Segmenting and Clustering Districts in the KT Postcode Area

Capstone Project – The Battle of Neighbourhoods

1. Introduction/Business Problem

London is one of the most expensive cities in the world to live in. With a population of roughly 9 million, London is made up of 33 boroughs. As one of the largest financial and culturally diverse cities in the world, more and more people are looking to locate into London. But due to high costs, people tend to buy properties on the suburbs of London.

This project will look to help property buyers identify the best area to buy a home in the KT postcode district of London also known as the Kingston Upon Thames postcode area. It is made up of 24 postcode districts within 19 post towns and covers part of southwestern Greater London and northern Surrey.

The investigation will identify the following:

- A comparison of property prices per bedroom by KT districts
    - And thus, allow a comparison of multiple districts with similar prices/bedroom by location
- A look at the most common venues in each borough

2. Data

Data on property prices were web scrapped from www.rightmove.co.uk which included the address, price, geo location (Lat and Long), number of bedrooms and bathrooms and type of property. This is shown in the following table:

	ID	Bedrooms	Bathrooms	PropertySubType	Latitude	Longitude	Amount
0	89682484	3	2	End of Terrace	51.341995	-0.28984	485000
1	102490706	4	2	Semi-Detached	51.395417	-0.28557	1000000
2	72967107	2	2	Penthouse	51.363893	-0.36371	1000000
3	103308818	4	0	Detached	51.3396	-0.51696	1000000
5	77571384	5	3	Semi-Detached	51.376815	-0.26222	1000000

Other listing features where also included in the data but were removed during the data cleaning stage as they were identified as unnecessary. The listings for the KT district resulted in 948 results after removing duplicates and missing values. As there was sufficient data it was decided not to replace the missing values as this may skew the mean of the results.

Using the Latitude and Longitude for each listing, the postcode for each property was gathered using Nominatim. This list was then compared against the list of KT postcodes which exist in the Kingston Upon Thames borough which was web scraped from Wikipedia. Property listings that were outside of the KT range where removed.

Using Foursquare data, a list of top 10 most common venues was created against each borough which included venues such as café, supermarkets and pubs. A K-means cluster algorithm was then used to group these boroughs into clusters based on similar venues.

3. Methodology

3.1 Cleaning Data

As the data was scraped from online, it was too raw to use immediately. Therefore, we had to prepare it before any data analysis could be performed. This process involved cleaning the data; either by removing missing values or imputing them, removing duplicates and results that may skew the overall mean. Then the process of feature extraction, where we extract only the information that is relevant to our data analysis. This would include removing unnecessary columns, extracting grouped data into their individual columns and potentially creating new columns based on existing data.

3.1.1 Duplicate Listing

The webpage's top results tend to be feature listings. These listings repeat on multiple pages and therefore it is necessary to remove these duplicates. This is achieved by identifying the that ID column should be unique. This process resulted in removing 19 duplicate results.

3.1.2 Auction properties

The data includes properties that are for auction. Usually listed auction prices do not represent the final sale value of the property and therefore all auction properties were removed. A total of 1 property was removed.

Non-buy properties

Non-buy properties, which are listings not for sale were also removed. No such properties existed in the latest dataset.

3.1.3 Missing values

Approximately 8% of property listing did not have the number of bathrooms listed. As this only accounts for a small amount of data, it was decided to remove these listings. A total of 80 properties were removed. However, an alternative could be to impute the results based on the mean number of bathrooms, or by comparing the number of bathrooms that other properties had based on other variables, such as the total price of the property or the number of bedrooms.

3.2 Feature Extraction

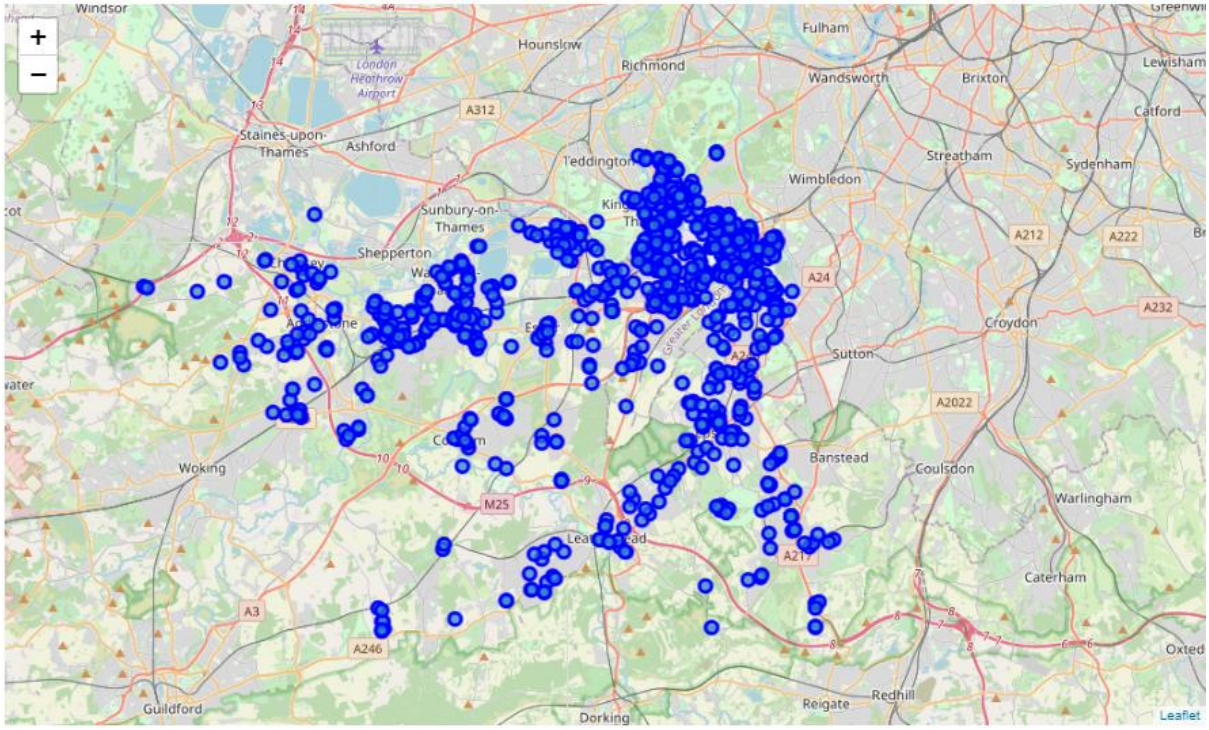
The data for each property listing included the Latitude, Longitude and sale price as dictionaries. These were extracted and placed into individual columns in the dataframe. The Lat and Long data were then used to acquire the postcode and city/town for each property using the Nominatim library. The postcode of each property was then checked against the list of KT postcodes that exist in the KT area.

The price per bedroom was created as a column which will allow us to make relative comparisons of different properties.

Furthermore, unnecessary columns were removed.

3.3 Processed Data – Property Listings Distribution

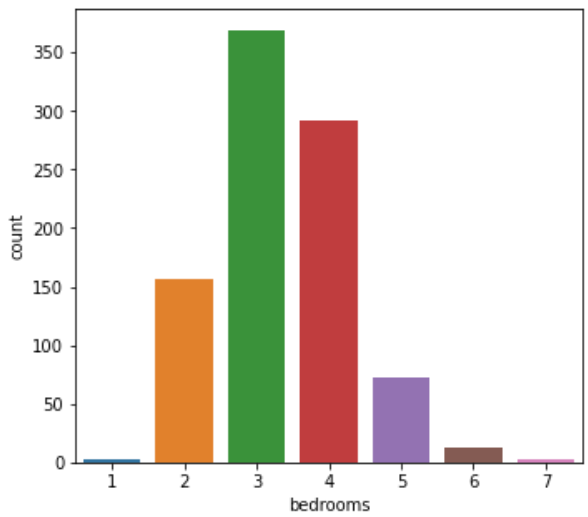
The following map shows the distribution of all property listings that will be used for analysis.

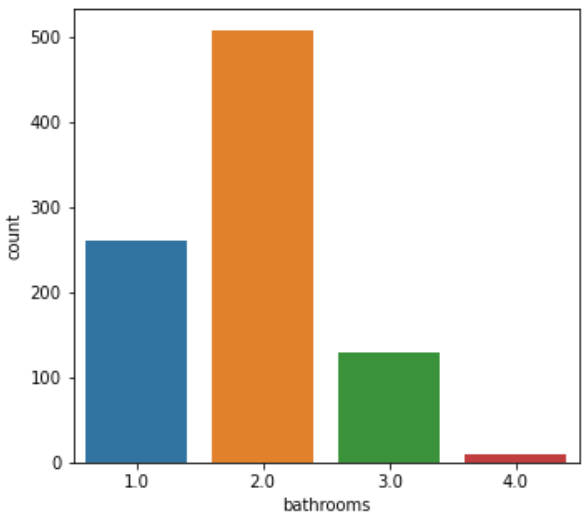


4. Results

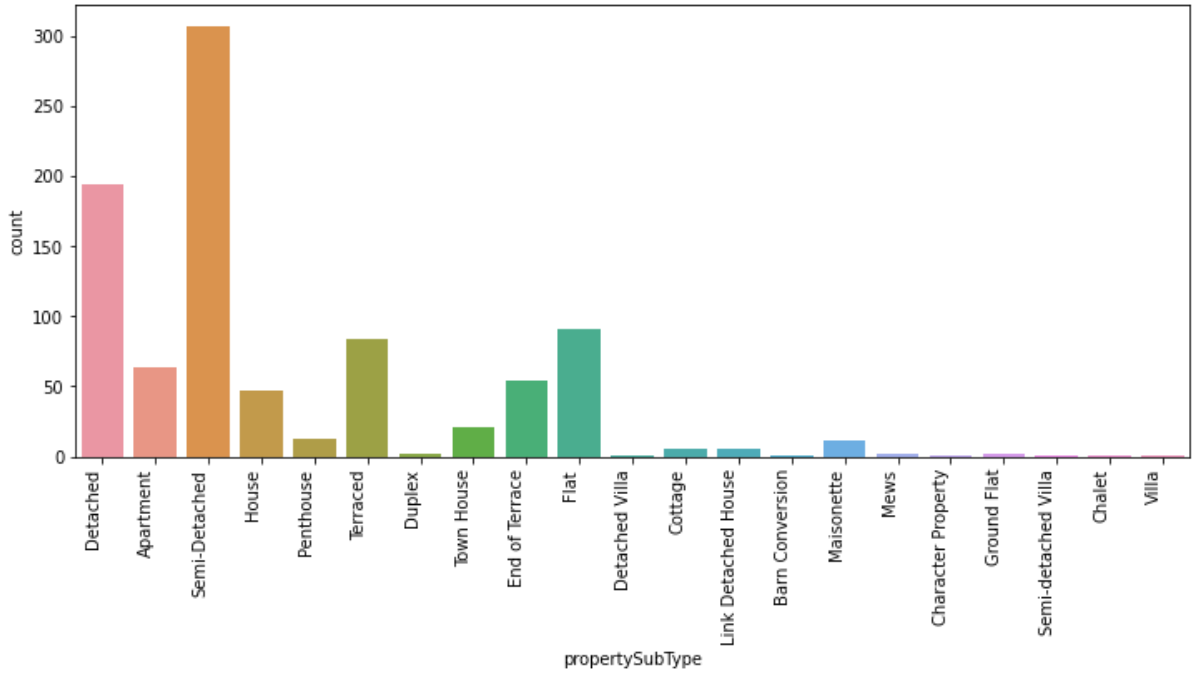
The following figures shows the distribution for the number of bedrooms and bathrooms for all the data that has been collected.

The results show that very few properties had less than 1 and more than 6 bedrooms. Furthermore, a majority of properties had 1-3 bathrooms.

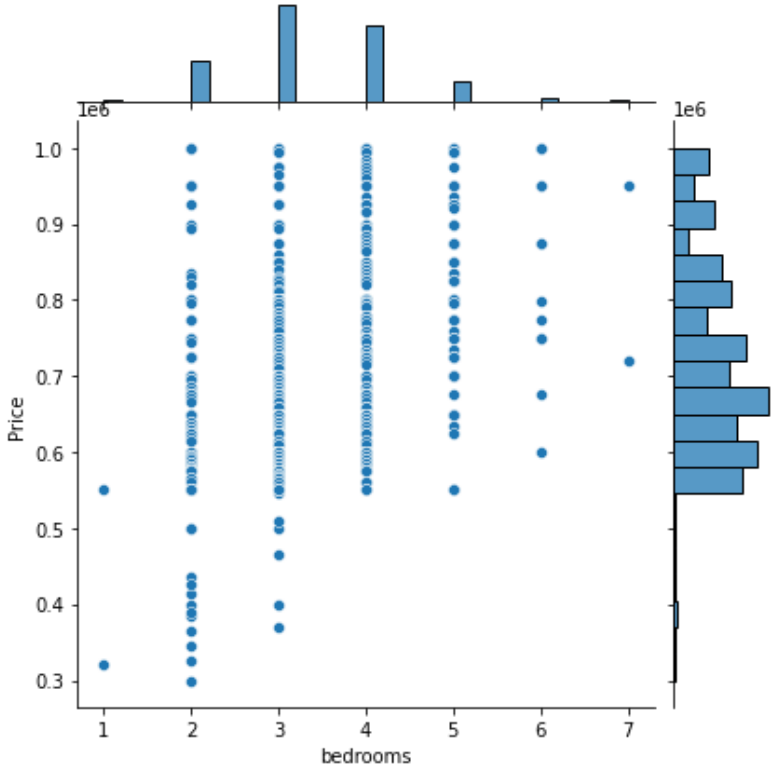




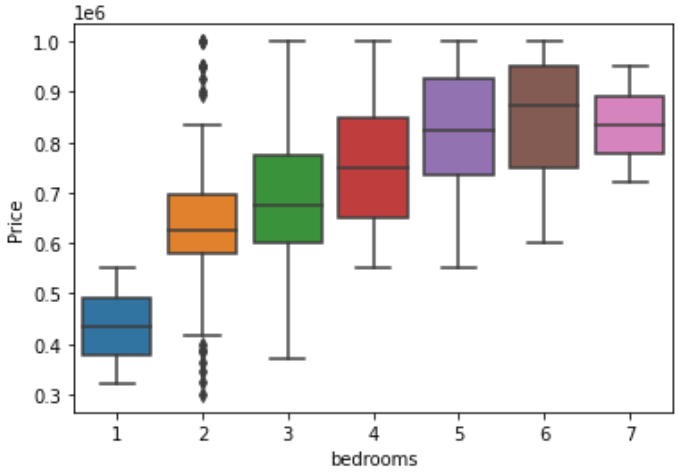
The following figure shows the distribution of the types of properties. A number of properties are of non-common property type. The unique property type of the property may have an effect on the overall price and may skew the results. Furthermore, there is minimal difference between a flat and Apartment. These properties could be merged. The "house" property type is also too generalised.



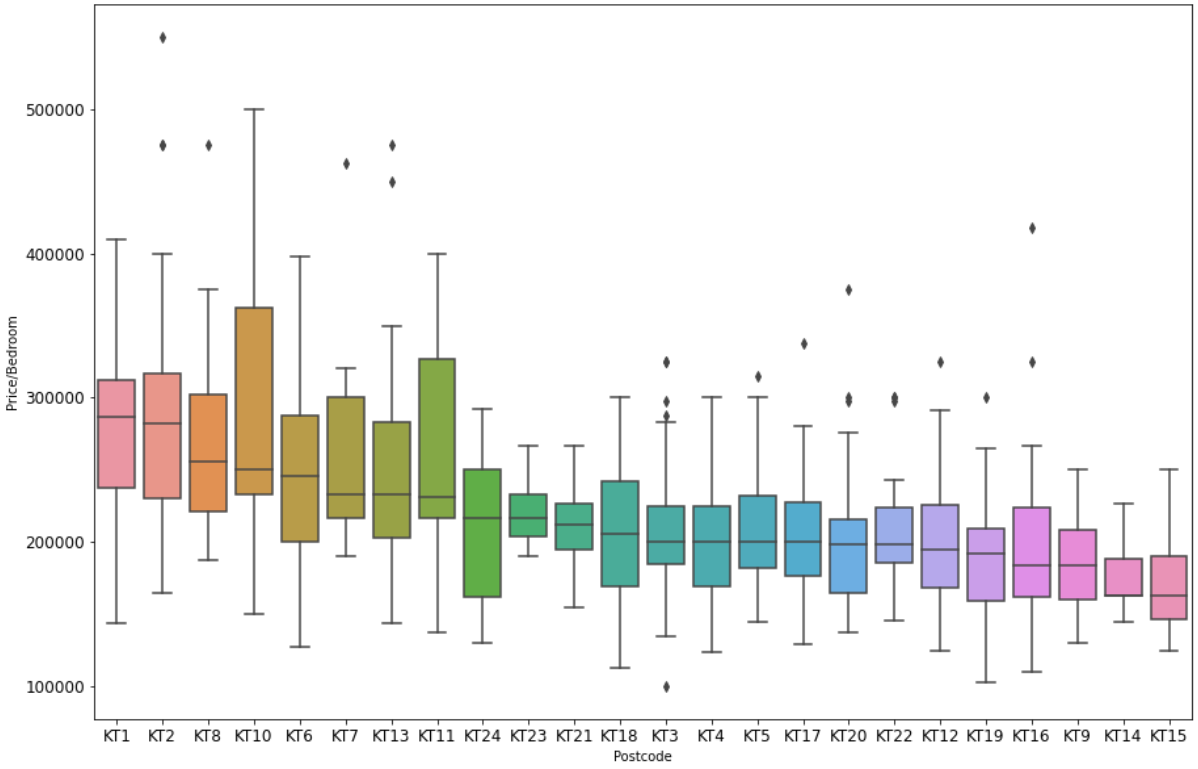
The following figure shows a majority of the house prices are between £550k and £750k.



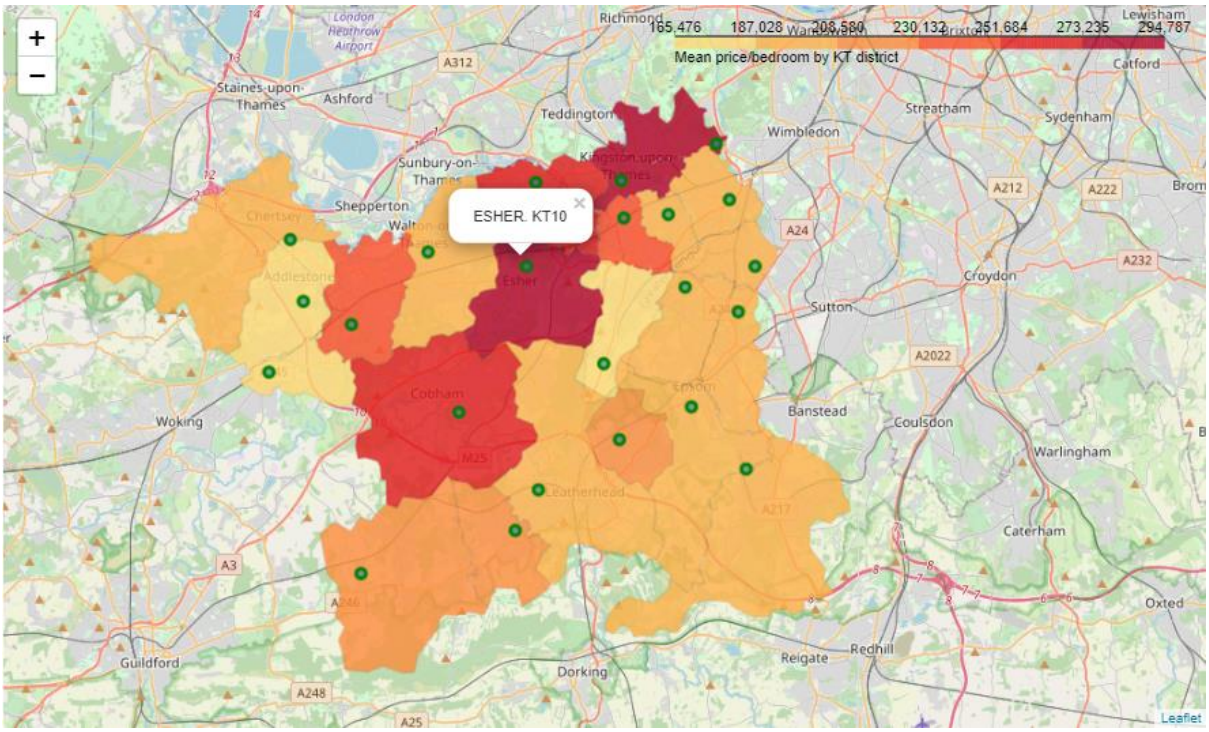
With increasing number of bedrooms, the overall price of the property increases. 2-bedroom properties had the greatest distribution in prices. Due to the small number of results for 6 and 7 bed properties, the prices may not be an accurate representation of the real world.



The following figure displays the average price/bedroom by postcode ordered by the median of the results. This data is better visualised in the map below.



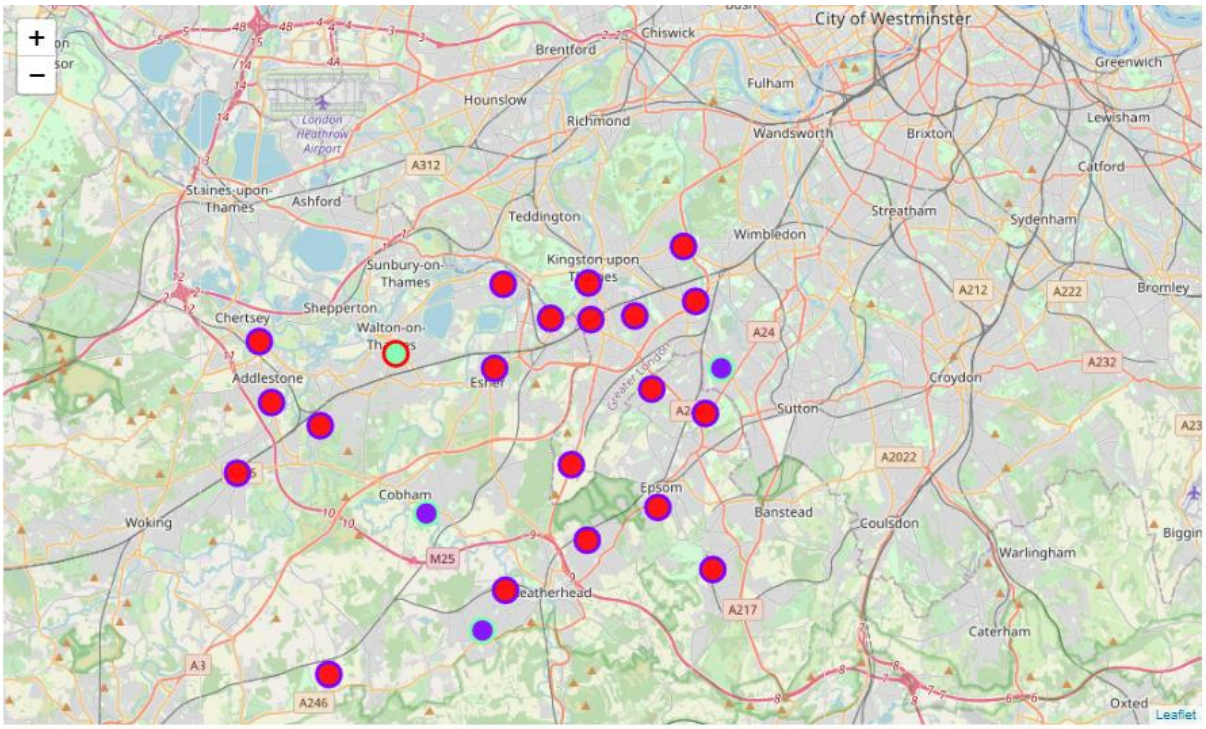
Visual representation of the average price/bedroom of properties by postcode area.



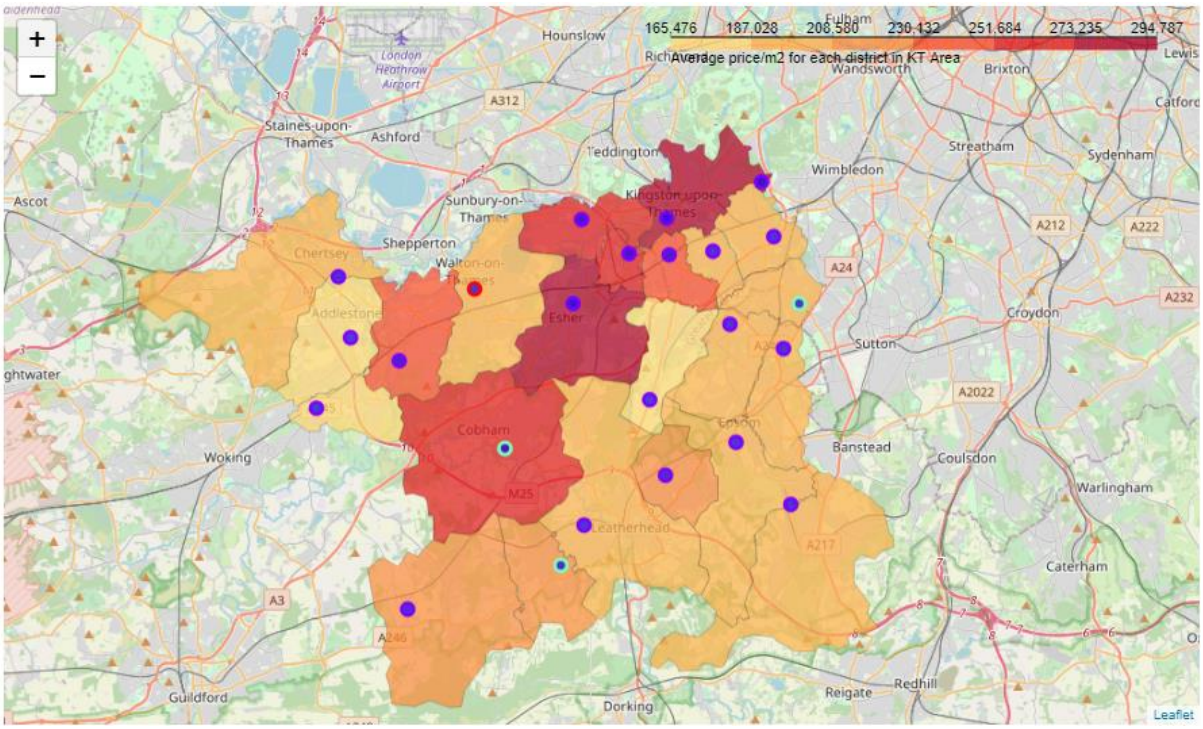
5. Discussion

Venue related information for each postcode district was collected from Foursquare. The top 10 popular venues were collated and K-means clustering was performed to cluster postcode areas together by similarity.

The following map is the result of the k-means clustering. The postcode areas were separated into 3 separate clusters. The map shows how a majority of the postcodes were similar to each other.



The map below allows us to see the clustering of postcodes by the average price/bedroom. The map shows postcodes with similar venue characteristics vary vastly in average price but also the fact that they are all located very close to each other.



6. Conclusion

From the above choropleth map, through the different shades of colour, a user can easily compare the average price/bedroom by district. For example, KT1 is more expensive than KT5 by approx. £70k, however both districts are located next to each other. Furthermore, postcodes of similar prices but geographically distanced can be compared together. The above map also shows how living closer to the city of London doesn't necessarily result in higher prices. However, it does seem that all of the higher priced postcodes are located next to each other in a line, represented by the darker red colour.