**British Property Sales – Where can you afford to buy a house?**

# Abstract

In recent years, we find upward trend of property price in almost all major cities of the world. People buy properties and the price of the property is an important attribute to consider before buying the property. Here we are going to discuss ‘British Property’ sales. The UK’s [Price Paid Data](https://www.gov.uk/government/collections/price-paid-data) contains virtually all of the country’s residential property sales, with only a [few exceptions](https://www.gov.uk/guidance/about-the-price-paid-data). Price Paid Data tracks property sales in England and Wales submitted to Land Registry for registration. The amount of time between the sale of a property and the registration of this information with Land Registry varies. It typically ranges between 2 weeks and 2 months.

The objective of this model is to suggest a place where an individual can buy an affordable house or invest money on properties in United Kingdom. And also an comparative study on how the property price of a region is getting changed year on year.

# Dataset

The data is derived from all residential property transactions registered with Land Registry. While properties are self-explanatory, transactions relate to the number of sales which have been returned. The two numbers will differ as a property can transact more than one in the time period searched. Therefore, there is a possibility that 100 properties returned with 105 transactions, meaning that one or more of the properties has changed ownership multiple times.

**Price Paid Data (PPD) excludes:**

* sales that have not been lodged with Land Registry.
* sales that were not for full market value. By this we mean, the sale of part of a property, a share of a property or the sale of a property at a discount. For examples a transfer between parties on divorce
* transfers, conveyances, assignments or leases at a premium with nominal rent which are:
  + ‘Right to buy’ sales at a discount
  + subject to an existing mortgage
  + to effect the sale of a share in a property
  + by way of a gift
  + under a compulsory purchase order
  + under a court order
  + to Trustees appointed under Deed of appointment
* Vesting Deeds Transmissions or Assents of more than one property

**Explanations of column headers in the PPD:** The goal is to develop a predictive algorithm that will help to determine the affordable property based on the following input variables except ‘Price’ (Output variable). Please keep in mind all attributes below may not influence the Price

| ***Data item*** | ***Explanation (where appropriate)*** |
| --- | --- |
| Transaction unique identifier | A reference number which is generated automatically recording each published sale. The number is unique and will change each time a sale is recorded. |
| Price | Sale price stated on the transfer deed. |
| Date of Transfer | Date when the sale was completed, as stated on the transfer deed. |
| Postcode | This is the postcode used at the time of the original transaction. Note that postcodes can be reallocated and these changes are not reflected in the Price Paid Dataset. |
| Property Type | D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other  Note that:  - end-of-terrace properties are included in the Terraced category above.  - ‘Other’ is only valid where the transaction relates to a property type that is not covered by existing values. |
| Old/New | Indicates the age of the property and applies to all price paid transactions, residential and non-residential. Y = a newly built property, N = an established residential building |
| Duration | Relates to the tenure: F = Freehold, L= Leasehold etc. Note that Land Registry does not record leases of 7 years or less in the Price Paid Dataset. |
| PAON | Primary Addressable Object Name. If there is a sub-building for example the building is divided into flats, see Secondary Addressable Object Name (SAON). |
| SAON | Secondary Addressable Object Name. If there is a sub-building, for example the building is divided into flats, there will be a SAON. |
| Street |  |
| Locality |  |
| Town/City |  |
| District |  |
| County |  |
| PPD Category Type | Indicates the type of Price Paid transaction. A = Standard Price Paid entry, includes single residential property sold for full market value. B = Additional Price Paid entry including transfers under a power of sale/repossessions, buy-to-lets (where they can be identified by a Mortgage) and transfers to non-private individuals.  Note that category B does not separately identify the transaction types stated. Land Registry has been collecting information on Category A transactions from January 1995. Category B transactions were identified from October 2013. |
| Record Status - monthly file only | Indicates additions, changes and deletions to the records.(see guide below). A = Addition C = Change D = Delete.  Note that where a transaction changes category type due to misallocation (as above) it will be deleted from the original category type and added to the correct category with a new transaction unique identifier. |

The date field - the date of the transfer deed lodged with us for registration.

**Additional Price Paid Data and VAT**

When the price paid data for a property transaction is captured, the price paid information quoted in the transfer deed lodged for registration is recorded. Some commercial property transactions lodged for registration quote the price paid and make a statement about value added tax (VAT). This statement may include the specific amount of VAT paid or may state that VAT is ‘excluded’ or ‘included’. There will also be occasions when VAT is not mentioned.

When the price paid quoted in a transfer deed differentiates between the consideration and the VAT, the price paid recorded will exclude VAT, eg if a transfer deed quotes the ‘price paid to be £100,000 plus VAT’ only £100,000 will be recorded. However when the consideration is inclusive of VAT, eg the ‘price paid is stated to be £100,000 inclusive of VAT, the transaction is excluded from the dataset. This is because decisions are not made on the percentage of VAT that should be applied to historic transactions.

The process for recording register entries for price paid remains unchanged and will continue to refer to VAT where appropriate.

**Sample Dataset**



Land Registry Price Paid Data contains on an average 100K records for each year. The dataset is a reliable source of house price information and consists of more than 24 million definitive records dating back to January 1995. Information on dataset is available in the below location.

<https://data.gov.uk/dataset/land-registry-monthly-price-paid-data>

# Libraries

pandas for

* + data loading, wrangling, cleaning and manipulation
  + feature selection and engineering
  + descriptive statistics

numpy for

* + array data structure, the primary input for classifiers
  + model comparison
  + matrix manipulation

matplotlib for

* + data visualization

### Data Wrangling and Cleaning

The data comes in the form of csv file which can be easily opened in MS Excel spreadsheets for viewing and can be loaded into pandas dataframes without much issues. All the data are formatted as unicode strings, some of which pandas handles automatically. For example, the ‘Date of Transfer’ are automatically read as pandas date-time objects. Some strings require a little more work to prepare them for the classifiers. ~~To deal with categorical data, dummy variables are created for each category using a patsy formula.~~ NaNs in String data field are handled by filling ‘blank’.

* The field ‘Transaction Unique Identifier’ is a GUID field, has been dropped as it is not used for any calculation.
* The time component from the date field has been dropped as it is not so relevant information. pd.to\_datetime(pp2016['Date Of Transfer'])
* Yes/No fields are converted in to binary integers (1/0).
* Add column ‘Headers’ in the actual Data files for easy data interpretation.

### Data Story