Predicting London House Prices

Kenny Hunt

1. Introduction

1.1 Background

The London property market is leading the way for regeneration, many property developers choose to invest in London because the stable growth of house prices will pretty much guarantee a solid return on investment. Investors and developers adding capital to areas of London has enabled the regeneration of many London boroughs, this has in turn made many areas in the suburbs of London more desirable to live, which attracts more businesses to trade. This cycle has lead to the constant growth of the London property market. For developers, residential buyers and property investors, London continues to be the flagship property investment hotspot in the UK.

There are many variables that can impact the price of a property in London; size, number of bedrooms, location, proximity to travel links, proximity to other amenities e.g. supermarkets, restaurants, bars.

Price varies wildly between different boroughs and areas of London, for this reason a single borough will be selected for the analysis, although the experiment will be applicable to all areas. The selected area for this experiment is the London Borough of Ealing. Ealing is generally recognised as a sought after place to own property, this is largely down to the significant amount of regeneration that has been taking place over the last decade throughout the entire borough, coupled with its superb transport links into the city.

1.2 Problem

The project will identify and analyse the impact of different variables - both structural and environmental - and the effects they have on the price of a house in London Borough of Ealing. Once the key variables are identified, the project aims to utilise this analysis to predict the sales price of future property sales based on both current and historical data.

1.3 Interest

Understanding the key drivers behind property value in a given area presents many potential benefits to property developers, buyers and investors and alike.

2. Data Acquisition and Cleaning

2.1 Data Sources

The project utilises the following data sources:

Name	Description	Extract Method
Zoopla (in lieu of foursquare)	● Current properties on sale in the selected area. ● Historical sales prices of properties sold in the selected area. ● Longitude and Latitude of London postcodes	API
Transport for London	● Locations of Transport Links	Web Scraper to CSV

2.2 Data Preparation

Data scraped from multiple sources was combined into one table for ease of analysis. There was a limitation on the number of records that could be exported through the Zoopla API at any one time, therefore the decision was taken to break the entire borough down into 7 separate postcodes and call the api separately for exact data set, then merge them together into one table.

The data set required steps to cleanse. Firstly there were a number of data quality errors;

2.2.1 Number of Bedrooms

Where the property for sale was a piece of land, this was regarded as a row with zero bedrooms. As it was believed that the number of bedrooms was influential on the total price of the property, it is reasonable to assume that any calculations would be skewed by these figures - therefore any row with zero bedrooms was dropped.

There were a number of properties containing more than 4 bedrooms which were available at very low prices - analysis evidenced that these properties were “shared equity” therefore were filtered from the dataset at source.

2.2.2 Nulls

The data set contained nulls in the post_town column only, these records were left alone as post town was consistent throughout the data set, therefore this feature will be dropped later on.

	%	Total # Records
post_town	3.0	21
street_name	0.0	0
property_type	0.0	0
price	0.0	0
outcode	0.0	0

Fig1: Shows the number and percentage of nulls in the dataset

2.2.3 Price

There were a number of properties with the price set as zero, these are effectively “price upon enquiry” properties, and were therefore removed from the analysis.

2.3 Feature Selection

After cleansing there were 653 records and 28 columns in the data. Many of the columns were deemed irrelevant for our analysis and will require dropping.

Feature	Action	Justification
agent_address	Drop	Irrelevant
outcode	Keep	
num_recepts	Keep	
first_published_date	Drop	Irrelevant
displayable_address	Keep	
details_url	Drop	URL Irrelevant
num_bedrooms	Keep	
price	Keep	Target Variable
post_town	Drop	Identical throughout analysis
price_modifier	Drop	Irrelevant
property_type	Keep	
street_name	Keep	
floor_plan	Drop	Irrelevant
image_url	Drop	URL Irrelevant
last_published_date	Drop	Date Irrelevant

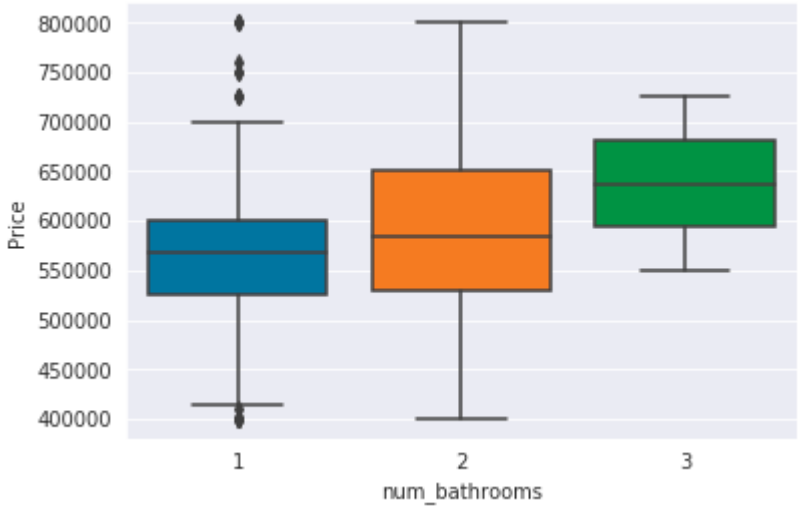
latitude	Keep	
listing_id	Drop	Primary key not needed
listing_status	Drop	Filtered to ‘Sale’ at source
longitude	Keep	
num_bathrooms	Keep	
country_code	Drop	Identical throughout analysis
agent_name	Drop	Irrelevant
agent_logo	Drop	Irrelevant
agent_phone	Drop	Irrelevant
category	Keep	
county	Drop	Identical throughout analysis
country	Drop	Identical throughout analysis

3. Exploratory Data Analysis

3.1 Relationship between property price and number of bathrooms

In order to visualise the relationship between number of bathrooms and price, the data was plotted into a box-plot. This is a great way to get an understanding of the data at a high level and identify if any patterns are emerging that require investigation.

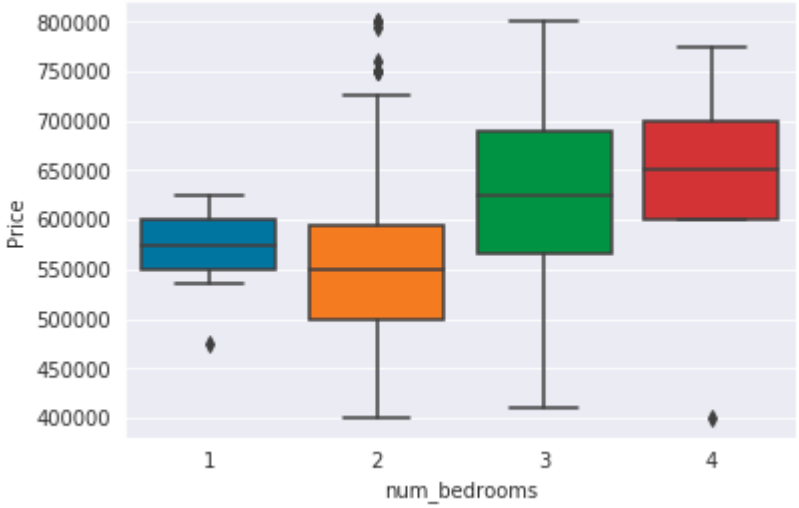
The initial observation was that there were a number of properties with a data quality error - bathrooms set to zero - these records were removed and the analysis was reproduced.



It's clear from the analysis that there is a strong linear relationship between number of bathrooms and total price of property. The average one-bathroom property costs approximately £560,000, whilst a property with an additional bathroom costs on average £590,000. This means a second bathroom costs on average £30,000! Similarly there is a large difference in price between 2 and 3 bathroom properties. I believe this is because once we start looking at properties with 3 bathrooms, the size of the property increases drastically - therefore an increase in price is expected.

3.2 Relationship between property price and number of bedrooms

These are the two most commonly expected variables to contain a linear relationship. I.e. we would expect the property with the most bedrooms to be the most expensive, and the property with the least number of bedrooms to be the least expensive. The analysis below indicates that we are partially correct in this hypothesis.



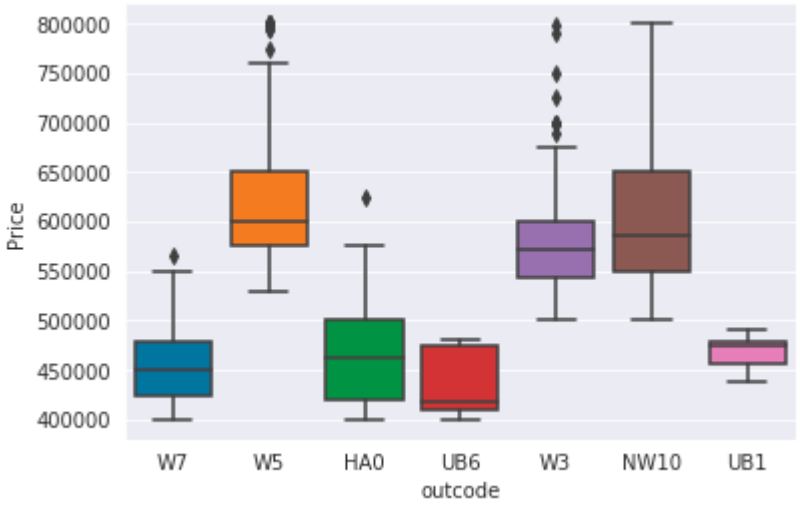
The important observation here was that the average price of a 1-bed property was higher than the average price of a two-bed property. This could be down to a number of reasons. In order to investigate further it was necessary to find out more about these specific one bed properties.

	category	displayable_address	latitude	longitude	num_bathrooms	num_bedrooms	num_recepts	outcode	price	property_type	street_name
69	Residential	Filmworks Walk, Ealing W5	51.513775	-0.306980	1	1	0	W5	624950.0	Flat	Filmworks Walk
70	Residential	Longfield Avenue, Dickens Yard W5	51.513740	-0.306087	1	1	1	W5	624000.0	Flat	Longfield Avenue
76	Residential	Vista House, 2 New Broadway W5	51.513306	-0.306348	1	1	1	W5	600000.0	Flat	Dickens Yard
83	Residential	Bond Street, Ealing W5	51.513775	-0.306980	1	1	0	W5	599950.0	Flat	Bond Street
102	Residential	Fitzroy House, Dickens Yard, Ealing W5	51.513520	-0.306917	1	1	1	W5	575000.0	Flat	Dickens Yard
103	Residential	Vista House, London W5	51.513740	-0.306087	1	1	0	W5	575000.0	Flat	Vista House
141	Residential	Northwick Road, London HA0	51.538740	-0.302232	1	1	0	HA0	475000.0	NaN	Northwick Road
235	Residential	Long Island House, 42 Warple Way, Acton W3	51.504204	-0.255261	1	1	1	W3	550000.0	Flat	Island House
238	Residential	Warple Way, London W3	51.504204	-0.255261	1	1	1	W3	550000.0	Flat	Warple Way
242	Residential	Long Island House, 42 Warple Way, London W3	51.504204	-0.255261	1	1	1	W3	550000.0	Flat	42 Warple Way London
247	Residential	Portal Way, London W3	51.522053	-0.261843	1	1	1	W3	536000.0	NaN	Portal Way

Its clear from the analysis that these properties were all located within the same postcodes, W3 and W5. The next piece of analysis will determine if location (or postcode) has an impact on a property price.

3.3 Relationship between property price and postcode

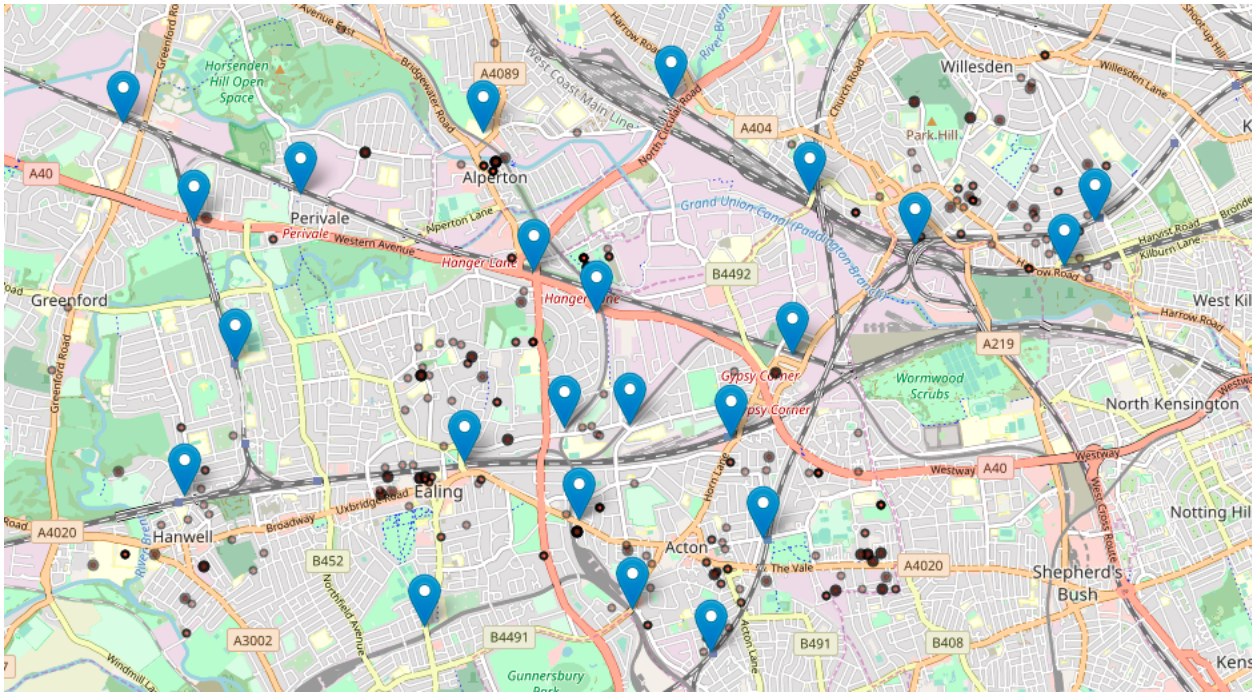
This piece of analysis indicates that there is clearly a correlation between postcode (outcode) and price of property. We observe that properties in W3, W5, NW10 are all significantly more expensive on average than properties in the other postcode districts.



This backs up the hypothesis made in the previous statement, that the one-bed properties are more expensive because our sample dataset contains one bed properties in expensive neighbourhoods.

3.4 Relationship between property price and location to transport links

In order to analyse adhoc the relationship between localisation to transport links and housing prices. In order to do this, CSV data was extracted from Transport for London and plotted on a Folium map (blue markers) alongside the individual properties in the analysis. On selecting individual group of properties that are close to and far away from tube and train stations, its apparent that properties closer to tube stations cost aproximately £40,000 more.



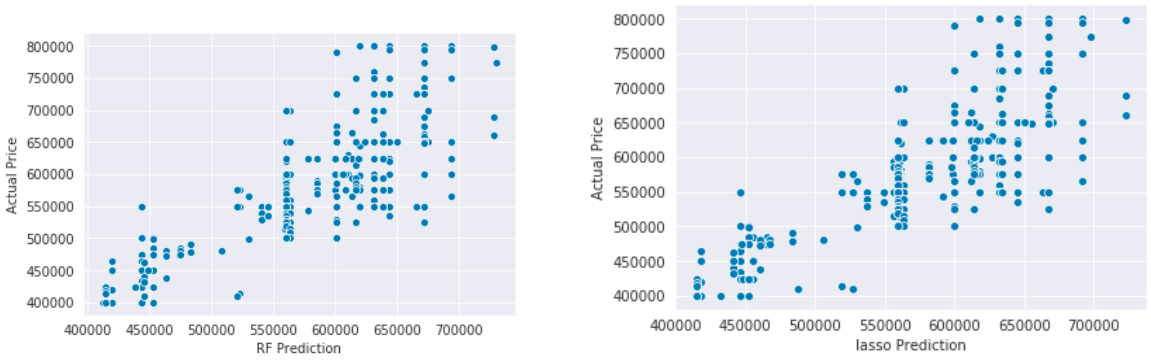
4. Predictive Modelling

4.1 Regression

In order to help predict current house prices based on the variables in the dataset, a machine learning model was selected and trained. Defining house prices based on a number of variables is a Regression problem as we are trying to predict a continuous variable based on a number of static variables.

4.2 Model Selection

Two different models were trained. A Lasso and a Random forest model. As visible in the below figures, the models performed similarly - they both performed adequately given the small size of data set provided. In order to make the models more accurate, they will require more data. We can visualise the performance of a regression algorithm by plotting the actual house prices provided in the original data set, by the predicted prices generated by the models created. Ideally the results would be a diagonal line.



Figs - Random Forest and Lasso Performance in Predicting House Prices

5. Conclusions

In conclusion the experiment was considered a success - valuable insight surrounding London house prices were gained, and the analysis can be re-applied to any housing dataset scraped from Zoopla - we can perform the same analysis for any part of London using the same python code.

6. Future applications

In order to take this experiment to the next level, it is required that more data is made available by Zoopla, this will allow for further analysis into the high level insights made in this experiment.

- What makes a specific postcode more expensive? Availability of amenities? Low crime?
- Are one beds generally more expensive that two beds? I believe a larger dataset would prove this analysis to be inconclusive surrounding number of bedrooms.
- What is the difference between the price of a new build or resale property?