ANALYSIS OF THE DEVELOPMENT AND GENTRIFICATION OF DISTRICTS OF GREATER LONDON

Tulsi Mehta

Abstract.

Keywords: London, Clustering, Foursquare, Development, Gentrification

1. Introduction

1.1. **Background.**

London is the capital and the largest city of England and is also considered one of the worlds most important global cities. Amassing approximately 9 million people (according to the 2011 census) across the City of London and the 32 other boroughs, London has a high population density which raises interest of the spatial projection of the population and trends of housing within different areas.

1.2. **Problem.**

The establishment of the 33 boroughs of Greater London is relatively modern in its history and span relatively large areas thus having a weaker sense of identity compared to their constituent districts. More familiarly known, there are 630 of these districts with each one pertaining to distinct quirks and characteristics of its residents and local businesses. The names of these areas are majoratively shared with the names of the electoral wards. These differences can exhibit themselves through various metrics- namely, the average housing prices and range of venues and businesses in the area. With the ever-increasing interest for property in the areas of Greater London, identifying the areas that are in infancy of development is key for those looking to invest in property and businesses looking to open in up-and-coming areas before the inevitable hike in prices.

Closely linked to the concept of regeneration of districts within London is the issue of gentrification and displacement of communities within these areas. Gentrification, as a term coined to describe the dramatic social changes, started to occur in London in the 1960s This can be traced by studying the changes in demographic, social funding, crime, education statistics, poverty and housing prices in areas known to have already undergone the process. Archetypes of this include Hackney in the late 2000s, Barnsbury in the 1970s and Brixton in the 2000s, to name a few.

1.3. **Interest.**

This report looks to analyse the districts of London in order to categorise them by their stage of development and synonymous gentrification in the interest of residents, local business owners and even local authorities. All parties outlined have a vested interest in the regeneration and hence property prices in their local area: residents looking to get on the property ladder, businesses looking to open, relocate or expand and local authorities needing to plan public investment.

2. Data

2.1. **Data Sources.**

The data acquired for this exploration comes primarily from Foursquare location data of venues within London districts. This initial list itself will be obtained by scraping a list of London areas (https://en.wikipedia.org/wiki/List_of_areas_of_London). This will enable the categorisation of each of the districts into their stages of development.
This data set outlined:

- **Location** - the name of the area/district/ward
- **London Borough** - its parent borough
- **Post Town** - its postal town
- **Postcode District** - the area postcode
- **Dial Code** - the area telephone dial code
- **OS Grid Reference** - grid reference identificiation

Secondary to this, records of average housing prices in each London ward between 1995 - 2017 from the London datastore (https://data.london.gov.uk/). This will be used to trace development and make predictions as to the future of these areas.

- **New Code** - area identification code
- **Ward Name**
- **Borough Name**
- Quarterly average housing prices from Dec 1995 - Dec 2017

2.2. **Data Cleaning.**

After scraping the data from the webpage this was the resulting dataframe:

	Location	London borough	Post town	Postcode district	Dial code	OS grid ref
0	Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	020	TQ205805
2	Addington	Croydon[8]	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon[8]	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728

Figure 1. List of London districts as scraped from webpage.

For purposes of only retaining the most relevant features from the datasets, the 'Post Town', 'Dial Code' and 'OS Grid Reference' columns were dropped as they would play no part in the retrieval of Foursquare location data. Also, inspection of the dataframe in Fig 1 shows clearly that reformatting of various columns was necessary as, for example, the London borough column shows residual information from internal references from the web page it was scraped from. Furthermore, in the case of the Location, borough and Postcode district columns, single data entries are filled with multiple values. This would not be conducive for the acquisition of longitude and latitude coordinates and therefore the columns must be cleaned. The method conducted in this instance saw the first listed value in the case of the Borough and Postcode being retained by dropping the surplus values. For the Location (renamed to District) column, the multiple values were separated and passed through the 'Ward' column of the second dataset. The values found to exist in the second dataset were retained in order to maximise the recovered data entries in further analysis. The final dataframe had 533 entries corresponding to each of the 'districts' of London explain in Sect 1.2.

The second data source was loaded and its resultant dataframe can be seen in Fig 2. Similar cleaning was done on this dataset dropping the first column and row and the resultant dataframe had 630 entries for each of the 630 London wards.

	New code	Ward name	Borough name	Year ending Dec 1995	Year ending Mar 1996	Year ending Jun 1996	Year ending Sep 1996	Year ending Dec 1996	Year ending Mar 1997	Year ending Jun 1997	Year ending Sep 1997	Year ending Dec 1997	Year ending Mar 1998	Year ending Jun 1998	Year ending Sep 1998	Year ending Dec 1998	Y end
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
1	E09000001	City of London	City of London	-	-	-	-	-	-	-	-	-	-	-	-	-	
2	E05000026	Abbey	Barking and Dagenham	51077.6	49868.8	49901.6	51935.1	50766.5	50189.5	47807.5	48004.6	51318.4	53046	56650.8	62273.5	62804.7	6342
3	E05000027	Alibon	Barking and Dagenham	45490.4	44701.5	44486	45894.1	46145.1	47019.2	48608.8	50142.5	51232.4	51968.8	52586.1	54027.2	56402.5	56
4	E05000028	Becontree	Barking and Dagenham	48947.6	49655.6	50129.2	51113.3	52333.8	54958.4	56108.9	56414.5	56734.9	55058.4	55249.1	56023.7	59766.9	62

Figure 2. Dataframe showing average house prices by ward from Dec 1995 to Dec 2017 (before cleaning)

The final step in preparing the data was to merge the two dataframes in Figs 1 and 2. The details of this will be discussed in Sect 3.1.

3. Methodology

3.1. **Data Retrieval.**

Using the GeoPy package, the longitude and latitude coordinates for the London districts in first dataset were obtained. In the case that the coordinates were not found, those areas were dropped from the dataframe which left 500 retained districts. The distribution of these areas can be seen in the map in Fig 3. Upon this visualisation, it was found that some of the coordinates were incorrectly identified far from their true positions (e.g. in America).

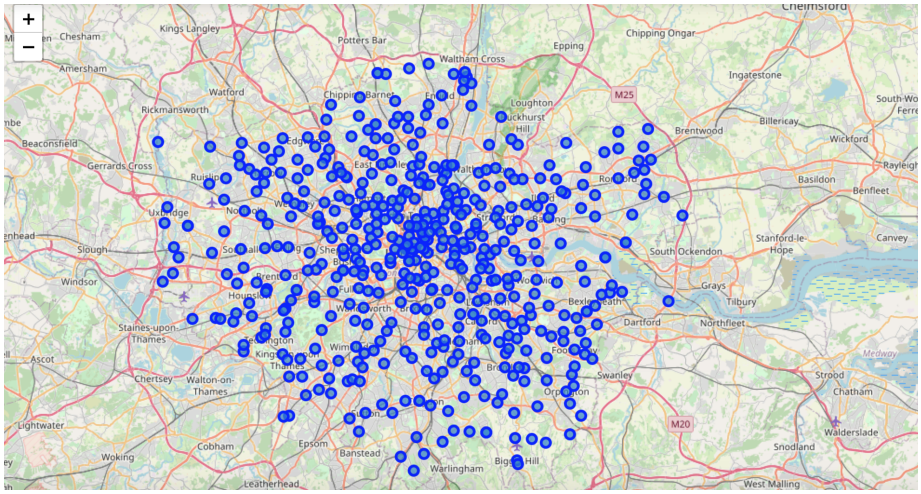

Figure 3. Visualisation of the distribution of 500 London districts

The two datasets were joined on the district/ward columns for each of the datasets and the resulting dataframe (seen in Fig 4) shows the successful matching of 194 London districts to wards. During the process of this dataframe merging, the data points that had erroneous coordinate identification were automatically dropped. The distribution of these retained districts are seen visualised in Fig 5. Comparing Figs 3 and 5 it is clear that the spatial distribution of the merged dataframe is far less dense (particularly in the central and western regions) however, it was decided that it comprehensively covers the footprint of London such that no further action was necessitated.

	District	Borough_x	Postcode	Latitude	Longitude	Borough_y	Year ending Dec 1995	Year ending Mar 1996	Year ending Jun 1996	Year ending Sep 1996	Year ending Dec 1996	Year ending Mar 1997	Year ending Jun 1997	Year ending Sep 1997	Year ending Dec 1997	Year ending Mar 1998
0	Abbey Wood	Bexley	SE2	51.495847	0.127708	Greenwich	46278.4	47470.2	46562.3	46827.1	49205.7	48997.7	49966.1	52283.7	54557.4	56360.2
1	Addiscombe	Croydon	CR0	51.367674	-0.095250	Croydon	58136.2	56418	56878.7	57114.7	56781.4	58149.8	60069.8	62740.2	64934.4	66555
2	Alperton	Brent	HA0	51.541077	-0.300885	Brent	57641.2	58597.4	58688.9	57154.9	57972.4	57824.6	60968.4	63026.6	67009.3	70068
3	Balham	Wandsworth	SW12	51.443352	-0.152869	Wandsworth	121877	120807	121691	135179	140407	144394	151338	153078	161158	171233
4	Barkingside	Redbridge	IG6	51.588247	0.080450	Redbridge	82587.8	84411.9	83820.8	87662	90260.4	89812.8	91806.3	91810.9	92705.8	96309.5
5	Barnehurst	Bexley	DA7	51.459186	0.148975	Bexley	67701.7	67574.3	68017.9	67797.9	68547.5	73211.9	74914.9	77721.2	78249.5	77848.8
6	Barnes	Richmond upon Thames	SW13	51.471896	-0.238744	Richmond upon Thames	229312	222985	228838	253730	245753	260901	295601	299955	308436	316411
8	Barnsbury	Islington	N1	51.538935	-0.114735	Islington	139615	127995	134750	143699	150238	162659	168360	173945	183266	175541
9	Bayswater	Westminster	W2	51.512273	-0.188244	Westminster	153905	163740	164753	170522	179176	179586	183065	195330	188953	193061

Figure 4. DataFrame showing the amalgamation of the location data and house price data for the 194 retained London districts.

Using the Foursquare API, data for 100 of the top venues within a 1000m radius of each district was collected. This returned a total of 298 unique venue categories. Subsequent clustering analysis of these districts using this data would prove cumbersome and inaccurate

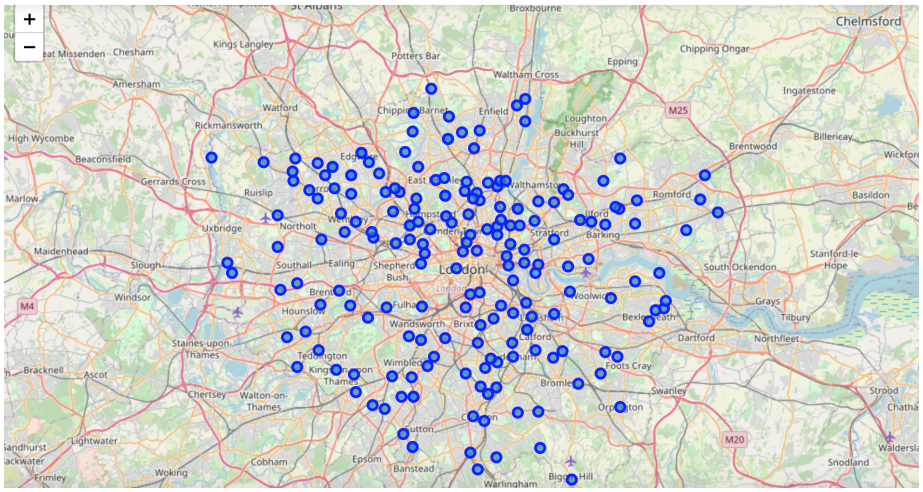

Figure 5. Visualation of the distribution of the 194 retained districts across the map of Greater London.

with this many variables upon which to account in the clustering as for example, this low level of venue category identification identified the difference between categories depicting Indian restaurants and Italian restaurants. Going forward, only the highest level of venue categorisation, which identified 10 distinct venue types (see Table 1), were to be considered.

Venue Category	Category ID
Arts & Entertainment	4d4b7104d754a06370d81259
College & University	4d4b7105d754a06372d81259
Event	4d4b7105d754a06373d81259
Food	4d4b7105d754a06374d81259
Nightlife Spot	4d4b7105d754a06376d81259
Outdoors & Recreation	4d4b7105d754a06377d81259
Professional & Other Places	4d4b7105d754a06375d81259
Residence	4e67e38e036454776db1fb3a
Shop & Service	4d4b7105d754a06378d81259
Travel & Transport	4d4b7105d754a06379d81259

Table 1. High level venue categories.

Foursquare API was used to obtain, for each district, how many venues from each category can be found within a 1000m radius of the area. These counts were then normalised using the min-max regime. The results of this can be seen in Fig 6.

	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	0.000	0.012048	0.0	0.013699	0.036585	0.011494	0.010526	0.04	0.030675	0.030769
1	0.040	0.072289	1.0	0.237443	0.152439	0.068966	0.400000	0.16	0.319018	0.207692
2	0.016	0.060241	0.0	0.045662	0.042683	0.034483	0.073684	0.12	0.098160	0.023077
3	0.096	0.072289	0.0	0.191781	0.115854	0.126437	0.231579	0.24	0.251534	0.069231
4	0.000	0.036145	0.0	0.041096	0.012195	0.034483	0.052632	0.12	0.073620	0.053846

Figure 6. DataFrame showing the scaled proportions of venues in districts of London from the top 10 venue categories.

3.2. **Exploratory Data Analysis.**

3.3. **Distribution of Venue Types.**

Preliminary data analysis consisted of visualisation of the distribution of venue categories across the whole dataset (thus across all of London. This box plot can be seen in Fig 7. From this graph we deduce that the most frequently occurring venue is that of Professional type as the median value for this category is the highest. Conversely the least frequently occurring venue type is that of the branch 'Event'. Since there is very few of this venue type it was dropped as its presence would likely skew later analysis. It is also noted that the Food, Professional, Residence and Shops Service categories have the largest spread of occurrences and are the most common venue type. There are many outliers towards the higher end of the distribution for every category which shows that there are districts which are very different to the majority.

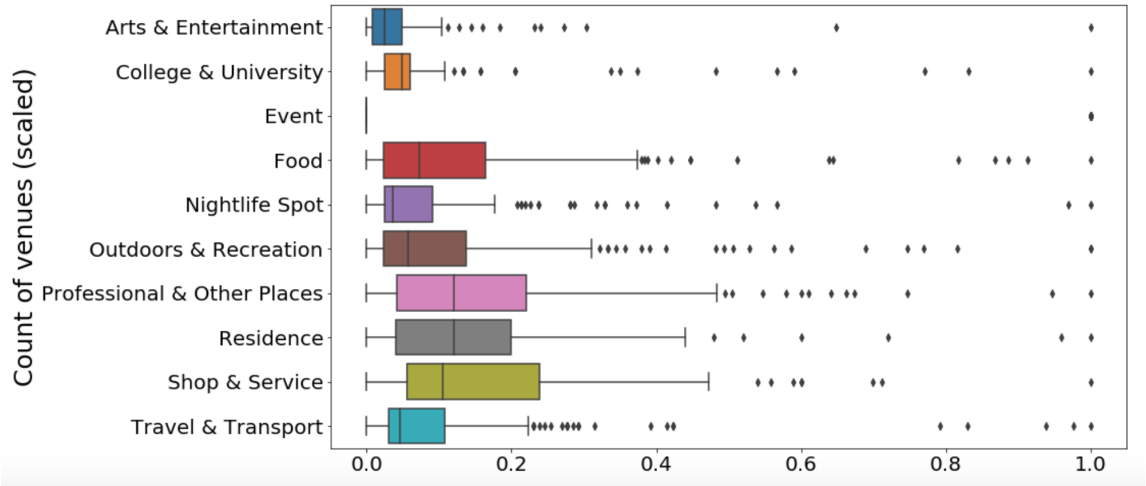

Figure 7. Boxplots showing the distribution of venue categories across all districts in London.

3.4. **Clustering of Districts.**

In order to attempt to classify the districts to distinguish between those pre- and post-

gentrification, the unsupervised machine learning algorithm called K-means clustering was employed to cluster the districts based upon their constituent venues. This partition based clustering algorithm divides the data into k number of non-overlapping spherical clusters.

The algorithm was run with k equal to 2, 3, 4 and 5 and the results compared. It was deduced that partitioning the data into just 2 classes was too crude and only identified a distinction between the innermost central London districts and the rest. In the case of using k=3 the clustering outcome was slightly better with the more peripheral suburban districts distinguished from the other Greater London districts. On the other hand using partitioning the data into 5 classes meant that some of the classes were too similar and it was unclear what variable the subdivision of these more similar classes was. Setting k=4 was found to produce optimal clustering results as will be discussed in Sect 4.

4. Results & Discussion

4.1. **K-means Clustering.**

A visualisation of the results of running the clustering algorithm with k=4 can be seen in Fig 8. Areas belonging to cluster 0 are depicted by red dots, those belonging to cluster 1 are purple, cluster 2 districts can be seen in blue and lastly cluster 3 areas are shown by green markers. An immediate observation would be that there seems to be a radial relationship between cluster type and distance from the centre of London with cluster 3 being the most central districts and cluster 0 being the furthest districts.

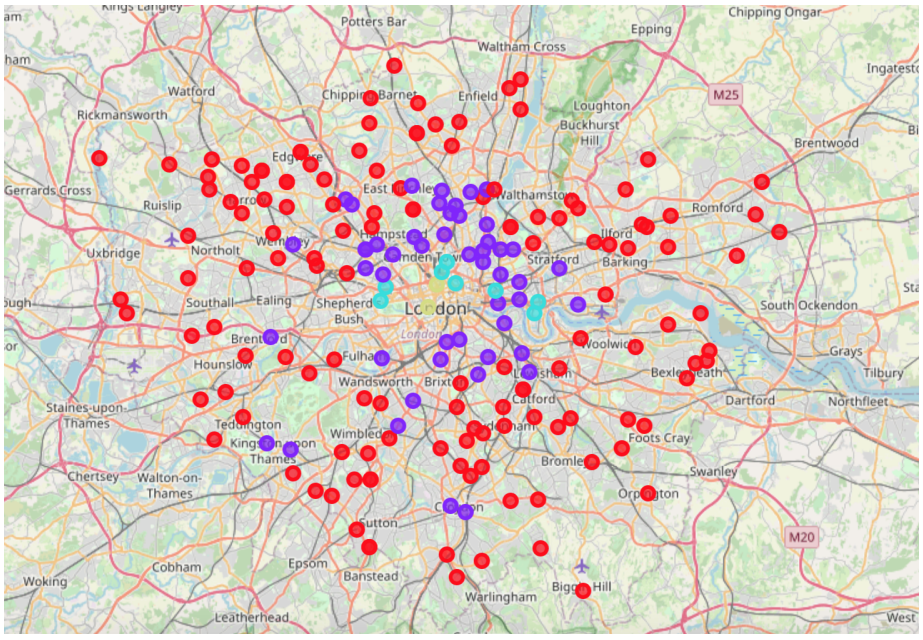

Figure 8. Map showing the London districts in their parent clusters. Cluster 0 is red, cluster 1 is purple, cluster 2is blue and cluster 3 is green.

A more detailed analysis of what exactly these clusters labels mean is exhibited through the box plots in Fig 9. It is clear that cluster 3, with its high proportions across all categories compared to the other clusters, represents the most developed of all districts in London. It is therefore no surprise that these are the districts found in the very heart of the city which will be the most built up and therefore busiest with the most venues.

Following this, cluster 2 is most distinctly identified by its relative higher number of residence type venues compared to that of cluster 3. Districts belonging to cluster 2 have a high number of venues in all categories with the highest belonging to Food, Outdoor & Recreation and Professional venues. Cluster 2 therefore is also identified as very developed with the difference that it has more residential venues.

Cluster 1 is seen to spread further away from the centre of London than the aforementioned clusters, especially in southern regions. Districts that fall in this cluster have more venues that are Professional and Shop & Service. Second to this they have a lot of venues in the the Food, Recreation and Residence categories. The box plot also shows that in Cluster 1 areas, you are less likely to find Arts & Entertainment, University and Nightlife venues. This suggests that these areas are less focused towards students in times were there is growing investment in word-wide universities and colleges in the capital city. Cluster 1 districts in this scheme are labelled as developing and perhaps are currently undergoing the process of gentrification. Interestingly, there appears to be a hub of cluster 1 areas in the northeast London in districts like Bethnal Green, Stoke Newington, Hakney and Dalston. These areas have been highlighted as areas that were the target of 'redevelopment' and in the process of gentrification in in recent history. This validates the accuracy of the model this analysis promotes.

Lastly, cluster 0 is seen to have the lowest number of venues across all categories compared to the other cluster groups. In this scheme, these regions are labels under-developed and in a 'pre-gentrification' state. The most common venue in this cluster group is residential which aligns with this definition. Intuitively, this matches with the expectation that the further from the city centre, the less developed an area would be. However, this cluster group, encompassing the majority of the suburban districts, is rather rudimentary as is shown by the number of outliers for the box plots in cluster 0 in every category. It is expected that there are sub-categories within cluster 0 which would better describe the development stage of these districts and another factor other than the venues in an area must be considered to appropriately label these areas.

4.2. **Average Housing Price.**

In order to attempt to investigate cluster 0 further, the data for housing price was utilised. The average house price from December 1995 to December 2017 was plotted for all cluster 0 districts and the results can be seen in Fig 10. Since cluster 0 was the largest of all the clusters with 135 member districts out of the original 194, the plot bares a lot of information.

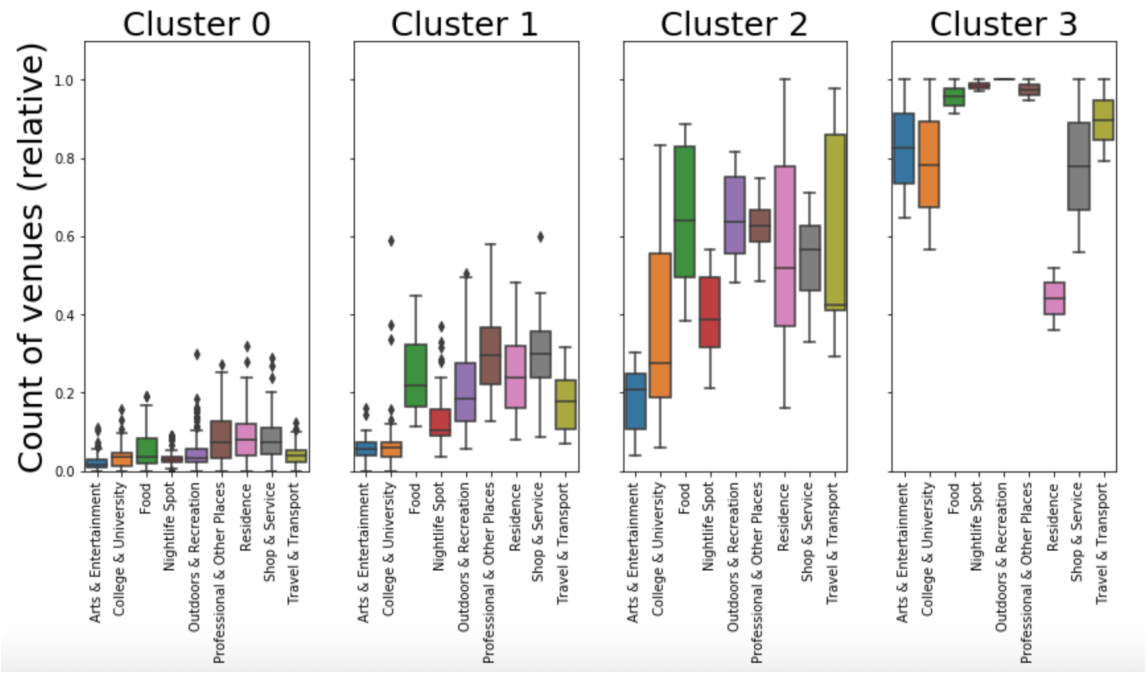

Figure 9. Box plots showing the breakdown of the population of venue categories within K-means clustering groups.

Through the density of lines on the plot it is clear that some lines reside consistently above the rest across the entire observational time span. Namely, this includes Totteridge, Kew and Highgate. These areas are notoriously affluent but characteristically very residential. This raises the concern that the prior analysis is blind to those London districts that have few venues by preference of retaining very residential communities rather than being due to causes of under-development.

To further divide this cluster, an upper limit of £500,000 in December 2017 was used to restrict how much the average price of a house can be in a district for the district to be classified as under-developed. The average house price of the resulting 78 districts are seen plotted in Fig 11. From inspection there is a clear trend in house prices increasing over the 12 years of observation with the view to this continuing. The names of the 78 from the original 194 London districts that are deemed as 'underdeveloped' by means of this analysis can be found in Table 2.

5. Conclusion

To conclude, a K-means clustering algorithm was used with Foursquare location data of London districts in order to divide the districts into 4 clusters: underdeveloped, developing, developed (residential) and developed (non-residential). This identified 135 of 194 districts

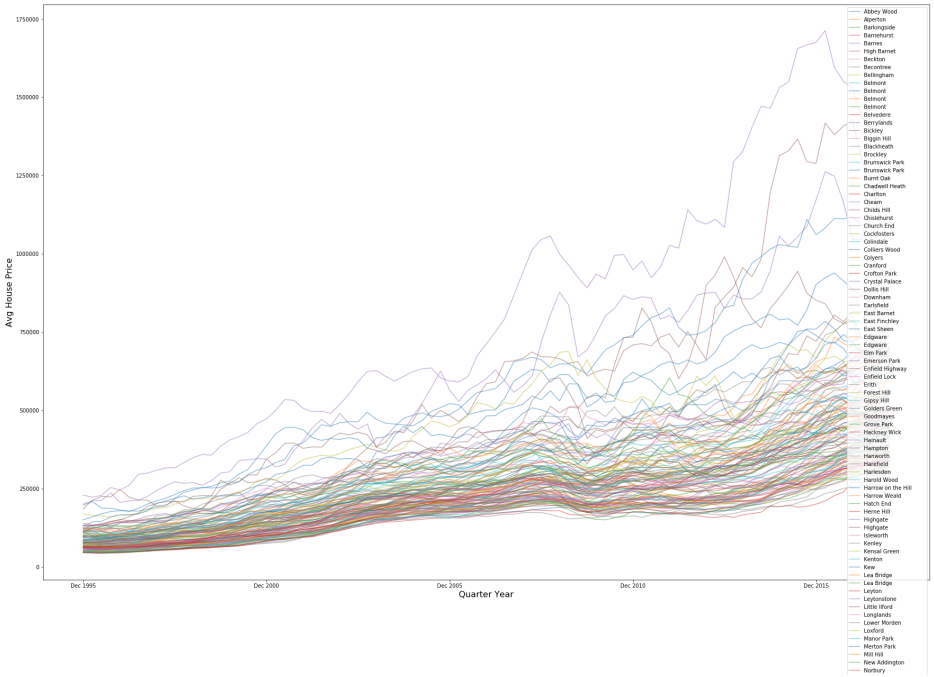

Figure 10. Average house price by London district per yearly quarter from Dec 1995 to Dec 2017

as underdeveloped. Within this category of underdeveloped it was expected that some of these districts were more developed than others. Data of average housing prices in these districts was used to further divide this category by placing a limit of £500,000 in December 2017 in order for the district to be classified as underdeveloped. This would suggest that these areas should be a priority for local councils in terms of local investment in these communities. Also, these areas bare the lowest housing prices and are of potential interest for local businesses and people looking to invest in areas that have a future of a development boom.

Since it was seen during this analysis that the inclusion of housing prices on top of the location data identified areas where the analysis failed to pick up on important nuances in area classification, further development on this analysis could see to consideration of a number of other data sources like demographic, crime rates, education, poverty and social funding of London districts in order to refine the classification of these areas by evaluating as many dependant factors as possible.

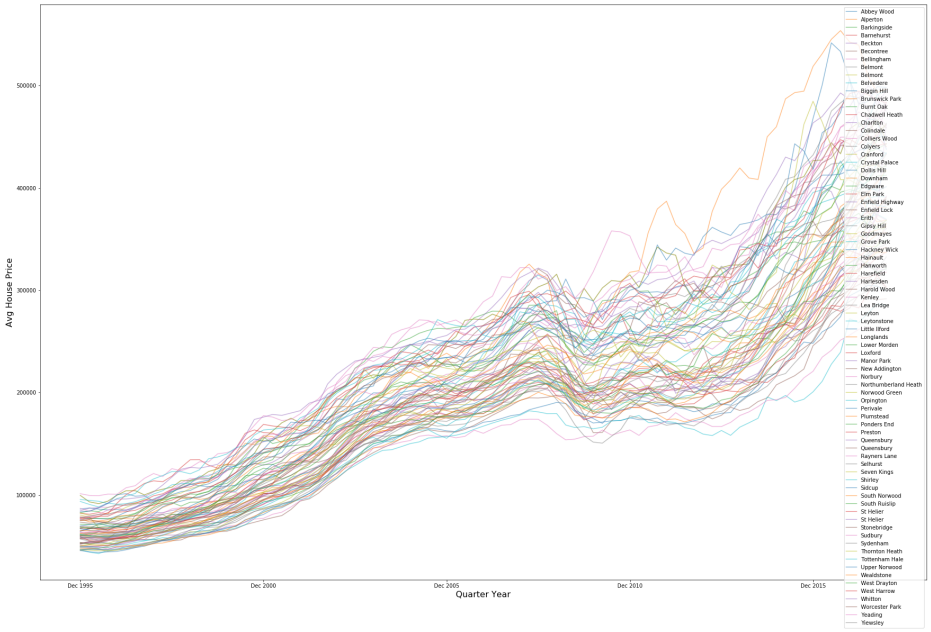

Figure 11. Average house price per yearly quarter from Dec 1995 to Dec 2017 for all districts with an average house price of less than £500,000 in Dec 2017.

Abbey Wood	Alperton	Barkingside	Barnehurst	Beckton
Becontree	Bellingham	Belmont	Belmont	Belvedere
Biggin Hill	Brunswick Park	Burnt Oak	Chadwell Heath	Charlton
Colindale	Colliers Wood	Colyers	Cranford	Crystal Palace
Dollis Hill	Downham	Edgware	Elm Park	Enfield Highway
Enfield Lock	Erith	Gipsy Hill	Goodmayes	Grove Park
Hackney Wick	Hainault	Hanworth	Harefield	Harlesden
Harold Wood	Kenley	Lea Bridge	Leyton	Leytonstone
Little Ilford	Longlands	Lower Morden	Loxford	Manor Park
New Addington	Norbury	Northumberland Heath	Norwood Green	Orpington
Perivale	Plumstead	Ponders End	Preston	Queensbury
Queensbury	Rayners Lane	Selhurst	Seven Kings	Shirley
Sidcup	South Norwood	South Ruislip	St Helier	St Helier
Stonebridge	Sudbury	Sydenham	Thornton Heath	Tottenham Hale
Upper Norwood	Wealdstone	West Drayton	West Harrow	Whitton
Worcester Park	Yeading	Yiewsley		

Table 2. List of London districts identified as underdeveloped