Applied Data Science Capstone
Assignment: Capstone Project - The Battle of Neighborhoods
Final Report

Stefano Paoli, November 2019

1. Introduction/Business Problem

A (hypothetical) client is a dentist who would like to open a new dental office in the New York area.

He has all the expertise and capital to open the clinic and is asking us to recommend one or more places where to establish the clinic.

The client has two requirements:

1) The business must be located in an area that presents good business opportunities and allow him to get enough patients
2) Ideally, the location would be (or not far from) a nice place to live for a family, as he would like to live near his work.

To address the first requirement, we look for locations that are relatively underserved, i.e. have a lower number of existing dental offices per capita. We also factor in economic data, like the median household income levels, as we can expect that a more affluent population creates better business conditions for a dental office.

To address the second requirement, we look at various aspects like: the presence of schools, parks, shopping and restaurants, and we look for lower levels of criminality.

2. Data

The primary source for this analysis is the Foursquare Places API, from which we get data about existing dental offices and neighborhood data like the presence of schools, shopping, parks etc.

For other demographic and social data we use government sources.

We will use the data to identify places that have the lowest level of dental office per capita, but the highest level of income. As a second step we will look at which one also offer the best family living environment by offering higher number of schools, shopping and restaurants, but lowest level of criminality.

The following section describes sources and data in more detail.

1.1.Data sources

Foursquare:

- We use the Venues endpoint group to get number of venues for the categories of interest: Dentist's Office, Arts & Entertainment, Parks, etc.

NYC Open Data:

- We use this source to get population data
- Source: https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp/data
- File: New_York_City_Population_By_Neighborhood_Tabulation_Areas.csv

NYU Furman Center:

- We use this source to get social and economic data
- Source: http://app.coredata.nyc
- Income: borough-medianhouseholdincome2018.csv
- Poverty rate: borough-povertyrate.csv
- Unemployment: borough-unemploymentrate.csv
- Crime rate: borough-seriouscrimerateper1000residents.csv
- Population density: borough-populationdensity1000personspersquaremile.csv

We may use additional sources for the small town in the NY area we want to use as comparative option. In general, we can use the official USA census data source https://data.census.gov/cedsci/.

1.2.Data cleaning

For **socioeconomic data** we used the data from the NYU Furman Center. As we were not able to find data at individual neighborhood for all parameters, we chose to use the data at borough level; we also verified that data is generally consistent within a borough.

The csv files were imported and merged into a single dataframe. We used the reference year 2017, as it was the one for which data for all parameters was available.

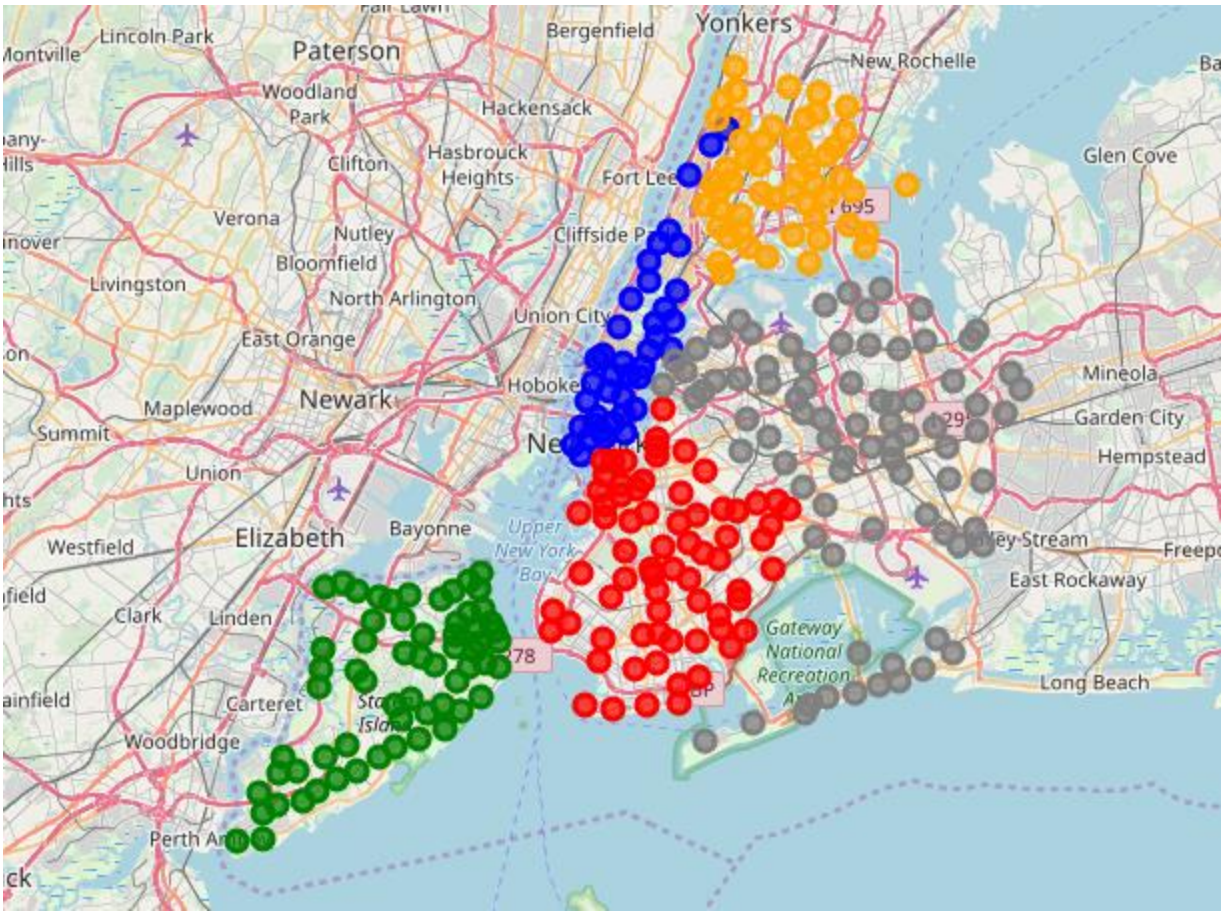
For **business and venues information** we used the Foursquare Paces API, specifically the Venues-Explore endpoint.

3. Methodology

The key point of this analysis was to identify neighborhoods that offer the best environment for a new dentist office. Therefore, we look for neighborhoods with low number of established businesses. We then compare with the other socioeconomic factors to identify the most attractive neighborhoods.

We load the file “newyork_data.json” that contains the names and coordinates of the 5 boroughs and 306 neighborhoods in NYC.

The following map of New York indicates each neighborhood that will be studied – using a different color for each borough.



From the sources and files indicated in previous section we build a dataframe that contains all **socioeconomic** data for the 5 boroughs.

```
nyc_socioeconomic
```

	Borough	median_household_income	pop_density	unemployment	crimerateper1000residents	poverty_rate
0	Manhattan	86693	73.479122	0.053543	16.458252	0.162231
1	Bronx	38110	34.985434	0.109212	14.552795	0.280333
2	Brooklyn	58027	37.941149	0.064241	11.281601	0.198106
3	Queens	65739	21.684484	0.051791	8.611700	0.121145
4	Staten Island	80711	8.241244	0.043181	5.970494	0.117867

We used Foursquare's API https://api.foursquare.com/v2/venues/explore to get the list of dentists in each neighborhoods. In the API we passed:

category = '4bf58dd8d48988d178941735' # 'Dentist's Office'

radius = 500

LIMIT = 40 # limit of number of venues returned by Foursquare API

time ='any'

day = 'any' # any day of the week, not current day

Note that, out of 306 neighborhoods, for 74 neighborhoods the search returned no results.

The search returned a total of 1542 dentist offices.

```
print(NYC_venues.shape)
NYC_venues.head()
```

```
(1542, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Co-op City	40.874294	-73.829939	Dental Group NY	40.875545	-73.829761	Dentist's Office
1	Co-op City	40.874294	-73.829939	Advanced Dental Group	40.875545	-73.829761	Dentist's Office
2	Co-op City	40.874294	-73.829939	Cohen Gentle Dental	40.871569	-73.830243	Dentist's Office
3	Co-op City	40.874294	-73.829939	Smile-Savers Pediatric Dentistry	40.877143	-73.828029	Dentist's Office
4	Co-op City	40.874294	-73.829939	City Smiles Dental	40.870210	-73.827829	Dentist's Office

We elaborated the dataframe to obtain a dataframe with list of neighborhood and number of dentists in each:

```
NYC_dentists_nr
```

	Neighborhood	Dentists
0	Allerton	4
1	Annadale	2
2	Arden Heights	3
3	Arrochar	2
4	Astoria	13
5	Auburndale	3
6	Bath Beach	7

At this point we used ML algorithm k-means to cluster the neighborhood into 5 clusters, according to the number of dentists. All neighborhoods with no dentist office (found in Foursquare) where identified as an additional cluster (cluster label 5). The picture below represents a sample:

```
NYC_merged.head(10)
```

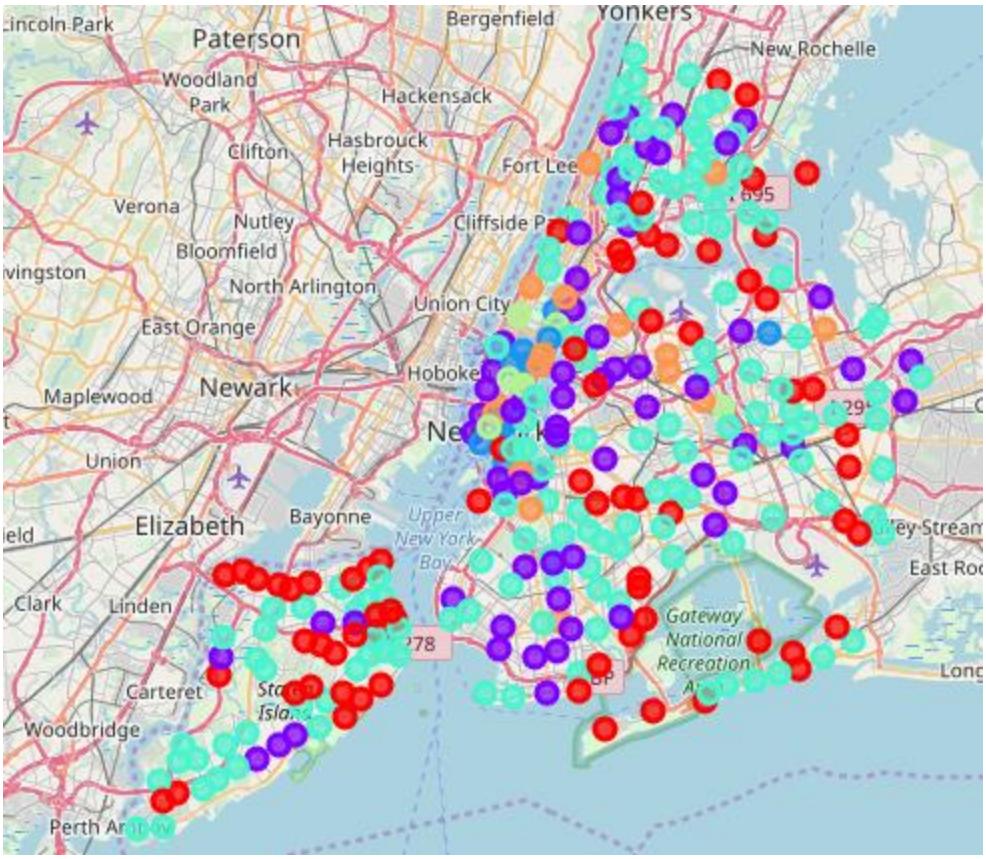
	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Dentists
0	Bronx	Wakefield	40.894705	-73.847201	5	0
1	Bronx	Co-op City	40.874294	-73.829939	0	5
2	Bronx	Eastchester	40.887556	-73.827806	5	0
3	Bronx	Fieldston	40.895437	-73.905643	2	1
4	Bronx	Riverdale	40.890834	-73.912585	2	4
5	Bronx	Kingsbridge	40.881687	-73.902818	2	4
6	Manhattan	Marble Hill	40.876551	-73.910660	0	5
7	Bronx	Woodlawn	40.898273	-73.867315	2	2
8	Bronx	Norwood	40.877224	-73.879391	0	7
9	Bronx	Williamsbridge	40.881039	-73.857446	2	1

The table below shows the statistics of the identified clusters.

```
NYC_merged[['Cluster Labels','Dentists']].groupby('Cluster Labels').describe()
```

	Dentists							
	count	mean	std	min	25%	50%	75%	max
Cluster Labels								
0	63.0	6.984127	1.853437	5.0	5.00	6.0	8.00	11.0
1	10.0	40.600000	2.412928	38.0	39.25	40.0	40.00	45.0
2	141.0	2.226950	1.110783	1.0	1.00	2.0	3.00	4.0
3	6.0	32.166667	2.316607	29.0	30.50	32.5	33.75	35.0
4	16.0	16.000000	2.804758	12.0	14.00	15.5	17.25	23.0
5	70.0	0.000000	0.000000	0.0	0.00	0.0	0.00	0.0

The clusters are well identified on the map below; red color identifies cluster 5 (with no dentists) and green color identifies cluster 2 (with mean 2.2 dentists per neighbordhood).



As we see form the count below, Staten Island and Queens contain the highest concentration of underserved neighborhoods:

```
clusters[2].Borough.value_counts()
```

```
Queens           38
Brooklyn         33
Bronx            32
Staten Island    31
Manhattan         7
Name: Borough, dtype: int64
```

```
clusters[5].Borough.value_counts()
```

```
Staten Island    26
Queens           19
Brooklyn         13
Bronx            11
Manhattan         1
Name: Borough, dtype: int64
```

From the socioeconomic data table we see that Staten Island represents an interesting option as it has the second highest household income, lowest poverty rate and lowest unemployment, which adds to the business opportunity in terms of family spending capabilities. It also shows the lowest crime rate, which is appealing from a family living perspective.

As our client has expressed preference for Staten Island, we focus on this borough for further analysis; the Staten Island's neighborhoods with no dentists are shown in the map below.

To analyze these neighborhoods from a family living perspective we queried Foursquare looking for Outdoors & Recreation venues:

categoryId = Outdoors & Recreation = 4d4b7105d754a06377d81259

We obtained the list of such venues for each neighborhood; the table below shows a sample:

Neighborhood	
Arlington	6
Clifton	5
Egbertville	3
Elm Park	2
Emerson Hill	1
Fox Hills	2
Grant City	6
Howland Hook	1
Lighthouse Hill	3

For each neighborhood we also identified the top venues like Playground, Park, Pool; sample below:

Neighborhood	Venue	Venue Category
St. George	Lt. Lia Playground	Playground
St. George	Fort Hill	Park
St. George	Maritime Hospital Quarantine Cemetery	Park
St. George	Barrett Triangle	Park
New Brighton	Bocce Courts	Park
New Brighton	Skyline Playground	Playground
New Brighton	Mahoney Park	Park
Rosebank	White playground	Park
Rosebank	DeMatti Park	Park
Todt Hill	St Francis Woodlands	Park
South Beach	Q Ave	Park
Port Richmond	Levy Playground	Playground
Mariner's Harbor	44 Bus Stop.	Playground
Travis	Dr Seuss Park!	Park

4. Results and Discussion

Our analysis identifies the neighborhoods in NYC that have the lowest number of established dental offices. When matched with other socioeconomic factors like population density and family income, it is possible to identify a subset of neighbors that offer the best opportunities for a new business.

The neighborhoods have also been analyzed to identify the ones that offer the best family environment for our client, by presenting a higher number of parks, playgrounds and sport venues.

We can offer to our client the list of neighborhoods to check out, with their specific characteristics. This list will be used by our client to execute a personal search and selection of the place of his preference.

5. Conclusion

Purpose of this project was to identify neighborhoods in NYC that would present the best business opportunity for a new business office.

The analysis has leveraged publicly available data to identify such neighborhoods. Out of 306 neighborhoods in NYC we identified 26 neighborhoods in Staten Island that represent the best match for our client preference. We have provided further information on environmental conditions of each neighborhood that will allow the client to easily select the place of his choice.