Capstone Project - Final Report

The business problem and the background of this project

The aim of this project is to find the best location in Central London to open a new Real Estate Office. London is a big and vibrant city, where lots of people dream to live in. Property in London could be very expensive, but there are many affluent people from all over the world looking to buy property either for personal use or as an investment. The real estate market moves fast and there are a number of real estate agencies, but we are planning to open an exclusive luxurious agency and we will try to find the best location for it based on property prices in the neighbourhood and the number of real estate agencies as we don't want an impression that our agency is just one of many.

Data and how it is going to be used to solve the problem

The UK government openly shares Price Paid Data and it includes information on all property sales in England and Wales that are sold for value and are lodged for registration. This data could be downloaded from HM Land Registry Price Paid Data. Data contains HM Land Registry data © Crown copyright and database right 2021. This data is licensed under the Open Government Licence v3.0. I am going to use property sales in 2020 to segment neighbourhoods in Central London (defined by postcodes starting with EC or WC) based on property price to choose the most luxurious areas.

FreeMapTools provides a list of UK post codes and their geographical coordinates latitude and longitude. We are going to use them to plot neighbourhoods with sold properties on a map. This data could be found at Download UK Postcodes with Latitude and Longitude.

Once I had the coordinates of each neighbourhood in Central London with some properties sold, I used a 1km radius from the centre point of each neighbourhood, defined by a postcode, to locate all Real Estate Offices nearby using FourSquare data.

I then counted the number of real estate offices in each neighbourhood and average property price in each neighbourhood. A k-cluster analysis was done following that to find the best areas for the new real estate office targeting premium property sellers and buyers.

Methodology

Data preparation and exploration

Property sales and geolocation data

We Download property sales data for England and Wales in 2020.

The data file contains a lot of information about a property and sale including price and location. We then create a subset of property sales data and name it *lon_price_subset* to simplify the dataset and select only what's needed:

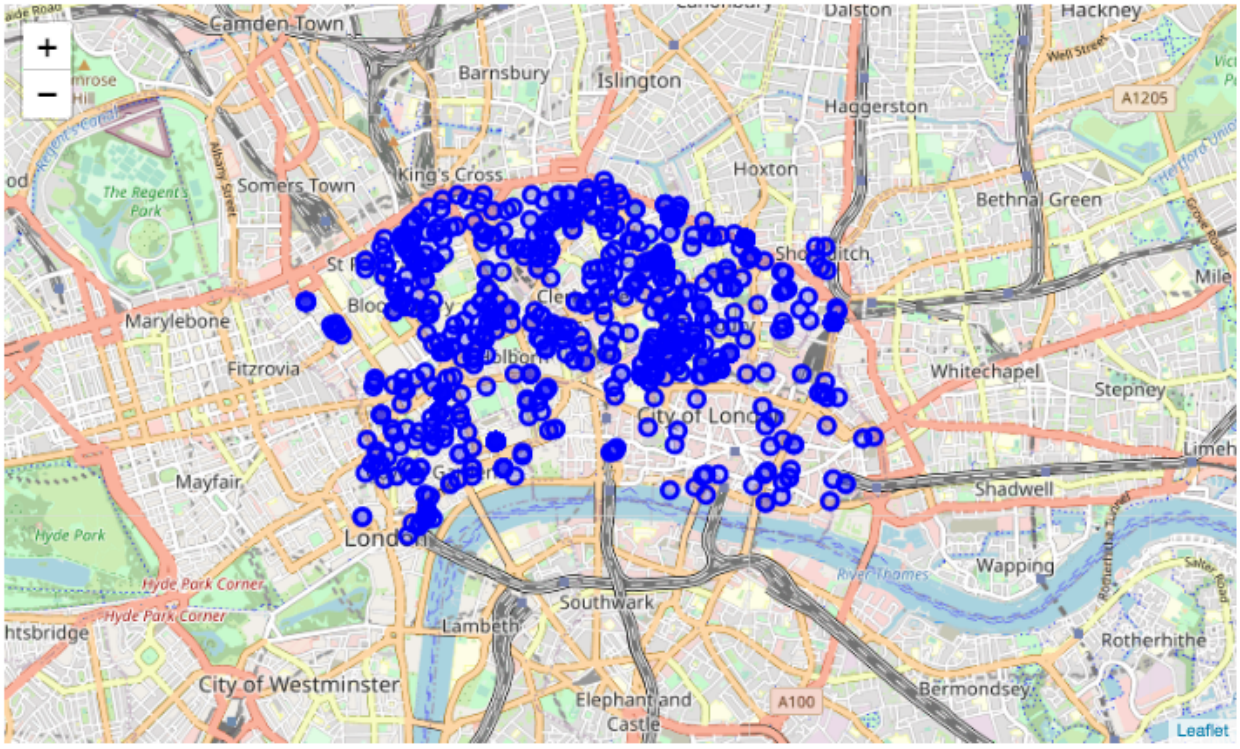
- County is Greater London and City is London to select only sales in London area
- Record Status is A as we are interested in new additions to the sales list only and not changes or deletions
- Postcodes starting with WC or EC to select Central London locations only
- Prices are greater than 5000 to filter out data that doesn't make sense pricewise.
- Select only the information we need for segmentation and clustering - **id, price, postcode, type and Borough**

We get latitude and longitude data for each postcode in the UK from FreeMapTools and merge the London property sales data subset with the geo data (latitude and longitude) to get coordinates for each postcode with property sales. We continue by getting some summary statistics of the newly created dataframe. It looks like all fields are fully populated and we have a good quality dataset to use for further analysis.

	id	price	postcode	type	Borough	latitude	longitude
count	1037	1.037000e+03	1037	1037	1037	1037.000000	1037.000000
unique	1037	NaN	481	4	5	NaN	NaN
top	{A2479555-145F-74C7-E053-6B04A8C0887D}	NaN	WC2A 2AT	F	ISLINGTON	NaN	NaN
freq	1	NaN	147	799	267	NaN	NaN
mean	NaN	3.523350e+06	NaN	NaN	NaN	51.521213	-0.105640
std	NaN	1.760576e+07	NaN	NaN	NaN	0.005871	0.014895
min	NaN	5.000000e+03	NaN	NaN	NaN	51.507431	-0.136397
25%	NaN	6.000000e+05	NaN	NaN	NaN	51.515733	-0.116228
50%	NaN	9.300000e+05	NaN	NaN	NaN	51.522031	-0.105987
75%	NaN	1.566330e+06	NaN	NaN	NaN	51.526403	-0.095988
max	NaN	3.034700e+08	NaN	NaN	NaN	51.531625	-0.074277

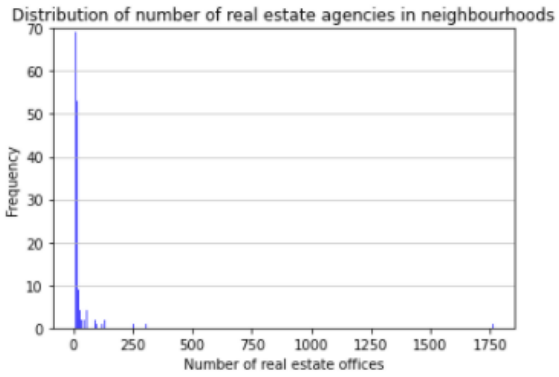
Visualisation of the sales data and getting real estate offices nearby

We then explored the neighborhoods with property sales in London. Create a map of London with sold properties superimposed on top. We then created a function to get all the venues in Real Estate Office category in London. Real Estate Office category is defined by categoryId=5032885091d4c4b30a586d66.



	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	EC1Y 0AD	51.523139	-0.095988	Felicity J Lord estate agents Clerkenwell	51.524567	-0.099469	Real Estate Office
1	EC1Y 0AD	51.523139	-0.095988	Felicity J Lord letting agents, Clerkenwell	51.524592	-0.099468	Real Estate Office
2	EC1Y 0AD	51.523139	-0.095988	Amdas Management	51.520367	-0.096080	Real Estate Office
3	EC1Y 0AD	51.523139	-0.095988	Foxtons Clerkenwell Estate Agents	51.522775	-0.100654	Real Estate Office
4	EC1Y 0AD	51.523139	-0.095988	Urban Spaces	51.522765	-0.101509	Real Estate Office

We then create a dataframe with the number of real estate offices near each neighbourhood and look at the distribution of the number of venues close by. We also calculate the median price for each neighbourhood as some neighbourhoods had a high number of sales and look at the distribution of prices.



We can see that both the number of real estate offices dataset and median prices are very skewed. We are going to split both price and number of offices variables into quantiles and use them for modelling to avoid those outliers having too much impact. We then merge the price data in each neighbourhood with geolocation data and the number of real estate offices nearby.

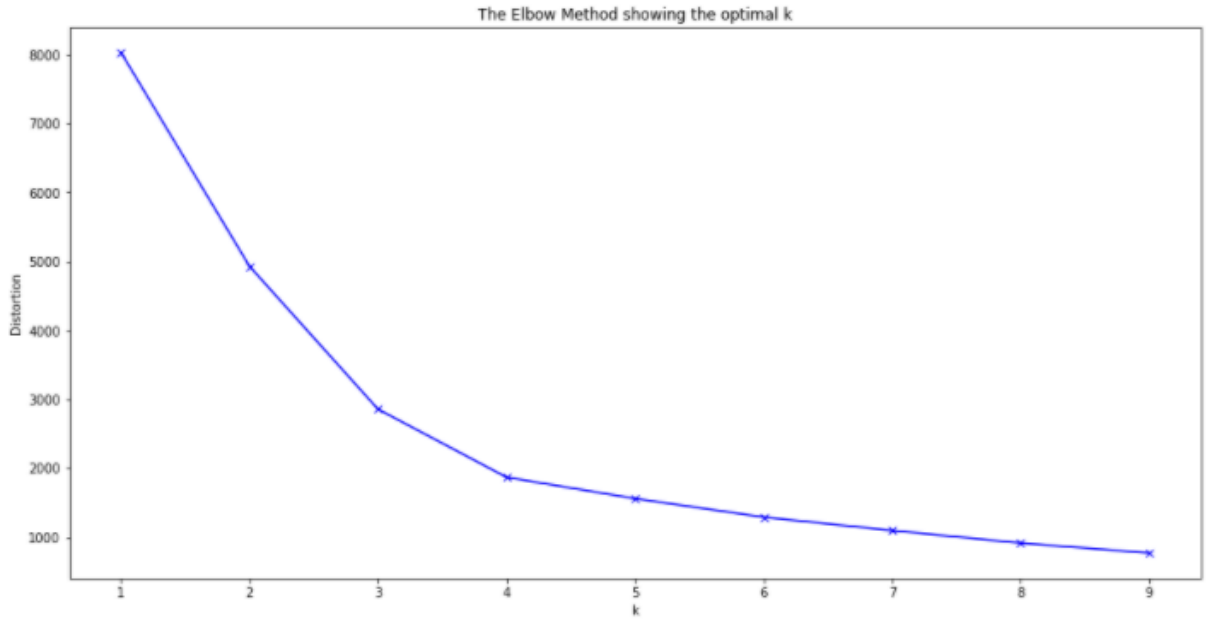
	Neighbourhood	Borough	price	latitude	longitude	price quantiles range	price quantiles	Venue	agencies quantiles range	agencies quantiles
0	EC1A 2DJ	CITY OF LONDON	14900000.0	51.516995	-0.102727	(7650000.0, 303470000.0]	9	16	(15.0, 17.0]	6
1	EC1A 4HU	CITY OF LONDON	746000.0	51.518236	-0.097184	(701112.0, 830000.0]	4	18	(17.0, 20.0]	7
2	EC1A 7AB	CITY OF LONDON	649950.0	51.519512	-0.098774	(575000.0, 701112.0]	3	17	(15.0, 17.0]	6
3	EC1A 7BB	CITY OF LONDON	1465000.0	51.517906	-0.098956	(1320000.0, 1990000.0]	7	36	(20.0, 36.0]	8
4	EC1A 7BD	CITY OF LONDON	1500000.0	51.518864	-0.098559	(1320000.0, 1990000.0]	7	17	(15.0, 17.0]	6
...	...	...	...	...	...	...	...	...	...	...
476	WC2R 1HA	CITY OF WESTMINSTER	1569500.0	51.512084	-0.118208	(1320000.0, 1990000.0]	7	18	(17.0, 20.0]	7
477	WC2R 1JA	CITY OF WESTMINSTER	1100000.0	51.511471	-0.118233	(985000.0, 1320000.0]	6	9	(7.0, 9.0]	1
478	WC2R 3DX	CITY OF WESTMINSTER	1195000.0	51.511505	-0.113562	(985000.0, 1320000.0]	6	10	(9.0, 11.0]	2
479	WC2R 3JF	CITY OF WESTMINSTER	1500000.0	51.512940	-0.112840	(1320000.0, 1990000.0]	7	10	(9.0, 11.0]	2
480	WC2R 3JJ	CITY OF WESTMINSTER	1500000.0	51.513140	-0.112840	(1320000.0, 1990000.0]	7	10	(9.0, 11.0]	2

481 rows × 10 columns

Model Building - KMeans

As there might be some similar neighbourhoods but we want one area belonging to on cluster only, we will be using KMeans Clustering Machine learning algorithm to cluster similar neighbourhoods together and then choose the best one for opening a new real estate office as discussed in the introduction section.

We use both KMeans elbows method and Silhouette score method to select the number of clusters. Both showed that the optimum is to choose 4 clusters. So, we created the final model and grouped neighbourhoods in 4 clusters.



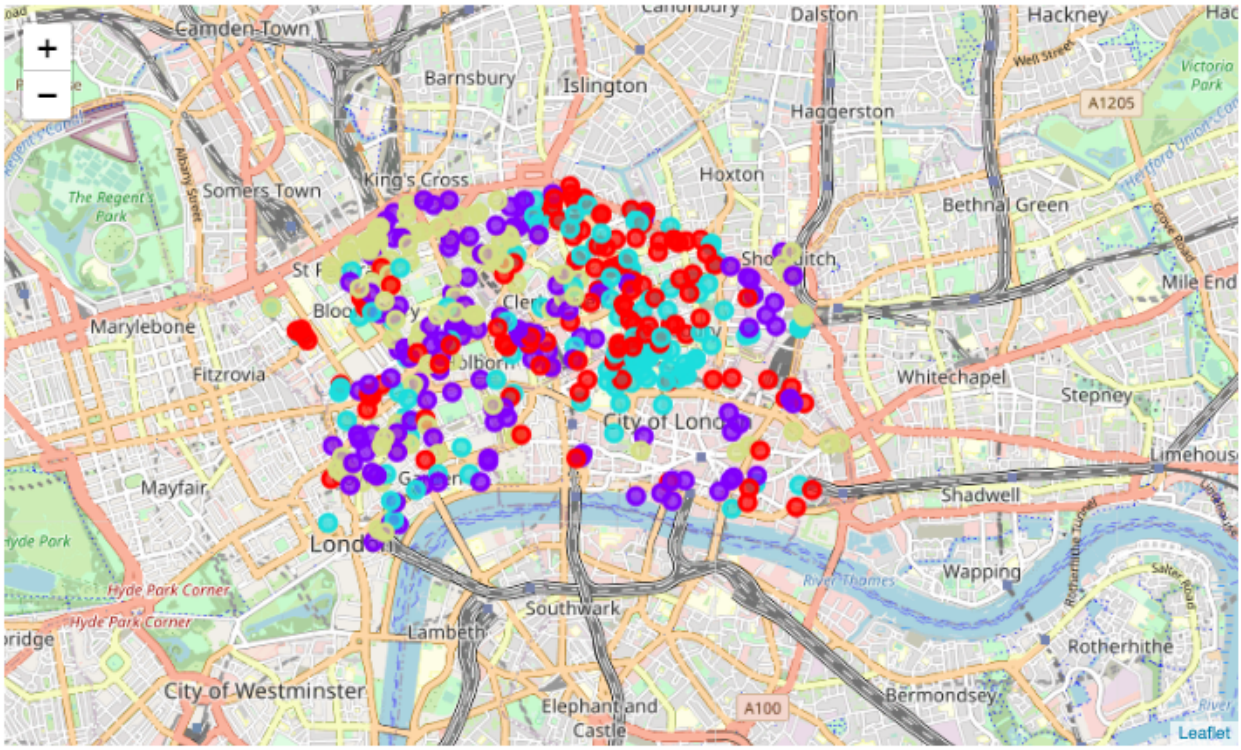

```
Silhouette Score for k= 2 is:  0.3664219055418742
Silhouette Score for k= 3 is:  0.4170669012274279
Silhouette Score for k= 4 is:  0.43639664955339147
Silhouette Score for k= 5 is:  0.4081160096499596
Silhouette Score for k= 6 is:  0.3851527934642992
Silhouette Score for k= 7 is:  0.4104884483411236
Silhouette Score for k= 8 is:  0.39430032515451724
Silhouette Score for k= 9 is:  0.3998430070285899
Silhouette Score for k= 10 is:  0.41133931447238786
```

A sample of the resulting dataframe with clustered neighbourhoods is provided below.

	Cluster Labels	Neighbourhood	Borough	price	latitude	longitude	price quantiles range	price quantiles	Venue	agencies quantiles range	agencies quantiles
0	2	EC1A 2DJ	CITY OF LONDON	14900000.0	51.516995	-0.102727	(7650000.0, 303470000.0]	9	16	(15.0, 17.0]	6
1	2	EC1A 4HU	CITY OF LONDON	746000.0	51.518236	-0.097184	(701112.0, 830000.0]	4	18	(17.0, 20.0]	7
2	0	EC1A 7AB	CITY OF LONDON	649950.0	51.519512	-0.098774	(575000.0, 701112.0]	3	17	(15.0, 17.0]	6
3	2	EC1A 7BB	CITY OF LONDON	1465000.0	51.517906	-0.098956	(1320000.0, 1990000.0]	7	36	(20.0, 36.0]	8
4	2	EC1A 7BD	CITY OF LONDON	1500000.0	51.518864	-0.098559	(1320000.0, 1990000.0]	7	17	(15.0, 17.0]	6

Results

Let's visualize the resulting clusters.



We have 4 clusters that differentiate neighbourhoods by property prices and density of real estate offices nearby. More information about each cluster is provided below and also visualised in a map of Central London above.

Cluster 1

	Cluster Labels	Neighbourhood	Borough	price	Venue	price quantiles	agencies quantiles
2	0	EC1A 7AB	CITY OF LONDON	649950.0	17	3	6
8	0	EC1A 7ES	CITY OF LONDON	550000.0	18	2	7
10	0	EC1A 9JR	CITY OF LONDON	562500.0	32	2	8
11	0	EC1A 9LS	CITY OF LONDON	562125.0	32	2	8
12	0	EC1A 9PN	CITY OF LONDON	701112.5	34	3	8
...	...	...	...	...	...	...	...
403	0	WC1X 8SF	CAMDEN	255000.0	20	0	7
430	0	WC2B 6SR	CAMDEN	27500.0	22	0	8
436	0	WC2E 7NU	CITY OF WESTMINSTER	548000.0	18	2	7
446	0	WC2H 7BA	CITY OF WESTMINSTER	20000.0	13	0	4
451	0	WC2H 8DY	CAMDEN	570000.0	13	2	4

120 rows × 7 columns

Cluster 2

	Cluster Labels	Neighbourhood	Borough	price	Venue	price quantiles	agencies quantiles
14	1	EC1M 3JB	CAMDEN	3637000.0	13	8	4
16	1	EC1M 4AJ	ISLINGTON	17600000.0	15	9	5
17	1	EC1M 4DG	ISLINGTON	29600000.0	15	9	5
22	1	EC1M 6EJ	ISLINGTON	13900000.0	15	9	5
30	1	EC1N 8DH	CAMDEN	4050000.0	14	8	4
...	...	...	...	...	...	...	...
473	1	WC2R 0NR	CITY OF WESTMINSTER	3099000.0	11	8	2
477	1	WC2R 1JA	CITY OF WESTMINSTER	1100000.0	9	6	1
478	1	WC2R 3DX	CITY OF WESTMINSTER	1195000.0	10	6	2
479	1	WC2R 3JF	CITY OF WESTMINSTER	1500000.0	10	7	2
480	1	WC2R 3JJ	CITY OF WESTMINSTER	1500000.0	10	7	2

130 rows × 7 columns

Cluster 3

	Cluster Labels	Neighbourhood	Borough	price	Venue	price quantiles	agencies quantiles
0	2	EC1A 2DJ	CITY OF LONDON	14900000.0	16	9	6
1	2	EC1A 4HU	CITY OF LONDON	746000.0	18	4	7
3	2	EC1A 7BB	CITY OF LONDON	1465000.0	36	7	8
4	2	EC1A 7BD	CITY OF LONDON	1500000.0	17	7	6
5	2	EC1A 7BF	CITY OF LONDON	1655000.0	54	7	9
...	...	...	...	...	...	...	...
469	2	WC2N 6LU	CITY OF WESTMINSTER	2750000.0	27	8	8
472	2	WC2R 0JF	CITY OF WESTMINSTER	26410177.0	22	9	8
474	2	WC2R 0PP	CITY OF WESTMINSTER	1108750.0	20	6	7
475	2	WC2R 1AB	CITY OF WESTMINSTER	1712500.0	66	7	9
476	2	WC2R 1HA	CITY OF WESTMINSTER	1569500.0	18	7	7

120 rows × 7 columns

Cluster 4

	Cluster Labels	Neighbourhood	Borough	price	Venue	price quantiles	agencies quantiles
41	3	EC1R 0DY	ISLINGTON	750000.0	11	4	2
44	3	EC1R 0HA	ISLINGTON	775000.0	10	4	2
45	3	EC1R 0HR	ISLINGTON	10267.0	11	0	2
46	3	EC1R 0JQ	ISLINGTON	800170.0	8	4	1
51	3	EC1R 1XJ	ISLINGTON	655000.0	12	3	3
...	...	...	...	...	...	...	...
462	3	WC2N 5HS	CITY OF WESTMINSTER	59175.0	8	0	1
463	3	WC2N 6AR	CITY OF WESTMINSTER	580000.0	8	3	1
465	3	WC2N 6EG	CITY OF WESTMINSTER	905000.0	8	5	1
466	3	WC2N 6HA	CITY OF WESTMINSTER	790000.0	9	4	1
470	3	WC2N 6NG	CITY OF WESTMINSTER	60000.0	7	0	0

111 rows × 7 columns

We have successfully clustered Central London neighbourhoods by property prices and density of real estate agencies:

- Cluster 1: Lower prices, high number of agencies
- Cluster 2: High prices, low number of agencies
- Cluster 3: High prices, high number of agencies
- Cluster 4: Low prices, low number of agencies

Discussion

We tried to find the best area in Central London to open a new luxury real estate agency. Central London is very expensive overall, but some neighbourhoods are more popular than others between real estate companies and property prices vary as well. In some neighbourhoods, we have more than 1000 real estate offices registered nearby and in some properties might cost tens and hundreds of millions. As such sales are very rare, it's better to concentrate on luxury class expensive property that are not one of a kind to ensure enough volume. We clustered all neighbourhoods based on property prices and number of real estate offices in close proximity using KMeans algorithm. Based on clustering results, the best areas to consider for the new office are in Cluster 2.

Conclusion

The purpose of this project was to find the best neighbourhood to open a new real estate agency. Even though there are lots of agencies already, there are places that are less dense and it is possible to find a niche to open an exclusive real estate agency and be successful. By combining various data sources and applying the K-Means algorithm, we managed to find the best areas for a new real estate agency. However, it's up to the stakeholders to decide what other criteria is also important to them. It might be also worth taking into account rental prices in the area and whether there are any other luxury class venues nearby before opening the agency.