Capstone Project - The Battle of Neighborhoods (Week 1-2)

Business Problem section

Background

According to Bloomberg News, the London Housing Market is in a rut. It is now facing a number of 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. More specifically, four overlooked cracks suggest that the London market may be in worse shape than many realize: hidden price falls, record-low sales, homebuilder exodus and tax hikes addressing overseas buyers of homes in England and Wales.

Business Problem

In this scenario, it is urgent to adopt machine learning tools in order to assist homebuyers clientele in London to make wise and effective decisions. As a result, the business problem we are currently posing is: how could we provide support to homebuyers clientele in to purchase a suitable real estate in London in this uncertain economic and financial scenario?
To solve this business problem, we are going to cluster London neighborhoods in order to recommend venues and the current average price of real estate where homebuyers can make a real estate investment. We will recommend profitable venues according to amenities and essential facilities surrounding such venues i.e. elementary schools, high schools, hospitals & grocery stores.

Data section

Data on London properties and the relative price paid data were extracted from the HM Land Registry (http://landregistry.data.gov.uk/). The following fields comprise the address data included in Price Paid Data: Postcode; PAON Primary Addressable Object Name. Typically the 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 dataframe 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.

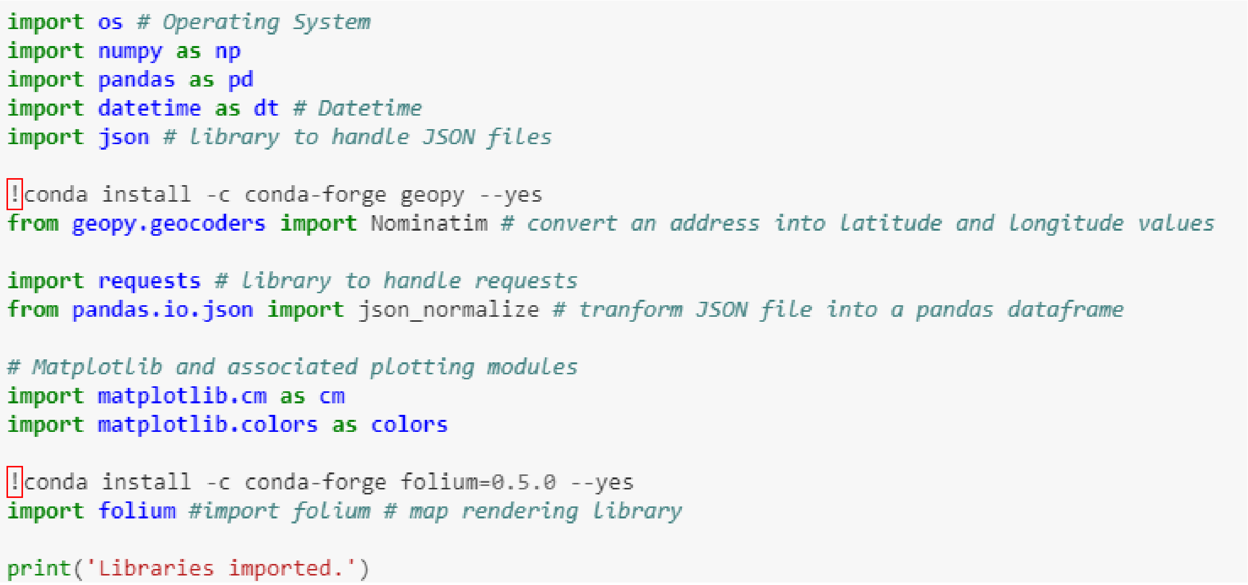
Methodology section

The Methodology section will describe the main components of our analysis and predication system. The Methodology section comprises four stages:

1. Collect Inspection Data
2. Explore and Understand Data
3. Data preparation and preprocessing
4. Modeling

1. Collect Inspection Data

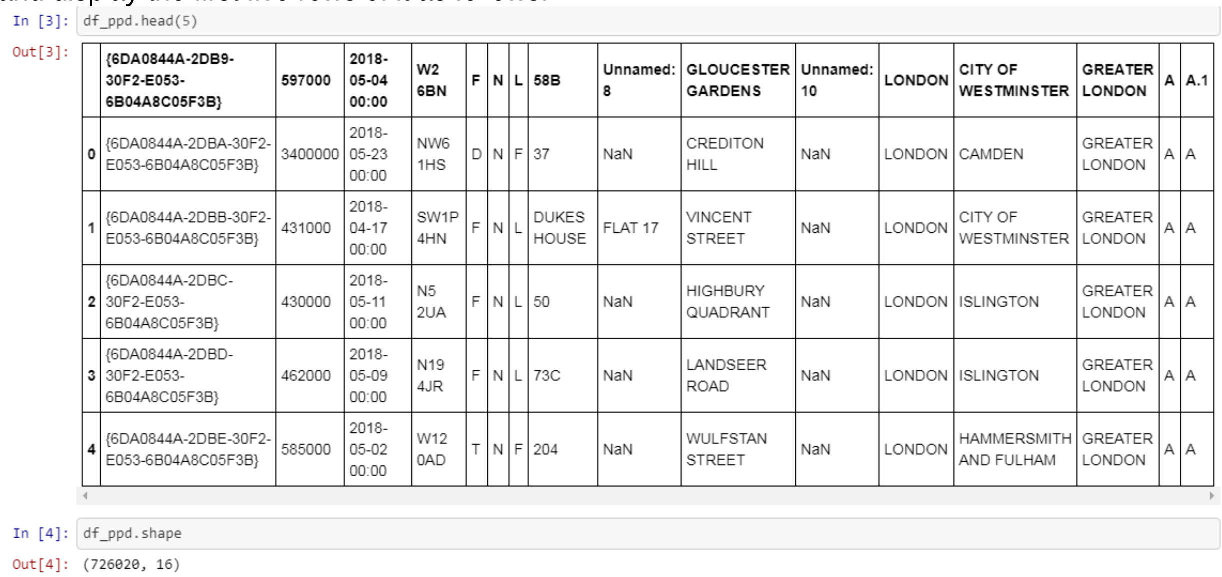
After importing the necessary libraries, we download the data from the HM Land Registry website as follows:



```
import os # Operating System
import numpy as np
import pandas as pd
import datetime as dt # Datetime
import json # library to handle JSON files

!conda install -c conda-forge geopy --yes
from geopy.geocoders import Nominatim # convert an address into latitude and longitude values

import requests # library to handle requests
from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors

!conda install -c conda-forge folium=0.5.0 --yes
import folium #import folium # map rendering library

print('Libraries imported.')
```

2. Explore and Understand Data

We read the dataset that we collected from the HM Land Registry website into a pandas' data frame and display the first five rows of it as follows:



```
In [3]: df_ppd.head(5)
```

Out[3]:

	{6DA0844A-2DB9-30F2-E053-6B04A8C05F3B}	597000	2018-05-04 00:00	W2 6BN	F	N	L	58B	Unnamed: 8	GLOUCESTER GARDENS	Unnamed: 10	LONDON	CITY OF WESTMINSTER	GREATER LONDON	A	A.1
0	{6DA0844A-2DBA-30F2-E053-6B04A8C05F3B}	3400000	2018-05-23 00:00	NW6 1HS	D	N	F	37	NaN	CREDITON HILL	NaN	LONDON	CAMDEN	GREATER LONDON	A	A
1	{6DA0844A-2DBB-30F2-E053-6B04A8C05F3B}	431000	2018-04-17 00:00	SW1P 4HN	F	N	L	DUKES HOUSE	FLAT 17	VINCENT STREET	NaN	LONDON	CITY OF WESTMINSTER	GREATER LONDON	A	A
2	{6DA0844A-2DBC-30F2-E053-6B04A8C05F3B}	430000	2018-05-11 00:00	N5 2UA	F	N	L	50	NaN	HIGHBURY QUADRANT	NaN	LONDON	ISLINGTON	GREATER LONDON	A	A
3	{6DA0844A-2DBD-30F2-E053-6B04A8C05F3B}	462000	2018-05-09 00:00	N19 4JR	F	N	L	73C	NaN	LANDSEER ROAD	NaN	LONDON	ISLINGTON	GREATER LONDON	A	A
4	{6DA0844A-2DBE-30F2-E053-6B04A8C05F3B}	585000	2018-05-02 00:00	W12 0AD	T	N	F	204	NaN	WULFSTAN STREET	NaN	LONDON	HAMMERSMITH AND FULHAM	GREATER LONDON	A	A

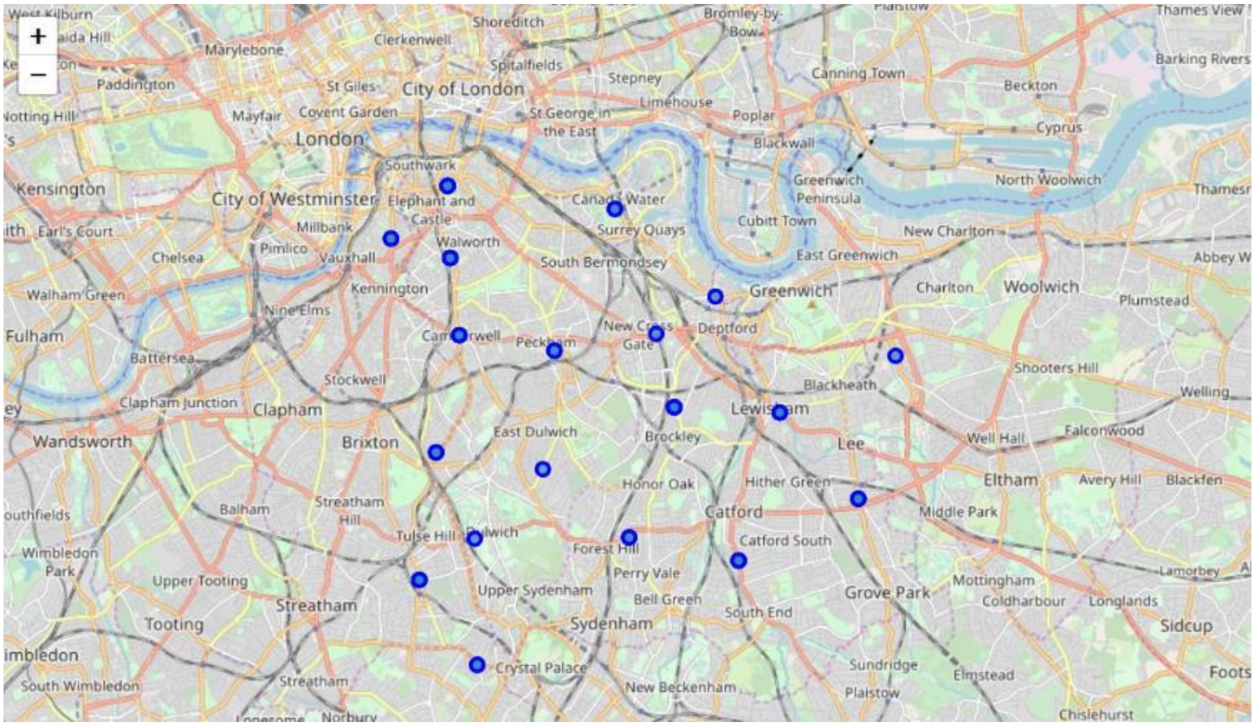
```
In [4]: df_ppd.shape

Out[4]: (726020, 16)
```

Our dataset consists of over 700000 rows and 16 columns. We will now prepare and preprocess data accordingly.

3. Data preparation and preprocessing

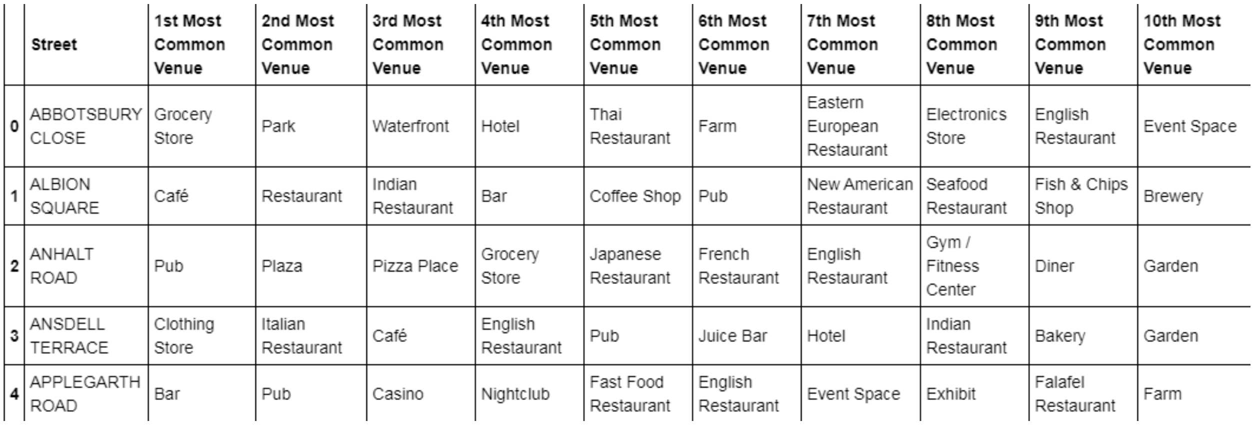
At this stage, we prepare our dataset for the modeling process, opting for the most suitable machine learning algorithm for our scope. Accordingly, we perform the following steps:

- Rename the column names
- Format the date column
- Sort data by date of sale
- Select data only for the city of London
- Make a list of street names in London
- Calculate the street-wise average price of the property
- Read the street-wise coordinates into a data frame, eliminating recurring word London from individual names
- Join the data to find the coordinates of locations which fit into client's budget
- Plot recommended locations on London map along with current market prices



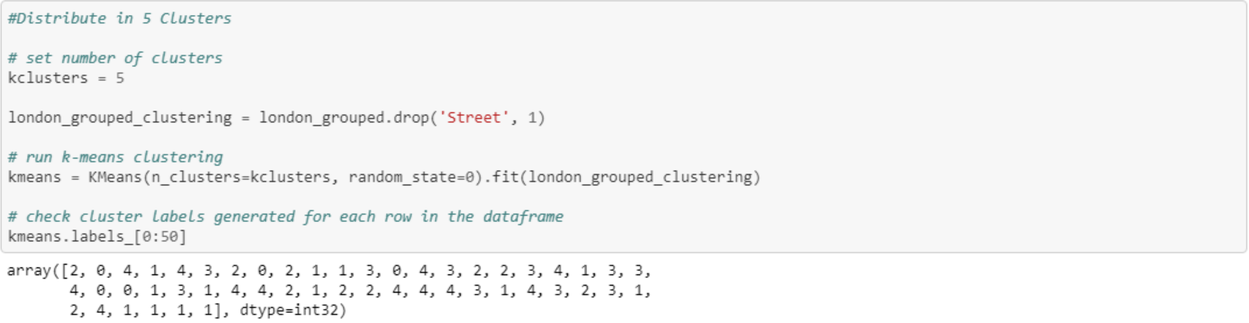
4. Modeling

After exploring the dataset and gaining insights into it, we are ready to use the clustering methodology to analyze real estates. We will use 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.



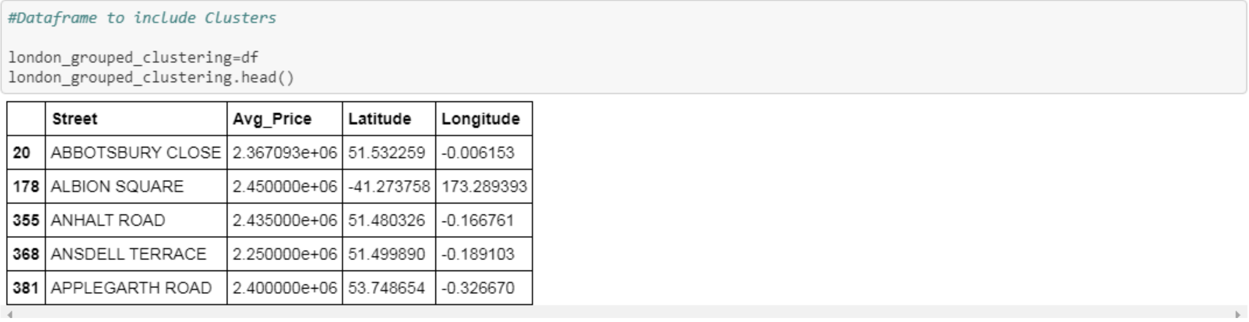
	Street	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	ABBOTSBURY CLOSE	Grocery Store	Park	Waterfront	Hotel	Thai Restaurant	Farm	Eastern European Restaurant	Electronics Store	English Restaurant	Event Space
1	ALBION SQUARE	Café	Restaurant	Indian Restaurant	Bar	Coffee Shop	Pub	New American Restaurant	Seafood Restaurant	Fish & Chips Shop	Brewery
2	ANHALT ROAD	Pub	Plaza	Pizza Place	Grocery Store	Japanese Restaurant	French Restaurant	English Restaurant	Gym / Fitness Center	Diner	Garden
3	ANSDELL TERRACE	Clothing Store	Italian Restaurant	Café	English Restaurant	Pub	Juice Bar	Hotel	Indian Restaurant	Bakery	Garden
4	APPLEGARTH ROAD	Bar	Pub	Casino	Nightclub	Fast Food Restaurant	English Restaurant	Event Space	Exhibit	Falafel Restaurant	Farm

After our inspection of venues/facilities/amenities nearby the most profitable real estate investments in London, we could begin by clustering properties by venues/facilities/amenities nearby.



```
#Distribute in 5 Clusters

# set number of clusters
kclusters = 5

london_grouped_clustering = london_grouped.drop('Street', 1)

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

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

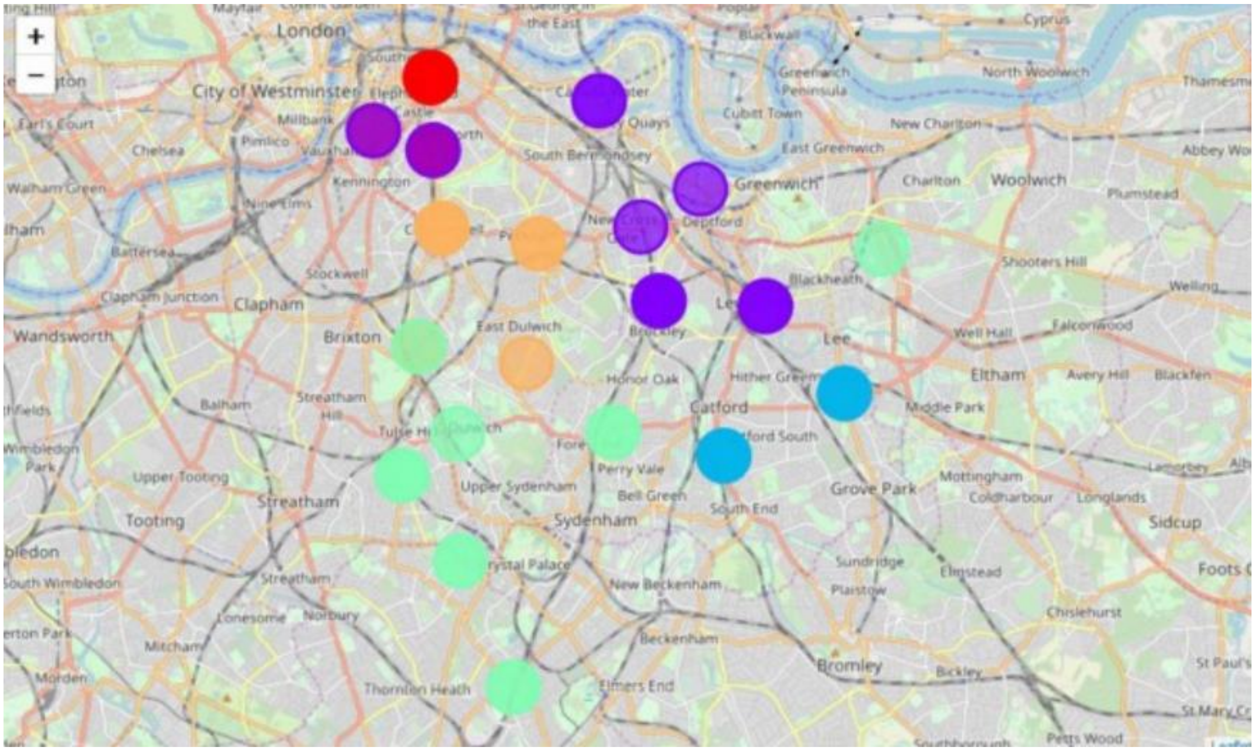
```
array([2, 0, 4, 1, 4, 3, 2, 0, 2, 1, 1, 3, 0, 4, 3, 2, 2, 3, 4, 1, 3, 3,
       4, 0, 0, 1, 3, 1, 4, 4, 2, 1, 2, 2, 4, 4, 4, 3, 1, 4, 3, 2, 3, 1,
       2, 4, 1, 1, 1, 1], dtype=int32)
```



```
#Dataframe to include Clusters

london_grouped_clustering=df
london_grouped_clustering.head()
```

	Street	Avg_Price	Latitude	Longitude
20	ABBOTSBURY CLOSE	2.367093e+06	51.532259	-0.006153
178	ALBION SQUARE	2.450000e+06	-41.273758	173.289393
355	ANHALT ROAD	2.435000e+06	51.480326	-0.166761
368	ANSDELL TERRACE	2.250000e+06	51.499890	-0.189103
381	APPLEGARTH ROAD	2.400000e+06	53.748654	-0.326670

Visualizing the Resulting Clusters — To visualize the clusters, we have the following:



Results and Discussion section

First of all, even though the London Housing Market may be in a rut, it is still an "ever-green" for business affairs.

We may discuss our results under two main perspectives.

First, we may examine them according to neighborhoods/London areas. 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.

Second, we may analyze our results according to the five clusters we have produced. 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.

Conclusion

To sum up, according to Bloomberg News, the London Housing Market is in a rut. It is now facing a number of 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 in order to assist homebuyers 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 homebuyers 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 neighborhoods in order 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, 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 preprocessing; 4. Modeling. In particular, in the modeling 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 neighborhoods/London areas. 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. Second, we analyzed 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.