**Coursera Capstone**

## Capstone Project - The Battle of Neighborhoods

**Author: Deborah Babu Daniel**

**Introduction**

* 1. **Background:**

Settling into a new city can be a very intimidating task. There are a lot of factors to be investigated including finances, location, amenities and so on. London is the capital of and largest city in England and the United Kingdom, with the largest municipal population in the European Union standing on the River Thames in the south-east of England, at the head of its 50-mile estuary leading to the North Sea, London has been a major settlement for two millennia.

This project aims to aid the people who are trying to settle down in London with a budget and looking for a set of amenities to be present to them, since buying a home can be a life altering decision.

* 1. **Problem Statement**

There is a tremendous amount of research that goes into buying a house. The value of the property and the proximity of venues of amenities to the house which is being bought can be very critical factors during the purchase of the house. Usually this involves a lot of driving around and checking out the neighborhoods. The Application of Data Science for this project allow me to perform this analysis from my home to make an educated decision without all the manual work/driving/phone calls.

**Approach:**

The approach followed in this project is as follows:

1) Data from Land registry of London is used to select the streets in which properties are within the budget of the user

2) Geographical data is added to the collected data and the data is categorised into a list of streets and their average pricing.

3) Foursquare API is used to collect the list of venues near the listed streets

4) The streets are clustered and analysed based on the data from foursquare API

5) A table of clustered data along with a map based on it is provided to the user to find the necessary details

**Methodology**

**Data Acquisition and cleaning**

Data Sources

* The value of properties in London based on transaction values available in Land Registry of London - <http://landregistry.data.gov.uk/>
* Geo Spatial data from the python library Geopy
* Foursquare Venue search based on London Zip Code as input for Clustering Analysis.

Data Cleaning

Data was downloaded from the UK Land Registry directly and saved off as a CSV from the site. It was uploaded to the IBM Watson Project as a CSV in files and imported using Python code. The data was striped down to the bare minimum of name, address and property value. The locations dataset was then filtered out based on the price range that was set.

The Geopy library was used to find the Latitude and longitude of the addresses which was in turn used as the input for the foursquare API. The venues near to the street’s coordinates were collected from the foursquare API and were cleaned from unnecessary data like the coordinates of the venues. The data with venue types was used to cluster the data based on K-Means clustering analysis.

Feature Selection:

The most critical features for this clustering analysis came from the following data attributes across data sources:

* London geospatial data from geopy.
* Nearby Venues and Venue Categories for each Zip Code from Foursquare API.

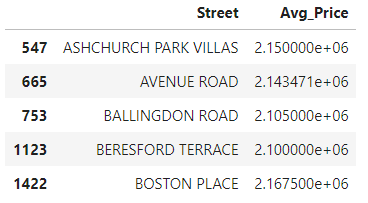
**Data Analysis**

The following libraries were included for the analysis of data collected from the above-mentioned sources:

* Pandas for data framework, reading csv data
* Requests for making http requests
* Sklearn for K-means clustering
* Geopy for geocoding the datasets
* Folium to make maps using the datasets

Pandas library is used to read the CSV file that is downloaded from the UK land registry website. Meaningful names are provided to the data obtained from the CSV. Of the all available data, We require the name of the property, complete address (from the street and locality columns of each item) and the pricing information from the available dataset.

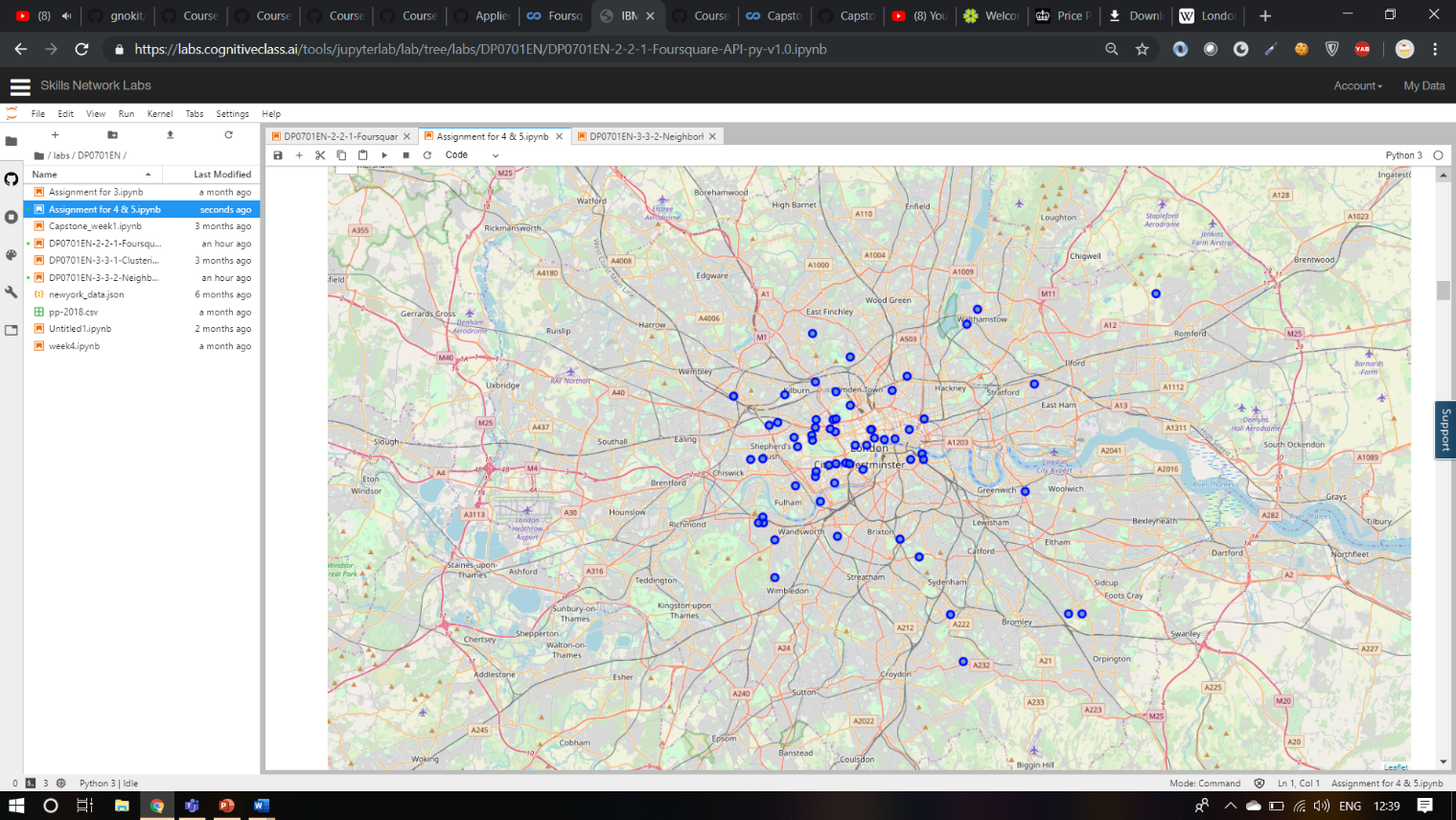
The individual properties are then grouped based on the streets they are in. The complete list of property is used to calculate the average price of properties on a street. This will help us narrow down the locations in which the user should be searching for the property based on the budget. The following table shows a sample of the list of first five streets having an average price between 2100000 and 2200000 Euros arranged in alphabetic order.



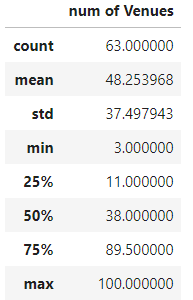
A total number of 63 streets were extracted from the dataset. Then, the geopy library is used to find the geolocation of the streets that are available in the dataset. This will be an input while the venues near the streets are found using the foursquare API. The table below shows the first five items in the dataset with the geospatial data is the set price range arranged in alphabetic order.



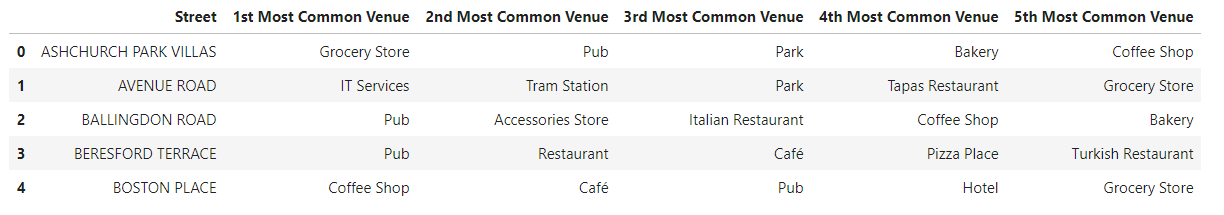
Folium library is used to mark the list of streets in a map with the average price in the marker. The dataset from the previous step is used in creating the map with the locations marked on



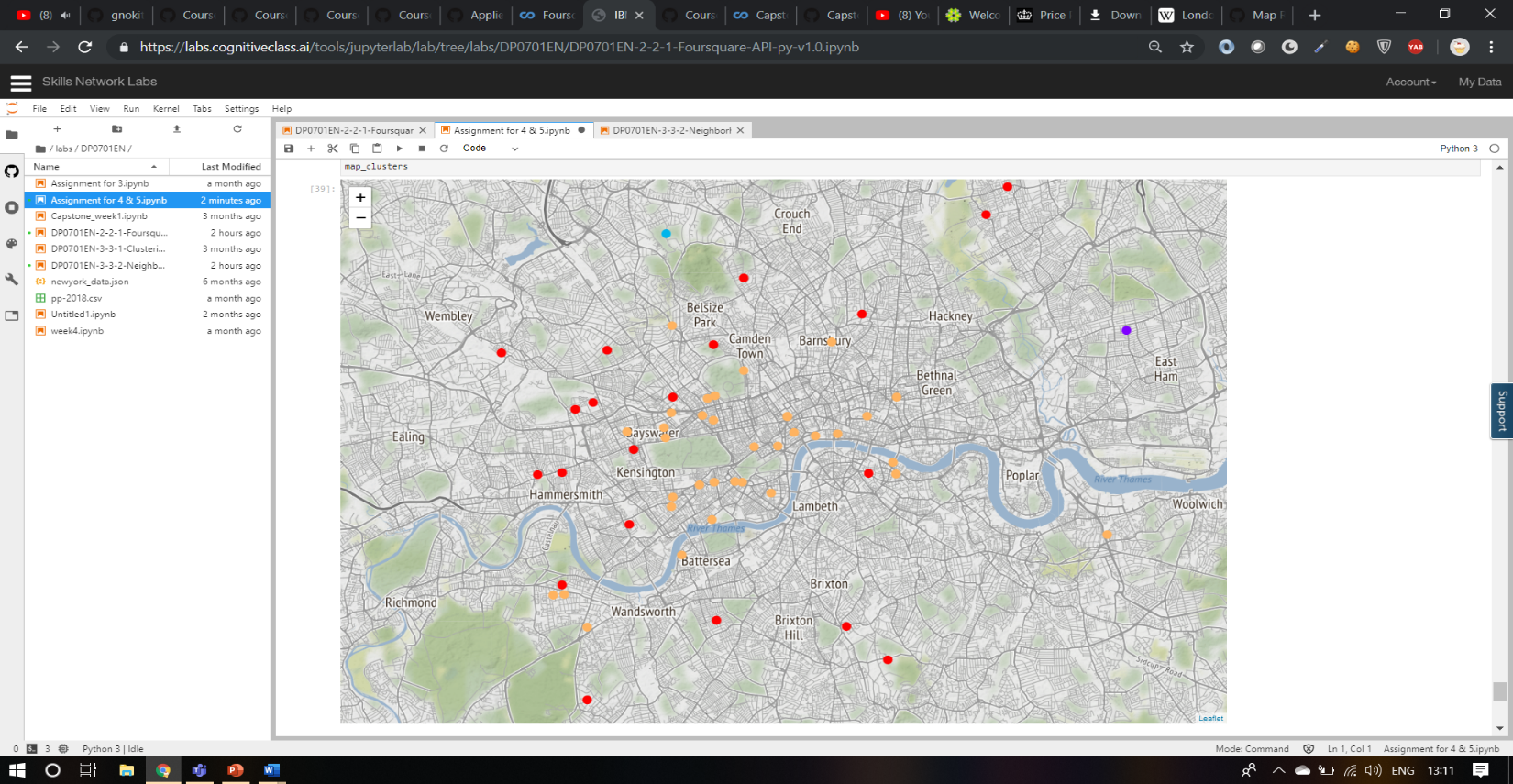
The list of streets and the coordinates are used as input for the foursquare API to find the nearby venues of the streets. While using the foursquare API, 3040 venues and about 265 unique venues were returned. A description of the entire datasets and with the 63 streets are as follows



One shot dataset of the 63 locations are made from the data obtained from the venue list obtained from the venue list obtained from the foursquare API. The top two common venues are obtained from the dataset that was previously created. A table showing the top five venues in the first five streets arranged in alphabetical order is shown below



The dataset is used to perform the K-mean clustering. The dataset is grouped into 5 clusters. The grouping is marked into the map showing the clusters. A map showing the cluster of locations based on the venues are shown below:



**Conclusion and discussion**

In current world, migration is a common but intimidating thing and people move from places to place and from time to time. This analysis gives us an idea of average price of houses in London, their locality and what their surroundings look like which can be used by people who wants to move to London. The location in clusters give an idea of the locations that the users are supposed to looked into.

**Future Study**

The crime history of the location can also be analysed before choosing the locations that the users are supposed to look into.