French Bakery in New York City

1. Introduction

1. Backgroud

New York is one the biggest cities in the world. It's both one of the largest economical center of the world and one of the most popular tourist attraction. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles and a gross metropolitan product (GMP) of $2.0 trillion makes New York one of the most popular places to invest and open business there. There are many bakeries in NYC, we will conclude where are the existing bakeries and also find locations where there are none. The report will explore in which neighbourhoods and boroughs of NYC the most bakeries are located and where to open new French bakery. The aim of this project is to find the best locations for French Bakery (pâtisserie) in the New York City area. Specifically, this report will be targeted to stakeholders who are interested in **opening a French Bakery in New York, USA.**

2. Problem

Problem: A Bakery is a place where someone can usually buy breakfast or some small-medium snack. French bakery specializes in French bread, pastries and sweets which allows it to run both as morning place to buy bread and also as a place to buy takeaway or get a small-medium lunch It's a competitive market so some assumptions have been made:

1. it would be good to open new bakery where there are none or small number of bakeries

2. as a hint for finding the best place it would be good to find locations with French Restaurants as bakery won't be the competitions to the restaurant (offers small takeaway meals) but would be located near other French joints - People in the area might be interested in additional French meals.

3. Interest

To recommend the correct location, Bakery Company Ltd has appointed me to lead of the Data Science team. The objective is to locate and recommend to the management which neighborhood of New York city will be best choice to start a bakery. The Management also expects to understand the rationale of the recommendations made. This would interest anyone who wants to start a new bakery in New York.

2. Data acquisition and cleaning

1. Data sources

Data

In order to find the answers for the problems provided we need:

1. a dataset that contains all 5 NYC boroughs and all of the neighborhoods (306) that exist in each borough as well as the the latitude and longitude coordinates of each neighborhood. NY CITY

DATA SET has all these information:
https://cocl.us/new_york_dataset
DataSet is in csv format and it's free and available to download. Data set can provide us the below information:

-Borough - borough name,
-Neighborhood - neighbourhood name,
-Latitude - latitute of the neighborhood,
-Longitude - longtitude of the neighborhood
which is enough to visualize Neighbourhoods and use them to query FourSqare data using API

Example data from the DataSet:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Confirm that data frame has 5 boroughs and 306 neighbourhoods:

```
print('The dataframe has {} boroughs and {} neighborhoods.'.format(
        len(neighborhoods['Borough'].unique()),
        neighborhoods.shape[0]
    )
)
```

```
The dataframe has 5 boroughs and 306 neighborhoods.
```

2. location of current venues (we will filter the Bakeries and French Restaurants). For that we can use FourSquare API:
https://api.foursquare.com/v2/venues/explore

FourSquare API can provide us the below information:
-VenueID (which can be used to identify Venue and get Venue details),
-Venue Name - name of the Venue,
-Venue Location - lat and lng of the Venue,
-Venue categories - category of the Venue, in our case it will be used to filter Bakeries and French Restaurants

We will use the API with our CLIENT_ID and SECRET to connect to FourSquare. For each neighborhood I will get the data using my own function def getNearbyVenues(names, latitudes, longitudes, boroughs) Data is returned as JSON with the below format:
["response"]['groups'][0]['items']['venue']
where ["venue"] contains:
-v['venue']['id']

-v['venue']['name']
-v['venue']['location']['lat']
-v['venue']['location']['lng']
-v['venue']['categories']

Example data from the DataSet:

```
NY_venues.head()
```

	ID	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	4c537892fd2ea593cb077a28	Bronx	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	4c783cef3badb1f7e4244b54	Bronx	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	5d5f5044d0ae1c0008f043c3	Bronx	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	4d33665fb6093704b80001e0	Bronx	Wakefield	40.894705	-73.847201	Subway	40.890468	-73.849152	Sandwich Place
4	508af256e4b0578944c87392	Bronx	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant	40.898083	-73.850259	Caribbean Restaurant

2. Data cleaning

Data downloaded from both 2.1.1 and 2.1.2 Data Sources do not require special cleaning. Only the missing values (nan/0) requires replacement on some steps. Examples:

Dropping neighbourhoods without bakeries (where count() is equal to NaN)

Exactly opposite – dropping neighbourhoods where value is not equal 0 to find neighbourhoods without any bakery

3. Feature selection

For the 2.1.1 data set all features regarding localization, name of the borough and name of the neighbourhood were selected as these are the only ones required to divide the city by neighbourhoods.

For the 2.1.2 data set I have selected: VenueID, VenueLocation, VenueCategory and VenueName. For future analysis and project development I have already selected the data that can be used in future models: number of likes of each Venue, number of Tips and Average Score of the Venue. Data has been saved in csv file to load it quickly in the future.

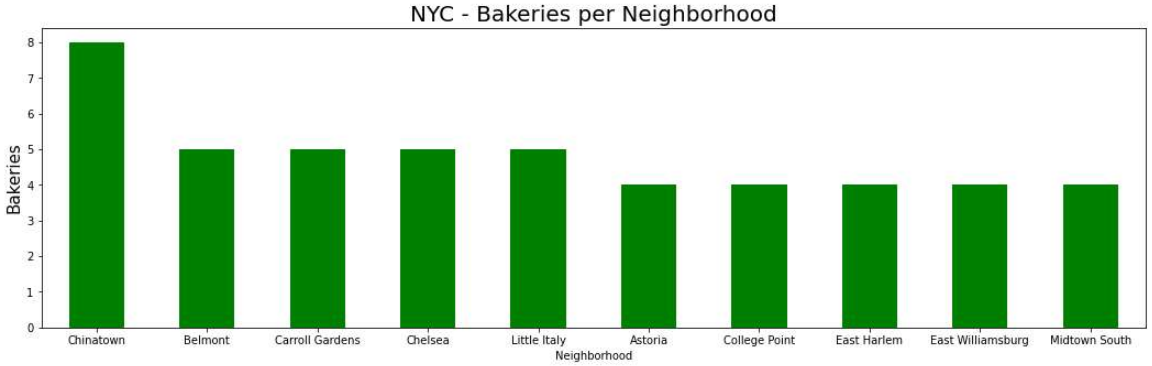
4. Methodology / Strategy

Our Strategy for problem solving will be:

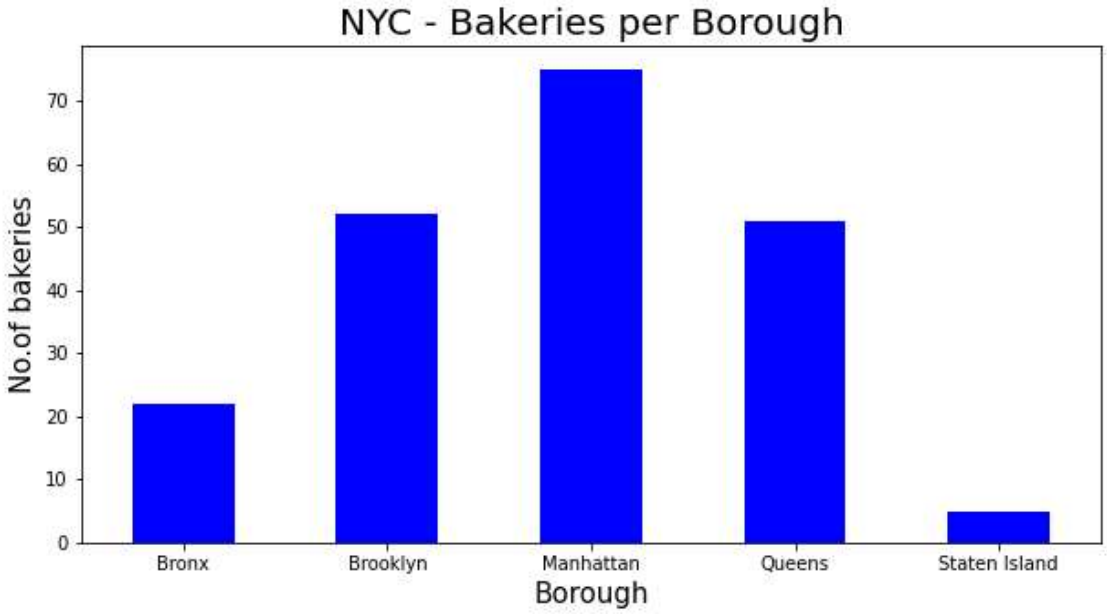
1. Download NY Data Set, clean and process the downloaded data and import it to Panda DataFrame
2. Visualize all the neighborhoods using Folium
3. Using FourSquare API localize all the bakeries in all the neighbourhoods.
4. Visualize the 'bakeries in the neighbourhoods' results (HeatMap)
5. Using FourSquare API data localize all the neighbourhoods without bakeries
6. Visualize the 'neighbourhoods without bakeries' results (HeatMap)
7. Using FourSquare API localize all the neighbourhoods with French Restaurants
8. Use k-means algorithm to group the data (bakeries and French Restaurants)
9. Analyze and choose the best neighbourhoods among clusters

3. Exploratory Data Analysis

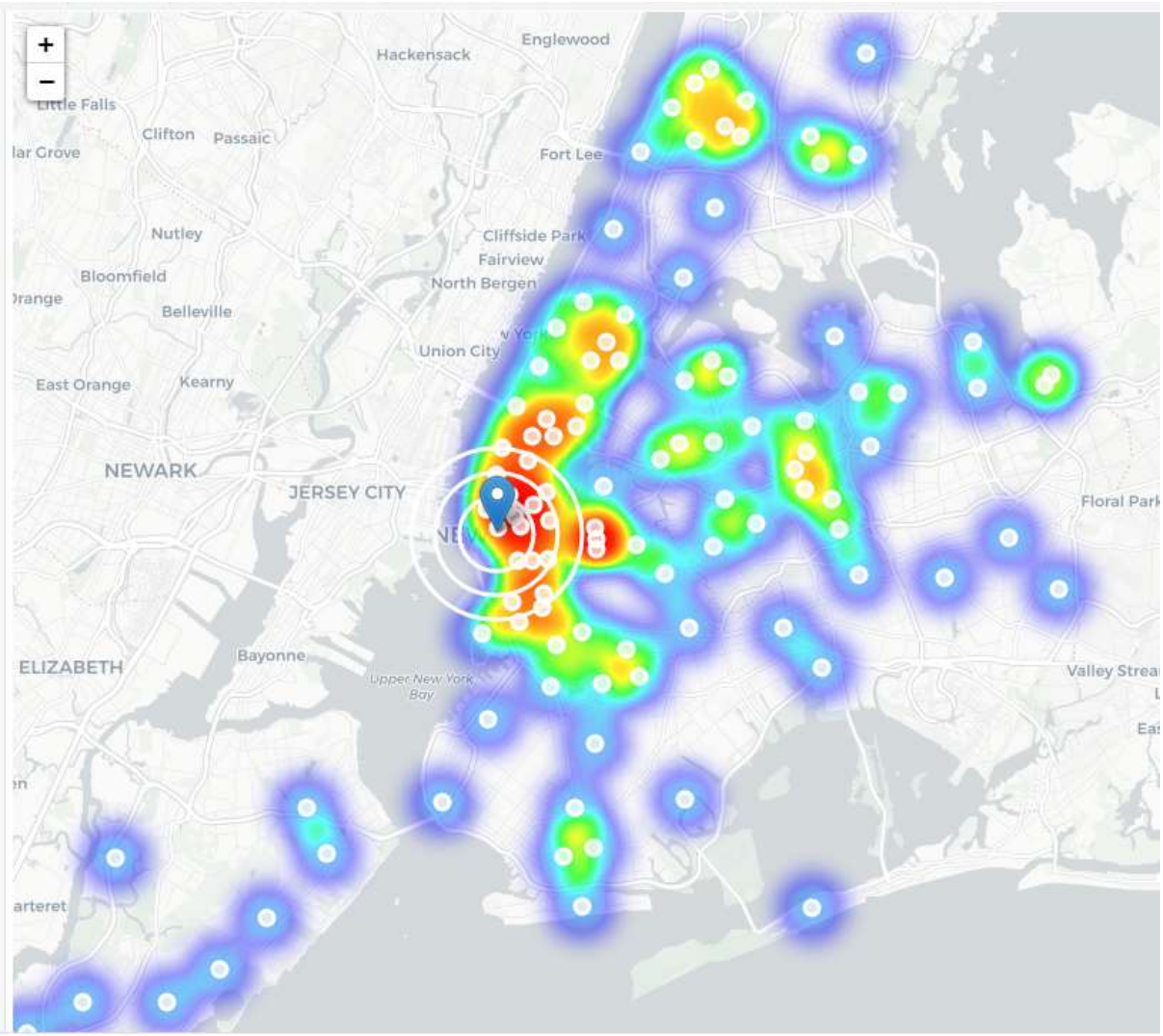
1. Bakeries per Neighbourhood (TOP10)



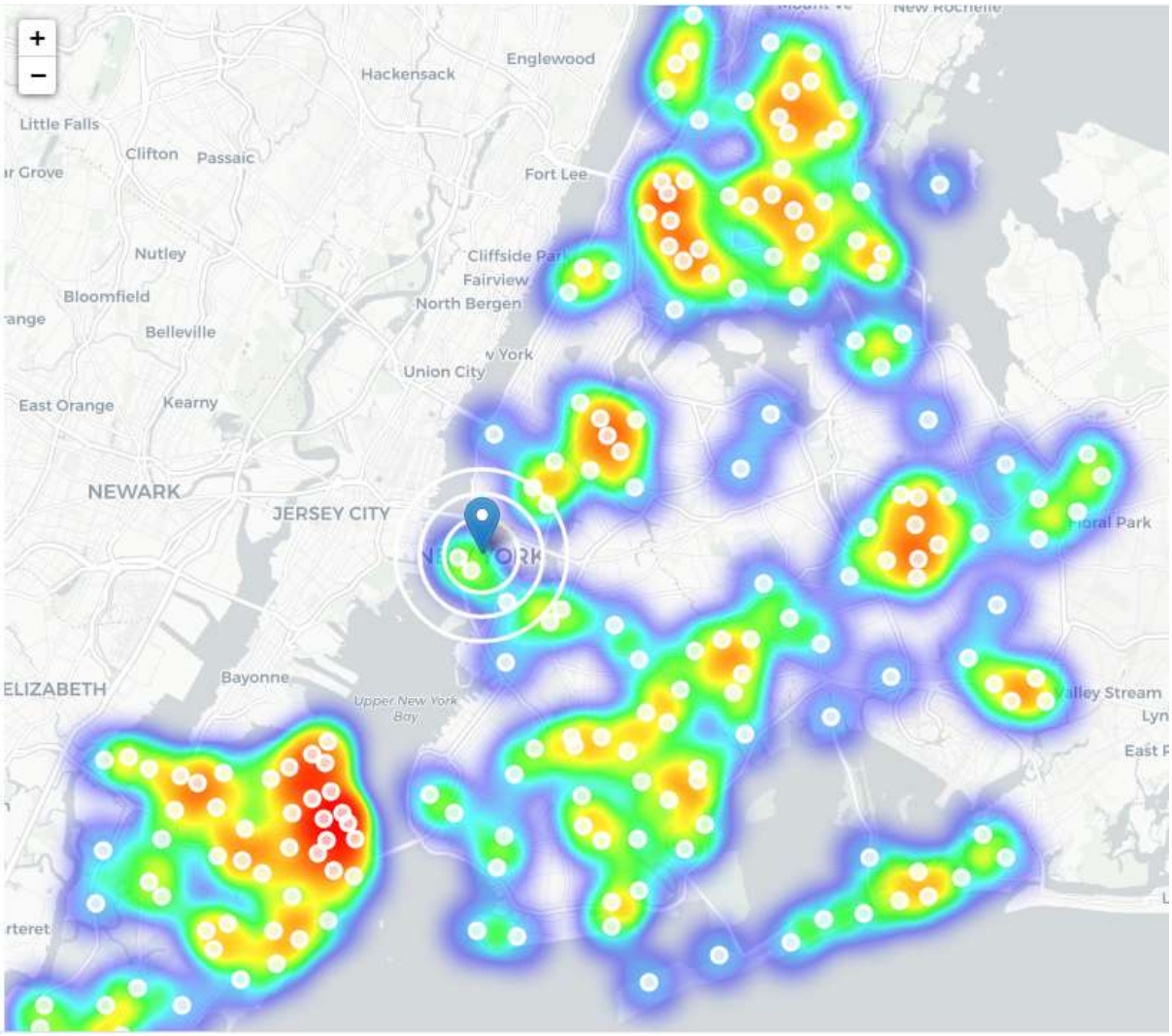
2. Bakeries per Borough



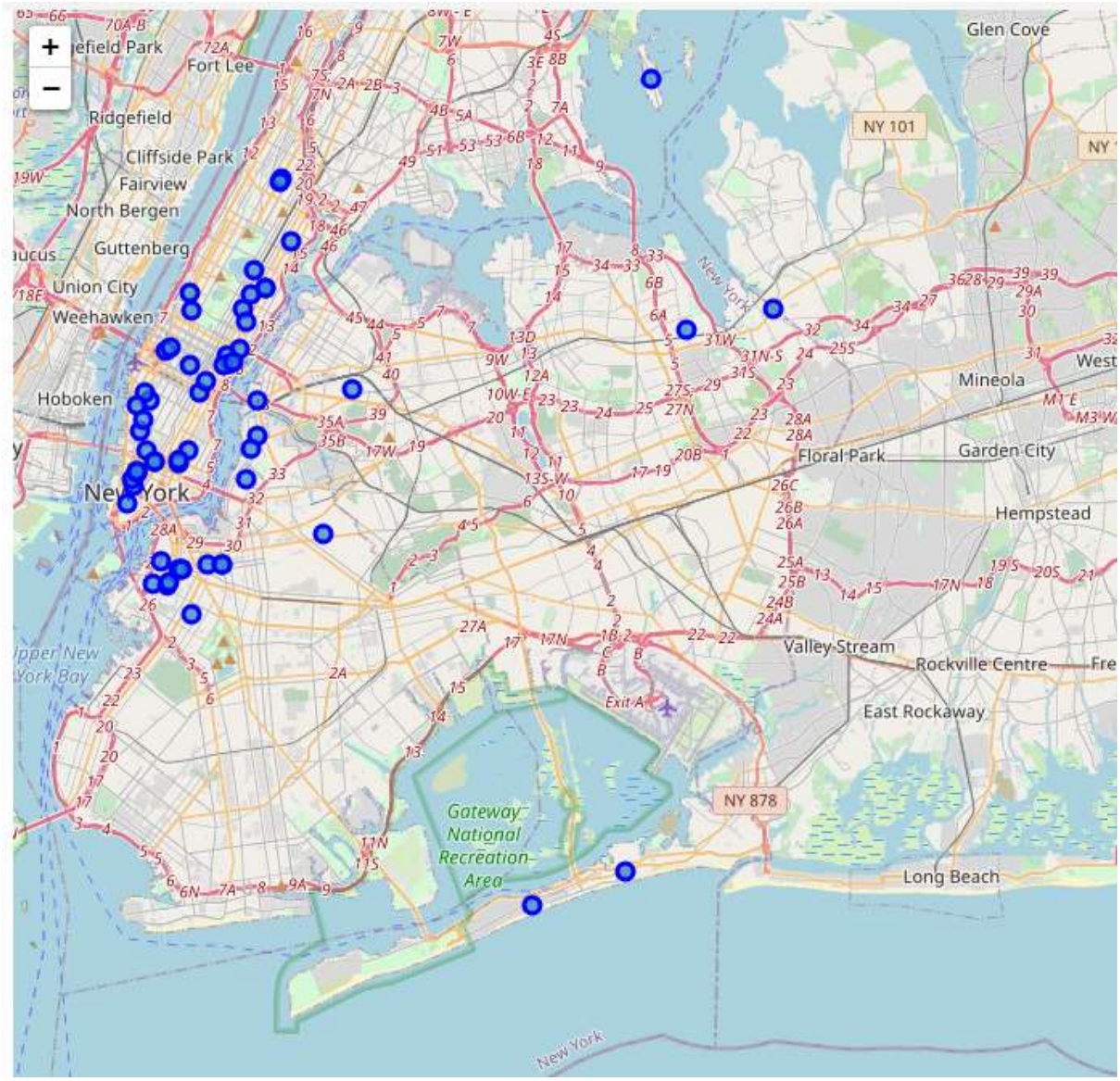
3. Neighbourhood with Bakeries (HeatMap)



4. Neighbourhood without Bakeries (HeatMap)



5. French restaurants in New York (Folium)



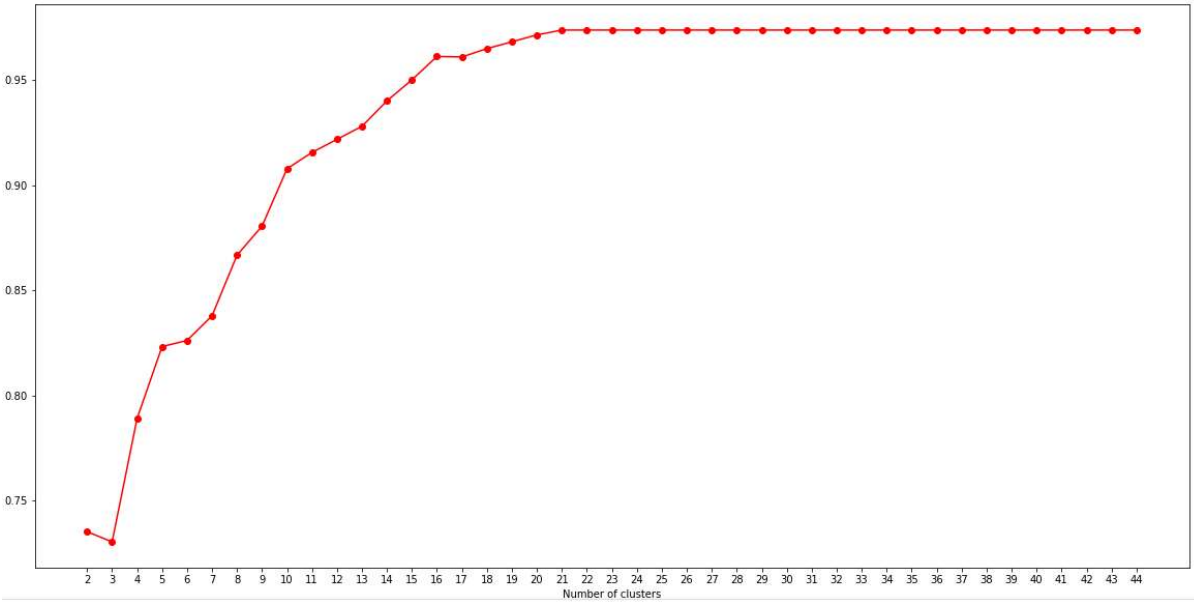
4. Data clustering

For the analysis of the data the k-means clustering has been selected as the most suitable to cluster the neighbourhoods in NYC.

“k-means clustering” is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid)

Data preparation for the clustering:

Vector with each neighbourhoods and its bakeries count and French restaurants count. Number of clusters has been selected based on silhouette_score:



Clusters numbers parameter has been set to 16 which gave us below results:

```
[306 rows x 2 columns]
[ 7 13  8  8 13  2 12 12 12 12 13 12 12 12 13  8  2  1  9  3  9  1 15 15
 15  3  3  3  9  9  9  9 14  1  1  9  9 14  1 14  3  3  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  6  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  4  4  4  4  4  4  4  4  0  4  4  4  4  4  4  4  4  4  4  4  4
  4  4  4  4  4  4  4  4  4  4  4  4  4  5  5 10 10 10 10 10 10 10 11 11
 11 11 11  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  4  4  4
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  4  0  4  4  4  4  4  4  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
```

Mean value of the cluster groups has been analyzed:

Labels	BakeriesCount	FrRestaurCount
0	0.000000	0.000000
1	4.200000	1.200000
2	0.000000	2.000000
3	2.000000	1.000000
4	1.000000	0.000000
5	2.000000	0.000000
6	8.000000	0.000000
7	2.000000	4.000000
8	5.000000	2.666667
9	0.000000	1.000000
10	4.142857	0.000000
11	3.000000	0.000000
12	2.142857	2.000000
13	1.000000	2.250000
14	3.000000	1.000000
15	1.000000	1.000000

As we assumed earlier that clusters with high number of French restaurants and small number of Bakeries can be potentially a good match we have analyzed two clusters: Cluster 2 and Cluster 13.

```
1  ny_neighborhood_with_bakeryCount_and_FrenchRest.loc[ny_neighborhood_with_bakeryCount_and_FrenchRest['Labels'] == 2]
```

1]:

	Borough	Neighborhood	Latitude	Longitude	BakeriesCount	FrRestaurCount	Labels
288	Manhattan	Central Harlem	40.815976	-73.943211	0.0	2.0	2
211	Brooklyn	Fort Greene	40.688527	-73.972906	0.0	2.0	2

```
1  ny_neighborhood_with_bakeryCount_and_FrenchRest.loc[ny_neighborhood_with_bakeryCount_and_FrenchRest['Labels'] == 13]
```

3]:

	Borough	Neighborhood	Latitude	Longitude	BakeriesCount	FrRestaurCount	Labels
103	Manhattan	Turtle Bay	40.752042	-73.967708	1.0	3.0	13
78	Manhattan	Clinton	40.759101	-73.996119	1.0	2.0	13
90	Brooklyn	Greenpoint	40.730201	-73.954241	1.0	2.0	13
72	Manhattan	Tribeca	40.721522	-74.010683	1.0	2.0	13

We have visualized both clusters on the map (blue – cluster 2, orange – cluster 13):

We have compare the clusters with the “no-bakery” HeatMap from 3.4 and selected the best location as Brooklyn, Fort Green as this area do not have any bakeries but have French Restaurants. Downtown and Dumbo which are quite far are the nearest locations with Bakeries (2 per neighbourhood).

5. Results/Conclusion

Manhattan and Brooklyn are the boroughs were French Restaurants are operating which give us a hint that there is a chance for French Bakery in the area. From the clustered data we have selected the clusters where number of bakeries is low (Cluster No 2 - No bakeries at all, Cluster No 13 - one bakery in the Neighbourhood).

We have also identified neighbourhoods without any bakeries where is a high change of success to open one. Largest areas according to our HeatMap is around Queens, Hilcrest and Staten Island, Fox Hills.

I would open French Bakery in Brooklyn, Fort Greene as this area do not have any bakeries but have French Restaurants. DownTown and Dumbo which are quite far are the nearest locations with Bakeries (2 per neighbourhood).

As a final note, all of the above analysis is depended on the adequacy and accuracy of Four Square data. A more comprehensive analysis and future work would need to incorporate data from other external databases.

6. Future directions

Data analysis was mostly done based on neighbourhoods analysis for bakeries and French restaurants. We could improve the clustering including the data from each Venue: ratings, likes and tips parameters in Data Preparation Process. Foursquare is not the most popular localization service in NY so in the future some other providers can be used (for example: google.com) to gather more accurate data.