Greater London Housing Prices

Nigel Brown

2020/10/13

Contents

1 Introduction

The goal of this project is to predict house prices in the Greater London Area based on the type of property and the area of London that the property resides in.

The dataset for the project is made from data sourced from HM Land Registry and FreeMapTools. This data has been pre-processed into a dataset titled lhd.rds. The data chosen is a 10 year period 2009-2019. The reason 2020 has been left out of the data is that is an incomplete year at the time of analysis. The data contains 4 different types of property detached; flats/maisonettes; semi-detached and terraced. It doesn't contain the number of rooms or the total area of the property, therefore a 2 bedroom flat is categorized the same as a 4 bedroom flat and a prediction can only be made on the property type of a given district.

The price paid data set from the HM Land Registry is joined together with spatial data to add latitude and longitude for the properties based on their postcodes. This will allow for mapping of the data.

2 Data Wrangling

As a first step data, that is in logical form (true/false) is converted into integers 1 or 0. Next all columns containing character data are factorized.

2.1 Features added to the data

Next a new feature is created based on the outward code contained in a postcode i.e the sub-district and another feature containing the number of times a property was observed to be sold in the 10 year time period is added.

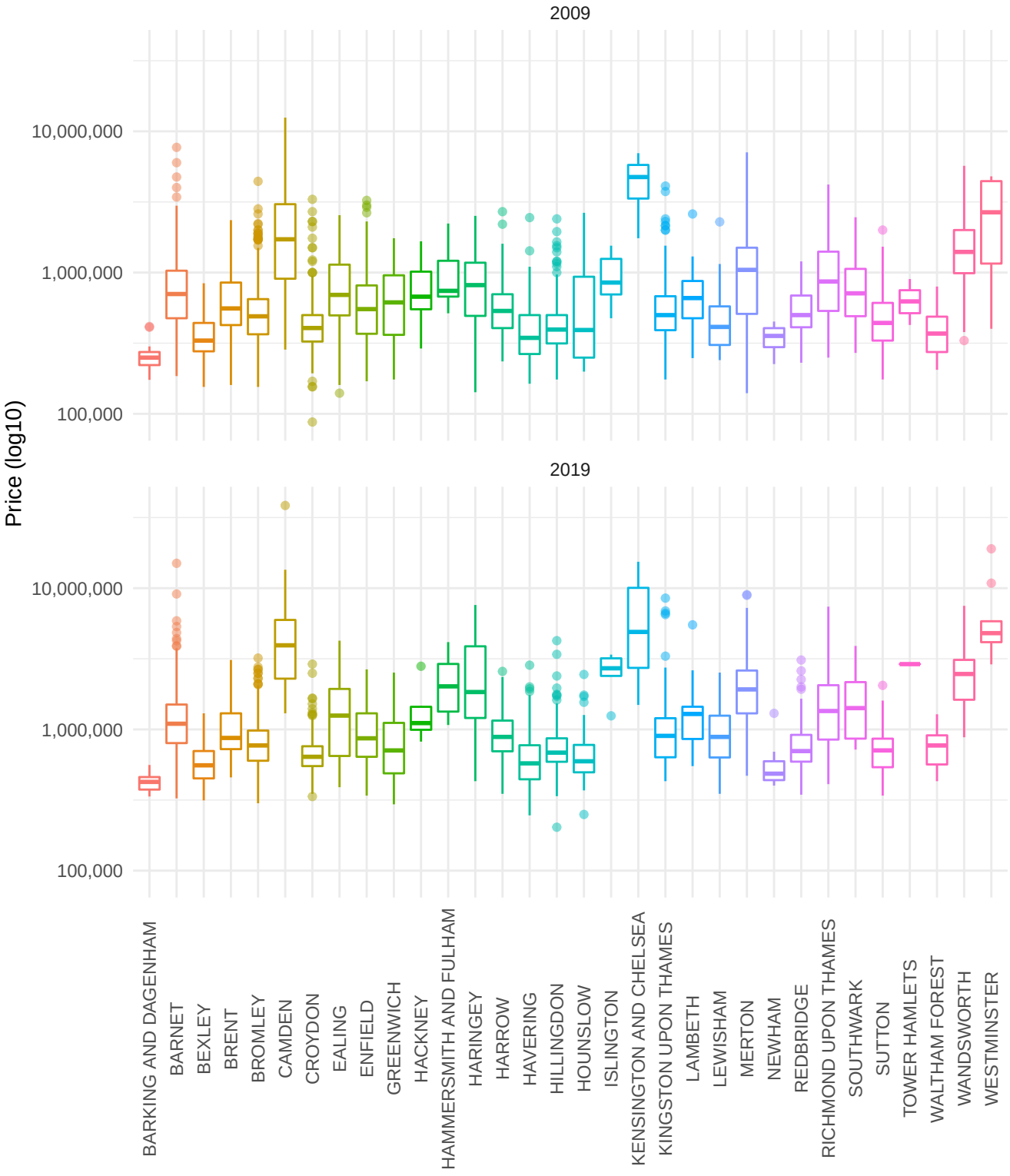
3 Exploritory Data Analysis

Now that the dataset is cleaned, an exploration of the data is performed. The data consists of 477382 rows of data, each row observes a single sale. The are 12 features in the dataset, with price being the outcome feature.

First the price movement for each property type is plotted

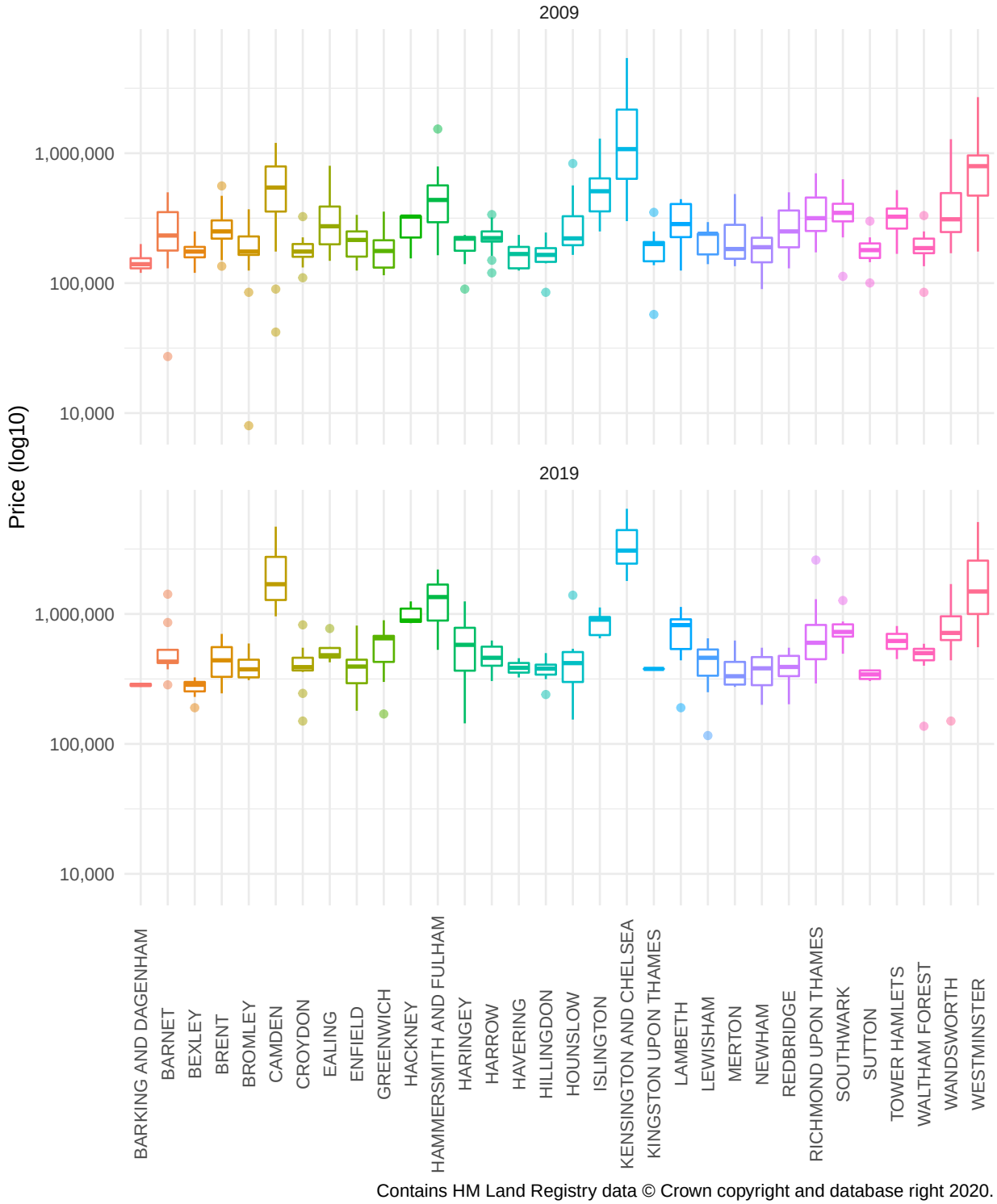
Movement of detached house prices per district 2009 − 2019

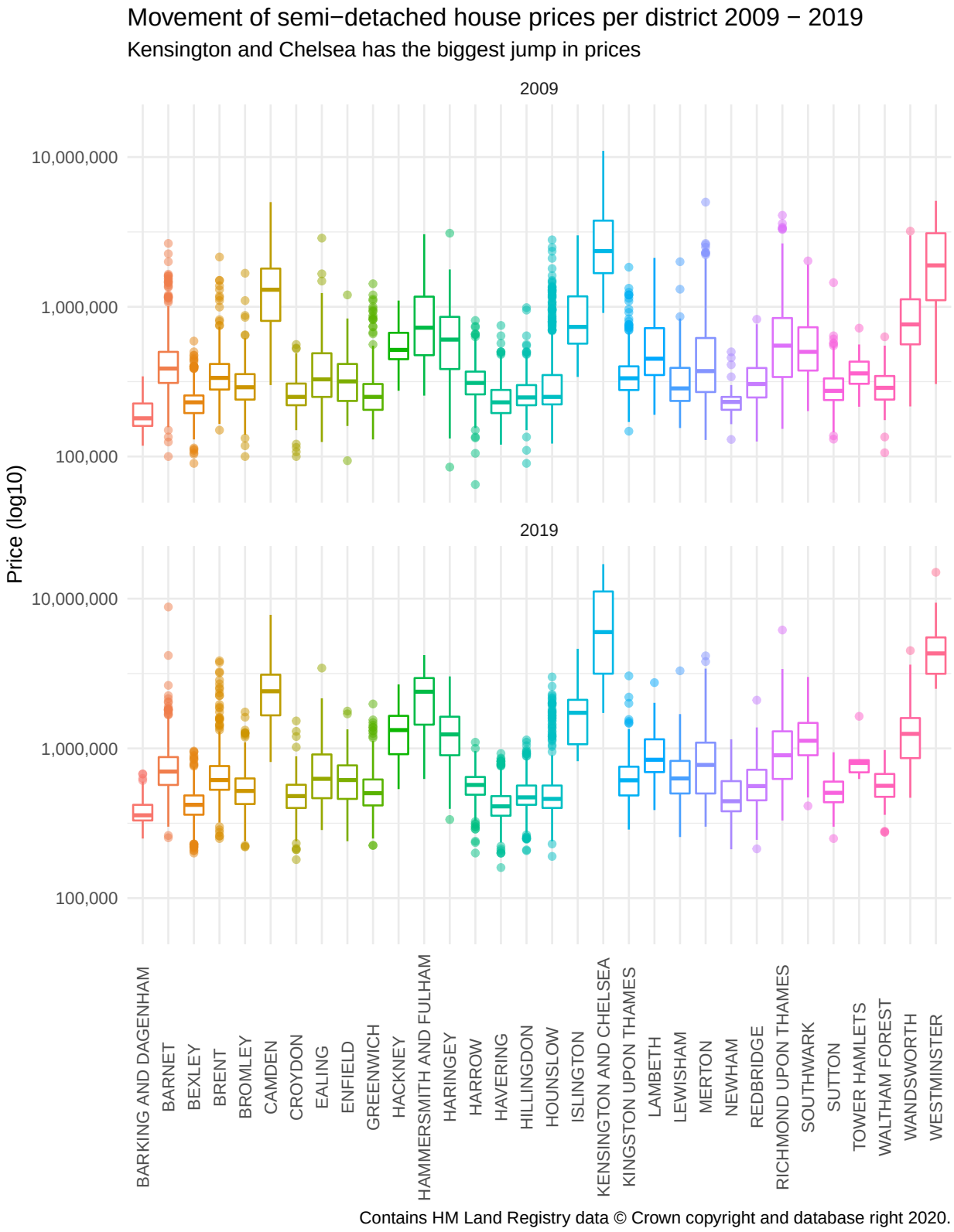
The spread of prices in Hounslow and Westminster has narrowed,
while in Kensington and Chelsea it has widened

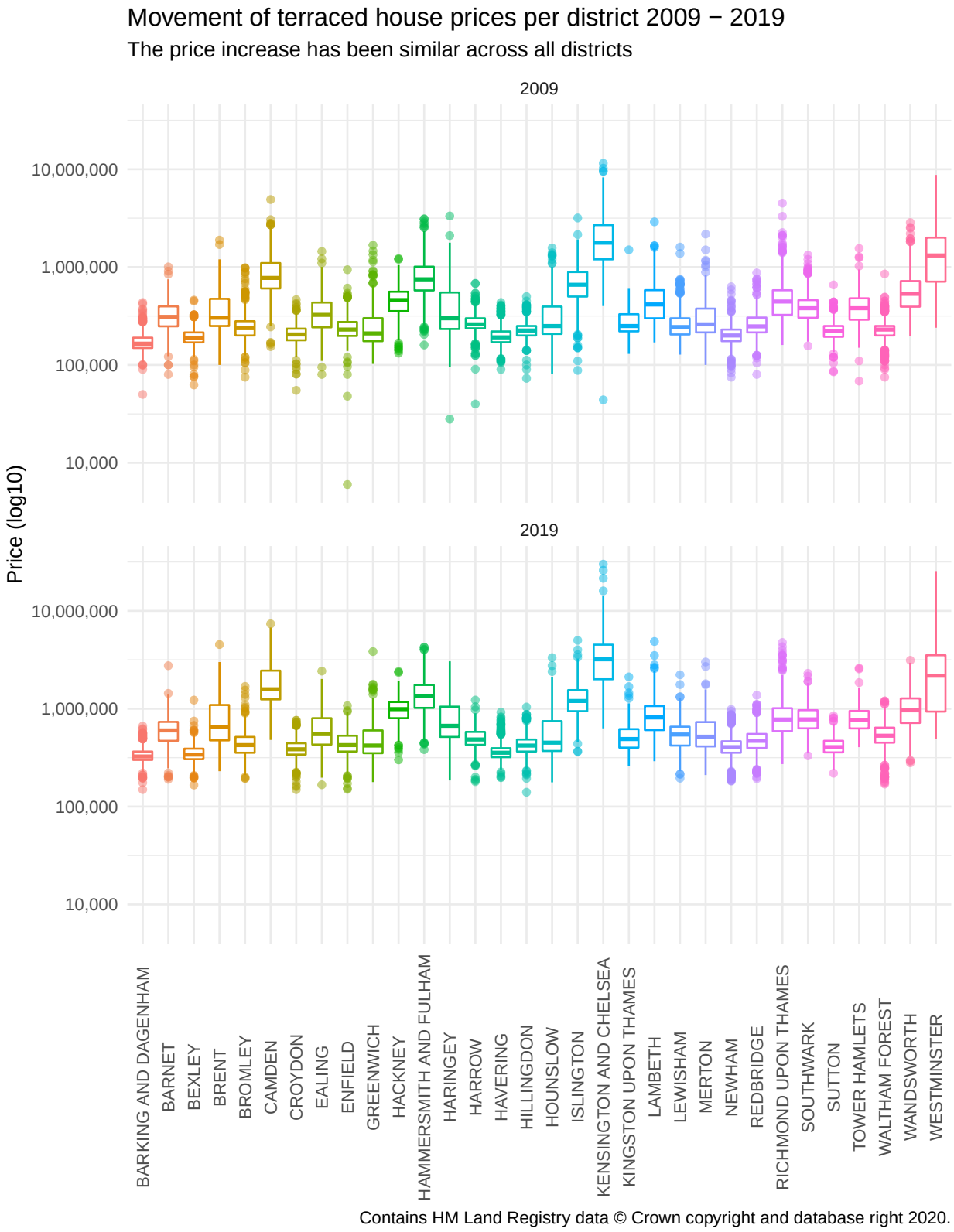


Movement of flat/masonette prices per district 2009 − 2019

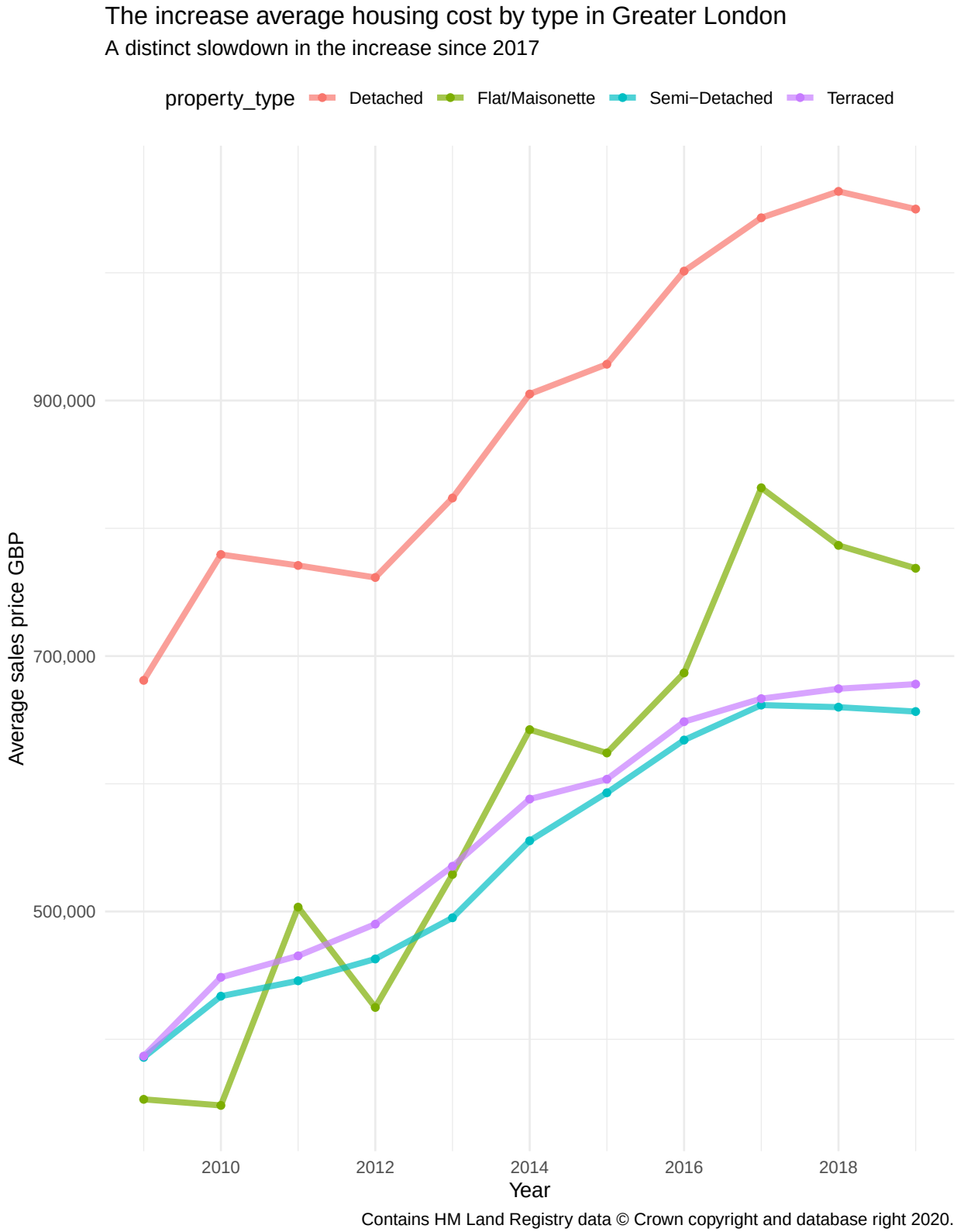
Camden & Kensigton and Chelsea have seen the largest price growth



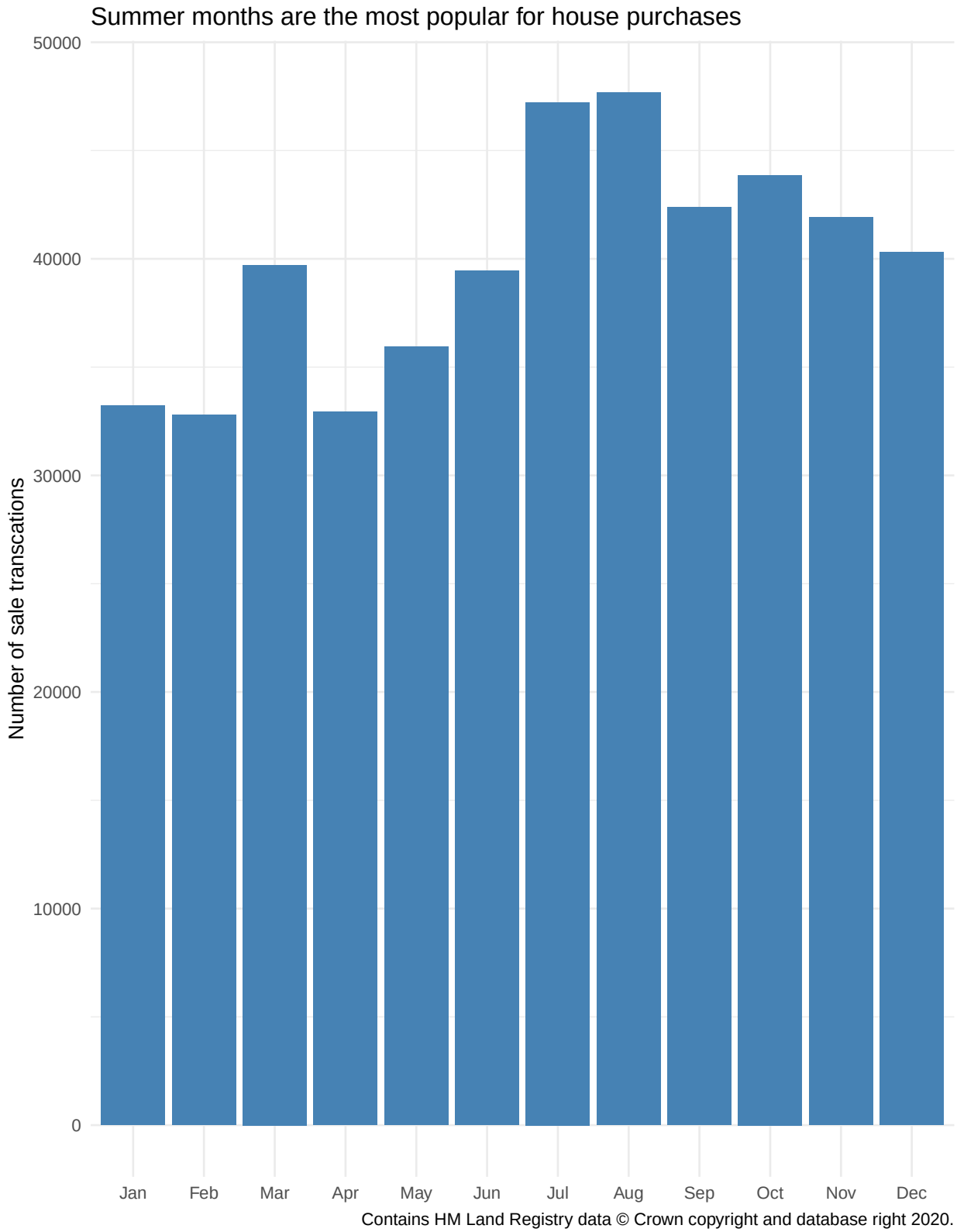

Movement of semi−detached house prices per district 2009 − 2019
Kensington and Chelsea has the biggest jump in prices
2009
2019
Price (log10)
10,000,000
1,000,000
100,000
BARKING AND DAGENHAM
BARNET
BEXLEY
BRENT
BROMLEY
CAMDEN
CROYDON
EALING
ENFIELD
GREENWICH
HACKNEY
HAMMERSMITH AND FULHAM
HARINGEY
HARROW
HAVERING
HILLINGDON
HOUNSLOW
ISLINGTON
KENSINGTON AND CHELSEA
KINGSTON UPON THAMES
LAMBETH
LEWISHAM
MERTON
NEWHAM
REDBRIDGE
RICHMOND UPON THAMES
SOUTHWARK
SUTTON
TOWER HAMLETS
WALTHAM FOREST
WANDSWORTH
WESTMINSTER
Contains HM Land Registry data © Crown copyright and database right 2020.


Movement of terraced house prices per district 2009 − 2019
The price increase has been similar across all districts
2009
2019
Price (log10)
10,000,000
1,000,000
100,000
10,000
BARKING AND DAGENHAM
BARNET
BEXLEY
BRENT
BROMLEY
CAMDEN
CROYDON
EALING
ENFIELD
GREENWICH
HACKNEY
HAMMERSMITH AND FULHAM
HARINGEY
HARROW
HAVERING
HILLINGDON
HOUNSLOW
ISLINGTON
KENSINGTON AND CHELSEA
KINGSTON UPON THAMES
LAMBETH
LEWISHAM
MERTON
NEWHAM
REDBRIDGE
RICHMOND UPON THAMES
SOUTHWARK
SUTTON
TOWER HAMLETS
WALTHAM FOREST
WANDSWORTH
WESTMINSTER
Contains HM Land Registry data © Crown copyright and database right 2020.

3.1 How does the avg price increase year on year ?



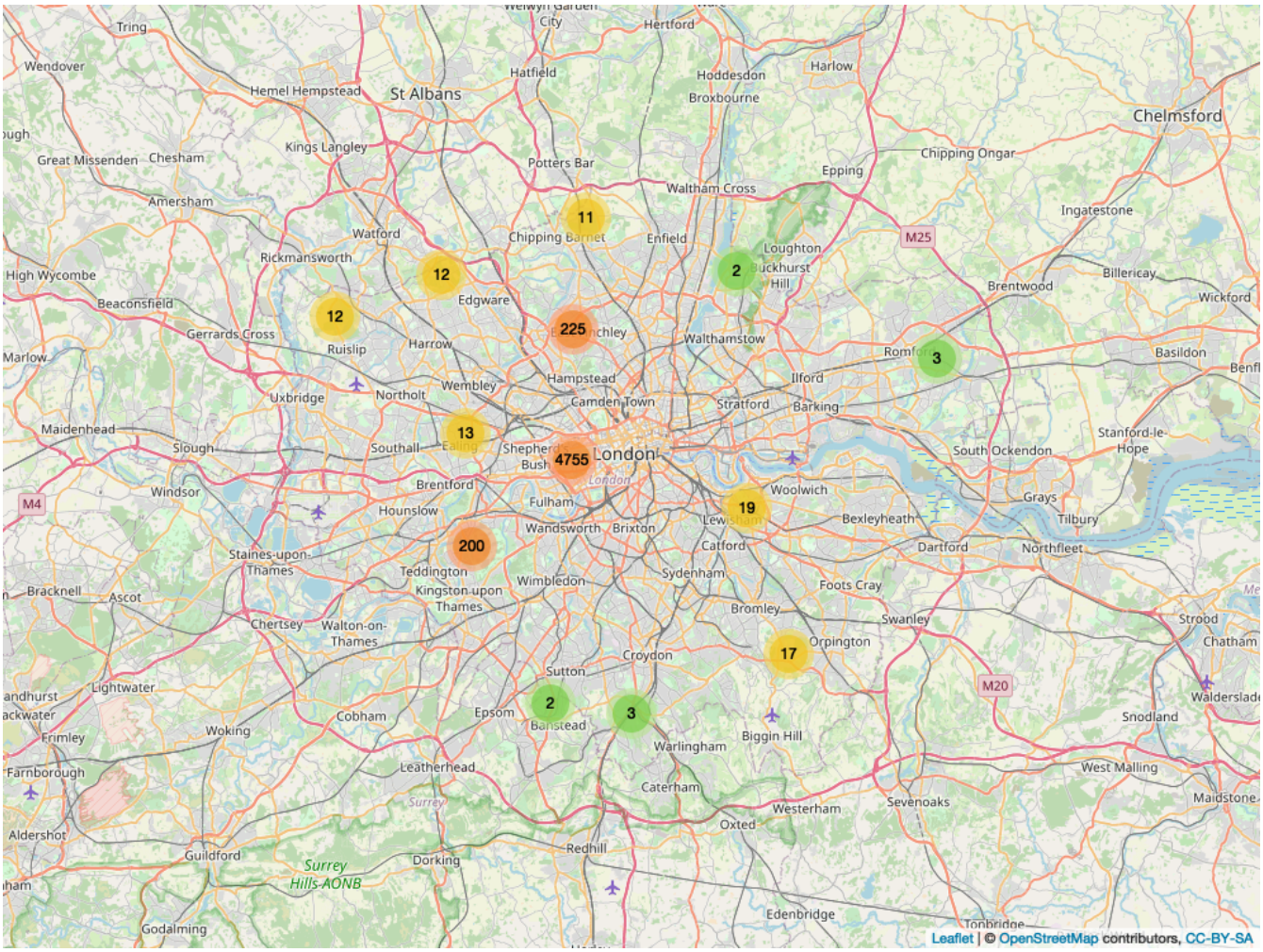
3.2 Is there a trend of which month most sales occur on ?



3.3 Which are the 10 most expensive districts based on the number of properties sold over 3 million GBP?

district	properties_greater_than_3_million_GBP
KENSINGTON AND CHELSEA	1927
WESTMINSTER	958
CAMDEN	648
MERTON	287
RICHMOND UPON THAMES	268
WANDSWORTH	225
BARNET	217
HAMMERSMITH AND FULHAM	211
HARINGEY	102
ISLINGTON	96

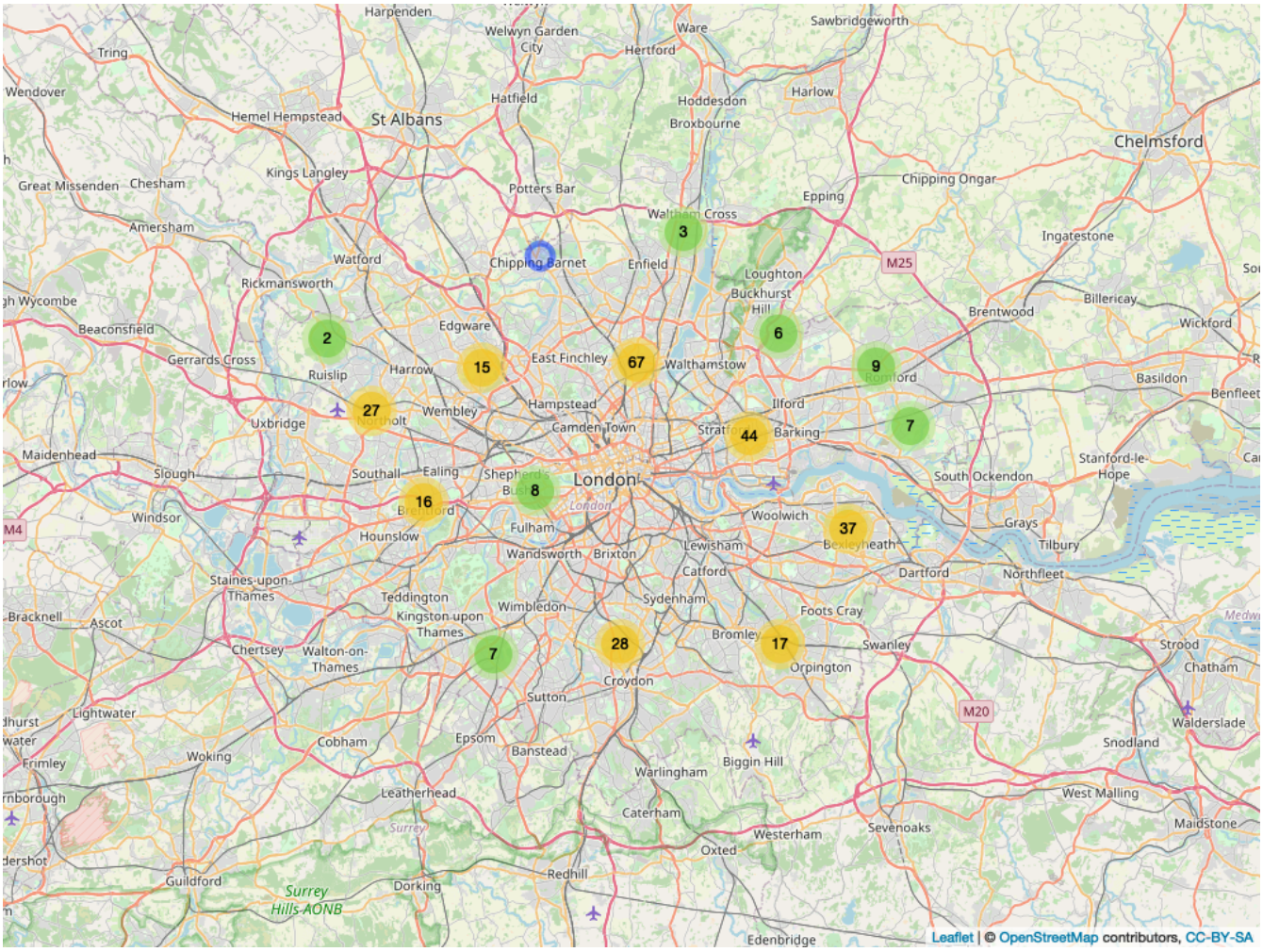
3.4 How are properties sold for greater than 3 million GBP clustered ?



3.5 Which are the 10 least expensive districts based on the number of properties under 100K GBP?

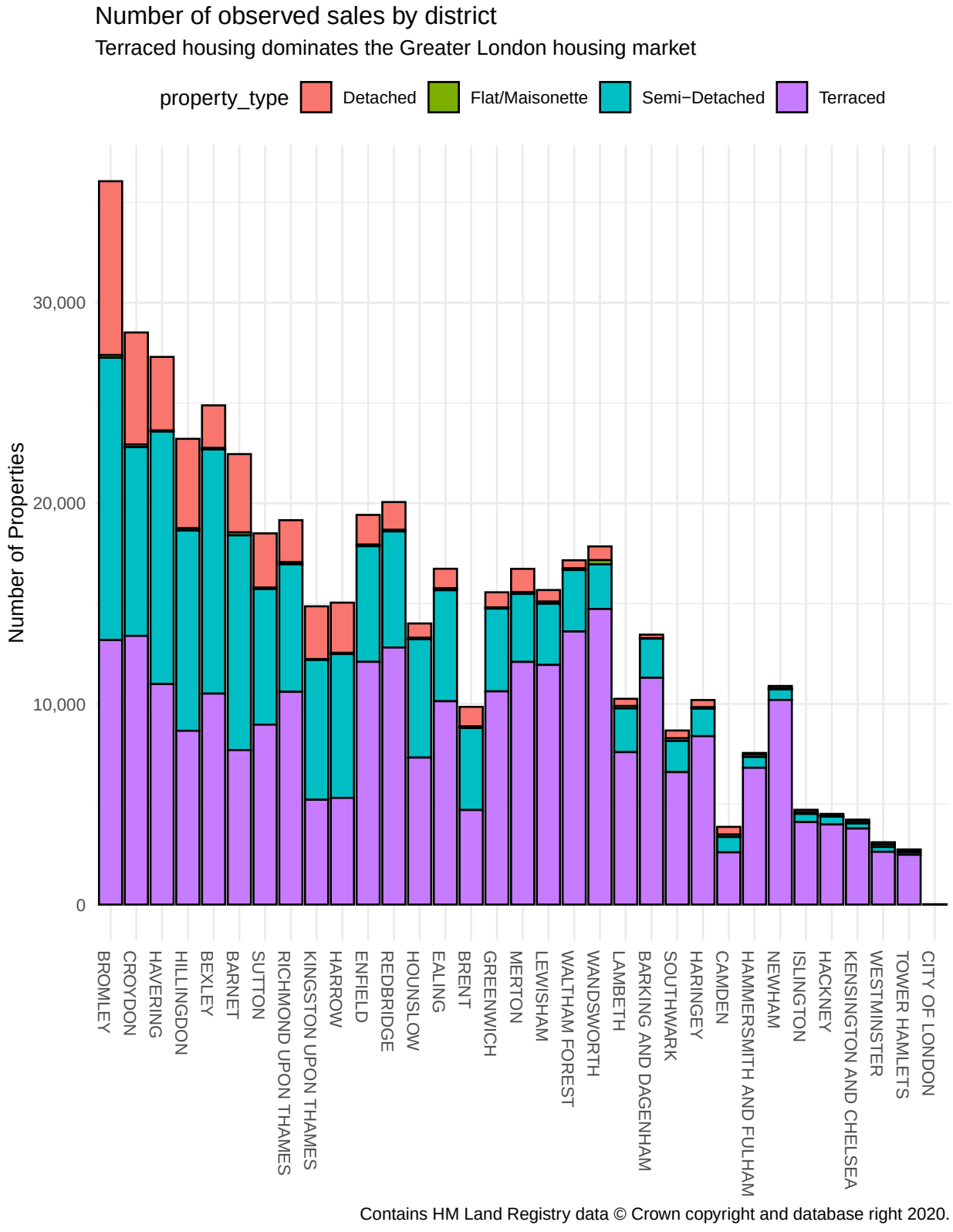
district	properties_less_than_100K_GBP
BEXLEY	31
WALTHAM FOREST	22
NEWHAM	21
ENFIELD	20
CROYDON	19
BROMLEY	17
HARROW	16
EALING	13
HARINGEY	12
BARNET	11

3.6 How are properties under 100k GBP clustered ?



As can be seen from the above analysis some districts contain both top end and bottom end properties, such as Barnet and Haringey.

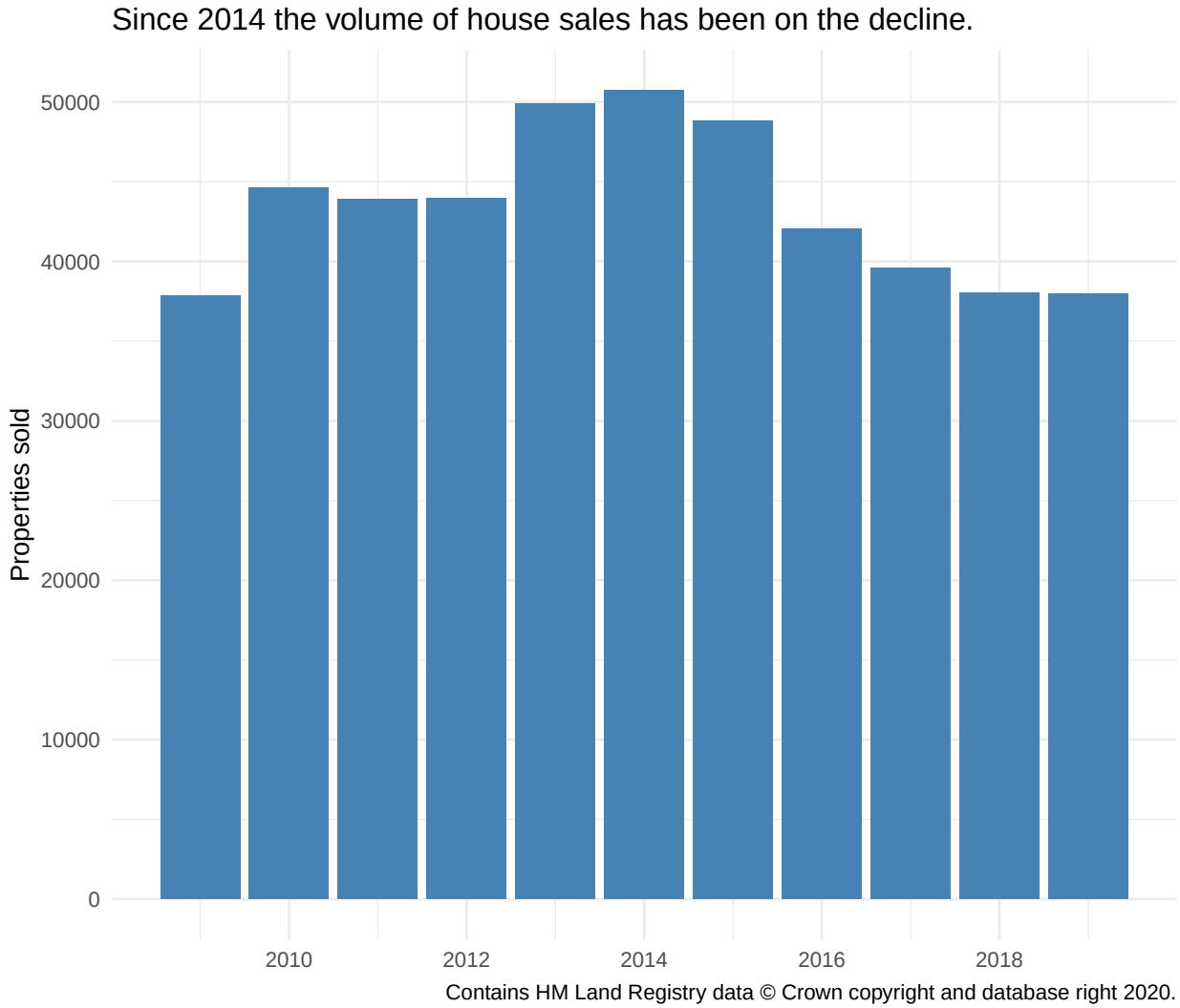
3.7 Are the sales evenly spread across Greater London ?



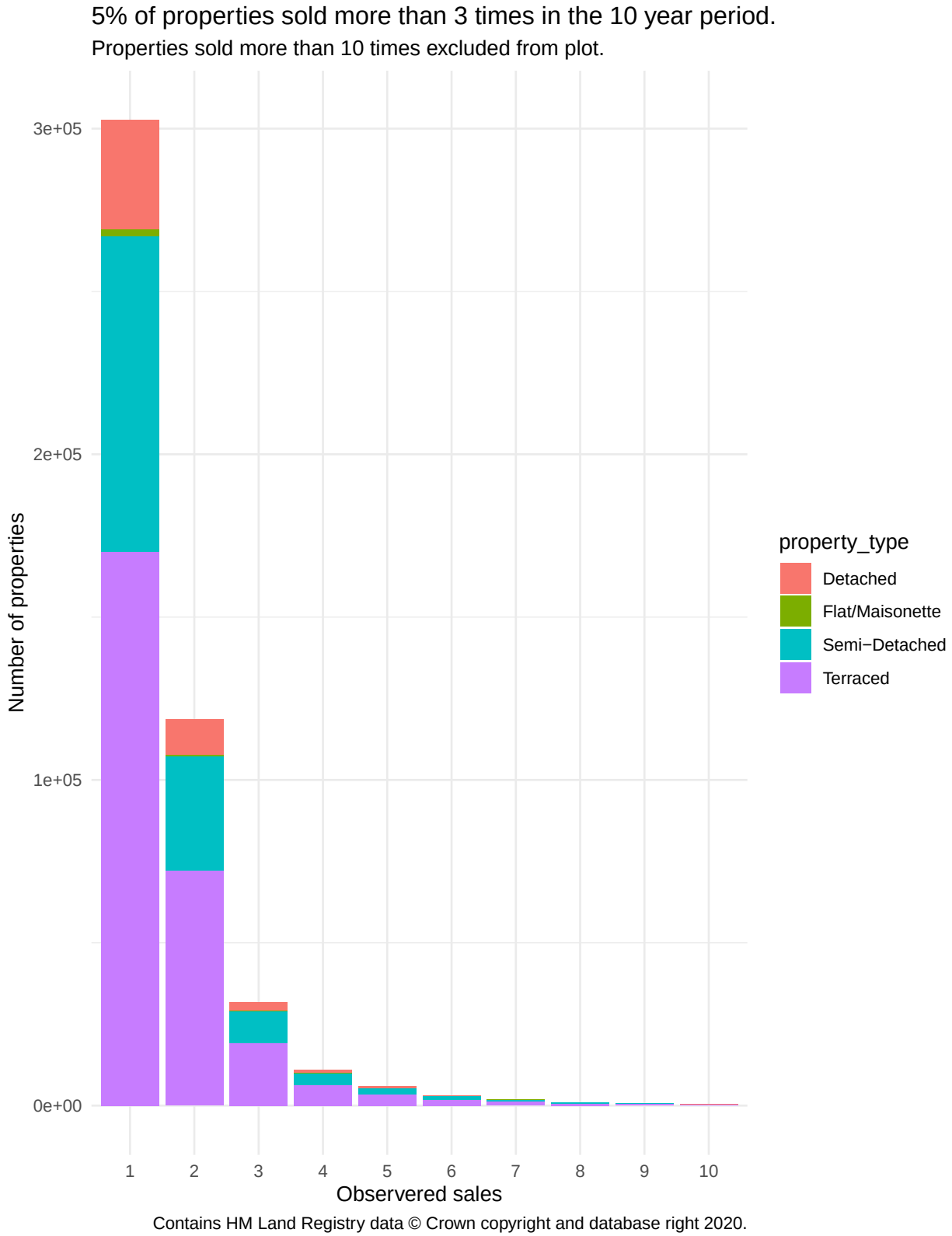
3.8 Which type of property sold the most ?

property_type	number_of_properties
Terraced	275212
Semi-Detached	149397
Detached	49792
Flat/Maisonette	2981

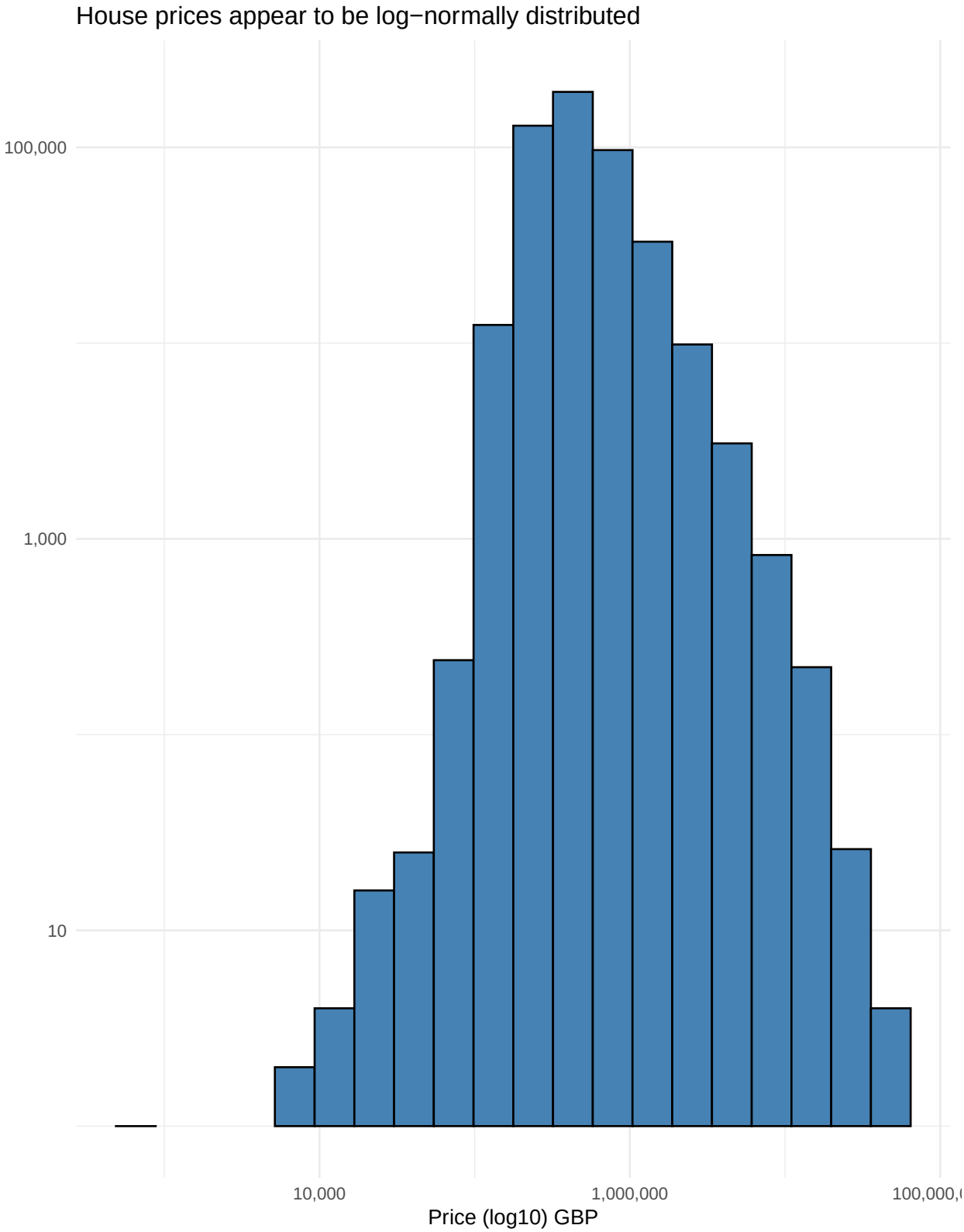
3.9 Which year had the most sales



3.10 How often do properties change ownership ?



3.11 How are house prices distributed ?



```r
glimpse(lhd)
```

```
## Rows: 477,382
## Columns: 12
## $ price             <int> 283000, 182000, 242000, 420000, 470000, 316250, 2...
## $ postcode          <fct> BR6 7DQ, TN16 3RA, SM4 4PT, NW2 1QA, BR2 7DD, BR5...
## $ property_type     <fct> T, T, T, S, S, S, S, S, T, T, T, T, T, T, S, T, S...
## $ new_build         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ district          <fct> BROMLEY, BROMLEY, MERTON, BARNET, BROMLEY, BROMLE...
## $ address           <fct> 6 CLIFTON CLOSE, 18 ST MARYS GREEN, 25 KINGSBRIDG...
## $ transaction_year  <int> 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2...
## $ transaction_month <int> 12, 9, 4, 9, 10, 6, 6, 2, 2, 2, 10, 9, 3, 7, 10, ...
## $ latitude          <dbl> 51.36033, 51.30605, 51.38303, 51.56470, 51.37647,...
## $ longitude         <dbl> 0.06765773, 0.02851643, -0.21894415, -0.21283547,...
## $ outward_code      <fct> BR6, TN1, SM4, NW2, BR2, BR5, SM2, N21, TW1, RM8,...
## $ num_of_sales      <int> 2, 1, 1, 1, 1, 1, 1, 1, 3, 8, 6, 1, 1, 2, 2, 2, 2...
```

```r
nlevels(lhd$outward_code)
```

```
## [1] 190
```

```r
nlevels(lhd$district)
```

```
## [1] 33
```

```r
nlevels(lhd$postcode)
```

```
## [1] 79137
```

```r
nlevels(lhd$address)
```

```
## [1] 377515
```

```r
nlevels(lhd$property_type)
```

```
## [1] 4
```

4 Methods/Analysis

The first step is to split the data into training and testing datasets. The split will be an 80/20 split with 80% of the data in the training set and the remaining 20% in the test set. The split of the data will be stratified by district, to ensure that the number of data points in the training data is equivalent to the proportions in the original data set.

5 Results

6 Conclusion

7 Disclaimers

8 Appendix