Segmenting and Clustering Districts in the KT Postcode Area

Capstone Project – The Battle of Neighbourhoods

Introduction

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

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

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.

Methodology

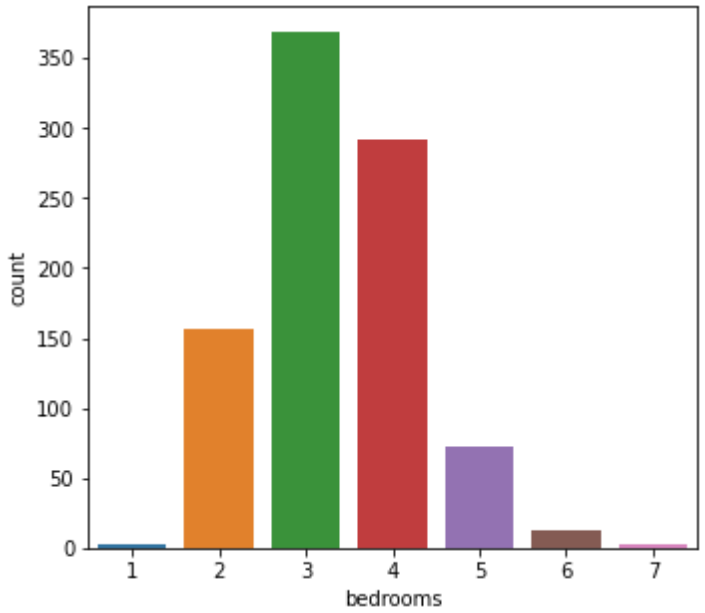
This stage involved cleaning the imported data and performing feature extraction so that in the end we have data to work with:

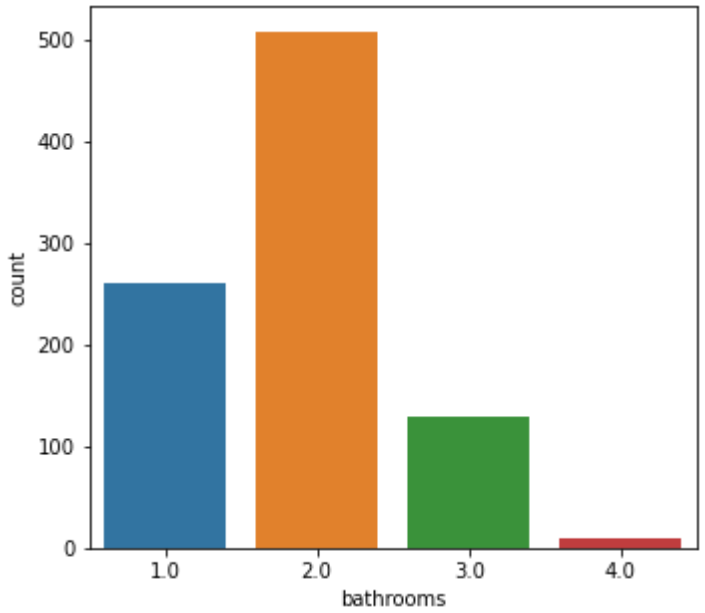
- Remove duplicates
- Remove missing values
- Remove unnecessary columns
- Drop results that contain outliers or results that would skew the data
- Extract grouped data and separate them into individual columns
- Create new features based on existing columns

Results

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

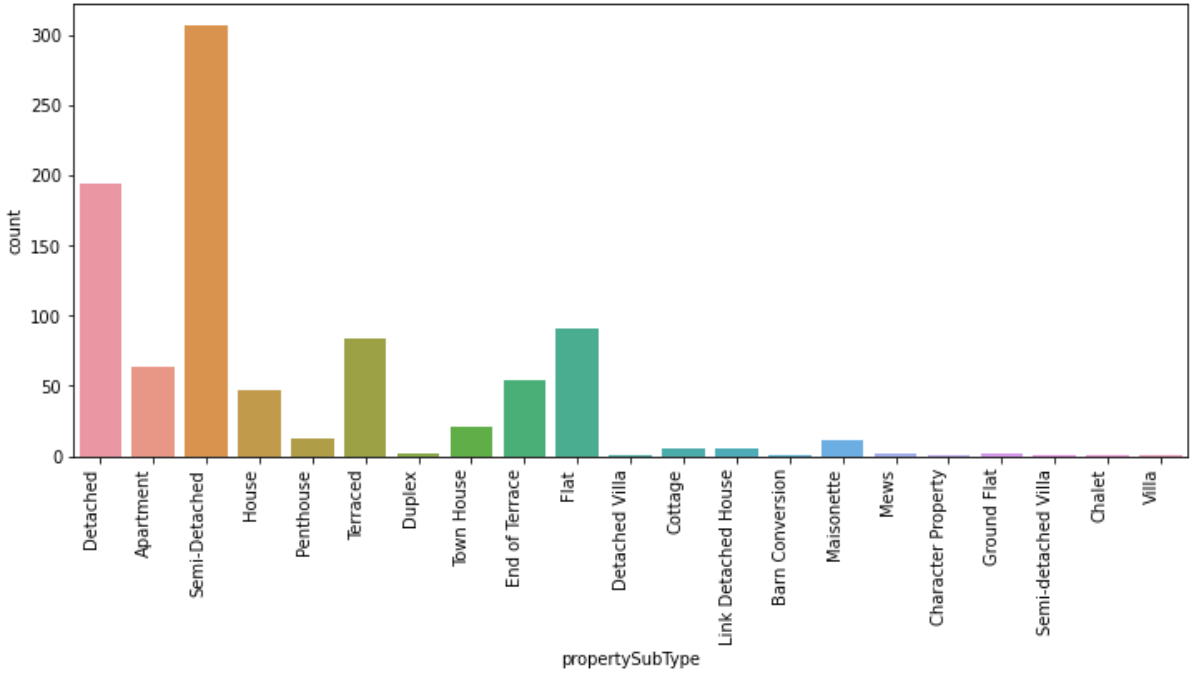
The results show that very few properties had less than 1 and more than 6 bedrooms. Furthermore, a majority of properties had 1-3 bathrooms.





Results

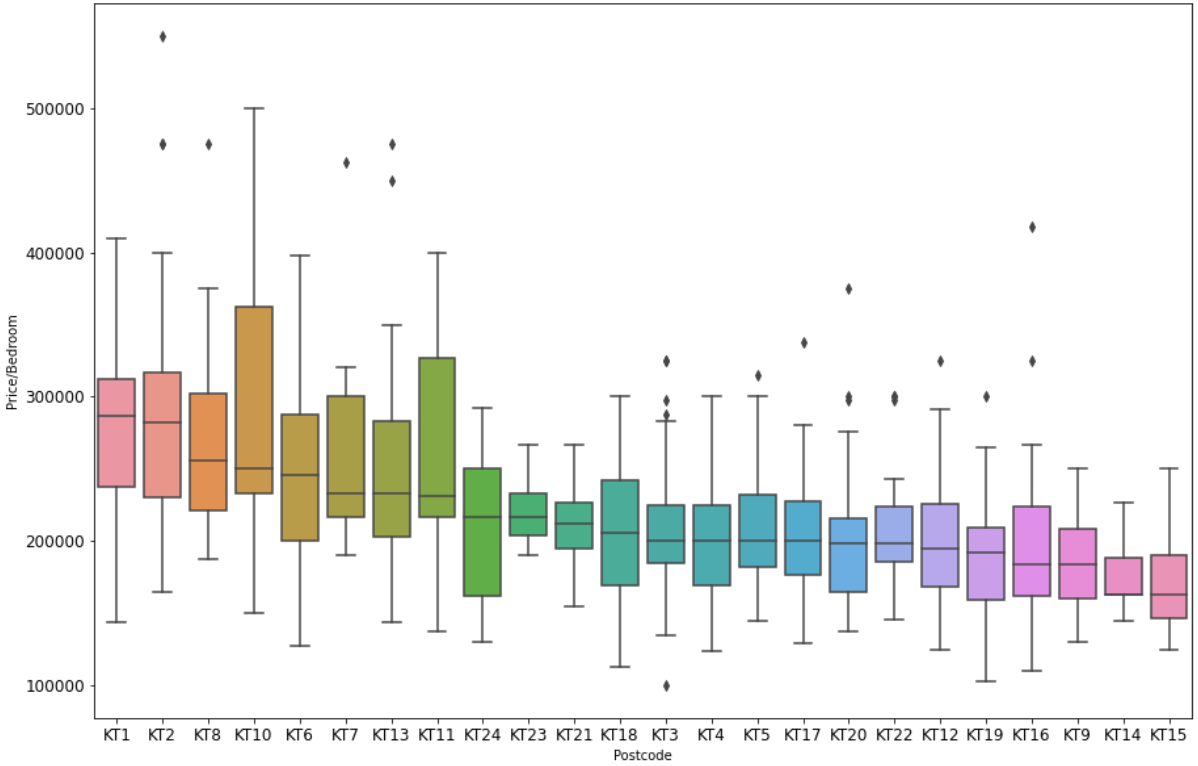
The following figure shows the distribution of the types of properties. A number of properties are of non-common property type. The unique property type of the property may have an effect on the overall price and may skew the results. Furthermore, there is minimal difference between a flat and Apartment.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

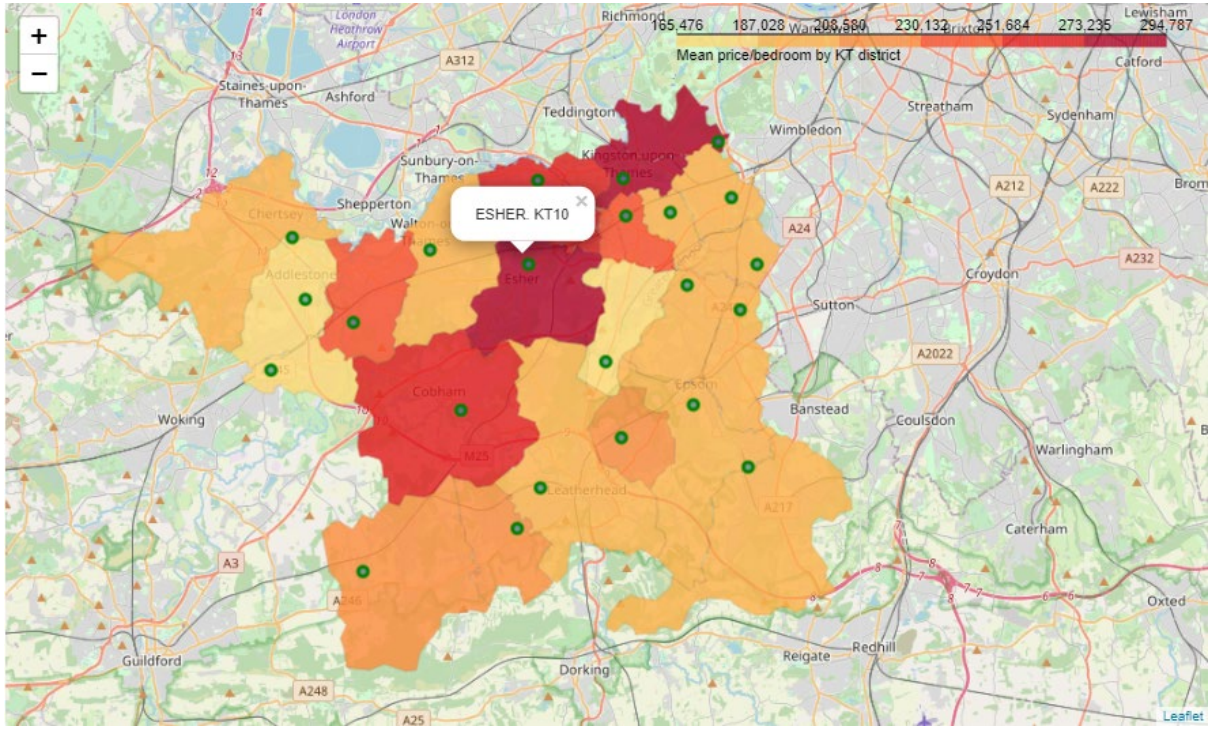


Results

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.

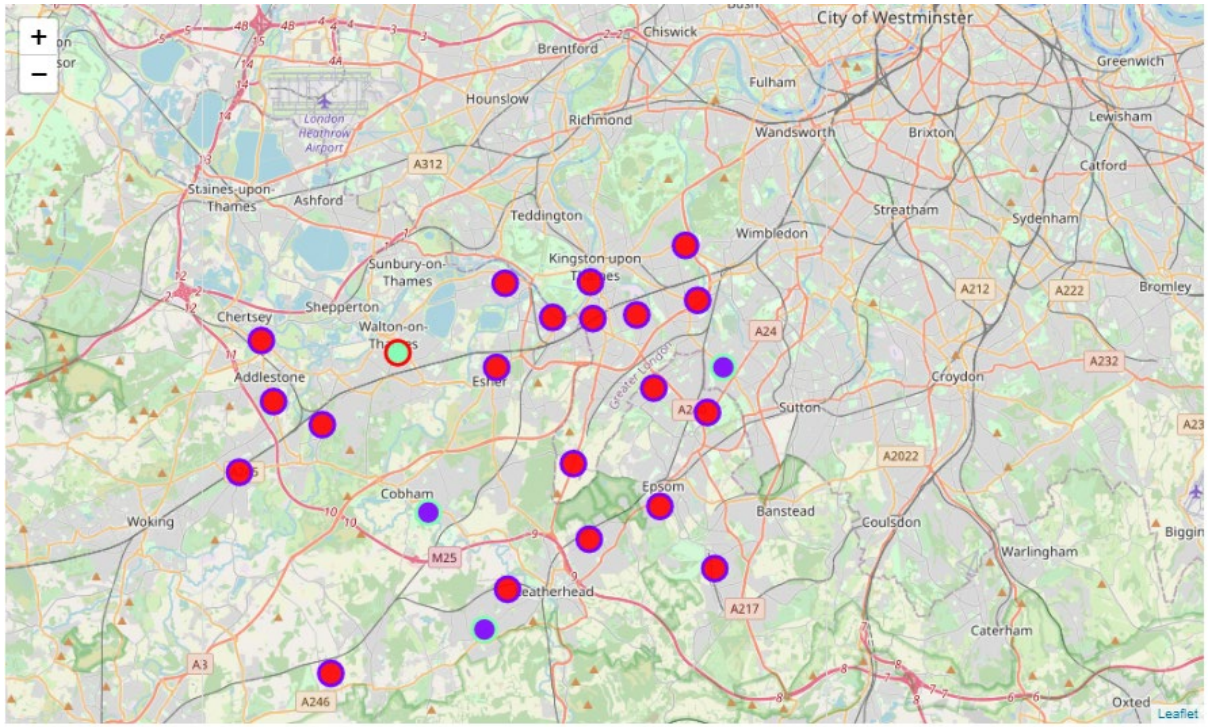




Results

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.



Conclusion

From the 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.