Neighbourhood insights for buying a home in Edinburgh

Sander Tanni

12th of March 2019

Coursera Applied Data Science Capstone Project

Contents

Introduction

Business problem

A major estate agent in Edinburgh would like to provide more information about neighbourhoods to their clients. Local amenities are very important in deciding to buy a home, in addition to the property itself. This information, however, is not provided by the estate agent to the same quality as details about the property itself. Presenting valuable insights about local amenities to potential buyers could attract more customers, particularly those new to the city.

The insights about local amenities should be provided to the home buyer in a format that can be directly used in making their decision. Firstly, the information should be easy and quick to understand. Secondly, it should allow intuitive comparison between available properties. Thirdly, it should be objective truth, based on statistics, and not a biased opinion.

This project aims to provide a solution for informing home buyers about the neighbourhoods in Edinburgh. We will achieve this by creating a neighbourhood categorisation tool that can be tailored for the specific types of properties the user is looking for.

Neighbourhood categorisation

Neighbourhoods will be categorised using k-means clustering based on the local amenities and identifying preference categories in the resulting clusters.

Data

Overview

For this project we will need data on venues and amenities across Edinburgh and property sale price data. We acquired the data on venues and amenities using Foursquare API. The sale price data was acquired using web scraping on Rightmove website.

The neighbourhoods were defined by their centres that were uniformly distributed positions across Edinburgh (Figure 1). Data listing venues of different categories and residential property sale prices was collected and their distance to each of the neighbourhood centres was computed.

These neighbourhoods with the resulting venue densities and property prices were the subject of the statistical and machine learning methods.

Postcode data

We used data from website doogal.co.uk, that contained the latitude and longitude of each postcode.

Defining neighbourhoods

We defined uniformly distributed points across Edinburgh to serve as centres for neighbourhoods. These points were arranged in a rectangular grid with 250 m spacing in 4 km radius of Edinburgh Castle. We looked up the postcode closest to the centre of each neighbourhood from the postcode data.

These neighbourhood centre coordinates and their most proximal postcodes were then used to collect data specific to each location. This allowed analysis of the combined spatial distributions of venues and home prices.

Neighbourhood spacing was used as search radius for venue and property price data. This ensures the whole Edinburgh area of interest is fully covered, as the search areas will be overlapping (Figure 1). Any duplicates of venues and property prices were discarded.

The neighbourhoods were defined arbitrarily to allow describing properties of different areas in Edinburgh based purely on geometric separation and not municipalities or other human defined regions.

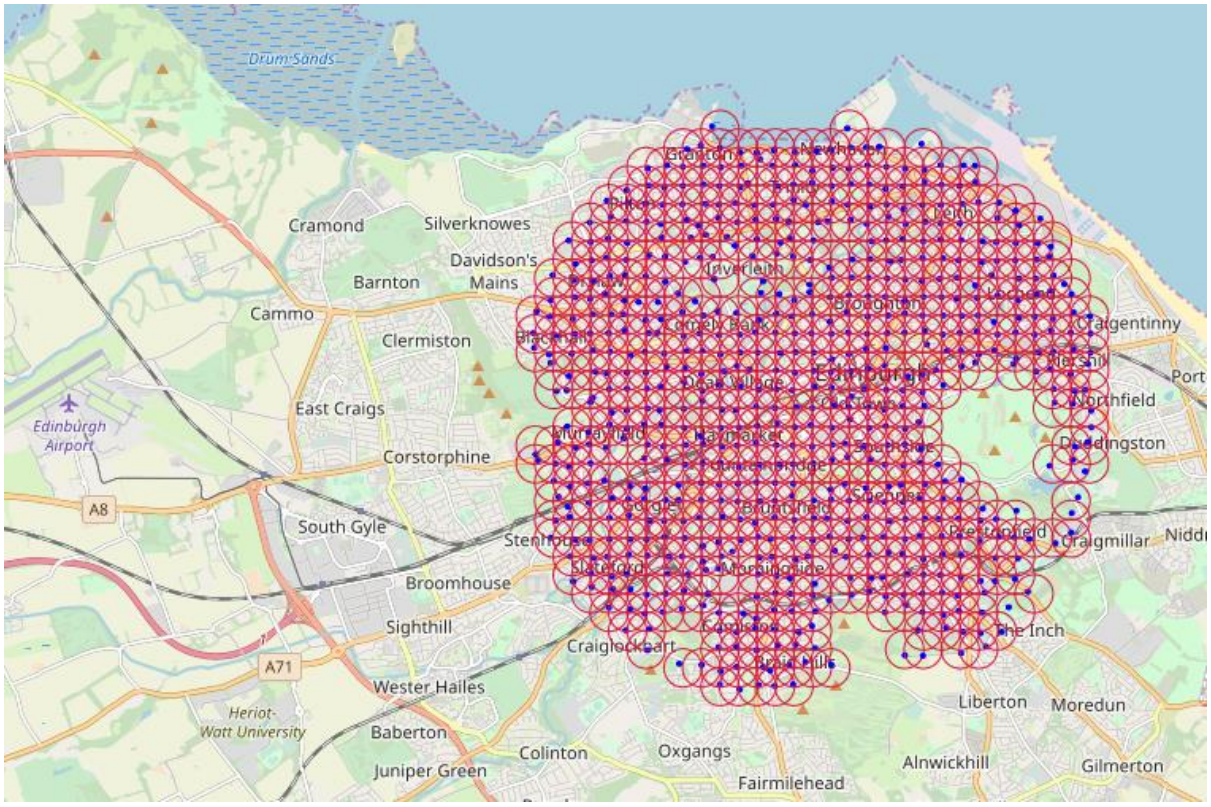

Figure 1 | Generated neighbourhoods marked with 250 m radius circles.

Foursquare venue data

We collected the identity and coordinates for all venues close to each neighbourhood centre for 9 different venue categories (Table 1). We used the coordinates of each venue to calculate its distance to all neighbourhood centres. These distances were used to find the venues nearby to each neighbourhood centre.

	total	per neighbourhood
Food	1659	2.50
Outdoors & Recreation	750	1.13
Nightlife Spot	710	1.07
Arts & Entertainment	658	0.99
Cafe	462	0.70
Food & Drink Shop	297	0.45
Bus Stop	290	0.44
Spiritual Center	144	0.22
Convenience Store	106	0.16

Table 1 | The total number and count per neighbourhood for each venue category.

Rightmove property sale price data

Rightmove is a major UK property website. They provide a list of sale price data going back several years.

For some of the properties on the list there is a link to a post on Rightmove website and information on the type of the property, including number of bedrooms. As the latter information is a major determinant of sale price, we only used data on properties where this information was available. This allowed more client preference specific estimation of mean property prices in neighbourhoods.

Rightmove provides the address for each property, including the postcode. We used Edinburgh postcode latitude and longitude dataset to approximate the latitude and longitude of each property. With these coordinates for each property we were able to calculate their distance to each of the neighbourhood centres. These distances were used to find the properties that were sold within nearby to each neighbourhood centre.

We found that by far the most commonly sold property type is 2 bedroom flat (Figure 2A). The next property types in decreasing order of sales are 1 bedroom flat, 3 bedroom flat, 3 bedroom house, 2 bedroom house and 4 bedroom house. Properties with more than 5 bedrooms are very rare in the dataset. The number of flats sold compared to houses is much higher for 1, 2 and 3 bedroom properties, while it is the other way round for properties with more bedrooms.

The mean price of flats and houses with same number of bedrooms is roughly the same (Figure 2B). This suggested that to maximise samples of sale prices in each neighbourhood, the prices of flats and houses could be combined. We used 2 bedroom properties as the example property type in this pilot study, as this is the most commonly sold property type.

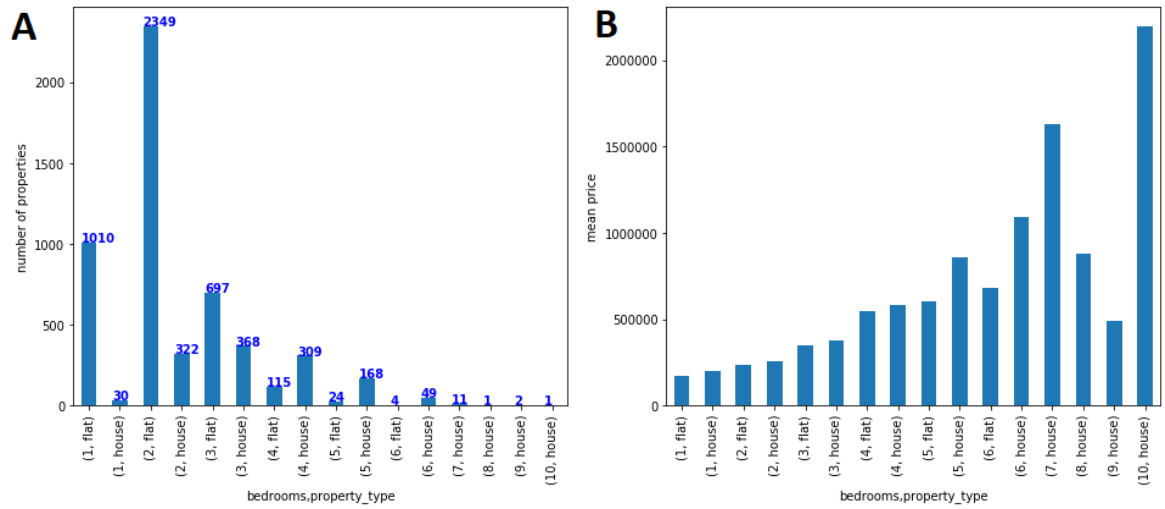

Figure 2 | A | Number of properties of different type in the dataset. B | Mean price of properties of different type.

Methods

To serve the aim of this project "to provide a solution for informing home buyers about the neighbourhoods in Edinburgh", we categorized each neighbourhood based on the number of nearby venues in each category and the local property price data.

Neighbourhood size

The Analysis section will cover data exploration was performed to decide on the radius of neighbourhoods to use. The size of the neighbourhoods needs to be large enough to be representative of local venue distributions, but not sacrifice too much spatial resolution.

Normalizing distributions

We used the log of the venue counts per neighbourhood, to make it more likely the clustering algorithm would separate neighbourhood categories also in cases of where the number of venues was low.

All data used for classification was normalized with z-score method. This was to ensure features with larger absolute values like number of Food places or property prices would not be the sole determinants of classification of ranking.

Clustering

We used k-means clustering as the numerical form of data permits using this simple and intuitive method. We experimented with different K values and made a subjective decision informed by statistical methods to pick the final K value.

Analysis

Neighbourhood size

To decide on the radius for finding venues and sold properties nearby to each neighbourhood, we investigated the distribution of venue counts across neighbourhood by category for different radii. In

all venue categories, with radius of 250 m, there was a very high proportion of neighbourhoods with zero venues. With 500 m radius the distributions were a bit broader and proportion of neighbourhoods with at least 1 venue was much larger, particularly for Outdoors & Recreation and Food & Drink Shop categories (Figure 3). With 750 m and 1000 m radius the distributions were broad in most categories, but at this level the spatial accuracy of the information would start to decay.

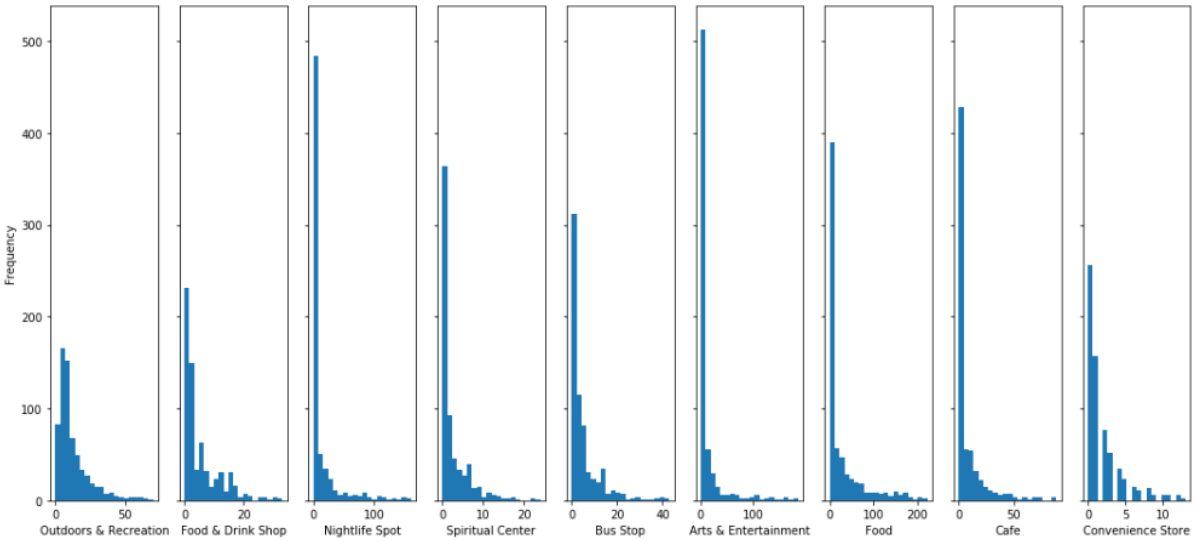

Figure 3 | Distribution of venue counts across neighbourhoods in different venue categories.

To assess the impact of the radius parameter on estimating property prices, we can inspect the number of neighbourhoods with insufficient data if the radius is set too small. We set a threshold for getting an acceptable estimate of property price for a specific criterion (location, number of bedrooms, flats or houses) to be at 5. We investigated in how many neighbourhoods it is possible to compute the mean price, given the threshold of 5 properties (Figure 4).

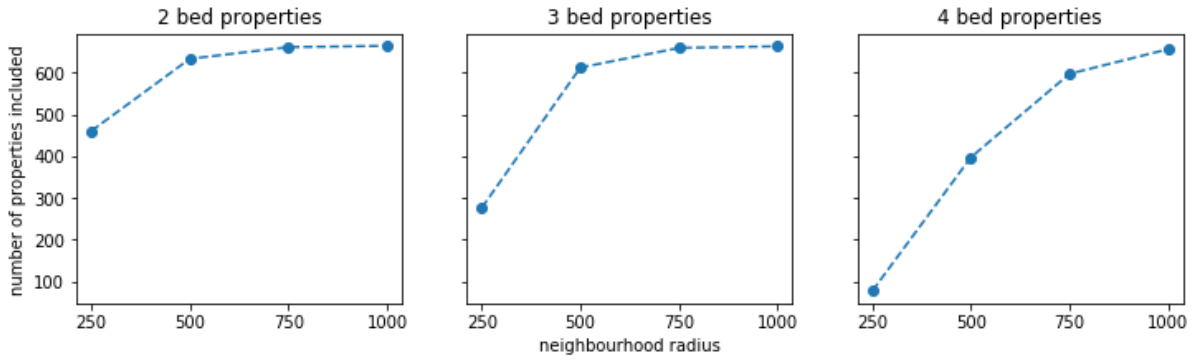

Figure 4 | The number of neighbourhoods that would be included in the analysis depending on property type given different values of neighbourhood radius.

With radius of 500, almost all neighbourhoods had enough properties for the mean price to be computed when 2 or 3 bed properties are considered. There are far fewer 4 bed properties sold each year and even with radius of 750 many of the neighbourhoods did not have enough data. From this we concluded that optimal values for radius are 500 and 750. We decided to use radius of 500 to maximise the spatial accuracy of our location data at the expense of sampling in some neighbourhoods.

Normalizing distributions

To make sure the clustering algorithm is not biased by any features unintentionally, we investigated the distribution of values for features across neighbourhoods. The distributions for all features but the Price were highly skewed towards low values (Figure 5). This would make clustering with k-means difficult, as the neighbourhoods with higher values in any category will be much more separated in the feature space. To account for this, we will use logarithmic values for all features apart from Price. The remaining issue was the large difference in absolute values in the features, particularly the Price category. This would result clustering being based mostly on Price. To avoid that, we used z-score method to normalize each feature (Figure 6).

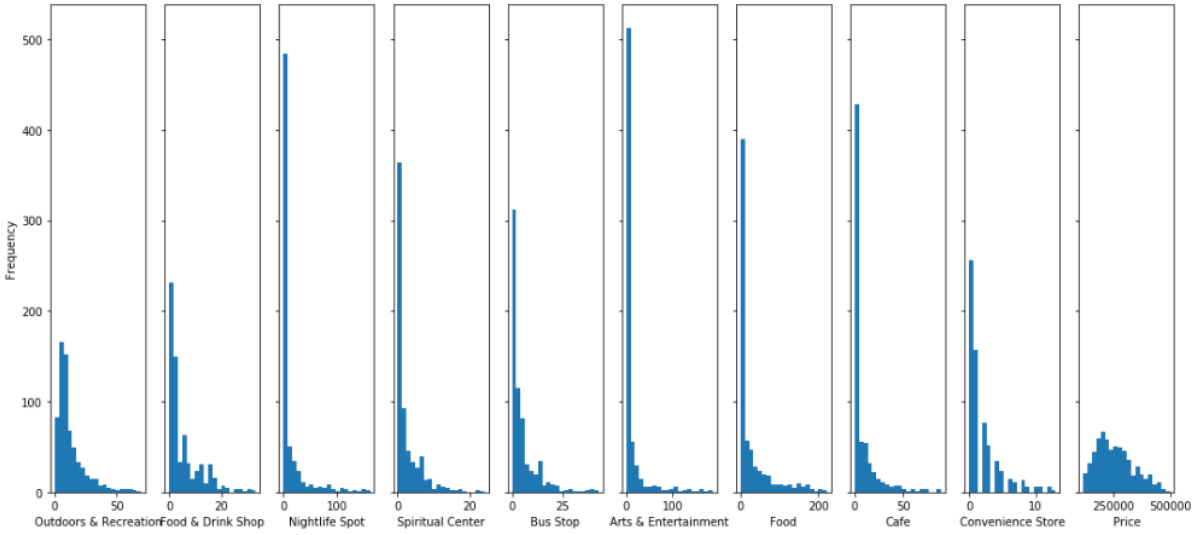

Figure 5 | Distribution of different features across neighbourhoods. The distributions are strongly skewed to low values for all venue categories. The absolute values also vary a lot, especially in the case of Price.

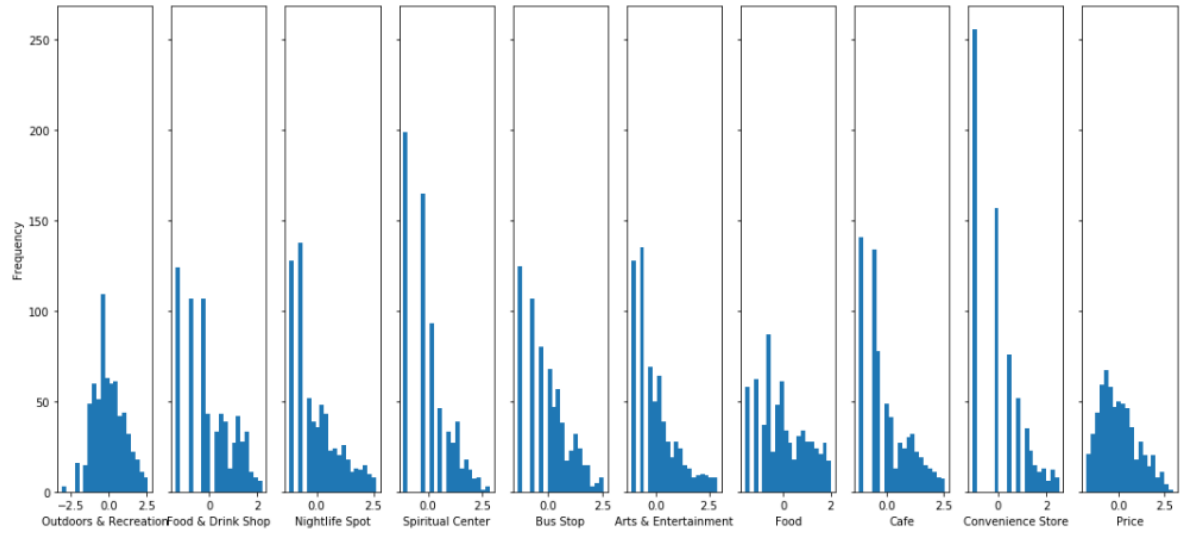

Figure 6 | Distribution of different features across neighbourhoods after normalizing.

Clustering

To decide on the k-value (n_clusters) to use for clustering, we used the Elbow Method with Inertia and Average Silhouette Method (Figure 7). The Inertia plot was not useful at all for deciding on the k-value,

as there was no clear “elbow”. On the other hand, the Silhouette Score plot showed a peak at k-value of 5, suggesting it would be a good pick.

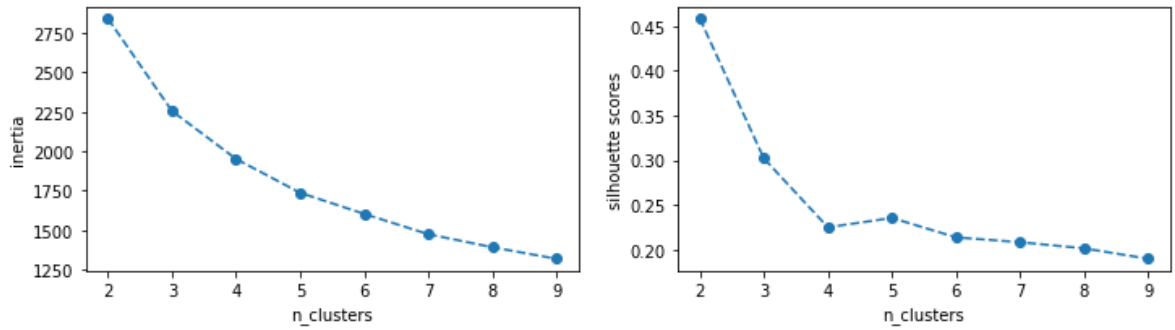

Figure 7 |Inertia (left) and average Silhouette score (right) for different k-values (n_clusters).

We also investigating the clustered neighbourhoods on a map along with the median feature values using results obtained with different k-values. Indeed, k-value of 5 yielded good segmentation of Edinburgh areas on the map and the median features for resulting clusters were informative. Interestingly, with k-value of 6 there was no clear over-clustering compared to k-value of 5. The clusters with k-value of 6 were similar as with k-value of 5, however, there was additional separation of outer neighbourhoods into categories that differ mainly in property prices. Considering our aim of informing new home buyers in their decisions, this was a clear advantage. We decided to use clusters obtained with k-value of 6 as the final output of our analysis.

Results

The results of our analysis show that it is possible to cluster areas in Edinburgh into 6 different neighbourhood categories. We collected data on several venue types and property price data across Edinburgh.

We found that by far the most common property type sold was a 2 bedroom flat. We also found that for properties with less than 5 bedrooms the mean price was similar for flats and houses. The most common venue category in the data we collected was Food places, which primarily comprise of restaurants.

The rest of the analysis focused on 2 bedroom properties (flats and houses), as this is the most often sold property type. This is enough to prove the value of this analysis for the aim of informing home buyers. However, the same analysis could easily be performed for a different property type (e.g. 1 or 3 bedroom properties), likely without changing any other parameters.

Exploratory analysis showed that the spatial density in this data was enough to estimate these features for most of Edinburgh using 500 m radius areas.

We clustered the resulting circular areas distributed across Edinburgh with k-means clustering into different number of clusters. We then chose 6 clusters as optimal for this purpose based on statistical measures and how intuitive the median features of each category would be to a potential user.

Clustering the areas into 6 clusters we found that these areas were arranged into larger contiguous areas in different parts of Edinburgh, which could be used to effectively inform users on where they

may wish to purchase their new home. The resulting spatial visualisation (Figure 8) and median feature values (Table 2) of these neighbourhood categories is the main result of this project.

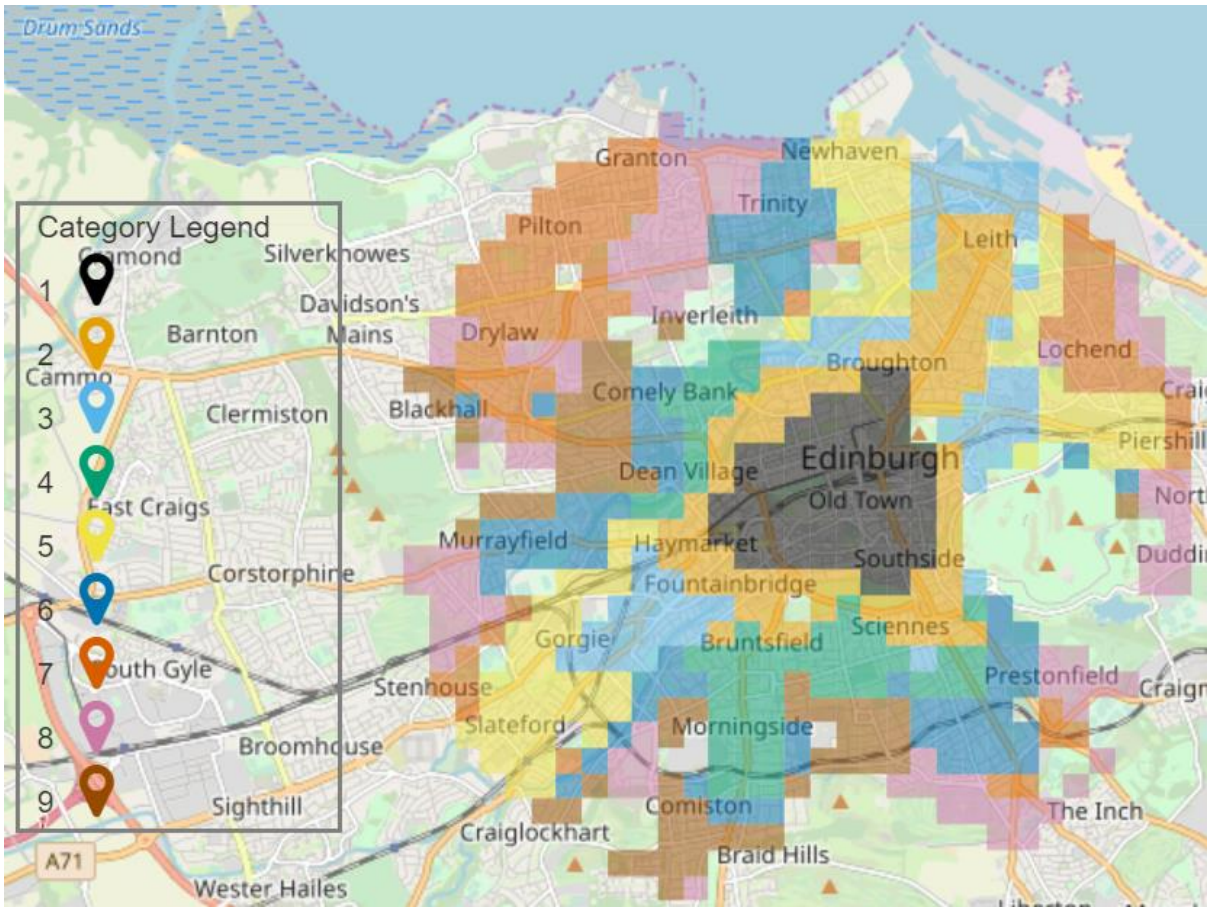

Figure 8 | Edinburgh areas included in the analysis marked by neighbourhood category.

Category	Outdoors & Recreation	Food & Drink Shop	Nightlife Spot	Spiritual Center	Bus Stop	Arts & Entertainment	Food	Cafe	Convenience Store	Price
1	40	18	85	11	20	73	155	43	8	320492
2	21	13	24	6	11.5	17.5	72	17	5	276951
3	16	5	9.5	3	8	11	30.5	8.5	2	227013
4	10	9	4	4	2	3	33	9	2	358400
5	9	4	6	1	4.5	4	11	2	1	204842
6	8	1	2	2	2	1	6	2	1	285002
7	6	2	1	0	3	1	5	1	0.5	162020
8	5	1	1	0	1	0	1	0	0	190337
9	3	0	0	0	0	1	1	1	0	317088

Table 2 | Median feature values (before normalizing) of neighbourhood categories.

The following is the description of the resulting neighbourhood categories for 2 bedroom properties.

1. City center neighbourhoods
    a. Located just around the Edinburgh Old Town and in the immediately neighouring areas.
    b. Highest number of venues of each category. Nightlife Spot, Arts & Entertainment, Food and Cafe venues are proportionally higher (around 3-4 times higher than Intermediate neighbourhoods) relative to other neighbourhood categories than other venue categories.

    c. The property prices are second highest of all neighbourhood categories, about 10 % lower than in category 3.
2. Intermediate neighbourhoods
    a. Located between the City center neighbourhoods and surrounding residential areas.
    b. The second highest number of venues in each category, except Spiritual Center. The median number of venues in all categories is much more similar to Old residential neighbourhoods than the City center, generally only about twice as high.
    c. The property prices are second highest of all neighbourhood categories, about 10 % lower than in category 3.
3. Old residential neighbourhoods
    a. Located mostly in South, but also in parts of North West, adjacent to Intermediate neighbourhoods.
    b. Similar number of venues to New residential areas in most categories, but considerably higher number of Food & Drink Shop, Food and Cafe venues.
    c. The property prices are highest in these neighbourhoods, approximately 60 % higher than in New residential areas.
4. New residential neighbourhoods
    a. Located adjacent to Intermediate neighbourhoods in South West (around Georgie) and North East (outskirts of Leith). A few areas in South East, slightly fruther out of the city than Old residential neighbourhoods, are also in this category.
    b. The number of venues is similar to Old residential neighbourhoods, with the exceptions mentioned above. There are far more venues in the New residential neighbourhoods than in Outer neighbourhoods.
    c. The property prices in neighbourhoods of this category are one of the lowest in Edinburgh.
5. General outer neighbourhoods
    a. Located at the outskirts of Edinburgh in North West, East, South East and West.
    b. There are very few venues, only 1 or 2 in 500 m radius in most categories, except for Outdoors & Recreation of which there are 6 in most neighbourhoods.
    c. The property prices in these neighbourhoods are the lowest in all of Edinburgh.
6. Expensive outer neighbourhoods
    a. Located at the outskirts of Edinburgh. Firstly, furtherest out in the South. Secondly, between the General outer neighbourhoods and the Old residential neighbourhoods in North West.
    b. The number of venues in these neighbourhoods is similar to General outer neighbourhoods.
    c. The property prcies in these neighbourhoods is one of the highest in Edinburgh.

Discussion

We acquired spatial data from Foursquare on venues of several categories and property prices from Rightmove for locations across Edinburgh. Using this data, we found six different neighbourhood categories and presented these areas on a map of Edinburgh.

We acquired the data independently for 664 locations, spaced out by 250 m, in 4 km radius of Edinburgh Castle. The data was acquired with a radius that ensured overlap between each of these data collection areas and this ensured we collected all the data available. Although, due to the 50-

venue limit in Foursquare queries we may have missed some venues in the most densely populated areas. The impact of this is likely negligible because the search areas overlapped extensively. Ideally, we would acquire complete datasets for Edinburgh without this workaround, but due to the limited interface with Foursquare and Rightmove datasets, this was the only option.

The collection of this data was itself a significant part of the output from this project. Whilst out of the scope of this project, much more could be investigated in this dataset, such as trends in property prices across years in different areas or associations between property prices and nearby venues.

Investigating the property prices, we found that there were many more samples for flats with two bedrooms than houses or any other number of bedrooms. This suggests that two-bedroom flats are the most commonly sold property type. As the aim of this project was to pilot a new tool for informing new home buyers on neighbourhoods, we focused on the most common property type. However, this same analysis can be performed focusing on another property type simply by specifying the property type in one place in the code. This makes it easy for the estate agents to put this tool into use.

To combine the venue and property price data in an intuitive format, we used the same 664 locations again as focal points across Edinburgh. We counted the number of venues and calculated mean property price at each of these focal points using a 500 m radius. We came to the value of 500 m by investigating the distributions of venue counts and checking the radius required for enough sampling of property prices. The 250 m spacing and 500 m radius of these focal areas means they overlap, therefore individual venues and properties can be used for calculating the features for multiple focal areas. This determines the interpretation of the table listing the features of each of the focal point (neighbourhood) categories. For example, at any position on the map marked as category 4, New residential neighbourhood, it is very likely to find 11 food places and 5 nightlife venues within 500 m, while the 2-bedroom properties within 500 m from that point are likely to cost around £214,000.

We did several steps of processing in order to find these exact clusters. Using the logarithm of venue counts was necessary to better separate categories of areas that all have low number of venues. We also used z-scoring of all features to ensure each feature contributes to resulting clusters equally. While outside the scope of this project, it would also be possible for the user to specifically bias the clustering for importance of some venue types, by multiplying that feature by a weighting factor. For example, they may enjoy outdoor activities and would like to find the best neighbourhoods mostly focusing on that. The pre-processing steps that we used were aimed for optimised for finding maximally generic neighbourhood categories.

The resulting clusters depended on a parameter that defined how many clusters we were looking for. The decision on the final parameter informed by statistics but rather subjective. The statistical methods were not very informative likely because the clusters were statistically difficult to separate. A larger dataset with higher spatial resolution would improve clustering and make deciding on the number of clusters more objective, but such a dataset was not possible the obtain with available sources. Regardless, the identified clusters are intuitive and would be useful for home buyers new to Edinburgh. Furthermore, the fact that the neighbourhoods in each category form contiguous regions on the map beyond the extent to which individual focal areas overlap suggest we have not over clustered the data.

When clustering with any number of clusters, we saw the same trend in median venue counts. The number of venues in 500 m radius generally decreases concurrently in all venue categories from one cluster to another. In other words, the relative intra-cluster venue counts were similar across clusters.

In the neighbourhood categories that are further from the centre the number of venues in any categories decreases rapidly.

By consulting with local estate agents in Edinburgh, it would likely be possible to describe the neighbourhood categories even better. For example, the Old residential neighbourhoods and the New residential neighbourhoods are in the same ballpark with number of venues, except for Food & Drink Shops, Food places and Cafes, of which there are far more in Old residential neighbourhoods. The property prices in Old residential neighbourhoods are the higher than in any other neighbourhood category. Based on this, it is likely that the Old residential neighbourhoods are well-developed areas, while the New residential neighbourhoods may be the areas considered to be still up and coming. Further analysis on property price changes over the last few years could be used to verify this.

The six different neighbourhood categories presented in the Results can provide effective guidance to home buyers new the Edinburgh. When considering buying a home in an area belonging to one of the neighbourhood categories, the buyer can easily see how many venues of certain category they will likely find within 500 m from their new home. They will also know what the common price for a specific property type is in each area. Furthermore, when they find that they like a certain part of the city, they can find out which neighbourhood category it belongs to and discover other parts of the city that are similar.

Conclusion

By collecting venue location data from Foursquare and spatial property price data from Rightmove, we were able to find six distinctive neighbourhood categories and localise these on the map of Edinburgh. The results show areas of Edinburgh that can be intuitively related to types of neighbourhoods in any city. With further analysis more specific conclusions could be made about many of the categories. The rich data acquired for this project could be used for additional analysis of the Edinburgh property market beyond clustering of neighbourhoods.