IBM Data Science Capstone Project - The Battle of Neighbourhoods

1. Introduction: Business Problem

Suitable New Store Location in London for an Indian Fashion Retailer

The Task at Hand:
A company is looking to open a clothing store based on Indian culture in London.
I've have been given a task of identifying the best location the open the store where the ***Asian*** population is maximum since most of the customers will be ***Indians, Pakistanis and Bangladeshis***.

Criteria:
Qualitative data from another retailer that they know, suggests that the best locations to open new fashion retail stores may not only be where other clothing is located but that the best places are in fact areas that are near ***Indian Restaurants, Cafés and Parks***. As **Asians** are very social people that frequent these places often, so opening new stores in these locations is becoming popular.

The analysis and recommendations for new store locations will focus on general neighbourhoods with these establishments, not on specific store addresses. Narrowing down the best district options derived from analysis allows for either further research to be conducted, advising agents of the chosen district, or on the ground searching for specific sites by the company's personnel.

Why Data?:
Without leveraging data to make decisions about new store locations, the company could spend countless hours walking around neighbourhoods, consulting many real estate agents with their own district biases, and end up opening in yet another location that is not ideal.

Data will provide better answers and better solutions to their task at hand.

Outcomes:
The goal is to identify the best neighbourhood to open new stores as part of the company's plan. The results will be translated to management in a simple form that will convey the data-driven analysis for the best locations to open stores.

2. Data Section

London is one of the most ethnically diverse cities in the world. At the 2011 census, London had a population of 8,173,941. Of this number, 44.9% were White British. 37% of the population were born outside the UK, including 24.5% born outside of Europe.

The demography of London is analysed by the Office for National Statistic and data is produced for each of the Greater London wards, the City of London and the 32 London boroughs, the Inner London and Outer London statistical sub-regions, each of the Parliamentary constituencies in London, and for all of Greater London as a whole.

For our fashion store problem, we will focus on the Boroughs of London and work on getting the data from all the Boroughs. There are 32 London Boroughs with a population of around 150,000 to 300,000.

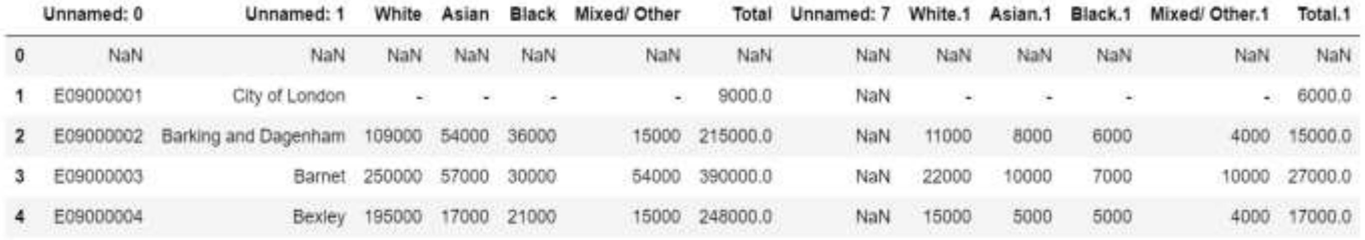
To solve our problem of finding a best location to open an Indian fashion store in London, we need to datasets based on various parameters such as :

1. List of **areas of London** available at: https://en.wikipedia.org/wiki/List_of_areas_of_London

2. The latitudes and longitudes of those areas which the done with the help of **geopy.geocoders** library in python
3. Population of target audience in all the **boroughs** of London based on their **ethnicity** at **London Datastore**, which is a free and open data-sharing portal where anyone can access data relating to the city. The data is available in XLS and CSV format, which we can download and can use as-is for solving our problem. https://data.london.gov.uk/dataset/ethnic-groups-borough

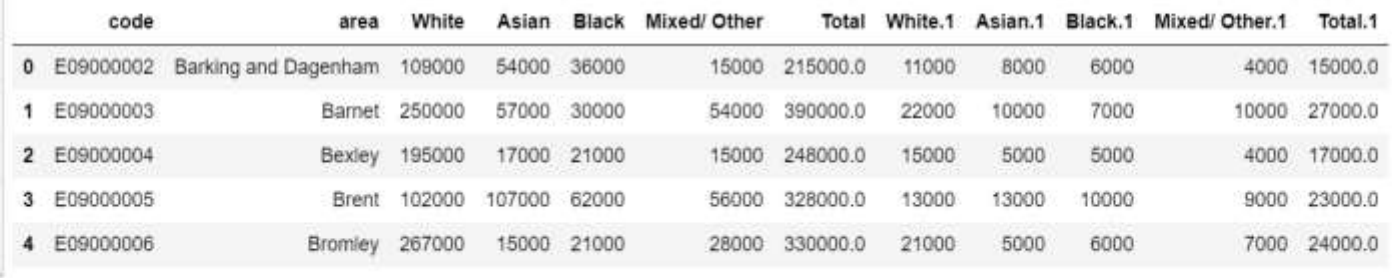
The cleansed data will then be used alongside **Foursquare data**, which is readily available. Foursquare location data will be leveraged to explore or compare neighborhoods around London.

The ethnicity data was imported from the source, but as can be seen, was not in the right format.



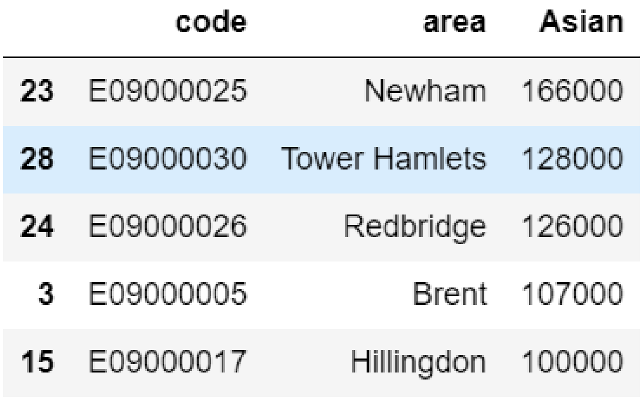
	Unnamed: 0	Unnamed: 1	White	Asian	Black	Mixed/ Other	Total	Unnamed: 7	White.1	Asian.1	Black.1	Mixed/ Other.1	Total.1
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	E09000001	City of London	-	-	-	-	9000.0	NaN	-	-	-	-	6000.0
2	E09000002	Barking and Dagenham	109000	54000	36000	15000	215000.0	NaN	11000	8000	6000	4000	15000.0
3	E09000003	Barnet	250000	57000	30000	54000	390000.0	NaN	22000	10000	7000	10000	27000.0
4	E09000004	Bexley	195000	17000	21000	15000	248000.0	NaN	15000	5000	5000	4000	17000.0

After some data wrangling and cleaning – renaming and dropping unnecessary columns - the dataframe was in a structure that could be used.



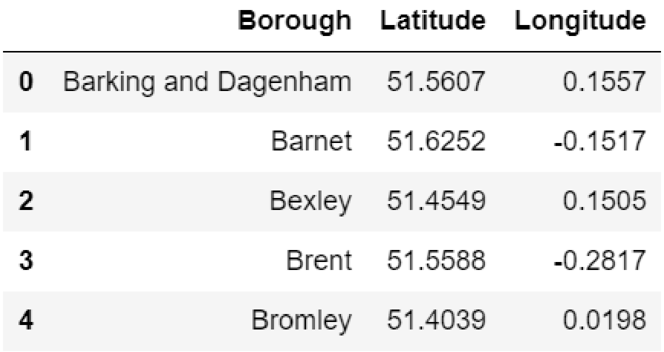
	code	area	White	Asian	Black	Mixed/ Other	Total	White.1	Asian.1	Black.1	Mixed/ Other.1	Total.1
0	E09000002	Barking and Dagenham	109000	54000	36000	15000	215000.0	11000	8000	6000	4000	15000.0
1	E09000003	Barnet	250000	57000	30000	54000	390000.0	22000	10000	7000	10000	27000.0
2	E09000004	Bexley	195000	17000	21000	15000	248000.0	15000	5000	5000	4000	17000.0
3	E09000005	Brent	102000	107000	62000	56000	328000.0	13000	13000	10000	9000	23000.0
4	E09000006	Bromley	267000	15000	21000	28000	330000.0	21000	5000	6000	7000	24000.0

Data sorted according to Asian population



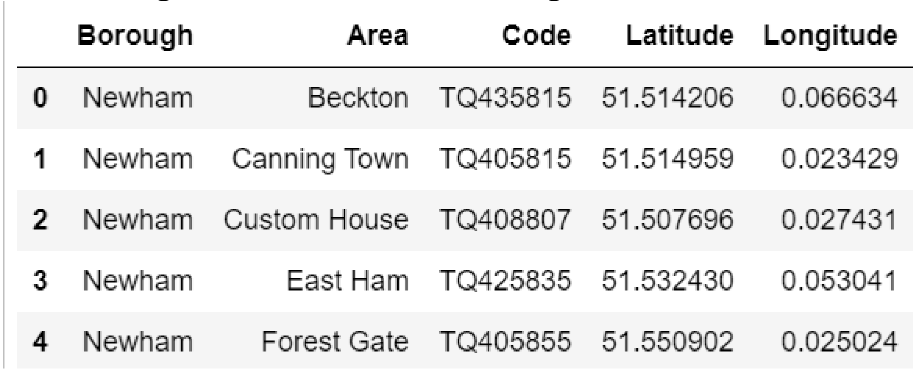
	code	area	Asian
23	E09000025	Newham	166000
28	E09000030	Tower Hamlets	128000
24	E09000026	Redbridge	126000
3	E09000005	Brent	107000
15	E09000017	Hillingdon	100000

The data for list boroughs taken from Wikipedia



	Borough	Latitude	Longitude
0	Barking and Dagenham	51.5607	0.1557
1	Barnet	51.6252	-0.1517
2	Bexley	51.4549	0.1505
3	Brent	51.5588	-0.2817
4	Bromley	51.4039	0.0198

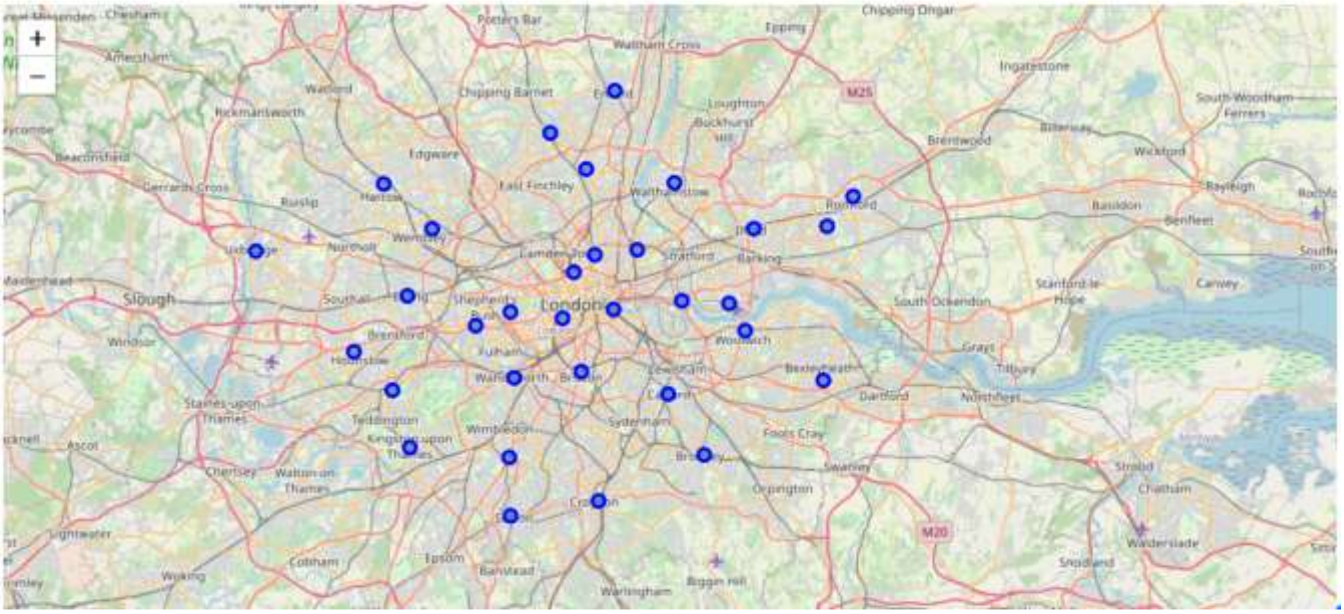
Data for neighborhoods in Newham borough



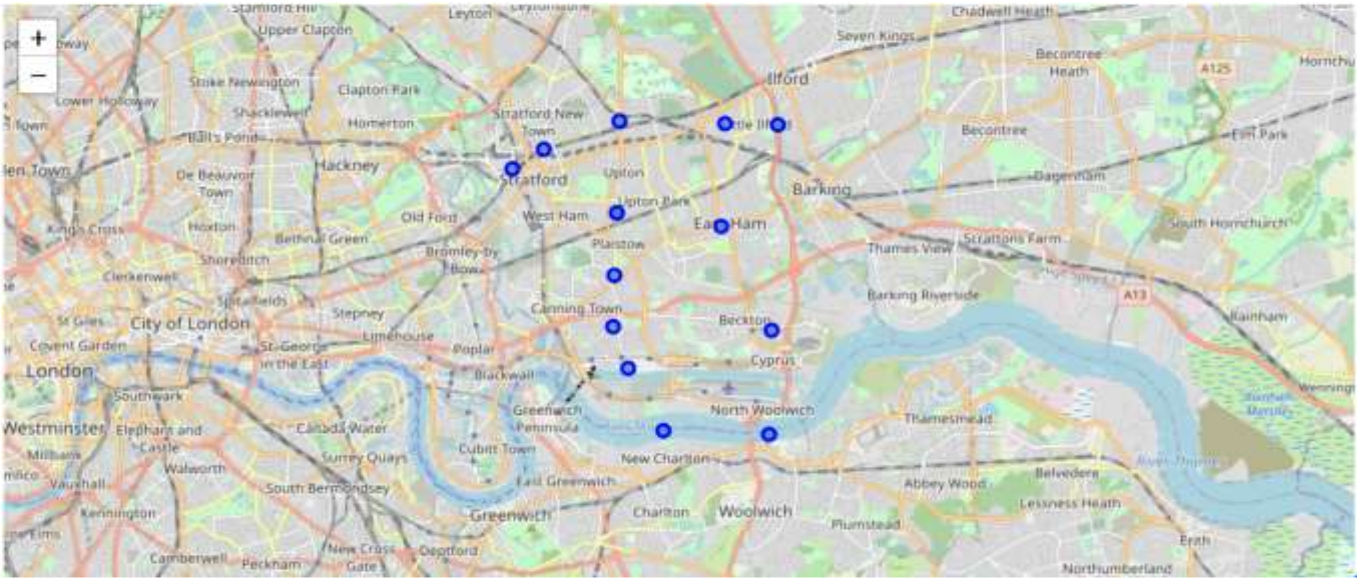
	Borough	Area	Code	Latitude	Longitude
0	Newham	Beckton	TQ435815	51.514206	0.066634
1	Newham	Canning Town	TQ405815	51.514959	0.023429
2	Newham	Custom House	TQ408807	51.507696	0.027431
3	Newham	East Ham	TQ425835	51.532430	0.053041
4	Newham	Forest Gate	TQ405855	51.550902	0.025024

3. Methodology and Exploratory Analysis

The GeoPy library is used to geocode location data and get location coordinates. It was imported here to get the latitude and longitude values of London. From this using folium I created a map of London with the location of all boroughs superimposed.



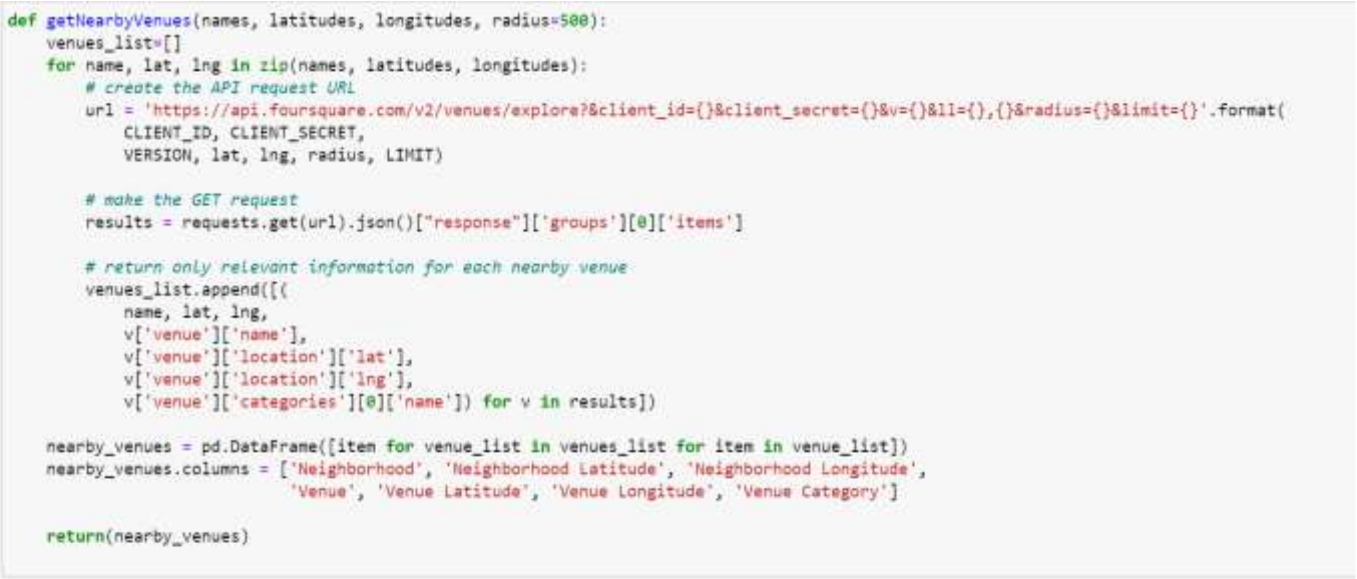
Visualizing the selected neighborhood of Newham borough as it has maximum Asian population



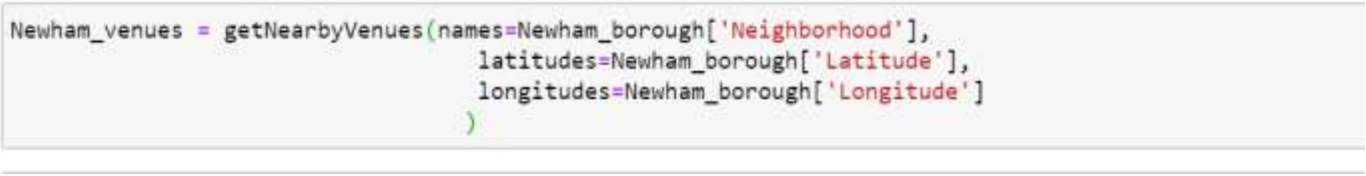
Foursquare

Use the Foursquare API to explore neighborhoods of borough Newham of London

After setting up the Foursquare API, we can explore the geolocation data. Exploratory data analysis allows us to look at what we are dealing with, and is this case, all the neighborhoods of Newham were explored.

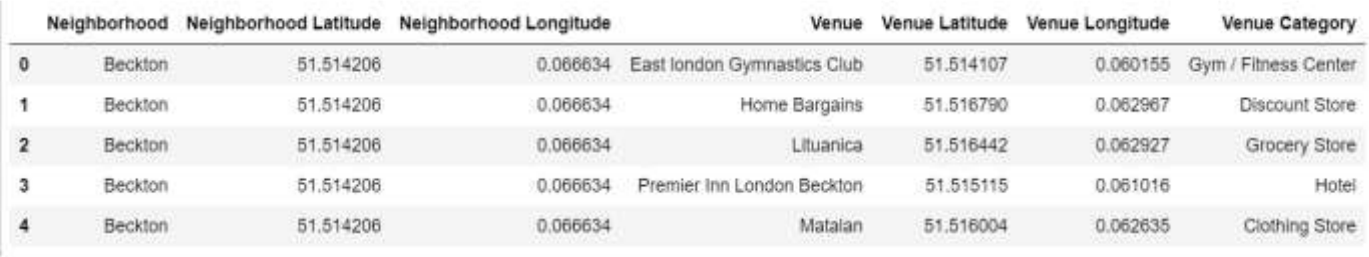


```
def getNearbyVenues(names, latitudes, longitudes, radius=500):
    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID, CLIENT_SECRET,
            VERSION, lat, lng, radius, LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name, lat, lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood', 'Neighborhood Latitude', 'Neighborhood Longitude',
                             'Venue', 'Venue Latitude', 'Venue Longitude', 'Venue Category']

    return(nearby_venues)
```



```
Newham_venues = getNearbyVenues(names=Newham_borough['Neighborhood'],
                                 latitudes=Newham_borough['Latitude'],
                                 longitudes=Newham_borough['Longitude']
                                )
```

And we’re able to generate a new dataframe with all of the nearby venues for all of the neighborhoods



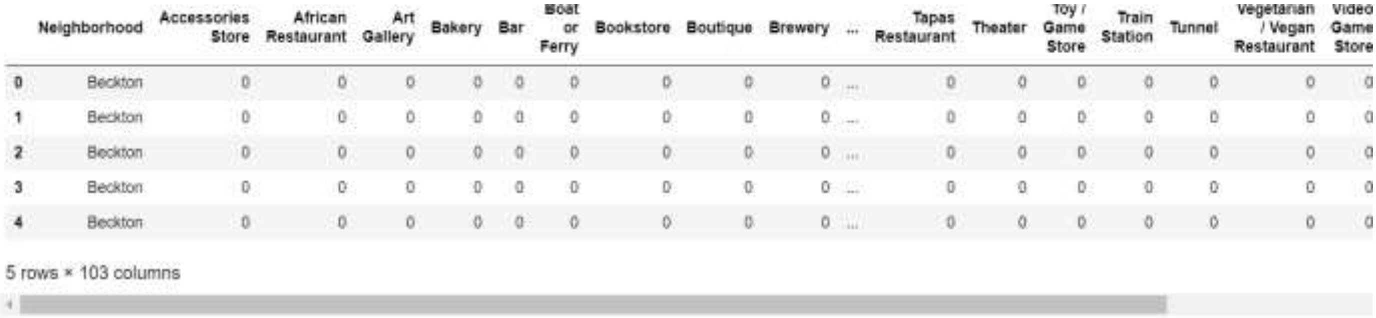
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Beckton	51.514206	0.066634	East london Gymnastics Club	51.514107	0.060155	Gym / Fitness Center
1	Beckton	51.514206	0.066634	Home Bargains	51.516790	0.062967	Discount Store
2	Beckton	51.514206	0.066634	Lituanica	51.516442	0.062927	Grocery Store
3	Beckton	51.514206	0.066634	Premier Inn London Beckton	51.515115	0.061016	Hotel
4	Beckton	51.514206	0.066634	Matalan	51.516004	0.062635	Clothing Store

With new dataframe from the data, it was then possible to check how many venues were returned for each neighbourhood, and it was possible to calculate how many unique venue categories there are.

This is a useful statistic in itself.

This all very useful to accomplish our task of finding the best location areas for new stores, but also is great data to have access to as a resource in future planning of new stores down the track.

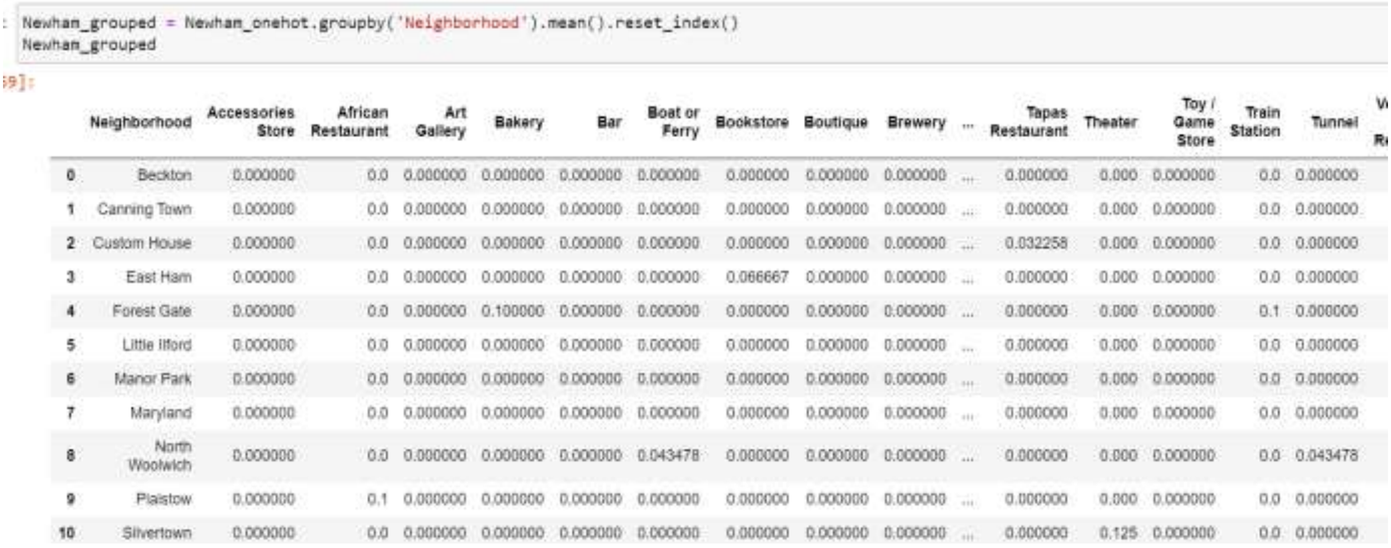
It was then possible to analyse each of the neighbourhoods from the results, displaying how many venues of each category were in each neighbourhood.



	Neighborhood	Accessories Store	African Restaurant	Art Gallery	Bakery	Bar	Boat or Ferry	Bookstore	Boutique	Brewery	...	Tapas Restaurant	Theater	Toy / Game Store	Train Station	Tunnel	Vegetarian / Vegan Restaurant	Video Game Store
0	Beckton	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
1	Beckton	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
2	Beckton	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
3	Beckton	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
4	Beckton	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0

5 rows × 103 columns

An important part of the process was to group rows by neighbourhood and take the mean of the frequency of occurrence of each category. This will be used in narrowing down the suitable neighbourhoods for new stores



```
Newham_grouped = Newham_onehot.groupby('Neighborhood').mean().reset_index()
Newham_grouped
```

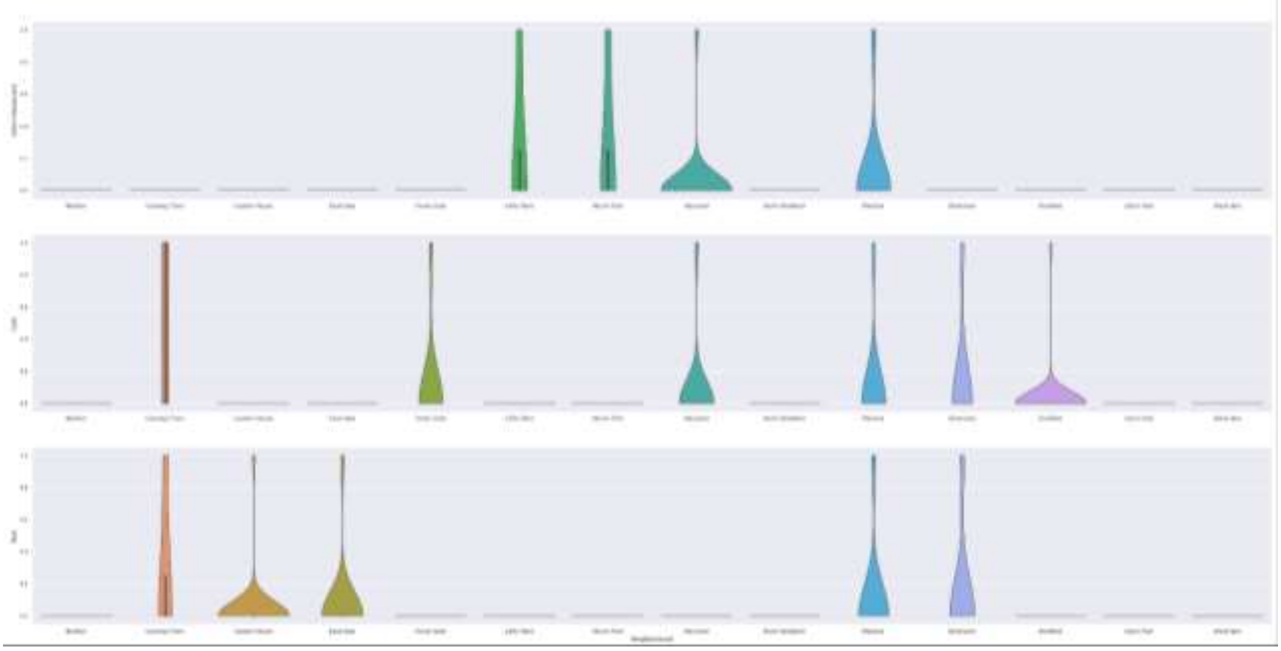
59]:

	Neighborhood	Accessories Store	African Restaurant	Art Gallery	Bakery	Bar	Boat or Ferry	Bookstore	Boutique	Brewery	...	Tapas Restaurant	Theater	Toy / Game Store	Train Station	Tunnel
0	Beckton	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000	0.000000	0.0	0.000000
1	Canning Town	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000	0.000000	0.0	0.000000
2	Custom House	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.032258	0.000	0.000000	0.0	0.000000
3	East Ham	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.066667	0.000000	0.000000	...	0.000000	0.000	0.000000	0.0	0.000000
4	Forest Gate	0.000000	0.0	0.000000	0.100000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000	0.000000	0.1	0.000000
5	Little Ilford	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000	0.000000	0.0	0.000000
6	Manor Park	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000	0.000000	0.0	0.000000
7	Maryland	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000	0.000000	0.0	0.000000
8	North Woolwich	0.000000	0.0	0.000000	0.000000	0.000000	0.043478	0.000000	0.000000	0.000000	...	0.000000	0.000	0.000000	0.0	0.043478
9	Plaistow	0.000000	0.1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000	0.000000	0.0	0.000000
10	Silvertown	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.125	0.000000	0.0	0.000000

Referring back to the original task - the business types criteria specified by the client! ***'Indian Restaurants', 'Cafés'*** *and* ***'Park'.***

Let's look at their frequency of occurrence for all the London neighbourhoods, isolating the categorical venues. These are the venue types that the client wants to have an abundant density of in the ideal store locations. I've used a violin plot from the seaborn library - it is a great way to visualise frequency distribution datasets, they display a density estimation of the underlying distribution.

Let’s see the results.



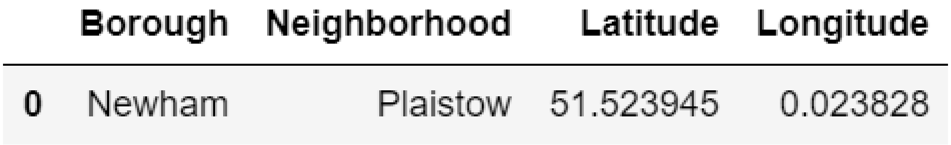
The Neighbourhoods

So as we can see from the analysis that the best neighbourhood to open new stores is **Plaistow** as according to the criteria it has the 3 specified venues in a great frequency **(Café**, **Indian restaurant** and **Parks**).

Other Neighbourhoods which satisfy atleast 2 criteria are:

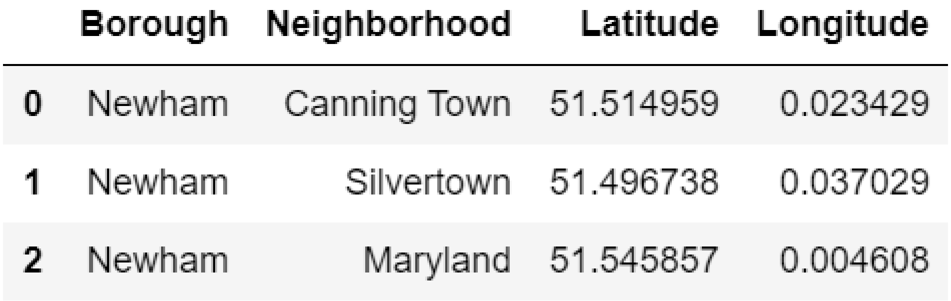
- **Canning Town**
- **Silvertown**
- **Maryland**

Moving the best prospective neighborhood in new dataframe



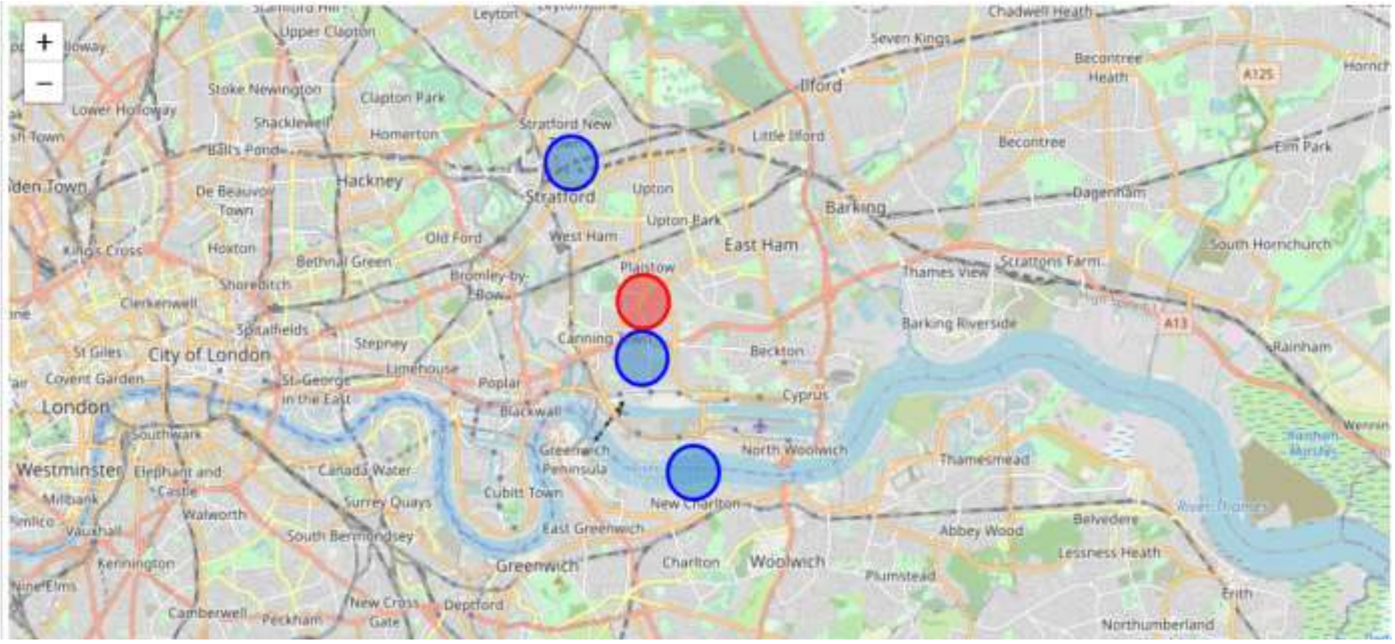
	Borough	Neighborhood	Latitude	Longitude
0	Newham	Plaistow	51.523945	0.023828

and the other alternatives neighborhoods in another data frame



	Borough	Neighborhood	Latitude	Longitude
0	Newham	Canning Town	51.514959	0.023429
1	Newham	Silvertown	51.496738	0.037029
2	Newham	Maryland	51.545857	0.004608

Visualizing the main neighborhood **Plaistow** in **Red**
The alternative neighbourhoods ***Canning Town, Silvertown*** *and* ***Maryland*** *in* ***Blue***



4. Result and Discussion

We have made inferences from the data in making the location recommendations, but that is exactly the point. There is no right or wrong answer or conclusion for the task at hand. The job of data analysis here is to steer a course for the location selection of new stores (i) to meet the criteria of being in neighbourhoods that are lively with abundant leisure venues, and (ii) to narrow the search down to just a few of the main areas that are best suited to match the criteria.

Results
So as we can see from the analysis the best neighborhood to open new store is **Plaistow** as according to the criteria it has the 3 specified venues in a great frequency (**Café, Indian restaurant and Parks**).

Other Neighobourhoods which satisfy atleast 2 criteria are:

- **Canning Town**
- **Silvertown**
- **Maryland**

5. Conclusion

There are many ways this analysis could have been performed based on different methodolgy and perhaps different data sources. I chose the method I selected as it was a straight forward way to narrow down the options, not complicating what is actually simple in many ways – meeting the the critera for the surrounding venues, and in my case, domain knowledge I have on the subject.

I originally intended to use the clustering algorithms to cluster the data, but as it progressed it became obvious that this only complicated the task at hand. The analysis and results are not an end point, but rather a starting point that will guide the next part of the process to find specific store locations. The next part will involve domain knowledge of the industry, and perhaps, of the city itself. But the data analysis and resulting recommendations have greatly narrowed down the best district options based on data and what we can infer from it.

Without leveraging data to make focussed decisions, the process could have been drawn out and resulted in new stores opening in sub-standard areas for this retailer. Data has helped to provide a better strategy and way forward, these data-driven decisions will lead to a better solution in the end.

Thanks for taking part in my Data Science journey!