A visual analysis of house prices in London

Madhu Sudha Subramani

Abstract—This report intends to analyse the house prices in London across all 33 boroughs and over a period of years. We use features like sales volume, house price index, average sales price, and percent of change in the index compared to the previous year to analyse the distributions of house prices and growth trends. Most of the analysis uses temporal and spatial visualisation along with the numeric data and a few simple charts like the histogram and line chart. The temporal visualisation is done with monthly and yearly time step according to the demand of each research question. Spatial visualisation is used to view the regions visually and understand which areas have what range of house prices and growth. For analysing the growth trend of the index values for all the boroughs and clustering boroughs based on the similar pattern, we used the KMeans time series clustering model with a soft dynamic time warping metric to calculate the distance between the two trends. Tableau is used for all the visualisations while python jupyter notebook is used with many libraries for all the data processing and computations.

1 Problem Statement

Buying a house is a big financial decision in one's lifetime, whether it is a first-time buyer or a repeat buyer. Therefore, extensive research must be carried out while considering many factors before investing such a huge amount. The two most important factors are location and budget. This report intends to analyse the house prices in London. To gain an understanding of both price and area in London, we developed the following research questions and attempted to find answers through spatial and temporal visualization.

1. How many new and old houses have been sold in London each year? And using annual sales data, find the boroughs where new house sales are higher.

2. Does the yearly average percentage change in the index follow a similar pattern in all the London boroughs?

3. What is the average price of different types of houses in London?

4. Which areas have higher and lower average sales prices in London compared to each other?

5. How has property value growth been in each borough over the years, and which boroughs have followed a similar pattern of growth?

To analyse these questions, the UK government's House Price Index dataset seemed suitable. This dataset contains details of all the areas in the UK, but we can filter out only 33 London boroughs for our analysis. There was another source file that contained the details of only London boroughs, but the data structure was not simple and the data was spread across different sheets.

2 State of the Art

First The focus of visual analytics is on analytical reasoning through the use of interactive visuals. Visual analytics is distinguished by the combination and interaction of visual and automatic analysis methods. It enables the analysis results to be refined and evaluated gradually. Patterns discovered using the visual method, for example, can aid in the refinement of the computational model. As a result, visual data exploration combined with model-based analysis can frequently result in improved analysis results (Sun et al., 2013).

Our research primarily examines the temporal and spatial variation of house prices in all the boroughs of London and also analyses the trends in the housing price index. According to (Chi et al., 2021), the spatio-temporal trends of the regional housing markets in England are represented using choropleth mapping. They used spatial and temporal patterns to analyse regional housing prices throughout England and the ripple effect emanating from London to other parts of the country. Similarly, we can use choropleth maps to visualise the spatial distribution of average house prices in London and also to view the boroughs clustered together based on similar index trends. Also, for temporal visualization, we can use both monthly and yearly temporal scales to view the distribution and analyse the variation of different features like the number of sales and price index change of all the boroughs in London.

According to the research paper (Abraham, Goetzmann and Wachter, 1994), they use clustering techniques to find the structural relationships between US housing markets and to create meaningful groups of cities based on the house price index fluctuations. They used the k-means clustering algorithm with the sum of squared Euclidean distances. According to (Chotirat, Ratanamahatana and Keogh, n.d.), calculating with Euclidean distance will result in a dissimilarity measure because it assumes the ith point in one sequence will be aligned with the ith point in the other. Calculating a more understandable distance is made possible by the non-linear, Dynamic Time Warped alignment. Therefore, we intend to use the dynamic time warp metric in the k-means clustering algorithm to find the optimal grouping of boroughs based on the house price index trend exhibited over the past few years. The temporal scale we use is the monthly index, as that is the minimum time step we have in our dataset.

3 PROPERTIES OF THE DATA

The dataset UK-HPI-full-file-2022-10.csv needed for this analysis is taken from the UK government website (GOV.UK, n.d.). There are a total of 139039 rows and 54 columns in the dataset, which covers data from 421 regions in the UK. The data has a monthly time interval, where some regions have data from January 1995 while others have data from January 2004. But as we are analysing only regions in London, we filter the data with the 33 boroughs. Now, the dataset consists of 11022 rows. There was another dataset with only London data but the format was not easily readable, instead filtering London boroughs from this dataset was an easier option. As there are many columns which we are not going to use, let us analyse only the data in the required columns, namely,

Date - the year and month of each data record in the format (01/01/2005),
RegionName - name of the London boroughs,
AreaCode - geographical code of London boroughs (useful for spatial visualization),
AveragePrice - average house price for a borough in each month,
Index - house price index for a borough in each month (January 2015 = 100),
12m%Change - the percentage change in the Average Price compared to the same period twelve months earlier,
SalesVolume - number of registered transactions for a borough in each month,
NewSalesVolume - number of registered transactions for a new house in a borough in each month,
OldSalesVolume - number of registered transactions for an old house in a borough in each month,
DetachedPrice, SemiDetachedPrice, TerracedPrice and FlatPrice - average house price for a particular property type (detached, semidetached, terraced and flat) for a borough in each month (GOV.UK, n.d.)

The main column values, from Date to Index, do not have any missing values. While there are some missing values in the other columns, they do not impact our analysis, as most of it needs aggregated values like the average. Also, the data for 2022 is not complete; we have some values until August and some until October. So, we can ignore the missing values. All the columns have float datatypes except Date, RegionName, and AreaCode, which are in object datatype. The datatype of Date is converted to datetime using format= '%d/%m/%Y'.

4 ANALYSIS

4.1 Approach

For this visual-based analysis report, we need to use both human intelligence and computational methods in conjunction to draw meaningful insights that help answer our research questions. The diagram (Figure 1) represents the sequential steps that need to be carried out, and all the actions mentioned in the yellow box highlight human intervention, while the actions mentioned in the green box use computational tools and methods.

The first step of the research is to define the goals by choosing the problem domain and preparing the research questions. The next step is to collect data and analyse it to see if the set is suitable to answer our research questions. We need to find all possible datasets, try to understand the variables, and select one or more datasets that suit our needs the best. If possible, refine the research questions based on the data available. Also, in the meantime, search for research papers that use the visual analytics approach to solve similar research questions as ours. Then, we select the appropriate visualisation techniques to answer each of our research questions. The next step is data preprocessing, which includes checking if all the values are in the correct format and analysing any missing values or outliers. Everything up to this point is done with basic human intelligence, and then the coding techniques to convert our Date column to a datetime type are applied. In our dataset, there is no need for any transformation, and the few missing values do not affect our analysis, as we are going to do the analysis on the aggregated values like mean and median. For our last research question, we need to use a time series clustering algorithm, and the input for that method should be in a specified format where the length of each time series should be the same and there should not be any missing values. We intend to use python jupyter notebook with libraries like numpy and pandas for basic processing.

The next step is to use the data visualization, and the tool selected for this is Tableau as it is very easy to use, dynamic, and produces crisp and detailed visual representations with minimal effort. To answer four of our five research questions, we tend to use histogram, line chart, temporal and spatial visualisations directly with the available data. For spatial visualisation in Tableau, we need to use the geographic shape file and join it with our dataset. The next step is to interpret the visualisation and try to answer our questions by changing the variables in the rows and columns, filtering, colouring and using aggregate measures.

For the final question, we cannot answer with direct visual representation, so we need to use clustering technique to group the trend of house price index of all the boroughs into groups and then visualise the results. The python library tslearn provides the class TimeSeriesKMeans which is to be used for clustering time series data. The silhouette_score function is used to validate the cluster result. So, for analysing the cluster results in both line chart and spatial representation, Matplotlib and geopandas libraries are to be used. Once the final result is obtained, it can be viewed in Tableau for more aesthetic appeal.

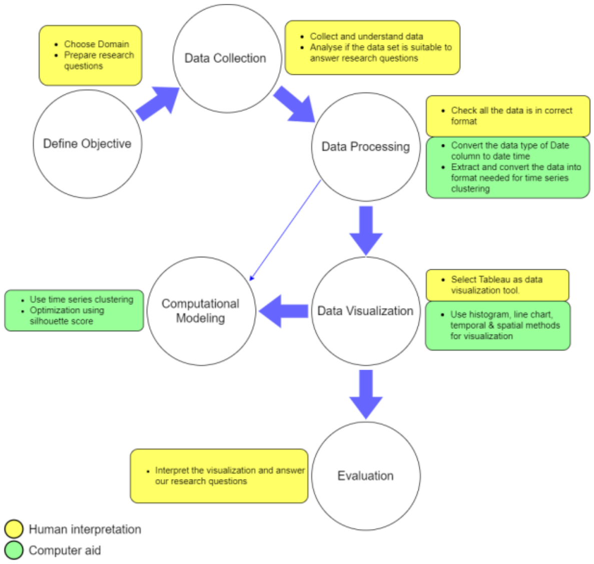

Figure 1. Diagram representing the approach of analysis.

4.2 Process

Let us find the answers to our research questions one by one using visual representations and modelling where simple visualisations are not enough.

1. How many new and old houses have been sold in London each year? And using annual sales data, find the boroughs where new house sales are higher.

For the first part of the question we use simple bar graph representation (Figure 2) that displays the total number of old and new houses sold in London over the period of time from 1995. From 1996 to 2007, the total number of houses sold was more than 130,000, and up to 10 percent of the total sales volume was new houses. The year 2002 saw the maximum sales volume with a total of 173,993, where 162,317 were old houses and 11,676 were new houses. After 2007, there was a sudden decline in the years 2008 and 2009 due to the great depression that we all know. Did the sales volume ever return to its previous state after that? Unfortunately, no. But the sales began to increase slowly and reached a total sales volume of around 122,000 in 2014. This is the highest peak since 2008. Then again, slowly, it started to decrease owing to known factors like the Brexit referendum in 2016, followed by the COVID pandemic in 2020. The sales again increased in 2021 to 107,800, where 7,023 were new houses and the remaining were old houses. For 2022, we don't have data for the last 4 months, so conclusions cannot be drawn for this year.

For the second part of the question, we used a temporal visualisation (Figure 3) across the data for the past 20 years, starting from 2002. This figure gives more details about the new house sales in London. Overall, the new house sales in Tower Hamlets are mostly the highest compared to other boroughs with over 1,000 sales. In 2016, it sold 2,066 new houses, which is the highest among all the boroughs in any given year. In 2017 and 2018, Newham saw the highest new house sales volume with around 1500 sales. Greenwich mostly stands second highest in the new house sales volume next to Tower Hamlets. Following paragraphs...

<1500 words, <

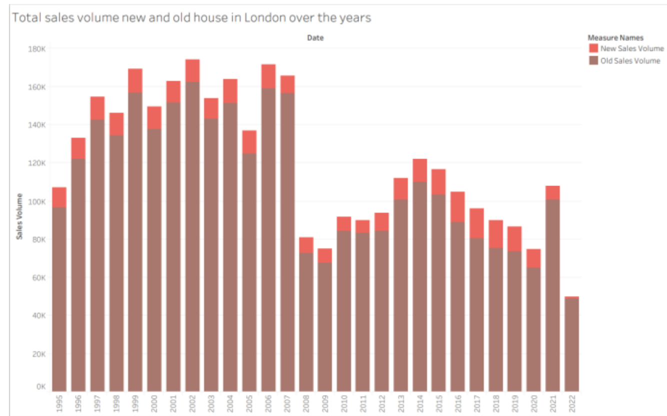

Figure 2 Total sales volume

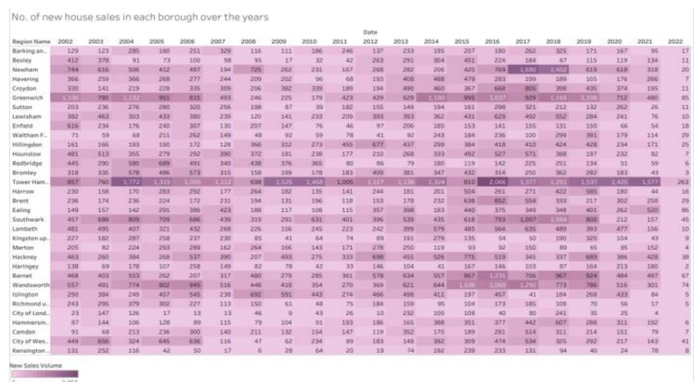

Figure 3 New house sales

2. Does the yearly average percentage change in the index follow a similar pattern in all the London boroughs?

To find an answer for this question, we use a temporal visualisation (Figure 4) annual percentage change in index for all the boroughs over the past 20 years from 2022. The years 2007 and 2014 saw a sharp increase in the index compared to the previous years by 10 to 20 percent. Some areas like City of London and Kensington saw an increase of more than 25 percent. In 2009 for the known reason, the index decreased by an average of 10 percent across all the boroughs. Again, there was a decrease in the index in years 2018 and 2019 in most of the boroughs. And the growth of the index is very slow from then onwards. Few boroughs like the City of London and Westminster have not gone back to the previous index it was in 2017. There are some random decreases in the yearly index here and there, which do not follow any pattern. Overall, there seems to be a slow growth in the index after 2016 with a decrease in 2018 and 2019.

Figure 4 Annual percentage change in index

3. What is the average price of different types of houses in London?

To find answer to this question, we use a simple line chart (Figure 5) to visualise the difference between the average prices of different house types. The median house price is calculated for each year to find the average price of the most sales. There are four different types of houses: detached, semi detached, terraced and flats. In 1995, the average price of a flat was around 60,000, terraced house was around 80,000, semi detached was around 100,000 and detached was more than 160,000. The terraced house was around 30 percent more than the flat while semi detached was around 25 percent more than terraced and detached was 50 percent more than the semi detached. In the recent years, the average price of a flat in London is nearly 400,000, terraced house has increased sharply from around 500,000 to 600,000 in the last couple of years. Similarly, the average price of semi detached house increased from around 650,000 to more than 750,000 in the past few years, while the detached house price increased from just under 1,000,000 to nearly 1,200,000. Now, the terraced house is around 50 percent more than the flat while semi detached is around 30 percent more than terraced and detached is still 50 percent more than the semi detached house.

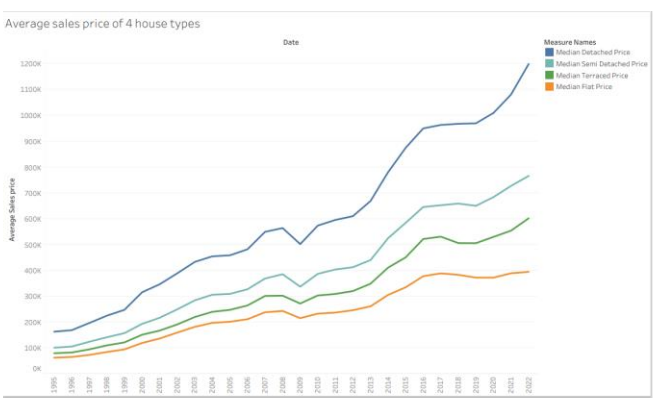

Figure 5 Average sales price

4. Which areas have higher and lower average sales prices in London compared to each other?

To answer this question, let us use a spatial visualisation (Figure 6), to visually view the areas where the prices are more or less comparatively. We choose to compare two maps, one with the average sales price (using the median) of each borough over the last 5 years and the other with data of last year. Both the maps do not show much difference and as expected, the east and outer London has lower average price, around 300,000 to 500,000 while the Kensington and Chelsea borough is incomparably high with around 3 times more value than most of the London. The remaining boroughs in the centre and the west ranges from around 600,000 to around 900,000.

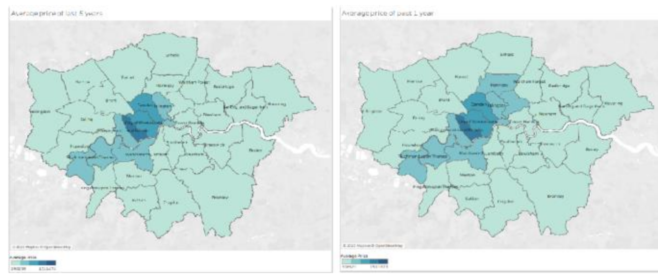
Figure 6 Average price of years

5. How has property value growth been in each borough over the years, and which boroughs have followed a similar pattern of growth?

For analysing the growth of the value of houses in each borough over the past few years, the index feature was chosen. And let us analyse the data from 2015, when the index was set to 100. If we plot a regular line graph for all the boroughs, it will look like figure(7), which is very difficult to interpret. As a result, we will use time series clustering to analyse the trend of the house price index and group the boroughs that follow a similar pattern. We choose to use the famous clustering algorithm, K Means and to implement it in our time-series data, we use the class TimeSeriesKMeans() from tslearn library. There are 3 types of metrics to calculate the distance between two time series: Euclidean distance, dynamic time warping distance (dtw), and soft dynamic time warping (softdtw).

Before choosing the metric, we have to select the features, preprocess the data, and convert it into the required format, where the time index is in columns and the series index is in rows. Therefore, as a first step, we create a DataFrame subset with just 3 columns: Date, AreaCode and Index, from the year 2015. Now, we use the pivot() function to change the single Date column into columns of each date value and then index the AreaCode. As all the values are in the same range, we do not need any transformation. Also, verifying if the length of each time series data is the same and there are no missing values. The euclidean metric measures the euclidean distance between the data points at the same time stamp throughout, and therefore it is not suitable for our data, as no trend lines are exactly the same. So, we choose to use the dtw metric and give the number of clusters as 3.

On analysing the division of clusters, it was not satisfying, and to further validate the cluster performance, we calculated the average silhouette coefficient using silhouette_score() function from the tslearn library. The score was very less with an average of 0.3 with different random states. Then tried with different clusters, 4 and 5 with the combination of different random states. The silhouette coefficient was better in cluster 4 with an average of 0.58 compared to cluster 5 which was less than that. To further improve the cluster performance, we tried another metric, softdtw with all the above mentioned combinations. The division with 4 clusters seemed to perform well both in dtw and softdtw. And finally the K means clustering with 4 clusters using softdtw metric and random state 1, resulted in the average silhouette coefficient of 0.65 and also the trend of all the boroughs in each cluster seemed to be similar.

On analysing the growth of house values across the boroughs from the year 2015 and grouping the boroughs which follows a similar trend using times series Means clustering, we found that the growth trend can be split into four groups. From the figures (8) and (9), it is seen that the three costliest boroughs in central London are highly volatile, and the highest peak does not go beyond 20 percent. The areas in the red have average growth of about 10 percent, and this is also volatile. The areas in both yellow and light green are growing over the years, from 100 to an average of about 130 and 150 respectively.

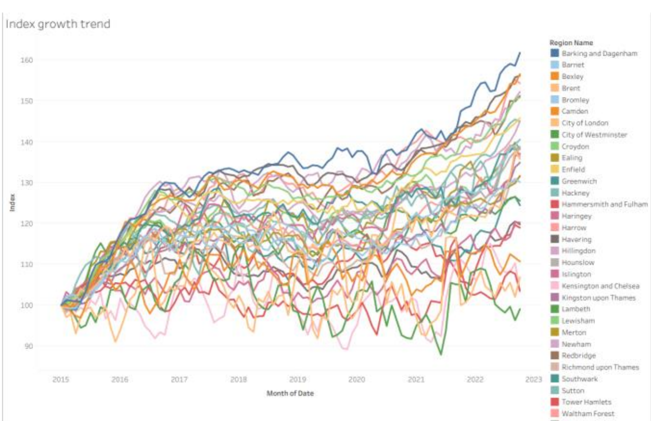

Figure 7 – Index growth trend

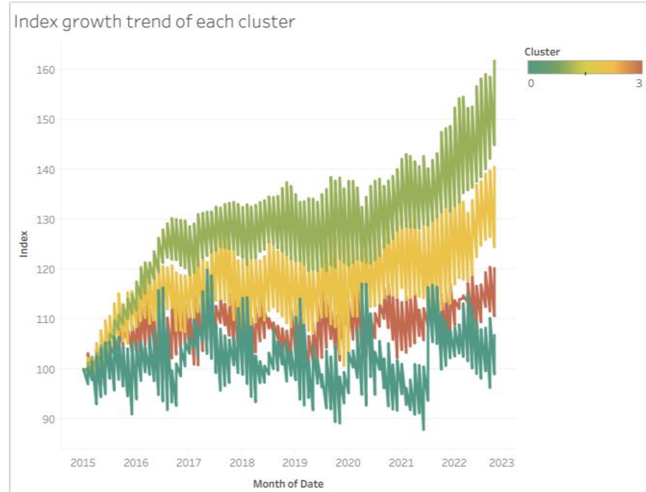

Figure 8 – Cluster wise Index growth trend

4.3 Results

To precisely answer our research questions, the sales of the new houses are higher in the borough of Tower Hamlets, followed by Greenwich and Newham. This means there are more new housing developments in these areas. The annual percentage change in the index compared to the previous year is not same for all the boroughs. Some may increase or decrease randomly across the boroughs. But overall, the change in the house price index depends on the economy and market value of a particular place.

In recent times, the difference between the house prices of different types of properties has increased to such a great extent that the average difference between a flat and a terraced house is 50 percent, between a terraced and semi-detached house is 30 percent, and between a semi-detached and detached house is around 50 percent. And the average prices of boroughs in the east and other outer regions are much lower than those in the west, while the central London area, particularly Kensington and Chelsea, is way more expensive than anywhere else.

From the figures (8) and (9), we have seen that the areas in yellow and light green are developing rapidly while the ones in red and green are highly volatile.

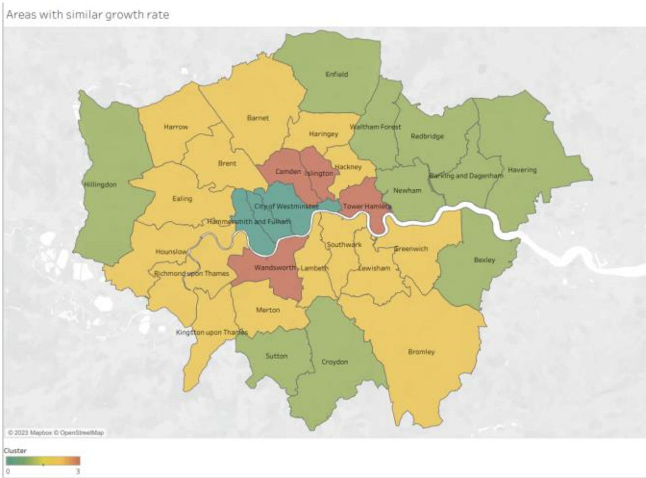

Figure 9 – Areas with similar growth based on cluster

5 Critical Reflection

I am satisfied to have found the use of KMeans clustering in time series data and the distinction of the first two groups were satisfying. But the third group was not very distinct and heavily overlapped with the other two groups near it. But, the overall silhouette score is 0.65, which is good but still this could be increased for the overlapping groups. Maybe this can be achieved by using different time series clustering and comparing the results with ours. The use of Tableau is very easy for simple, temporal and spatial representations of data that do not need processing or computation. Therefore, all the data preprocessing and modelling techniques were all done in python and only the results were exported to Tableau and visualised.

Regarding the research questions, there could be more detailed questions like the sales volume and location of the different types of houses. To answer these questions, there is another dataset named "house prices paid," which has the transaction details of each and every transaction in London. This can be integrated with our dataset, and insights can be gained. Also, in our dataset, we actually have details not just for London but for the whole of the UK. Our research questions can be extended and applied to analyse all the regions across the UK. Another important research question that I thought was important was to forecast the index values and predict the house prices for all the different regions. This requires a lot of input details and analysis, which cannot be done with this time duration of the report's analysis.

Table of word counts

Problem statement	247
State of the art	385
Properties of the data	364
Analysis: Approach	535
Analysis: Process	1501
Analysis: Results	205
Critical reflection	266

References

[1] Sun, G.-D., Wu, Y.-C., Liang, R.-H. and Liu, S.-X. (2013). A Survey of Visual Analytics Techniques and Applications: State-of-the-Art Research and Future Challenges. Journal of Computer Science and Technology, 28(5), pp.852–867. doi:10.1007/s11390-013-1383-8.

[2] Chi, B., Dennett, A., Oléron-Evans, T. and Morphet, R. (2021). Delineating the Spatio-Temporal Pattern of House Price Variation by Local Authority in England: 2009 to 2016. Geographical Analysis. doi:10.1111/gean.12287

[3] Abraham, J.M., Goetzmann, W.N. and Wachter, S.M. (1994). Homogeneous Groupings of Metropolitan Housing Markets. Journal of Housing Economics, 3(3), pp.186–206. doi:10.1006/jhec.1994.1008.

[4] Chotirat, A., Ratanamahatana, E. and Keogh (n.d.). Making Time-series Classification More Accurate Using Learned Constraints. [online] Available at: https://www.cs.ucr.edu/~eamonn/RATANAMC.pdf.

[5] GOV.UK. (n.d.). UK House Price Index: data downloads October 2022. [online] Available at: https://www.gov.uk/government/statistical-data-sets/uk-house-price-index-data-downloads-october-2022?utm_medium=GOV.UK&utm_source=summary&utm_campaign=UK_HPI_Summary&utm_term=9.30_14_12_22&utm_content=download_data [Accessed 10 Jan. 2023].

[6] GOV.UK. (n.d.). UK House Price Index: data downloads October 2022. [online] Available at: https://www.gov.uk/government/statistical-data-sets/uk-house-price-index-data-downloads-october-2022?utm_medium=GOV.UK&utm_source=summary&utm_campaign=UK_HPI_Summary&utm_term=9.30_14_12_22&utm_content=download_data [Accessed 10 Jan. 2023].