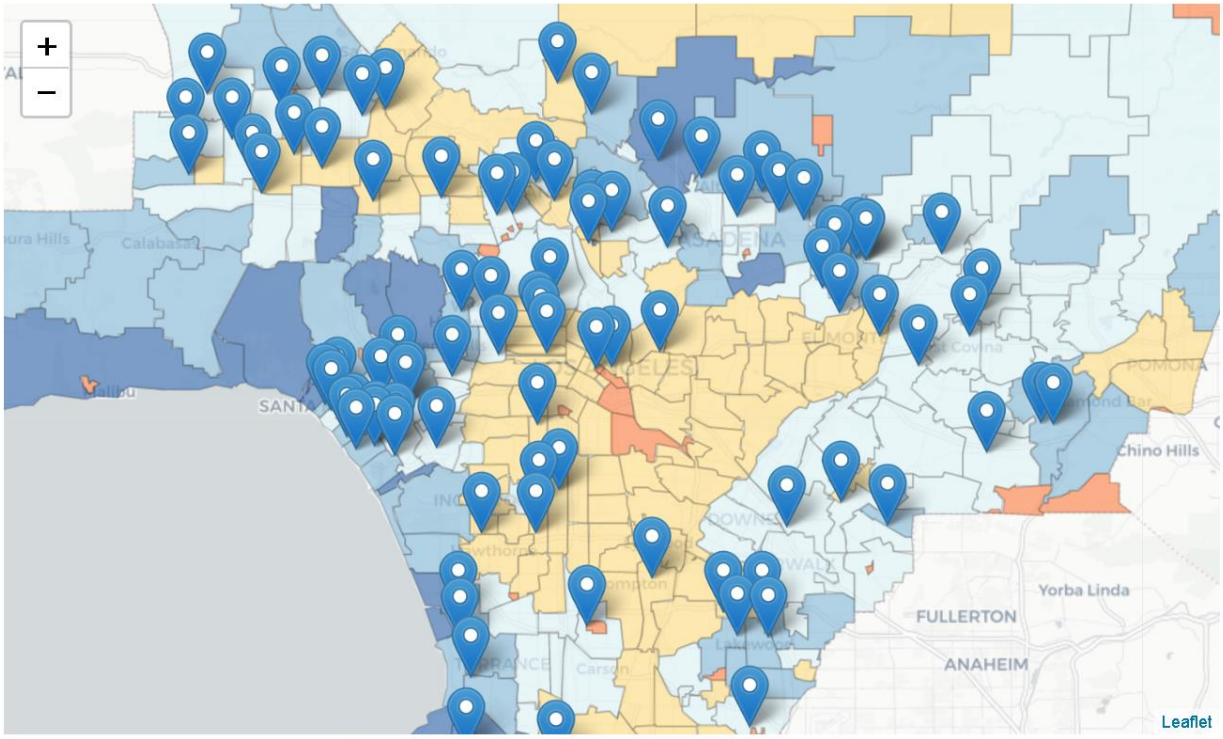
```
df_venues = df_venues[(df_venues["City"] != 'Anaheim') & (df_venues["City"] != 'Fullerton') & (df_venues["City"] != 'Buena Park') & (df_venues["City"] != 'Los Alamitos') & (df_venues["City"] !=  'Cypress') & (df_venues["City"] !=  'La Palma') & (df_venues["City"] !=  'Garden Grove')]
df_venues = df_venues.reset_index()
df_venues
```

	index	Venue Name	Latitude	Longitude	City
0	0	Acting with Alisha Los Angles top youth acting...	34.053122	-118.244798	Los Angeles
1	1	Alcazar Child Development Center at USC	34.063663	-118.201448	Los Angeles
2	2	Lily Preschool	34.062897	-118.302483	Los Angeles
3	3	Sae Ssak Preschool	34.074281	-118.307548	Los Angeles
4	4	Ethel Education	34.079216	-118.310126	Los Angeles
5	5	First United Methodist Preschool	34.188041	-118.312322	Burbank
6	6	Lake Avenue Preschool	34.162839	-118.132846	Pasadena
7	7	Creative Angels Preschool/Kindergarten	34.102546	-118.299750	Los Angeles
8	8	Creative Learning Academy	34.010646	-118.309754	Los Angeles
9	9	Paper Pinecone	34.061932	-118.345413	Los Angeles
10	10	Growing Years Children's Academy	34.143772	-118.264443	Glendale
11	11	Eagle Rock Montessori School	34.139726	-118.195589	Los Angeles
12	12	West Hollywood Children's Academy	34.089510	-118.351900	Los Angeles
13	13	Cedar Montessori	34.151737	-118.244236	Glendale
14	14	South Vermont KinderCare	33.963073	-118.291166	Los Angeles
15	15	Bully Free TV	34.152887	-118.262115	Glendale
16	16	Bright Horizons at South Figueroa Street	34.052086	-118.258127	Los Angeles
17	17	Three Little Stars Learning Center	33.954144	-118.309357	Los Angeles
18	18	Gan Yaffa Preschool	34.045337	-118.386094	Los Angeles
19	19	Montessori Academy	34.174359	-118.294744	Burbank
20	20	Rising Stars Academy	33.931370	-118.311820	Los Angeles
21	21	Palisades Preschool	34.024860	-118.498238	Santa Monica
22	22	Media Center Montessori Preschool	34.164430	-118.345959	Burbank
23	23	My Friend's Montessori Preschool Coop	33.993305	-118.399007	Culver City
24	24	Olive Tree Learning Centers	34.191200	-118.165184	Altadena
25	25	The Sand Castle Preschool	34.204307	-118.203495	La Cañada Flintridge
26	26	The University Parents Nursery School (UPNS)	34.024793	-118.427069	Los Angeles
27	27	Playfactory Preschool	34.111217	-118.057877	Temple City
28	28	Teremok Preschool	34.253029	-118.602754	Chatsworth
29	29	Pasadena Preschool Academy	34.166790	-118.096260	Pasadena
30	30	Blue Oak Creative Schoolhouse	33.997030	-118.435514	Los Angeles

Result

Now, let's plot the preschool location with a pin on the map to analyze if it has relationship with each city's estimated median income.

```
for lat, lng, vname in zip(df_venues['Latitude'], df_venues['Longitude'], df_venues['Venue Name
    label = folium.Popup(vname, parse_html=True)
    folium.Marker(
    [lat,lng],
    popup = label,
    ).add_to(MAP)

MAP
```



As expected, it turns out that there are fewer childcare facilities in the city of Los Angeles and low-income areas, primarily downtown. Conversely, there are more childcare facilities in the higher-income seaside and suburban areas.

Education Opportunity Ration per children

I want to explore another possibility of children's educational opportunities varies based on the area that they live in. Suppose the income of the city and the number of school locations are in a positive correlation. (Positive correlation is a relationship between two variables in which both variables move in tandem—that is, in the same direction.)

I imported the population data from the city of Los Angeles. There is no exact number of 3 to 5 years children population data available anywhere. So, I used the Age 0 - 15 population and divided it by 5, assuming the population demographic is equally spread out.

```
with open('Estimated_Population_Age.csv') as popu:
    population = pd.read_csv(popu)

population.tail()
```

	Census_Tract	FIPS	CITYNAME	Service_Area	Age_0_15	Age_16_18	Age_19_20	Age_21_25	Age_26_59	Age_
2797	980028	44000	Los Angeles city - Westchester	5	0	0	0	0	3	
2798	980030	22412	El Segundo city	8	0	0	0	0	0	
2799	980031	44000	Los Angeles city - San Pedro	8	8	8	8	52	918	
2800	980033	43000	Long Beach city	8	2	1	1	2	33	
2801	990300	43000	Long Beach city	8	0	0	0	0	0	

5 rows × 21 columns

```
d = df_venues['City'].value_counts().to_dict()
df_num = pd.DataFrame(list(d.items()), columns=['City', 'Number'])

df_num['Estimated 2-5YRs Population'] = 0
```

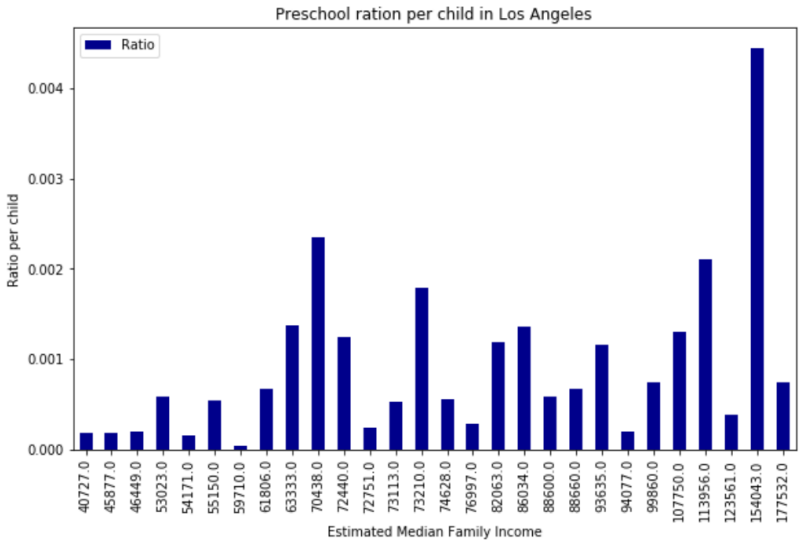
```
i=0
for city, EP in zip(df_num['City'], df_num['Estimated 2-5YRs Population']):
    EP = ((sum([pop2 for pop1, pop2 in zip(population['CITYNAME'], population['Age_0_15']) if p
    df_num.at[i, 'Estimated 2-5YRs Population'] = EP
    i+=1

df_num['Ratio'] = df_num['Number']/df_num['Estimated 2-5YRs Population']
df_final = df_num.sort_values(by=["Ratio"],ascending=False)

drop_index2 = df_final.index[df_final["Estimated 2-5YRs Population"] == 0]
df_final = df_final.drop(drop_index2)

df_final
```

```
d= pd.read_html('https://en.wikipedia.org/wiki/List_of_California_locations_by_income', match = 'Acampo')

# The dataframe will consist of three columns: PostalCode, Borough, and Neighborhood
df1 = d[0]

df2 = df1.drop(['Population[1]', 'Populationdensity[1][2][note 1]', 'Per capita income[3]', 'Median household income
[6]'], 1)
df2 = df2.rename(columns={'County/ies[note 2]': "county", "Place": "City", "Median family income[5]": "median family in
come"})

drop_index1 = df2.index[df2["median family income"] == '[7]']
df3 = df2.drop(drop_index1)

df4 = df3[df3['county'].isin(['Los Angeles'])].copy()
df4["median family income"] = df4["median family income"].map(lambda x: x.lstrip('$'))
df4["median family income"] = df4["median family income"].map(lambda x: float(x.replace(',', '')))
df4 = df4.drop_duplicates(subset=['City']).reset_index()
df4 = df4.drop(['index', 'county'], 1)
df4
```

```
df_plot = pd.merge(df_final, df4, how="inner", on = "City")
df_plot = df_plot.drop(index=0)
df_plot = df_plot.drop(['City', 'Number', 'Estimated 2-5YRs Population'], 1)
df_plot = df_plot.sort_values(by=['median family income'])
df_plot
```

	Ratio	median family income
25	0.000191	40727.0
26	0.000186	45877.0
24	0.000197	46449.0
15	0.000589	53023.0
27	0.000157	54171.0
18	0.000551	55150.0
28	0.000050	59710.0
14	0.000669	61806.0
5	0.001377	63333.0
2	0.002353	70438.0
8	0.001250	72440.0
22	0.000248	72751.0
19	0.000527	73113.0
4	0.001794	73210.0
17	0.000558	74628.0
21	0.000290	76997.0
9	0.001193	82063.0
6	0.001355	86034.0
16	0.000585	88600.0
13	0.000678	88660.0
10	0.001162	93635.0
23	0.000202	94077.0
12	0.000740	99860.0
7	0.001299	107750.0
3	0.002103	113956.0
20	0.000386	123561.0
1	0.004444	154043.0
11	0.000748	177532.0

```
df_plot.plot(kind='bar', x='median family income', y='Ratio', figsize=(10, 6), color='darkblue')
plt.title('Preschool ration per child in Los Angeles')
plt.xlabel('Estimated Median Family Income')
plt.ylabel('Ratio per child')
plt.show()
```



As a result, the results suggest that early childhood education facilities are relatively more prevalent in middle-class and upper-class areas.

Discussion

Everyone knows for a fact that while Los Angeles is a large city, people are segregated by race and wealth. This attempt used only minimal indices and data, so the causes and precise categorization of this problem are remained open to study. However, we have shown that access to education depends on the area where one lives and wealth.

Of course, more complex explanations may be required in practice, as historical context and political factors also need to be taken into account. Besides, some clustering and classification studies may try very different approaches. It is essential to note that not all classification methods will produce the results as this study.

I used the K means algorithm as part of this clustering study. As an extension of this study, it may be possible to use linear regressions to predict simple future predictions, taking into account population growth rates, economic indices, and inflation, which is currently being led by the US government.

This study's results could influence future development plans for the city and the decision making process for education plans.

Conclusion

The gap between the rich and the poor has become a hot topic that symbolizes division and inequality in the United States today, and it is getting people's attention.

I do not mean to criticize American-style capitalism, but I believe that our children's education should be ethical and equal. Humanity and morality must not be sacrificed to market forces.

I hope this study to be used by city officials and local legislators to bring about some positive results.