An investigation of the influence of location and property characteristics in Swansea, UK

Overview

An investigation of the effect of the details (or characteristics) of a property (e.g., room size) and its nearness to different location venues (e.g. shops), on the property prices in Swansea, UK.
The Swansea area is characterised by details of its households, both property and location details. This information is used to show the importance these aspects have on property price and create house price predictions. Cluster analysis is performed to look at the characteristics of different areas.

The investigation is an exploratory analysis of the capabilities of machine learning and the foursquare location service (https://foursquare.com/). The report is produced for the final submission of the IBM Data Science Professional Certificate [https://www.coursera.org/professional-certificates/ibm-data-science]. Some of the data and python files involved in the analysis are given at https://github.com/dMaterialia/SwanseaProperty. Some other links involving interactive plots are given within the report.

Keywords: Swansea, Property, FourSquare, IBM Data Science, Machine Learning, Clustering, Random Forrest Model

Contents

1 Introduction

Swansea is a city in Wales in the United Kingdom. It is the second largest city in Wales and the twenty-fifth largest city in the United Kingdom, with a population of 241,300 in 2014 [1]. The city is located by the sea, on the Bristol Channel. Within the Swansea region there is a mixture of location types including seaside, maritime, residential, shopping, industrial, and agricultural (Figure 1).

The city's location relative to the sea and the nearest cities in Wales and England have an important influence on the city. The nearest city is Cardiff the capital of Wales, 68 km to the west, whereas to the east St. Davids is the nearest city at 110 km and to the north St Asaph at 220 km away is the nearest city. The population density in the surrounding regions is therefore relatively low compared to the rest of the UK (Figure 2) and Swansea is a relatively remote location to other places in the UK.

Figure 1. Images of Swansea. Sources various.

Within the Swansea city region being investigated the nature of the individual regions varies considerably. To the west is the Gower peninsula a coastal region which is designated as an Area of Outstanding Natural Beauty [2]. Whereas, to the north and east are regions with a high density of industrial estates. There are also areas known for their nightlife including Uplands and Mumbles as well as the city centre region. The Swansea marina district is a thriving residential area near the city centre consisting of a high density of flats, which replaced the previous industrial area that had declined from 1970s onwards [3]. In addition, the student population has an important influence on the city, with a high-density student population to the west of the city in the Uplands and Brynmill areas. In between these areas are a variety of residential areas, with varying deprivation and property prices.

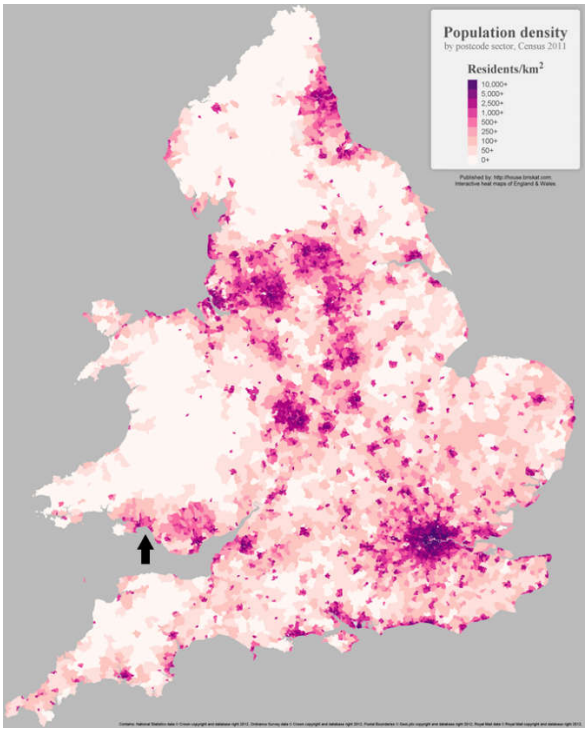


Figure 2. Population density of the UK, arrow highlights Swansea. Image from [4].

In the property world, the cliche is that the three most important factors in the desirability of a house are "location, location, location" [5]. The location of Swansea relative to the rest of the UK and the location of the parts of Swansea to its local geography (e.g., the beach, city centre, marina) have shaped the nature of the areas within it. One of the questions that this work seeks to answer is how the current local geography dictates the property prices. How true is “location, location, location” and can this effect be quantified relative to the details of a property? What are the different aspects about location that matter? And to what extent? What can be used to guide property investment or government planning? This can be done by considering what details may change in the future and what cannot (e.g., distance to a beach versus distance to local shop).

It should be noted that socio-economics or crime statistics are purposely ignored, this makes the problem simpler but is also done because this is not what is being investigated. However, the simplicity of the data used may be at a detriment in terms of predictions. These details could be relatively easily included, as deprivation is included in one of the datasets and the crime statistics are readily available. However, property price and deprivation or crime statistics are well linked. Hence, including these variables may mask other effects of location or property details.

The work will try to address:

- How much of the property price can be ascertained from details of the property, such as number of rooms, alone?
- And can the influence of location on price be quantified based on the businesses, facilities and environment situated nearby?

Property prices will be used as a target value, as an indication of the property's desirability. The other values will be used as parameters to understand this.

2 Data and Methods

2.1 Data

The two main sources of data in this report are the UK Census data and the foursquare.com database. In addition, the website doogal.co.uk and UK data service are also used.

2.1.1 Census Data

The UK census is a data collection exercise in the United Kingdom which occurs every 10 years. Participation in the census is compulsory and is used to obtain statistics on the UK's economy and society, which is then used to assist the planning and allocation of resources, policy-making and decision-making [6]. Details of the census can be accessed at the website [7]. Mainly to maintain anonymity of participants, the census data is separated into areas which are unique to the census. The areas are larger than single postal code regions, e.g., SA2 0DE, but smaller than the first digits of the postcode e.g., SA2 or district regions such as Uplands in Swansea.

In the Swansea area there are ~9,000 unique postcodes. Whereas for the UK census areas there are ~1,000 regions in Swansea. Most analysis will be done on these census areas to allow data to be merged from different sources.

As shown in Figure 3 lots of information can be obtained from the census data. Based on the remit used here, only 3 of these datasets were used: 1) KS101EW- Usual resident population, 2) KS401EW- Dwellings, household spaces and accommodation type, 3) KS403EW- Rooms, bedrooms, and central heating. Within these files only some of the data is used. The categories extracted are shown in Table 1.

Table 1. Categories used from Census data about property details.

No. of People	Area (hect.)	Density (ppl/hect.)	Rooms Per House	Bedrooms Per House	% No central heat	No. of Dwell.s	% Detached	% Semi Detached	% Terraced	% Flats	% Mobile

Figure 3. Screen shot of the Census data page used [8].

2.1.2 Doogal data

There is some cross-over between with the Census data as some of the same data sources are used. But Doogal will mainly be used for:

- To match postcodes to census ID

- Location data- latitude and longitude
- Property price information

2.1.3 Four Square

For location information FourSquare [9] will be used. Instead of using the explore function, several key locations have been identified that can characterise an area. These include:

- Beach
- Parks
- Pubs
- Supermarkets
- Doctors
- etc

Approximately 20 categories will be used, which will mean ~20, 000 calls to foursquare. The reason for this method is that the explore approach was not able to uniquely identify regions, for example large cluster sizes were needed. More detail is provided in the methods section.

2.1.4 Location Polygons

Location polygons were found from the UK data service [10] this allows plotting of Choropleth Maps. Choropleth Maps are location plots where different regions are colour coded based on a parameter such as house price. To do this polygons of the regions are required, normally in in .KML or .shp file formats. A slight modification of the Census polygons was done to convert .shp files to .json files for use in Folium which is shown in the file "*createSwanseaJSON_fromGov.ipynb*".

2.2 Methodology

All analysis is done using Python 3.8. Both Spyder and JupyterLabs are used. Spyder to create .py function files which are then called by the notebooks in JupyterLabs for visualisation of the results. Most of the scripts and data can be accessed on my GitHub page for this project [11]. The interactive nature of folium plots are difficult to share, the best way around this is to use nbviewer [12] to share notebooks put on github.

2.2.1 Property Prices

The property data exists as sales data for each post code over several years. Since, property price is the key variable in this report some consideration of how to deal with this is needed. Given inflation of property prices, this data is not ideal. Data sources such as Zoopla which create an expected price of a property; however, this cannot be used here as they forbid webscraping. Instead, to use the property sales data property inflation will have to be accounted for. A simple way to do this, and that used here, is to take the average over a period after the financial crash and before a current upsurge in prices when prices were approximately level (Figure 4). Between 2008 and 2019 is used.

A secondary data filtering was attempted, whereby the house prices where adjusted based on the percentage of house types in the Census data. i.e., Avg. House Price = %Terraced x Avg. Terrace price + %Detached x Avg. Detached price + etc. However, given the sparsity in some areas for property sales it gave undue weight to some individual house sales. For this reason, the overall average was used. A more refined procedure accounting for the number of sales in each property type that uses general trends in prices in different property types across Swansea would be an improvement. In addition, a weighting of predictions based on confidence of property prices would be another improvement.

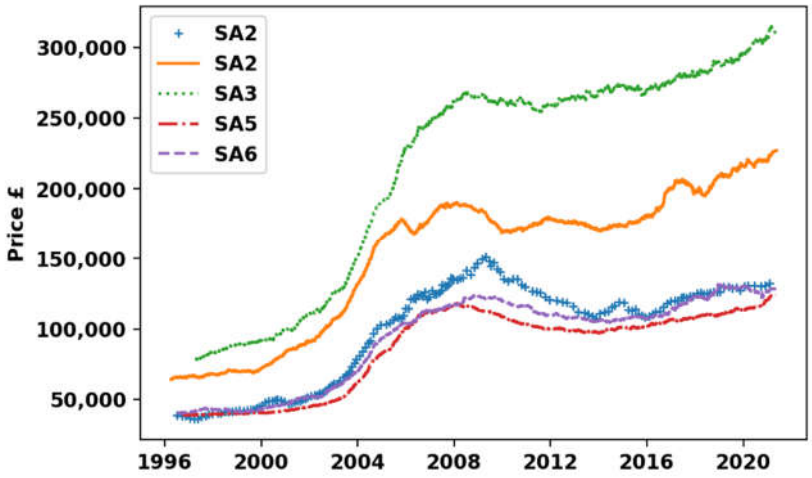


Figure 4. Change in property price for different post code areas. All property sales are used, and a moving average is used.

As is shown in Figure 5. There is a considerable variation in property prices across Swansea.

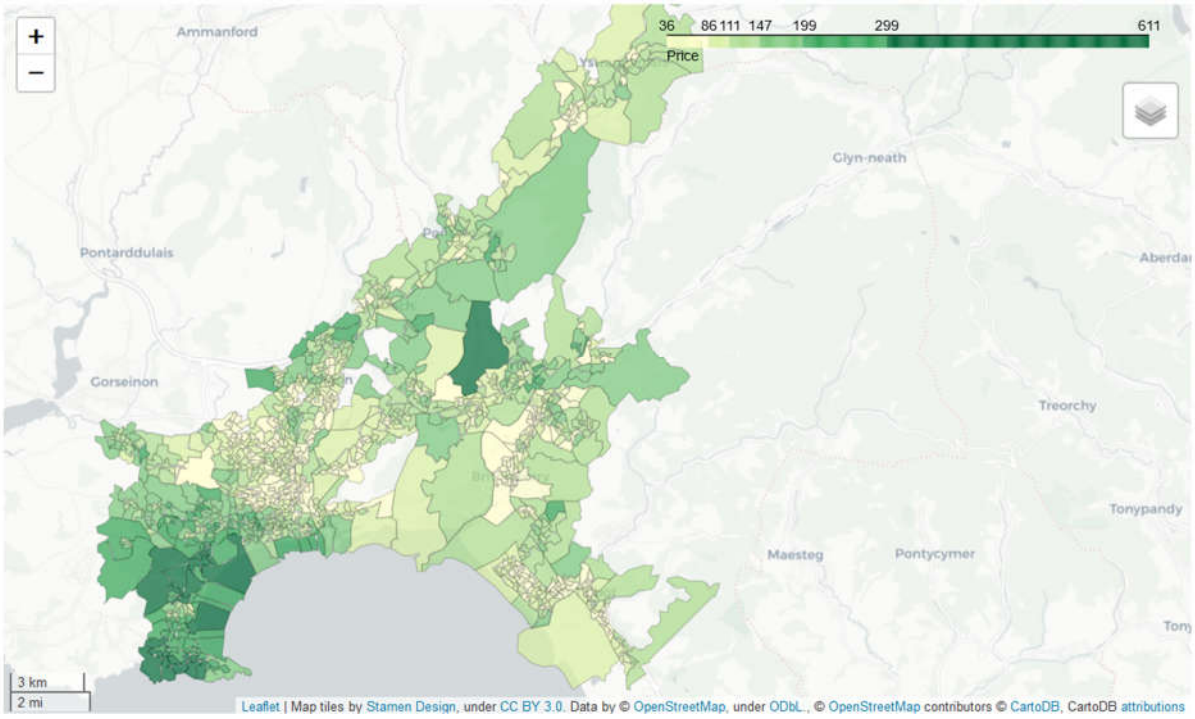


Figure 5. Price of properties in Swansea Census regions.

2.2.2 Selected Areas

To help break-down the data, five regions within Swansea have been selected to provide additional information. These are shown in Figure 6.

- The Langland region is situated to the west of the city on the Gower peninsula close to a beach.
- The Penderry region ranks low in the Census deprivation index, situated to the north of the region.
- The Morriston North region is a middling region within Swansea, situated in the north of the region and close to the M5 motorway.

- Uplands is one of the most popular areas outside the centre for its nightlife and shops, it also has a high student density.
- The Swansea Marina area is a relatively new region within Swansea, emerging after the deindustrialisation of the area, consisting mainly of flats. The Maritime region chosen is one of the more expensive Census regions in this area.

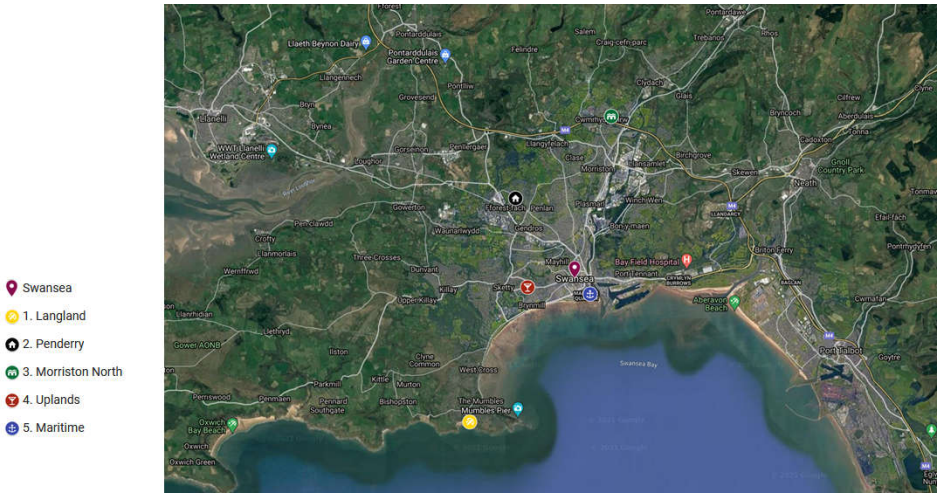


Figure 6. Map of Swansea area. The city centre and locations of post-code regions that will be looked at in more detail are shown.

2.2.3 FourSquare

Foursquare is used to get information about venues close to the census regions. Venues include a wide range of places including shops, offices, beaches, sports centres (a full list is found here [13]. The venues are separated with a venue hierarchy, so for example for a football stadium, is in the stadium category which is itself in the Arts & Entertainment category. One way this hierarchy caused problems was in separating out restaurants from fast food venues, where each specific venue had to be selected.

Foursquare has two main venue searches:

- Search: returns a list of venues near a given location [14]
- Explore: returns a list of recommended venues near a given location [15]

Both have some customization on the criteria searched, based on the type of venue as well as other criteria. This is particularly the case for 'explore' where time of day/week, popularity of venue and price can be considered. Venue categories can be limited both by providing a list of categories and in 'explore' only by specifying a section which includes: "food, drinks, coffee, shops, arts, outdoors, sights, trending, nextVenues (venues frequently visited after a given venue), or topPicks (a mix of recommendations generated without a query from the user)" [15].

Two methods are used, one using 'search' to get selected venues near the census regions, and a second using 'explore' using the different section categories. 'Search' is used for supervised machine learning to predict property price. And 'explore' for unsupervised learning to categorise clusters based on location.

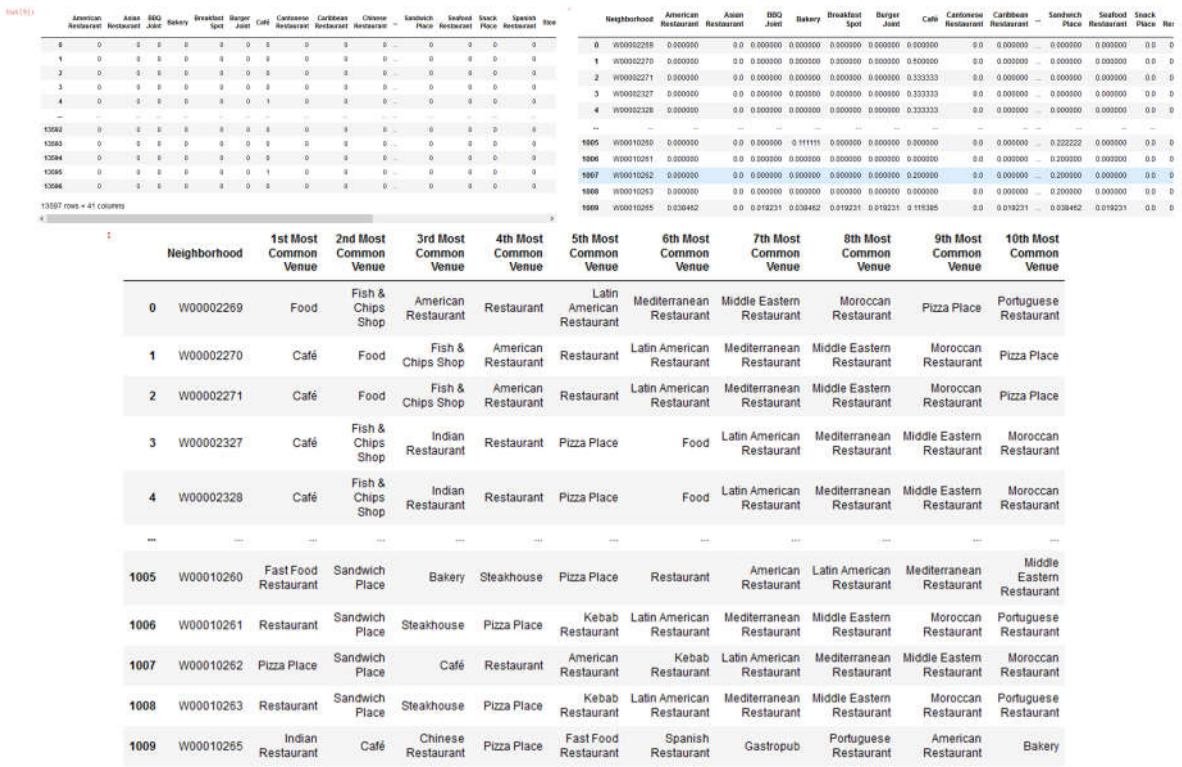
For the 'search' call 24 venue categories were chosen (shown in Table 2) and either the distance to this venue type (given as Distance in the table) or how many are within a given radius of the census region (given as Frequency in the table). The distance metric is the radius over which these are determined. Further details of this processing is in the file "ClusteringByLocation.py" in the function "fit_Clusters".

Table 2. The different venue categories used in the 'search' call. 'Distance' means that the distance to a venue is measured, whereas 'Frequency' gives the number of those venues within the given radius.

Category	Beach	Park	Sports	Marina	Supermarket	Foodstore	Restaurant	Takeaway	Bar	Nightclub	School	University
Measurement type	Distance	Frequency	Frequency	Distance	Distance	Frequency	Frequency	Frequency	Frequency	Frequency	Distance	Distance
Distance metric (m)	15000	700	700	15000	15000	300	700	700	700	700	15000	15000

Category	Betting shop	Post-office	Doctor	Hospital	Transport (e.g. bus station)	Spiritual centre	Hotel	Waste disposal	Office	Industry	Shopping centre	Pawnshop
Measurement type	Frequency	Distance	Distance	Distance	Distance	Distance	Frequency	Distance	Frequency	Frequency	Distance	Frequency
Distance metric (m)	700	15000	15000	15000	15000	15000	700	15000	700	5000	15000	700

For the 'explore' call each of the 'Sections' are used with a search radius of 800 m, the time and day were set as 'any', and sortByPopularity as 1 (i.e. venues at any time sorted by popularity). The data requires additional processing to incorporate into a machine learning model. First one hot encoding is used to convert the categorical data about venue details (a list of venue types for each location) to a form as shown in Figure 7, whereby a 1 indicates the venue category exists. Then for each location the rows are added and normalised, so that the number in a row represents the fraction of that venue type found in that area. So, in the figure area W00010265 has 3.8% American restaurants, 1.9% BBQ joints. This is the data used in the model. To help with categorising the data is then converted to a list of the most common venue types, as shown in the figure. Further details of this processing is in the file "ClusteringByLocation.py" in the function "Clustering".



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	W00002269	Food	Fish & Chips Shop	American Restaurant	Restaurant	Latin American Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Moroccan Restaurant	Pizza Place	Portuguese Restaurant
1	W00002270	Café	Food	Fish & Chips Shop	American Restaurant	Restaurant	Latin American Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Moroccan Restaurant	Pizza Place
2	W00002271	Café	Food	Fish & Chips Shop	American Restaurant	Restaurant	Latin American Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Moroccan Restaurant	Pizza Place
3	W00002327	Café	Fish & Chips Shop	Indian Restaurant	Restaurant	Pizza Place	Food	Latin American Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Moroccan Restaurant
4	W00002328	Café	Fish & Chips Shop	Indian Restaurant	Restaurant	Pizza Place	Food	Latin American Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Moroccan Restaurant
...	...	...	...	...	...	...	...	...	...	...	...
1005	W00010260	Fast Food Restaurant	Sandwich Place	Bakery	Steakhouse	Pizza Place	Restaurant	American Restaurant	Latin American Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant
1006	W00010261	Restaurant	Sandwich Place	Steakhouse	Pizza Place	Kebab Restaurant	Latin American Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Moroccan Restaurant	Portuguese Restaurant
1007	W00010262	Pizza Place	Sandwich Place	Café	Restaurant	American Restaurant	Kebab Restaurant	Latin American Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Moroccan Restaurant
1008	W00010263	Restaurant	Sandwich Place	Steakhouse	Pizza Place	Kebab Restaurant	Latin American Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Moroccan Restaurant	Portuguese Restaurant
1009	W00010265	Indian Restaurant	Café	Chinese Restaurant	Pizza Place	Fast Food Restaurant	Spanish Restaurant	Gastropub	Portuguese Restaurant	American Restaurant	Bakery

Figure 7. Output from data processing after details of venues have been determined. Top-left, one hot encoding, top-right convert to a fractional values for modelling, bottom, covert to help categorising and plotting.

2.2.4 Swansea Data Plots

Plots of Swansea to visualise the results were performed with Folium [16]. Two types of plots were used Choropleth maps, where the region is coloured based on a scalar quantity, and Cluster plots, where the colour of marker shows what cluster an area is in.

For the latter two colour arrays (rainbow and rainbow2) were created for the inner and outer colour of the markers using seaborn, as follows:

```
# set color scheme for the clusters
N = kclusters+5#add a bit as color returns to same
```

```
import seaborn as sns
import copy
import random
colors_array = sns.color_palette('hls', N)#husl hls Paired
rainbow = [colors.rgb2hex(i) for i in colors_array]
rainbow2=copy.copy(rainbow)
random.shuffle(rainbow)
```

2.2.5 Machine Learning Algorithms

Both supervised and unsupervised models were used.
Figure 8 displays a handy guide for choosing a machine learning algorithm, taken from [17]. For the trained models, we want to predict a value. Therefore, as shown in the figure a linear regression model is a fast method to use. For comparison a Decision Forest Regression is used, which has a high accuracy whilst also having fast training times. The third method used for supervised learning, k-nearest neighbour regression, was chosen because the algorithm is like the one used for the unsupervised learning, and may provide different insights in the data. For the unsupervised learning model, a model was wanted that could cluster data and reveal structures. For this there is one clear choice from the figure, K-means clustering.

The supervised models used property price as a target. For all models, data was split into test and train datasets with 70% of the data used for training. All reported errors in results are from applying the model to the test dataset. A linear regression model was used on the property details, but due to its accuracy compared to other methods was not used elsewhere. A random forest regressor model (RandomForestRegressor) and a K nearest neighbours regressor (KNeighborsRegressor) are the main models used for supervised modelling. For the RandomForestRegressor the number of trees in the forest is determined by finding the minimum mean absolute error using a range of values for the number of trees (up to 600). For the KNeighborsRegressor the number of nearest neighbours was found using GridSearch for neighbours up to 15 (values normally found were ~4).
An unsupervised model was used on the location data to cluster different areas. For this a K-Means clustering (Kmeans) model was used. To determine the number of clusters to form, the ‘elbow method’ was used. This was done by calculating the square of Euclidean distance of each point from its cluster centre for different number of clusters. The two are plotted against each other and the value chosen is where there is a noticeable flattening in the curve (the elbow).
All models used are from scikit-learn [18].

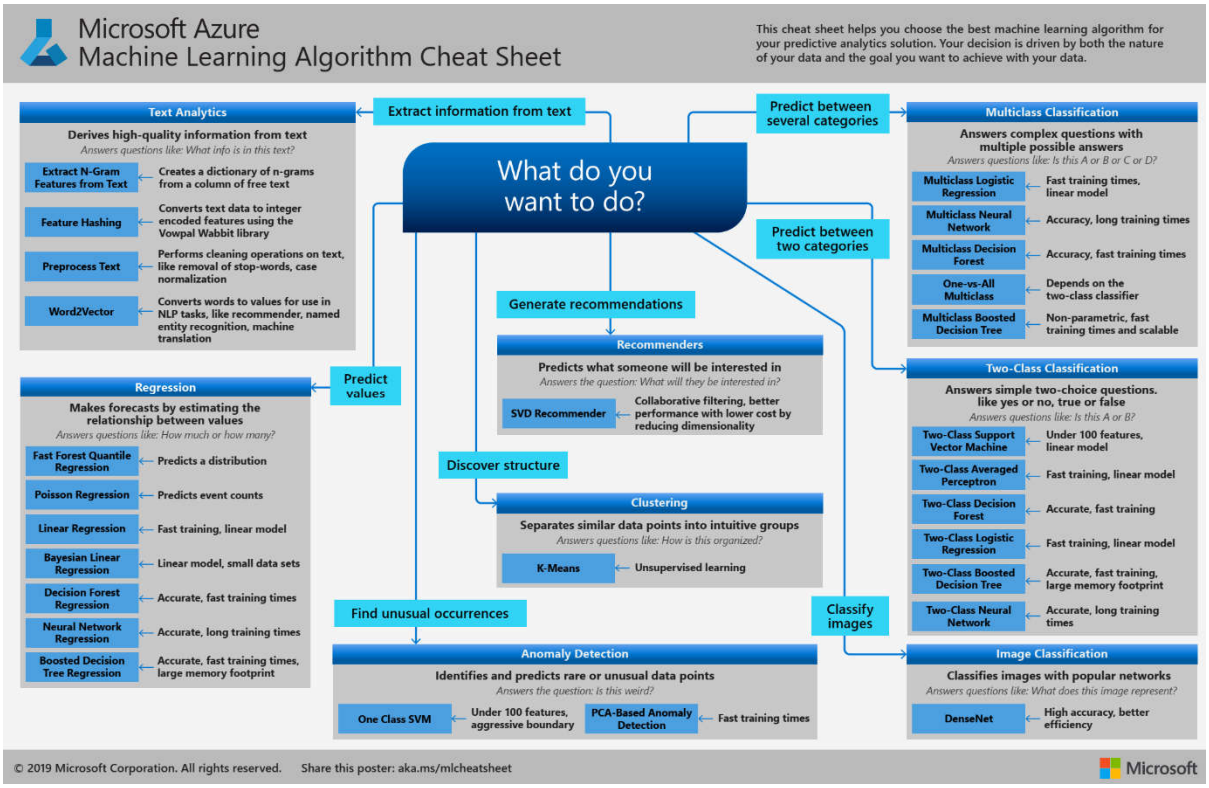


Figure 8. Machine learning algorithm cheat sheet, taken from [17].

2.2.6 Post processing

To help understand the results of the supervised learning models more and find out about the effect of location and number of rooms an additional processing is performed. The way this is done is by altering the data slightly and giving it back to the model.

For example, to understand the location effect on property prices, the average property details are given to the model alongside the region's location details. This can be seen in the function "locationeffect" in "MachineLearnBig.py". Or to quantify the effect of the number of rooms by changing the number of rooms whilst keeping other parameters the same.

3 Results

3.1 Property Details

An overview of the property details data is shown in Table 3 and for the selected locations. For each parameter there is a wide range in values, many variables vary by a factor of more than 2 between maximum and minimum values. The deprivation index, a measure of how deprived an area is using multiple sources such as employment, social class, and availability of cars [19], is also included here but not used in analysis (the higher the value the less deprived). The selected locations also show a range of values for the parameters. The properties in Mumbles have an average price over five times that in Penderry, and two times higher than the Uplands and Maritime regions. There is also a mix in property types, Maritime is mainly flats, whereas Penderry and Morriston are mainly semi-detached and Mumbles mainly detached. Variations in the density of people in the areas also changes significantly, being lowest in the Maritime district which is around seven times denser than Penderry and Mumbles districts.

The 12 parameters on the left are used as input for the model with a target of the average price.

Table 3. Information of the details of properties in the different Census regions within Swansea using the ".describe" function of pandas.

	No. of People	Area (hect.)	Density (ppl/hect.)	Rooms Per House	Bedrooms Per House	% No central heat	No. of Dwell.s	% Detached	% SemiD	% Terraced	% Flats	% Mobile	Deprivation Index	Price
mean	299.7	299.7	26.2	5.5	2.8	1.3	137.1	19.3	37.1	28.5	14.9	0.2	855.3	£125,848
std	70.6	70.6	105.9	0.8	0.4	1.5	23.2	21.8	25.4	26.1	21.6	2.3	611.5	£61,445
min	119.0	119.0	0.8	2.1	1.2	0.0	50.0	0.0	0.0	0.0	0.0	0.0	16.0	£35,000
25%	261.0	261.0	4.9	4.9	2.5	0.0	124.0	3.4	14.4	5.6	0.9	0.0	290.0	£85,000
50%	292.0	292.0	7.6	5.5	2.8	0.8	133.0	9.3	35.1	21.4	5.0	0.0	744.0	£110,000
75%	326.0	326.0	14.5	5.9	3.0	1.8	145.0	29.8	56.7	44.7	19.8	0.0	1456.0	£146,000
max	994.0	994.0	2336.5	8.2	4.1	11.9	290.0	97.8	95.3	96.2	100.0	46.2	1906.0	£611,000

Table 4. Average property details of the selected regions. Note that a higher deprivation index indicates less deprivation of a region.

Area	GEOGRAPY CODE	No. of People	Density (ppl/hect.)	Rooms Per House	Bedrooms Per House	% No central heat	No. of Dwell.s	% Detached	% SemiD	% Terraced	% Flats	% Mobile	Deprivation Index	Price
Mumbles And Newton SA3 4N	W00004422	285	20.6	7.96	3.78	0.85	123	91.06	4.88	1.63	0.0	2.44	1760	£611,000
Penderry SA5 5E	W00004441	353	13.9	5.03	2.59	0.74	138	15.22	63.04	10.87	10.87	0.00	31	£66,000
Morriston North SA6 6Q	W00004348	286	7.26	5.51	2.97	0.82	124	24.19	74.19	1.61	0.00	0.00	1618	£126,000
Uplands SA2 0A	W00004643	398	3.1	6.28	3.43	2.19	138	4.35	3.62	72.46	18.84	0.72	1541	£182,000
Maritime quarter SA1 1S	W00010253	185	1.63	4.08	2.23	0.98	117	0.85	4.27	17.09	76.92	0.85	1034	£181,000

The influence that the parameters have on property price can be highlighted by scatter plots and the Pearson correlation coefficient. The Pearson correlation gives the following relationships: 0.5 to 1 strong correlation, 0.3 to 0.49 moderate correlation, 0 to 0.29 low correlation.

For the property data, scatter plots with Pearson correlation displayed in the tile are shown in Figure 9. Most of the property detail variables have a low correlation with price. The exceptions to this are: (1) Number of rooms and bedroom which have a high positive correlation (i.e., more rooms more expensive properties), (2) The percentage of detached properties, positively correlated with price, and (3) the percentage of terraced properties, which is negatively correlated with price and unlike the other two only moderately correlated.

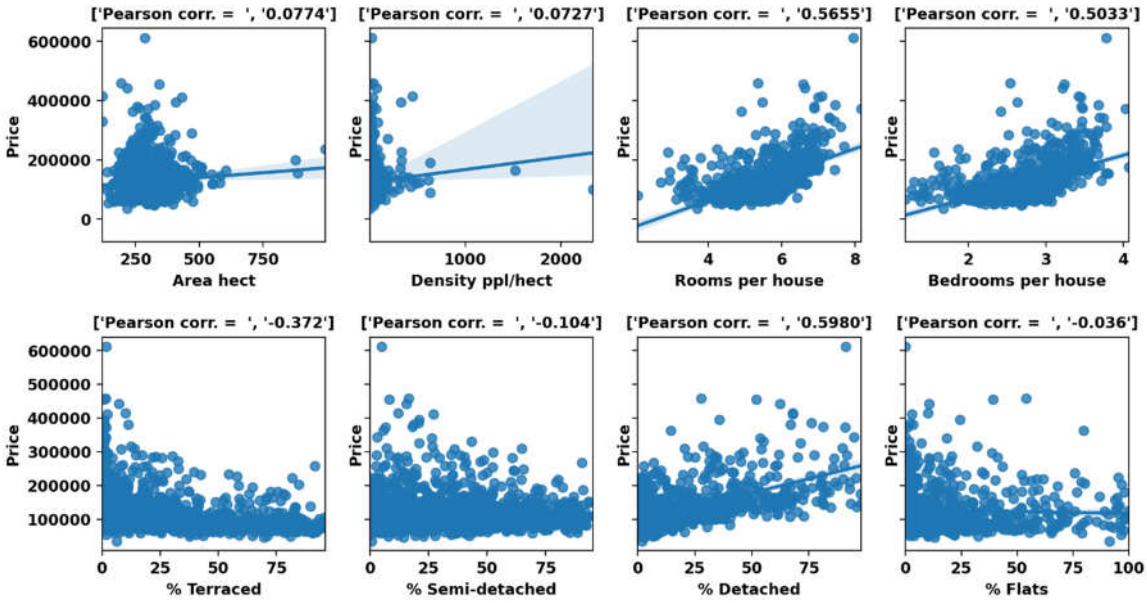


Figure 9. Correlation plots of the different details of properties used compared with the averaged property prices.

Using the property details and house prices, predictions on the property prices using different models were made. The results of this analysis are shown in Table 5.

The first model used was a linear regression model, the parameters obtained being shown in Table 6. The results of the model are consistent with the correlation plots, although a little difficult to interpret for the house types. But it is notable that the model estimates a gain of £36,000 per extra room in a property.

With the linear regression model using the property data the mean absolute error (MAE) in property price is £27,000 and the variance score is 0.58. For the KNN regressor model the values are improved for MAE at £24,900 but lower for variance at 0.53. The best model is the Random Forrest Regressor, which has MAE of £24,600 and variance of 0.57.

Based on these results a reasonably accurate predictor of property price can be made from the property details alone. However, there is more to property prices than property details alone. This is illustrated in Figure 10, which shows the difference between property prices and predictions using the Random Forrest model. The figure shows that some areas have a marked difference in property price between model and actual. The regions in the Gower in the west (with negative price difference) suggest the importance of location.

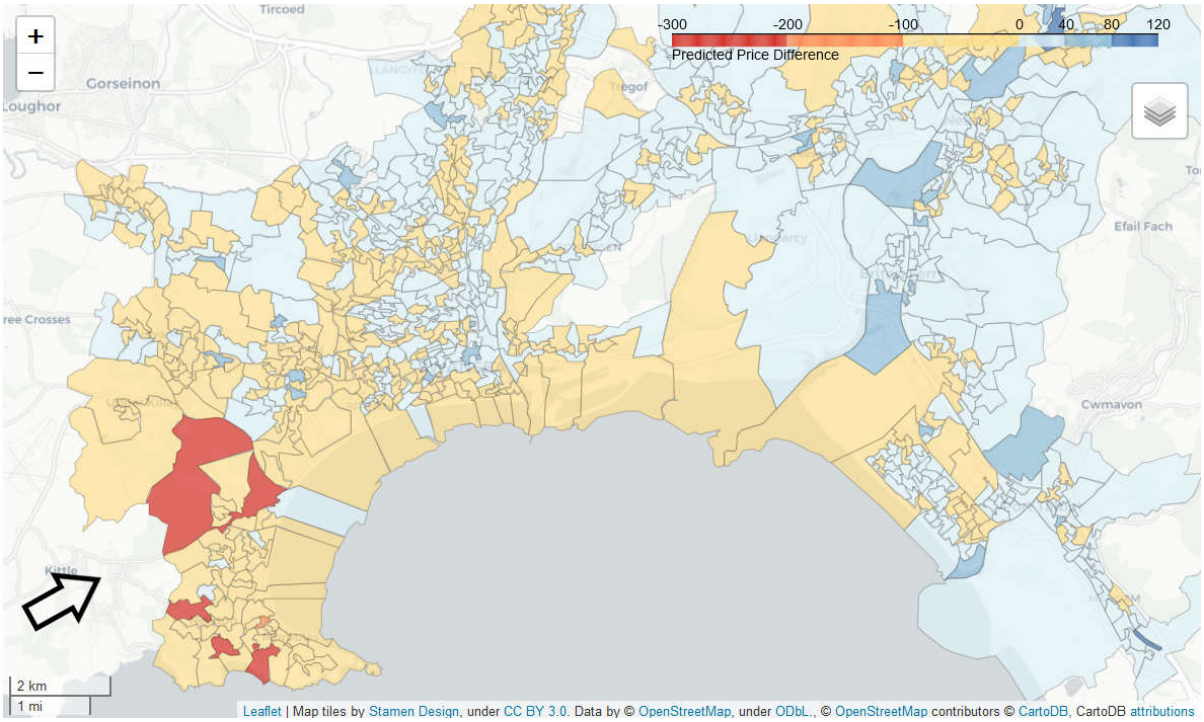


Figure 10. Map of Swansea showing the difference in Census regions between predicted property price based on property details. Using the RF model. A negative value indicates the model underestimates the measured property price.

Table 5. Details of models to predict the property price in Swansea. The mean absolute error (MAE), variance score and details of the parameters used are shown. The models are performed for details about the property alone (Property), details about the location alone (Location), and details about both location and property details (Location + Property). Three different machine learning models are used.

	Mean absolute error (MAE) (£ 000s)			Variance score			Parameter Used (n_estimators / Neighbors)		
	Property	Location	Location + Property	Property	Location	Location + Property	Property	Location	Location + Property
Linear Regression	27.0			0.58					
KNN Regressor Using the Search function	24.9	24.0	19.8	0.53	0.66	0.73	7	3	2
Random Forest Regressor Using the Search function	24.6	22.8	17.8	0.57	0.70	0.78	90	270	480
Random Forest Regressor Latitude & Longitude		22.6			0.72			90	
Random Forest Regressor Explore topPicks		30.5			0.42			540	
Random Forest Regressor Explore trending		30.0			0.43			360	
Random Forest Regressor Explore shops		27.8			0.54				

Table 6. Results of using a linear regression on the property details to predict property prices. The property price is given as: intercept + parameter multiplied by value. E.g. 7000 + 37000x5 + -2100 + …. For a 5 room terraced property.

Intercept	Area hect.	Density (ppl/hect.)	Rooms per house	Bedrooms per house	% Central heating	% Detached	% Semi	% Terraced	% Flat
7,000	-65	-1,200	37,000	36,000	2,200	-1,000	-2,000	-2,100	-520

3.2 Location details

To examine the influence of location, several approaches are taken. Given that the Random Forrest Regressor model was more accurate in its prediction than the other two, this model will be used here.

As a first step the longitude and latitude of each region are used as the variables to predict price (recall that 70% of the data is used for training and 30% for testing). As shown in Table 5, with only this position data, this model predicts property prices with a lower mean absolute error than the equivalent RF model using the property details (£22,600 versus £24,900 for RF property details). Therefore, this confirms location as the number one classifier on a property and the mantra: "Location, Location, Location".

This description of location (Latitude and Longitude) would be of no use in predicting property prices in different areas outside Swansea, or for understanding the reason for the location effect. Instead, to try to quantify this location effect in Swansea the FourSquare website is used to determine details about the census regions. Two approaches are used:

(1) Finding the distance and density of several selected features from each Census region using the FourSquare 'search' function [14],

(2) Finding the features nearby a region based on different categories using the FourSquare 'explore' function [15].

3.2.1 Locations Using Search

From the first of these categories, the correlation of different categories is shown in Figure 11. The Pearson correlation gives the following relationships: 0.5 to 1 strong correlation, 0.3 to 0.49 moderate correlation, 0 to 0.29 low correlation. Most categories have a low correlation. Only supermarket distance has a strong relationship (a positive one, the further away the higher the property prices) and only industry distance has a moderate correlation (a negative one lower house prices closer to industry). Both shopping centre and Schools have a positive correlation at ~0.29, just under the moderate definition. This may be a sign that the categories, or how they were implemented (e.g. distance/frequency), are not the best choice.

The clustering based on these categories is shown in Figure 13. As may be expected due to the nature of the data, the clustering occurs over distinct regions of 1-3 km in size. Although, there are some exceptions to this (e.g. Cluster 29 and cluster 2 where there are 2 separate regions).

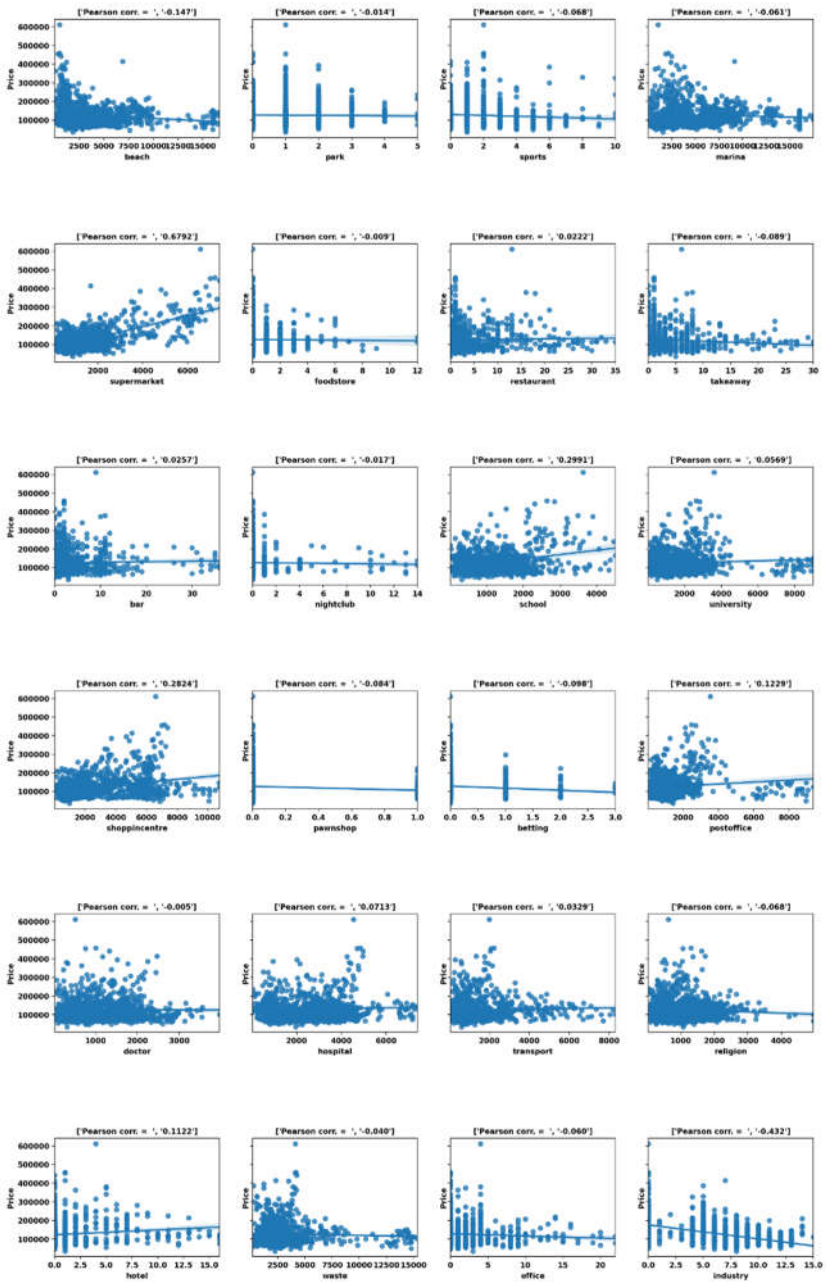


Figure 11. Correlation plots of the different location details of properties used compared with the averaged property prices.

When the location data is used to predict property prices the predictions (see Table 5) are better than using the property details alone. However, they are also equivalent to those using just the longitude and latitude information. Furthermore, if a 50:50 split between training and test data is used the Longitude-Latitude method outperforms the venue-location based predictions. This would again suggest the venue-location data is not getting to the bottom of the reason of differences in locations. It could either be that the location-venue data is unable to understand the differences and that other data including deprivation and crime statistics may need to be included or that the implementation of the foursquare data needs improvement. The large negative areas in the predicted price difference plots on the West of Swansea found using the property details (Figure 10), are reduced when using the location details (Figure 12) but still exist.

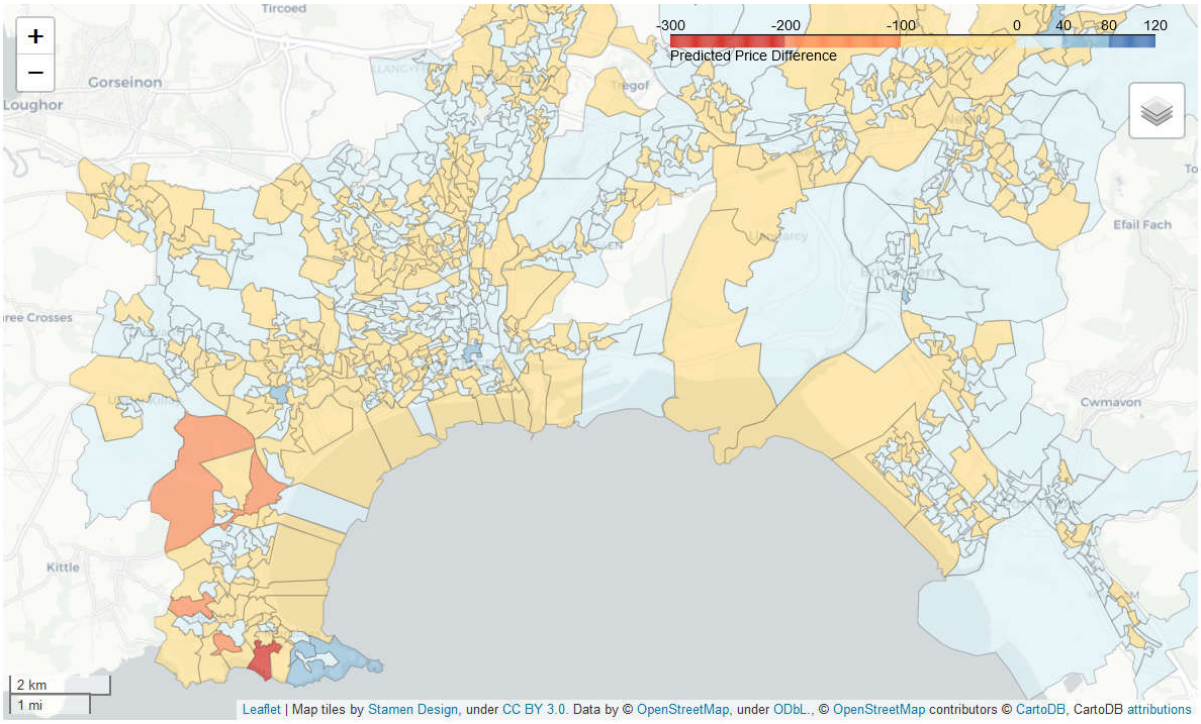


Figure 12. Map of Swansea showing the difference in Census regions between predicted property price based on location using the 'search' function. Using the RF model. A negative value indicates the model underestimates the measured property price.

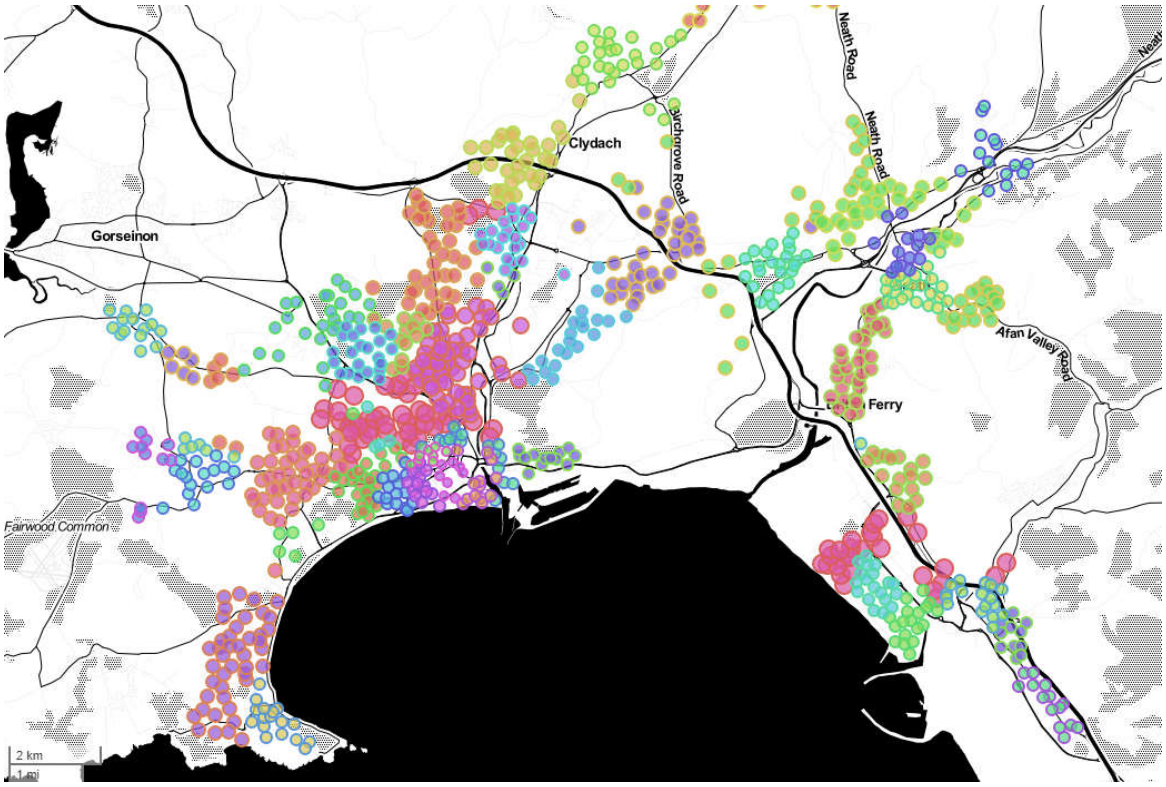


Figure 13. A cluster analysis of locations based on the location details from the 'search' category.

3.2.2 Locations Using Explore

The location details from explore using the section "topPicks", "trending" and "shops" were used for predicting the house price. As is shown in Table 5, the models give the largest mean absolute error of all the models. The data was not meant primarily for this, it was intended for unsupervised Cluster analysis, and so it makes sense that the model is poor at predictions.

The same data was used to predict different Clusters. One of these plots is shown in Figure 14 using the section as 'trending'. The main feature of these plots is the interactivity, which cannot be displayed using an image file. Instead, the interactive plots are found on nbviewer [20]. This gives details about what the clusters represent and to check whether two markers that look the same are the same cluster, when the cluster size is high different clusters can look the same.

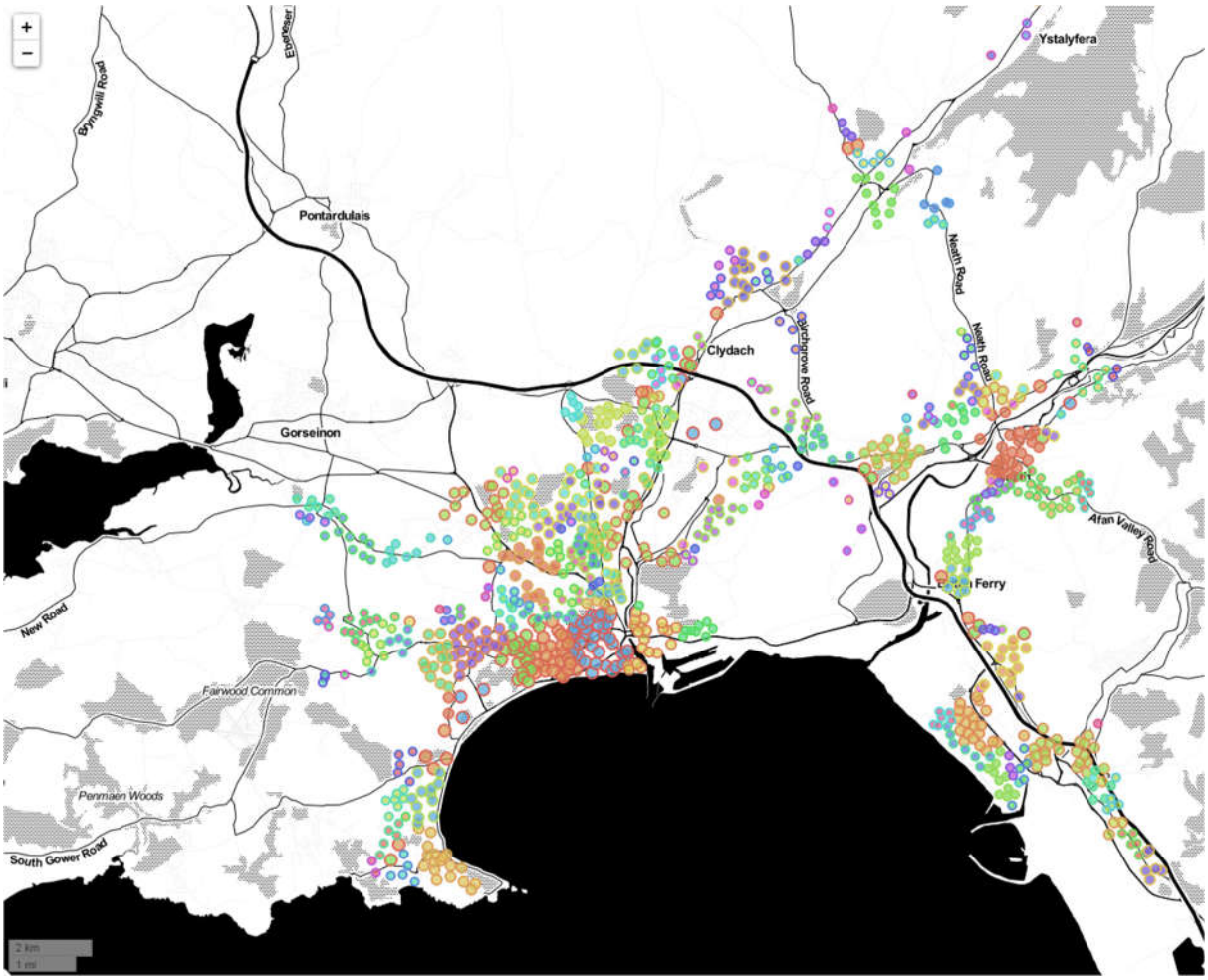


Figure 14. A cluster analysis of locations based on the location details from the 'explore category.

3.3 Location and Property details

House price predictions were made using the property details and the location information from the 'Search' call and are shown in Table 5. As would be expected the mean absolute error is lower than the individual models (£17,800 versus £22,800 for location alone), and the variance score is higher (0.78 compared to 0.70 respectively). The error is still relatively high at £17,800, which indicates there is more details to the property price than those given to the model. To illustrate these differences between predictions, the difference between predicted and actual house prices is shown in Figure 15.

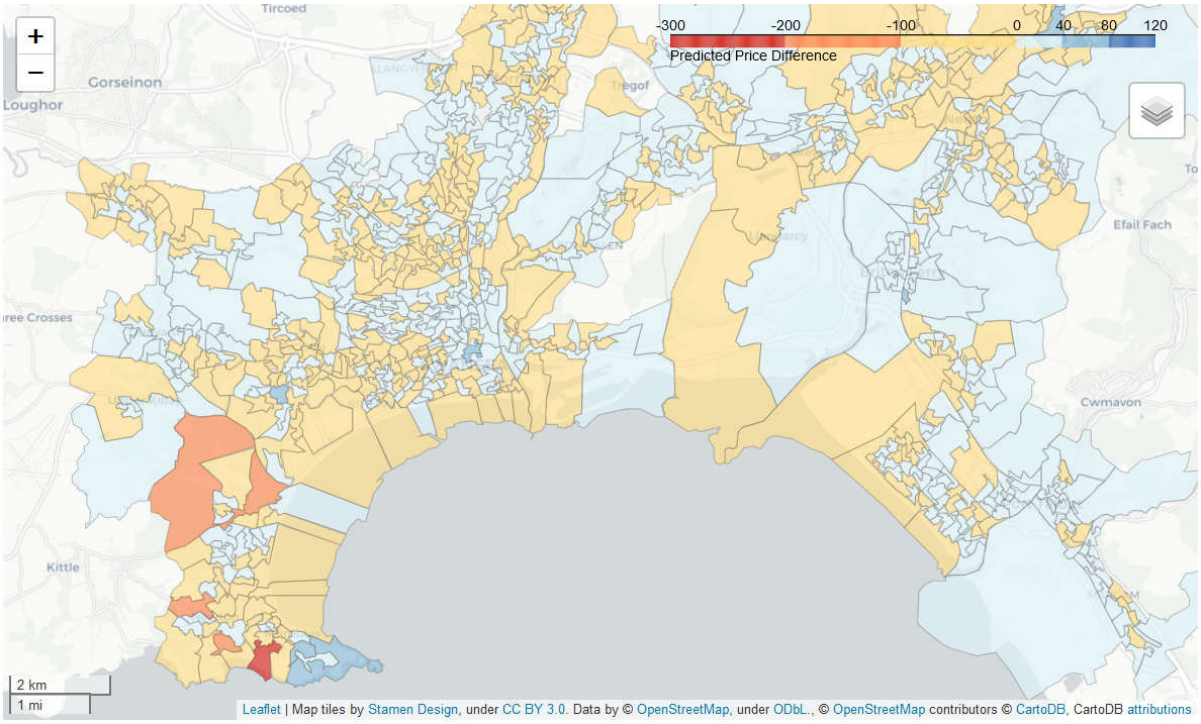


Figure 15. Map of Swansea showing the difference in Census regions between predicted property price based on property details and location details. Using the RF model. A negative value indicates the model underestimates the measured property price.

4 Discussion

4.1 The Selected Areas

To help understand the data, more detail of the selected regions is shown below.

Table 7. Details about the selected areas, including model predictions and the most common venues in different categories from clustering.

Area	GEOGRAPY CODE	House Price	Property Details Predictions	Location Predictions	Location and Property Predictions	No. of rooms effect	Location effect
Mumbles And Newton SA3 4Q	W00004422	£611,000	£328,000	£360,000	£355,000	£90,000	£162,000
Penderry SA5 5E	W00004441	£66,000	£110,000	£74,000	£103,000	£41,000	-£7,800
Morriston North SA6 6Q	W00004348	£126,000	£127,000	£134,000	£127,000	£16,000	£1,000
Uplands SA2 0A	W00004643	£182,000	£173,000	£178,000	£179,000	£14,000	£19,000
Maritime quarter SA1 1S	W00010253	£181,000	£160,000	£178,000	£173,000	£16,000	£80,000

Area	GEOGRAPY CODE	Trending 1st	Trending 2nd	Trending 3rd	Trending 4th	Trending 5th
Mumbles And Newton SA3 4Q	W00004422	Café	Coffee Shop	Restaurant	Grocery Store	Clothing Store
Penderry SA5 5E	W00004441	Furniture / Home Store	Pizza Place	Clothing Store	Coffee Shop	Shopping Plaza
Morriston North SA6 6Q	W00004348	Forest	Pub	Gym / Fitness Center	American Restaurant	Other Repair Shop
Uplands SA2 0A	W00004643	Fast Food Restaurant	Pub	Park	Sushi Restaurant	Pool Hall
Maritime quarter SA1 1S	W00010253	Pub	Grocery Store	Supermarket	Bar	Harbor / Marina

Area	Sights 1st	Sights 2nd	Sights 3rd	Sights 4th	Sights 5th
Mumbles And Newton SA3 4Q	Café	Coffee Shop	Restaurant	Fish & Chips Shop	Mediterranean Restaurant
Penderry SA5 5E	Furniture / Home Store	Arts & Crafts Store	Grocery Store	Clothing Store	Electronics Store
Morriston North SA6 6Q	Gym / Fitness Center	Pub	Grocery Store	Forest	Picnic Area
Uplands SA2 0A	Park	Pub	Fast Food Restaurant	Diner	Sandwich Place
Maritime quarter SA1 1S	Pub	Grocery Store	Supermarket	Bar	Movie Theater

Area	Outdoors 1st	Outdoors 2nd	Outdoors 3rd	Outdoors 4th	Outdoors 5th
Mumbles And Newton SA3 4Q	Beach	Indoor Play Area	Park	Playground	Athletics & Sports
Penderry SA5 5E	Gym / Fitness Center	Athletics & Sports	Mountain	Paintball Field	Park
Morriston North SA6 6Q	Forest	Gym / Fitness Center	Scenic Lookout	Paintball Field	Park
Uplands SA2 0A	Park	Athletics & Sports	Basketball Court	Nature Preserve	Paintball Field
Maritime quarter SA1 1S	Waterfront	Harbor / Marina	Botanical Garden	Plaza	Castle

Area	Shops 1st	Shops 2nd	Shops 3rd	Shops 4th	Shops 5th
Mumbles And Newton SA3 4Q	Clothing Store	Grocery Store	Furniture / Home Store	Bridal Shop	Construction & Landscaping
Penderry SA5 5E	Furniture / Home Store	Clothing Store	Shopping Plaza	Electronics Store	Cosmetics Shop
Morriston North SA6 6Q	Newsagent	Grocery Store	Accessories Store	Mobility Store	Paper / Office Supplies Store
Uplands SA2 0A	Grocery Store	Pharmacy	Laundromat	Business Service	Newsagent
Maritime quarter SA1 1S	Clothing Store	Grocery Store	Supermarket	Newsagent	Event Service

4.2 Effect of number of rooms

One way to increase a property's value that is available to most home-owners is increasing the number of rooms in a property, which can be done by converting an attic or building an extension. Decisions on doing this will be influenced by several factors, and one of the most important is the influence it will have on the property sales price. An indication of the effect of the number of rooms was found by using the Random Forrest (RF) model and shown in Figure 16.

To calculate the value, for each location a range of number of rooms is given to the model whilst maintain all the other features (property details and location), which gives the change in the price of a property with number of rooms (Figure 17). The RF model is based around the use of the sigmoid function, which switches between two values. Due to this the price increase per extra room is not easily found, so instead the difference between the lower and upper house price values is taken and plotted in the figure.

The data should be seen as a guide for homeowners in the regions considering increasing the rooms, or for assessing whether to buy a property. The main trend seen in the figure, which may be expected, is that there is a reasonable correlation between the average property price in a region and this room effect.

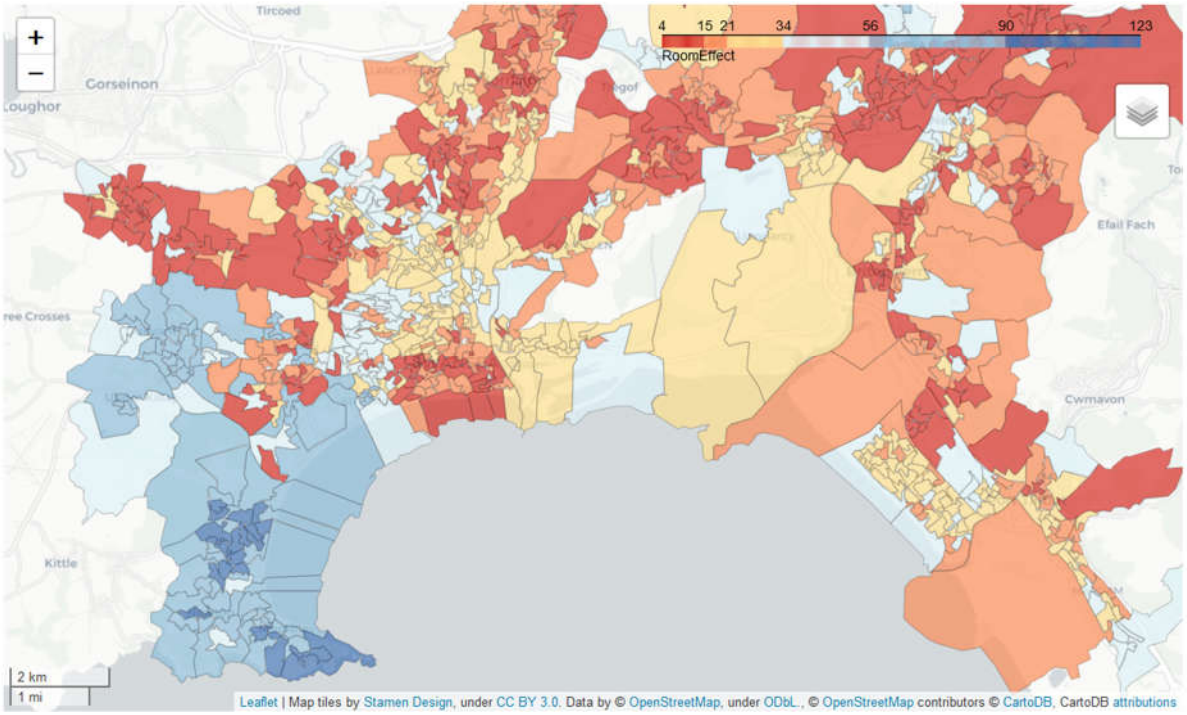


Figure 16. Map of Swansea showing the difference in Census regions for the effect of the number of rooms in a property. Using the RF model.

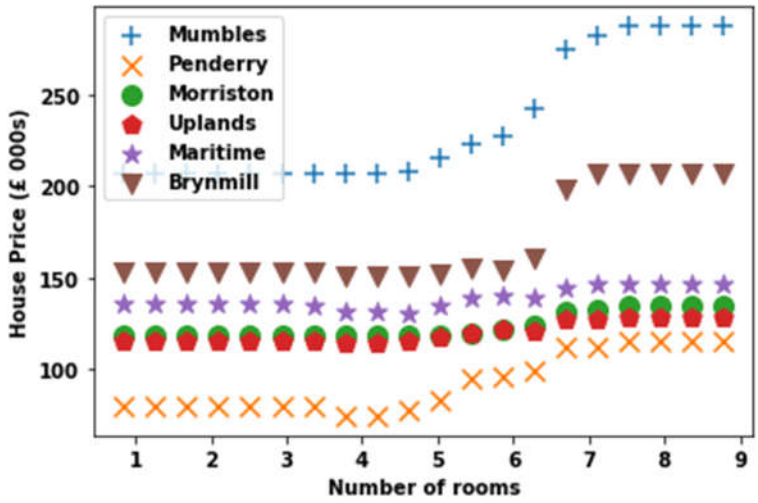


Figure 17. The effect of number of rooms on property prices in the selected regions. The property effect is the difference between the maximum and minimum prices.

4.3 Quantifying the Location Effect

In previous sections it was shown that the location of a property has a greater influence on the price than its property details. To quantify the location effect, a value was obtained of the difference in the average houses price based on its location (Figure 18). This again uses the RF model but this time the property details taken for each region are the average across the districts and using their own location details. The value displayed is the average house price minus the price of the average property in that location.

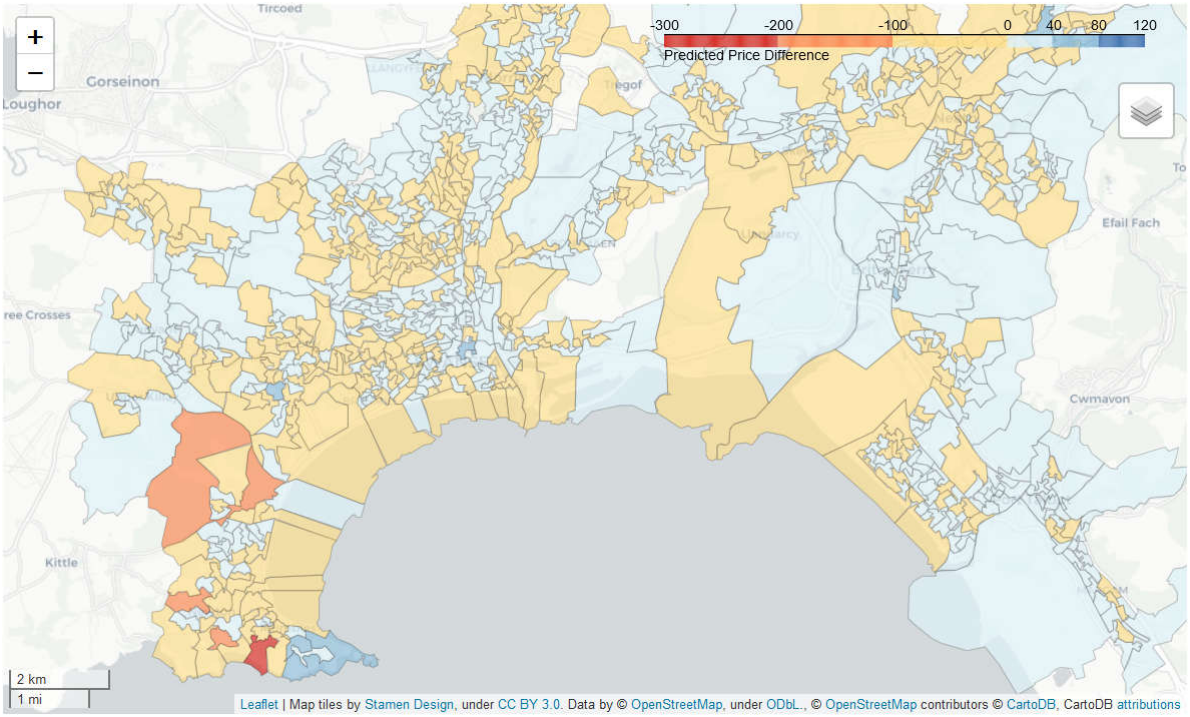


Figure 18. Map of Swansea showing the difference in Census regions for the effect of the number of location of a property. Using the RF model. The value displayed is the average house price minus the price of the average property in the individual locations. Lower values indicate expensive locations.

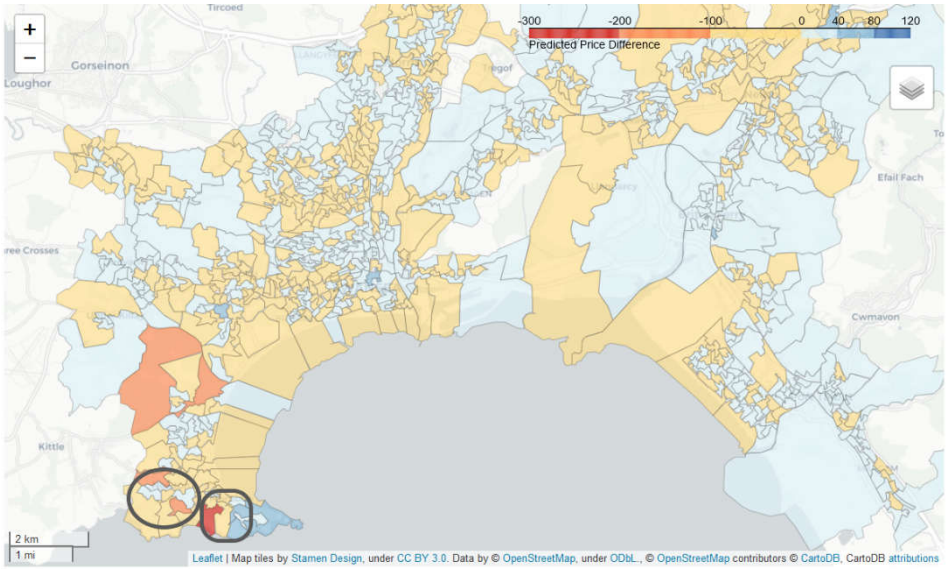


Figure 19. Map of Swansea showing the difference in Census regions between predicted property price based on property details and location details. Using the RF model. A negative value indicates the model underestimates the measured property price. Circles show neighbouring regions with large variations in accuracy of property price predictions.

4.4 Is the model and data a good predictor of property prices?

The predictions of the property details and location model, have an error of £18,000. The error can be low in some areas but considerable in others. If the areas with large property prices were excluded, Gower and Marina, the model may be reasonably accurate but with these areas included it is a poor predictor. The level of predictions is of the level of someone with a reasonable understanding of Swansea property prices but would not compete with valuations from estate agents or from those on Zoopla [22].

The success of the model could be greatly improved by incorporation of crime statistics and deprivation, both of which are readily available. However, refinement is probably required on the property details and perhaps more importantly the location details provided. It is worth recalling that using the latitude and longitude performs slightly better than using the location information from the 'search' function. So better incorporation of the location details using foursquare or something similar would greatly improve the model.
These modifications could include including the star rating of venues, taking account of the surrounding ruralness, distance to beach and not just one of several recognised beach venues along Swansea Bay. Alternatively, splitting the model into two, one for the Gower and another for the rest could help. In addition, A comparison of the model on a different area, such as Cardiff, would help to understand the underlying trends about what contributes to the location effect.
The data and models are instead best used as a guide for potential buyers/renters or current homeowners.

4.5 Getting more out of clustering

The way the clustering and 'explore' location details are used could be refined to get more from this data. In this form they act as a series of maps where differences in the area can be explored from the interactive folium maps. One way to do this would be to convert them into an algorithm that suggests areas to a user based on the wight they put on certain venue types. For example, a user is asked how important different venue types are to them and the model suggests certain areas.

5 Conclusion

A model to predict property prices in the Swansea region based on details of the property and location relative to nearby venues, was shown to be reasonably accurate. Although, large errors particularly in the West were found in house price predictions.
When broken down to using either the property details or the location, the location was found to be a more accurate indicator of property prices. The characterising of location using the distance and frequency of nearby venues appears to be a significant limitation on the model, as just using longitude and latitude information in the model gave more accurate results. This detail along with not accounting for crime or deprivation would significantly improve predictions. But because of this, and other factors, the reasons why a location is desirable or not has not been established.
Clustering locations based on nearby venues offers a good visual tool for understanding an area. But to make it effective for potential homeowners or others exploring Swansea housing requires more work.

6 References

[1] https://en.wikipedia.org/wiki/Swansea
[2] https://en.wikipedia.org/wiki/Gower_Peninsula
[3] https://en.wikipedia.org/wiki/Swansea_Marina
[4] https://medium.com/@briskat/england-wales-population-density-heat-map-26a28a2b6091
[5] https://en.wikipedia.org/wiki/Location,_Location,_Location
[6] https://en.wikipedia.org/wiki/2011_United_Kingdom_census
[7] https://www.nomisweb.co.uk/census/2011
[8] https://www.nomisweb.co.uk/census/2011/bulk/r2_2
[9] https://foursquare.com/
[10] https://borders.ukdataservice.ac.uk/easy_download.html
[11] https://github.com/dMaterialia/SwanseaProperty
[12] https://nbviewer.jupyter.org/
[13] https://developer.foursquare.com/docs/build-with-foursquare/categories/
[14] https://developer.foursquare.com/docs/api-reference/venues/search/
[15] https://developer.foursquare.com/docs/api-reference/venues/explore/
[16] https://python-visualization.github.io/folium/
[17] https://docs.microsoft.com/en-us/azure/machine-learning/how-to-select-algorithms
[18] https://scikit-learn.org/stable
[19] https://census.ukdataservice.ac.uk/get-data/related/deprivation.aspx
[20]https://nbviewer.jupyter.org/github/dMaterialia/SwanseaProperty/blob/main/ClusterPlots.ipynb
[21]https://nbviewer.jupyter.org/github/dMaterialia/SwanseaProperty/blob/main/ChoroPlotting.ipynb
[22] https://www.zoopla.co.uk/