Coursera Capstone Final Project

Battle of neighborhoods

Introduction

- Our focus stays on the New York City. We have already began to explore and cluster the neighborhoods, so lets take a challenge of a real business problem in this city.
- As we know, the New York City has 5 boroughs: Manhattan, Brooklyn, Queens, The Bronx, and Staten Island. My borough of interest is Brooklyn. It is the most populous borough. Brooklyn is known for its economical, cultural, social, and ethnic diversity and distinct neighborhoods. There are more than 60 neighborhoods in Brooklyn, so it will be interesting and challenging to cluster Brooklyn and compare the neighborhoods one with another in order to solve our business case.

Business Problem

- Let us assume there is a very successful home interior shop in Brooklyn, Neighborhood Greenpoint, called "Home of the brave". The owner has an intention (and all the needed resources) to open 2 more shops in Brooklyn in different neighborhoods but he is not sure with the new locations . Let us assume there is a very successful home interior shop in Brooklyn, Neighborhood Greenpoint, called "Home of the brave". The owner has an intention (and all the needed resources) to open 2 more shops in Brooklyn in different neighborhoods but he is not sure with new locations.
- Our task will be to find the most appropriate 2 Neighborhoods in Brooklyn where the business will repeat the previous success.

Data sources

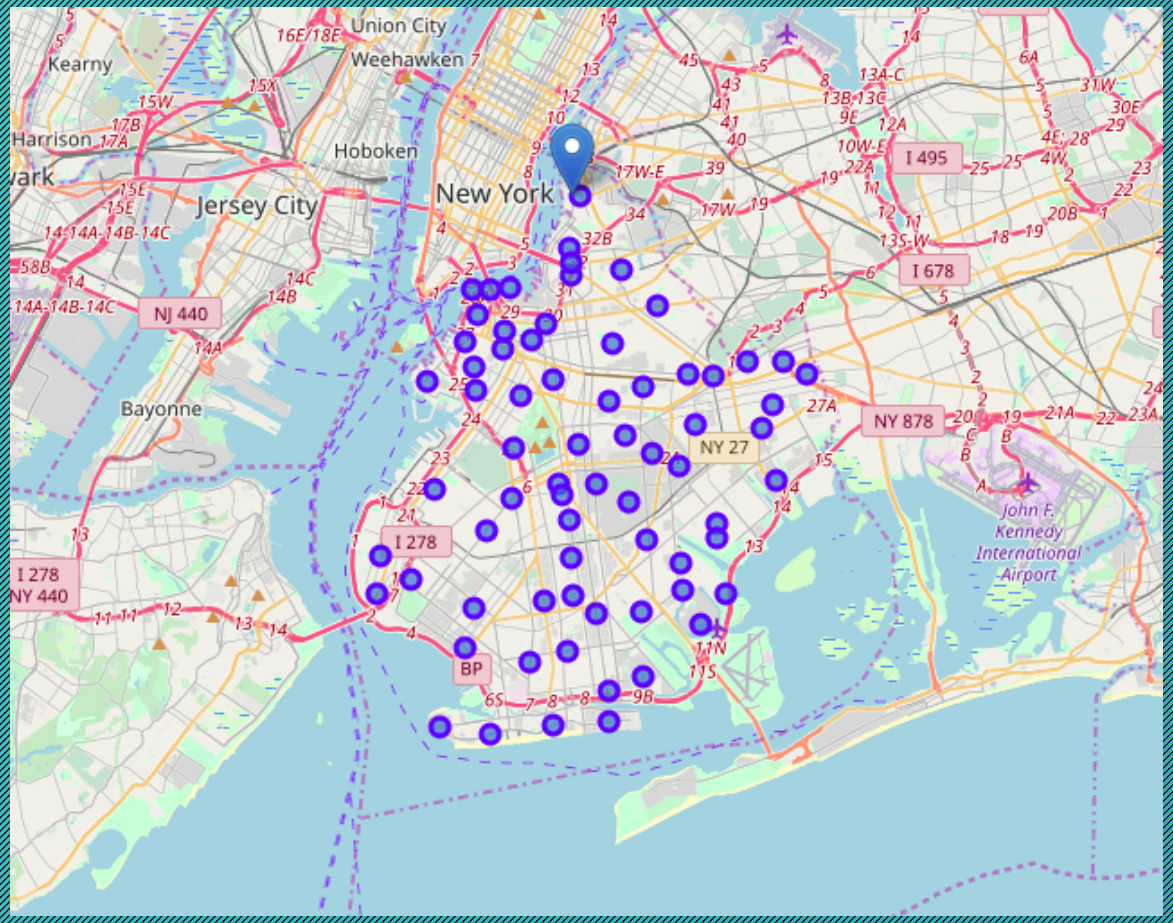
- 1. Coordinates of all 5 boroughs of New York and their corresponding neighborhoods, available for free on https://geo.nyu.edu/catalog/nyu_2451_34572
- 2. Foursquare location data, following requests:
- https://api.foursquare.com/v2/venues/search
- https://api.foursquare.com/v2/venues/explore
- 3. The data from http://www.city-data.com/nbmaps/neigh-Brooklyn-New-York.html , from year 2016 contains population, area, income, housing prices information for Brooklyn by neighborhood. This data will build our main data set that will be used for clustering.

1. Get the data set with coordinates of neighborhoods in Brooklyn:

	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471

```
In [7]: Brooklyn_df.shape
Out[7]: (70, 4)
```

2. Visualize the neighborhoods of Brooklyn on Folium map



3. Get the main dataset with neighborhood‘s features:

	Neighborhood	Area sq.m.	Population	Pop. density	Median household income	Median rent	Number males	Number females	Median age males	Median age females
0	Bath Beach	0.376	18331	48805	67622	1204	9155	9176	36.4	40.2
1	Bay Ridge	1.779	83083	46696	63178	1269	39460	43622	38.9	41.5
2	Bedford Stuyvesant	2.782	178627	64212	45546	1112	83031	95596	31.4	35.0
3	Bensonhurst	5.788	259876	44899	50788	1146	127175	132701	36.0	38.2
4	Bergen Beach	1.072	3619	3377	76466	1646	1740	1879	41.4	43.0

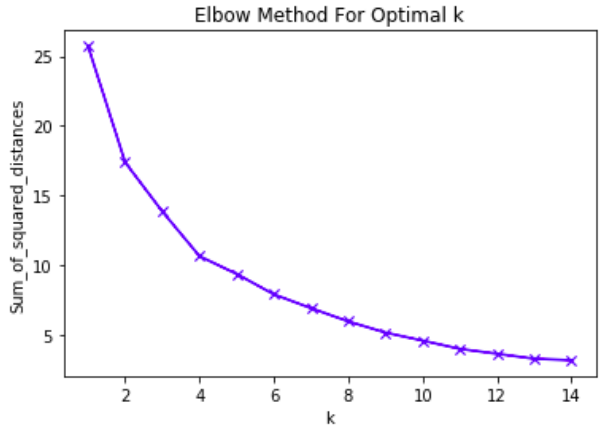
4. Get the feature „Restaurants per Neighborhood“ from Foursquare and merge 2 dataframes together

	Neighborhood	Area sq.m.	Population	Pop. density	Median household income	Median rent	Number males	Number females	Median age males	Median age females	Restaurants per 1000 residents
0	Bath Beach	0.376	18331.0	48805.0	67622.0	1204.0	9155.0	9176.0	36.4	40.2	0.818286
1	Bay Ridge	1.779	83083.0	46696.0	63178.0	1269.0	39460.0	43622.0	38.9	41.5	0.337012
2	Bedford Stuyvesant	2.782	178627.0	64212.0	45546.0	1112.0	83031.0	95596.0	31.4	35.0	0.011197
3	Bensonhurst	5.788	259876.0	44899.0	50788.0	1146.0	127175.0	132701.0	36.0	38.2	0.034632
4	Bergen Beach	1.072	3619.0	3377.0	76466.0	1646.0	1740.0	1879.0	41.4	43.0	0.000000
5	Boerum Hill	0.431	26928.0	62491.0	104214.0	1785.0	12748.0	14786.0	34.7	36.7	0.594177

5. Find optimal number of clusters with Elbow method. It results in 6 clusters.

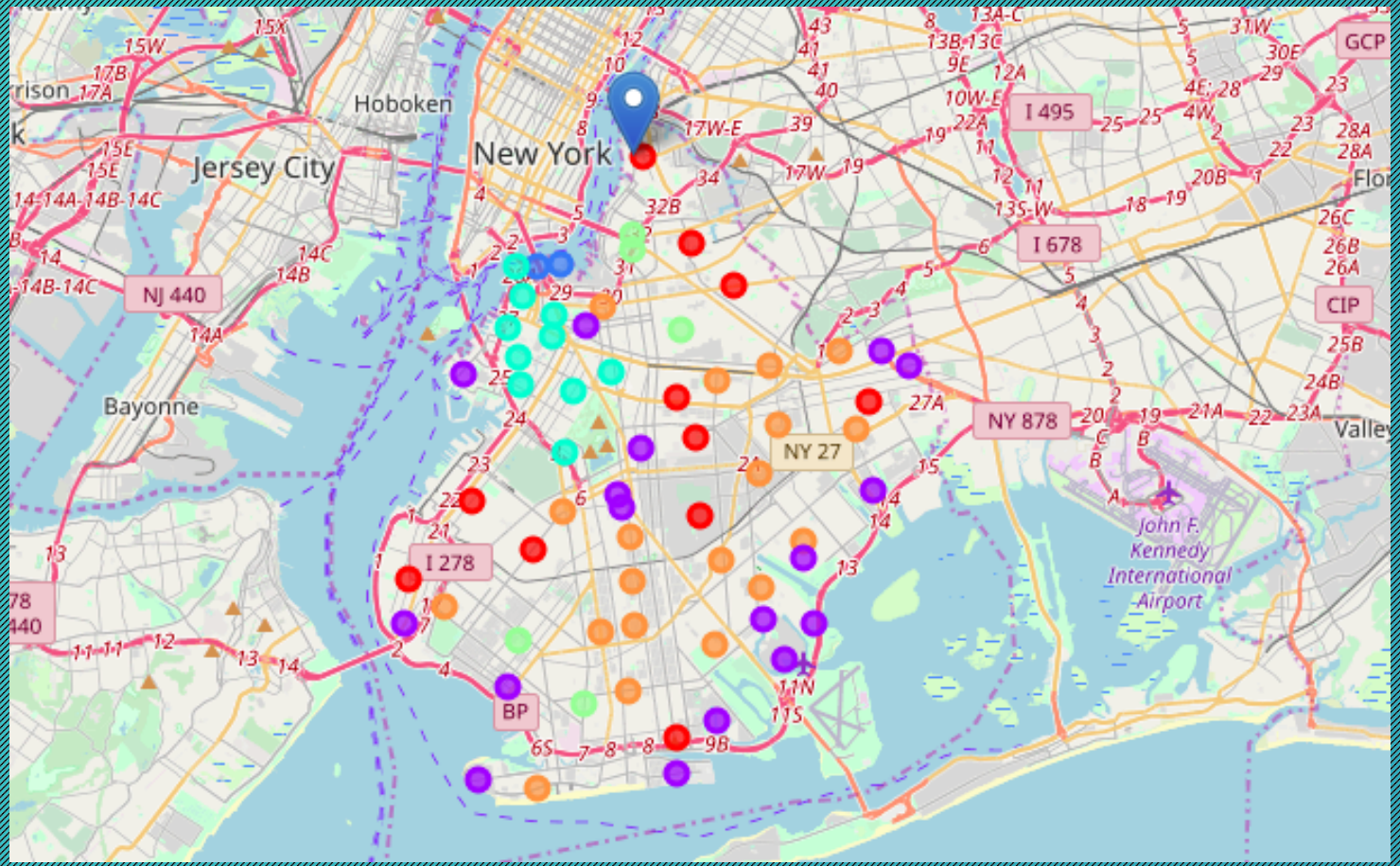
In [32]:
```
#find the sum of squared distances:
Sum_of_squared_distances = []
K = range(1,15)
for k in K:
    km = KMeans(n_clusters=k)
    km = km.fit(data_transformed)
    Sum_of_squared_distances.append(km.inertia_)
```

In [33]:
```
#plot the results
plt.plot(K, Sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('Sum_of_squared_distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```



The Elbow is at 6, so it is the optimal k for our data set.

6. Apply K-Means clustering to the dataframe and visualize it:



7. Find the cluster where the neighborhood Greenpoint lays. It is 0:

	Neighborhood	Area sq.m.	Population	Pop. density	Median household income	Median rent	Number males	Number females	Median age males	Median age females	Restaurants per 1000 residents	Cluster labels	Borough	Latitude	Longitude
1	Bay Ridge	1.779	83083.0	46696.0	63178.0	1269.0	39460.0	43622.0	38.9	41.5	0.337012	0	Brooklyn	40.625801	-74.030621
6	Borough Park	2.071	135597.0	65487.0	40212.0	1163.0	68904.0	66693.0	28.2	30.1	0.036874	0	Brooklyn	40.633131	-73.990498
9	Bushwick	1.305	85392.0	65452.0	42470.0	1178.0	42185.0	43206.0	29.7	33.0	0.210793	0	Brooklyn	40.698116	-73.925258
16	Crown Heights	1.418	85886.0	60570.0	45776.0	1020.0	38794.0	47092.0	31.4	36.5	0.000000	0	Brooklyn	40.670829	-73.943291
22	East Flatbush	2.887	132692.0	45969.0	45630.0	1107.0	62013.0	74444.0	34.8	40.7	0.022609	0	Brooklyn	40.641718	-73.936103
23	East New York	1.867	89017.0	47678.0	36773.0	1010.0	40300.0	48717.0	29.6	34.5	0.056169	0	Brooklyn	40.669926	-73.880699
24	East Williamsburg	2.508	96265.0	38382.0	58314.0	1496.0	47334.0	49119.0	30.9	33.5	0.124656	0	Brooklyn	40.708492	-73.938858
55	Sheepshead Bay	3.108	126369.0	40660.0	55781.0	1133.0	60880.0	65489.0	37.8	41.1	0.055393	0	Brooklyn	40.586890	-73.943186
57	Sunset Park	1.581	91456.0	57830.0	48323.0	1241.0	46869.0	44586.0	32.4	34.1	0.076540	0	Brooklyn	40.645103	-74.010316
63	Wingate	1.885	111322.0	59051.0	42056.0	1096.0	49938.0	61383.0	32.3	37.9	0.026949	0	Brooklyn	40.660947	-

8. Find the neighborhoods from cluster 0, that already have interior shops. They will be removed from the list:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	
0	Borough Park	40.633131	-73.990498	AJ Madison	40.641880	-73.984512	Furniture / Home Store
1	East Flatbush	40.641718	-73.936103	HomeGoods	40.631514	-73.946310	Furniture / Home Store
2	East Flatbush	40.641718	-73.936103	Courts (Furniture, Electronics, & Appliances)	40.650529	-73.950617	Furniture / Home Store
3	East New York	40.669926	-73.880699	Pier 1 Imports	40.653062	-73.872661	Furniture / Home Store
4	Sunset Park	40.645103	-74.010316	Cost Plus World Market	40.659293	-74.004411	Furniture / Home Store

9. There still 7 neighborhoods. It makes sense to find the mean from median household income and then pick up neighborhoods that lay above the mean

	Neighborhood	Area sq.m.	Population	Pop. density	Median household income	Median rent	Number males	Number females	Median age males	Median age females	Restaurants per 1000 residents	Cluster labels	Borough	Latitude	Longitude
0	Bay Ridge	1.779	83083.0	46696.0	63178.0	1269.0	39460.0	43622.0	38.9	41.5	0.337012	0	Brooklyn	40.625801	-74.030621
1	Borough Park	2.071	135597.0	65487.0	40212.0	1163.0	68904.0	66693.0	28.2	30.1	0.036874	0	Brooklyn	40.633131	-73.990498
2	Bushwick	1.305	85392.0	65452.0	42470.0	1178.0	42185.0	43206.0	29.7	33.0	0.210793	0	Brooklyn	40.698116	-73.925258
3	Crown Heights	1.418	85886.0	60570.0	45776.0	1020.0	38794.0	47092.0	31.4	36.5	0.000000	0	Brooklyn	40.670829	-73.943291
4	East Williamsburg	2.508	96265.0	38382.0	58314.0	1496.0	47334.0	49119.0	30.9	33.5	0.124656	0	Brooklyn	40.708492	-73.938858
5	Sheepshead Bay	3.108	126369.0	40660.0	55781.0	1133.0	60880.0	65489.0	37.8	41.1	0.055393	0	Brooklyn	40.586890	-73.943186
6	Wingate	1.885	111322.0	59051.0	42056.0	1096.0	49938.0	61383.0	32.3	37.9	0.026949	0	Brooklyn	40.660947	-73.937187

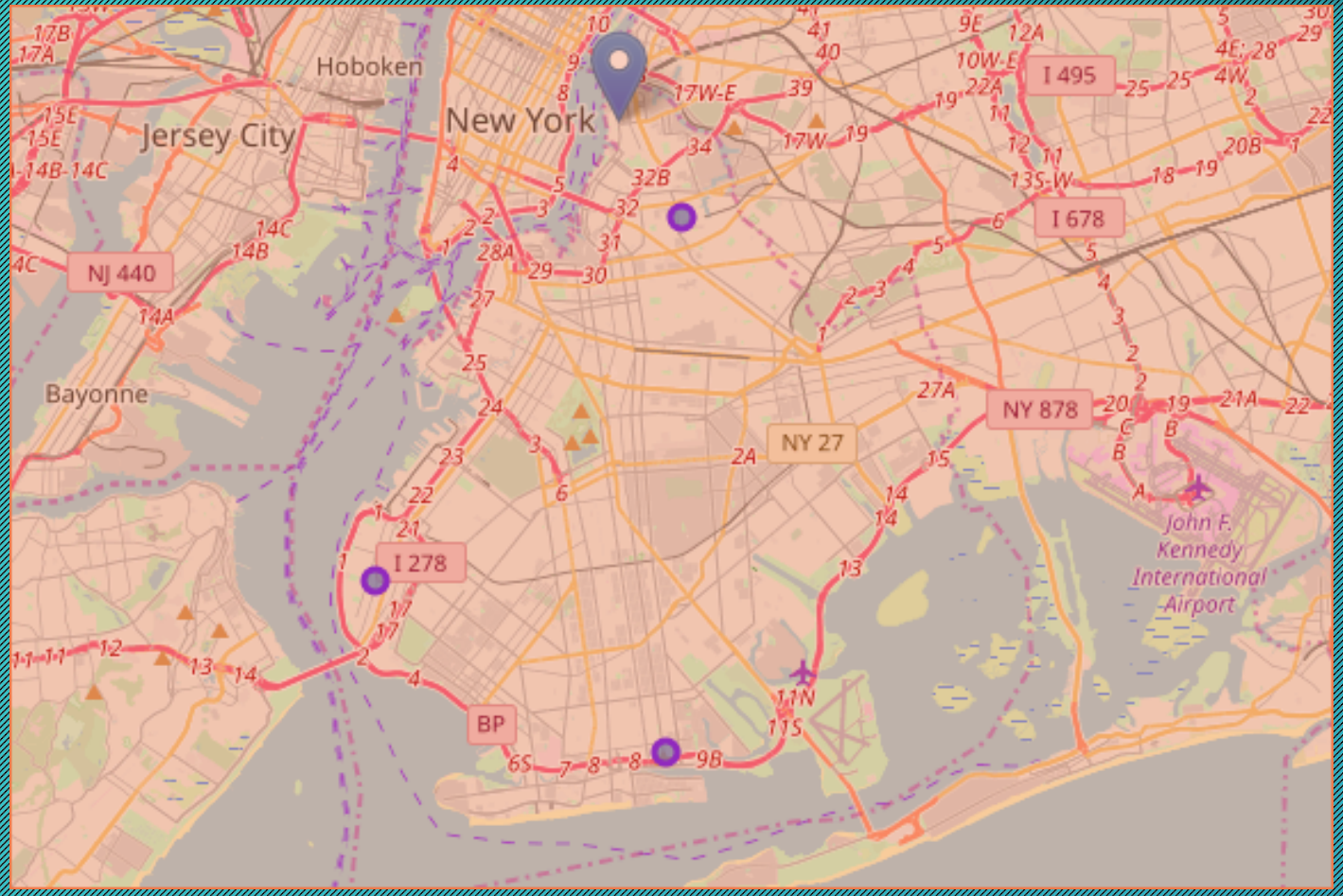
10. Find the mean of the column „Household income“. It is $49683

	index	Area sq.m.	Population	Pop. density	Median household income	Median rent	Number males	Number females	Median age males	Median age females	Restaurants per 1000 residents	Cluster labels	Latitude	Longitude
count	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.0	7.000000	7.000000
mean	24.857143	2.010571	103416.285714	53756.857143	49683.857143	1193.571429	49642.142857	53800.571429	32.742857	36.228571	0.113097	0.0	40.654887	-73.958414
std	24.558967	0.628994	21298.258815	11594.460494	9210.119189	153.647711	11399.494469	10352.919087	4.057445	4.278128	0.121907	0.0	0.042771	0.037939
min	1.000000	1.305000	83083.000000	38382.000000	40212.000000	1020.000000	38794.000000	43206.000000	28.200000	30.100000	0.000000	0.0	40.586890	-74.030621
25%	7.500000	1.598500	85639.000000	43678.000000	42263.000000	1114.500000	40822.500000	45357.000000	30.300000	33.250000	0.031911	0.0	40.629466	-73.966895
50%	16.000000	1.885000	96265.000000	59051.000000	45776.000000	1163.000000	47334.000000	49119.000000	31.400000	36.500000	0.055393	0.0	40.660947	-73.943186
75%	39.500000	2.289500	118845.500000	63011.000000	57047.500000	1223.500000	55409.000000	63436.000000	35.050000	39.500000	0.167724	0.0	40.684473	-73.938022
max	63.000000	3.108000	135597.000000	65487.000000	63178.000000	1496.000000	68904.000000	66693.000000	38.900000	41.500000	0.337012	0.0	40.	

11. Display the neighborhoods with household income above the mean:

	Neighborhood	Area sq.m.	Population	Pop. density	Median household income	Median rent	Number males	Number females	Median age males	Median age females	Restaurants per 1000 residents	Cluster labels	Borough	Latitude	Longitude
0	Bay Ridge	1.779	83083.0	46696.0	63178.0	1269.0	39460.0	43622.0	38.9	41.5	0.337012	0	Brooklyn	40.625801	-74.030621
4	East Williamsburg	2.508	96265.0	38382.0	58314.0	1496.0	47334.0	49119.0	30.9	33.5	0.124656	0	Brooklyn	40.708492	-73.938858
5	Sheepshead Bay	3.108	126369.0	40660.0	55781.0	1133.0	60880.0	65489.0	37.8	41.1	0.055393	0	Brooklyn	40.586890	-73.943186

12. Visualize the neighborhoods on Folium map and choose the most remote:



13. Results: Bay Ridge and Sheephead Bay

- East Williamsburg is located too close to Greenpoint - this location will not provide significant extension of customer audience.
- So now we have our final choice - Bay Ridge and Sheephead Bay are the best possible neighborhoods for 2 new shops "Home of the brave", what we will recommend to the owner. This 2 neighborhoods lay in the same cluster as Greenpoint, have the highest income level pro household in cluster and they are also at most remote from Greenpoint.

14. Recommendations:

- The neighborhoods Bay Ridge and Sheephead Bay are the best possible neighborhoods for 2 new shops "Home of the brave".
- Not each neighborhood in Brooklyn has an interior shop, so there are more possibilities to extend the business without enter to the market place competition.
- In the meantime it makes sense to use the advertisement in the other neighborhoods, that are located nearby in order to reach a bigger customer audience.
- Location for the new shops must be chosen with regard to convenient transport connection – for example nearby metro stations - in order to make the shops easy reachable.
- The pricing and the range of goods in the shops could be adapted to the neighborhoods' features - on the basis of mean household income and average age of the customers.
- One old and two new shops are located in different parts of the Brooklyn, far away one from another. It gives a good chance to reach a bigger customer audience and make the retail net "Home of the brave" well-known in Brooklyn.

The Project is completed.
Thank you for your attention!