Location Analysis for Positioning Medical Practice in Greater London Area

Asanka Pannipitiya – IBM Capstone Project – April 2020

1. Introduction

In the real estate market, whether it be for purchasing a home or investment property, it's all about location, location and location. There is no difference in the logic when it comes to starting a medical practice (or any other business venture for that matter), finding best suitable place would be the key for success. In this study, we will look at few characteristics of wards in Greater London area to determine the best wards for starting a new medical centre.

In order to identify potential locations for a medical centre, we will collect few important statistics on each ward in Greater London area. Two main parameters to look at would be concentration of the population that we are trying to target and competition, meaning what other medical practices already exist in the area. There could be few secondary factors, which might be of importance, such as rate of population growth, young (0 – 15 year) or elderly (65+ year) population percentage, median household income, etc. For instance, if we intent to start a practice with paediatric-related services, it would be wise to focus on communities where parents with young children are moving.

Findings from this study can be used for zeroing on neighbourhoods for the business purposes from the providers' perspective, however, there are could be other use-cases as well. For instance, as a patient or a member of general public, one could be interest in localities those having better medical coverage in terms of availability of number medical practices within reasonable distance. Especially, with the current pandemic situation of COVID-19, living in such a local area where you have sufficient services available for medical care would be something that one might be particularly interested in. It is my view that, findings from this study can be equally applicable in those scenarios and provide valuable insights to desired parties to make an educated decision.

2. Data Gathering and Manipulation

2.1 Data Description and Sources

For this study, we analyse 625 wards in Greater London area. One of the best portal for free and open data relating to the capital city is the London Datastore (https://data.london.gov.uk/). Site has over 700 datasets to assist various users for range of applications purposes.

As mentioned earlier, in this study, we will look at number of features for each ward in Greater London. We assume these features would assist in clustering the wards in such a way that it would ultimately help us in picking up few suitable suburbs, from the list of 600+ wards, for the purpose of initiating a medical practice.

Specific features that we use for this analysis is given in the below table with brief explanation for such consideration and the source of those data.

Feature	Justifications for Selection
Population Density	With the increased number of people in the neighbourhood, proportionately, there would be more people, who may require medical assistance. Further to that, with the increased population in an area, there is also an elevated likelihood in community transmitted diseases, which is very much proved to be valid with the current COVID-19 situation in the world.

	Source: Ward profiles and atlas in London Datastore provide a range of demographic and related data for each ward in Grater London (https://data.london.gov.uk/dataset/ward-profiles-and-atlas). We use *Population – 2015* and *Area – Square Kilometres* columns to extract the population density.
Population Growth Coefficient	Certain wards have higher population growth rate depending on the underlying situations in those areas. For instance, newly developed neighbourhood with significant proportion of young families would have higher growth rate compared to established neighbourhood. So, even though the current population density may not be higher, there is an increased tendency that the area would become dense in reasonable future, which might work preferably for the business purposes. **Source:** Land area and population density figures for wards and boroughs are available in London Datastore for 2001 - 2050 (https://data.london.gov.uk/dataset/land-area-and-population-density-ward-and-borough). Growth coefficient will be calculated from 2011-2020 population data, where 2011 figures are from census data. In here, we use *sklearn library – Linear Regression model* to calculate the growth coefficient from 2010-2020 population data.
% of 65+ years in the Population	It is part of life that when we grow old, we find problems in our health and need medical assistance more frequently than the young adults. So, it is safe to assume that if the locality has higher percentage of elderly people, then there could be more people, who would be seeking medical care in that neighbourhood. **Source:** Ward profiles and atlas in London Datastore provide a range of demographic and related data for each ward in Grater London (https://data.london.gov.uk/dataset/ward-profiles-and-atlas). We use the column named; *Older people aged 65+ - 2015* for our analysis.
% of 0-15 years in the Population	Similar to elderly population, young kids equally need medical care in their early life. As such, if a neighbourhood has comparably higher percentage of kids, it might be better suited for starting a medical practice. **Source:** Ward profiles and atlas in London Datastore provide a range of demographic and related data for each ward in Grater London (https://data.london.gov.uk/dataset/ward-profiles-and-atlas). We use the column named; *Children aged 0-15 - 2015* for our analysis.
Existing medical centre count	Competition is a critical factor to consider about when starting any business. So, if a given ward already has significant number of established medical centres compared to population density, it might not be a better locality to consider.

	Source: This is where we use Foursquare API to fetch details about medical practices. API support number of category identifiers (https://developer.foursquare.com/docs/build-with-foursquare/categories), which would assist in getting suitable venues for the query.

Further to our initial analysis of clustering the local wards by their feature statistics and filtering out handful of suburbs, which would be the ideal candidates for our medical centre, we then use Folium and Geopandas for visualizing these on Greater London area map. We will use both maps with markers and Choropleth thematic maps for this purpose. Geopandas allows to create standalone choropleth maps without many other dependencies. For the shape files that Geopandas required for mapping is readily available in London Datastore. We found number of source files with significant level of details in following location, https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london.

Figure 1: Greater London Map with Boroughs

2.2 Data Manipulation and Cleaning

Data downloaded from London datastore contains significant amount of information, but it does need cleaning and manipulation to mould it to the format that we wanted. In below diagram, we illustrate the process flow of data highlighting the main steps in our pre-processing stage. Output of this stage produces a final data frame which contains all the required columns for our clustering analysis.

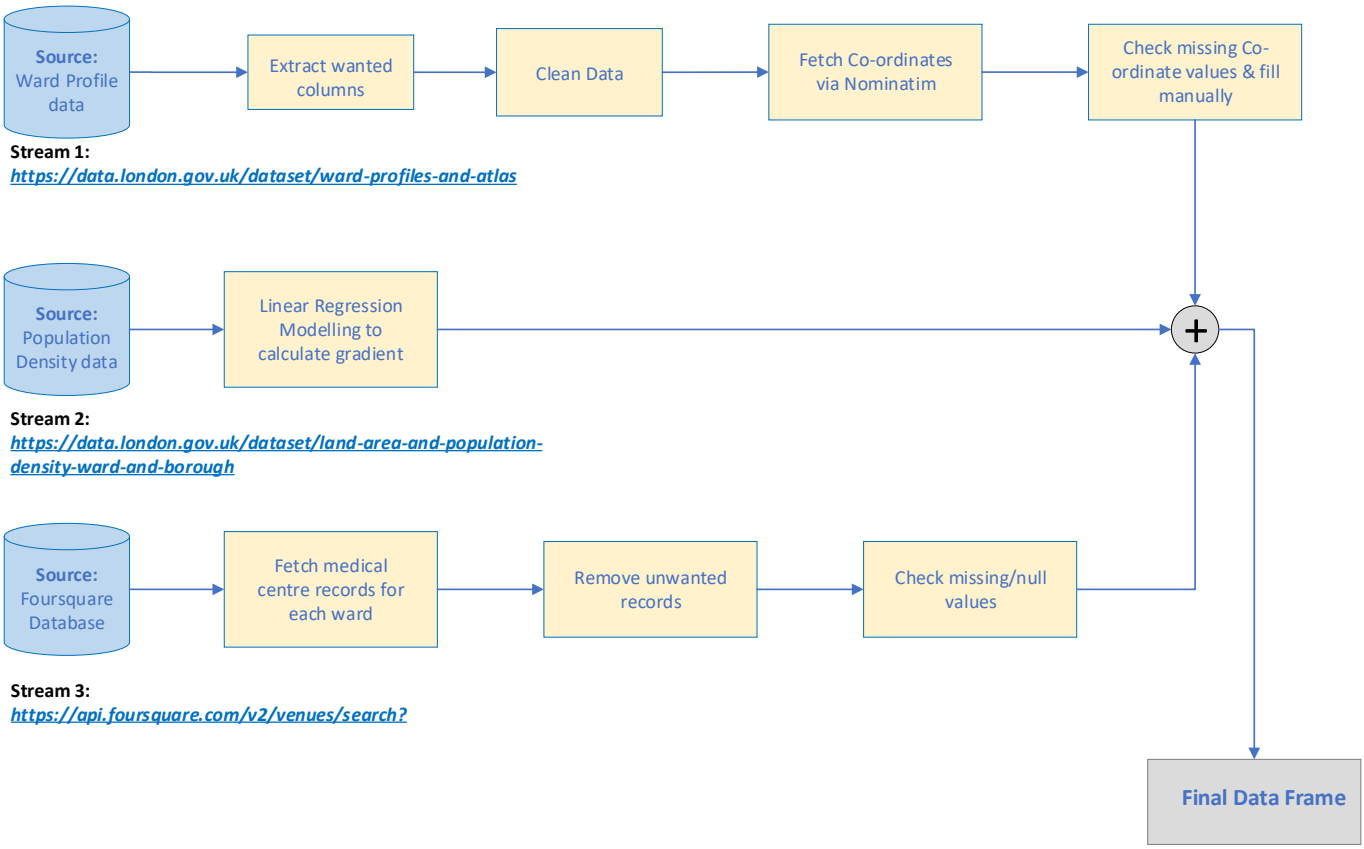

Figure 2: Data Flow Diagram

Let's discuss briefly about each stream depicted in above diagram.

Stream 1: Ward profile dataset in London datastore does provide considerable piece of information about each local ward in Greater London area. This dataset has 660 rows of entries with total of 67 columns, providing basic figures like population density, area to intricate figures like cars per household, crime rate, employment rates, etc. As mentioned in the earlier sections, we only interested on few features for our analysis. So, first, we extract those required columns from the imported dataset.

Looking at all the entries in this dataset, we can identify that records 625 to 659 (35 entries) correspond to details of the boroughs and not related to individual wards. Hence, we removed those rows from our dataset. Further to that, when we look at the first row of data, it contains details of City of London. However, City of London itself is a borough, which contains 25 smaller wards within it. Yet, our dataset contains details of the entire borough. Considering the unavailability of the individual ward data for City of London and, those wards being relatively small compared to rest of the wards, we consider this row of City of London as a single ward in Greater London.

Another thing to notice is the content of the Ward name column in the dataset (refer Figure 3). Two main things to notice here. First, there are two parts in this column, it follows the format: ***Borough – Ward name***. Second, some wards fall into two boroughs, so they are named in the format: ***Borough 1 and Borough 2 – Ward name***. This would create problems when fetching co-ordinates for each of these wards. So, we split this column into two separate columns at the hyphen, named, Borough and Ward.

Then, for the case of two boroughs in one ward, we split that again at the word *'and',* use the first borough name only for our records. This modification allows us to fetch geo co-ordinates through Python package, Nominatim.

Finally, we use Nominatim package in Python to get the co-ordinates for each local ward in our dataset. After this stage, we check the completeness of the data by checking the null values. In our test, we had 32 local wards with missing co-ordinates, this is primarily due to the naming issues such as ambiguity in the ward name. As this is just 5% of our dataset, we manually extract co-ordinates via Google maps and merged them with our original dataset.

Ward name	*Old code*	*New code*	Female life expectancy - 2009-13	% children in reception year who are obese - 2011/12 to 2013/14	% children in year 6 who are obese- 2011/12 to 2013/14	Rate of All Ambulance Incidents per 1,000 population - 2014	Rates of ambulance call outs for alcohol related illness - 2014	Number Killed or Seriously Injured on the roads - 2014
City of London	00AA	E09000001	88.6	11.1	23.2	140.0	19.34	57
Barking and Dagenham - Abbey	00ABFX	E05000026	83.9	13.3	24.7	157.3	1.3	2
Barking and Dagenham - Alibon	00ABFY	E05000027	80.6	10.0	26.0	139.8	0.9	1
Barking and Dagenham - Becontree	00ABFZ	E05000028	79.3	12.3	29.3	130.1	0.6	3
Barking and Dagenham - Chadwell Heath	00ABGA	E05000029	82.2	12.8	24.6	139.1	0.8	6
Barking and Dagenham - Eastbrook	00ABGB	E05000030	81.1	12.6	21.6	147.8	0.7	2
Barking and Dagenham - Eastbury	00ABGC	E05000031	84.9	11.5	26.5	113.0	0.4	5
Barking and Dagenham - Gascoigne	00ABGD	E05000032	81.1	16.2	29.0	136.1	0.8	2
Barking and Dagenham - Goresbrook	00ABGE	E05000033	82.8	16.2	25.2	136.6	0.7	1

Figure 3: Extract from Ward Profile dataset

Stream 2: Population density dataset contains GLA 2016-based population projections (housing-led model) from 2001 to 2050. This dataset has data in cleaned format, which does not require much pre-processing. Only additional processing that we require in this stream is to calculate the population growth rate from the 2001 – 2050 records. For that purpose, we use linear regression modelling in scikit-learn module in Python. For each ward in our stream 1 dataset, we get population records for 50-year span and fit a linear regression model on that. Gradient coefficient for each ward would give the growth factor for the ward.

Stream 3: In this stage, we fetch information about existing medical centres for each local ward in our stream 1 dataset via Foursquare API. This has pre-defined venue categories and respective identifiers, (refer https://developer.foursquare.com/docs/build-with-foursquare/categories/). What we interest here is medical centres, and they have a category identifier, *4bf58dd8d48988d104941735*. However, under medical centre category, there are multiple sub-categories, such as. emergency rooms, hospitals, medical labs, etc. We particularly be interested in medical centres and Doctor's offices. So, we will filter them out by looking at category field (plural name) for each medical record. For data completeness, we check the missing/null values in the dataset, but in our case, we were unable to find any null values for all 625 local neighbourhoods.

Final part of this stage is to combine all these data into one single Pandas data frame for further processing. So, we add the two columns we extracted at Stream 2 and 3 to our data frame created in Stream 1. In the next section, we do a preliminary data analysis on distribution and statistical figures to understand this dataset that we have created.

3. Exploratory Data Analysis

In this section, we look at descriptive statistics on the features and investigate relationship between existing medical centre count and each feature.

3.1 Population Density

Let's first see how the distribution of the population density in the local wards. Histogram for population density shows, just over 51% of the wards have population density between 3000 and 9000. Yet, there are quite a few local wards towards 9000+, and looking at the box plot for the whole Grater London area, there are few outlier suburbs as well.

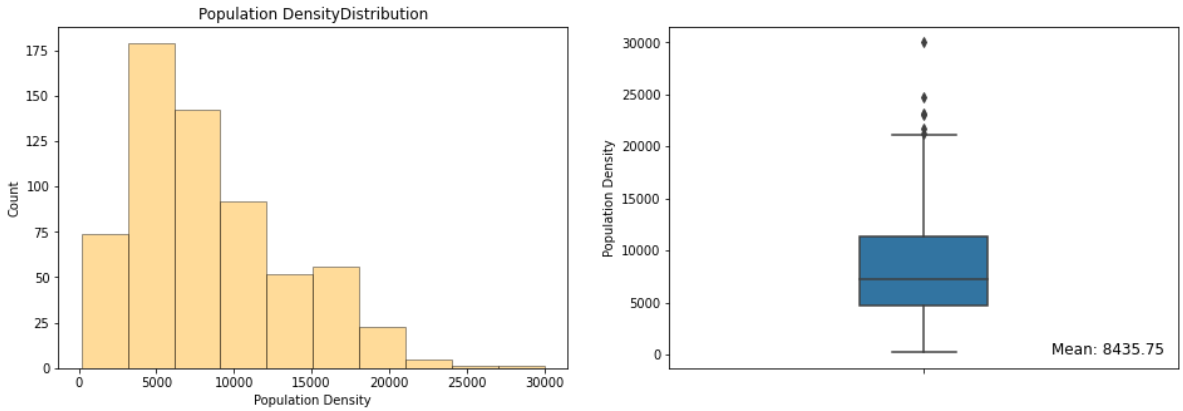

Figure 4: Population Density Distribution

To understand dense boroughs in the area, we use another box plot, but it will now show the population density variable *within* borough for local wards. As can be seen from the plot below, Westminster borough has one local ward, which has the highest population density in Greater London area, i.e., the one with 30,000 people. Across all the boroughs, most of them have at least one or many wards with lower population density (< 2000 peoples/ward).

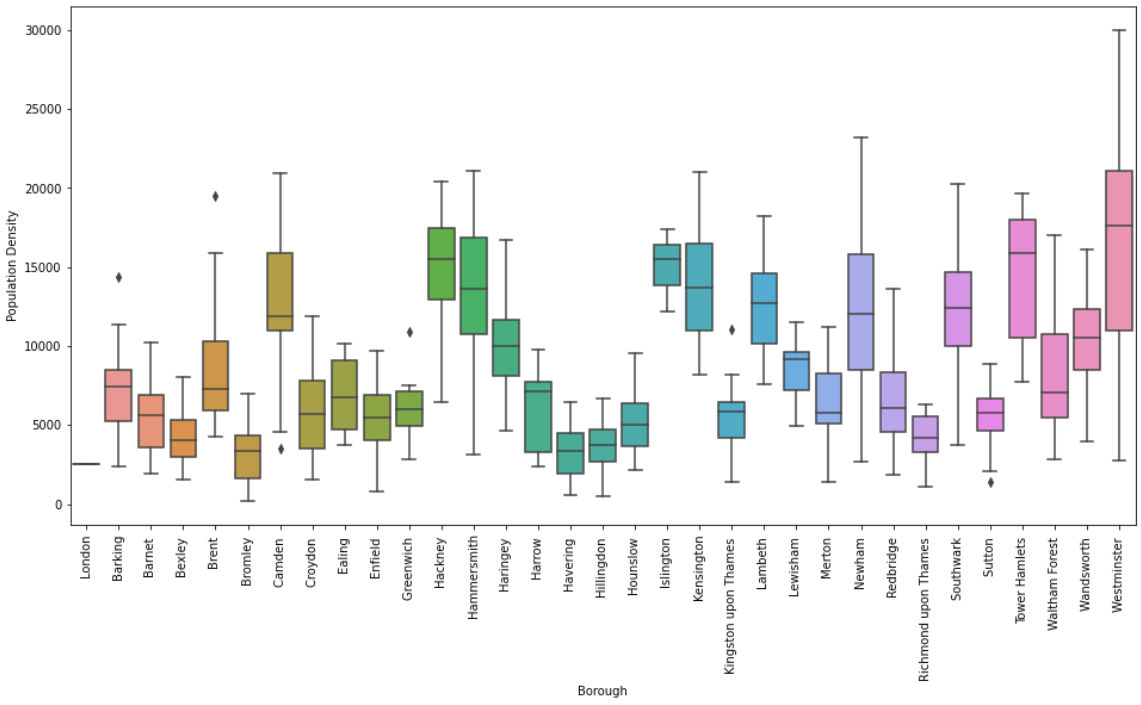

Figure 5: Population Density variation in Boroughs

3.2 Medical Centre Count

Looking at the numbers in our final data frame, we can see that our search has fetched total of 8,154 medical centres across 625 local wards in Greater London area. It is important to stress here that, for *neighbouring* wards, our search may have captured the same medical centre through Foursquares API. So, this total count across the entire set of wards does not necessarily mean, it is total *unique* number of medical centres.

There are some areas with high count of medical centres, but they don't have higher population density, and opposite is true for some localities too.

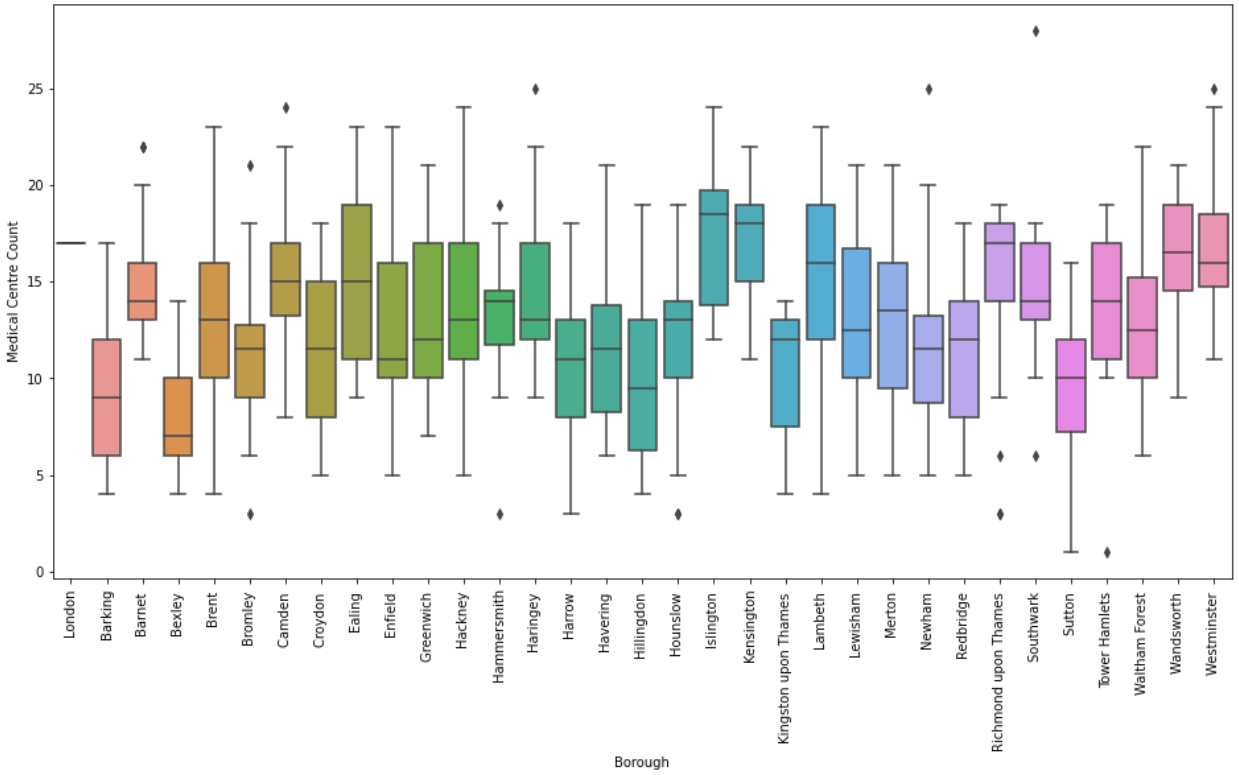

Figure 6: Medical Centre Variation in Boroughs

If we look at the histogram for the medical centre count, we can see that it is generally hover between 10 – 20 centres for a given local ward. There is only one outlier suburb, which has relatively high number of medical centres (28) within its neighbourhood.

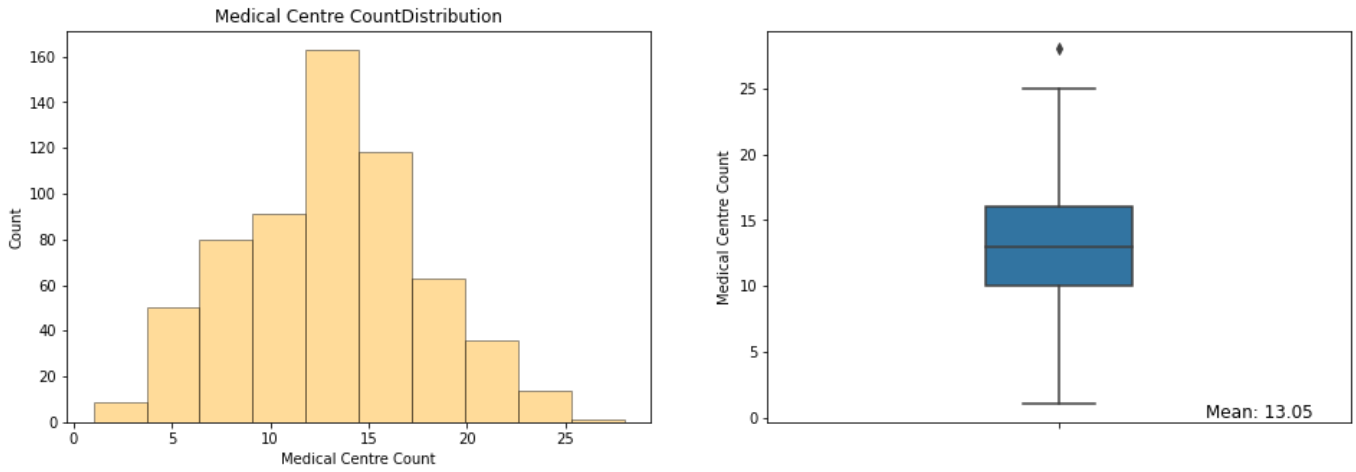

Figure 7: Medical Centre Count Distribution

3.3 Population Growth Coefficient

Population growth coefficient is calculated using the linear regression model on our population figures for 50-years, GLA 2016-based population projections (housing-led model) for all 625 wards. When we look at the distribution, most of the wards have population growth at the lower end of the spectrum. In fact, there are some suburbs with negative growth coefficient. It would be interesting to study the underlying characteristics of those suburbs compared to others to understand the behaviour better, but we will leave that to a separate discussion.

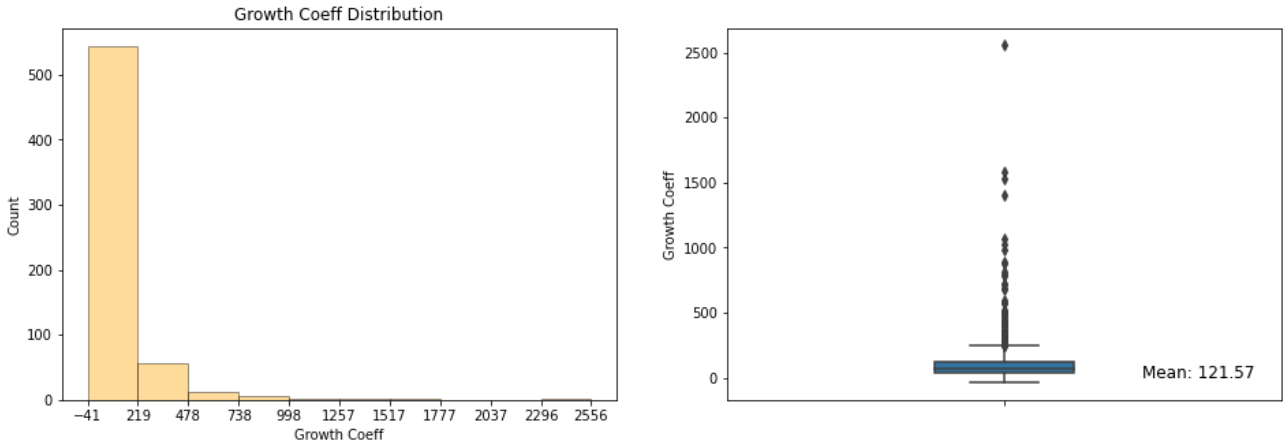

Figure 8: Population Growth Coefficient Distribution

3.4 Percentage of all children (0-15 years) and older people (65+ years)

We briefly discuss on these two features, children and older people distribution among the local wards. As can be seen from the figures, most of the wards have children percentage range from 17% to 22% and older people percentage between 5.75% - 14%.

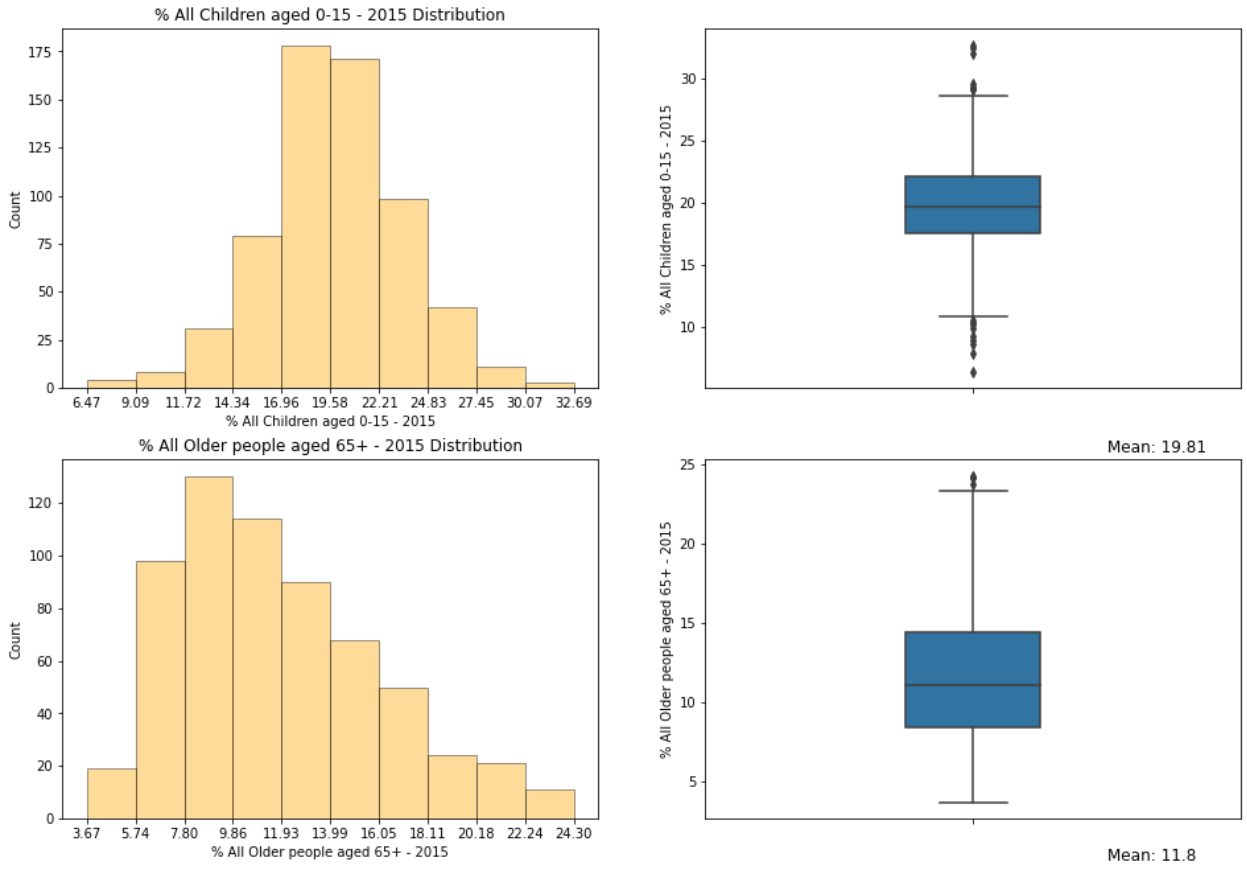

Figure 9: Children and Adult Percentage Distribution

As anticipated, in general, there is a slight correlation between population density and medical centre count. This is not the case with other features and medical centre count. See the below regression plots (Figure 10) on this. We will see further on Growth coefficient, but it is clear from the below regression plot that most of the growth factors are between 0 and 100 people/year. There is not much correlation with this feature and medical centre count. It is interesting to note that our hypothesis of having higher percentage of older population or kids in the local ward would be favourable for medical centre placement. Looking at the regression plots, this does not seem to be the case.

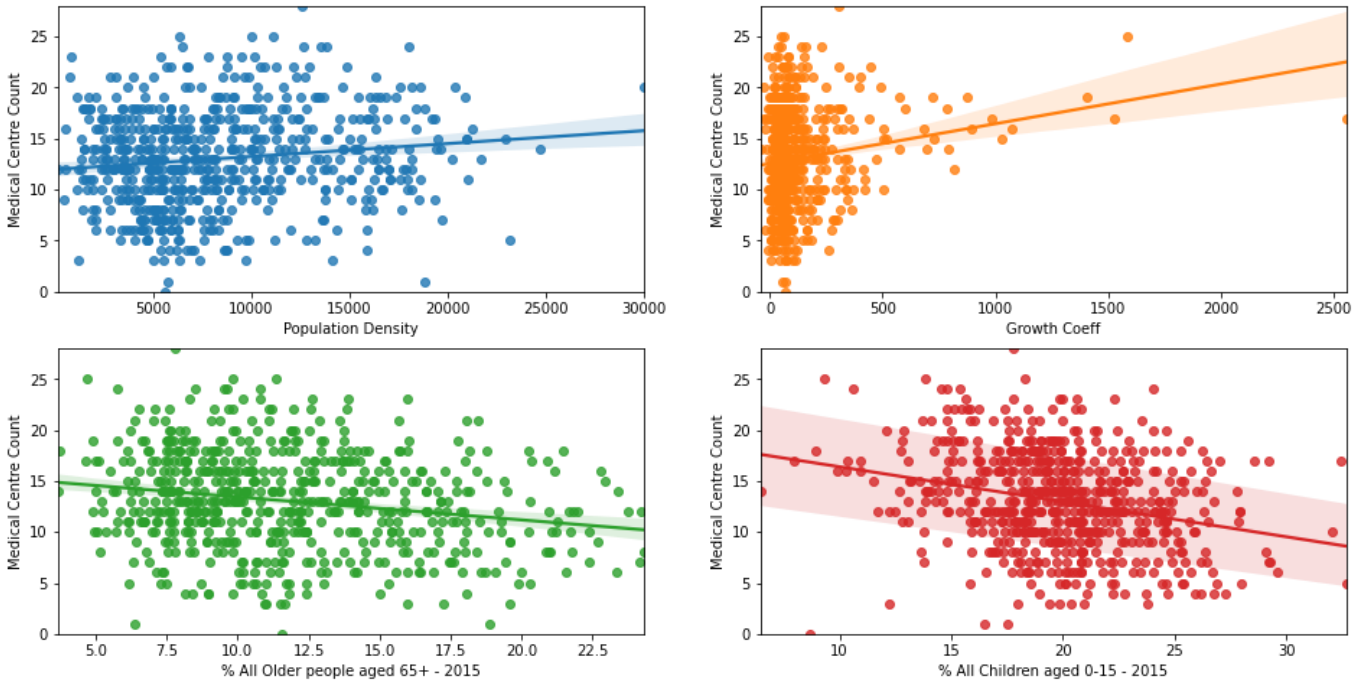

Figure 10: Regression Plots for Medical Centre Count

4. Machine Learning Approach

In this section, we perform clustering of the local wards based on the features that we discussed above. It is important to note that, we do not include the feature, medical centre count for this clustering. We first look at the preferred suburbs based on the other 4 features. Then, we look at the medical centre count on each cluster to determine best local wards for our business purpose.

First, we use scikit-learn library to perform scaling on our features. It is important to do scaling before feeding the features to algorithm. As clustering techniques use Euclidean distance to identify the neighbours, without scaling, features with large values, such as population density would dominate the distance metric compared to other features, like children/adult population percentage in the ward.

4.1 Algorithm Used and Parameter Tuning

There are number of clustering algorithms readily implemented on scikit-learn package in Python. We have tried three common approaches, which are K-means, agglomerative and DBSCAN clustering algorithms. Looking at the cluster allocations on all three approaches, we decided to proceed with the K-means algorithm.

Important parameter to determine in K-means approach is the number of clusters to be used, which is the value K. For this, we plot the sum of squared distance (SSD) of samples to their closest cluster centre with varying cluster numbers. Then, we use *elbow-method* to pick the best K-value for the dataset clustering. Figure 11 below shows the variation of the SSD with the cluster numbers and looking at the decrease of the SSD and gradient variation, we pick ***K = 4*** as our best cluster number.

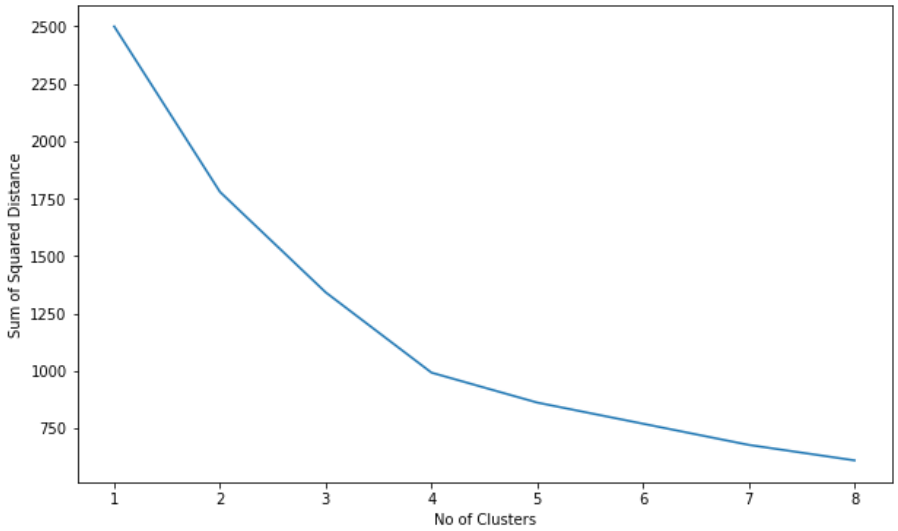

Figure 11: SSD vs Cluster Numbers

4.2 Cluster Analysis

Here, we first look at the mean values for all the features for each of the four clusters. We also illustrate box plots for each feature separated by clusters.

	Population Density	Growth Coeff	% All Older people aged 65+ - 2015	% All Children aged 0-15 - 2015	Medical Centre Count	Count
Cluster-0	4494.954294	64.854493	16.786284	18.613186	12.250000	204
Cluster-1	14354.709188	99.592653	8.690110	17.038291	14.748634	183
Cluster-2	7316.774520	119.181886	10.149063	23.100405	12.045455	220
Cluster-3	6598.366225	1017.007281	6.913234	21.187281	17.000000	18

Looking at the numbers, we have summarized following characteristics about our clusters.

Cluster – 0: Low-Density & Low-Growth suburbs

These is the second highest number of suburbs (204 wards) in Grater London area. Characteristics of these suburbs close to more of established neighbourhoods, where you have less population density, higher old (65+) age population and low growth factor. Even the young kid population is comparatively lower in these suburbs.

Cluster – 1: High-density suburbs

These are the local wards with high population density. Red verticals in population density box plot shows (Figure 12), significantly higher population density than the mean London population density for almost all of the suburbs in Cluster 1.

Cluster – 2: Average suburbs

Just over one-third of the wards (220) are assigned to this cluster. Looking at the statistics of this cluster, it appears to be the cluster that contain wards with average statistics. In each box plot, orange vertical is corresponding to cluster 2. Blue solid line across each graph shows the mean value across all the 625 wards for the respective feature. As can be seen, orange vertical bar is more or less at the mean level for each feature.

Cluster – 3: High-Growth/Booming suburbs

These local wards have very high population growth rate. Comparatively, they have lower older (65+) age population and lower population density, suggesting they could be new suburbs with young families, as they have higher percentage of young kids in the population as well.

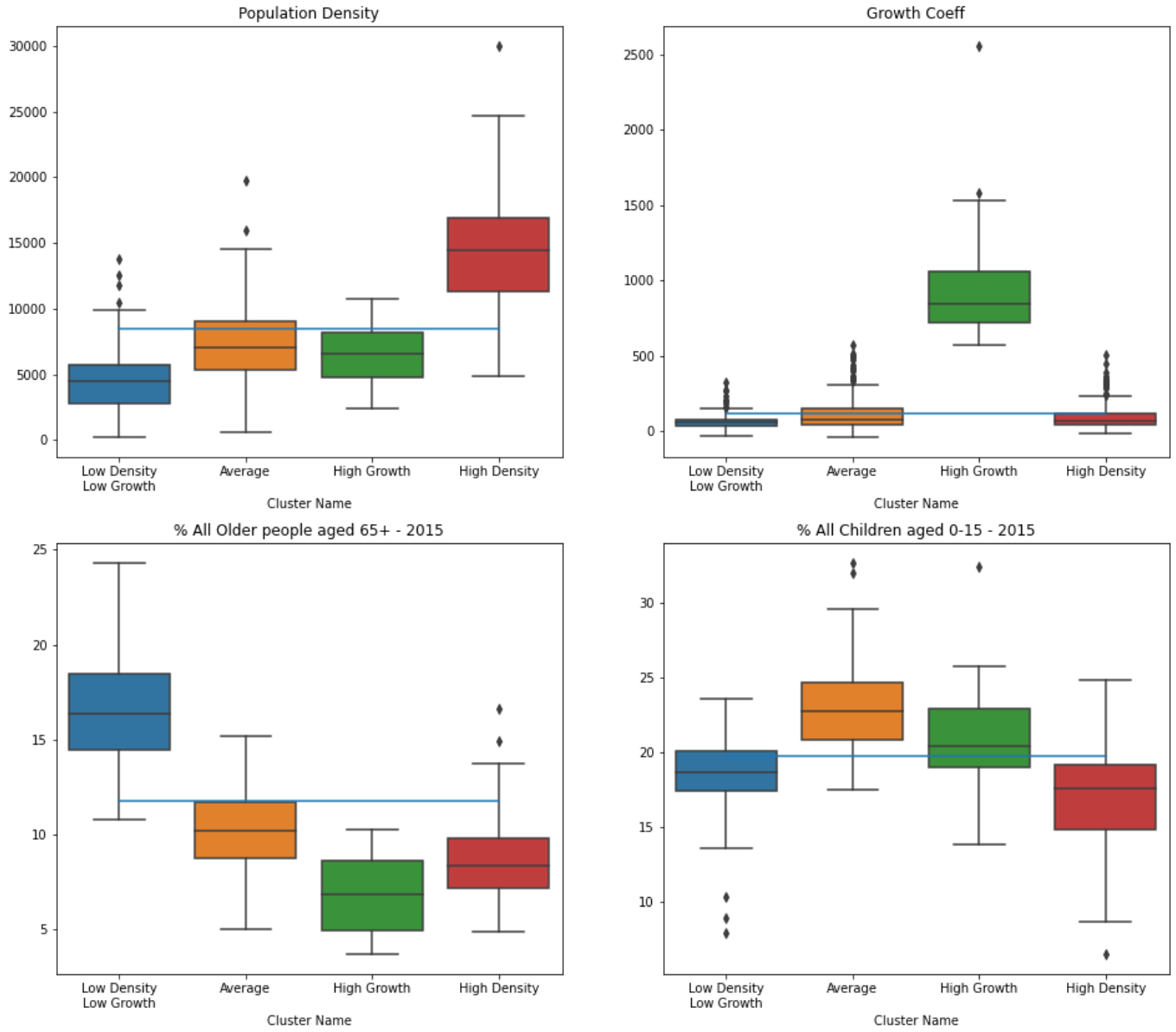

Figure 12: Feature Statistics for each Cluster

We now investigate the medical centre count variation within each cluster. Plotted below in Figure 13 is both box and swarm plots for medical centre distribution in each cluster. In the figure, each grey dot is a local ward with respective medical centre count as shown in the vertical axis.

One interesting characteristic with the *High-Growth* wards is that they tend to have considerably higher medical centre count compared to other clusters. Looking at the numbers given in the table above, mean medical centre count for this class is 17, and minimum value is 12, these are comparatively higher numbers with respect to other clusters.

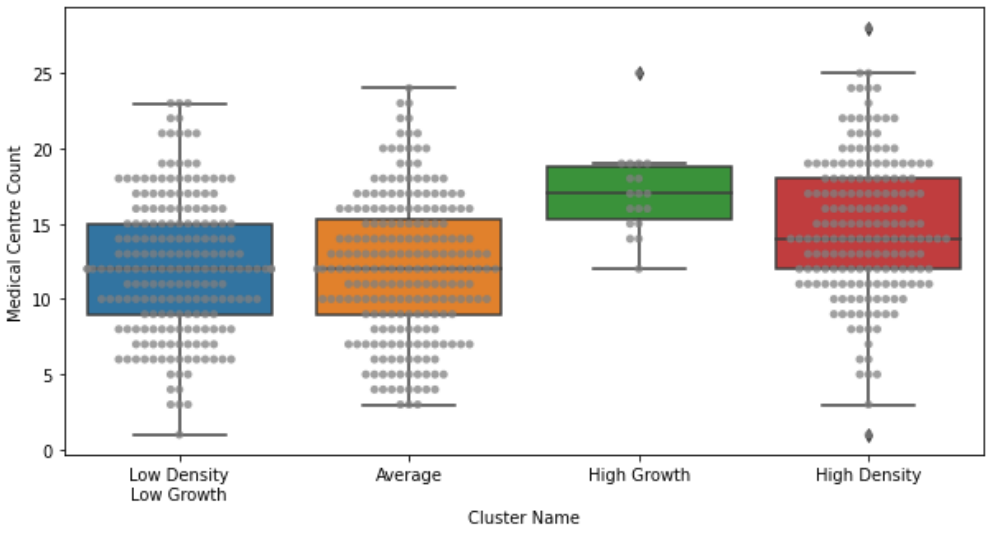

Figure 13: Medical Centre Count Distribution within Clusters

4.3 Cluster Visualization in Choropleth Map

We use Geopandas for creating a choropleth map of Greater London area with our clusters. Geopandas data frames are somewhat like Pandas data frames. Best of all, Geopandas allows to create standalone choropleth maps without many other dependencies. For the shape files that Geopandas required for mapping is readily available in London Datastore. We found number of source files with significant level of details in following location, https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london.

One important point to note here (which was also mentioned at the start of this project) is that our starting dataset (ward profiles dataset) has City of London as one local ward, but it is in fact a borough. So, in our shape file (mapping data), we have 25 wards those belongs to City of London borough, but we don't have individual feature details for all these 25 wards in our dataset. Hence, what we do is, we use same figures for City of London to all 25 wards in mapping Geopandas data frame.

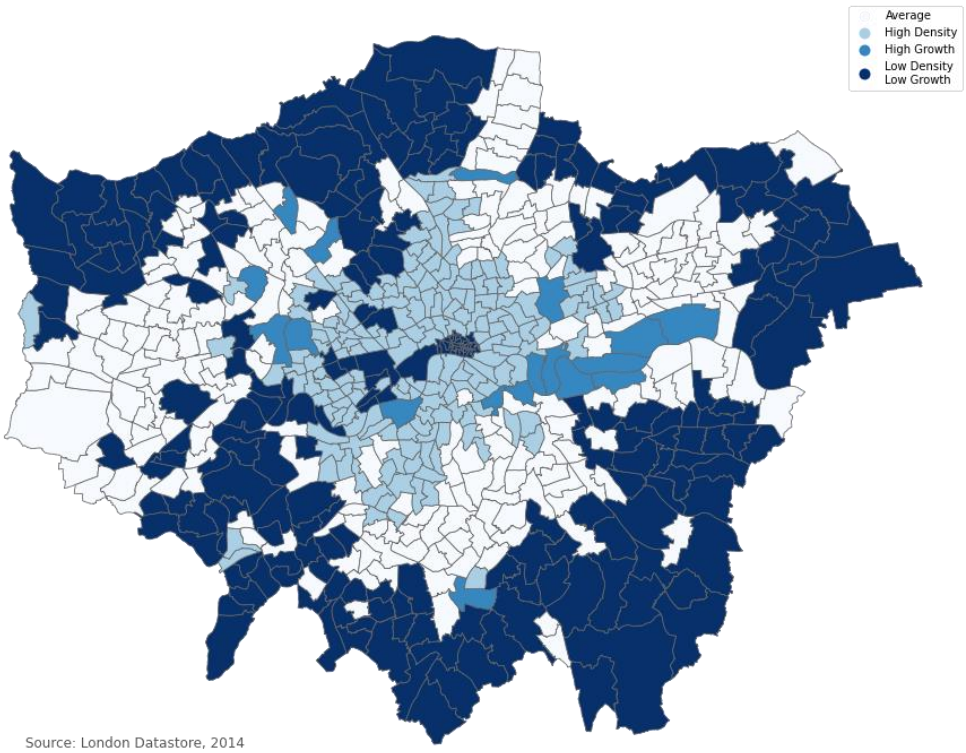

Figure 14: Local Ward Classification based on Population Features in Greater London Area

It is clear from the choropleth map that, most of the low-density/low-growth suburbs are at the fringe of the Greater London area. Most of the high-density wards are closer to City of London, yet, it is interesting to note that City of London itself has been classified as a low growth suburb. One possible explanation to this would be the industrialization and possibly high number of commercial premises in the area, resulting lower number of residential properties and population count per square meter.

5. Results and Discussion

5.1 Filtering Criteria for Local Wards and Outcomes

For the purpose of our business idea, which is to pick few candidate wards for starting a new medical practice, we look further into each cluster in this section. We discard low-density/low-growth local wards from our analysis because they do not seem to provide any profitable perspective for the business.

To further filter out the wards to pick handful of localities for positioning our practice, we follow below logic:

- Cluster 1, 2 and 3, named *Average*, *High-growth* and *High-density* wards are most suitable for the purpose.
- Within those clusters, wards positioned *above the mean value* of respective population features are considered. Here, we only consider three features in our dataset, i.e., population density, percentage of children population and population growth coefficient.
- Next, we look at the competition in those neighbourhoods. We consider only suburbs those reside below the 25th percentile with respect to medical centre count.

Note, we disregard 65+ old population from our filtering criteria, assuming that we are interested in a paediatric-related medical centre, in which case, it is more important to consider on the population density, growth rate and (0-15) year old percentages in the local ward.

Below are the 10 local wards that fall into our criteria. There are 5 wards with *High Density* and another 5 with *Average* category.

	Label	New code	Borough	Ward	Population - 2015	Area - Square Kilometres	Population Density	Growth Coeff	% All Older people aged 65+ - 2015	% All Children aged 0-15 - 2015	Medical Centre Count
0	High Density	E05000420	Lambeth	Coldharbour	17500	1.2	14583.333333	108.574484	6.091777	20.241157	11.0
1	High Density	E05000480	Newham	East Ham Central	16200	1.0	16200.000000	171.955910	7.529804	22.422633	8.0
2	High Density	E05000535	Southwark	Camberwell Green	15950	1.0	15950.000000	168.438274	6.969602	21.297399	10.0
3	High Density	E05000550	Southwark	South Bermondsey	14400	1.0	14400.000000	162.376548	7.374775	19.765506	6.0
4	High Density	E05000578	Tower Hamlets	Bromley	19750	1.1	17954.545455	338.839681	4.957924	24.632465	10.0
5	Average	E05000246	Hackney	Springfield	12800	1.0	12800.000000	181.457411	7.475759	32.694714	5.0
6	Average	E05000498	Redbridge	Chadwell	14850	1.5	9900.000000	142.317167	10.076788	25.414253	5.0
7	Average	E05000507	Redbridge	Loxford	17700	1.3	13615.384615	283.987711	6.871824	29.192547	7.0
8	Average	E05000510	Redbridge	Newbury	17450	2.1	8309.523810	310.135929	9.808640	24.808067	8.0
9	Average	E05000512	Redbridge	Seven Kings	16950	2.1	8071.428571	280.408912	8.996519	24.877588	7.0

Figure 15: Desired Local Wards with their Statistics

5.2 Desired Local Wards Visualization in Choropleth Map with Markers

In this section, we visualize those filtered local wards on a Greater London area map. In the choropleth, we use binary colour scheme, where unshaded areas represent the desired local wards. We again use Foursquare API to fetch medical centre details for those 10 desired wards, which we will then use to create markers on the map. This will allow us to further zoom into each local ward and see where existing medical centres are in the neighbourhood.

In Figures below, we can see 84 medical centres on the map, those are merged into two clusters. Zoomed in excerpt from the map shows 3 local wards with separated clusters and markers showing existing medical centre positions.

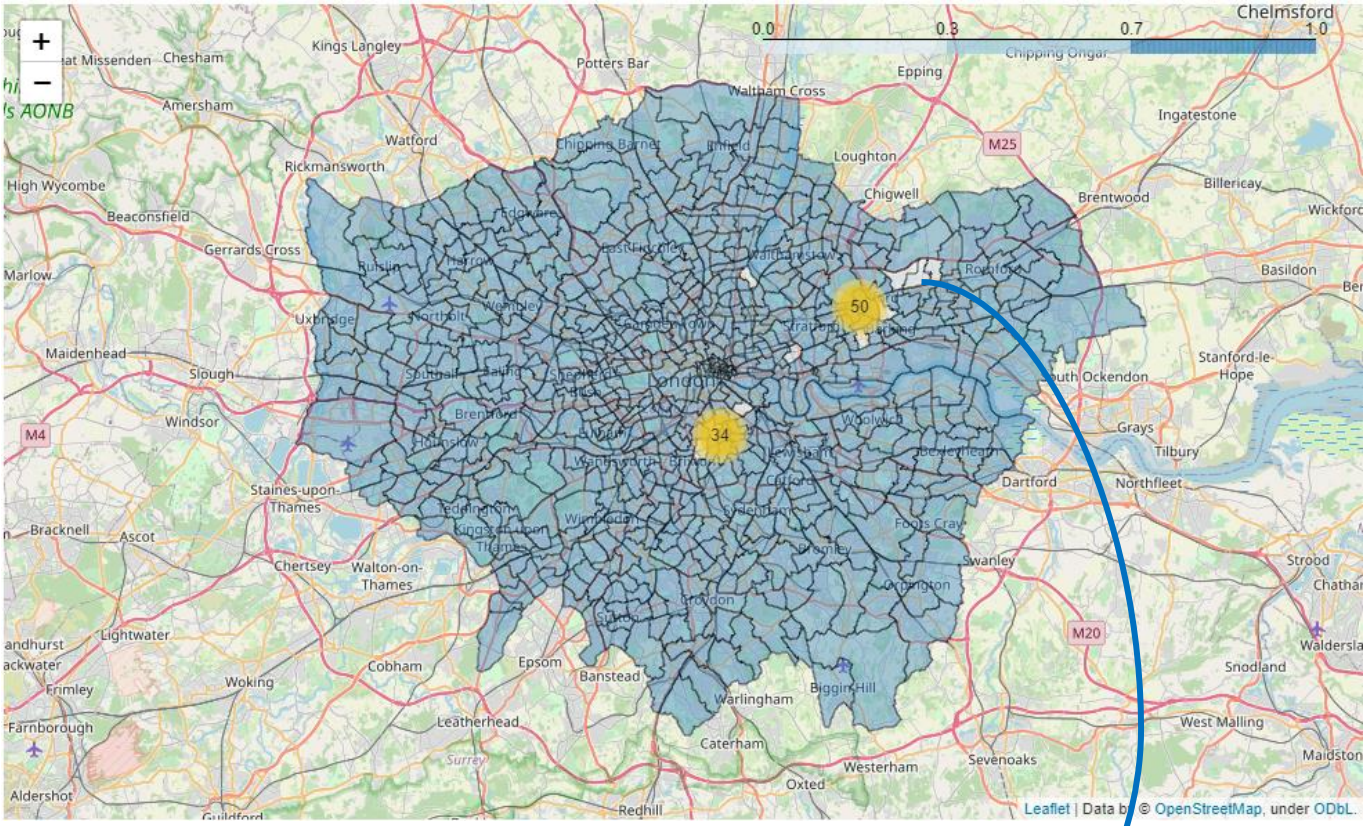

Figure 16: Desired Local Wards (unshaded) in Map.

Note: Numbers on the map show merged marker which represent medical centres in the neighbourhood.

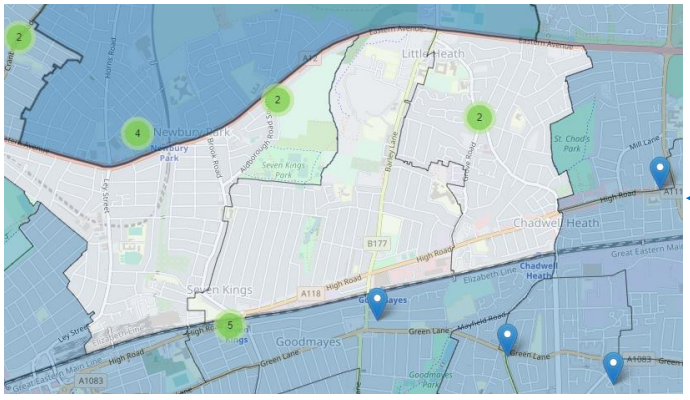

Figure 17: Zoomed in Map showing local wards, Newbury, Seven Kings and Chadwell.

We will not go into street level details in each local ward for suggesting a specific location for a medical centre. However, just for the sake of illustrate the idea, we pick up one ward and investigate bit more into details here.

Local ward of our interest here is *Bromley – Tower Hamlets.* As can be seen from the snippet, this has very high population density and relatively high growth coefficient and percentage of children in the population. Markers on the map show the location of those 10 medical practices in the neighbourhood. Looking at where the existing practices are, we can suggest following location encircled in purple for a new medical centre. Obviously, there are other factors to be considered by any business owner, for instance, commercial rental prices, government or local council restrictions and closeness to public transport, etc.

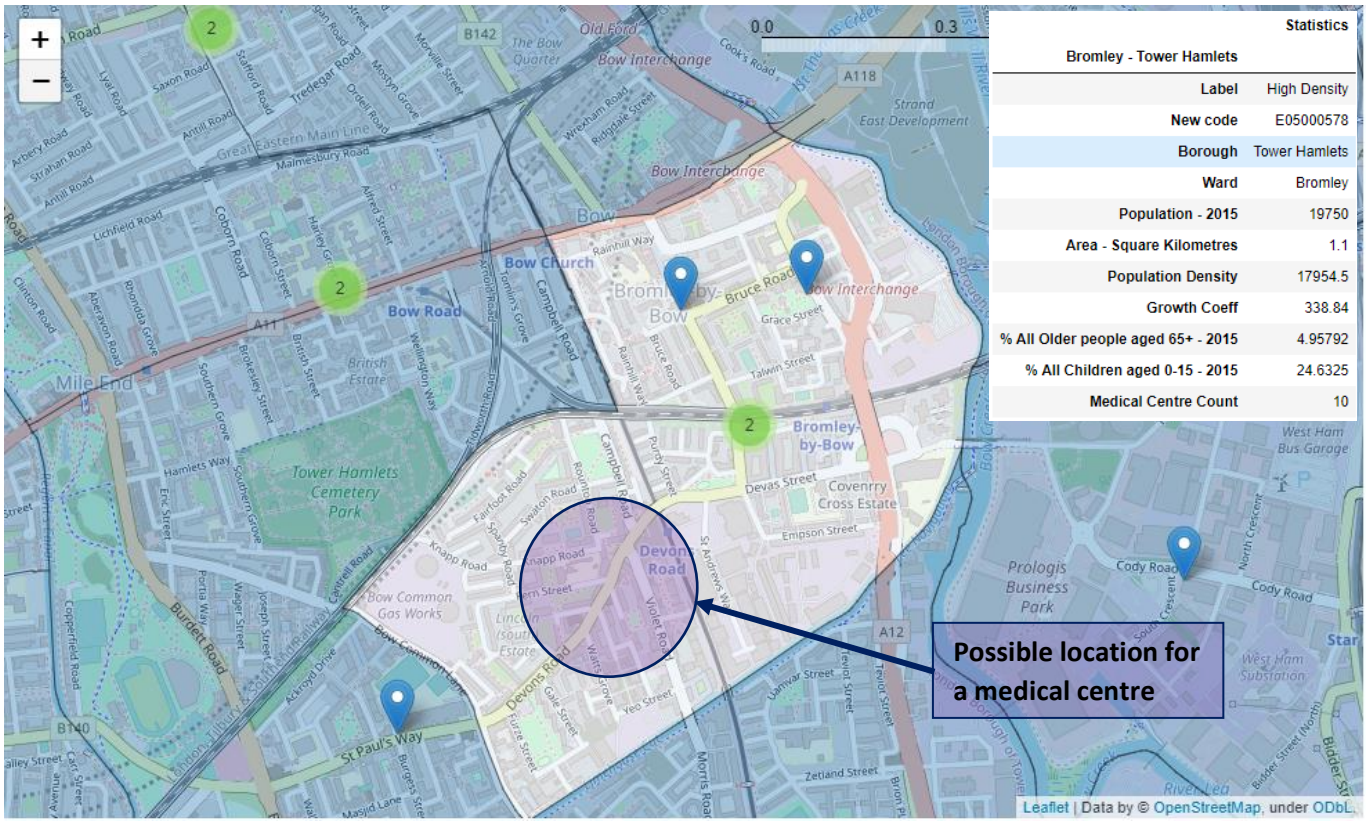

Figure 18: Bromley, Tower Hamlets Medical Centre Locations

6. Conclusion

In this study, we looked at 625 local wards in Greater London area to figure out suitable neighbourhoods for positioning a new medical centre. We primarily used datasets available in London Datastore and Foursquare API to get ward profile data for each local ward and existing medical centre details. After looking at descriptive statistics for each population related features, we segmented 625 wards into 4 different clusters using K-means clustering algorithm. Looking at the feature distribution, we identified that our machine learning algorithm has created clusters, which demonstrate characteristics similar to high density, high growth, low density/growth and average suburb profiles. We then investigated medical centre distribution on each cluster and then defined a further optimization approach to pick handful of local wards that fits into our business proposal. Based on that, we have identified 10 wards in Greater London area. We used Folium and Geopandas libraries in Python to generate choropleth maps and then used markers to show the medical centres on the map.

References

1. London Datastore, (https://data.london.gov.uk/)
2. Foursquare API, (https://developer.foursquare.com/)
3. Geopandas Library, (https://geopandas.org/mapping.html)
4. Folium Library, (https://python-visualization.github.io/folium/modules.html)
5. https://towardsdatascience.com/