RECOMMENDATION FOR BUYERS
FOR REAL ESTATE INVESTMENT IN LONDON

Nikky Priya

April 15, 2019

1. Intoduction

1.1 Background

According to bbc.com, London-focused estate agent Foxtons has swung to a loss and says the housing market in the capital is

in a prolonged downturn.Foxtons said annual revenues fell 5% to £111.5m, with the weakness in property sales being offset

slightly by a resilient lettings performance.Measures of consumer confidence weakened around the turn of the year and

surveyors reported a further fall in new buyer enquiries over the same period.

While the number of properties coming onto the market also slowed, this doesn't appear to have been enough to prevent a modest shift in the balance of demand and supply in favour of buyers in recent months.

1.2 Problem

Hidden price falls, record-low sales, Tax changes, Brexit uncertainty, higher house prices are making it increasingly harder for people to buy homes in the U.K. capital.

To provide support to homebuyers' clientele in to purchase a suitable real estate in London in this uncertain economic and financial scenario.

1.3 Interest

People who want to buy real estate in London City within affordable price and with essential venue and amenities.

2. Data acquisition and cleaning

2.1 Data sources

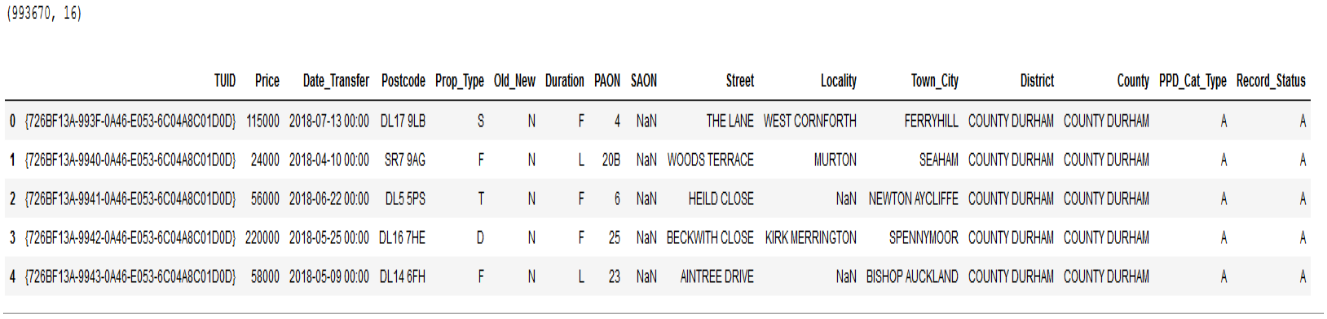
Data required for this problem is London properties details and the relative price paid data. Data was extracted from the HM Land Registry.Address data (http://prod2.publicdata.landregistry.gov.uk.s3-website-eu-west-1.amazonaws.com/pp-2018.csv) included in Price Paid Data: Postcode; PAON Primary Addressable Object Name.House number or name; SAON Secondary Addressable Object Name. If there is a sub-building, for example, the building is divided into flats, there will be a SAON; Street; Locality; Town/City; District; County.

To explore and target recommended locations across different venues according to the presence of amenities and essential facilities, we will access data through Foursquare API interface and arrange them as a data frame for visualization. By merging data on London properties and the relative price paid data from the HM Land Registry and data on amenities and essential facilities surrounding such properties from Foursquare API interface, we will be able to recommend profitable real estate investments. We will use

Clustering technique once the neighbourhood data is extracted using Foursquare API, this will help in understanding the locations with good amenities housing property in reasonable price.

2.2 Data cleaning and Feature Selection

Date field in the data from HM Land Registry was formatted to date type to filter the dataset using year as all the data before 2016 was not considered to make the model relevant for recent transactions.



```
(993670, 16)
```

	TUID	Price	Date_Transfer	Postcode	Prop_Type	Old_New	Duration	PAON	SAON	Street	Locality	Town_City	District	County	PPD_Cat_Type	Record_Status
0	{726BF13A-993F-0A46-E053-6C04A8C01D0D}	115000	2018-07-13 00:00	DL17 9LB	S	N	F	4	NaN	THE LANE	WEST CORNFORTH	FERRYHILL	COUNTY DURHAM	COUNTY DURHAM	A	A
1	{726BF13A-9940-0A46-E053-6C04A8C01D0D}	24000	2018-04-10 00:00	SR7 9AG	F	N	L	20B	NaN	WOODS TERRACE	MURTON	SEAHAM	COUNTY DURHAM	COUNTY DURHAM	A	A
2	{726BF13A-9941-0A46-E053-6C04A8C01D0D}	56000	2018-06-22 00:00	DL5 5PS	T	N	F	6	NaN	HEILD CLOSE	NaN	NEWTON AYCLIFFE	COUNTY DURHAM	COUNTY DURHAM	A	A
3	{726BF13A-9942-0A46-E053-6C04A8C01D0D}	220000	2018-05-25 00:00	DL16 7HE	D	N	F	25	NaN	BECKWITH CLOSE	KIRK MERRINGTON	SPENNYMOOR	COUNTY DURHAM	COUNTY DURHAM	A	A
4	{726BF13A-9943-0A46-E053-6C04A8C01D0D}	58000	2018-05-09 00:00	DL14 6FH	F	N	L	23	NaN	AINTREE DRIVE	NaN	BISHOP AUCKLAND	COUNTY DURHAM	COUNTY DURHAM	A	A

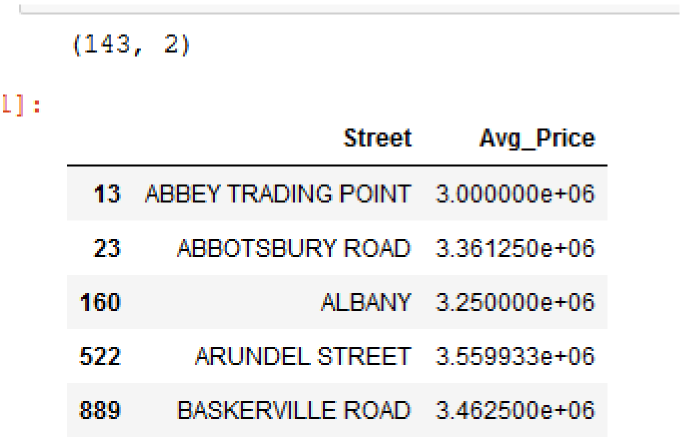
As the column name was missing in the dataset, we manually provided the relevant names.

As our target is to get the estimate only for London city, we filtered out the data with 'Town_City' column.

Now the dataset has only London data available. As we want to check the housing price in neighbourhood within London, only 'Street' and 'Price' column data is relevant for us.

So, we just take these two-column data and save in data frame to proceed with the analysis.

We grouped the streets and populated the average price Street wise in the data frame with two columns "Avg_Price" and "Street".We want to analyse the data within affordable average price in range of 3000000 to 3750000.



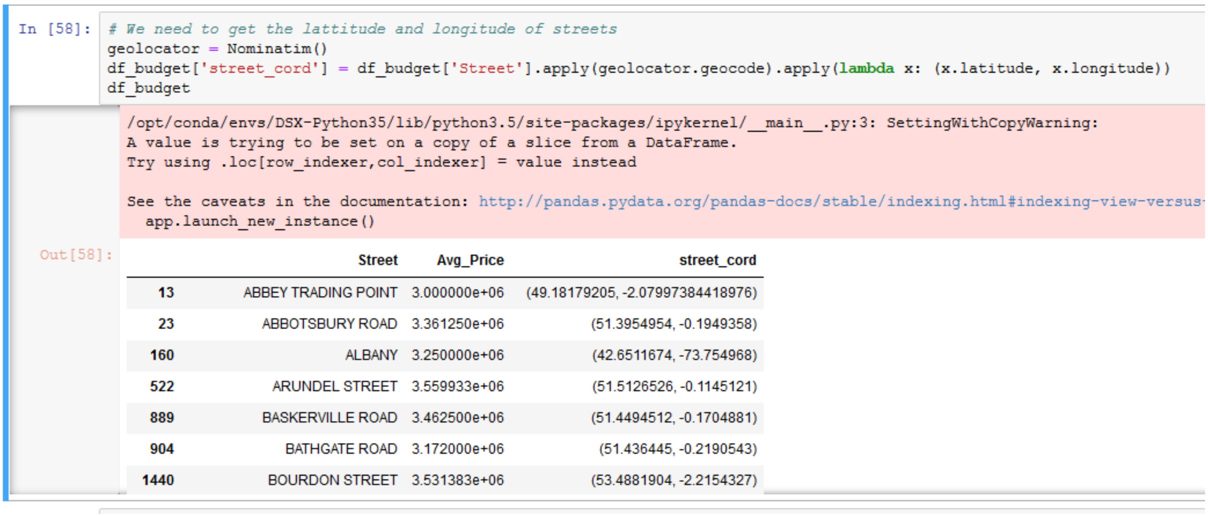
```
(143, 2)
```

l]:

	Street	Avg_Price
13	ABBEY TRADING POINT	3.000000e+06
23	ABBOTSBURY ROAD	3.361250e+06
160	ALBANY	3.250000e+06
522	ARUNDEL STREET	3.559933e+06
889	BASKERVILLE ROAD	3.462500e+06

3. Data Preparation with Longitude and latitude of Streets

Longitude and Latitude of the Streets in above data frame was pulled using geocoder and geolocator as below.



In [58]:
```
# We need to get the lattitude and longitude of streets
geolocator = Nominatim()
df_budget['street_cord'] = df_budget['Street'].apply(geolocator.geocode).apply(lambda x: (x.latitude, x.longitude))
df_budget
```

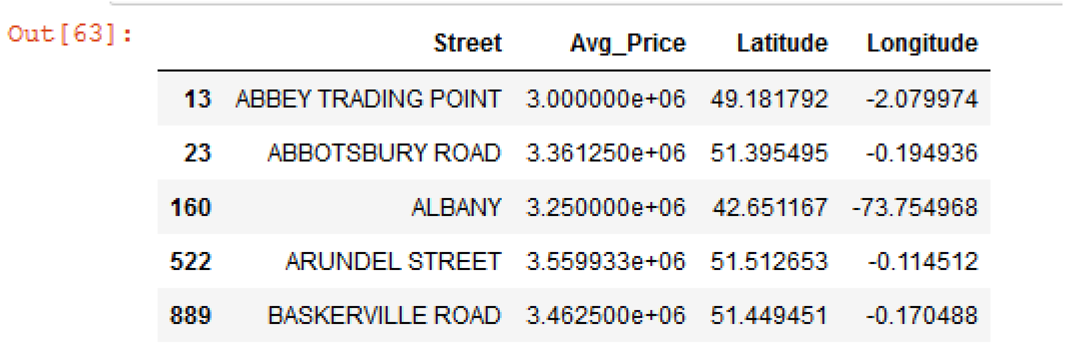
```
/opt/conda/envs/DSX-Python35/lib/python3.5/site-packages/ipykernel/__main__.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-
  app.launch_new_instance()
```

Out[58]:

	Street	Avg_Price	street_cord
13	ABBEY TRADING POINT	3.000000e+06	(49.18179205, -2.07997384418976)
23	ABBOTSBURY ROAD	3.361250e+06	(51.3954954, -0.1949358)
160	ALBANY	3.250000e+06	(42.6511674, -73.754968)
522	ARUNDEL STREET	3.559933e+06	(51.5126526, -0.1145121)
889	BASKERVILLE ROAD	3.462500e+06	(51.4494512, -0.1704881)
904	BATHGATE ROAD	3.172000e+06	(51.436445, -0.2190543)
1440	BOURDON STREET	3.531383e+06	(53.4881904, -2.2154327)

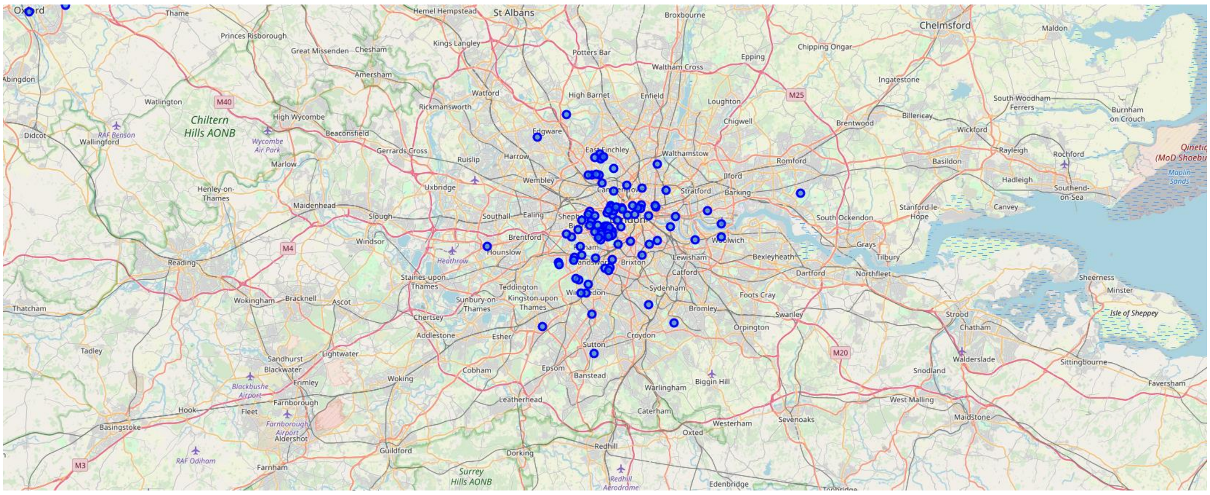
Longitude and Latitude data were separated and stored in different column names in the data frame.



Out[63]:

	Street	Avg_Price	Latitude	Longitude
13	ABBEY TRADING POINT	3.000000e+06	49.181792	-2.079974
23	ABBOTSBURY ROAD	3.361250e+06	51.395495	-0.194936
160	ALBANY	3.250000e+06	42.651167	-73.754968
522	ARUNDEL STREET	3.559933e+06	51.512653	-0.114512
889	BASKERVILLE ROAD	3.462500e+06	51.449451	-0.170488

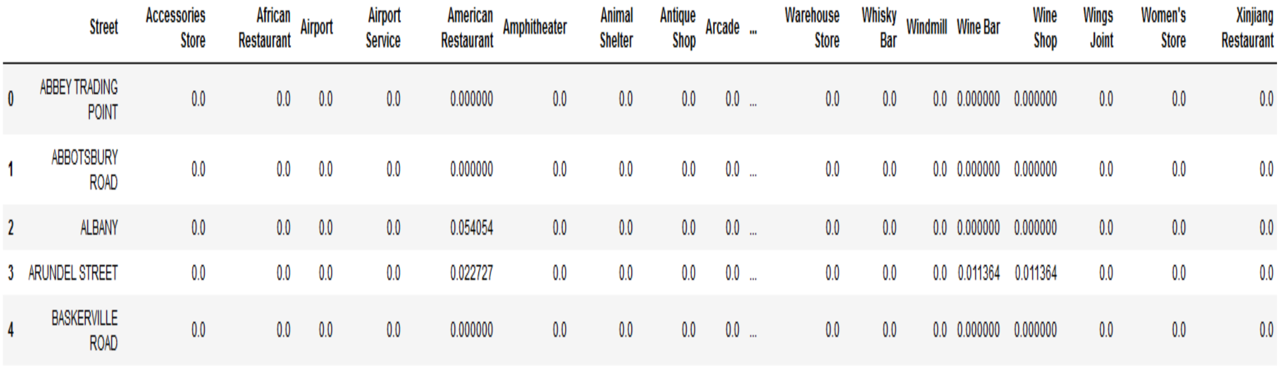
We created the map of streets in London to understand how the housing assets are spread with the selected price range.



4. FourSquare API for further Analysis

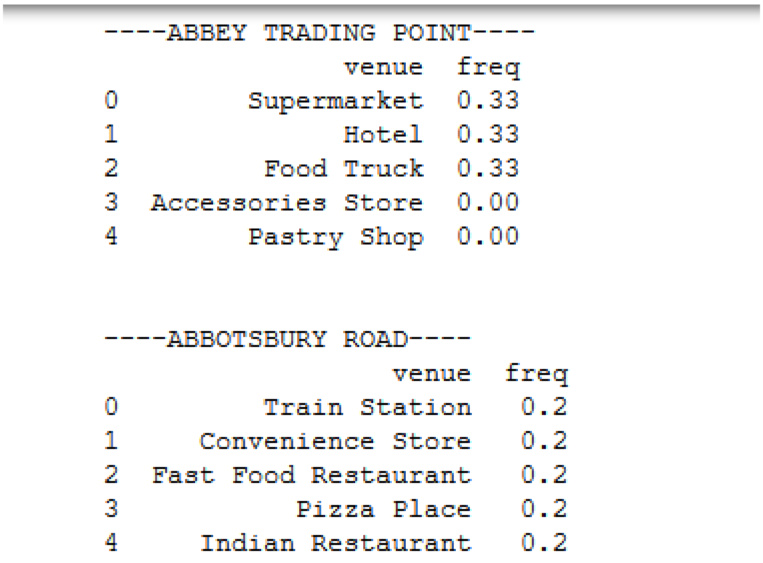
We used FourSquare API to complete the second part of our analysis to understand how and where the buyers can get required essentials like school, restaurant, office, sports, groceries etc.

We pulled the neighbourhood venue details from the FourSquare API for all the streets in our data-frame. Total 5970 rows were generated for venues, we again grouped the data frame by Streets and analysed the data based on venue Category. To perform Clustering based on venues we need to perform one-hot encoding on Venue Category to convert it from categorical to numerical format.



	Street	Accessories Store	African Restaurant	Airport	Airport Service	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Arcade	...	Warehouse Store	Whisky Bar	Windmill	Wine Bar	Wine Shop	Wings Joint	Women's Store	Xinjiang Restaurant
0	ABBEY TRADING POINT	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0
1	ABBOTSBURY ROAD	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0
2	ALBANY	0.0	0.0	0.0	0.0	0.054054	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0
3	ARUNDEL STREET	0.0	0.0	0.0	0.0	0.022727	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.011364	0.011364	0.0	0.0	0.0
4	BASKERVILLE ROAD	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0

We also analysed top 5 venues as below.

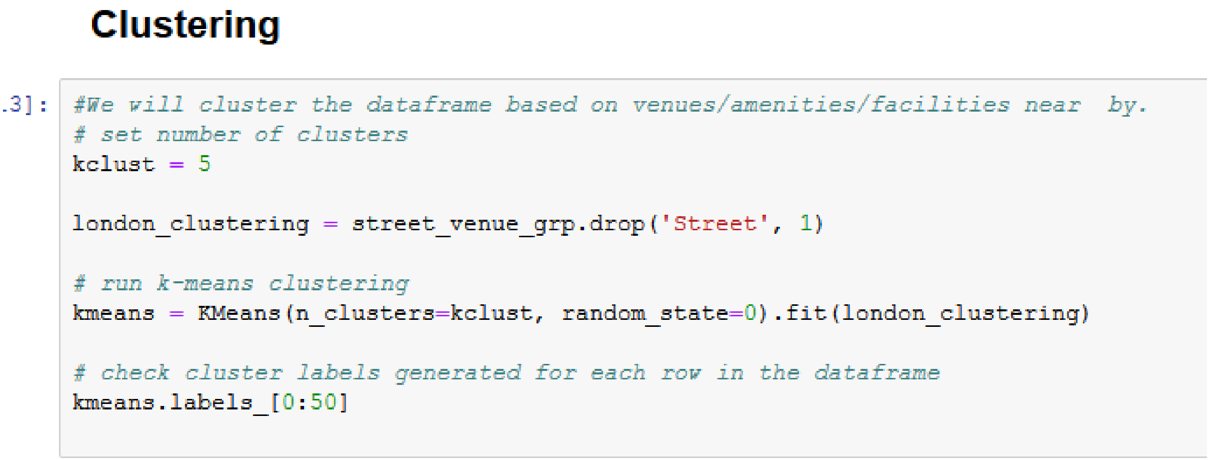


```
----ABBEY TRADING POINT----
             venue  freq
0      Supermarket  0.33
1            Hotel  0.33
2       Food Truck  0.33
3  Accessories Store  0.00
4      Pastry Shop  0.00


----ABBOTSBURY ROAD----
                  venue  freq
0         Train Station   0.2
1     Convenience Store   0.2
2  Fast Food Restaurant   0.2
3          Pizza Place    0.2
4     Indian Restaurant   0.2
```

5. **Clustering**:

We performed K-means clustering on data as below.



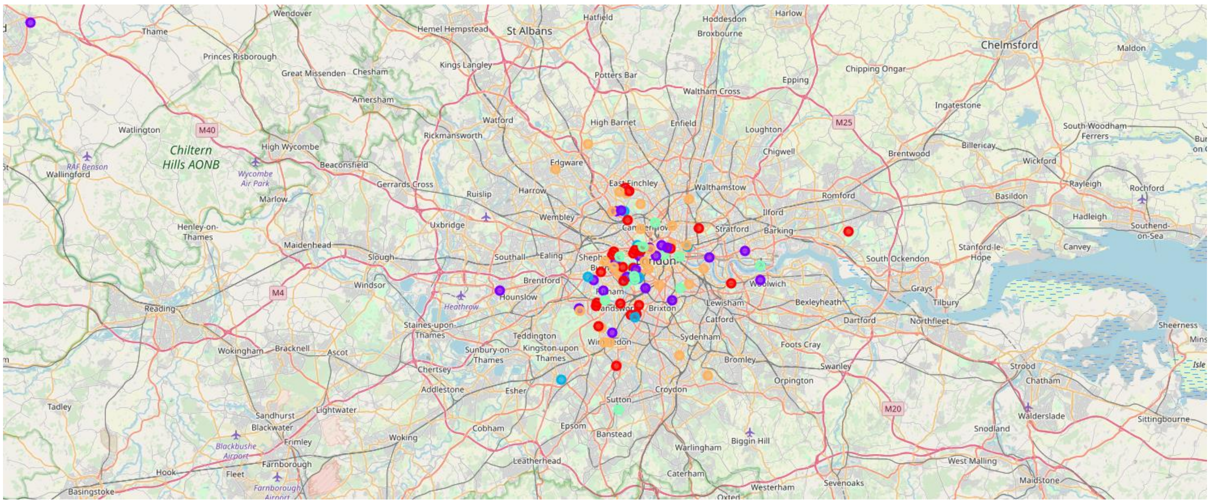
Clustering

```
.3]:  #We will cluster the dataframe based on venues/amenities/facilities near  by.
      # set number of clusters
      kclust = 5

      london_clustering = street_venue_grp.drop('Street', 1)

      # run k-means clustering
      kmeans = KMeans(n_clusters=kclust, random_state=0).fit(london_clustering)

      # check cluster labels generated for each row in the dataframe
      kmeans.labels_[0:50]
```



6. Results and Discussions

Even though the London Housing Market may be in downfall, there is still much opportunity and scope in it.

We have analysed the data in two aspects, one analysis is based on the affordable prices grouped by street/neighbourhood areas and another is by clustering the locations based on the venues and amenities.

Analysis on Neighbourhood areas in London: It is interesting to note that, although West London (Notting Hill, Kensington, Chelsea, Marylebone) and North-West London (Hampsted) might be considered highly profitable venues to purchase a real estate according to amenities and essential facilities surrounding such venues i.e. elementary schools, high schools, hospitals & grocery stores, South-West London (Wandsworth, Balham) and North-West London (Isliington) are arising as next future elite venues with a wide range of amenities and facilities. Accordingly, one might target under-priced real estates in these areas of London in order to make a business affair.

Analysis by Clusters based on amenities and venues nearby: Even though, all clusters could praise an optimal range of facilities and amenities, we have found two main patterns. The first pattern we are referring to, i.e. Clusters 0, 2 and 4, may target home buyers prone to live in 'green' areas with parks, waterfronts. Instead, the second pattern we are referring to, i.e. Clusters 1 and 3, may target individuals who love pubs, theatres and soccer.

7. Conclusion

To conclude, the housing market in the capital (London) is in a prolonged downturn It is now facing several different headwinds, including the prospect of higher taxes and a warning from the Bank of England that U.K. home values could fall as much as 30 percent in the event of a disorderly exit from the European Union. In this scenario, it is urgent to adopt machine learning tools to assist homebuyer’s clientele in London to make wise and effective decisions. As a result, the business problem we were posing was: how could we provide support to homebuyer’s clientele in to purchase a suitable real estate in London in this uncertain economic and financial scenario?

To solve this business problem, we clustered London neighbourhoods to recommend venues and the current average price of real estate where homebuyers can make a real estate investment. We recommended profitable venues according to amenities and essential facilities surrounding such venues i.e. elementary schools, high schools, sports club, restaurant, hospitals & grocery stores.

First, we gathered data on London properties and the relative price paid data were extracted from the HM Land Registry (http://landregistry.data.gov.uk/). Moreover, to explore and target recommended locations across different venues according to the presence of amenities and essential facilities, we accessed data through FourSquare API interface and arranged them as a data frame for visualization. By merging data on London properties and the relative price paid data from the HM Land Registry and data on amenities and essential facilities surrounding such properties from FourSquare API interface, we were able to recommend profitable real estate investments.

Second, The Methodology section comprised four stages: 1. Collect Inspection Data; 2. Explore and Understand Data; 3. Data preparation and pre-processing; 4. Modelling. In particular, in the modelling section, we used the k-means clustering technique as it is fast and efficient in terms of computational cost, is highly flexible to account for mutations in real estate market in London and is accurate.

Finally, we drew the conclusion that even though the London Housing Market may be in a rut, it is still an "ever-green" for business affairs. We discussed our results under two main perspectives. First, we examined them according to neighbourhoods/London areas. although West London (Notting Hill, Kensington, Chelsea, Marylebone) and North-West London (Hampstead) might be considered highly profitable venues to purchase a real estate according to amenities and essential facilities surrounding such venues i.e. elementary schools, high schools, hospitals & grocery stores, South-West London (Wandsworth, Balham) and North-West London (Islington) are arising as next future elite venues with a wide range of amenities and facilities. Accordingly, one might target under-priced real estates in these areas of London to make a business affair. Second, we analysed our results according to the five clusters we produced. While Clusters 0, 2 and 4 may target home buyers prone to live in 'green' areas with parks, waterfronts, Clusters 1 and 3 may target individuals who love pubs, theatres and soccer.