Which London Borough Will See the Greatest Increase in House Prices Over the Next Five Years?

Thesis by
Matthew Aitchison

In Partial Fulfillment of the Requirements for the
Degree of
Applied Data Science and Statistics

UNIVERSITY OF EXETER
Exeter, Devon

2021

ABSTRACT

Over the past 40 years, the 32 boroughs which make up Greater London have seen great fluctuating appreciation in house prices, which I can begin to explain the interborough house price disparity with the aid of past literature. Due to the number of boroughs, I cluster all 32 into four clusters via traditional euclidean k-means clustering and k-means longitudinal clustering which produced identical clustering subsets when optimised. With the more manageable four timeseries objects, I could apply several models of varying success in order to forecast the next five years. These included the basic benchmark methods of the average forecast, the Naïve method, and with drift. The final model I concluded with was an ARIMA model for each cluster predicting the next five years of the House Price Index for that cluster of boroughs. By calculating the percentage change between the March 2021 House Prices Index and my March 2026 predicted value, we discover the cluster which contains Kensington and Chelsea, and Westminster has the largest price increase of an estimated 16.7%. The risk associated with a five-year investment reveals the best risk vs reward cluster prediction is attributed to the cluster which contains the Northwestern boroughs.

TABLE OF CONTENTS

LIST OF PLOTS

LIST OF TABLES

Chapter 1

INTRODUCTION

"Your house is your biggest asset", the phrase echoed by financial experts. Your house is the majority of your financial net worth, an asset that has seen great fluctuating appreciation over a long period of time. As of the end of 2020, 63% of households own the house they live in [16]. The average house price in England has greatly increased, particularly in London, over the past few decades.

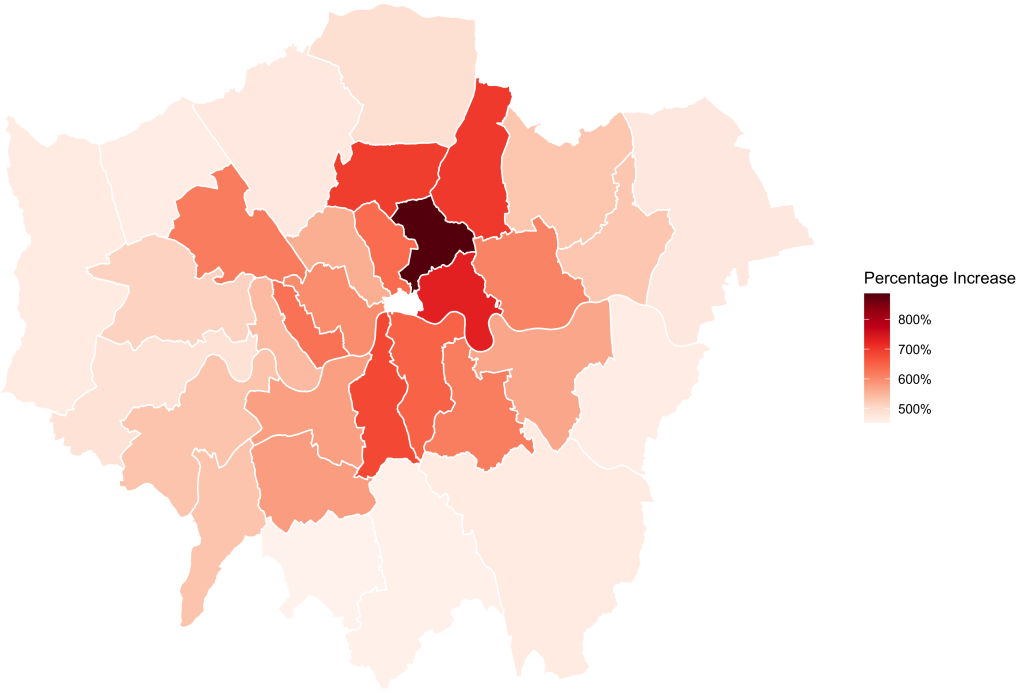

Figure 1.1: Percentage Increase in House Price for each London borough between 1995 and 2021

My initial aim for this project is to investigate how house prices in different boroughs move together and then clustering based on similar trends. This leads to my second primary objective, forecasting future house prices to determine which borough is predicted to have the greatest increase in property value over the next 5 years and which exhibits the greatest risk to reward ratio. This is achieved harnessing 26 years worth of complete sold price data, supplied by the HM Land Registry.

Greater London is made up of 32 boroughs, each governed by their own council. The figure above (1.1) displays a map of London boroughs, illustrating the increase in average house price percentage between 1995 and 2021. Between 150,000 and 300,000 people reside in each of the boroughs, where Inner London boroughs tend to be smaller in both population and area, compared to their Outer London counterpart. The map also highlights a small section of white in the very center, representing the City of London. Despite it's name and location, it is only defined as one of London's local authority districts so does not hold the status of a London Borough.

London has an estimated population of 9,435,622 [5], which is significantly higher than any other city in the UK. Having one governing body would be highly inefficient for managing the diverse areas seen in London, which all have different needs and interests. Tourist hubs like Covent Garden and the West End could not be treated like the primarily residential suburbs of East Finchley or Blackheath. This prompted the need to divide London into distinct boroughs, of which satisfy several criteria. In late 1950s, when the new borough system was being developed, it was decided that a proper borough must have a population greater than 200,000 and must operate like a city in its own right, as to having a center for shops and services as well as unbroken infrastructure connecting boroughs to one another. It was decided that dividing London into 32 boroughs was an appropriate number and that number has not been changed since.

The property landscape for each borough is very distinct. Areas of Inner London have allocated the majority of their land to the financial services industry. Few people live centrally due to exceptionally high house prices, instead commuting to work by the robust Transport for London network. House prices moving away from the epicenter continue to grow thanks to the increases population and limited supply. As a result, it is reasonable to assume there is a clear spatial element to house pricing. The further from the center of London, the cheaper housing gets.

The state of the economy, income, interest rates, changes in population; all have an impact on the housing market. As with all markets, the relationship between supply and demand determines the hedonic value of houses. In this project I'll forecast the future of house prices based on the monthly House Price Index for each borough, starting from 1995 until present. This

Although limited research has been conducted on London house prices on a borough level there are many factors which would cause boroughs to perform differently. Council tax varies between boroughs. For example, a Band D property in

Wandsworth has a council tax charge of £839.10 [6]. That compares with a council tax bill of £2056.47 [12], (nearly 2.5 times more) for a Band D property in the neighbouring borough of Kingston-upon-Thames. This could drive down the asking prices for houses since future buyers will have to factor this amount when budgeting for their new home.

The figure below (1.2) displays sold house prices from 1995 up until 2021, separated by borough. This is the time frame I will be investigating since 1995 is the first year the HM Land Registry has a value for the UK House Price Index. Unsurprisingly all boroughs follow a similar trend where the effect of economic factors have on the demand of housing. The clear dip in 2008 following the 2008 financial crisis where recession and rise in unemployment deterred many from buying. The rising population and shortage of properties, particularly in London, has been responsible for the steady increase in house prices.

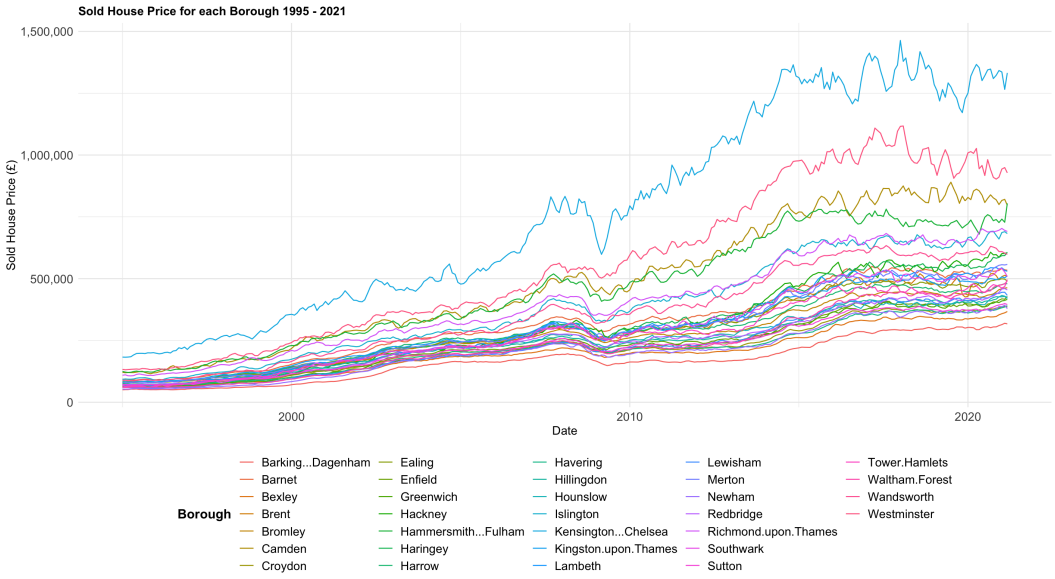

Figure 1.2: Sold House Prices for each month between 1995 and 2021, broken down by borough

The figure (1.2) also illustrates the similarity between different boroughs. The bottom of the graph is densely packed with overlapping lines, suggesting a clustering would effectively reduce the scale of any analysis. But this would also open the opportunity to discuss the difference between these boroughs and higher performing clusters like Kensington and Chelsea. Analysing trends of a few clusters is more reasonable based on the scope of this project than all 33 boroughs. It would be

superfluous to analyse and forecast each borough individually when I can remove a cluster if it lacks investment potential.

This research is valuable to those who are looking to buy property in London and seek advice as to which parts of London would see the greatest potential increase in value. This includes investors and home buyers alike since any party with a stake in property should see it as an investment they can yield maximum returns. Choosing to purchase real estate in the best performing borough could return up to a 40 percent increase in value over another borough.

The economic disruption of the Covid-19 global pandemic and the United Kingdom leaving the European Union has made forecasting future house prices very unreliable. However, house prices in London have always recovered after a maximum of five years, which is quicker than other parts of the UK. This was following the 2008 credit crunch, which is considered to be far more damaging to the market than the pandemic or Brexit.

Chapter 2

LITERATURE REVIEW

To forecast the future of house prices, we must first grasp the key factors we attribute to rising house prices, especially that of which are unique to London or similar cities.

Gentrification is considered to be the leading force in the increased price disparity in house prices. Historically, areas of inner London such as Chelsea, Camden Town and Notting Hill were the first to see an upwards change. [21]. In respect to London Boroughs, it has been commented that gentrification has been the gradual process of change to the social class seen in central London [25]. In part, the inner city was historically structured around manufacturing but in the past 60 years, it has shifted to one of finance, an industry that has increased separation between socio-economic groups. More recently, the London government compared the Gross Value Added (GVA) for each of the London boroughs in the years 1997 and 2014. GVA is the measure of the contribution made by a region to an economy. It was found that the proportion of the total GVA. In 1997, the majority of London boroughs were making equal contributions, other than the center and a few fringe boroughs. More recently, in 2014, the figures show a greater separation in recorded GVA. Now just Hillingdon is the only fringe borough to be grouped as one of the highest contributors along with a growing central influence, as Lambeth joins the group of highest contributors. This leaves more of the other boroughs contributing relatively less to the total output of London [20]. It would be valuable to compare the GVA every twenty years because I suspect the city has grown in wealth and the fringe boroughs have slowly declined since the conversion from manufacturing to financial services. However, with such a limited amount of historic GVA data, it would be unwise to make any predictions without having access to the GLA data sets. Therefore, it is implied that the majority of the industry takes place in the center, taking business from the rest of London.

A more recent publication identifies the difference in land use by distance from the epicenter[44]. "In 2017, Manufacturing was concentrated in Outer London (accounting for 78.3% of all London's manufacturing) while Inner London produced 95.3% of London's GVA in Financial and insurance activities". This socio-economic disparity is the driving force for gentrification, as London continues to grow, the traditional industries will most likely be forced out of London since they require

more space, which the city cannot support.

Many analysts forecast future house prices, with huge departments dedicated to predicting year-end as well as 5-year house price changes. Typically, estate agents like Savills will publish predictions for 'prime central London' and 'outer prime London', rather than on a borough level. For example, Savills predicts a 21.5% five-year price forecast for prime central London [7]. The boundaries of 'prime central London' are not well defined in the literature, as it is an arbitrary collection of well-performing, expensive parts of London. Presumably, this is to produce the most attractive possible figure, whilst still retaining accuracy. Academic research papers will often categorise the boroughs into Inner and Outer London when forecasting since there is a distinction made by the London Government Act 1963 [1], when the new borough system was devised.

Obviously, it is in the best interest of letting estate agents to withhold their forecasting methods, so we must look to academic research. It is common to consider different variables attributing to house prices when making predictions. One of the simpler approaches to forecasting house prices in London is produced by J. Stuart Wabe (1971)[47]. By building a regression model to explain the increase in house prices, he predicts house prices based on ten explanatory variables, including a mixture of location variables and house parameters. The most valuable takeaway from this paper is the very large constant term, which indicated the price for the plot for different parts of London. Comparing these values would give a baseline indication of how much a property is worth in each borough, which is more relevant for my analysis. Removing the explanatory variables would be justified for relatively outdated London data since most coefficients are not as influential, such as those impacted by the previously mentioned shift in London industries. Despite a high R - squared value (0.90), we should be cautious of an inflated value due to the data being grouped. By examining the zero-order correlation matrix, we can see that many of the variables are highly correlated. This is expected, for example, rail journey and cost of rail fare is of course correlated. The paper ultimately highlights the factors with the biggest influence on house price is associated with distance but it is more complex than just accessibility, rather the environment surrounding the area is ultimately most important. As I previously mentioned, each borough acts as its own city, with its own unique environment, impacting the house prices of boroughs individually.

Described as the first notable study of prediction using house price data to estimate

an autoregressive integrated moving average model of differenced log prices was by Rayburn, Devaney and Evans (1987) [41].

Another, more sophisticated approach to forecasting is to consider if house price changes in one location can not only be predicted by their own history but also by the changes in house prices in nearby locations. The Journal of Housing Research examined the spatial and temporal housing price relationships in the US from 1975 through 1994 [39]. This pairwise approach was developed further by [28].

Chapter 3

METHODOLOGY

3.1 Methods of Clustering

I will allocate the boroughs into a number of different clusters, with different methods of allocation. This meticulous approach to clustering should yield appropriate and proven cluster combinations. Ideally you would not need to cluster a collection of time series graphs but since 32 boroughs are too many to model individually, clustering scales down to meet the scope and time restraints of the project. Furthermore, clustering gives us an understanding of which boroughs act similarly and we can identify connections between certain boroughs and not others.

K-Means Clustering

K-Means is one of the most widely used and perhaps the simplest unsupervised algorithms to solve the clustering problems. Using this algorithm, we classify a given data set through a certain number of predetermined clusters or "k" clusters. Each cluster is assigned a designated cluster center and they are placed as much as possible far away from each other. Subsequently, each point belonging gets associated with it to the nearest centroid till no point is left unassigned. Once it is done, the centers are re-calculated and the above steps are repeated. The algorithm converges at a point where the centroids cannot move any further.

Total within-cluster variation is the sum of squared Euclidean distances between data points and appropriate centroid, as defined below:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{3.1}$$

where x_i is the mean value of the points assigned to the cluster μ_k and x_i is the data point belonging to the cluster C_k

We intend on minimizing the sum of squares (SS) distance by assigning each observation (x_i) to the cluster with the smallest distance to the cluster center (μ_k)

The total within-cluster variation as follows:

$$totalwithiness = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{3.2}$$

The Total within-cluster sum of square is the metric of compactness, a proxy for goodness, so we want to minimize it.

There are three main methods used to determine the optimum number of clusters. However finding this partition is a subjective matter, with no definitive answer, somewhat dependant on the method used. The first is the the Elbow method which plots the total within-cluster sum of squares (WSS) against the number of clusters. The lower the WSS, the better but you want to choose a number of clusters which would not improve much if another was added. This is the 'elbow' of the graph. The other two method of optimising the number of clusters is the silhouette method and the gap statistic.

Clustering Using the Kml Package

Using the K-means for Longtitudinal Data package (kml), I can identify the optimum number of clusters of all thirty two boroughs. Despite the name, it considers trajectories. KmL uses the object class ClusterLongData, converted from a dataframe. The algorithm KmL will build a partition - a subgroup of

Kml is a "hill-climbing" algorithm, that meaning the algorithm starts with an initial solution then makes iterative attempts to find a better solution through incremental change. The specificity of this algorithm sees it always converging towards a maximum however this may be a local maximum rather than a global, so we run the algorithm several times, select the best solution and thus increasing the chance of getting a quality partition. similar to the simple k - means method described before, the number of clusters is not known beforehand but can be calculated afterwards using clues provided by the result.

The kml package displays an interactive panel where you can view each clustering combination then export the cluster of your choice. There are several factors to consider when deciding the most useful cluster arrangement:

- The Calinski-Harabasz Index. On the left of the plot is a sorting scale which indicates the 'quality' by the Calinski-Harabasz Index, the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters.

This is properly defined below (3.3) This directly relates to the standard concept of a cluster, where the more dense and well separated a cluster is, the greater the ratio value is.

- Size of clusters. For example if a cluster only contained one or two boroughs, it would not be particularly helpful. Kml shows these values at the top of the graph. Furthermore, as the number of clusters increase, the reliability of each cluster decreases.

- The borough trajectories, as displayed in different colours. This involves choosing the partition which displays the borough clusters moving with similar trends.

The Calinski-Harabasz Index

For a set of data E of size n_E which has been clustered into k clusters, the score s is defined as the ratio of the between-clusters dispersion mean and the within-cluster dispersion:

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1} \tag{3.3}$$

Where $tr(B_k)$ is trace of the between group dispersion matrix and $tr(W_k)$ is the trace of the within-cluster dispersion matrix.

Both of the mentioned clustering methods are similar in that a predetermined number of clusters is chosen but the criteria each use to determine how to subset the elements into those cluster groups utilise different ratios. Using the results of these methods, you can compare allocation assess how effective they both are.

3.2 Methods of Forecasting

The predictability of any event depends primarily on three factors. How much data is available, how well we understand what influences the probability and whether the forecasts can act on themselves. So does time series forecasting house prices satisfy these factors?

Firstly, we have 315 house price index values from 1995 until March 2021, which is sufficient since the rule of thumb is more than 50, ideally 150 [13]. Having monthly data, we can investigate and feature seasonality in our model. Although we have a strong theoretical understanding of what factors influence house price changes, the weighting for the vast explanatory variables is still a point of discussion. Furthermore the amount of influence these characteristics are always changing making predication more challenging. That being said, fore forecasting, our theoretical understanding is sufficient to make reasonable forecasts. Lastly, forecasts of house prices have a direct effect on the house prices themselves. When forecasts publish an increase in house prices, people may adjust the price of their house inline with these predictions, thus the forecasts are self-fulfilling. This often creates a bubble, which will eventually burst once the real prices have been realised and adjusted.

Predictor variables such as population densities, strength of economy and interest rates are often useful in time series forecasting however, there are too many factors affecting house prices and we do not fully understand the system at hand. So I will just use forecasting methods based solely on historical sale price data.

Before any modeling can take place, some initial exploratory analysis should be conducted on the time series data to identify any patterns, trends, potential seasonality or cycles. Utilising a few different models is common and then there are a number of methods we can use to properly evaluate the accuracy of the forecast produced.

ARIMA Model Forecasting

A popular and widely used statistical method for time series forecasting is the ARIMA model. Appropriate for my dataset based on sold house prices, this is an example of univariate time series forecasting, that being predicting the future based entirely on past values.

Auto Regressive Integrated Moving Average recognised to have the acronym ARIMA, is a class of model that captures a suite of different standard temporal structures based on it's own past values. Since the house price data is available every month for years, we can also model for seasonal composition.

We can break down the name ARIMA into three parts:

- AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.
- I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
- MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Once a model is produced, calculate the percentage change over the five years of predictions to identify which cluster will increase the most.

Chapter 4

DATA

The two main data sets made available by the HM Land Registry with records of house prices from 1995 until the present are:

- Price Data. Published by the HM Land Registry, this data set includes the price paid for every house in the UK since 1995. With such variables as price paid, address and house type. [43]
- House Price Index (UK HPI). The UK HPI represents growth in average house prices within a geographic area. The UK HPI applies a hedonic regression model that utilises the various sources of data on property price (for example the Price Paid dataset) and attributes to produce up-to-date estimates of the change in house prices each period. [42]

The UK HPI is built on a hedonic regression model, which estimates the various factors' influence on the house price. [8] These include the characteristics of the house, for instance size, condition, style, and as the surrounding area, if it is situated near schools or a tube station for example. The price stands as the dependant variable, influenced by the independent values, in this instance, the characteristics and surrounding area. These are the weights that buyers place on the average house price, to produce the HPI. The transaction data used for the HPI was comprehensive, however, there are very few characteristic variables required to improve the accuracy and complexity of the hedonic regression model.

In different periods, we can expect a reasonable amount of fluctuation and variability so price data alone would not suffice. As stated in the HDI report, the two other data sets introduced in computing the model are 'Valuation Office Agency Council Tax Valuation list' [2], a comprehensive list of attributes for every property and the 'Acorn classification'[48], a data set containing spatial information about the area. Combined these three data sets have been wrangled into a regression model. Since this is not raw data, it has been manipulated introducing a certain degree of risk. The UK HPI represents growth in average house prices within a geographic

area however this may differ from the growth of individual properties within that geography.

Noting that the true sold house price data has far greater inherent precision, containing sold house prices for individual houses however that is not necessary, this would only be appropriate if I was analysing on a street, postcode or town level.

The HM Land Registry is responsible for collecting all residential housing transfers and producing the UK House Price Index. The UK HPI represents growth in average house prices within a geographic area. Constructed mainly from the aforementioned price data from HM Land Registry. Since it is only once the purchase has been registered that it is fed into the index. As a result, caution is advised when looking at the most recent sold prices since they can be revised. Taking the sold house price data up until March will negate any risk of revisions.

At the borough level, a 3-month average has been applied. Usually, this is done to remove some volatility but this is almost unnecessary due to the volume of houses sold each month in each borough. At borough level, the time series has not been seasonally adjusted nor has it been adjusted for inflation. Rather the The UK HPI measures nominal house price[37].

Of course there are different sources of house price statistics outside the ones made available by the HM Land Registry. Companies like Rightmove use their own data, that being listed price rather than sold price [9]. House prices are famously 'sticky' so we would expect some undetermined lag, which is not present with the UK HPI. Banks like Nationwide and Halifax utilise mortgage approval data [23], which means means a lot of houses are not included. This is not a large concern since the approximation for each borough would be reasonably accurate the volume of purchases each month for each borough.

Chapter 5

CLUSTERING

5.1 K-Means Clustering

Simple K-Means Clustering Method

Starting with the popular Hartigan-Wong algorithm variation of K-means clustering, which minimises intra-cluster Euclidean distances, I sorted the boroughs into a predetermined number of clusters. In the figure below 5.1 I have the results of the three different methods of optimising the number of clusters.

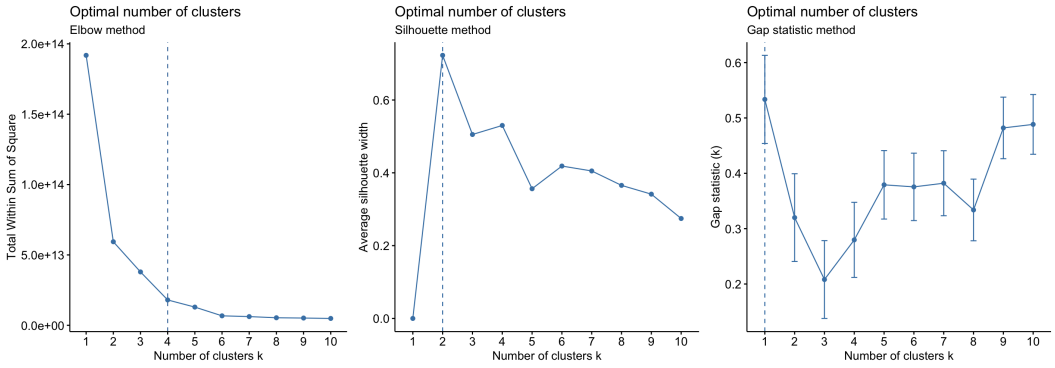

Figure 5.1: Optimum number of clusters

As is common with deciding the number of cluster to use, these three methods all give different optimum values. Considering the different results, I decided four clusters is appropriate. I identified this using the elbow method and has the second highest value using the silhouette method. I ignored the result of the gap statistic since it suggested a cluster of one, which is not a cluster. To ensure I made the right decision, I tested the allocation of clusters when using both 2 and 4 clusters, which produced a cluster with 2 boroughs in it and another with 30. I deemed this unsuitable since I believe that is an example of being overly sensitive to outliers, that being Kensington & Chelsea and Westminster.

The figure (5.2) below shows the result of clustering by k-means for four clusters. The light blue colour represents the sold house prices fro each borough and the dark blue line indicates the average centroid for that cluster.

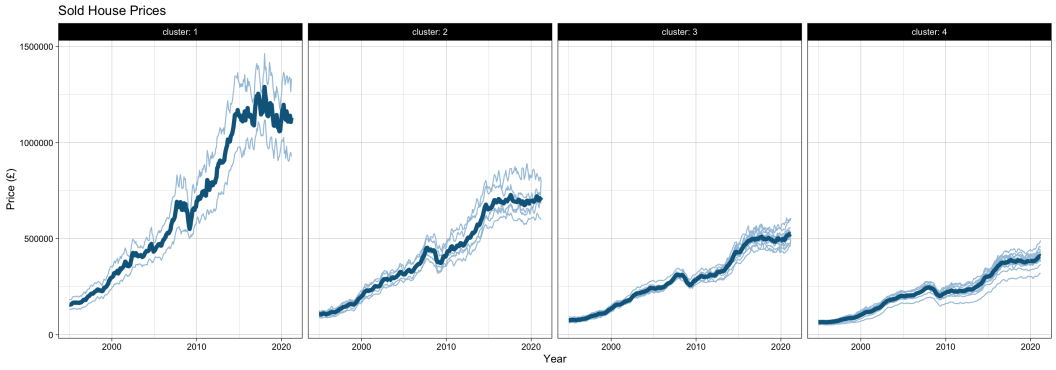

Figure 5.2: Sold House Prices for each Borough - Simply Clustered by K-Means.

From the figure above (5.2) we can see the quality of clusters 3 and 4 are significantly greater than that of cluster 1 and 2. One of the limitations of k-means clustering is how sensitive it is to outliers, that being Kensington & Chelsea, as per (5.1). This method indicates nothing about how the prices move together but rather the distance between points at each data point. This is evidence a more sophisticated approach must be taken.

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Kensington and Chelsea	Camden	Barnet	Barking and Dagenham
Westminster	Hammersmith and Fulham	Brent	Bexley
	Islington	Bromley	Croydon
	Richmond Upon Thames	Ealing	Enfield
	Wandsworth	Hackney	Greenwich
		Haringey	Havering
		Harrow	Hounslow
		Kingston Upon Thames	Lewisham
		Lambeth	Newham
		Merton	Redbridge
		Southwark	Sutton
		Tower Hamlets	Waltham Forest

Table 5.1: Table to show which borough belongs to which cluster seen in figure 4.1.

The kml Package

As described in the methodology, using the kml package, I can scan through and select the best cluster allocation which is based on the Calinksi-Harabasz Index, the size of the clusters and the trajectories which support my analysis. In this, I found four clusters to be most appropriate on the grounds that re-rolling the algorithm gave relatively consistent values for the Calinksi-Harabasz Index. The Calinksi-Harabasz Index is greatest when 6 clusters are used but that reduces some clusters to just one borough per cluster compared to four clusters. Alternatively, when k = 2, the two boroughs, Kensington & Chelsea and Westminster, make up one cluster and the others make up the second. Therefore, selecting the cluster size to be four appears to be optimum, which is very consistent with the results of the simple K-means clustering method

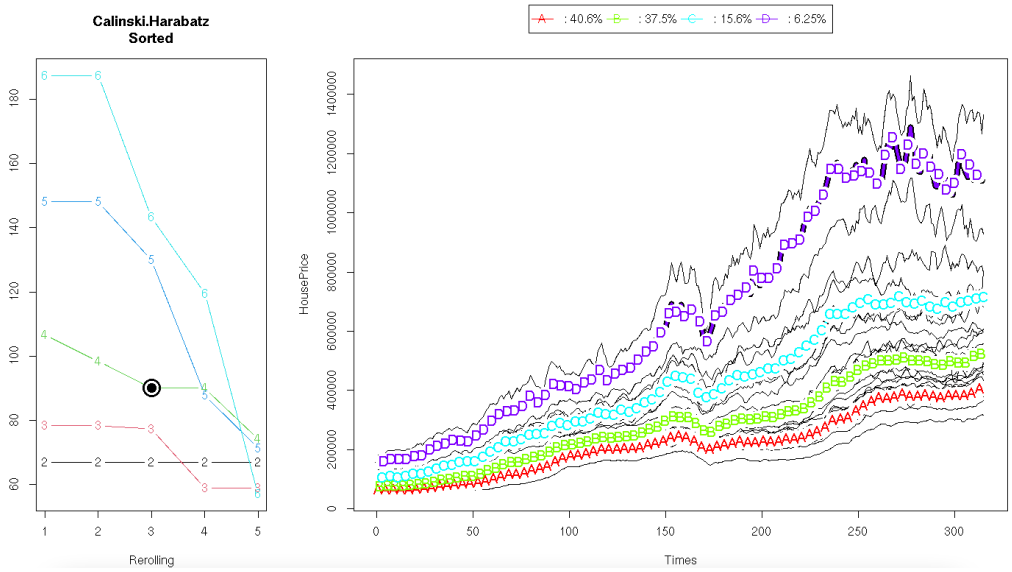

Figure 5.3: KML Cluster Results

The kml method allocates the boroughs into identical subsets as the ones produced when clustering by the simple k - means approach detailed above. This gives me confidence that these are reasonable clusters and hence I will use these clusters for forecasting.

It may be beneficial in understanding the relationship between boroughs if we look at the spatial patterns associated with the clustered boroughs. The four clusters have been colour coded on a map of London which reinforces our understanding from the literature review in regard to disparity between parts of London. The most expensive boroughs, are neighbours and the second most expensive cluster occupies the boroughs just outside the centre. Lastly, clusters 3 and 4 are situated in the fringes of London.

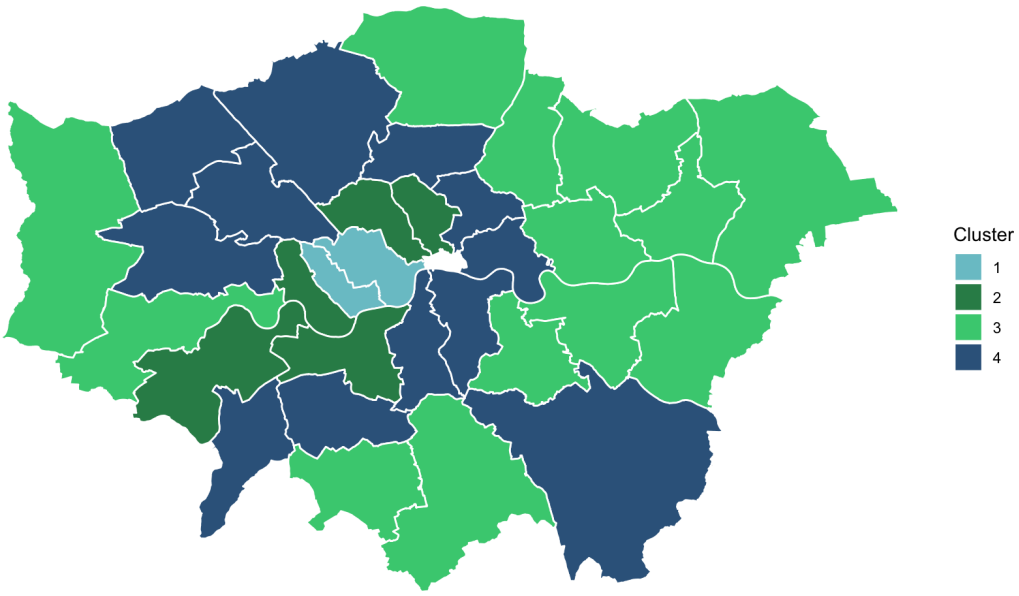

Figure 5.4: Heatmap of London Boroughs by clusters from fig. 4.1

Chapter 6

TIME SERIES MODELLING

6.1 Initial Analysis

Before forecasting it is important to get a sense of each time series and what explained. I have produced a graph below which displays all four clustered boroughs from 1995 to 2021 (6.1). All boroughs follow a similar trend although Kensington and Chelsea consistently have the highest sold house prices. The most noticeable dip in house prices is following the 2008 financial crisis. Between 2007-2012 house prices fell as a result of the credit crunch, leading to a decline in bank lending, as well as houses being overvalued in the prior boom, thus few first-time buyers could afford to buy. Finally, the recession and rise in employment discouraged many from buying which continued the following years although the house prices increased universally due to the ongoing shortage of supply and rising population.

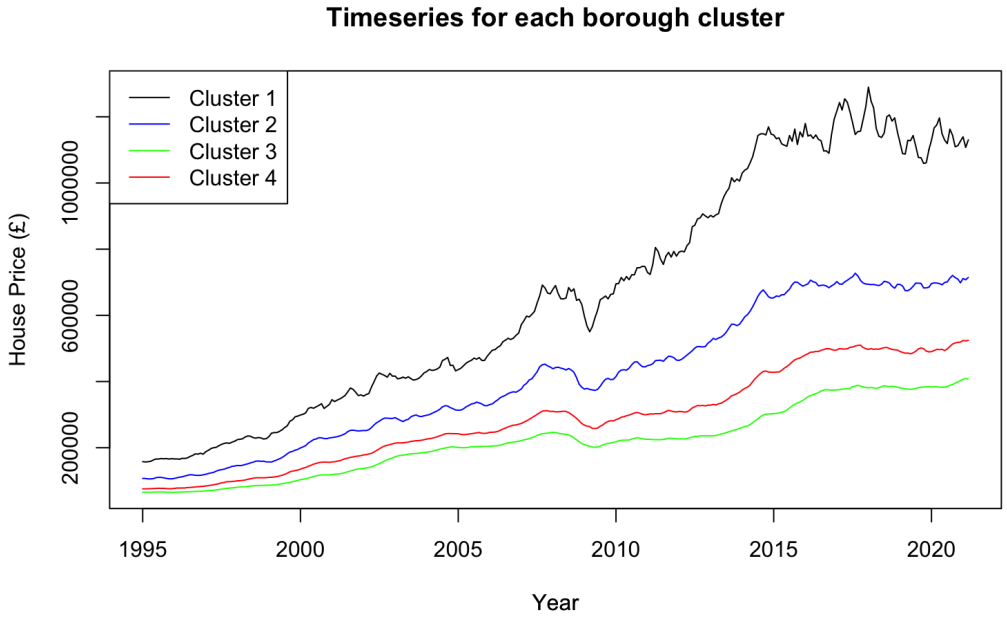

Figure 6.1: Timeseries for each borough cluster

Looking at the house prices, it is worth noting that cluster 1 (containing the two most expensive boroughs) took the largest relative decline but also recovered the quickest. The more volatile house price might be more challenging to forecast unless there is

some identifiable cycle. Looking at each cluster from the raw timeseries plots, we can see little to no seasonality, which is not surprising since people will more often hold onto their house rather than sell at a little below its market value. Furthermore, any fluctuations in the data are partially smoothed over in the clustering process. This is most noticeable when comparing cluster 1 with clusters 3 and 4, as these larger clusters appear to have been smoothed over. By comparing the changes in house prices for all four clusters over each year, we can see that there is little variation of house prices depending on which month the were sold in.

Before forecasting it is standard to make some adjustment to the data but in this instance, there is little we can do. The obvious suggestion would be to adjust for inflation however that needlessly makes things more complicated, considering which index to use and so forth. Instead, final model will factor the upwards trend present for each time series.

By decomposing each time series to isolate the trend line, as seen in the figure below (6.2), we can see an issue with cluster 1. Since it contains the fewest boroughs and the greatest within cluster dispersion, the trend line representing it is far more unpredictable unlike the other clusters which appear smoothed by the averaging out affect in the cluster stage.

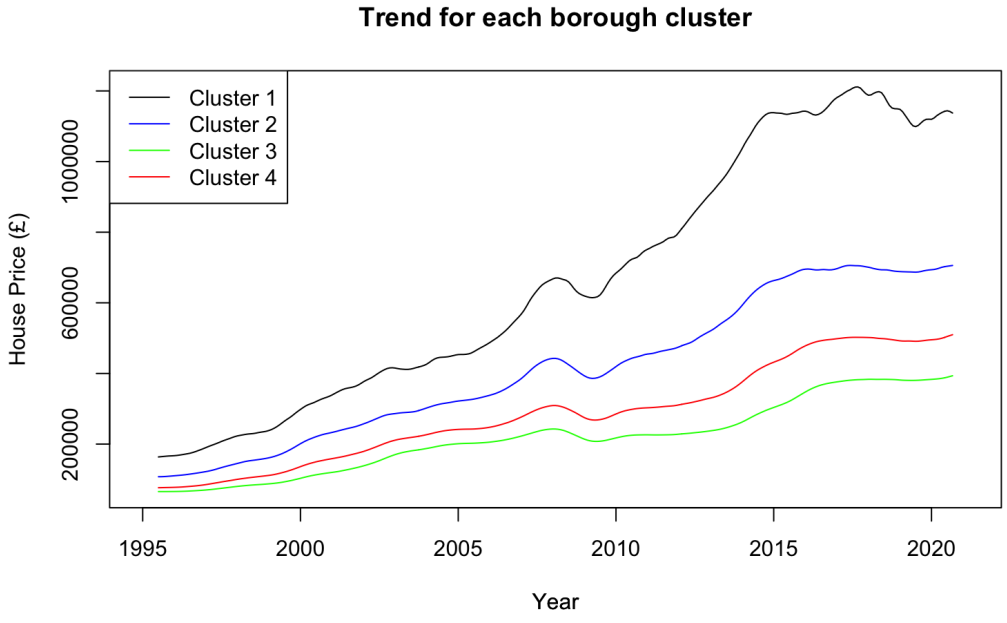

Figure 6.2: Trend for Each Borough Cluster

Typically a benchmark forecast is the average method, the naïve method or drift method. Since all four lines exhibit a slight upwards trend in the most recent months, we could expect a fairly consistent forecast from the drift method or the naïve method because a potential forecast would reflect the general trend of the years 1995 till 2016. We can rule out the average method for the whole data set since the all house prices have increased. By looking at the residual plots for each cluster when using the naïve method, we can see. The naïve method unsurprisingly fitted best for cluster 2 and 3 as they had the most linear increase in house price.

6.2 ARIMA models

By building an ARIMA model we have the results of the parameters below (Table 6.1).

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
ARIMA	(3,1,0)	(0,1,3)(0,0,2)[12]	(2,1,2)(0,0,1)[12]	(5,1,0)(0,0,2)[12]
With Drift?	Yes	Yes	Yes	Yes
Q*	52.191	40.440	42.401	50.386
p-value	0.0001	0.0018	0.0010	0.00002

Table 6.1: Table to show modelled cluster results

The results of the Ljung-Box test show that, for each cluster, the the residuals are distinguishable from a white noise series.

Below is the final ARIMA model Forecasts for each borough (6.3). The dark blue line represents the expected predicted value, the dark blue shading represents an 80% confidence interval and the dark grey shading represents the limits of the 95% confidence interval. There is a trend of, the more expensive the area, the greater the confidence intervals are. However, as I discussed previously, the clustering process may have superficially removed the variability of the individual trends.

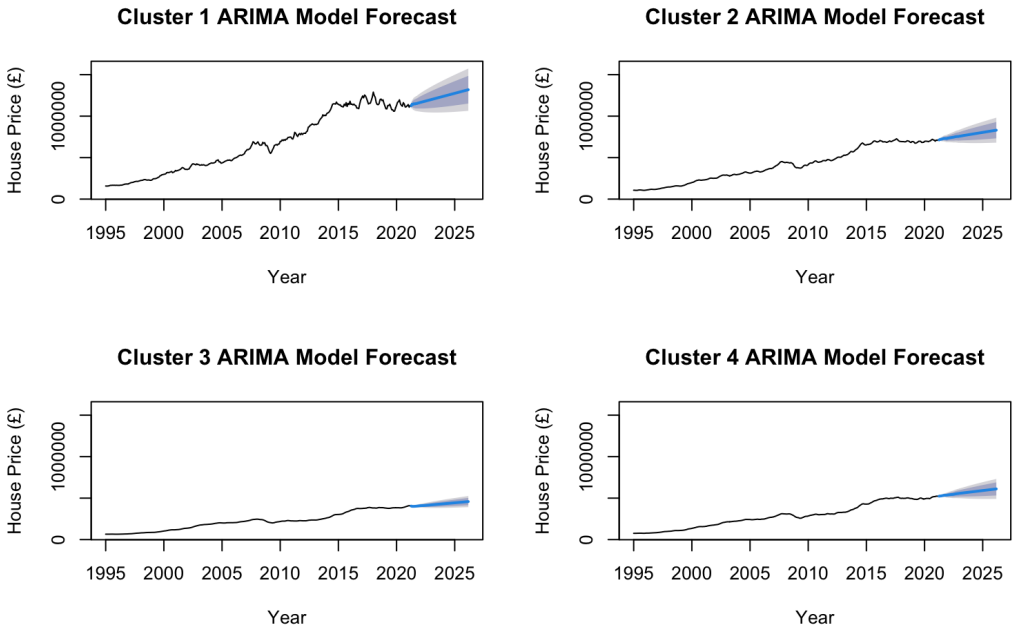

Figure 6.3: Clustered ARIMA models

Below (Table 6.2) shows the percentage increase over a five year period for each cluster, where the top three best performers had very similar percentage increases. So as per my forecasting, the boroughs expected to increase the most over the next five years are Kensington and Chelsea and Westminster.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
March 2021	£1,130,064	£714,542	£407,668	£524,382
March 2026	£1,318,623	£830,789	£458,208	£610,842
Predicted 5-Year Percentage Increase	16.7%	16.3%	12.4%	16.5%

Table 6.2: Table to show Predicted Price Changes

Chapter 7

CONCLUSION

This project was developed to ultimately answer 'Which London Borough House Prices will increase the Most?'. In doing so, I clustered all 32 boroughs into four clusters via traditional euclidean k-means clustering as well as k-means longitudinal clustering, producing an identical cluster subsets when optimised, indicating that these allocations were robust. If I had more time on this project it would obviously be useful to skip the clustering and produce forecasts for all the boroughs individually but that is beyond the scope of this project. Furthermore, if I were to repeat this project with more time, I would utilise transaction data and incorporate as many spatial variables to back up the time series forecasts.

With the more manageable four timeseries objects, I could apply several models of varying success in order to forecast the next five years. These included the basic benchmark methods of the average forecast, the Naïve method, and with drift. The final model I concluded with was an ARIMA model for each cluster predicting the next five years of the House Price Index for that cluster of boroughs. By calculating the percentage change between the March 2021 House Prices Index and my March 2026 predicted value, we discover the cluster which contains Kensington and Chelsea, and Westminster has the largest price increase of an estimated 16.7%. The risk associated with a five-year investment reveals the best risk vs reward cluster prediction is attributed to the cluster which contains the Northwestern boroughs.

The results of the forecast were slightly underwhelming. Analysis aside, my initial exploration into borough house prices indicated a quantifiable growth in gentrified boroughs. Despite this, it is still the Kensington and Chelsea which . Perhaps if this analysis was repeated in fifteen years, with updated HPI data we could produce more accurate forecasts and see gentrification in full effect.

My confidence in my forecasts, considering the future economic impact of Covid and Brexit after we saw the affect the 2008 credit crunch had on the price of houses, is. slightly wavered. That being said, I trust the price of central London properties will always recover due to the fixed limited supply.

I will have to come back to these forecasts in 5 years to see how well they performed.

BIBLIOGRAPHY

[1] 2011. URL: https://www.legislation.gov.uk/ukpga/1963/33/section/1.

[2] Dec. 2014. URL: https://www.gov.uk/government/organisations/valuation-office-agency/about/statistics.

[3] Sept. 2020. URL: https://www.buyassociation.co.uk/2020/09/07/latest-on-london-property-market-house-prices-and-investment/.

[4] Mar. 2020. URL: https://www.buyassociation.co.uk/2020/03/05/coronavirus-sparked-market-panic-investors-seek-safe-haven/.

[5] 2021. URL: https://worldpopulationreview.com/world-cities/london-population.

[6] 2021. URL: https://www.kfh.co.uk/south-west-london-and-surrey/wandsworth-london-borough/council-tax.

[7] 2021. URL: https://www.savills.co.uk/research_articles/229130/317586-0.

[8] 2021. URL: https://www.investopedia.com/terms/h/hedonicpricing.asp.

[9] Aug. 2021. URL: https://www.rightmove.co.uk/news/content/uploads/2021/08/Rightmove-House-Price-Index-16th-August-FINAL-NE.pdf.

[10] Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. 2017. URL: https://CRAN.R-project.org/package=gridExtra.

[11] Alex Bogin and William Doerner. "Property Renovations and Their Impact on House Price Index Construction". In: *Journal of Real Estate Research* 41 (Sept. 2019), pp. 249–283. DOI: 10.1080/10835547.2019.12091526.

[12] Royal Borough. *Council tax bands – www.kingston.gov.uk*. 2021. URL: https://www.kingston.gov.uk/council-tax/council-tax-bands/1.

[13] G. E. P. Box and G. C. Tiao. "Intervention Analysis with Applications to Economic and Environmental Problems". In: *Journal of the American Statistical Association* 70.349 (1975), pp. 70–79. ISSN: 01621459. URL: http://www.jstor.org/stable/2285379.

[14] Andrew P Bray and David M Diez. *OIdata: Data sets and supplements (OpenIntro)*. R package version 1.0. 2012. URL: https://CRAN.R-project.org/package=OIdata.

[15] Peter Ellis. *ggseas: 'stats' for Seasonal Adjustment on the Fly with 'ggplot2'*. R package version 0.5.4. 2018. URL: https://CRAN.R-project.org/package=ggseas.

[16] England. *Home ownership*. Feb. 2020. URL: https://www.ethnicity-facts-figures.service.gov.uk/housing/owning-and-renting/home-ownership/latest.

[17] Tal Galili. "dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering". In: *Bioinformatics* (2015). DOI: 10.1093/bioinformatics/btv428. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/22/3718/17122682/btv428.pdf. URL: https://academic.oup.com/bioinformatics/article/31/22/3718/240978/%20dendextend-an-R-package-for-visualizing-adjusting.

[18] Christopher Gandrud. *DataCombine: Tools for Easily Combining and Cleaning Data Sets*. R package version 0.2.21. 2016. URL: https://CRAN.R-project.org/package=DataCombine.

[19] Christophe Genolini et al. "kml and kml3d: R Packages to Cluster Longitudinal Data". In: *Journal of Statistical Software* 65.4 (2015), pp. 1–34. URL: http://www.jstatsoft.org/v65/i04/.

[20] Aaron Girardi and Joel Marsden. "A description of London's economy". In: (Mar. 2017).

[21] Ruth Glass. *London: aspects of change*. 3. MacGibbon & Kee, 1964.

[22] Garrett Grolemund and Hadley Wickham. "Dates and Times Made Easy with lubridate". In: *Journal of Statistical Software* 40.3 (2011), pp. 1–25. URL: http://www.jstatsoft.org/v40/i03/.

[23] *halifax house price index (hhpi) 2019 index manual for hhpi model introduced in 2019*. Sept. 2019. URL: https://content.markitcdn.com/corporate/Company/Files/DownloadDocument?cmsId=a60faba2343f486caf1a7abc0408

[24] Chris Hamnett. "Spatially Displaced Demand and the Changing Geography of House Prices in London, 1995–2006". In: *Housing Studies* 24.3 (2009), pp. 301–320. DOI: 10.1080/02673030902814580. eprint: https://doi.org/10.1080/02673030902814580. URL: https://doi.org/10.1080/02673030902814580.

[25] Chris Hamnett. *Unequal city: London in the global arena*. Routledge, 2004.

[26] Marco Helbich et al. "Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria". In: *Urban Studies* 51 (July 2013), pp. 1–22. DOI: 10.1177/0042098013492234.

[27] Toby Dylan Hocking. *directlabels: Direct Labels for Multicolor Plots*. R package version 2021.1.13. 2021. URL: https://CRAN.R-project.org/package=directlabels.

[28] Mark J Holmes, Jesús Otero, and Theodore Panagiotidis. "Climbing the property ladder: An analysis of market integration in London property prices". In: *Urban Studies* 55.12 (2018), pp. 2660–2681. DOI: 10.1177/0042098017692303. eprint: https://doi.org/10.1177/0042098017692303. URL: https://doi.org/10.1177/0042098017692303.

[29] Rob Hyndman et al. *forecast: Forecasting functions for time series and linear models*. R package version 8.14. 2021. URL: https://pkg.robjhyndman.com/forecast/.

[30] Alboukadel Kassambara and Fabian Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7. 2020. URL: https://CRAN.R-project.org/package=factoextra.

[31] Stephen Law. "Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London". In: *Cities* 60 (Feb. 2017), pp. 166–179. DOI: 10.1016/j.cities.2016.08.008.

[32] Sifei Lu et al. "A hybrid regression technique for house prices prediction". In: Dec. 2017, pp. 319–323. DOI: 10.1109/IEEM.2017.8289904.

[33] Martin Maechler et al. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.0 — For new features, see the 'Changelog' file (in the package source). 2019.

[34] Pablo Montero and José A. Vilar. "TSclust: An R Package for Time Series Clustering". In: *Journal of Statistical Software* 62.1 (2014), pp. 1–43. URL: http://www.jstatsoft.org/v62/i01/.

[35] Usue Mori, Alexander Mendiburu, and Jose A. Lozano. "Distance Measures for Time Series in R: The TSdist Package". In: *R journal* 8.2 (2016), pp. 451–459. URL: https://journal.r-project.org/archive/2016/RJ-2016-058/index.html.

[36] Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2. 2014. URL: https://CRAN.R-project.org/package=RColorBrewer.

[37] Tejvan Pettinger. *Definition of the housing market - Economics Help*. 2013. URL: https://www.economicshelp.org/blog/glossary/definition-of-the-housing-market/#:~:text=Definitions%20%5C%20related%5C%20to%5C%20housing%5C%20market,real%5C%20house%5C%20prices%5C%20rose%5C%208%5C%25.

[38] Tejvan Pettinger. *Factors that affect the housing market - Economics Help*. Oct. 2019. URL: https://www.economicshelp.org/blog/377/housing/factors-that-affect-the-housing-market/.

[39] Henry O. Pollakowski and Traci S. Ray. "Housing Price Diffusion Patterns at Different Aggregation Levels: An Examination of Housing Market Efficiency". In: *Journal of Housing Research* 8.1 (1997), pp. 107–124. ISSN: 10527001. URL: http://www.jstor.org/stable/24833634.

[40] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[41] William Rayburn, Michael Devaney, and Richard Evans. "A Test of Weak-Form Efficiency in Residential Real Estate Returns". In: *Real Estate Economics* 15.3 (Sept. 1987), pp. 220–233. DOI: 10.1111/1540-6229.00429. URL: https://ideas.repec.org/a/bla/reesec/v15y1987i3p220-233.html.

[42] HM Land Registry. *About the UK House Price Index*. June 2016. URL: https://www.gov.uk/government/publications/about-the-uk-house-price-index/about-the-uk-house-price-index.

[43] HM Land Registry. *Price Paid Data*. July 2014. URL: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads.

[44] Daryl Rozario. *Regional, sub-regional and local Gross Value Added estimates for London, 1998-2017 – London Datastore*. Feb. 2019. URL: https://data.london.gov.uk/blog/regional-sub-regional-and-local-gross-value-added-estimates-for-london-1998-2017/.

[45] Bob Rudis, Ben Bolker, and Jan Schulz. *ggalt: Extra Coordinate Systems, 'Geoms', Statistical Transformations, Scales and Fonts for 'ggplot2'*. R package version 0.4.0. 2017. URL: https://CRAN.R-project.org/package=ggalt.

[46] Adrian Trapletti and Kurt Hornik. *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-48. 2020. URL: https://CRAN.R-project.org/package=tseries.

[47] J. Stuart Wabe. "A Study of House Prices as a means of Establishing the Value of Journey Time, the Rate of Time Preference and the Valuation of some Aspects of Environment in the London Metropolitan Region". In: *Applied Economics* 3.4 (1971), pp. 247–255. DOI: 10.1080/00036847100000012. eprint: https://doi.org/10.1080/00036847100000012. URL: https://doi.org/10.1080/00036847100000012.

[48] *What is Acorn?* 2013. URL: https://acorn.caci.co.uk/what-is-acorn.

[49] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[50] Hadley Wickham. “Reshaping Data with the reshape Package”. In: *Journal of Statistical Software* 21.12 (2007), pp. 1–20. URL: http://www.jstatsoft.org/v21/i12/.

[51] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0. 2019. URL: https://CRAN.R-project.org/package=stringr.

[52] Hadley Wickham and Dana Seidel. *scales: Scale Functions for Visualization*. R package version 1.1.1. 2020. URL: https://CRAN.R-project.org/package=scales.

[53] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.3. 2021. URL: https://CRAN.R-project.org/package=dplyr.

[54] Hadley Wickham et al. “Welcome to the tidyverse”. In: *Journal of Open Source Software* 4.43 (2019), p. 1686. DOI: 10.21105/joss.01686.

[55] Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.4. 2021. URL: https://CRAN.R-project.org/package=kableExtra.

Appendix A

APPENDIX

A.1 Data

The House Price Index dataset I used is for this project can be found on the HM Land registry website: https://www.gov.uk/government/publications/about-the-uk-house-price-index/about-the-uk-house-price-index [9]

A.2 Figures

Libraries

```
# Packages ----
# Loading in libraries
library(tidyverse)
library(dplyr)
library(ggplot2)
library(RColorBrewer)
library(OIdata)
library(kableExtra)
library(TSdist)
library(cluster)     # clustering algorithms
library(factoextra) # clustering visualization
library(dendextend) # for comparing two dendrograms
library(ggseas)
library(DataCombine)
library(scales)
library(tseries)
library(forecast)
library(kml)
library(ggalt)
library(lubridate)
library(reshape2)
library(directlabels)
library(gridExtra)
library(stringr)
```

```
library(TSclust)
set.seed(1234)
```

Figure 1.1: Percentage Increase in House Price for each London borough between 1995 and 2021

```
# Data Wrangling House price index for London ------
## Loading data set
HousePriceIndex <- read.csv("Data/UK_House_price_index.csv")
HousePriceIndex
## Selecting only the London Boroughs and renaming Data column
AveragePrice <- HousePriceIndex[-1, c(1,3:34)] %>%
  rename(Date = X)
# Convert to dataframe
AveragePrice <- as.data.frame(apply(AveragePrice, 2, as.numeric))

# Percentage Change Heatmap ------
## Selecting first and last column
AveragePrice1_315 <- AveragePrice[c(1, 315),2:34]
## Transposing dataframe
tAveragePrice1_315 <- t(AveragePrice1_315)
## Naming the first and last columns "Start" and "Finish" respectfully
colnames(tAveragePrice1_315) <- c("Start", "Finish")

## Adding Percentage Change Column
tAveragePrice1_315 <- within(as.data.frame(tAveragePrice1_315),
PercentChange <- ((Finish - Start)/Start)*100)

##Combining borough coordinates with percentage change data
tAveragePrice1_315 <- cbind(Borough = rownames(tAveragePrice1_315),
tAveragePrice1_315)
rownames(tAveragePrice1_315) <- 1:nrow(tAveragePrice1_315)

# Assigning variable 'Names" with list of borough names
Names <- as.data.frame(unique(london_boroughs$name))

# Assigning Names variable to the first and last entry in the time series
PercentageChange <- cbind(Names, tAveragePrice1_315[2:33,4]) %>%
  rename(Borough = "unique(london_boroughs$name)",
  PercentageChange = "tAveragePrice1_315[2:33,␣4]")
LB <- london_boroughs %>%
  rename(Borough = name)
HeatMap <- merge(LB,PercentageChange,by="Borough",all=TRUE)


# Create Heatmap
ggplot() +
  geom_polygon(data=HeatMap,
               aes(x=x, y=y, group = Borough, fill = PercentageChange),
               colour = "white") +
  scale_fill_gradientn(colours = brewer.pal(9, "Reds"),
                       name = "Percentage␣Increase",
                       labels = c("500%", "600%", "700%", "800%", "900%"),
```

```
                        breaks = c(500, 600, 700, 800, 900)
  ) +
  geom_sf_label_repel(data = HeatMap, aes(label = Borough), force = 50) +
  # ggtitle("Percentage Increase House Price Since 1995") +
  theme_void()


ggsave("Heatmap.png", plot = last_plot(), scale = 1, dpi = 400)
```

Figure 1.2: Sold House Prices for each month between 1995 and 2021, broken down by borough

```
# Data Wrangling House price index for London ------
## Loading data set
HousePriceIndex <- read.csv("Data/UK_House_price_index.csv")
HousePriceIndex
## Selecting only the London Boroughs and renaming Data column
AveragePrice <- HousePriceIndex[-1, c(1,3:34)] %>%
  rename(Date = X)
# Convert to dataframe
AveragePrice <- as.data.frame(apply(AveragePrice, 2, as.numeric))

# Percentage Change Heatmap ------
## Selecting first and last column
AveragePrice1_315 <- AveragePrice[c(1, 315),2:34]
## Transposing dataframe
tAveragePrice1_315 <- t(AveragePrice1_315)
## Naming the first and last columns "Start" and "Finish" respectfully
colnames(tAveragePrice1_315) <- c("Start", "Finish")

## Adding Percentage Change Column
tAveragePrice1_315 <- within(as.data.frame(tAveragePrice1_315),
PercentChange <- ((Finish - Start)/Start)*100)

##Combining borough coordinates with percentage change data
tAveragePrice1_315 <- cbind(Borough = rownames(tAveragePrice1_315),
tAveragePrice1_315)
rownames(tAveragePrice1_315) <- 1:nrow(tAveragePrice1_315)

# Assigning variable 'Names" with list of borough names
Names <- as.data.frame(unique(london_boroughs$name))

# Assigning Names variable to the first and last entry in the time series
PercentageChange <- cbind(Names, tAveragePrice1_315[2:33,4]) %>%
  rename(Borough = "unique(london_boroughs$name)",
  PercentageChange = "tAveragePrice1_315[2:33,␣4]")
LB <- london_boroughs %>%
  rename(Borough = name)
HeatMap <- merge(LB, PercentageChange ,by="Borough",all=TRUE)


# Create Heatmap
ggplot() +
  geom_polygon(data=HeatMap,
               aes(x=x, y=y, group = Borough, fill = PercentageChange),
               colour = "white") +
```

```
  scale_fill_gradientn(colours = brewer.pal(9, "Reds"),
                       name = "Percentage Increase",
                       labels = c("500%", "600%", "700%", "800%", "900%"),
                       breaks = c(500, 600, 700, 800, 900)
  ) +
  geom_sf_label_repel(data = HeatMap, aes(label = Borough), force = 50) +
  # ggtitle("Percentage Increase House Price Since 1995") +
  theme_void()


ggsave("Heatmap.png", plot = last_plot(), scale = 1, dpi = 400)
```

Figure 5.1: Optimum Number of clusters

```
# Standard K means clustering without scaling ----
## Selecting only the London Boroughs and renaming Data column
AveragePrice <- HousePriceIndex[-1, c(1,3:34)] %>%
  rename(Date = X)
# Convert to dataframe
AveragePrice <- as.data.frame(apply(AveragePrice, 2, as.numeric))
# Set to integer
AveragePrice[] <- lapply(AveragePrice, as.integer)
# Transpose dataframe, select data from Jan 1995 to Mar 2021
tAveragePrice <- t(AveragePrice)[,1:315]
# Rename columns to original column names
tAveragePrice <- cbind(Borough = rownames(tAveragePrice), tAveragePrice)
rownames(tAveragePrice) <- 1:nrow(tAveragePrice)
colnames(tAveragePrice) <- c("Borough",1:315)
tAveragePrice <- tAveragePrice[2:33,]

# Assign all numeric columns
tAveragePrice <- tAveragePrice %>%
  as.data.frame() %>%
  mutate_at(vars(-Borough), as.numeric)

wss <- map_dbl(1:7, ~{kmeans(select(tAveragePrice, -Borough), .,
nstart=1,iter.max = 50 )$tot.withinss})
n_clust <- 1:7

# Graph each optimisation method to compare
elbow <- fviz_nbclust(select(tAveragePrice, -Borough), kmeans,
method = "wss") +
  geom_vline(xintercept = 4, linetype = 2, color = "#5482b0") +
  labs(subtitle = "Elbow method")

silhouette <- fviz_nbclust(select(tAveragePrice, -Borough),
                           kmeans,
                           method = "silhouette") +
  labs(subtitle = "Silhouette method")

gap_stat <- fviz_nbclust(select(tAveragePrice, -Borough),
                         kmeans,
                         nstart = 25,
                         method = "gap_stat",
                         nboot = 500) +
```

```
  labs(subtitle = "Gap statistic method")

grid.arrange(elbow, silhouette, gap_stat, ncol = 3)
```

Figure 5.2: Sold House Prices for each Borough - Simply Clustered by K-Means.

```
# Standard K means clustering without scaling ----
## Selecting only the London Boroughs and renaming Data column
AveragePrice <- HousePriceIndex[-1, c(1,3:34)] %>%
  rename(Date = X)
# Convert to dataframe
AveragePrice <- as.data.frame(apply(AveragePrice, 2, as.numeric))
# Set to integer
AveragePrice[] <- lapply(AveragePrice, as.integer)
# Transpose dataframe, select data from Jan 1995 to Mar 2021
tAveragePrice <- t(AveragePrice)[,1:315]
# Rename columns to original column names
tAveragePrice <- cbind(Borough = rownames(tAveragePrice), tAveragePrice)
rownames(tAveragePrice) <- 1:nrow(tAveragePrice)
colnames(tAveragePrice) <- c("Borough",1:315)
tAveragePrice <- tAveragePrice[2:33,]

# Assign all numeric columns
tAveragePrice <- tAveragePrice %>%
  as.data.frame() %>%
  mutate_at(vars(-Borough), as.numeric)

wss <- map_dbl(1:7, ~{kmeans(select(tAveragePrice, -Borough), .,
nstart=1,iter.max = 50 )$tot.withinss})
n_clust <- 1:7

# Graph each optimisation method to compare
elbow <- fviz_nbclust(select(tAveragePrice, -Borough), kmeans,
method = "wss") +
  geom_vline(xintercept = 4, linetype = 2, color = "#5482b0") +
  labs(subtitle = "Elbow method")

silhouette <- fviz_nbclust(select(tAveragePrice, -Borough),
                           kmeans,
                           method = "silhouette") +
  labs(subtitle = "Silhouette method")

gap_stat <- fviz_nbclust(select(tAveragePrice, -Borough),
                         kmeans,
                         nstart = 25,
                         method = "gap_stat",
                         nboot = 500) +
  labs(subtitle = "Gap statistic method")

grid.arrange(elbow, silhouette, gap_stat, ncol = 3)


clusters <- kmeans(select(tAveragePrice, -Borough), centers = 4)

(centers <- rownames_to_column(as.data.frame(clusters$centers), "cluster"))
```

```
tAveragePrice <- tAveragePrice %>%
  mutate(cluster = clusters$cluster)

Years <- format(seq(as.Date("1995-1-1"), as.Date("2021-3-1"),
by = "months"), format="%d-%m-%Y")

colnames_tAveragePrice <- as.character(c("Borough",Years,"cluster"))
colnames(tAveragePrice) <- colnames_tAveragePrice


tAveragePrice_long <- tAveragePrice %>%
  pivot_longer(cols=c(-Borough, -cluster), names_to = "Date",
  values_to = "Price")

tAveragePrice_long$Date <- as.Date(tAveragePrice_long$Date ,
format="%d-%m-%Y")

colnames_centers <- as.character(c("cluster",Years))
colnames(centers) <- colnames_centers


centers_long <- centers %>%
  pivot_longer(cols = -cluster, names_to = "Date", values_to = "Price")


centers_long$Date <- as.Date(centers_long$Date, format="%d-%m-%Y")


ggplot() +
  geom_line(data = tAveragePrice_long,
  aes(y = Price, x = Date, group = Borough), colour = "#a3c4dc") +
  facet_wrap(~cluster, nrow = 1) +
  geom_line(data = centers_long,
  aes(y = Price, x = Date, group = cluster), col = "#0e668b", size = 2) +
  labs(title = "Sold␣House␣Prices") +
  theme(plot.title = element_text(hjust = 0.5)) +
  facet_grid(~ cluster, labeller = label_both) +
  xlab("Year") +
  ylab("Price") +
  theme_linedraw()
```

Figure 5.3: Kml Cluster Results

```
HousePriceIndex <- read.csv("Data/UK_House_price_index.csv")
AveragePrice <- HousePriceIndex[-1, c(1,3:34)] %>%
  rename(Date = X)

# Convert to dataframe
AveragePrice <- as.data.frame(apply(AveragePrice, 2, as.numeric))

# ensure data is in a data frame (cld also accepts a matrix but
no other data type)
AveragePricedf <- t(as.data.frame(AveragePrice))[2:33,1:315]
```

```
# create clusterLongData object
AveragePricedf <- kml::cld(AveragePricedf, timeInData = 1:315, maxNA = 0,
varNames = "HousePrice")

# inspect object
class(AveragePricedf)
AveragePricedf
# run kml with 2-6 clusters and five redrawings for each
kml::kml(AveragePricedf, nbRedrawing = 5)

# on Mac or Linux you may need to enable X11 console first

if (.Platform$OS.type != "windows") {
  X11(type = "Xlib")
}

# run choice

kml::choice(AveragePricedf)
AveragePricedfClusters <- read.csv("AveragePricedf-C4-3-Clusters.csv")
AveragePricedfClusters <-
str_sub(AveragePricedfClusters$idAll.clusters,-1,-1)
AveragePricedfClusters <- cbind(Names, AveragePricedfClusters)

AveragePricedfClusters[order(AveragePricedfClusters$AveragePricedfClusters),]
```

Figure 5.4: Heatmap of London Boroughs by clusters from fig. 4.1

```
## Clustered Map ------

ClusteredBoroughs <- tAveragePrice[,c(1,317)]

ClusteredBoroughs <- cbind(Names, ClusteredBoroughs)[,c(1,3)] %>%
  rename(Borough = "unique(london_boroughs$name)", Cluster = "cluster")

ClusteredBoroughs <- merge(LB, ClusteredBoroughs,by="Borough",all=TRUE)


ClusteredBoroughs[] <- lapply(ClusteredBoroughs, as.character)


ClusteredBoroughs <- transform(ClusteredBoroughs, x = as.numeric(x),
y = as.numeric(y))


aggregate(ClusteredBoroughs$x, list(ClusteredBoroughs$Borough), FUN=mean)

ClusteredBoroughs %>%
  group_by(Borough) %>%
  summarise_at(vars(x), list(x = mean))


aggregate(x = ClusteredBoroughs$x,                # Specify data column
          by = list(ClusteredBoroughs$Borough),
```

```
          # Specify group indicator
          FUN = mean)

# Create Clustered map
p <- ggplot() +
  geom_polygon(data=ClusteredBoroughs,
               aes(x=x, y=y, group = Borough, fill = as.character(Cluster)),
               colour = "white") +
  # geom_text(data=centroids.df,
  aes(label = , x = Longitude, y = Latitude))+
  scale_fill_manual(values = c("cadetblue3",
  "seagreen", "seagreen3", "steelblue4"), name = "Cluster") +
  # ggtitle("Clustered House Prices") +
  theme_void()

direct.label(p,"Borough")
```

Figure 6.1: Timeseries for each borough cluster

```
#### Clustered Time Series
## Cluster 1 ----
{Cluster1_ts <- centers %>%
  t()
Cluster1_ts <- as.numeric(Cluster1_ts[2:316,1])
Cluster1_ts <- ts(data = Cluster1_ts,
start=c(1995, 1), end=c(2021, 3), frequency=12)

Cluster2_ts <- centers %>%
    t()
  Cluster2_ts <- as.numeric(Cluster2_ts[2:316,2])
  Cluster2_ts <- ts(data = Cluster2_ts, start=c(1995, 1),
  end=c(2021, 3), frequency=12)

Cluster3_ts <- centers %>%
    t()
  Cluster3_ts <- as.numeric(Cluster3_ts[2:316,3])
  Cluster3_ts <- ts(data = Cluster3_ts, start=c(1995, 1),
  end=c(2021, 3), frequency=12)

Cluster4_ts <- centers %>%
    t()
  Cluster4_ts <- as.numeric(Cluster4_ts[2:316,4])
  Cluster4_ts <- ts(data = Cluster4_ts, start=c(1995, 1), end=c(2021, 3), frequency=12)
}

# Plot Raw Time Series for Each Cluster
ts.plot(Cluster1_ts,Cluster2_ts,Cluster3_ts,Cluster4_ts,
        gpars = list(col = c("black","blue","green","red")), xlab = "Year")
legend("topleft", legend = c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4" ),
       col = c("black","blue","green","red"), lty = 1)
title(main = "Timeseries for each borough cluster", xlab = "Year", ylab = "House Price" )
```

Figure 6.2: Trend for each borough cluster

```
# Timeseries decomposition
Cluster1_trend <- decompose(Cluster1_ts)$trend
Cluster2_trend <- decompose(Cluster2_ts)$trend
Cluster3_trend <- decompose(Cluster3_ts)$trend
Cluster4_trend <- decompose(Cluster4_ts)$trend



ts.plot(Cluster1_trend,Cluster2_trend,Cluster3_trend,Cluster4_trend,
        gpars = list(col = c("black","blue","green","red")), xlab = "Year")
legend("topleft",
legend = c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4" ),
                            col = c("black","blue","green","red"), lty = 1
       )
title(main = "Trend for each borough cluster",
xlab = "Year", ylab = "House Price" )
```

Figure 6.3: Clustered ARIMA models

```
## Fitting ARIMA model forecasts
Cluster1_fit <- auto.arima(Cluster1_ts)
Cluster2_fit <- auto.arima(Cluster2_ts)
Cluster3_fit <- auto.arima(Cluster3_ts)
Cluster4_fit <- auto.arima(Cluster4_ts)

plot(forecast(Cluster1_fit, 60),
     main = "Cluster 1 ARIMA Model Forecast",
     xlab = "Year",
     ylab = "House Price",
     ylim = c(60000,1600000)
       )
plot(forecast(Cluster2_fit, 60),
     main = "Cluster 2 ARIMA Model Forecast",
     xlab = "Year",
     ylab = "House Price",
     ylim = c(60000,1600000)
     )
plot(forecast(Cluster3_fit, 60),
     main = "Cluster 3 ARIMA Model Forecast",
     xlab = "Year",
     ylab = "House Price",
     ylim = c(60000,1600000)
     )
plot(forecast(Cluster4_fit,60),
     main = "Cluster 4 ARIMA Model Forecast",
     xlab = "Year",
     ylab = "House Price",
     ylim = c(60000,1600000)
     )
```