2018

SPATIAL ANALYSIS OF LONDON BOROUGH PROFILES & THEIR INFLUENCE ON HOUSE PRICES

DIVYA BALASUBRAMANIAN

CITY, UNIVERSITY OF LONDON

23 December 2018

1. Motivation, Data and Research Questions:

1.1. Motivation for Study

London is the prime hub for the working population of the United Kingdom (UK). Increasing number of people aim to buy a house and shift into a location that would be both affordable as well as meet other personal requirements, which makes this process difficult. The Government also launched the Help to Buy Scheme in 2013 to improve the buying capacity of the section of the society who are buying their first home. However, London has been in an overall housing crisis and in May 2018, the Mayor of London launched the London Housing Strategy to address this (*Greater London Authority, 2018*). Apart from affordability, there are several factors including accessibility to public transport and safety included as part of the improvement strategy. There are 33 boroughs in London (Fig.1) each having a different profile. Hence, it is a possibility that several factors influence the London housing domain. This study focuses on identifying if and how London borough profiles (henceforth: LBP), that include geographic, demographic and other features, influence [i] London house prices and [ii] boroughs to exhibit spatially similar behaviour.

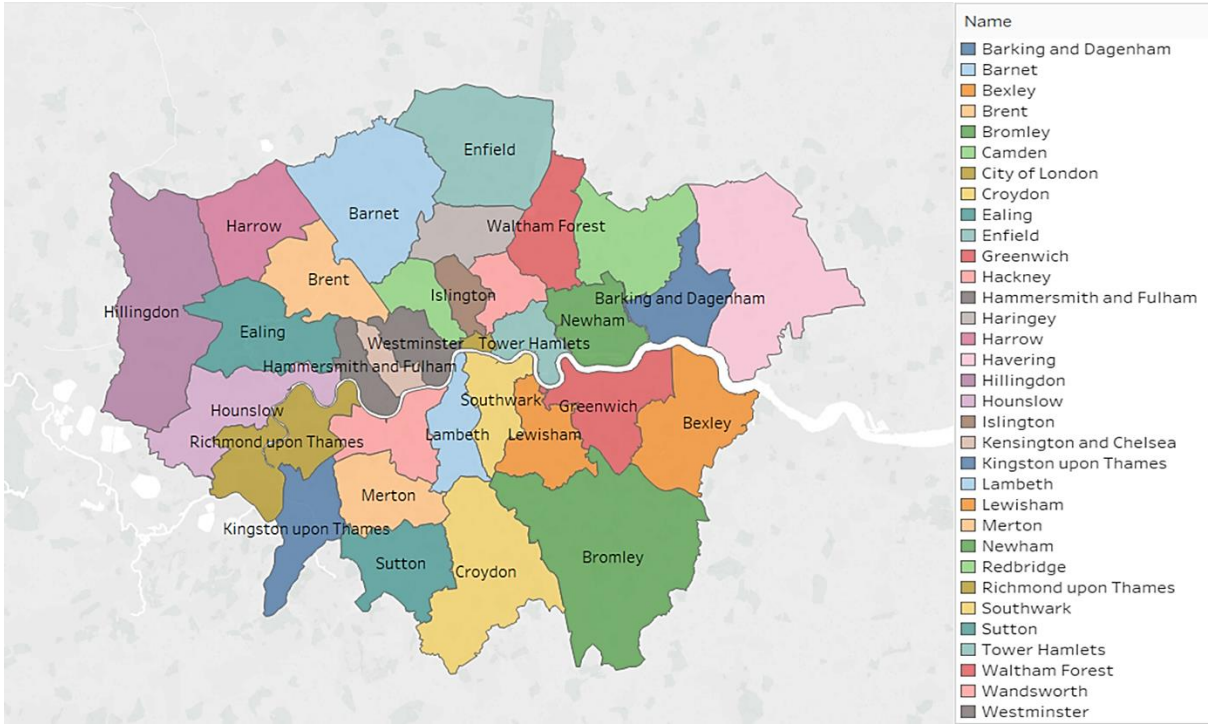

Fig.1: Choropleth map* with 33 boroughs of London included

1.2. Data and Suitability

This analysis works on a combination of data sources given its wide approach. LBP information consisting of 87 features across 33 London boroughs and the London house prices by borough from 2008 to 2018 were collected from London Datastore (*London Datastore, 2018*). Further data on critical information such as commute time (*Office for National Statistics, 2018*) and median employment income (*Nomis, 2018*) by borough were also added to the collation. All the data was combined and wrangled thoroughly to ensure it is suitable for further analysis. Feature derivation using statistical calculations were performed to derive features appropriate for the purpose of this analysis. For example, median house prices (henceforth: MHP) across the years considered were calculated to

make necessary comparisons. MHP were preferred over mean house prices due to its central tendency, as in the latter case the outliers would skew the data or create bias (*QFREB, 2018*). To feed curiosity, additional data on year-on-year supply of dwellings (houses) by borough were collected to perform extended exploratory data analysis (*Gov.uk, 2018*).

1.3. Research Questions

This analysis aims to answer the below questions:

- Do features of LBP influence house prices?
- Do we notice spatially similar behaviour among boroughs considering influential LBP features?

2. Tasks and Approach:

This analysis was divided into three main tasks that used combinations of computational and visual analytics techniques.

2.1. Task 1: Exploring the Data

In this task, the primary aim is to understand the relationships between house prices and LBP features. As a first step, the MHP from 2008-2018 were first visualised using line graphs to analyse if boroughs exhibited evident trends. Next, using domain knowledge, features which seemed to have a logical connection to house prices were explored using various visual techniques. For example, percentage area that is greenspace was compared to MHP using a choropleth map. It was an effective visualisation to easily identify if there is a direct impact as would be expected in an ideal scenario. Scatter plots were used to analyse if there is a similarity between MHP and median employment income in each borough. The exploratory data analysis was critical to identify subsequent analytical approaches. Although there were some interesting revelations about features and their influence on house prices in each borough, it was preferred that a systematic approach is taken to select features for deep dive analysis.

2.2. Task 2: Selecting Features from LBP to inspect if they influence MHP

The aim of this task is to answer the first question, if features of LBP influence house prices. For this, a correlation heatmap was plotted to identify features that had the highest correlation coefficients with MHP. Correlation matrices derive correlation coefficients which represent linear interdependence between variables. Correlation coefficients are in the range of -1 to 1. In this dataset the highest coefficient with MHP was 0.68, hence a threshold of -0.4 or lesser and 0.4 or greater was chosen for further investigation. Negative correlation coefficients were also selected as some features, such as commute time, are expected to have a negative correlation with house prices. I.e. lower the commute time, higher the house price. Following this, a set of eight features were selected for further investigation. The selected features were individually normalised to ensure further analysis is not affected due to outliers in each sample range. Linear regression analysis and regression plots were run on the selected features to validate if they indeed have a linear relationship with MHP. The p-value obtained from these regression analyses was used to validate the relationship [7].

2.3. Task 3: Investigate for Spatial Similarities in Borough Behaviours based on Influential Features

The aim of this task is to answer the second question, i.e. if there are spatial similarities in behaviours among boroughs, considering the influential LBP features. In order to effectively observe this, k-means clustering algorithm was utilised [8]. There were two approaches taken to facilitate this section of the analysis. The first approach involved applying k-means clustering on the features selected in the Task 2. The clusters were then projected on a choropleth map to visually represent the boroughs clustered together. This enabled isolation of sections for further investigation. The second approach involved application of Principal Component Analysis (PCA) on the LBP dataset to combat the curse of dimensionality [9]. PCA produces linear combinations of variables that have the strongest correlation. The principal components obtained from PCA were then fed into the k-means clustering algorithm to obtain new clusters. The number of clusters were maintained the same as that found ideal in the previous approach. These were then projected on a choropleth map once again. The choropleth maps from both approaches were then compared and analysed further. Two approaches were considered as the features in each case were selected using different computational techniques and it would be beneficial to compare the relevance of both to answer the second question.

3. Analytical Steps:

3.1. Task 1: Explore the data

The ideal starting point was to identify if there are borough-specific trends with respect to MHP. As seen in Fig.2, each borough reflects a consistent unique pattern except for City of London which has shown a drop in the MHP for 2018 compared to 2017. The MHP for all the boroughs plummeted in 2009 compared to 2008 during the financial crisis. Similarly post EU referendum results in 2016, all the boroughs have shown a minimal increase in MHP in 2017. This gives a clear pattern that MHP of all boroughs react similarly to economic implications.

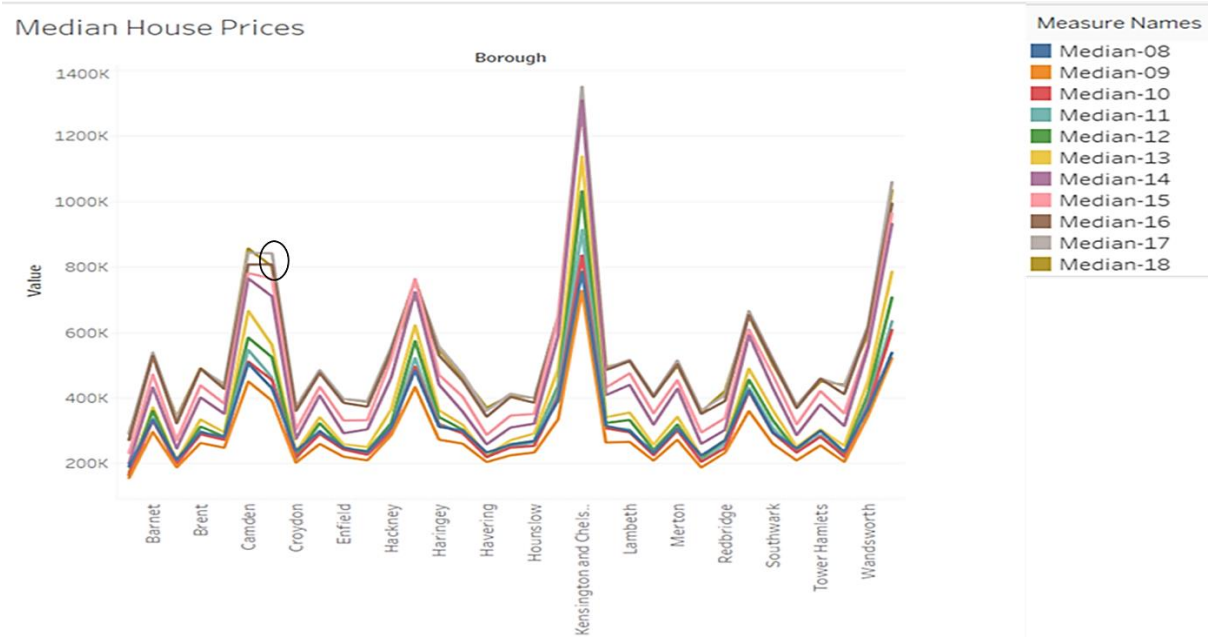

Fig.2: Median House Prices plotted from 2008-2018 for all London Boroughs

Following this, using domain knowledge, the features that were expected to have logical impact on MHP were identified and explored using visual techniques. The following list include examples of features examined, techniques used and corresponding findings.

- Crime rate vs MHP
  The crime rate and MHP were visualised on packed bubbles. While the colour depicts the extent of crime, the size of the bubbles depicts the MHP. It is expected that a higher crime rate would lower the MHP. However, as seen in Fig.3, there is no apparent trend between the total crime and MHP.
- Greenspace vs MHP
  Percentage of area that is greenspace is a feature of LBP. A choropleth map was used to visually analyse if the greenspace had an impact on the MHP. While the deepness of green colour depicts the amount of green space, MHP was labelled in each of the boroughs. The MHP were expected to be higher with greater percentage of greenspace. As seen in Fig.4, the boroughs do not reflect a direct relationship between greenspace and MHP.
- Median Employment Income vs MHP
  The median employment income is an important factor that is expected to determine MHP in boroughs. A scatter plot (Fig.5) projected to visualise this, highlighted a linear pattern with a few exceptions. The outliers including City of London and Kensington and Chelsea bring to notice that there are other factors playing into defining the MHP apart from median employment income.

Fig.3. Packed bubbles: size depicting MHP and colour depicting range of crime rate (Unnamed bubbles top to down: Waltham Forest, Barking and Dagenham, Greenwich)

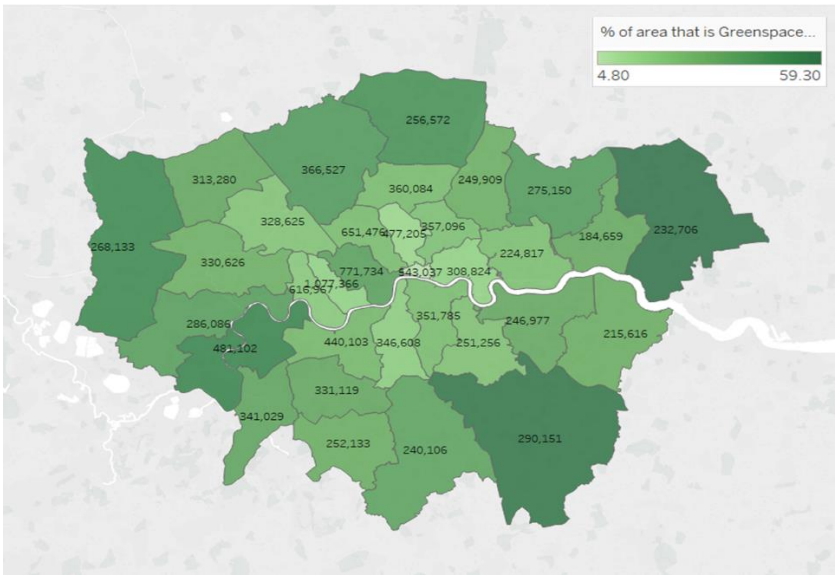

Fig.4. Choropleth map* depicting Greenspace (in deepness of green) and MHP labelled

Fig.5. Scatter plot of Median Employment Income vs MHP

Further combinations of features tested for correlation with MHP using visual techniques such as tree maps, bar graphs and line graphs have been included in the appendix.

The exploratory data analysis revealed that although some boroughs exhibit the expected pattern, majority of boroughs are influenced by a combination of features to determine their MHP. This was an ideal prompt to proceed to Task 2.

3.2. Task 2: Selecting Features from LBP to inspect if they influence MHP

The correlation heatmap in Fig.6 was plotted to isolate features with correlation coefficients of values -0.4 or lesser and 0.4 or greater for further analysis. The diagonal correlation plot in Fig.7 shows the list of features selected with their corresponding correlation coefficients.

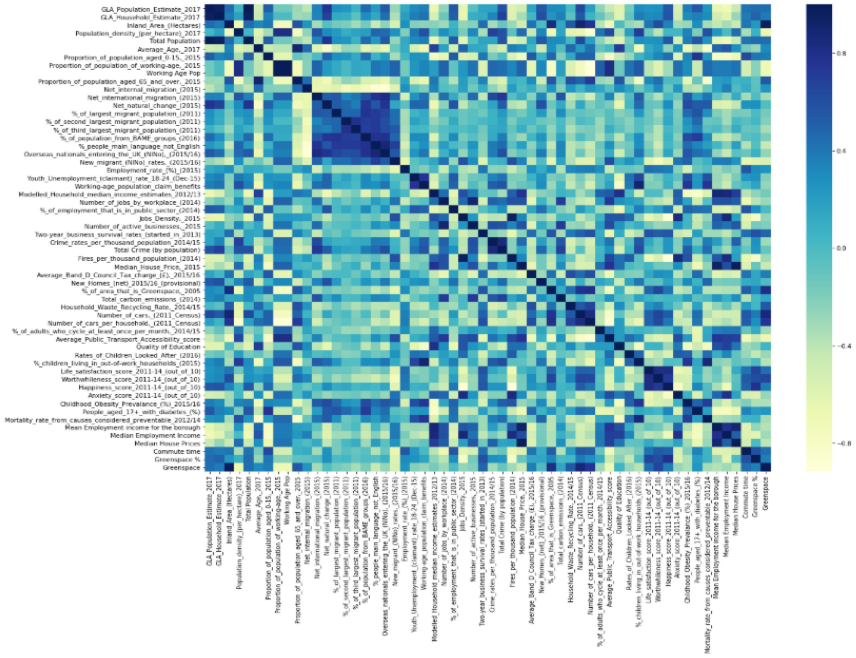

Fig.6. Correlation heatmap of London Borough profiles with MHP

Fig.7. Diagonal correlation plot of selected features

Where relevant, features were then individually normalised using MinMaxScaler from sklearn prior to further processing. Individual normalisation was undertaken as the measure of each feature is different.

Linear Regression analysis was run on each of the selected features and MHP to verify if there is a linear relationship between them and to determine the p-value which would define the statistical significance of the regression analysis. The accepted threshold of $p < 0.05$ [7] was considered to validate if there is a true linear relationship between the features and MHP. Table 1 lists the p-value in each case and Fig.8 shows the fitted line on a scatter plot between the features and MHP.

Feature	p-value
Number of active businesses	0.0007
Commute time	-0.003
Median Employment Income	0.002
Quality of Education	0.019
Public transport accessibility	0.00001
Working Age Population	0.012
Total Population	0.019
Number of Jobs	0.0017

Table 1: p-values following Linear Regression analysis of selected features with MHP

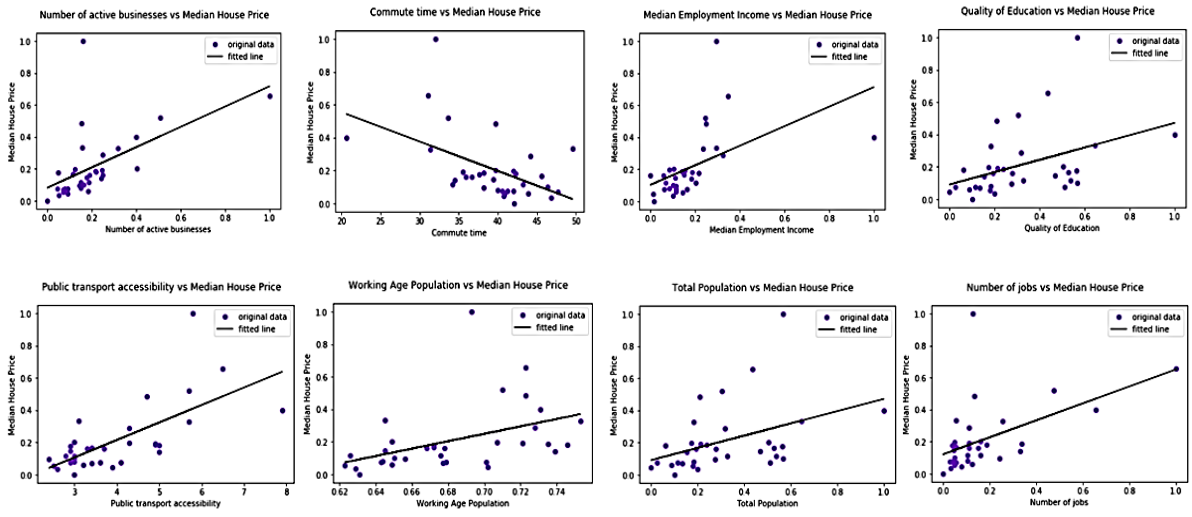

Fig.8. Scatter plot with fitted line following Linear Regression analysis of selected features with MHP

As seen in Table 1, the p-value was below the expected threshold for all the features selected and prove that there is a linear relationship between the selected features and MHP.

This task answers the first question that although at varying levels, features of the LBP do indeed have an influence on the MHP.

3.3. Task 3: Investigate for Spatial Similarities in Borough Behaviours based on Influential Features

As suggested in Section 2.3 above, there were two approaches taken to investigate spatial similarities considering LBP features.

Approach 1:

This approach involved application of an unsupervised machine learning model, k-means clustering using Euclidean distances on the features selected following regression analyses in the previous task. The selection of the number of clusters was an iterative process using human judgement assessing the data as well as optimal cluster parameters (discussed further in Section 5 below), settling finally with 5 clusters. The clusters were then projected on choropleth maps to visually analyse the partitions.

Fig.9 shows the scenarios of 3 and 4 clusters and Fig.10 shows the 5-clusters scenario. The 5-clusters scenario not only separates out Central London (i.e. City of London and Westminster) into individual clusters but also distinguishes Kensington and Chelsea from boroughs of cluster 2 which have lower median incomes. The findings are discussed in Section 4.

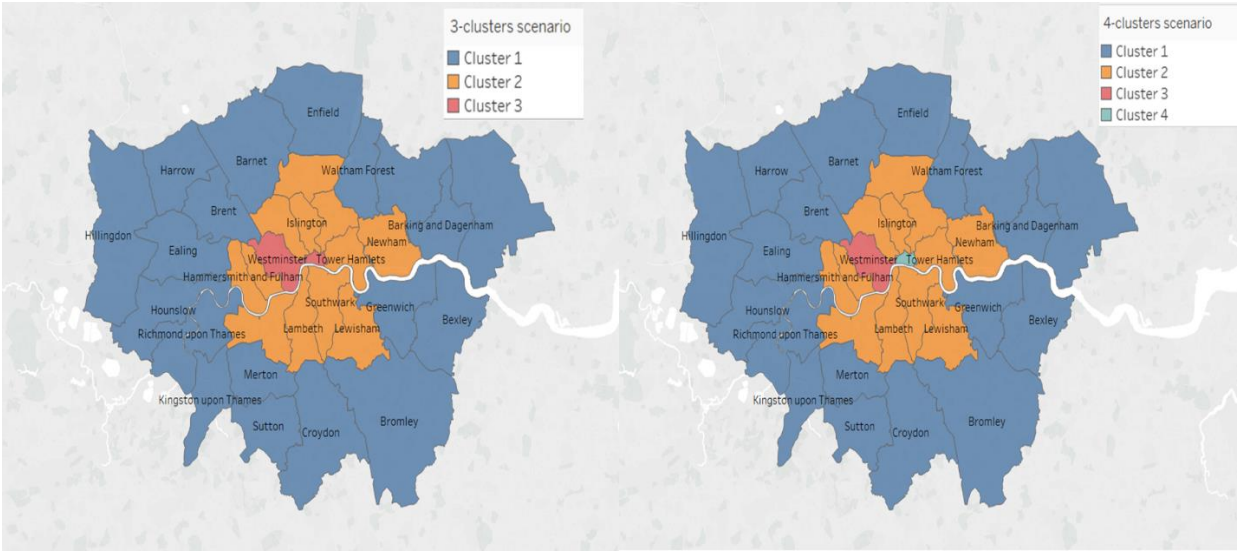

Fig.9. Choropleth maps* of 3-clusters scenario (left) and 4-clusters scenario (right)

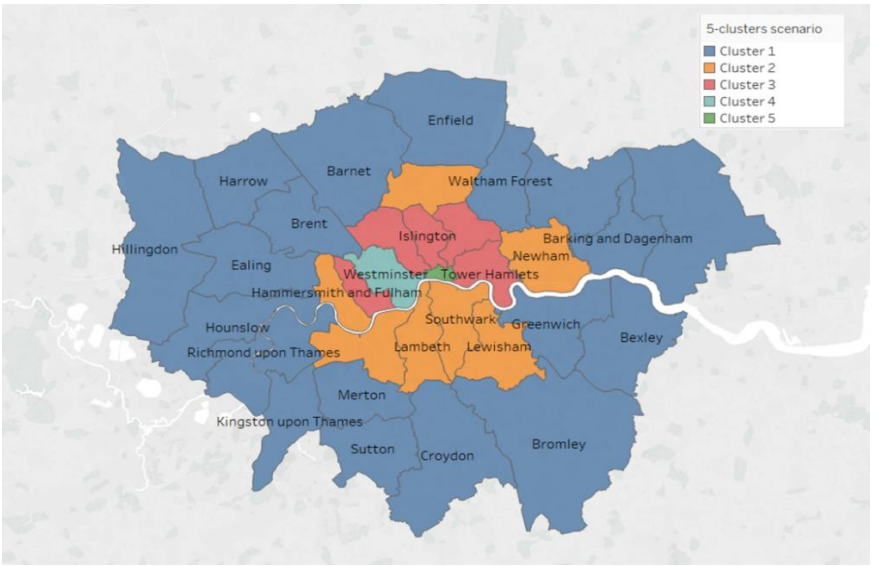

Fig.10. Choropleth map* of 5-clusters scenario

Approach 2:

The second approach uses PCA which is a linear dimensionality reduction method. Given the large number of features associated with this dataset, it was the ideal method to project the data to a lower dimensional space. The data was standardised using sklearn's scale function. The number of components were decided based on the explained variance ratio. It was preferred to obtain an explained variance ratio, i.e. the variance percentage explained by selected components, greater than 90%. The number of components were hence determined iteratively. Table 2 shows the explained variance ratio in each case tested.

Number of components	Explained variance ratio
2	34.6%
5	56.23%
10	74.3%
15	86.0%
20	93.2%

Table 2: Number of components and explained variance ratio combinations from PCA

The principal components obtained were then fed into a k-means clustering algorithm. The number of clusters were maintained the same as approach 1 to enable an unbiased comparison. The clusters were then projected on a choropleth map to visualise the results (Fig.11).

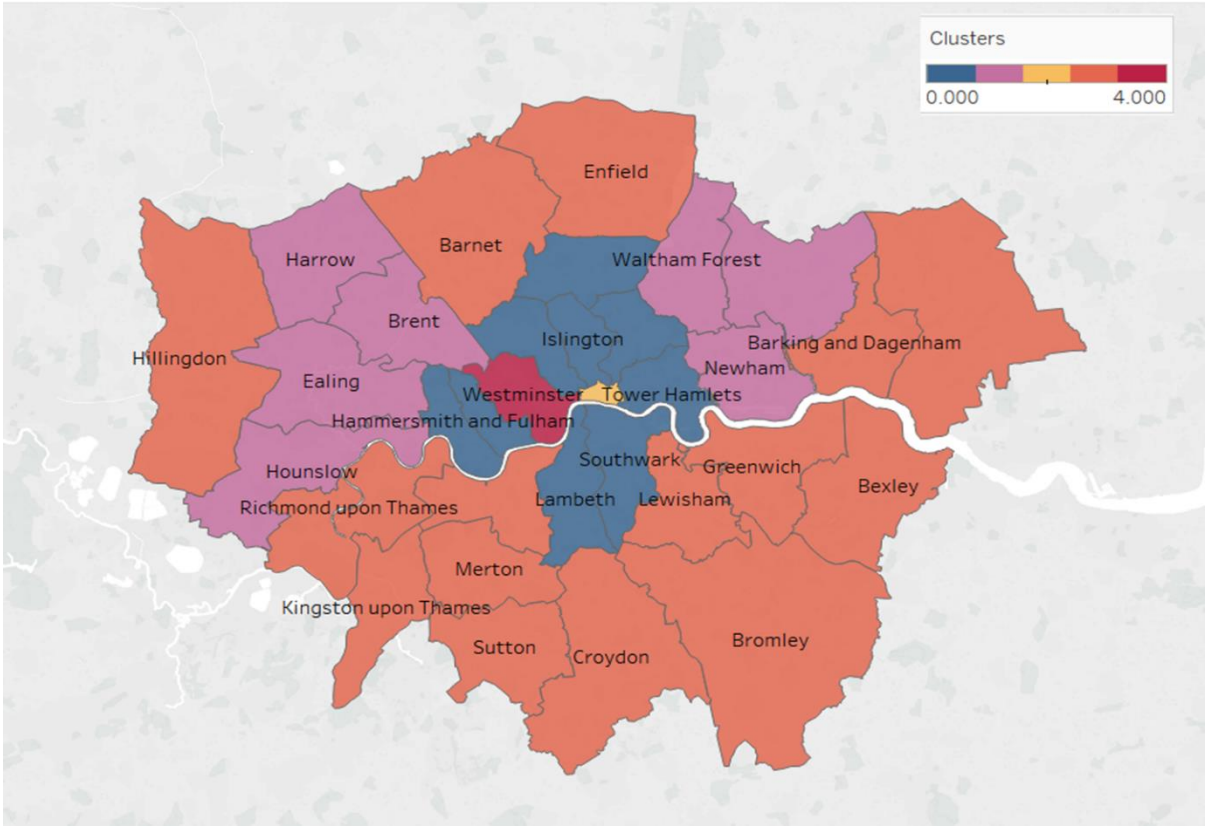

Fig.11. Choropleth map* projecting clusters following k-means on principal components

4. Findings:

Task 1 had a significant impact in defining the strategy for this analysis to answer the research questions. Exploring the data found that each feature had a disparate effect on the MHP. It was also seen that boroughs exhibited patterns of spatial similarity considering combinations of LBP features. For example, both Lambeth and Haringey had a mean commute time of approx. 42 minutes and similar quality of education, although they varied in MHP.

Analysis in Task 2 led to findings on the first research question, i.e. if features of LBP influence house prices. The application of linear regression enabled identification of linear effect of selected LBP

features on MHP, although to noticeably varying degrees. It validates that LBP features do in fact affect MHP. Since boroughs are diverse, several factors concerning the borough play into formulating the MHP for that region.

The two approaches undertaken in Task 3 are used to validate the second research question, i.e. if spatially similar behaviour is noticed among boroughs considering influential features of LBP. The clusters from the first approach (Fig.10) depict areas of similarity using the eight features selected in Task 2. Central London which includes City of London and Westminster were isolated into unique clusters partitioned from other regions. Domain knowledge validates that this is an expected scenario. City of London and Westminster are spatially similar, in that, both have high number of jobs and low commute time. However, Westminster has significantly larger population due to which it may have been clustered separately. Similarly, cluster 3 depicted commonality in accessibility to public transport score and average commute time.

In the second approach as well, the clusters isolated Central London like approach 1. The other clusters, however, consisted of overall different groups of boroughs although with some commonalities to those in approach 1. For example, Kensington and Chelsea and Camden were in the same cluster in both approaches, however approach 2 clustered Hammersmith and Fulham into the same cluster. Fig.17 and Fig.18 in the appendix list the top ten features considered in principal components 1 and 2. It can be noticed that, the only common feature considered in both the approaches was median employment income.

Although in both cases the number and selection of features varied, it is validated that there are indeed spatial similarities among boroughs considering LBP features. However, there are multiple combinations of features where the similarities are noticed and cannot be generalised.

5. Critical Reflection:

The analysis of LBP using visual and computational techniques has strengthened my knowledge and skill sets in this subject. Besides, the techniques further aided in understanding borough behaviours both with respect to their influence on house prices as well as spatial similarities. The implications of the findings of this report are that:

- London house prices are influenced by LBP features
- Boroughs exhibit spatially similar behaviours considering combinations of features

The exploration of data was useful to visualise the magnitude of differences exhibited by the boroughs for each feature considered. The use of visual techniques such as choropleths, tree maps and graphs expedited the process.

Additional exploratory analysis was also performed to test if there is a trend between demand-supply of dwellings and MHP. The test hypothesis was that if there is an increase in supply, there would be a decrease in MHP and vice versa. The percentage variance of dwellings added between 2015-16 and 2016-17 were projected on choropleth maps and compared to identify boroughs with large differences. These were then investigated against variance in percentage of MHP in those corresponding years. The findings however rejected the test hypothesis. The MHP in each borough are influenced differently by the supply of dwellings. The analysis is recorded in Fig.15 and Fig.16 in the appendix.

Linear regression analysis was ideal to confirm influence of selected features on the house prices. Although the p-values strengthened the relevance of statistical significance in the regression models,

as indicated in [10], the limitation is that p-values reflect the summary statistics of the studied samples only. Hence, for an individual interested in house prices, the p-values of the LBP features and MHP cannot be generalised as a measure to ascertain the degree of LBP influence. Were the scope of this project widened, it would perhaps be beneficial to apply multiple linear regression analysis considering all the influential features to visualise the impact as a combination.

Clustering was critical in visualising the partitioning of boroughs based on LBP features considered. The iterative nature of clustering provides an opportunity to enable various scenarios to be projected. The most ideal number of clusters, in this case, had to be finalised using a combination of human judgement and optimal parameters. The parameters studied included *between-group sum of squares* which measures the separation between clusters and *within-group sum of squares* which provides the level of cohesion within a cluster. Both parameters were most favourable in the 5-clusters scenario. The variance explained by the 5-clusters model was 72% as against 39% in the 3-clusters model and 67% in the 4-clusters model. However, the variance cannot be used as the only parameter in verifying our ideal scenario as the ratio (*between-group sum of squares/total sum of squares*) tends to increase with increasing number of clusters. Here, the human judgement with knowledge of the dataset and domain played in to affirm the 5-clusters model being the ideal scenario for this analysis. Increasing the number of clusters did not significantly differentiate the partitioning. The 5-clusters model separated Central London into two separate clusters and differentiated Kensington and Chelsea borough from those with larger difference in median employment income as seen in the 4-clusters model. Deep dive into the clusters allowed identification that although the boroughs clustered together exhibited similarities in selective feature combinations, there were outliers in some cases. This confirmed that boroughs do exhibit spatially similar behaviours however it would be dependent on the feature or combinations of features pertaining to requirements of an individual.

Application of PCA allowed a varied approach to study spatially similar behaviours. Although PCA is beneficial for dimensionality reduction, it takes factors based on linear combinations rather than relevance to requirements. It is seen that clustering of the principal components isolated Central London like the previous approach. There were also intersections in terms of boroughs partitioned into the remaining clusters however the clusters themselves were distinct from the previous approach. On investigating, it was found that the features selected in this case varied. This re-confirms our finding that different combinations of features influence the spatially similar behaviour exhibited by boroughs, differently.

These results can be utilised by borough councils and property agents to promote the respective areas. For example, Tower hamlets council can target the sale of houses to individuals earning a salary of around £27,300 and prefer an average commute time of 35 minutes. Borough councils can also use the impact of features of LBP to enhance diversity to improve their overall positioning.

The visual and computational techniques used in this analysis can be utilised in varied domains such as travel, retail or healthcare. Linear regression can extensively be used in all three domains for predictions of travellers, sales or patients of specific medical categories in future quarters or years. Choropleth maps are suitable for most spatial analysis as they visually represent the features across geographic locations. Clustering is an easily adaptable technique that can be used in domains such as travel or retail for customer segmentation. The accommodation industry which is tightly linked to the travel industry can benefit particularly from clustering to personalise their offering and elevate customer satisfaction. The healthcare industry would benefit from PCA as it is a heavily data-driven space with large amounts of data collected on a regular basis. PCA can help in focussed analysis in this field leading to better service provisions for patients.

In future analysis, it would be interesting to collect further details on the housing market such as sales volumes, house price index, types of houses and mortgage rates to investigate to what extent LBP features impact this market. It would be beneficial to perform an in-depth exploration of various combinations of features and observe the spatial similarities revealed by boroughs. This would be essential from the customer's perspective, as they would then have access to information of boroughs exhibiting favourable behaviours based on their preferences.

*Note: *Contains National Statistics data © Crown copyright and database right [2015] and Contains Ordnance Survey data © Crown copyright and database right [2015]*

REFERENCES:

[1] Greater London Authority (2018) *London Housing Strategy*

Available at: https://www.london.gov.uk/sites/default/files/2018_lhs_easy_read_fa.pdf

[2] London Datastore (2018) *London Borough Profiles and Atlas*

Available at: https://data.london.gov.uk/dataset/london-borough-profiles

(Accessed: 1 December 2018)

[3] London Datastore (2018) *UK House Price Index*

Available at: https://data.london.gov.uk/dataset/uk-house-price-index

(Accessed: 1 December 2018)

[4] Office for National Statistics (2018) *Travel to work methods and the time it takes to commute from home to work, Labour Force Survey, 2007 to 2016*

Available at:
https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/labourproductivity/adhocs/008005traveltoworkmethodsandthetimeittakestocommutefromhometoworklabourforcesurvey2007to2016

(Accessed: 1 December 2018)

[5] Nomis (2018) *annual survey of hours and earnings - workplace analysis*

Available at:
https://www.nomisweb.co.uk/query/construct/summary.asp?mode=construct&version=0&dataset=99&Session_GUID=%7b81BC49CC-FBE3-44B8-AEBA-40FCBA61818F%7d

(Accessed: 1 December 2018)

[6] QFREB (2018) *Why Median Price Rather Than Average Price*

Available at: http://www.fciq.ca/pdf/Carrefour/definitions/en/prix_median_a.pdf

(Accessed: 4 December 2018)

[7] Sedgwick, P. 2014, "Understanding P values", BMJ : British Medical Journal, vol. 349, no. jul11 3, pp. g4550-g4550.

[8] Dudeni, N., Holloway, J., Makhanya, S. & Koen,R. (2013) 'CLUSTERING OF HOUSING AND HOUSEHOLD PATTERNS USING 2011 POPULATION CENSUS', *55th Annual Conference of the South African Statistical Association for 2013*, Available at:
https://www.researchgate.net/publication/272158795_CLUSTERING_OF_HOUSING_AND_HOUSEHOLD_PATTERNS_USING_2011_POPULATION_CENSUS

[9] Abdi, H. & Williams, L.J. 2010, "Principal component analysis: Principal component analysis", Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 4, pp. 433-459.

[10] Dick, F. & Tevaearai, H. 2015, "Significance and Limitations of the p Value", European Journal of Vascular & Endovascular Surgery, vol. 50, no. 6, pp. 815-815.

[11] Gov.uk (2018) *Live tables on housing supply: net additional dwellings*

Available at: https://www.gov.uk/government/statistical-data-sets/live-tables-on-net-supply-of-housing

(Accessed: 15 December 2018)

APPENDIX

Task 1:

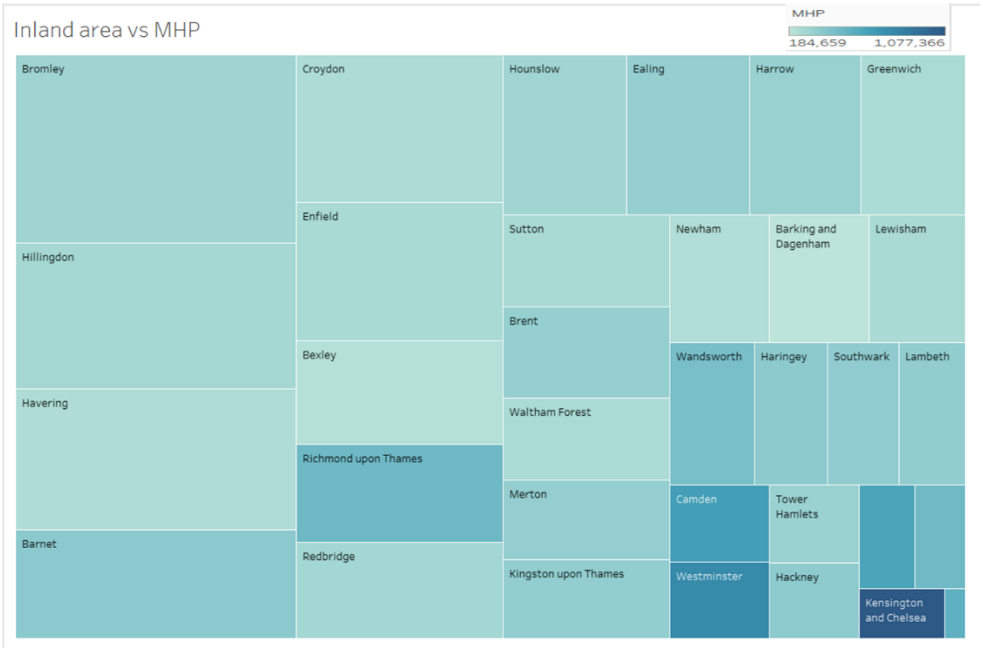

Fig.12. Tree map of Inland area of boroughs (by size of grid) and MHP (by colour)

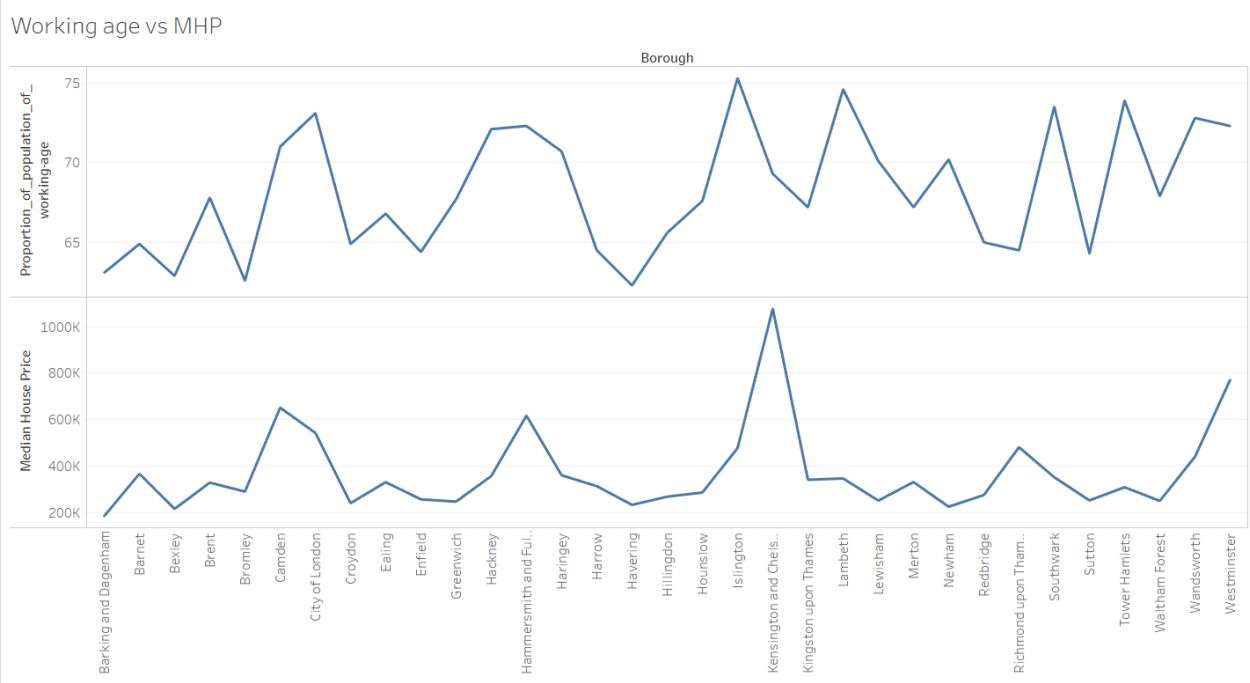

Fig.13. Line graph comparing percentage of working age population with MHP by borough

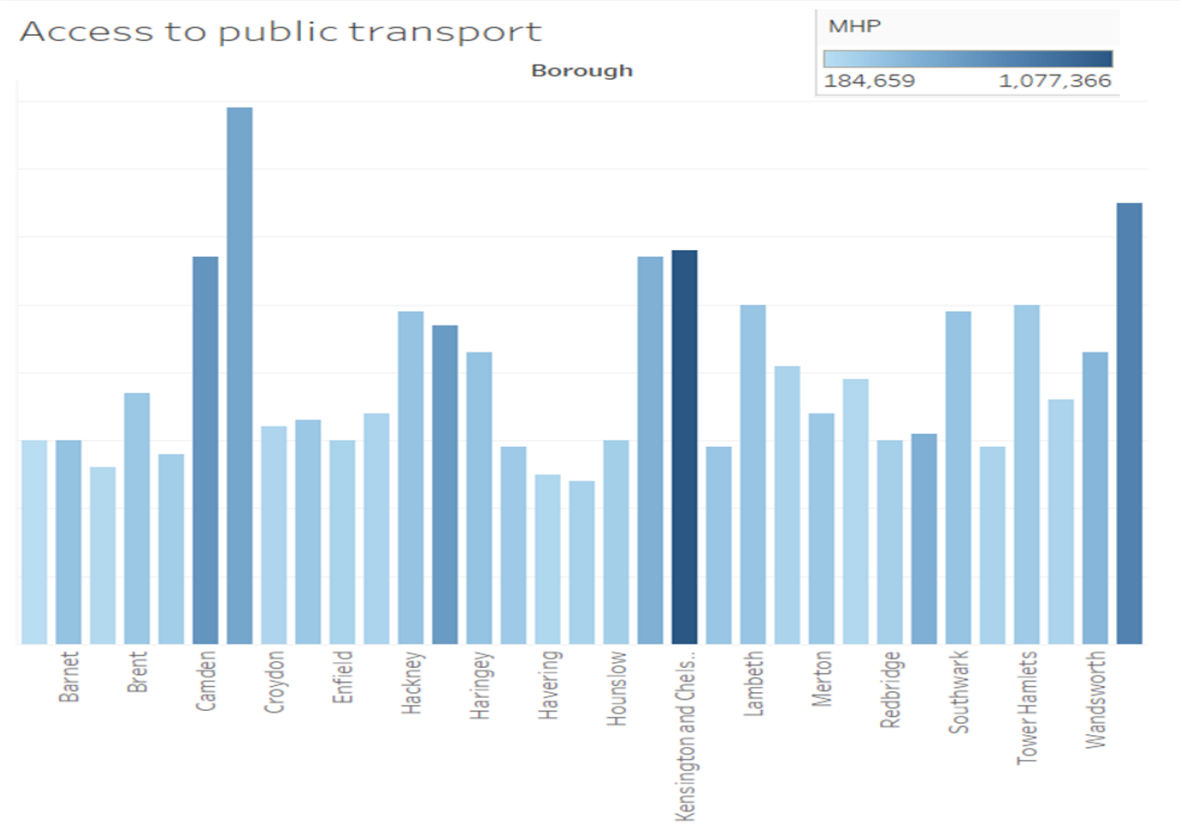

Fig.14. Bar graph of accessibility to public transport by borough coloured by MHP for comparison

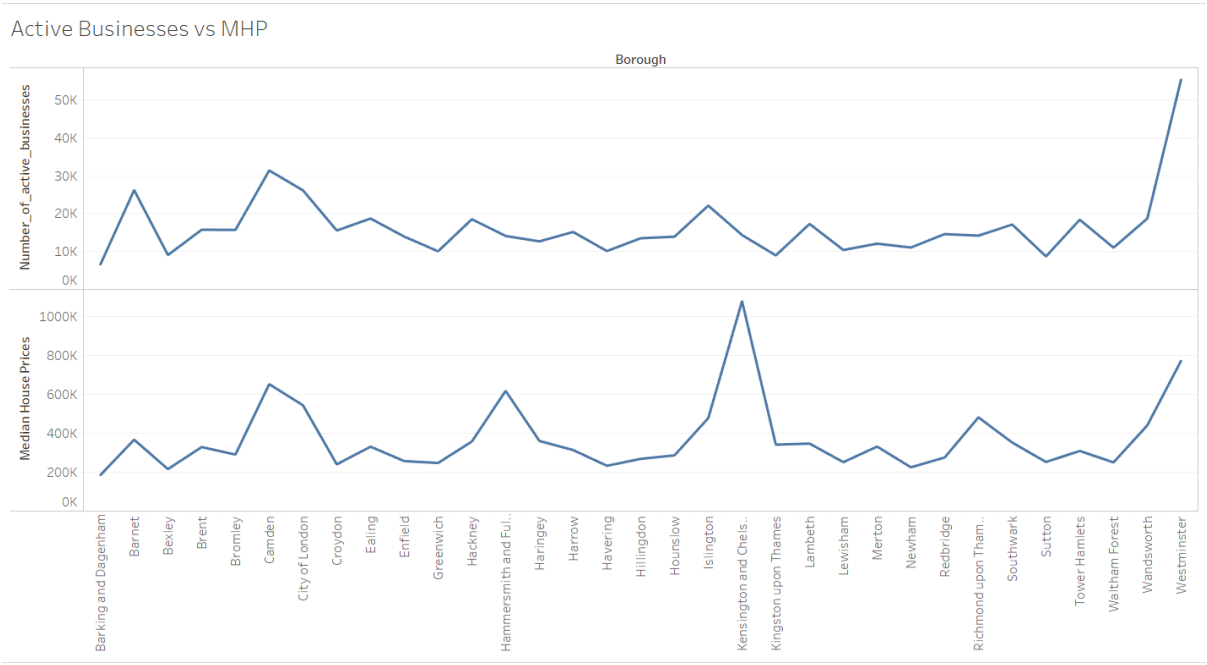

Fig.14. Line graph comparing number of active businesses in each borough with MHP

Additional Exploratory Analysis:

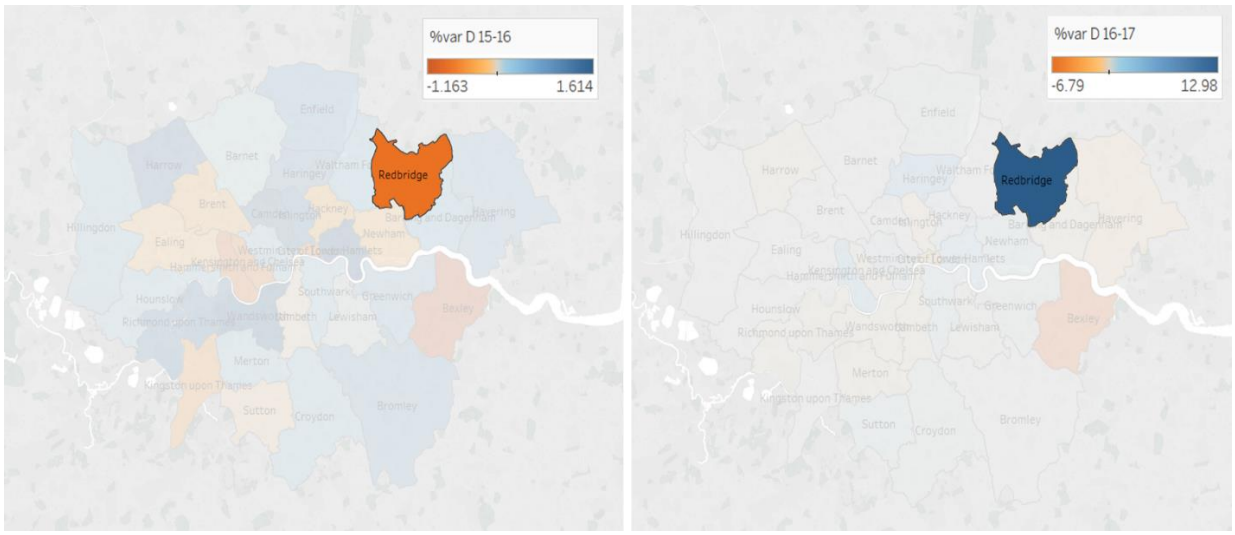

Fig.15. Redbridge has seen a significant different in variance of dwellings added between 2015-16 and 2016-17

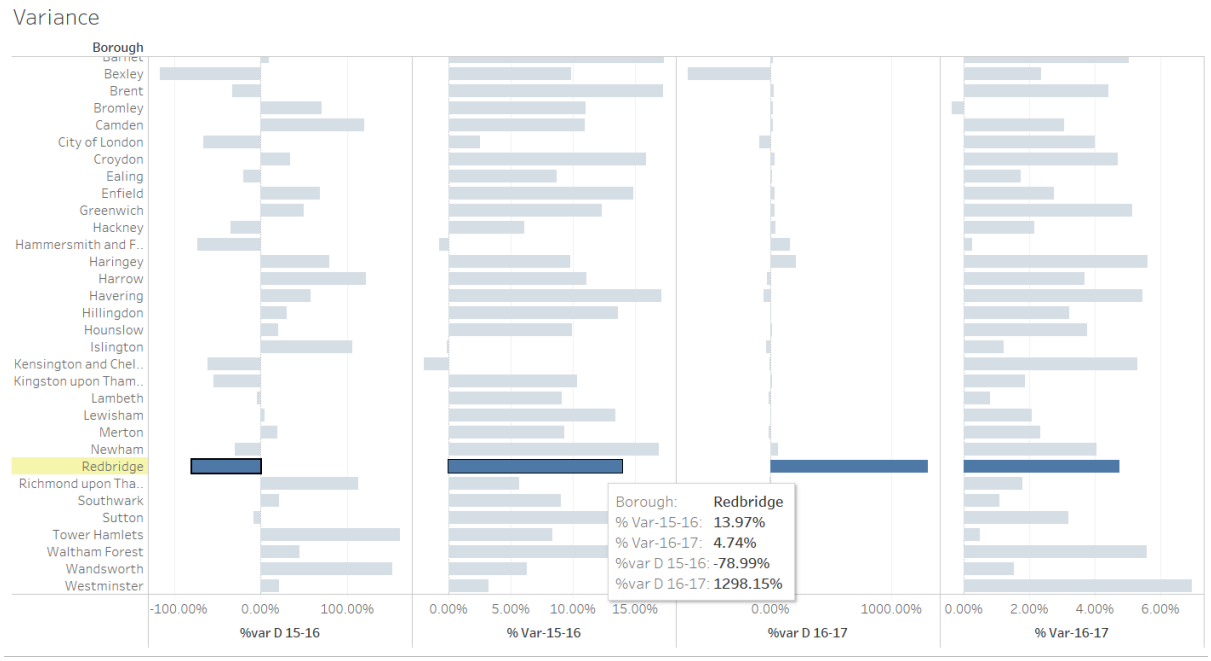

Fig.16. Although Redbridge followed the hypothesis in 2015-16, the large increase in dwellings in 2016-17 led to an increase in MHP, thereby rejecting the hypothesis.

Task 3:

```
Column " Median Employment Income " has a loading of:  0.23385992593960153
Column " Jobs_Density,_2015 " has a loading of:  0.23370503614515803
Column " Turnout_at_2014_local_elections " has a loading of:  -0.21615099982094416
Column " GLA_Population_Estimate_2017 " has a loading of:  -0.20449390614138616
Column " Fires_per_thousand_population_(2014) " has a loading of:  0.19993528972339855
Column " Anxiety_score_2011-14_(out_of_10) " has a loading of:  0.19811683137358657
Column " %_working-age_with_a_disability_(2015) " has a loading of:  -0.19296810214019858
Column " GLA_Household_Estimate_2017 " has a loading of:  -0.1921946826525807
Column " Mean Employment income for the borough " has a loading of:  0.1917617991055805
Column " Average_Band_D_Council_Tax_charge_(£),_2015/16 " has a loading of:  -0.1808001298029517
```

Fig.17. Top 10 features of principal component 1

```
Column " Net_international_migration_(2015) " has a loading of:  0.23535787842898756
Column " Number_of_cars_per_household,_(2011_Census) " has a loading of:  -0.22724350289240725
Column " Overseas_nationals_entering_the_UK_(NINo),_(2015/16) " has a loading of:  0.22261573291900358
Column " Homes_Owned_outright,_(2014)_% " has a loading of:  -0.21715517903319867
Column " %_people_aged_3+_whose_main_language_is_not_English_(2011_Census) " has a loading of:  0.21207808839175005
Column " Rented_from_Local_Authority_or_Housing_Association,_(2014)_% " has a loading of:  0.20416118385335388
Column " %_children_living_in_out-of-work_households_(2015) " has a loading of:  0.1800618025731458
Column " Worthwhileness_score_2011-14_(out_of_10) " has a loading of:  -0.17900365429569282
Column " Inland_Area_(Hectares) " has a loading of:  -0.17778946978207255
Column " Net_internal_migration_(2015) " has a loading of:  -0.17651780525752633
```

Fig.18. Top 10 features of principal component 2