HM Land Registry Open Data Report

# **Introduction**

This document serves to show a brief inspection of data provided by HM Land Registry Open Data. The project initially involved cleaning the data available, including archived data for transaction, which suffered from misnamed files and csv files without header information.

Next, the defining questions were drafted. In an ideal world the defining questions would be to use the data to estimate the worth of each property in the UK, using a machine learning algorithm and some auxiliary datasets that have some information about homes that have not changed deeds in the past 25 years, the economic status of the UK at the time of sale, as well as physical characteristics such as dimension, number of bedrooms, etc. I find this question interesting as it would be invaluable to an internet based property broker to provide an accurate estimate of what a potential customer’s property is worth when looking for the price range of another house to move to.

However, to make the best use of available resources, a much simpler question was asked. How does house pricing differ between the 4 countries that make up the UK? This question was explored through analysis of how the average house price has changed across the UK over the past 50 years. Then a look at what the mean price of housing types across the UK was done as well. Finally we drill down to the county level to see how changes in the number of sales have been affected alongside changes in population.

Assumptions that were made when making mean calculations over time periods that the difference in the number of homes sold per month were negligible, which is quite crude and could be alleviated through using the “sales volume” column of the House Price Index (HPI) to weight the averaging.

## Issues Faced

Two main obstacles came up when processing the information. The main one was the unexpectedly large amount of time it took to reformat archived data so that it could pair up with the production data. Although I did not use this data I feel it was a good scenario for me to show an ability to clean data.

The second issue was that I was treating the HPI dataset as being a list of counties due to misunderstanding what the header information was referencing when reading its description on the website. This resulted in a quick change of direction to looking at countries in the UK so that some analysis could be shown, with the time remaining.

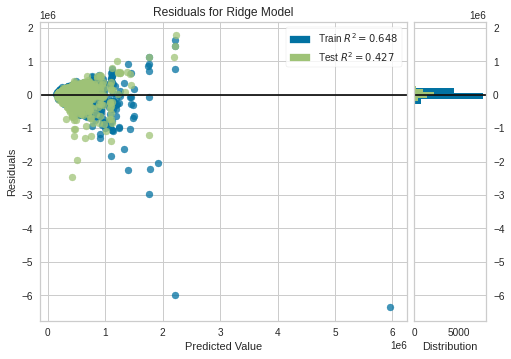
# Machine Learning

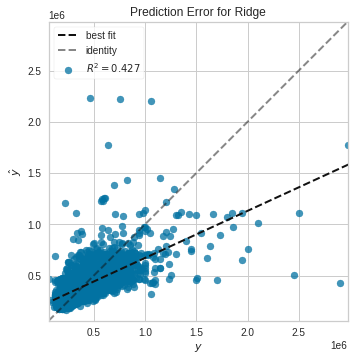
Using PyCaret, a reasonable regression model was chosen to model the freehold detached housing market in Southampton between 2010 and 2021. This subset of the overall price paid dataset was chosen due to it’s size being easily modelled, and redundant information such as constants (like city and council) as well as noise (such as primary and secondary number numbering) being removed to simplify the model further.

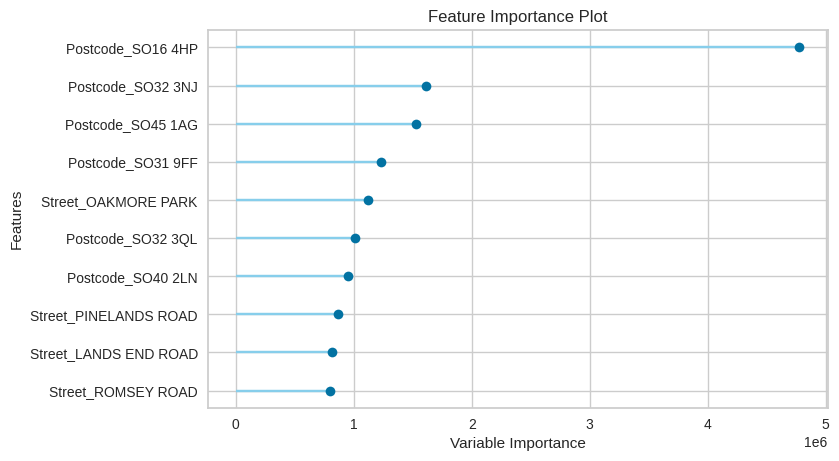
The machine learning model used was Ridge Regression, as although Bayesian Ridge was overall more accurate (pycaret comare\_models output below), the difference was very small compared to the time spent on each fold.

|  | **Model** | **MAE** | **MSE** | **RMSE** | **R2** | **RMSLE** | **MAPE** | **TT (Sec)** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **br** | Bayesian Ridge | 91440.8694 | 38606536191.4087 | 194100.4931 | 0.2940 | 0.3025 | 0.2456 | 120.0200 |
| **ridge** | Ridge Regression | 91450.1979 | 38684581888.0000 | 194313.6771 | 0.2926 | 0.3029 | 0.2471 | 1.1133 |
| **lasso** | Lasso Regression | 92469.0026 | 41403433642.6667 | 200518.4062 | 0.2460 | 0.3052 | 0.2392 | 11.5133 |

The training results were not incredibly impressive, with the validation scoring an R2 of only 0.427. As is expected, out of the 3 columns; postcode, street name, and date of transaction, the postcode was the most influential. This differed from prior models that scored worse, likely due to having too many dimensions and thus becoming very shallow to be able to process, which resulted in the date taking prime spots.

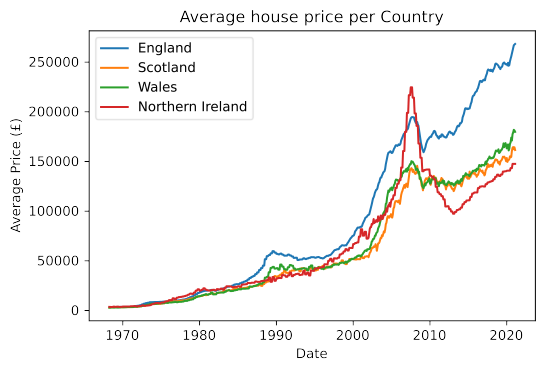
Figure 1

Figure 2

Figure 3

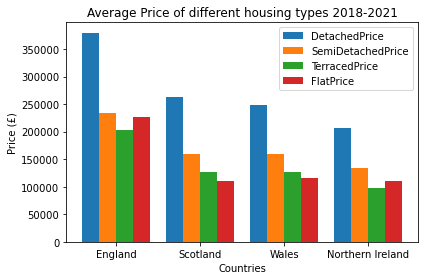
# Analysis

The monthly average house price in the UK shows England starting to break away from the other countries in the late ‘80s, with Northern Ireland briefly overtaking England right before the 2008 recession. At the 2008 recession all countries saw a rapid decrease in property value after a few years of rapid growth. After the recession, England has grown much faster than the other countries, with all countries seeing a sharp increase in property value during 2020, possibly due to the covid pandemic.

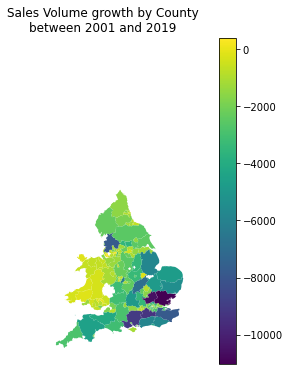
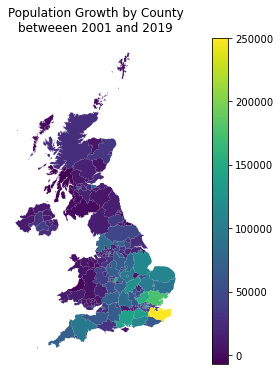
Figure 4: Average house price between January 1970 and February 2021

When breaking down the different kinds of property into detached, semi-detached, terraced, and flat, it is revealed that despite all countries having detached homes as their most expensive property type, England’s non-detatched homes cost nearly as much as detached homes in Scotland and Wales, and all but terraced housing costs more than detached homes in Northern Ireland.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Country** | **Detached (£)** | **Semi-Detached (£)** | **Terraced (£)** | **Flat (£)** |
| England | 379456 | 234352 | 202467 | 225878 |
| Scotland | 263394 | 159851 | 127513 | 109938 |
| Wales | 248478 | 159054 | 127623 | 115050 |
| Northern Ireland | 206817 | 134412 | 97624 | 109666 |

Figure 5: Distribution of property types per country in the UK, exact values seen in the table above.

A map of the UK (figure 3) demonstrates the change in population per county between 2001 and 2019 (figure 3a). To the right is the change in property value between 2001 and 2019 (figure 3b).



Although the UK has a growing population across all 4 countries, when looking at England and Wales the volume of properties sold in 2019 was lower in most counties despite this population growth.

# Conclusion

The regression model was a partial success, although the R2 value leaves much room for improvement. With extra dimensions likely to lead to higher levels of accuracy, such as using GDP values or population growth to provide more material for the regression algorithm to work with.

The data shows that on average England is a much more expensive place to buy a property than the other countries in the UK. Although further research would be able to show any skew in the data such as homes near London tending to cost much more than similarly sized homes in the West of England. Along these lines, an interesting area for further research would be to investigate how property value decreases with distance from a major city, which could be achieved using the Price Paid dataset using postcode information to estimate the distance from different cities around the England and Wales.

Finally, the decrease in house sales despite a growth in population likely means that there is an increase in renting properties now than in 2001, possibly due to more people in a county driving up the cost of buying, and thus making renting a more lucrative business.