Final assignment - Capstone project

Examining London’s wards

Introduction

My project is about finding a proper way to buy a house in London. Instead of just pick a nice one, I'd like to find the best one according to my preferences. I am going to use Foursquare and clustering to solve the problem and get the neighborhood where I might to look around to live.

Also, I have preferences, like

- least populated area where
- the average price is lower than £350.000
- it should be safe for sure, and
- it would be just nice if the people around were not too old

So, I have strict requests to find the neighborhood where I should start the search.

To solve the problem, I would create a dataset which will contain useful information about London's districts, and I will use this one to get data from Foursquare.

Those who wants to buy a new house might be interested on the method I applied here.

Data

Administrative areas

The table (requested from https://www.doogal.co.uk/AdministrativeAreas.php) contains the postcode data filtered by administrative area that is Districts. The dataset contain the codes of every districts and their geographical data i.e. Latitude and Longitude.

London postcodes

The table requested from https://www.doogal.co.uk/london_postcodes.php

This is a complete list of London postcode districts with their Ordnance Survey coordinates and longitude and latitude. It seems cool but notice the Latitude and Longitude data are go with the Postcode which is a bit deep level for this examination.
On the other hand, it contains the data of **District Code**, **Ward Code**, **District**, **Ward**, **Rural/urban**. Those data would be useful:

- District Code and the Ward Code for merging with other datasets,
- Rural/urban for the analysis

Ward level geographical data

The table requested from http://geoportal.statistics.gov.uk/datasets/wards-december-2018fullclipped-boundaries-gb/data

To get geographical data about wards. It does not contain ward level data so I could merge it with the *London postcodes* dataset

London borough profile

The table requested from https://data.london.gov.uk/dataset/london-borough-profiles

Borough profile displays data for that borough, plus either Inner or Outer London, London and a national comparator (usually England where data is available). The data is set out across 11 themes covering most of the key indicators relating to demographic, economic, social and environmental data.

I would mainly use **GLA Population Estimate 2017**, **GLA Household Estimate 2017**, **Population density (per hectare) 2017**, **Average Age, 2017** columns to help me to get closer to the desirable new House.

Recorded crime summary

The table requested from https://data.london.gov.uk/dataset/recorded_crime_summary

This data counts the number of crimes at three different geographic levels of London (borough, ward, LSOA) per month, according to crime type. I will not show all the data provided, because I will summarize them i.e. I need only one number per wards to help me compare the safety of wards

House Price Index - HPI

The table requested from https://data.london.gov.uk/dataset/uk-house-price-index

The UK House Price Index (UK HPI) captures changes in the value of residential properties.

The UK HPI uses sales data collected on residential housing transactions, whether for cash or with a mortgage.

The **Average price** and the **Sales volume** are available from this table which I definitely need to solve the problem

---

Get the neighborhood

Once I collected all the needed data and created the dataset I want, I will start to examine it to find the proper neighborhood which is:

- least crowded, so least populated
- the average price is lower than £350.000
- safe
- the area is youthful

Once I get the neighborhood which is close enough to my expectation (filtered data) I will use the data to make a map and then to make clusters to see the venues.

Based on the filtered data and the venues I am going to have the idea where to start the searching.

Methodology

In this project I am going to make my effort to detecting the correspondent area of London with relatively low crime volume, populated mostly with youth and for sure with a convenient price. I am using Foursquare to find a family friendly **Ward** with regarding venues.

I will limit the analysis to area ~1km around the **Ward** I found.

In first step I have collected the required **data: Location data** (Postcodes, Latitude, Longitude), **Borough profile, Crime data** and **House Price Index**.

Get Administrative areas

Since we need only three columns from it, I extract *District Code, Latitude, Longitude* Also I rename those columns to show they belong to Districts

Get London postcodes then check the dataset

Create a shorter dataset by get just the relevant columns and roll them up, so I have only one row for each Ward

Get Ward level geographical data

Since there are a lot of columns I do not need right now, I am going to keep only *wd18cd*, *lat*, *long* columns and rename them as *Ward Code*, *Ward Latitude* and *Ward Longitude*. *Ward Code* will be used as the key to merge with df_full dataset

Get London borough profile

The data is on the 'Data' sheet in an excel file. Also, after an examination I realized I do not need all of the columns but just a few so I grab only those. The first row contains nothing so just skip it. I need to rename the *New code* column too, which is our *District Code*

Get Recorded crime summary

I will use only the data of year 2017 (it starts with 201702) to be consistent with the other data. We have *WardCode* in there but after I tried it I realized it is not as good to use to merge as the column *Borough*, so I keep that one. Also, I do not need to know this time what kind of crime has been committed so i just create a **Total** column for them. To create the **Total** column I am using *.loc[]* and aggregate all the columns containing crime data. The crime dataset did not contain any row on *City of London*, so I got **NaN** for these rows. I do not want to get rid of these rows so I decided to replace those **NaN**-s with the mean Also I need each borough only once, so I aggregate the rows as well.

Get House Price Index

I will get the data in two steps:

- in the first step I will get the *average price* which is on the sheet Average price
- in the second step I get the *sales volume* from the Sales volume sheet

From those excel sheets I need only the **Districts**, their **Districts Code** and the close data of 2017 which is **2017-12-01**

Then transpose the dataset and rename the columns

After I got the data, I merge all og them into a new dataframe called df_full. First I use df_london dataset as a basic so when I am using **pd.merge()** on them I connect the tables with a *left join*.

In the second step I am filtering the dataframe so I will find the **Ward** which attributes are close enough to my expectation.

In the third step I will use clustering technique to find differences among the Wards. I will use those clusters to decide where should I start to explore the neighborhood personally. I will present map of all such locations as well.

Analysis / Exploration

Okay, that was all the data I needed. I put together so let's check them out.

Let's start with the location of the **Wards** Observe the table we got:

1. GLA Population Estimate 2017 ○ as you can guess, Urban city is the most populated one however surprise, surprise, not the rural site is the least populated but, the conurbation

2. Population density (per hectare) 2017 ○ That one is interesting: I expected the city to have the highest density on population, but I was wrong. However, the hamlets are the least crowded area

3. Average Age, 2017 ○ Conurbation again. Those areas have the youth
4. Crime volume in 2017 ○ Most of the crimes are committed in the conurbation area. The presence of the police might have more patrol in the inner city or maybe the conurbation area is just too big to supervise

5. AVG price on 2017-12-01 ○ I assume that a lot of people move to conurbation recently. This can be the explanation of the high average price, also it can explain the density

6. Sales volume on 2017-12-01 ○ Despite the comments above the conurbation has the least sales volume

There are a lot of questions above which are out of our topic now, because the history is not important right now, however it would worth a proper examination.

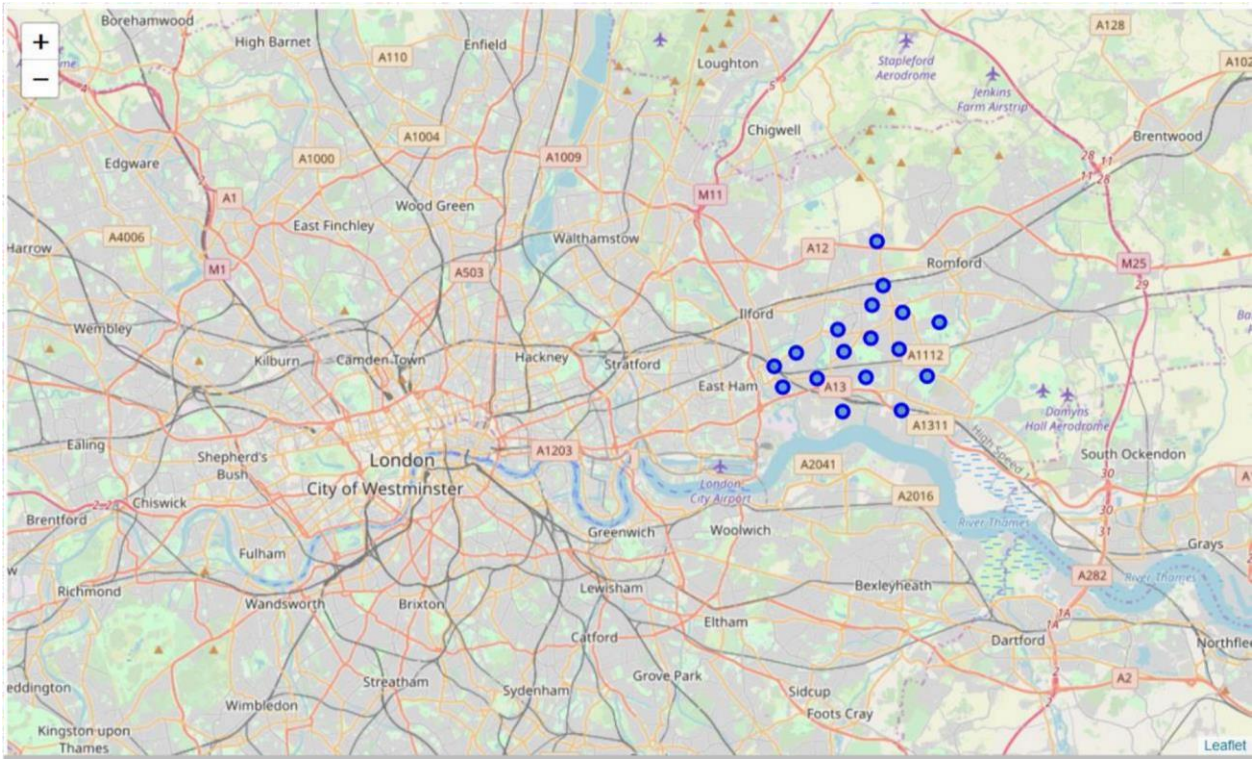
Now, let's focus on why we are here.

Let's apply our requests on the dataset

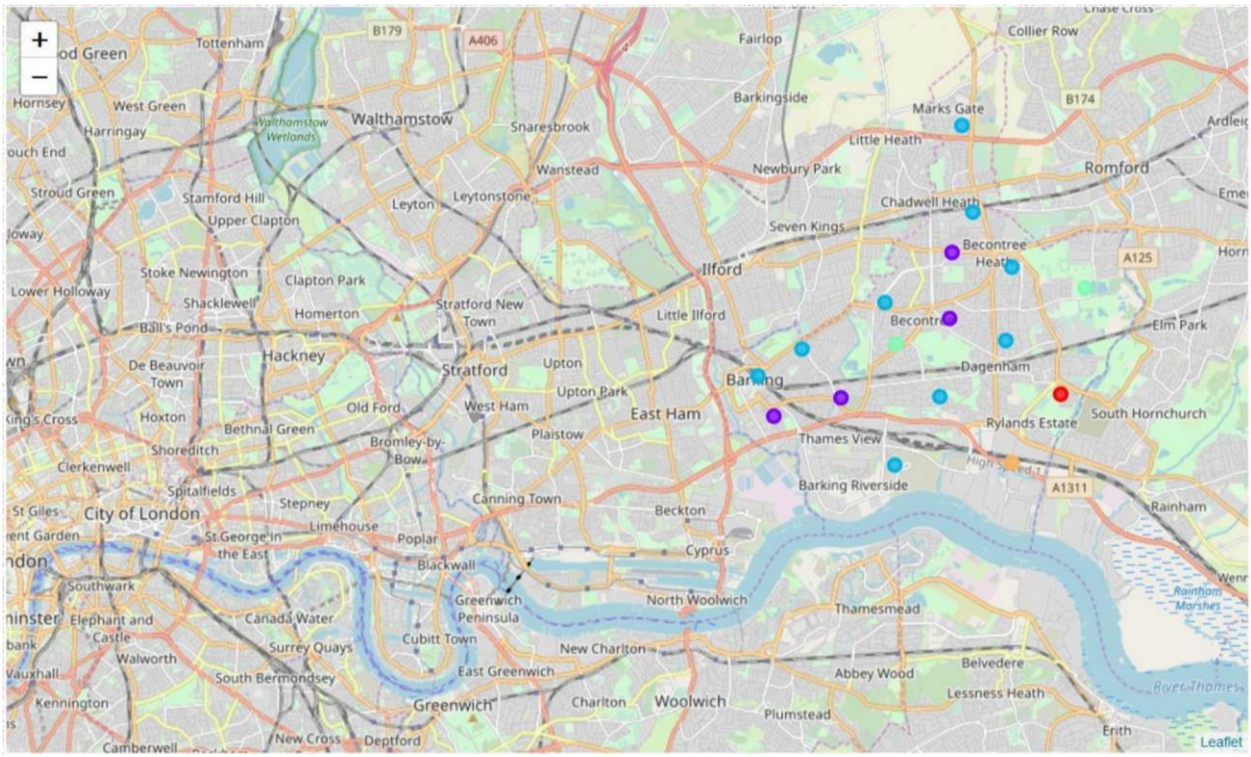
Mapping London

I am using geopy library to get the latitude and longitude values of London.

In order to define an instance of the geocoder, I define a user_agent. I will name our agent *UK_explorer*, as shown below.



Also, I visualize the resulting clusters



Results and Discussion

My examination contains data coming from different aspects.

During the analysis I found that

- the most populated areas are Cities; however, the conurbation is the least which was against my expectations. I expected to the rural site would be the least populated areas. There is a chance that country people are moving closer to the cities but the city itself is already too crowded or expensive to them
- that can be the reason why my other expectation was wrong again. Not the City is which have the highest population density among the observed areas. Maybe they are too expensive to live maybe it has another reason, the find the reason was not part of this project. Hamlets are still the least crowded area, no surprise on this.
- When it comes the average age of the observed areas, I found something which can reinforce my assumption taken before. The conurbation area has the youth. So, it can mean that the younger generation is moving there from the rural.
- No surprise: a lot of people are out there in the conurbation so most of the crimes are committed in here. The presence of the police might have more patrol in the inner city or maybe the conurbation area is just too big to supervise.
- I assume that a lot of people move to conurbation recently. This can be the explanation of the high average price, also it can explain the density
- Despite the observations made before the conurbation area has the least sales volume.

The clusters enlightened the question from another point of view.

Since I filtered the dataframe before, all the clusters are very similar, however we can find the differences. Some clusters seem to be more approachable with public transport, and the others are more "picnic" friendly.

Conclusion

There are a lot of questions out there which were not covered by my analysis but that is okay.

I just wanted to find the area which attributes looks good enough to check the field personally.

Based on the data and the clusters information I would check **Cluster 1** and **Cluster 2**. According the data and the venues information they seem work to me.

They both approachable with public transport which is necessary to me. Also, they have places which can make the weekdays easier.
Right now, I am satisfied with the results so I would check **Cluster 1** at first.