COURSERA

Applied Data Science Capstone Project

"Analysing the Suitability of Neighbourhoods in New York for Young Families to Live In"

Joe O'Reilly

23.03.2021

Contents

1. Introduction:

1.1 Background:

The aim of this project is to find the most suitable neighbourhood in New York for a family with young children to live in. The neighbourhood should have some amenities that the parents would benefit from as well as some amenities for the children to avail of.

1.2 Problem Statement:

The parents of the family would like to live within walking distance of at least one pub, bar or restaurant. For the purpose of this study, walking distance is considered to be within 1,500 metres.

The children would like to live in a neighbourhood that is close to one or more of the following amenities:

- A Sports Club
- A Park
- A Playground
- A Toy / Game Store
- A Movie Theater (Cinema)

2. Data Sources:

2.1 Geospatial Data:

The primary dataset that will be used in this analysis is available at the following link:

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

This json file contains a list of the neighbourhoods in New York along with the location data (latitude, longitude). This json file will be read into a Pandas dataframe for the analysis. The project will then use this location data in conjunction with the Foursquare API to explore which amenities/venues are nearby each neighbourhood.

2.2 Foursquare Data:

We will use this data from Foursquare to filter the neighbourhoods of New York to ensure that only neighbourhoods that are suitable for both the parents and the children are considered in our analysis. To do this, we will only consider neighbourhoods within 1,500m of either a Pub, Bar or Restaurants. This should satisfy the parents' wants. We will also only consider neighbourhoods that are within a similar distance of either a Sports Club, Park, Playground, Toy Store or Cinema to satisfy the childrens' needs.

We will then perform k-means clustering on the filtered data to cluster the neighbourhoods.

3. Data Treatment/ Filtering:

Firstly, the json geospatial data was filtered to include the 'features' key only as this contains the info that we want to use. All other keys are dropped as they are irrelevant to our analysis.

```
[4]: neighborhoods_data = newyork_data['features']
```

This updated neighbourhood data was then transformed into a Pandas dataframe. An empty dataframe was created and the neighbourhood data was insert into this empty dataframe containing the 4 fields that we want to use for our analysis.

```
[5]: # define the dataframe columns
column_names = ['Borough', 'Neighborhood', 'Latitude', 'Longitude']

# instantiate the dataframe
neighborhoods = pd.DataFrame(columns=column_names)
```

We then used the Foursquare API to get a list of the top 30 venues (by setting LIMIT = 30) within a distance of 1,500 metres of the neighbourhood's coordinates (by setting radius = 1500).

```
[13]: LIMIT = 30

radius = 1500 # define radius

url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    neighborhood_latitude,
    neighborhood_longitude,
    radius,
    LIMIT)
url # display URL
```

We then reduced the size of our working dataset by only including data on the following venues/ amenities:

- Park
- Supermarket
- Playground
- Toy / Game Store
- Movie Theater
- School
- Sports Club
- Irish Pub

```
NY_venues_Filtered = NY_venues[NY_venues['Venue Category'].isin(["Park",
                                "Supermarket",
                                "Playground",
                                "Toy / Game Store",
                                "Movie Theater",
                                "School",
                                "Sports Club",
                                "Pub",
                                "Bar",
                                "Restaurant"])]
NY_venues_Filtered.head()
```

Having reduced the number of columns of our working dataset above, we then applied further filters to reduce the number of records (rows) in the dataset before we begin our analysis. The filters applied ensure that both the parents and childrens' needs are met. I.e. it is within walking distance of a pub, bar or restaurant while also being close to facilities for the children, as outlined in the Problem Statement.

```
#Filter the data to suit the requirements for the parents. i.e. must be close to a Restaurant, Bar or Pub
NY_Parents = NY_grouped[(NY_grouped['Restaurant'] > 0) | (NY_grouped['Bar'] > 0) | (NY_grouped['Pub'] > 0)]

#Filter the data to suit the requirements for the Children. i.e. must be close to a Park, Playground, Cinema, Sports Club
#or Toy Store
NY_Children = NY_grouped[(NY_grouped['Park'] > 0) | (NY_grouped['Playground'] > 0) | (NY_grouped['Movie Theater']>0) |
                         (NY_grouped['Sports Club']>0) | (NY_grouped['Toy / Game Store']>0)]

#Inner join the 2 dataframes to find the neighborhoods that are suitable for both Parents & Children
NY_Parents_Children = pd.merge(NY_Parents,NY_Children[[]],how='inner',on=['Neighborhood'])
print("There are: {} neighborhoods that meet the minimum requirements".format(NY_Parents_Children.shape[0]))
```

4. Exploratory Data Analysis:

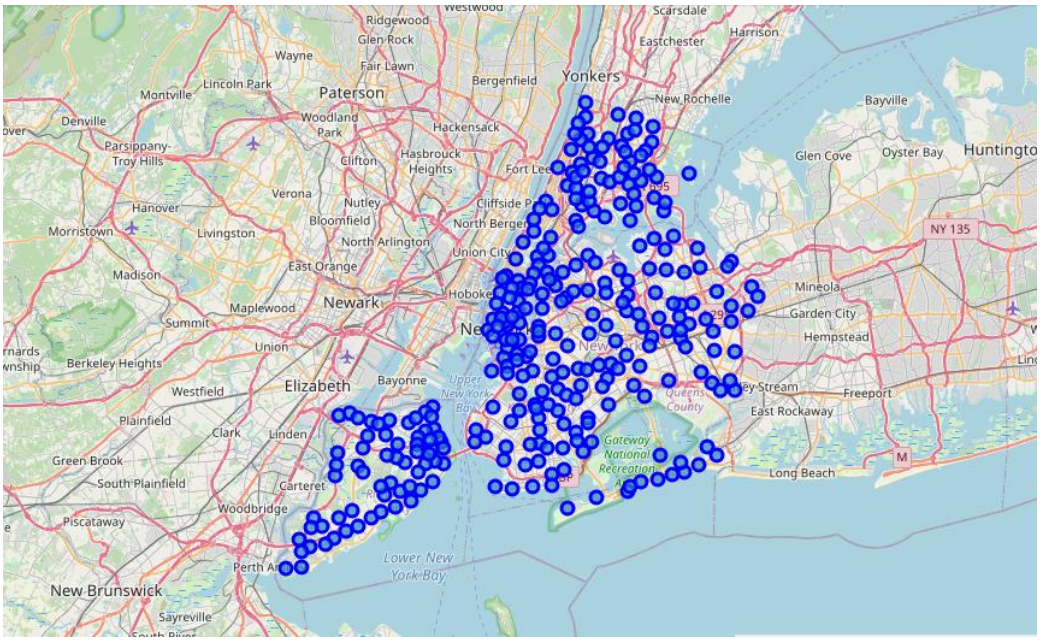
4.1 Mapping with Folium:

The first step of our analysis was to create a map centred on New York, assigning each neighbourhood a marker.

```
# create map of New York using latitude and longitude values
map_newyork = folium.Map(location=[latitude, longitude], zoom_start=10)

# add markers to map
for lat, lng, borough, neighborhood in zip(neighborhoods['Latitude'], neighborhoods['Longitude'], neighborhoods['Borough'],
                                           neighborhoods['Neighborhood']):
    label = '{}, {}'.format(neighborhood, borough)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_newyork)

map_newyork
```



4.2 K-Means clustering:

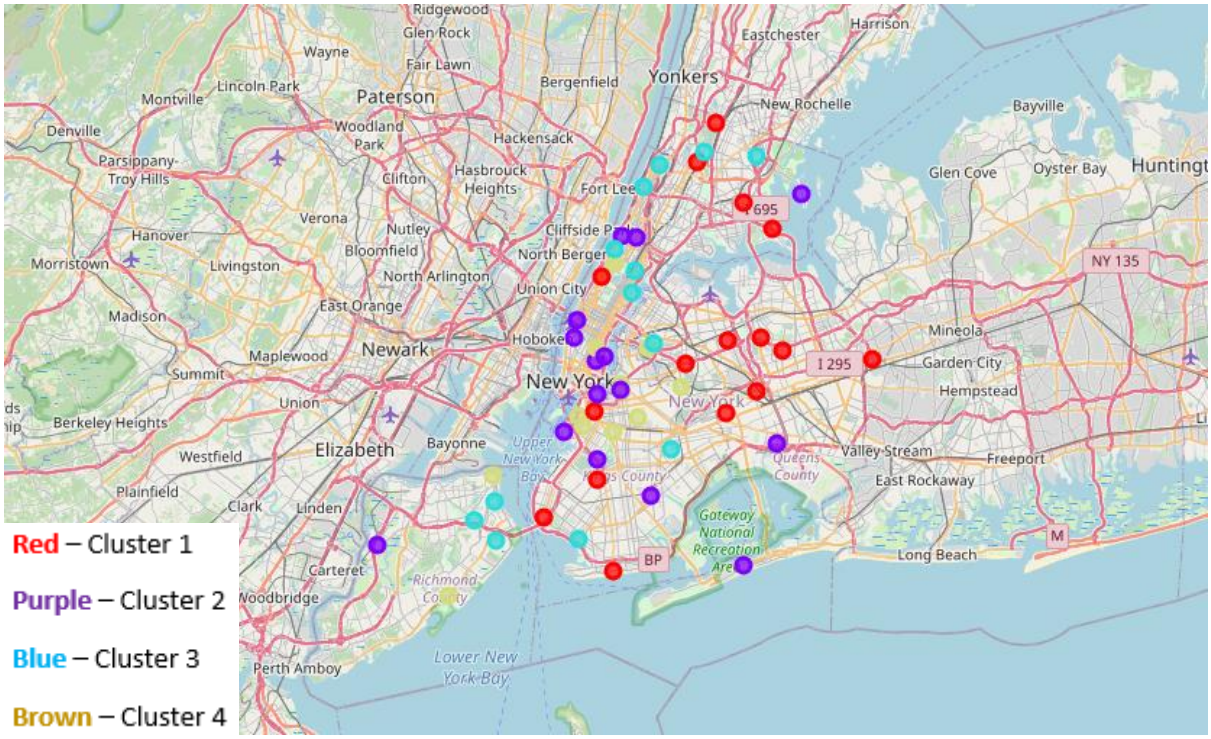
4.2.1 Mapping with Folium:

We then applied K-Means clustering to cluster the neighbourhoods into 4 groups/clusters (by setting n_clusters = 4).

```
#Aplpy Kmeans Clustering
k_means_1 = KMeans(init="k-means++", n_clusters = 4, n_init = 12)
k_means_1.fit(NY_Parents_Children)
labels = k_means_1.labels_

NY_Parents_Children["Labels"] = labels
NY_Parents_Children.head()
```

We then used folium again to create another map visualisation. This time, each suitable neighbourhood is assigned a marker colour to signify the different groups/ clusters.



4.2.2 Clustered Data:

Neighborhood	Bar	Movie Theater	Park	Playground	Pub	Restaurant	School	Sports Club	Supermarket	Toy / Game Store	Labels
Bath Beach	0.0	0.0	0.500000	0.000000	0.000000	0.500000	0.000000	0.0	0.000000	0.0	0
Co-op City	0.0	0.0	0.500000	0.000000	0.000000	0.500000	0.000000	0.0	0.000000	0.0	0
East Harlem	0.0	0.0	0.500000	0.000000	0.000000	0.500000	0.000000	0.0	0.000000	0.0	0
Grasmere	0.0	0.0	0.500000	0.000000	0.000000	0.500000	0.000000	0.0	0.000000	0.0	0
Inwood	0.2	0.0	0.400000	0.000000	0.000000	0.400000	0.000000	0.0	0.000000	0.0	0
Morningside Heights	0.0	0.0	0.750000	0.000000	0.250000	0.000000	0.000000	0.0	0.000000	0.0	0
Norwood	0.0	0.0	0.750000	0.000000	0.000000	0.250000	0.000000	0.0	0.000000	0.0	0
Roosevelt Island	0.0	0.0	0.333333	0.166667	0.000000	0.166667	0.166667	0.0	0.166667	0.0	0
Sunnyside	0.0	0.0	0.333333	0.000000	0.333333	0.333333	0.000000	0.0	0.000000	0.0	0
Washington Heights	0.0	0.0	0.666667	0.000000	0.000000	0.333333	0.000000	0.0	0.000000	0.0	0
Woodhaven	0.2	0.0	0.400000	0.000000	0.000000	0.200000	0.000000	0.0	0.200000	0.0	0
Yorkville	0.0	0.0	0.666667	0.000000	0.333333	0.000000	0.000000	0.0	0.000000	0.0	0

Data - Cluster 1

Neighborhood	Bar	Movie Theater	Park	Playground	Pub	Restaurant	School	Sports Club	Supermarket	Toy / Game Store	Labels
Auburndale	0.333333	0.000000	0.000000	0.000000	0.0	0.000000	0.0	0.0	0.333333	0.333333	1
Blissville	0.333333	0.333333	0.000000	0.000000	0.0	0.333333	0.0	0.0	0.000000	0.000000	1
Brighton Beach	0.000000	0.000000	0.000000	0.250000	0.0	0.500000	0.0	0.0	0.250000	0.000000	1
Carroll Gardens	0.400000	0.000000	0.200000	0.200000	0.0	0.200000	0.0	0.0	0.000000	0.000000	1
Corona	0.000000	0.000000	0.250000	0.000000	0.0	0.250000	0.0	0.0	0.500000	0.000000	1
Downtown	0.000000	0.333333	0.000000	0.000000	0.0	0.333333	0.0	0.0	0.000000	0.333333	1
Fort Hamilton	0.200000	0.000000	0.200000	0.000000	0.2	0.200000	0.0	0.0	0.200000	0.000000	1
Gramercy	0.250000	0.000000	0.250000	0.250000	0.0	0.250000	0.0	0.0	0.000000	0.000000	1
Kew Gardens Hills	0.000000	0.000000	0.000000	0.500000	0.0	0.500000	0.0	0.0	0.000000	0.000000	1
Ocean Parkway	0.000000	0.000000	0.000000	0.333333	0.0	0.333333	0.0	0.0	0.333333	0.000000	1
Pomonok	0.333333	0.000000	0.000000	0.333333	0.0	0.000000	0.0	0.0	0.333333	0.000000	1
Queensboro Hill	0.250000	0.000000	0.250000	0.250000	0.0	0.000000	0.0	0.0	0.250000	0.000000	1
Ridgewood	0.333333	0.000000	0.000000	0.333333	0.0	0.333333	0.0	0.0	0.000000	0.000000	1
South Side	0.333333	0.000000	0.000000	0.000000	0.0	0.333333	0.0	0.0	0.000000	0.333333	1
Stapleton	0.333333	0.000000	0.333333	0.000000	0.0	0.333333	0.0	0.0	0.000000	0.000000	1

Data - Cluster 2

Neighborhood	Bar	Movie Theater	Park	Playground	Pub	Restaurant	School	Sports Club	Supermarket	Toy / Game Store	Labels
Bedford Stuyvesant	0.500000	0.0	0.250000	0.250000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
Central Harlem	0.500000	0.0	0.500000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
Chelsea	0.500000	0.0	0.500000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
City Island	0.500000	0.0	0.500000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
Cobble Hill	0.666667	0.0	0.000000	0.333333	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
East Village	0.500000	0.0	0.500000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
Edgewater Park	0.333333	0.0	0.333333	0.000000	0.333333	0.0	0.0	0.0	0.000000	0.000000	2
Flatlands	0.500000	0.0	0.500000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
Hudson Yards	0.333333	0.0	0.333333	0.000000	0.000000	0.0	0.0	0.0	0.333333	0.000000	2
Manhattan Valley	0.500000	0.0	0.250000	0.250000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
Manhattanville	0.333333	0.0	0.333333	0.000000	0.000000	0.0	0.0	0.0	0.333333	0.000000	2
Prospect Heights	0.800000	0.0	0.000000	0.200000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
Red Hook	0.500000	0.0	0.333333	0.000000	0.166667	0.0	0.0	0.0	0.000000	0.000000	2
South Ozone Park	0.400000	0.0	0.600000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
St. George	0.666667	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.333333	2
Stuyvesant Town	0.500000	0.0	0.500000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
Vinegar Hill	0.500000	0.0	0.500000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
Williamsburg	0.500000	0.0	0.500000	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2
Windsor Terrace	0.333333	0.0	0.666667	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	2

Data - Cluster 3

Neighborhood	Bar	Movie Theater	Park	Playground	Pub	Restaurant	School	Sports Club	Supermarket	Toy / Game Store	Labels
Bedford Park	0.000000	0.000000	0.250	0.00	0.250000	0.0	0.0	0.0	0.50	0.0	3
Bellerose	0.000000	0.000000	0.000	0.50	0.500000	0.0	0.0	0.0	0.00	0.0	3
Kensington	0.000000	0.000000	0.250	0.25	0.250000	0.0	0.0	0.0	0.25	0.0	3
Kew Gardens	0.250000	0.250000	0.000	0.00	0.250000	0.0	0.0	0.0	0.25	0.0	3
Maspeth	0.000000	0.000000	0.000	0.50	0.500000	0.0	0.0	0.0	0.00	0.0	3
Upper West Side	0.333333	0.333333	0.000	0.00	0.333333	0.0	0.0	0.0	0.00	0.0	3
Westchester Square	0.250000	0.000000	0.250	0.00	0.250000	0.0	0.0	0.0	0.25	0.0	3
Woodlawn	0.125000	0.000000	0.125	0.25	0.500000	0.0	0.0	0.0	0.00	0.0	3

Data - Cluster 4

4.3 Cluster Analysis:

The 4 clusters are then analysed to determine the advantages and disadvantages that neighbourhoods in each group/cluster have.

```
#Examine clusters

C1 = NY_Parents_Children[NY_Parents_Children["Labels"] ==0]
C2 = NY_Parents_Children[NY_Parents_Children["Labels"] ==1]
C3 = NY_Parents_Children[NY_Parents_Children["Labels"] ==2]
C4 = NY_Parents_Children[NY_Parents_Children["Labels"] ==3]

print("There are {} neighborhoods in cluster 1".format(C1.shape[0]))
print("There are {} neighborhoods in cluster 2".format(C2.shape[0]))
print("There are {} neighborhoods in cluster 3".format(C3.shape[0]))
print("There are {} neighborhoods in cluster 4".format(C4.shape[0]))
```

5. Conclusions:

It is clearly visible from the folium generated map, that there are many neighbourhoods in New York City that meet the minimum requirements outlined to raise a young family. These neighbourhoods are in all areas of the city (North, East, South and West).

We have grouped/clustered the 54 neighbourhoods into 4 clusters (cluster 1, cluster 2, cluster 3 and cluster 4). Each cluster (group of neighbourhoods) has some advantages and disadvantages relative to the other clusters.

Neighbourhoods that are in cluster 1 have a relatively large number of restaurants and parks compared to the other clusters. However there are not many pubs or restaurants in these neighbourhoods. There are no sports clubs, Toy Stores or Movie Theatres in these neighbourhoods.

Neighbourhoods that are in cluster 2 have a relatively large number of restaurants and bars compared to the other clusters. These neighbourhoods might be better suited to families with older children, where parks and playgrounds are not as important.

Neighbourhoods that are in cluster 3 have a relatively large number of parks and bars compared to the other clusters. This means that it has amenities suitable for both adults and children. Another advantage of cluster 3 neighbourhoods is that there are a large amount of suitable neighbourhoods (19) relative to the other clusters – cluster 1 (12), cluster 2 (15) and cluster 3 (8). This would likely make it easier to find a suitable home in these neighbourhoods.

Neighbourhoods in cluster 4 have a higher density of pubs compared to the other clusters. However, there are not many other desired amenities in these areas.

To conclude, there are 54 suitable neighbourhoods in New York City that meet the minimum requirements to raise a young family. Of these 54 neighbourhoods, the 19 that were grouped in cluster 3 appear to be the most suitable in my opinion, for the reasons already outlined in this report.

6. Future Study:

The initial analysis presented for this project does not consider house prices in the analysis. However, it may be worth including another dataset containing historical regional house price data, as that would obviously have a big impact on the decision of which neighbourhood to live in.