Applied Data Science Capstone Project: Housing Rental Prices & Venues Data Analysis of London

Georgios Mittas

Athens, Greece

Introduction

In the last 3 months, many of my private plans were cancelled due to the Corona virus lock-down situation. I used some of the additional free time to understand the fundamentals of data science and learn how to practically use various tools and methods such as IBM Watson Studio, Jupyter Notebook, MySQL and Python with its libraries, such as pandas, folium, numpy, geopy, Scikit learn etc. In my current position as investment analyst, I do not aim to regularly program machine learning applications myself, but knowing the basics is totally useful to work in this environment. I would like to recommend this series of courses: the IBM Data Science Professional Certificate: https://www.coursera.org/professional-certificates/ibm-data-science

This article was written as part of final capstone project for IBM Data Science Professional Certification in Coursera, in which many of the tools and methods learned throughout the recent months are applied. This project deals with discussing the neighborhoods of London, capital of United Kingdom. It would specifically help individuals & investors planning to invest in real estate housing market, in terms of rental income and social venues attributes. Hope you like it.

Business Problem

London is the capital and largest city of England and the United Kingdom. It has a diverse range of people and cultures, and more than 300 languages are spoken in the region. Its estimated mid-2018 municipal population (corresponding to Greater London) was roughly 9 million, which made it the third-most populous city in Europe. London has some of the highest real estate prices in the world.As of 2015 the residential property in London is worth $2.2 trillion – same value as that of Brazil's annual GDP.The city has the highest property prices of any European city according to the Office for National Statistics and the European Office of Statistics. On average the price per square meter in central London is €24,252. This is higher than the property prices in other G8 European capital cities.

As from a local resident point of view we want to invest in such places were the housing rental prices are low, the facilities (shops, restaurants, parks, Hotels, etc.) and social venues are nearby. Keeping above things in mind it is very difficult for an individual to find such place in such big city and gather this much information. When we consider all these problems, we can create a map and information chart where the real estate rental values index is placed on London and each district is clustered according to the venue density.

Data Description

- London has multiple neighborhoods. I found the List of areas of London with its boroughs and postal codes from Wikipedia (https://en.wikipedia.org/wiki/List_of_areas_of_London).
- In order to obtain the venue details in each neighborhood, Foursquare API is used (https://developer.foursquare.com).
- For housing prices a website, where latest London house prices are available with postal codes distribution, is selected (https://propertydata.co.uk/cities/london).
- For chloropleth maps construction, a .geojson file of London is used. (https://joshuaboyd1.carto.com/tables/london_boroughs_proper/public)

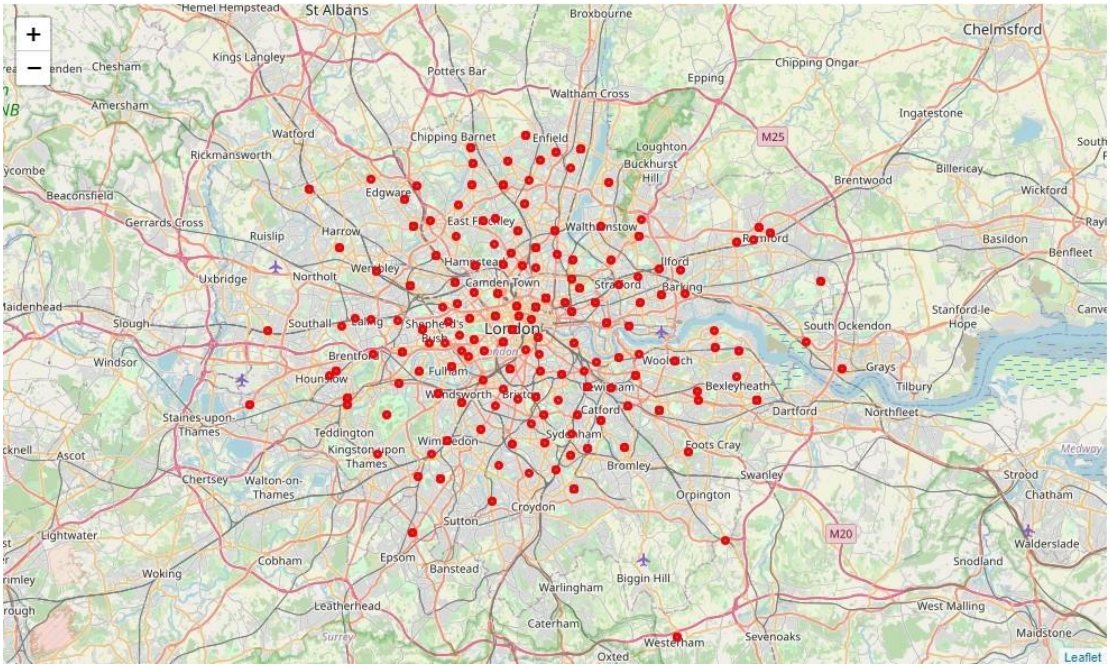
Methodology

In this section, I will describe the data analysis and how I process the data to create a solid and accurate data frame.

Starting out, I loaded 2 data tables from 2 sources Wikipedia & Property_data using the wikipedia page and the pandas read functions. I removed all the hyperlinks, nulls values and then get rid of unnecessary columns. The 'Avg Rent' (£ monthly / per unit) column contains string so I processed it to make integer by removing pound sign and comma. After "cleaning" the two tables I performed inner join and merge them into a unified data frame. Finally by using geocoder library, I load the coordinates of each location and add them as columns in my data frame. The result is the following data frame (top 5 rows):

	Location	London_borough	Postcode district	Avg rent	Latitude	Longitude
0	Abbey Wood	Bexley	SE2	1182.088333	51.492450	0.121270
1	Crossness	Bexley	SE2	1182.088333	51.492450	0.121270
2	West Heath	Bexley	SE2	1182.088333	51.492450	0.121270
3	Acton	Ealing	W3	1439.957750	51.513240	-0.267460
4	Addington	Croydon	CR0	1125.042750	51.384755	-0.051498

I used python **folium** library to visualize geographic details of London and its boroughs and I created a map of London with boroughs superimposed on top. I used latitude and longitude values to get the visual as below:



I utilized the Foursquare API to explore the boroughs and segment them. I designed the limit as **100 venues** and the radius **1,000 meters** for each borough from their given coordinates. The result was a list of 17,895 venues all over London city, which came from 317 unique categories, such as Supermarket, Grocery Store, Hotel, Park etc. Here is a head of the list Venues name, category and coordinates information from Foursquare API.

	name	categories	lat	lng
0	Lesnes Abbey	Historic Site	51.489526	0.125839
1	Sainsbury's	Supermarket	51.492826	0.120524
2	Lidl	Supermarket	51.496152	0.118417
3	Abbey Wood Railway Station (ABW)	Train Station	51.490825	0.123432
4	Co-op Food	Grocery Store	51.487650	0.113490

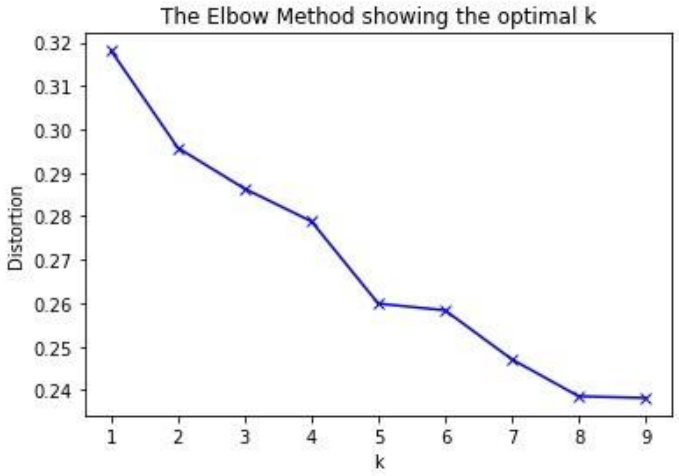
This information is used to create a data frame in which you can see the most common venue types for each city district. Specifically, by using the Foursquare API in conjunction with the created datasets, a table of the 10 most common venues in London neighborhoods is generated.

	Neighborhood	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	Abbey Wood	Supermarket	Train Station	Convenience Store	Coffee Shop	Historic Site	Zoo Exhibit	Fish Market	Farmers Market	Fast Food Restaurant	Filipino Restaurant
1	Acton	Grocery Store	Breakfast Spot	Indian Restaurant	Park	Train Station	Zoo Exhibit	Flea Market	Fast Food Restaurant	Filipino Restaurant	Film Studio
2	Addington	Grocery Store	Fish & Chips Shop	Art Gallery	Zoo Exhibit	Flower Shop	Fast Food Restaurant	Filipino Restaurant	Film Studio	Fish Market	Flea Market
3	Addiscombe	Grocery Store	Fish & Chips Shop	Art Gallery	Zoo Exhibit	Flower Shop	Fast Food Restaurant	Filipino Restaurant	Film Studio	Fish Market	Flea Market
4	Albany Park	Hotel	Monument / Landmark	Plaza	Theater	Garden	Pub	Boutique	Japanese Restaurant	Sandwich Place	Cocktail Bar

Clustering

We have common venue categories in boroughs. For this reason unsupervised learning K-means algorithm is used to cluster the boroughs. K-Means algorithm is one of the most common cluster methods of unsupervised learning.

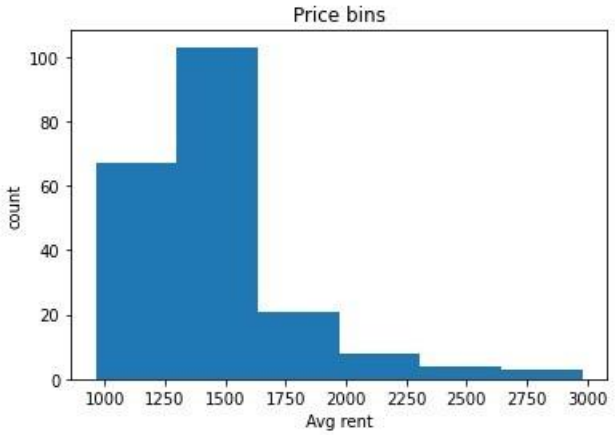
To begin with, I will run K-Means to cluster the boroughs into 6 clusters because when I analyze the K-Means with elbow method it ensured me the 6 degree for optimum k of the K-Means.



As a result, a merged table with cluster labels for each borough was created. After examination of each cluster attributes, each cluster was labeled, as follows:

- Cluster 0 : “Hotels and Social Activities”
- Cluster 1 : “Mixed Social Activities”
- Cluster 2 : “Specific uses”
- Cluster 3 : “Pubs and cafes”
- Cluster 4 : “Sports and Parks”
- Cluster 5 : “Restaurants and Bars”

We can also examine the frequency of average housing rental prices (£ monthly / per unit) for every borough of London, in different ranges. Thus, histogram can help to visualization:



As it seems in above histogram, the ranges of average monthly rental prices for every borough of London can be defined as below:

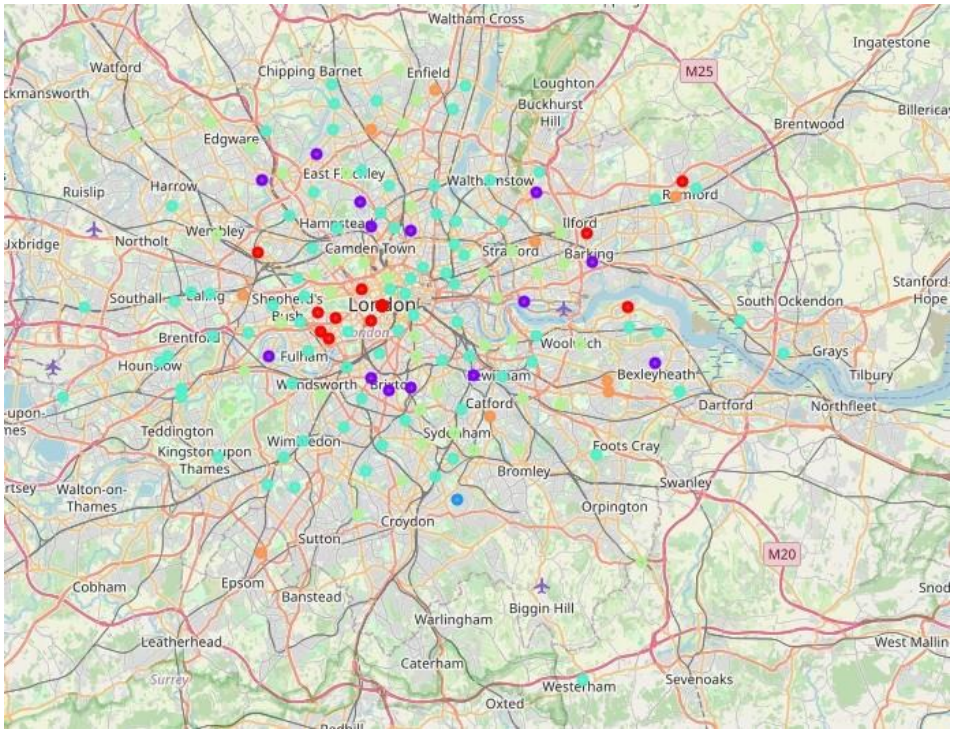
- < £1,300 : “Very Low”
- £1,300 –1,650 : “Low”
- £1,650 –2,000 : “Average”
- £2,000 –2,300 : “Above Average”
- £2,300 –2,650 : “High ”
- > £2,650 : “Very High”

The new variables with related cluster information are merged in our main **master table**. We can now examine Price & Cluster categories as the last two in below table, for each neighborhood of London.

	Location	London_borough	Postcode district	Avg rent	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Price-Categories	Cluster-Category
0	Abbey Wood	Bexley	SE2	1182	51.492450	0.121270	3	Supermarket	Platform	Train Station	Convenience Store	Coffee Shop	Very Low	Pubs and cafes
1	Acton	Ealing	W3	1439	51.513240	-0.267460	5	Grocery Store	Indian Restaurant	Train Station	Breakfast Spot	Park	Low	Restaurants and Bars
2	Addington	Croydon	CR0	1125	51.384755	-0.051498	2	Construction & Landscaping	Grocery Store	Zoo Exhibit	Fish & Chips Shop	Fabric Shop	Very Low	Specific uses
3	Albany Park	Bexley	DA5	1333	51.506420	-0.127210	0	Hotel	Garden	Theater	Plaza	Monument / Landmark	Low	Hotels and Social Activities
4	Aldborough Hatch	Redbridge	IG2	1249	51.506420	-0.127210	0	Hotel	Garden	Theater	Plaza	Monument / Landmark	Very Low	Hotels and Social Activities

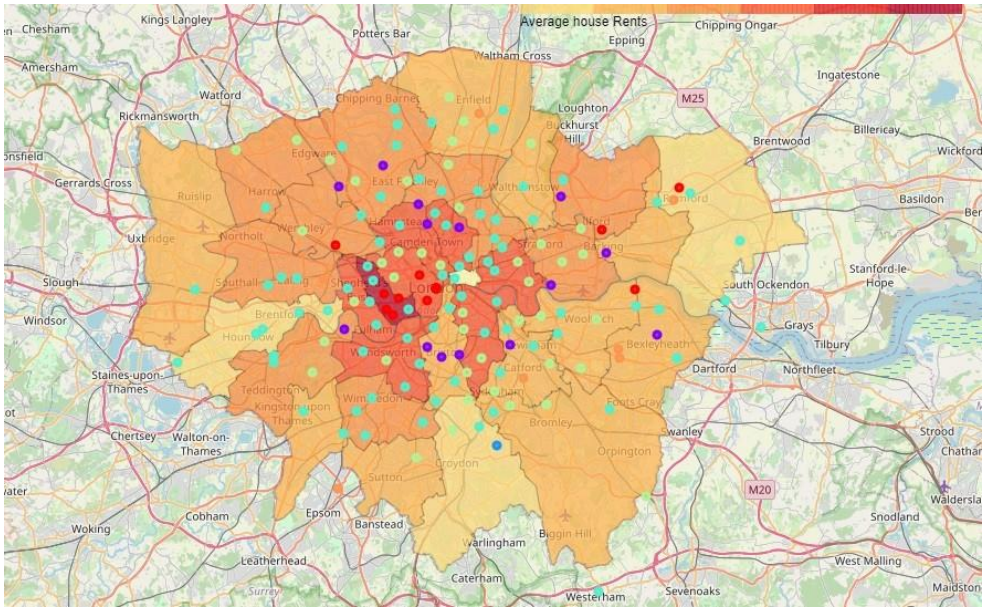
Results

Cluster labels are used to show the city boroughs of London, marked with a cluster-specific color on a map, using folium library:



- *Red (Hotels & Social Activities)*
- *Purple (Mixed Social Activities)*
- *Blue (Specific uses)*
- *Cyan (Pubs and cafes)*
- *Green (Sports and Parks)*
- *Orange (Restaurants and Bars)*

Next the average house rental pricing are visualized on the map, by using Folio chloropleth map and clusters on the top. Every marker in the map has also its cluster & borough name:



I came to the result that the rental house prices in the downtown with Hotels and Social venues nearby are very high you can clearly visualize in the map above while in the suburbs and the neighborhoods away from the city center have low prices but the facilities are also good. Most of low price neighborhoods are close to restaurants, pubs, sports facilities etc. Some Boroughs such as Westminster & Kensington and Chelsea have very high house rental prices. Ealing, Croydon, and Sutton Boroughs have very low house prices but they have also good venues to visit nearby.

Discussion

The first thing that comes to my mind is that for typical ways of scraping, cleaning, handling, transforming and visualizing data, all the tools are simply there. We just have to get to know the available open source packages, learn how to use them. What I find amazing is that mostly all of them are free and a simple notebook computer is enough. All the rest is concentrated, creative, interesting, sometimes hard work, patience and searching for tips, examples, explanations in the web, especially in forums. With these tools, many exciting data science use cases can be created, for all kinds of projects.

Conclusion

More and more people are turning to big cities to start a business or work. For this reason, people can achieve better outcomes and real estate investment decisions through their access to the platforms where such information is presented.

Not only for investors but also city managers can manage the city more regularly by using similar data analysis.

References & sources

A number of publications have inspired this project and helped me develop the skills to run this analysis and the coding. It is common to borrow a few fragments of code, as long as you fully understand them, change and apply them to your needs, and name the source - at least if we are talking about non-commercial use for qualification purposes. These sources are:

The courses of the IBM Data Science Professional Certificate itself and the plethora of hours I spent with them: https://www.coursera.org/professional-certificates/ibm-data-science

The examples for outstanding solutions for the capstone project mentioned on coursera and others I found in the web where also helpful: https://www.linkedin.com/pulse/housing-sales-prices-venues-data-analysis-ofistanbul-sercan-y%C4%B1ld%C4%B1z/, https://towardsdatascience.com/my-capstone-project-real-estate-prices-venues-data-analysis-of-london-c936c0bc4b1d, https://www.linkedin.com/pulse/applied-data-science-capstone-project-restaurant-wagner-mba/?articleId=6670274875946622976 and more.

Of course,a number of cheat sheets I found on github, kaggle, stack overflow etc.

The notebook is available to the public on my github repository. Since this is my first full length data science project, please be kind. Feel free to contact me if you have any questions, corrections or comments.