Coursera Capstone Project

Project Name

- The Battle of Neighborhoods

London House Price Index: Price Paid Data

Submitted by: Vedant Dave

Electrical and Computer Engineering
University of Windsor

Problem Statement:

With the extreme trend of the migration people of all kind always want to migrate to some better place for the living. UK is one of the best countries in the world to live with all kind of facility. But UK is also expensive in term of cost and when the matter is about buying the property in UK. Then consideration of the Price trend factor is much important for any individual.

With varying budget needs, People find it very hard to make the remarkable place in the big UK city like London. The exact place near the work area and other facility is more important than other consideration. The Data-Science method of K-Mean Clustering can give us the better idea to choose the ideal place nearer to our own priorities and their importance. As per the data and some recent half decade economic factors like Brexit, current housing market struggles to get exact idea about predicting the future of Housing price and profit margin in Property Sectors. Current Real-Estates and invertors are unpredictable about the profitable return. The Prices are vary according to city areas and the facility improvement. The trend of past and the rate of development also play a crucial role in this situation.

A Potential Client must want the perfect Knowledge about the current growing trend in market as well as ongoing price for conscious Decision. Furthermore He / She must like to consider the several factors of facilities, accommodations, Hotels, Local Businesses growth, other Property Prices etc.

Solution for this have been considered form Data-Science Technique called KMeans Clustering of Neighborhood. It may give some trustful weightage to unfailing factors and open door for reliable decision.

Target Audience:

Potential Clients are looking for the suitable Property as per their need and investors, who are searching for the good returns in current market condition.

Stakeholders:

1. Government of UK
2. Sellers
3. Buyers
4. Real estate agents

Data Section:

Following sources of data are used while executing the Capstone Project for the solution of the Current Business Problem.

[1] HM Land Registry: Price Paid Data

- **Data title: -**
  Open Data published by Government of UK under the section **HM Land Registry: Price Paid Data**

- **Type of data:**
  CSV file (Comma separated Value)

- **Duration:**
  August 2018 data

- **Description of the dataset: -**
  Price Paid Data includes information on all property sales in England and Wales that are sold for full market value and are lodged with us for registration.

  The dataset includes the transactions received at HM Land Registry in the period from the first to the last day of August 2018.

- **LINK:**
  This dataset was downloaded and later hosted on https://labs.cognitiveclass.ai/ for ease of use.

- **Source: -**
  http://landregistry.data.gov.uk/

[2] Google Maps Geocoding API

- **Data title: -**
  Google Maps Geocoding API

- **Type of data: -**
  JSON

- **Duration: -**
  N/A

- **Description of the data:**
  Location coordinates obtained by G Maps API calls.

Location Information obtained from Price Paid Dataset is used to obtain the location coordinates from Google Maps.

A separate Python script has been developed to extract the unique street names, district names from the Price Paid Dataset and embed those in the G Maps API calls to obtain the required information.

- **Source: -**
  Google Cloud Platform/ Google Maps

[3] Foursquare location data

- **Data title: -**
  Foursquare location data

- **Type of data: -**
  JSON

- **Duration: -**
  N/A

- **Description of the data: -**
  Location coordinates obtained by Foursquare API calls.

  To determine the proximity of various amenities as per the client's requirement, Foursquare location data is used.

- **Source: -**
  https://foursquare.com/

3. Methodology

The data are devided into the state, national and local region sets so it is easy to analyze the data. The data sets are in the form of csv file with containing some error and the misiing data. So we first need to clean the data set.

Price Paid Dataset contains the sale prices of properties in England and Wales submitted to HM Land Registry for registration. This is an open dataset which is hosted on http://landregistry.data.gov.uk/. This data is updated monthly and is made available from 1995.

This project is focused on investigating the most recent market prices of Property in the city of

London and recommend various locations where the prospective client can buy a property based upon his/ her budget.

Methodology steps:

The automated script developed as a part of this project does the following: -

1. First in the script import all the require libraries such as pandas, numpy, matplotlib and folium. We need to import the JSON file of the particular region and requests statement for query analysis.

2. First, pass the necessary data from the price paid dataset which includes the transactions received at HM Land Registry in the period from the first to the last day of August 2018. Load that data with the local directory saving path method, which save in jupetor notebook local folder.

3. Then we need to give understandable label to the columns and it is gone by the list method of python, in perfect order of the dataset. But we only require two columns for the initial analysis.

4. The data is cleansed and any data of sales agreements which predates 2016 is dropped from the dataset.

5. The data is further condensed by selecting it only for the city of London which is area of choice in this project. And our main focus is for the certain limit of price range.

6. Unique "Street names" in the city of London where recent transactions for sale of property were done are filtered from the dataset.

7. Location coordinates (latitude, longitude) of these street names are fetched by making API calls to Google Maps. A separate one-time Python script was developed to fetch this data and store it in a CSV file. By merging two datasets we can easily filtered out the data for the require street name.

8. The average price of property on each of these streets is determined by taking a mean on recent transactions of sale of property on respective streets.

9. Based upon the budget of the client, the current average prices are compared and all recommendations for the locations are made by plotting them on map of London. The locations popups are labelled with the respective street names and their average property price.

10. The recommended locations are further fed into Foursquare API calls to determine various venues in proximity to them. All reported venues are then tabulated and presented to the user.

11. Important facilities like Hospitals, Grocery stores, Elementary schools, High Schools are searched in vicinity of each location and then reported in a tabular form to the user.

To conduct a similar analysis for any other city in UK or Wales, the automated script has been written to accommodate a change in following two sections :

1. City/ Town
2. Budget of the client

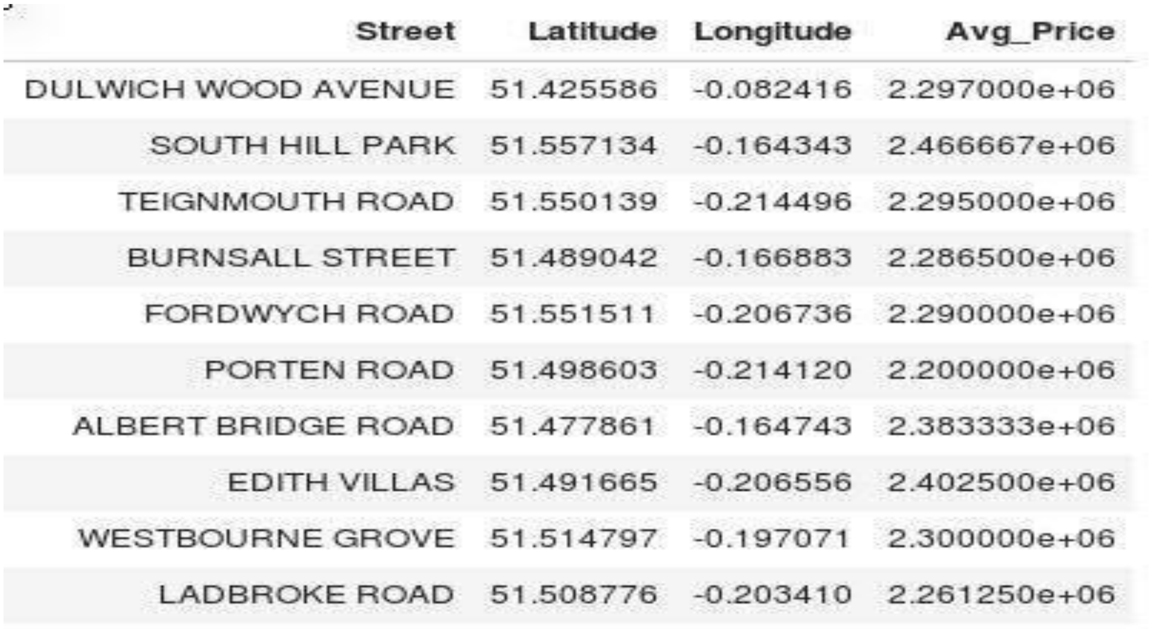
Such changes can be made with minimal effort and would generate the recommended locations to buy a property in the city of choice. The Number of the locations according to the street name give clear idea about the number of the facilities and the average price and distance of accommodation from the facility. Upon running the exploratory data analysis for city of London with a hypothetical budget of GBP 2.2 Million – GBP 2.5 Million, the machine learning algorithm recommends 39 streets in London where the prospective client can choose to buy the property as per the current market prices.

A list of such locations is presented to the user with location coordinates and most recent average prices.

Results:

By analysis of the expletory data analysis on the city of London, we can say the distribution of the 2.2 to 2.5 million GBP data. The machine learning algorithm actually differentiate the machine learning algorithm and get the analysis as following resultant datasets.

A list of such locations is presented to the user with location coordinates and most recent average prices



Street	Latitude	Longitude	Avg_Price
DULWICH WOOD AVENUE	51.425586	-0.082416	2.297000e+06
SOUTH HILL PARK	51.557134	-0.164343	2.466667e+06
TEIGNMOUTH ROAD	51.550139	-0.214496	2.295000e+06
BURNSALL STREET	51.489042	-0.166883	2.286500e+06
FORDWYCH ROAD	51.551511	-0.206736	2.290000e+06
PORTEN ROAD	51.498603	-0.214120	2.200000e+06
ALBERT BRIDGE ROAD	51.477861	-0.164743	2.383333e+06
EDITH VILLAS	51.491665	-0.206556	2.402500e+06
WESTBOURNE GROVE	51.514797	-0.197071	2.300000e+06
LADBROKE ROAD	51.508776	-0.203410	2.261250e+06

Figure: Initial require data separated from big dataset

According to above table analyze that most of the values are near to each other individual average price. When we plot the above table data on the map with folium tool box. Even by use of the tablue we can generate the live map with ticking up the location on map.

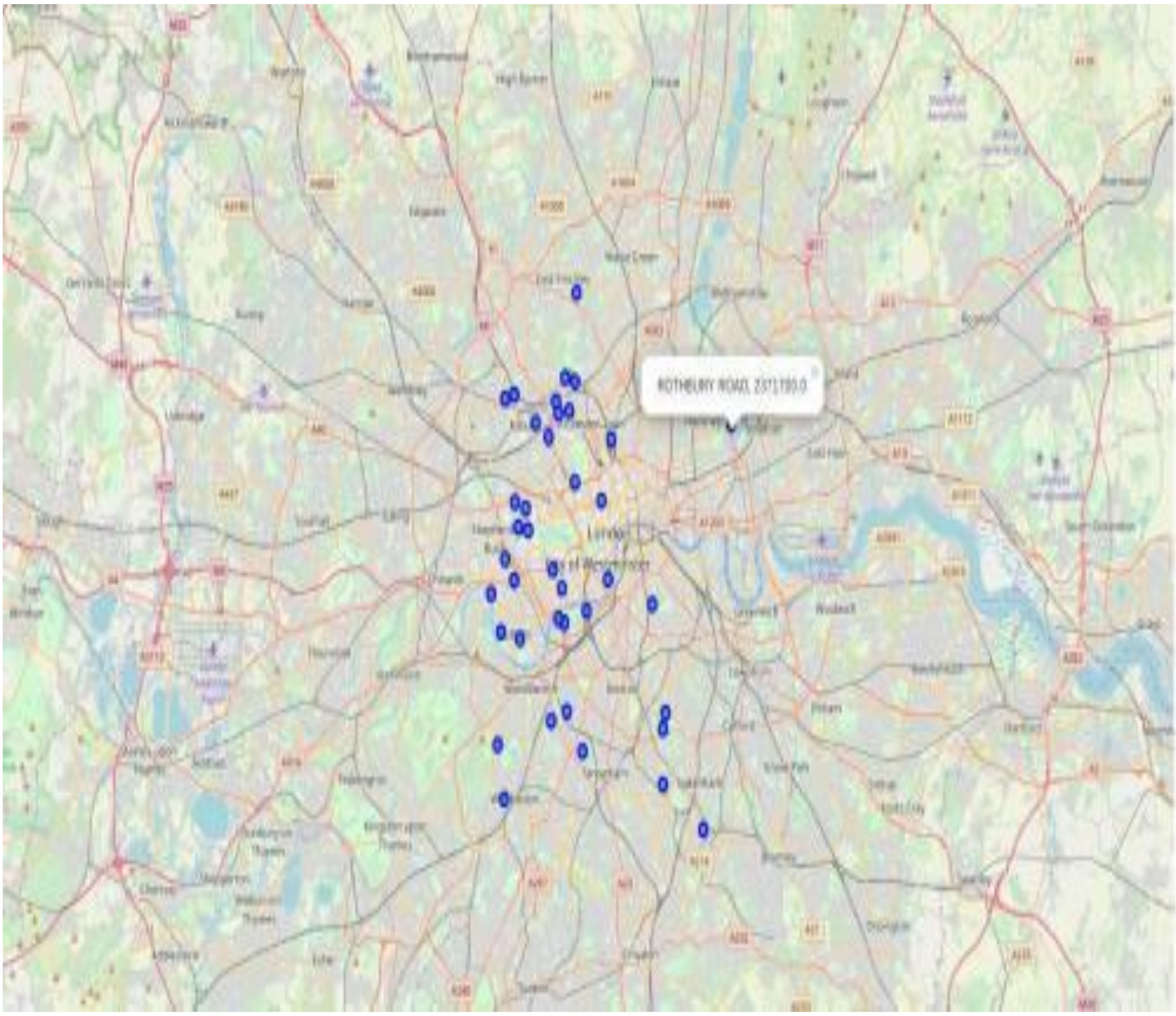
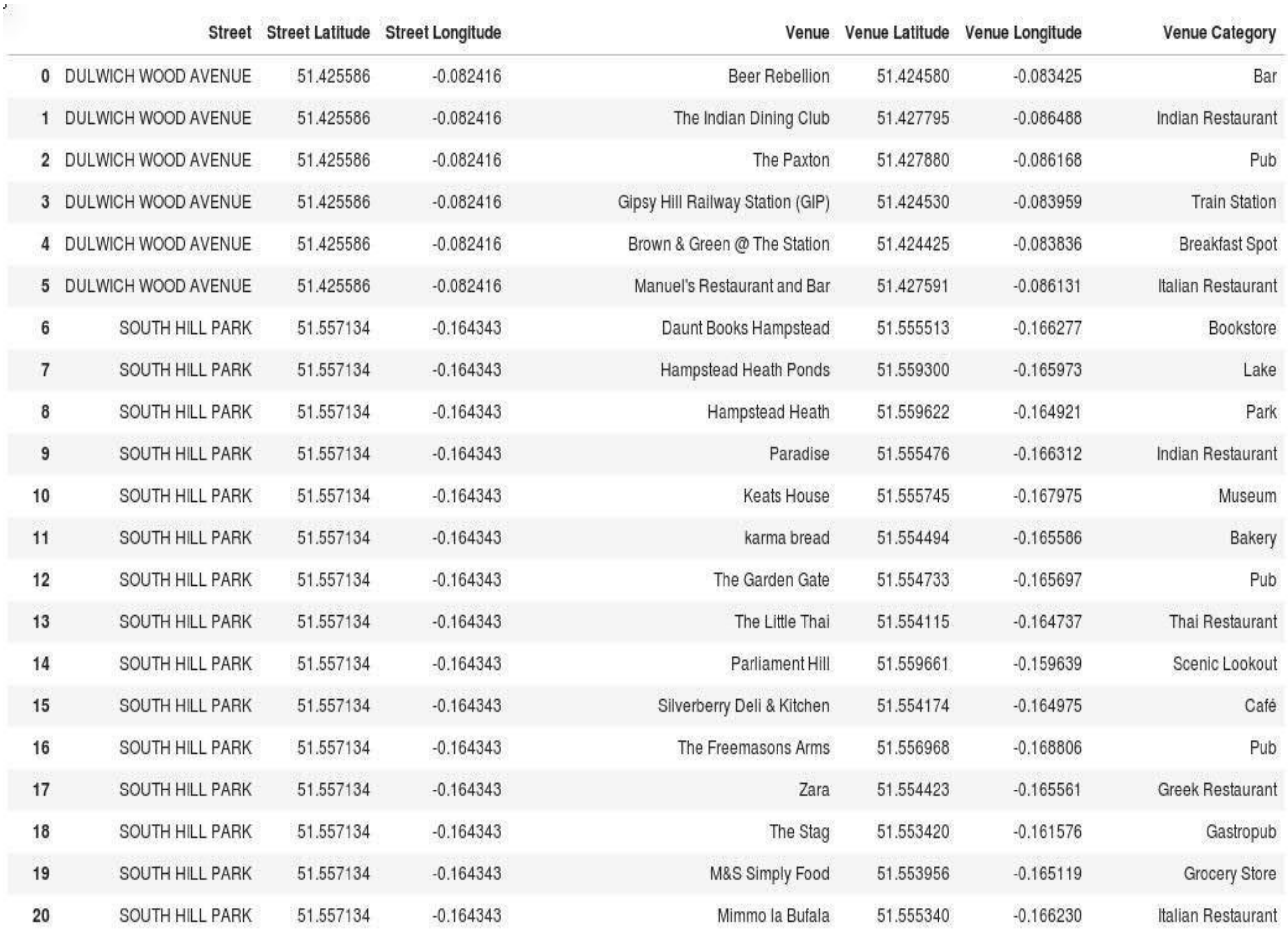


Fig = Plotting Diagram of folium dataset

Further, following venues are enlisted for the user to make an informed decision while choosing a location.

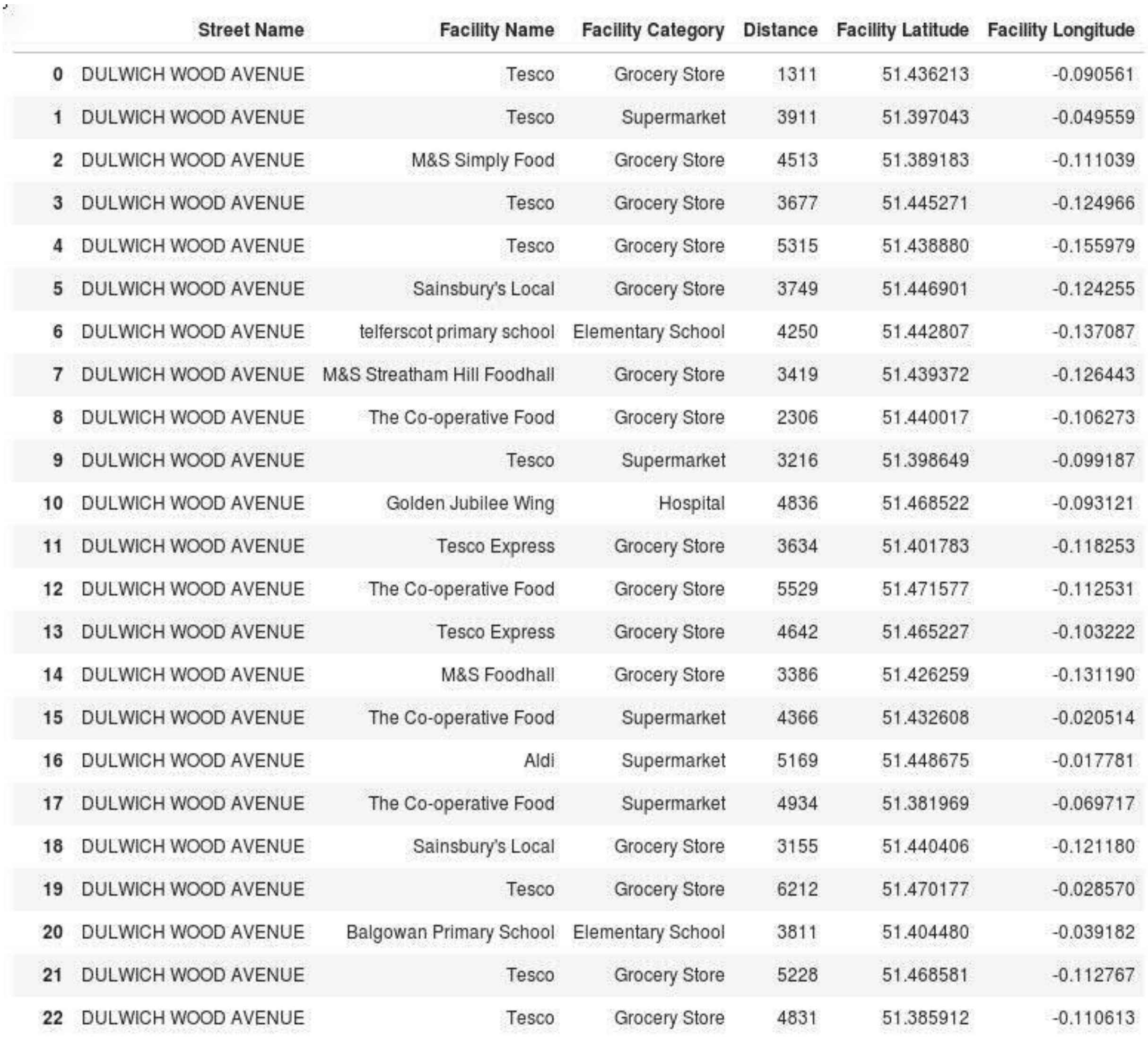
Important facilities are also presented to the user in a tabulated format to take care of his familial needs.



	Street	Street Latitude	Street Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	DULWICH WOOD AVENUE	51.425586	-0.082416	Beer Rebellion	51.424580	-0.083425	Bar
1	DULWICH WOOD AVENUE	51.425586	-0.082416	The Indian Dining Club	51.427795	-0.086488	Indian Restaurant
2	DULWICH WOOD AVENUE	51.425586	-0.082416	The Paxton	51.427880	-0.086168	Pub
3	DULWICH WOOD AVENUE	51.425586	-0.082416	Gipsy Hill Railway Station (GIP)	51.424530	-0.083959	Train Station
4	DULWICH WOOD AVENUE	51.425586	-0.082416	Brown & Green @ The Station	51.424425	-0.083836	Breakfast Spot
5	DULWICH WOOD AVENUE	51.425586	-0.082416	Manuel's Restaurant and Bar	51.427591	-0.086131	Italian Restaurant
6	SOUTH HILL PARK	51.557134	-0.164343	Daunt Books Hampstead	51.555513	-0.166277	Bookstore
7	SOUTH HILL PARK	51.557134	-0.164343	Hampstead Heath Ponds	51.559300	-0.165973	Lake
8	SOUTH HILL PARK	51.557134	-0.164343	Hampstead Heath	51.559622	-0.164921	Park
9	SOUTH HILL PARK	51.557134	-0.164343	Paradise	51.555476	-0.166312	Indian Restaurant
10	SOUTH HILL PARK	51.557134	-0.164343	Keats House	51.555745	-0.167975	Museum
11	SOUTH HILL PARK	51.557134	-0.164343	karma bread	51.554494	-0.165586	Bakery
12	SOUTH HILL PARK	51.557134	-0.164343	The Garden Gate	51.554733	-0.165697	Pub
13	SOUTH HILL PARK	51.557134	-0.164343	The Little Thai	51.554115	-0.164737	Thai Restaurant
14	SOUTH HILL PARK	51.557134	-0.164343	Parliament Hill	51.559661	-0.159639	Scenic Lookout
15	SOUTH HILL PARK	51.557134	-0.164343	Silverberry Deli & Kitchen	51.554174	-0.164975	Café
16	SOUTH HILL PARK	51.557134	-0.164343	The Freemasons Arms	51.556968	-0.168806	Pub
17	SOUTH HILL PARK	51.557134	-0.164343	Zara	51.554423	-0.165561	Greek Restaurant
18	SOUTH HILL PARK	51.557134	-0.164343	The Stag	51.553420	-0.161576	Gastropub
19	SOUTH HILL PARK	51.557134	-0.164343	M&S Simply Food	51.553956	-0.165119	Grocery Store
20	SOUTH HILL PARK	51.557134	-0.164343	Mimmo la Bufala	51.555340	-0.166230	Italian Restaurant

Figure: Result of Final Merge files of cleansing street data and Location data file

The results are full of the unrecognized pattern and based on the clear understating of dataset.



	Street Name	Facility Name	Facility Category	Distance	Facility Latitude	Facility Longitude
0	DULWICH WOOD AVENUE	Tesco	Grocery Store	1311	51.436213	-0.090561
1	DULWICH WOOD AVENUE	Tesco	Supermarket	3911	51.397043	-0.049559
2	DULWICH WOOD AVENUE	M&S Simply Food	Grocery Store	4513	51.389183	-0.111039
3	DULWICH WOOD AVENUE	Tesco	Grocery Store	3677	51.445271	-0.124966
4	DULWICH WOOD AVENUE	Tesco	Grocery Store	5315	51.438880	-0.155979
5	DULWICH WOOD AVENUE	Sainsbury's Local	Grocery Store	3749	51.446901	-0.124255
6	DULWICH WOOD AVENUE	telferscot primary school	Elementary School	4250	51.442807	-0.137087
7	DULWICH WOOD AVENUE	M&S Streatham Hill Foodhall	Grocery Store	3419	51.439372	-0.126443
8	DULWICH WOOD AVENUE	The Co-operative Food	Grocery Store	2306	51.440017	-0.106273
9	DULWICH WOOD AVENUE	Tesco	Supermarket	3216	51.398649	-0.099187
10	DULWICH WOOD AVENUE	Golden Jubilee Wing	Hospital	4836	51.468522	-0.093121
11	DULWICH WOOD AVENUE	Tesco Express	Grocery Store	3634	51.401783	-0.118253
12	DULWICH WOOD AVENUE	The Co-operative Food	Grocery Store	5529	51.471577	-0.112531
13	DULWICH WOOD AVENUE	Tesco Express	Grocery Store	4642	51.465227	-0.103222
14	DULWICH WOOD AVENUE	M&S Foodhall	Grocery Store	3386	51.426259	-0.131190
15	DULWICH WOOD AVENUE	The Co-operative Food	Supermarket	4366	51.432608	-0.020514
16	DULWICH WOOD AVENUE	Aldi	Supermarket	5169	51.448675	-0.017781
17	DULWICH WOOD AVENUE	The Co-operative Food	Supermarket	4934	51.381969	-0.069717
18	DULWICH WOOD AVENUE	Sainsbury's Local	Grocery Store	3155	51.440406	-0.121180
19	DULWICH WOOD AVENUE	Tesco	Grocery Store	6212	51.470177	-0.028570
20	DULWICH WOOD AVENUE	Balgowan Primary School	Elementary School	3811	51.404480	-0.039182
21	DULWICH WOOD AVENUE	Tesco	Grocery Store	5228	51.468581	-0.112767
22	DULWICH WOOD AVENUE	Tesco	Grocery Store	4831	51.385912	-0.110613

Figure: Distance between the Facility and the accommodation street

5. Discussion

Based on the analysis there are two main factors which affect the choice of the buyer are the housing price and the facilities requirement. In case of the facility, the distance between the facilities and accommodation (street location) is deciding factor.

The final analyzed result enlist 39 places where a prospective client can buy the property. Such choices are vary according to the personal requirement and family facility demand.

Few possible cases are:

1. A prospective client with elders in the family would be inclined to choose a location where hospitals and grocery stores are near.

2. A prospective client with kids in the family would choose a location where elementary and high schools are close-by. He would also like to choose a place with parks and other venues to accommodate his family are in the close vicinity.

   As an Example: if the small kid is in the family then ‘elementary school is more important. And in machine leaning it require more weightage. In Examples there are two option of schools from the Dulwich street. The distance measurement from other street location will give better ideas on selection

3. A bachelor would be inclined to choose a property which has pubs, bars, entertainment places close to the property.

6. Conclusion

The decision of a buyer is influenced by the familial needs, personal biases. So, based upon the findings summarized in the results and discussion sections, following conclusions can be made: -

1. While making recommendations to a prospective client, it is imperative to know his/ her immediate needs and requirements besides the budget. This would help to catch his/ her attention.

2. Knowledge about the most recent market prices can be very helpful for the client and can help him take a decision.