**Capstone Project**

# **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

# Data

Data on property prices were web scrapped from [www.rightmove.co.uk](http://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:



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.

# Methodology

* Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

# Results

* Results section where you discuss the results.

# Discussion

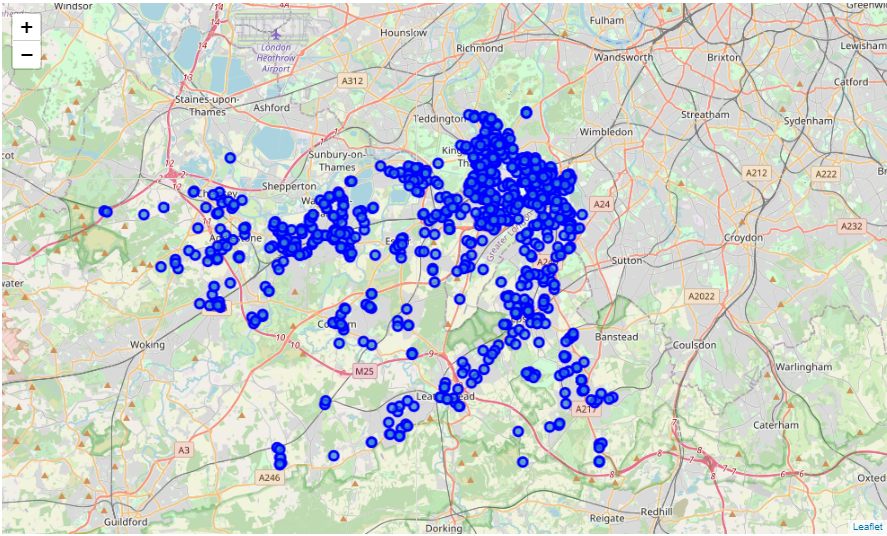
* Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

# Conclusion

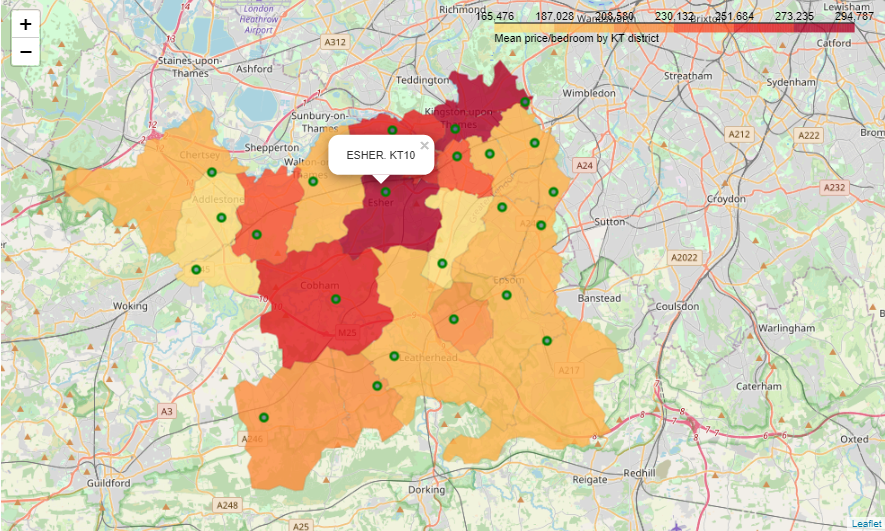
* Conclusion section where you conclude the report.

Graphs

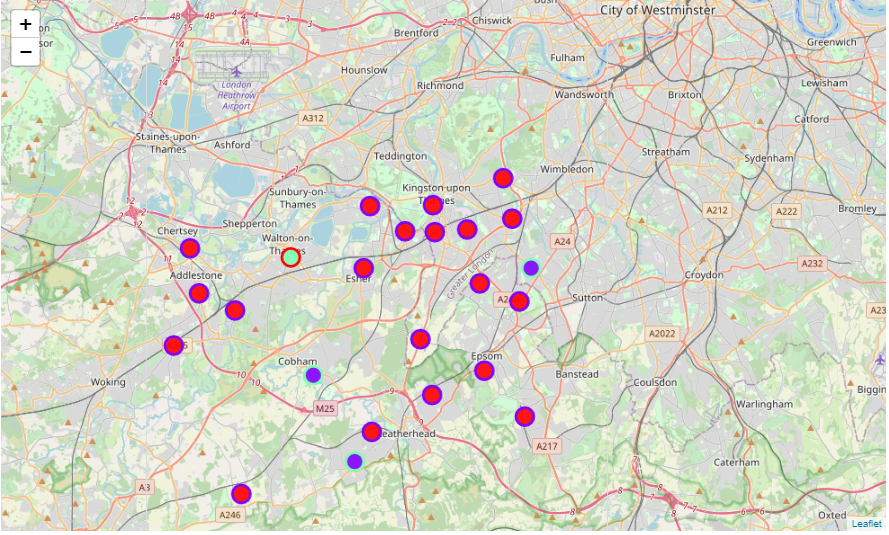
Out 43, map 1



Out 47, map 2



Out 79 mpa 3



Out 77 map 4

