BATTLE OF THE NEIGHBOURHOODS

INTRODUCTION

A new start-up company would like to open their offices in either NYC. They would like to have a good look at the quality of life in different districts of Brooklyn.

Some of the features that we would look at to compare the different districts/neighbourhoods would be the number neighbourhoods withe accessibility to transport, shopping , cafés/restaurants, sports venues, childrens' facilities and the cost of living, namely real estate.

We will use the use Foursquare API to collect the top venues of the different neighbourhoods, use K_means to cluster the neighbourhoods and then compare the similarities and dissimilarities of the different neighbourhoods of Brooklyn. We will then get an idea of the real estate prices (residential and commercial) for the clusters. This will help us get the in-depth information we need to determine which neighbourhood will be the best fit for our clients to open their new offices.

DATA

The data was acquired via the following internet links that provide the postal codes, names of neighbourhoods and their geographical coordinates. We used the Foursquare API to obtain information about the venues in these neighbourhoods to do our analysis & comparison and help us to reach a conclusion.

The NYC data was obtained from the following link and we will use Python and it's Pandas library to obtain the info we need and put it to a data frame that we can use to obtain venues from the Foursquare API.

NYC neighbourhood data: https://geo.nyu.edu/catalog/nyu_2451_34572

Once we obtained the necessary information about the clusters, we narrowed down our analysis to 2 clusters with the highest number of venues and varieties of venues. We looked at their residential and commercial real estate prices and did comparisons.

The real estate data was obtained from the site.
https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page

The data was then cleaned and filtered to obtain the dataset necessary for Brooklyn and its neighbourhoods. There was quite a bit of missing information in this dataset, so we used the mean price to replace the missing values. Once we had the real estate

dataset in order, we then merged it with the neighbourhood datasets to obtain the average real estate prices for the 2 clusters we are interested in.

METHODOLOGY

Creating clusters using K_ means clustering

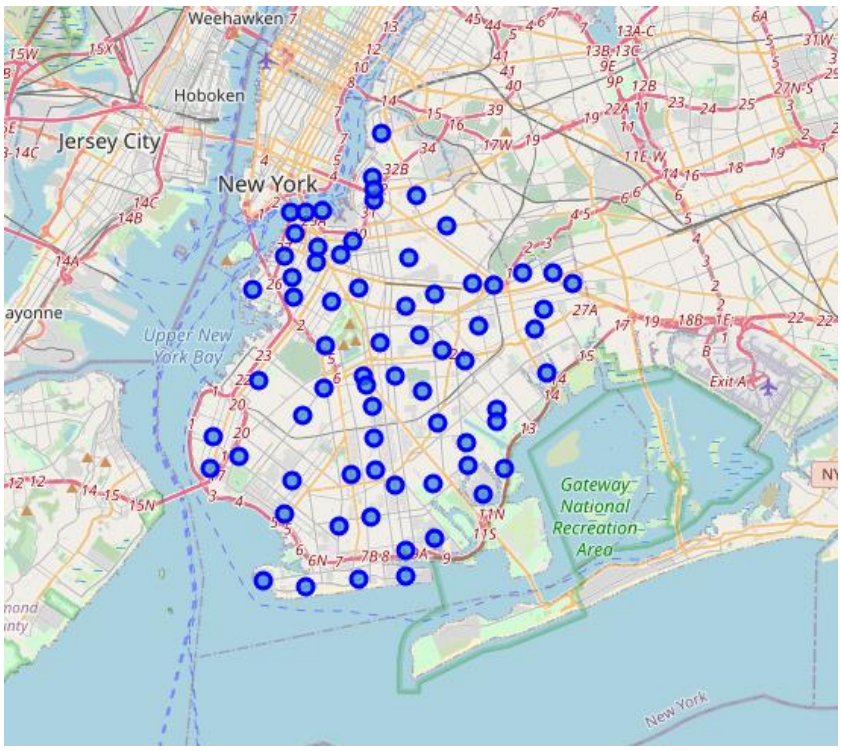
Using the above link, we scrapped the neighbourhood data for NYC and filtered it down to create a data frame that will include only the districts or neighbourhoods in Brooklyn and their geographical coordinates.

```
There are  70  districts or neighbourhoods 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

Brooklyn data frame showing the first 5 neighbourhoods out of the 70 neighbourhoods in Brooklyn

We then used Folium to create a map showing locations of these neighbourhoods.

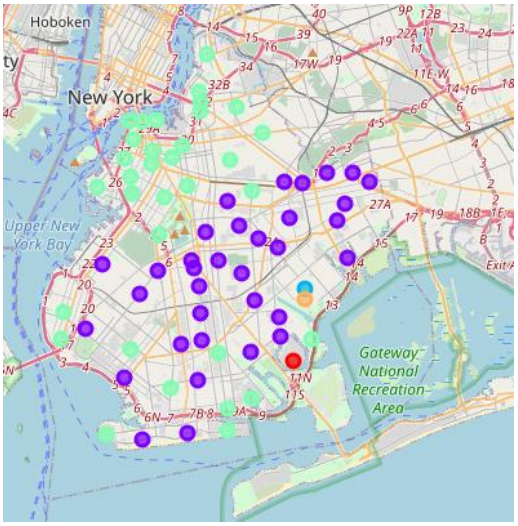


Locations of the different neighbourhoods in Brooklyn

Next, we used the Foursquare API to get the top 100 venues in Brooklyn within a radius of 500m. We found that there 6 neighbourhoods returned 100 venues (our limit) and that there are 289 different categories of venues in Brooklyn. We used one hot coding to get the unique venues in each neighbourhood and grouped them by the means of these categories. We also got the most common venues for each neighbourhood. We used this information and K_ means to divide the neighbourhoods into 5 clusters.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bath Beach	Pharmacy	Dessert Shop	Gas Station	Bubble Tea Shop	Cantonese Restaurant
1	Bay Ridge	Italian Restaurant	Spa	Pizza Place	Bar	American Restaurant
2	Bedford Stuyvesant	Coffee Shop	Café	Pizza Place	Bar	Deli / Bodega
3	Bensonhurst	Chinese Restaurant	Italian Restaurant	Park	Sushi Restaurant	Pizza Place
4	Bergen Beach	Harbor / Marina	Hockey Field	Park	Baseball Field	Athletics & Sports

5 most common venues in the first 5 neighbourhoods



5 clusters of neighbourhoods that were obtained using K means clustering

We then narrowed down our analysis to the 2 clusters (the turquoise and violet clusters) with the highest number of neighbourhoods. We did some exploratory data analysis to find out what these clusters had to offer. We used the venues data and the clusters (2 and 4) to find the count of the count of the most common venues in these clusters.

Real Estate price analysis using exploratory data analysis.

We used the real estate data set to obtain the data for Brooklyn and got the average price for residential and commercial real estate sales in Brooklyn neighbourhoods. Once we obtained this, we then merged our data with the 2 neighbourhood clusters (2 &4) and obtain just the average real estate price in these 2 clusters.

	Cluster	Avg Price
0	Cluster 2	940792.00
1	Cluster 4	1369005.59

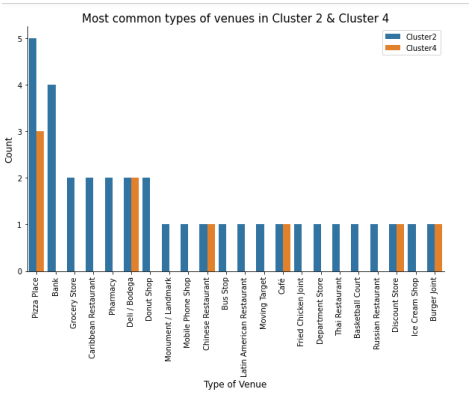
Average real estate price in USD of the neighbourhoods in cluster 2 and cluster 4

RESULTS

In the 5 clusters we created using K meaning clustering we found that,

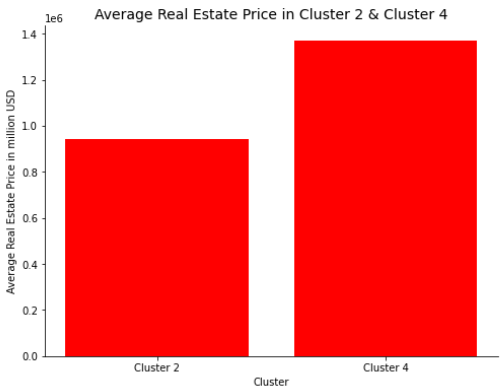
- cluster 1 has 1 neighbourhood
- cluster 2 has 35 neighbourhoods
- cluster 3 has 1 neighbourhood
- cluster 4 has 32 neighbourhoods
- cluster 5 has 1 neighbourhood

We concentrated the next part of our analysis on clusters 2 and 4. We got the count for the most common venues for these 2 clusters and the results are shown the bar graph below.



This indicates that in cluster 2, there is more choice of venues. Restaurants, transport, Banks, pharmacies, sports venues, etc.

Our results for the real estate prices for these 2 clusters is below.



The graph indicates that real estate prices are lower in cluster 2 than cluster 4.

DISCUSSION

From the above analysis, we can see that cluster 2 which contains 35 different neighbourhoods in Brooklyn has a higher potential to provide a better quality of life to its residents. It has more venues than cluster 4 (32 neighbourhoods) and a lot more variety of venues. These varieties include essential venues like transport, banks, grocery shopping, pharmacies etc and non-essential venues like important landmarks and phone shops. From the real estate price analysis, we can see that the average real estate price in the cluster 2 neighbourhoods are almost 25% less than that of the neighbourhoods in cluster 4.

CONCLUSION

Based on the above observations, I would recommend that the best choice of location to open your new offices would a neighbourhood in cluster 2. With the lower real estate prices and large variety for activities, one of these neighbourhoods in cluster 2 will provide a better quality of life than any other neighbourhood in Brooklyn.