Deprivation and Venue Data Analysis of London

Varnan Goodwin

February 2020

Contents

1. Introduction

1.1 Background

London, the capital of England and the United Kingdom, is a 21st-century city with history stretching back to Roman times. London is considered to be one of the world's most important global cities and has been termed the world's most popular for work city. The metro population in 2020 is estimated to be as much as 9.30 million according to the UN's World Urbanization Prospects. London has a diverse range of people and cultures, and more than 300 languages are spoken.

UN's World Urbanization Prospects. London has a diverse range of people and cultures, and more than 300 languages are spoken.

Investigators are looking to grab the opportunities in the investment friendly city of London in 2020, and the rich and mighty are seeking to build their mansions in the wealthiest parts of London. Despite there being greater economic opportunity, there are people who have gone through great hardship through 2019 or are struggling to get by. A significant number of local people continue to face persistent inequalities and are disproportionately affected by poverty, unemployment, long term health conditions and welfare dependency. There are also groups who are in work but are struggling with the rising cost of living as well as job and wage insecurity. Some areas have changed immeasurably over the last decade. Once neglected, now desirable places for business. However, as is typically the way with inner-city gentrification, there are many that have been left behind – particularly among poorer communities.

1.2 Business Problem

Charities/ NGOs/ local authorities and people with compassionate hearts are looking for ways to reach out to the poor & needy communities and help them out in their struggles. Though there are number of different statistics available for information and research, it is difficult to obtain information at one place to satisfy these needs, get useful data in front of the right people in the right format that can be used to help make decisions.

This project aim to analyse and identify most deprived neighbourhoods for development in London using data science methodology and machine learning techniques.

1.3 Target Audience

The results and recommendation of the projects will be useful for Charities, NGOs and local authorities to identify most deprived neighbourhoods and initiate, organise and carry out appropriate development projects for the affected areas.

2. Data

2.1 Data used to solve the problem

Deprivation describes the lack of material benefits, such as a job, income, decent home and education that are generally considered to be necessary in a society. Relative levels of deprivation are a crucial determinant of 'need' for many of the services that local authorities provide. Deprivation is the key driver of need in many demand-led services and a key cost driver in the current local government funding formulas and significantly affects the distribution of funding. Measures of deprivation are used in adjusting funding for Social services, Children's services, Environmental services, Fire and Rescue services and the police.

Index of Multiple Deprivation (**IMD**) draws together information from the following seven sources of official data, known as 'domains', to produce an overall measure of relative deprivation between one area and another

- Income Deprivation (22.5%)
- Employment Deprivation (22.5%)
- Education, Skills and Training Deprivation (13.5%)
- Health Deprivation and Disability (13.5%)
- Crime (9.3%)
- Barriers to Housing and Services (9.3%)
- Living Environment Deprivation (9.3%)

As IMD helps us to identify most deprived neighbourhoods that require development, this will be the main data used for this study and I acquired the all required data from the following sources

- **Postal data with IMD** from London data store

  Info. including Postcode, Latitude, Longitude, Ward, District, Constituency and IMD

  Along with IMD data we will also use Homelessness and recorded crime data for the analysis

- Venues from **Foursquare API** for clustering on neighbourhoods

  Massive data set accurate location data

- Geometry coordinates London geojson from GitHub

- List of London sub-regions from Wikipedia for the analysis by sub-regions of London (Central, East, North, South, West)

2.2 Data cleansing

- **Postal data with IMD** from London data store – London-postcodes had old post codes as well . Due the size of data set could not import it. I was able to import it after deleteing the old post codes

    1. **I call this csv file** London_postcodes _InUse

       The shape of the file is (178344, 27)

       It also has number of info. Which are not relevant to our study namely 'Easting','Northing','Altitude','London zone', 'Quality', 'User Type','Distance to station' etc

    2. I then created a new data frame with relevant features for our study

       The crime data set has 24 months of crime details of major and minor offences by districts

       I summaried to get the 2019 crime told per district which is relevant for our study

2.3 Feature selection

Though IMD is our main factor for identifying the deprived areas for development, the additional information, namely, Homelessness and Crime are very useful when prioritising the development work. Also, it's useful to analyse whether the deprived areas have high homelessness and Crime

Selected features:

Ward, Latitude, Longitude, District, Constituency, Index of Multiple Deprivation (IMD) from London postcode data set

District and Homeless from Homeless data set

Districts and Crime stats for 2018, 2019 from Crime data set

.

3. Methodology

- We have collected the required **data:**
    - Location (latitude & longitude) and Index of Multiple Deprivation (**IMD)** for each Ward in London.
    - Homeless and Crime statistics of the district the Ward belongs to.
    - Sub-region of the Ward and District

- The next step is to analyse the data sets to summarise their main characteristics with visual methods. This is called Exploratory data analysis (EDA)

- The last step will be to identify cluster of most deprived Wards of the London districts that requires development. For this project, we are going to focus on Sports development as the development type.

3.1 Exploratory data analysis (EDA)

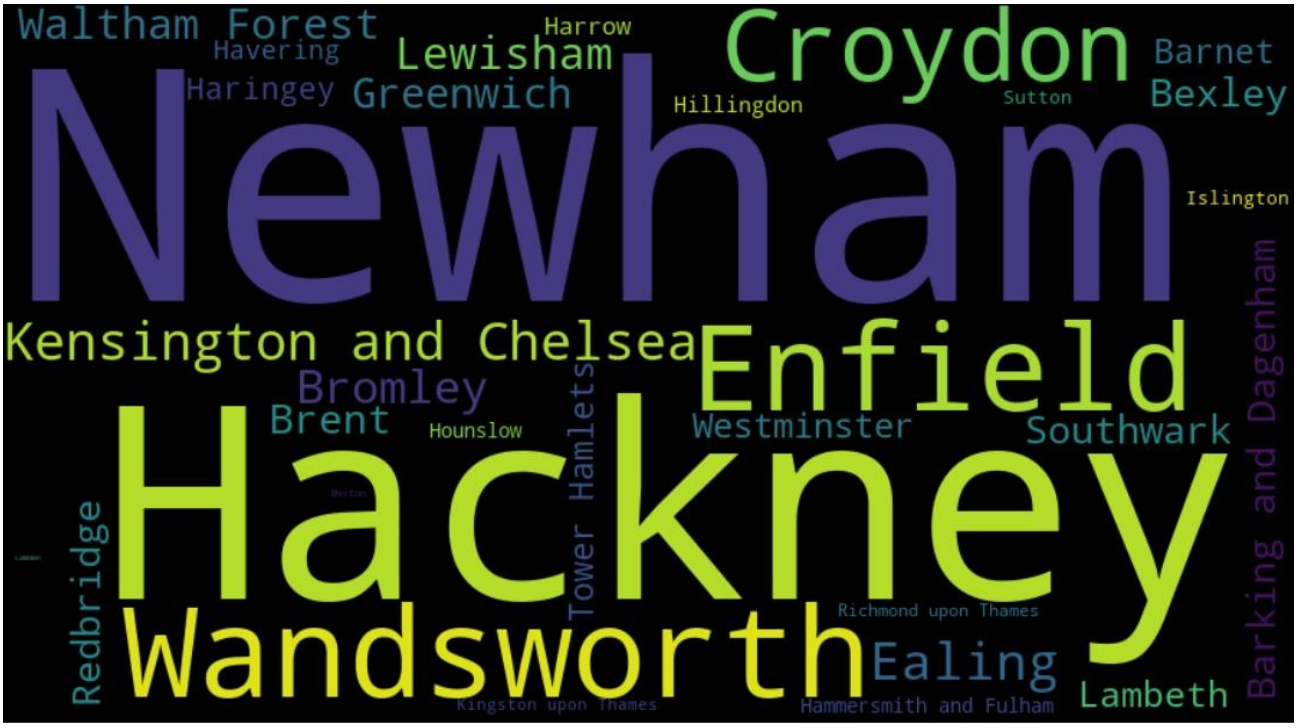
After groping the data by districts I performed the following

- Ward cloud analysis
- Regression plots
- Bar charts
- Box charts
- Histograms

3.1.1 Word Cloud Analysis - Index of Multiple Deprivations (IMD)

We can see from above plot that the wealthiest (IMD high) London districts are 'Richmond upon Thames, City of London, Kingston upon Thames. The most deprived (low IMD) districts are Barking and Dagenham, Hackney, Islington, Newham

3.1.2 Word Cloud Analysis Homelessness



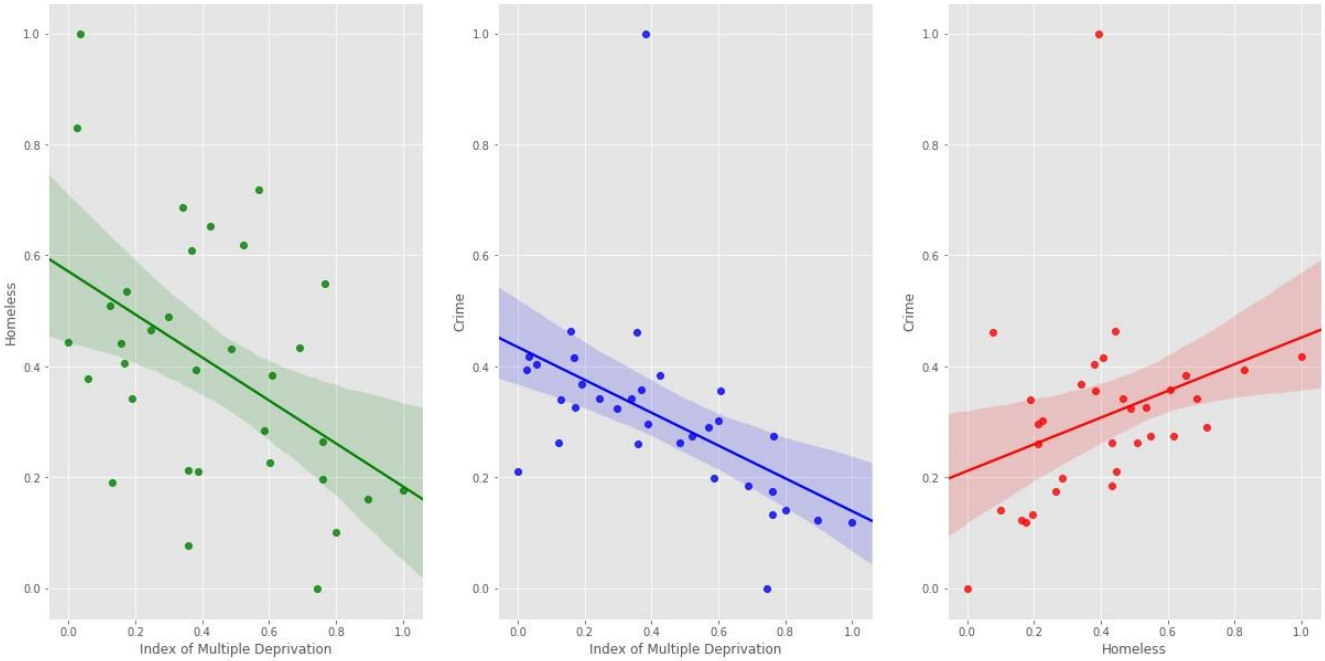
The most affected wars are Hackney, Newham and Wandsworth

3.1.1 Word Cloud Analysis Crime

Highest numbers of crimes were committed in the Westminister district and this could be due to the fact it's a popular tourist area. Hackney and Islington are among the high crime area

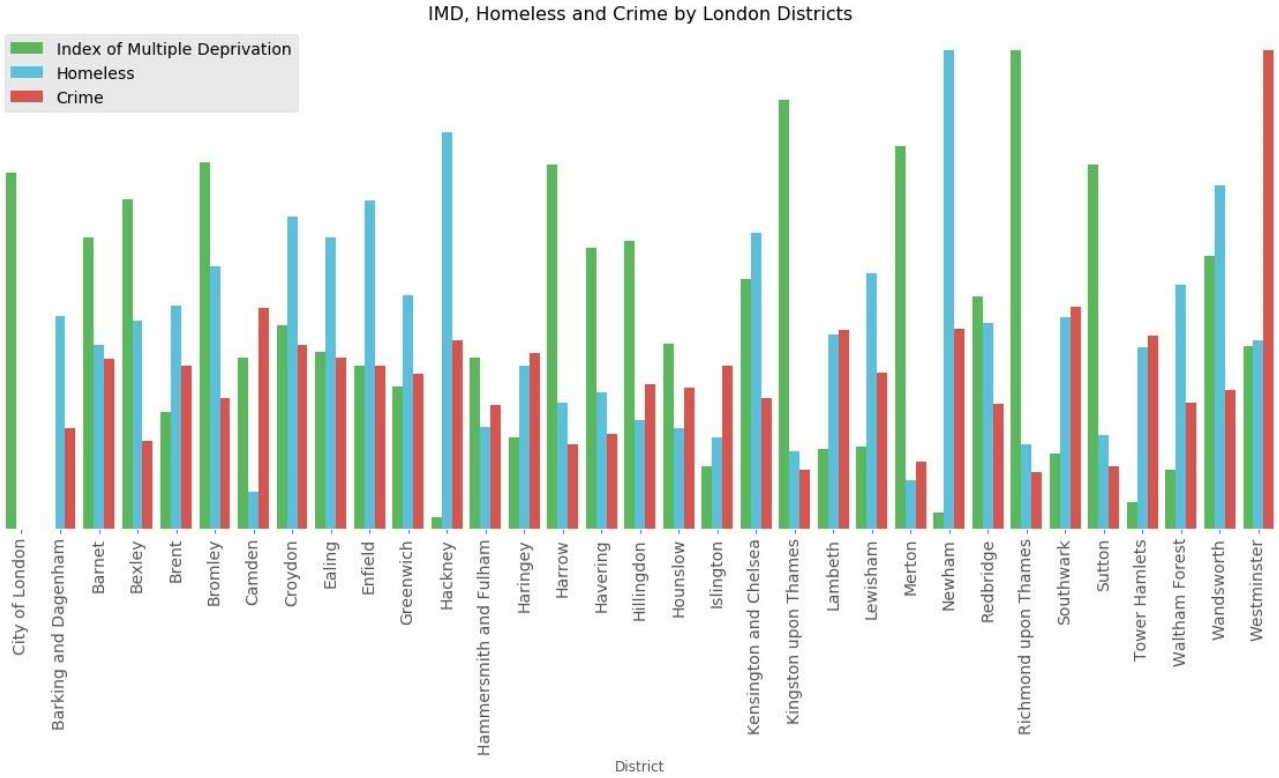
3.1.3 Regression plot of IMD, Homeless and Crime by District

As shown below is the strong negative linear regression IMD and Crime. As stated in the Data section, this could be the fact that 9.3% of the IMD is crime data

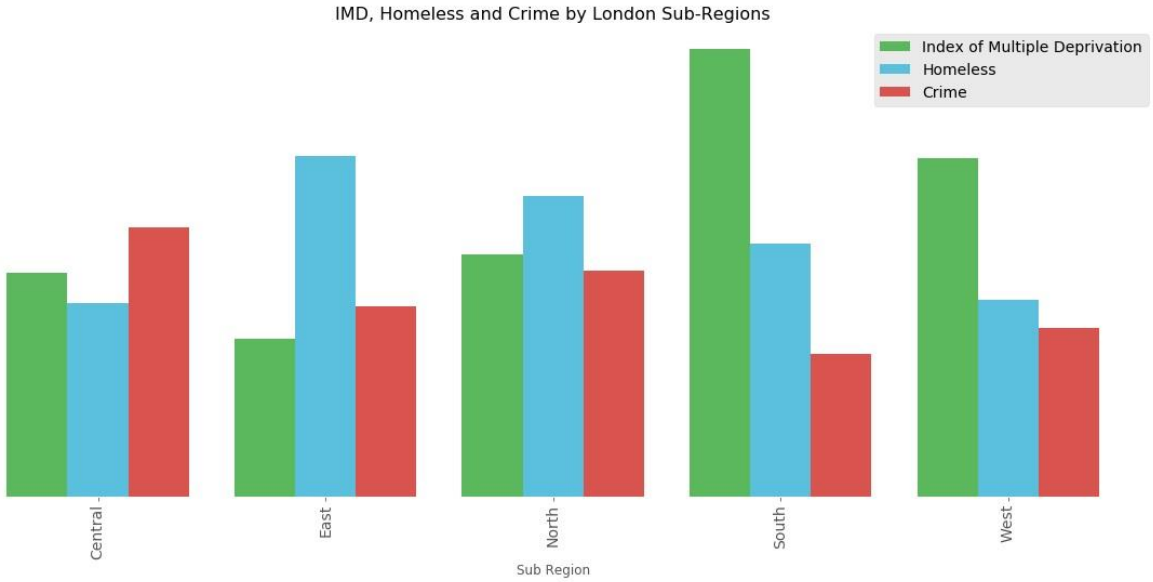


3.1.4 Bar chart analysis of IMD, Homeless and Crime by District

After normalising the data, I plotted a bar chart for IMD, Homeless and Crime data by districts and it confirms the findings from the three Word cloud analysis



3.1.4 Bar chart analysis of IMD, Homeless and Crime by Sub-regions

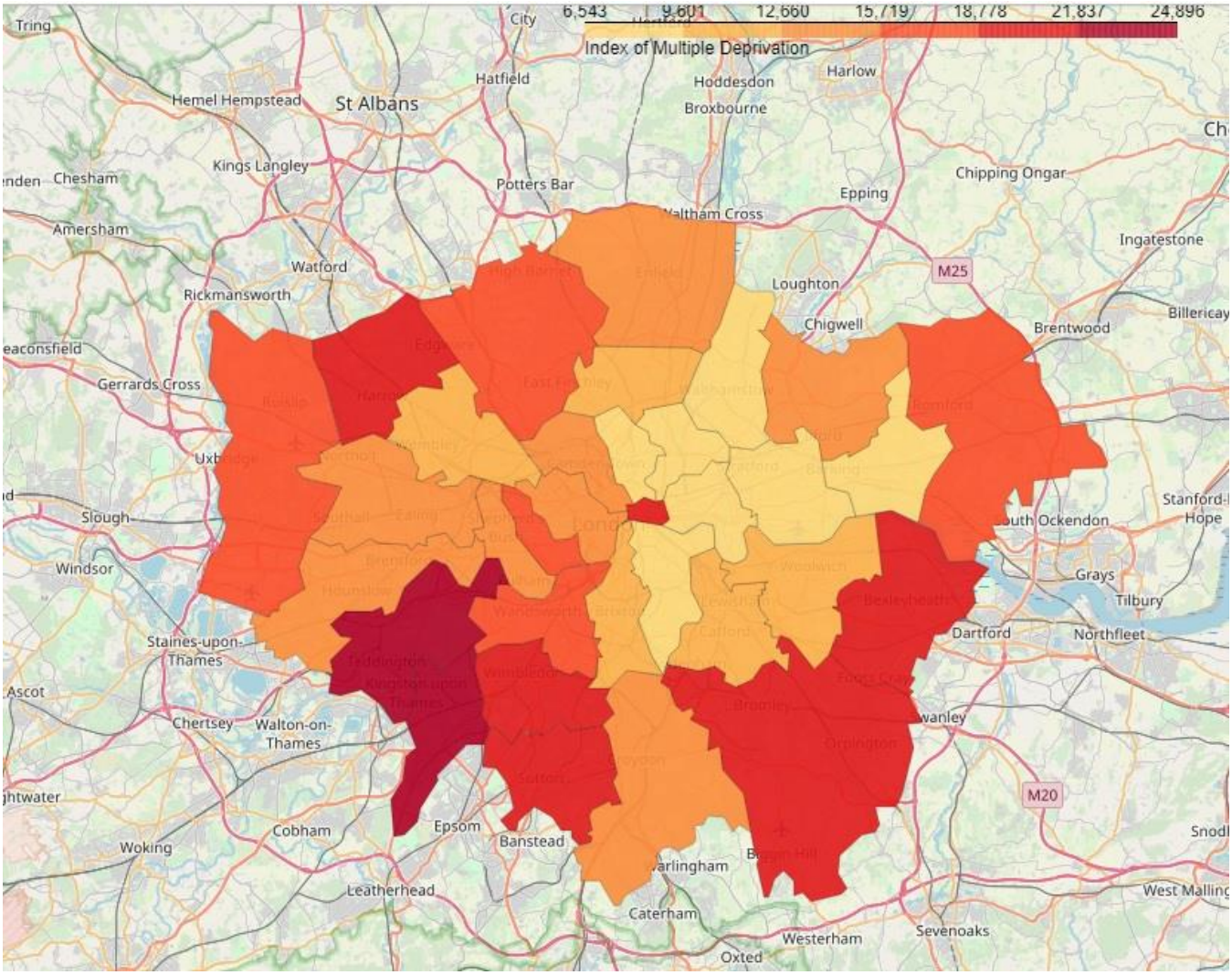


From the chart we can see East London sub-region is the most affected sub-region

3.1.5 Choropleth Index of Multiple Deprivation (IMD)

I used python **folium** library to visualize geographic details of London districts in a choropleth map - a thematic map in which areas are shaded or patterned in proportion to the measurement of the statistical variable being displayed on the map.

I used the geo json file form GitHub to create the areas/boundaries that match the London districts in the data file. The IMD legend is displayed in the upper right corner

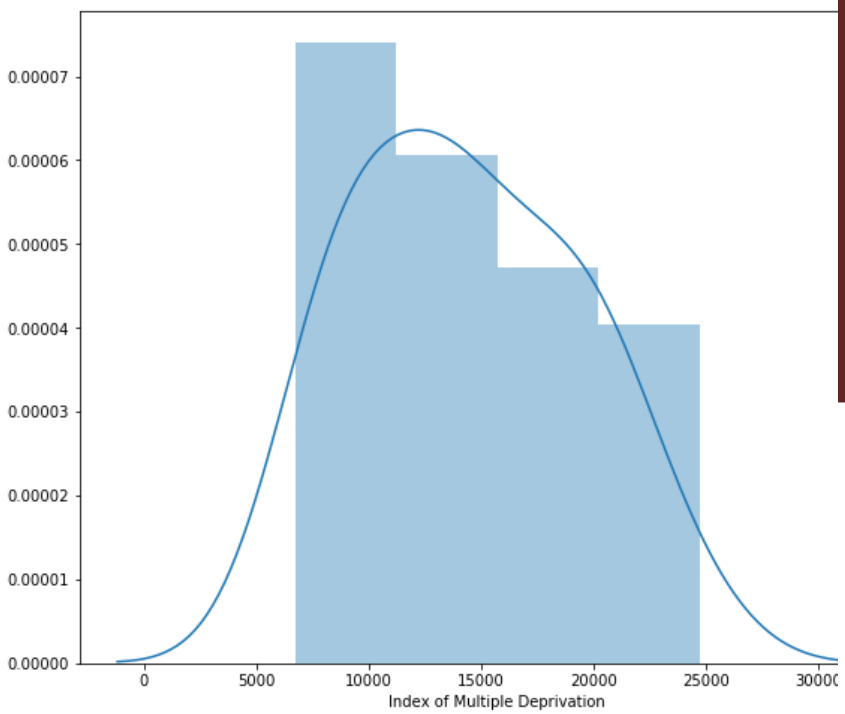


The Darker areas are the prosperous districts while the light shaded ones are the deprived districts that we need to explore more.

3.1.6 Histogram analysis of IMD

```
# plot boxplot with seaborn
plt.figure(figsize=(10,8))
sns.distplot(LondonDistrict_df['Index of Multiple Deprivation'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5927d4dc18>
```



Descriptive Statistics

count	33.000000
mean	14362.550517
std	4995.885040
min	6722.442784
25%	9826.488235
50%	13597.264217
75%	17657.859982
max	24715.858600

Name: Index of Multiple Deprivations

I then categorised the IMD values as High, MidHigh, Mid, MidLow and Low based on histogram analysis and descriptive statistics

	District	Sub-Region	Latitude	Longitude	Index of Multiple Deprivation	IMDCategory	Homeless	Crime2018	Crime2019
0	City of London	Central	51.514622	-0.092233	20105.585622	High	7	3063	3255
1	Barking and Dagenham	East	51.546853	0.126620	6722.442784	Low	512	16740	18530
2	Barnet	North	51.608680	-0.206189	17657.859982	MidHigh	444	25855	28899
3	Bexley	East	51.460000	0.135774	19129.949361	High	500	13904	16619
4	Brent	West	51.555335	-0.259383	11132.975313	MidLow	536	28416	27960

This helped me to identify IMD high districts and IMD Low districts (most deprived)

IMD high districts
'Bexley', 'Bromley', 'City of London', 'Harrow', 'Kingston upon Thames', 'Merton', 'Richmond upon Thames', 'Sutton'

IMD Low districts (most deprived)
'Barking and Dagenham', 'Hackney', 'Islington', 'Lambeth', 'Newham', 'Southwark', 'Tower Hamlets', 'Waltham Forest'

We will focus just the most deprived districts (IMD Low districts) for our analysis as these are ones development projects. Also, we will choose Sport development as the development type for this project

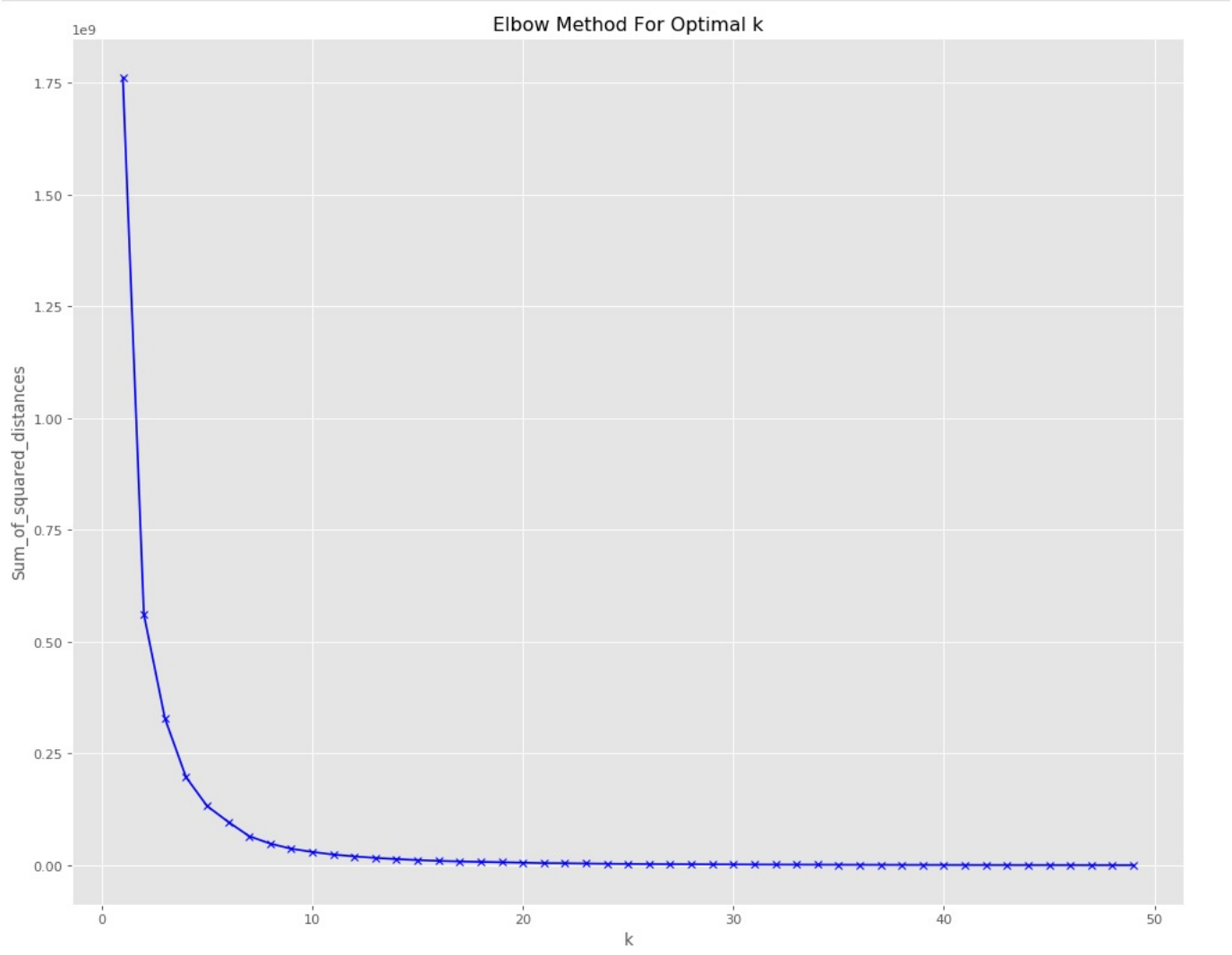
3.2 Clustering Analysis

3.2.1 K-means

I used unsupervised learning K-means algorithm to cluster the Wards. K-Means algorithm is one of the most common cluster method of unsupervised learning. It accomplishes this using a simple conception of what the optimal clustering looks like:

- The "cluster center" is the arithmetic mean of all the points belonging to the cluster.
- Each point is closer to its own cluster center than to other cluster centers.

The Elbow method for optimum K is 5 as shown below, hence we will cluster the Wards into 5 clusters



3.1.6 The Foursquare API

I utilized the Foursquare API to explore the Wards and segment them. I set the limit as 500 and the radius 1000 meter for each Ward

1005 venues were returned by Foursquare and there were 315 unique categories.

I set the num_top_venues to 20 and chose mainly chose the sports category.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	16th Most Common Venue	17th Most Common Venue
0	Abbey	Supermarket	Grocery Store	Metro Station	Gym	Park	Pharmacy	Yoga Studio	Event Space	Gym / Fitness Center	Golf Course	Go Kart Track	Garden Center	Film Studio	Cycle Studio	Home Service	Cricket Ground	Convenience Store
1	Alibon	Metro Station	Supermarket	Sporting Goods Shop	Bus Stop	Grocery Store	Convenience Store	Yoga Studio	Event Space	Gym / Fitness Center	Gym	Golf Course	Go Kart Track	Garden Center	Film Studio	Cricket Ground	Cycle Studio	Home Service
2	Barnsbury	Grocery Store	Café	Supermarket	Music Venue	Park	Yoga Studio	Event Space	Gym / Fitness Center	Gym	Golf Course	Go Kart Track	Garden Center	Film Studio	Cycle Studio	Home Service	Cricket Ground	Convenience Store
3	Beckton	Supermarket	Grocery Store	Gym / Fitness Center	Park	Yoga Studio	Event Space	Gym Pool	Gym	Golf Course	Go Kart Track	Garden Center	Film Studio	Cycle Studio	IT Services	Cricket Ground	Convenience Store	Climbing Gym
4	Becontree	Supermarket	Gym	Yoga Studio	Cycle Studio	Gym Pool	Gym / Fitness Center	Grocery Store	Golf Course	Go Kart Track	Garden Center	Film Studio	Event Space	Cricket Ground	IT Services	Convenience Store	Climbing Gym	Campground

Below is the merged table with cluster labels

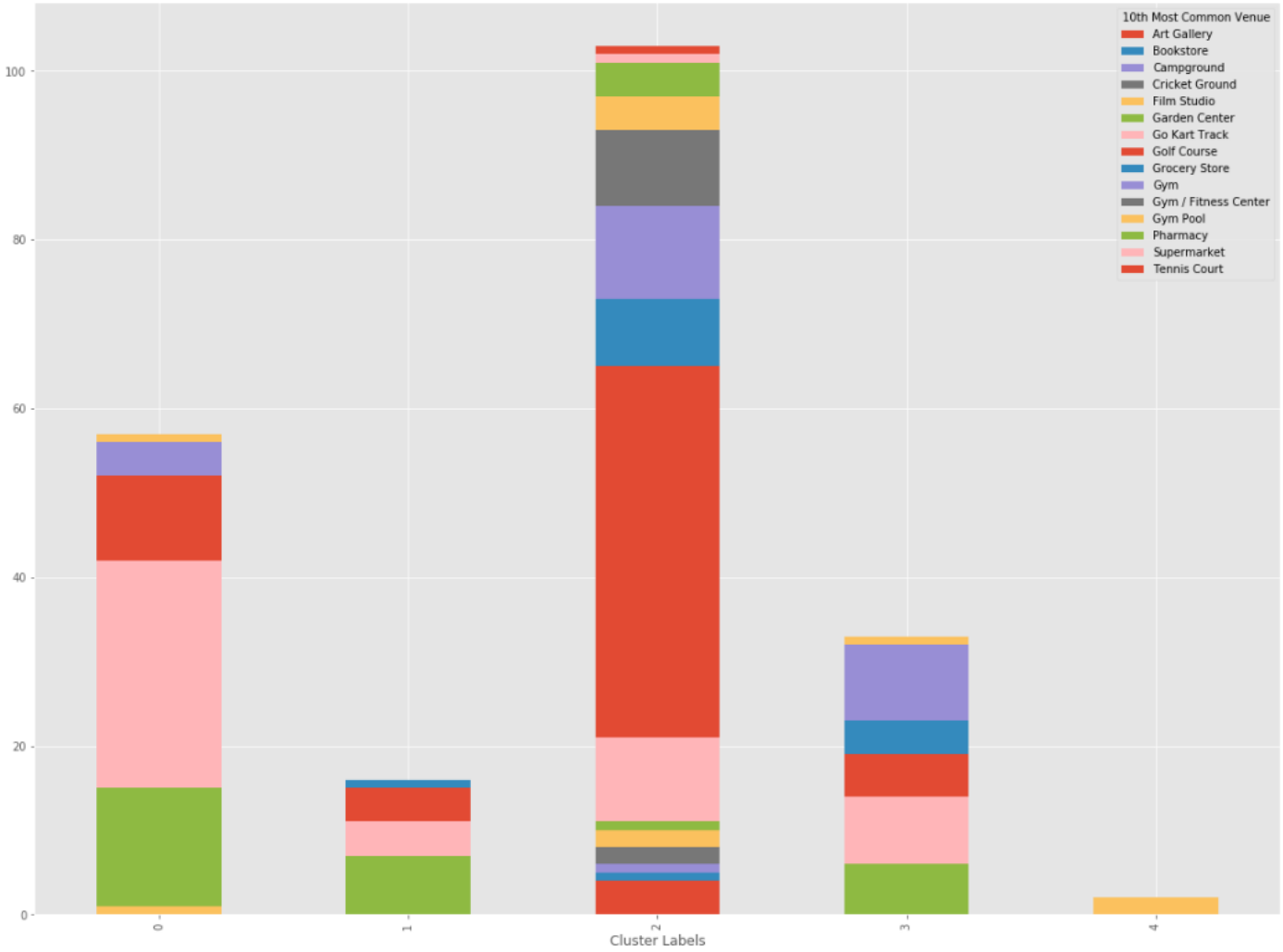
	Ward	Latitude	Longitude	Index of Multiple Deprivation	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue
0	Abbey	51.539509	0.078658	6719.834101	0	Supermarket	Grocery Store	Metro Station	Gym	Park	Pharmacy	Yoga Studio	Event Space	Gym / Fitness Center	Golf Course	Go Kart Track	Garden Center	Film Studio	Cycle Studio
1	Alibon	51.545772	0.149366	5045.475524	3	Metro Station	Supermarket	Sporting Goods Shop	Bus Stop	Grocery Store	Convenience Store	Yoga Studio	Event Space	Gym / Fitness Center	Gym	Golf Course	Go Kart Track	Garden Center	Film Studio
2	Barnsbury	51.537210	-0.111114	9783.516556	2	Grocery Store	Café	Supermarket	Music Venue	Park	Yoga Studio	Event Space	Gym / Fitness Center	Gym	Golf Course	Go Kart Track	Garden Center	Film Studio	Cycle Studio
3	Beckton	51.513422	0.060282	8324.954198	0	Supermarket	Grocery Store	Gym / Fitness Center	Park	Yoga Studio	Event Space	Gym Pool	Gym	Golf Course	Go Kart Track	Garden Center	Film Studio	Cycle Studio	IT Services
4	Becontree	51.554173	0.118678	5992.940217	0	Supermarket	Gym	Yoga Studio	Cycle Studio	Gym Pool	Gym / Fitness Center	Grocery Store	Golf Course	Go Kart Track	Garden Center	Film Studio	Event Space	Cricket Ground	IT Services

I then added an additional information which combines District, Sub-Region, Homeless and Crime to the df . Below are the first 5 rows. This will be displayed on the map that we plot and will be very useful.

District-SubRegion-Homeless-Crime
Barking and Dagenham,East,512,18530
Tower Hamlets,East,437,32412
Hackney,East,949,31702
Islington,Central,223,27860

Feature engineering:

I added a new feature called SportsDevReq? Y/N. If there are no spots categories in the top 12 venues then the Wards requires Sports development.



By examining all 20 most common venues individually we can come to a conclusion that there is no significant difference between the clusters in terms of the venue categories - except for Cluster 4, which only has two venues. Therefore, we could not give any meaningful names for the clusters

We can see from the below Cluster 4 has only two Wards
Clusters value counts
2 79
0 47
3 31
1 14
4 2

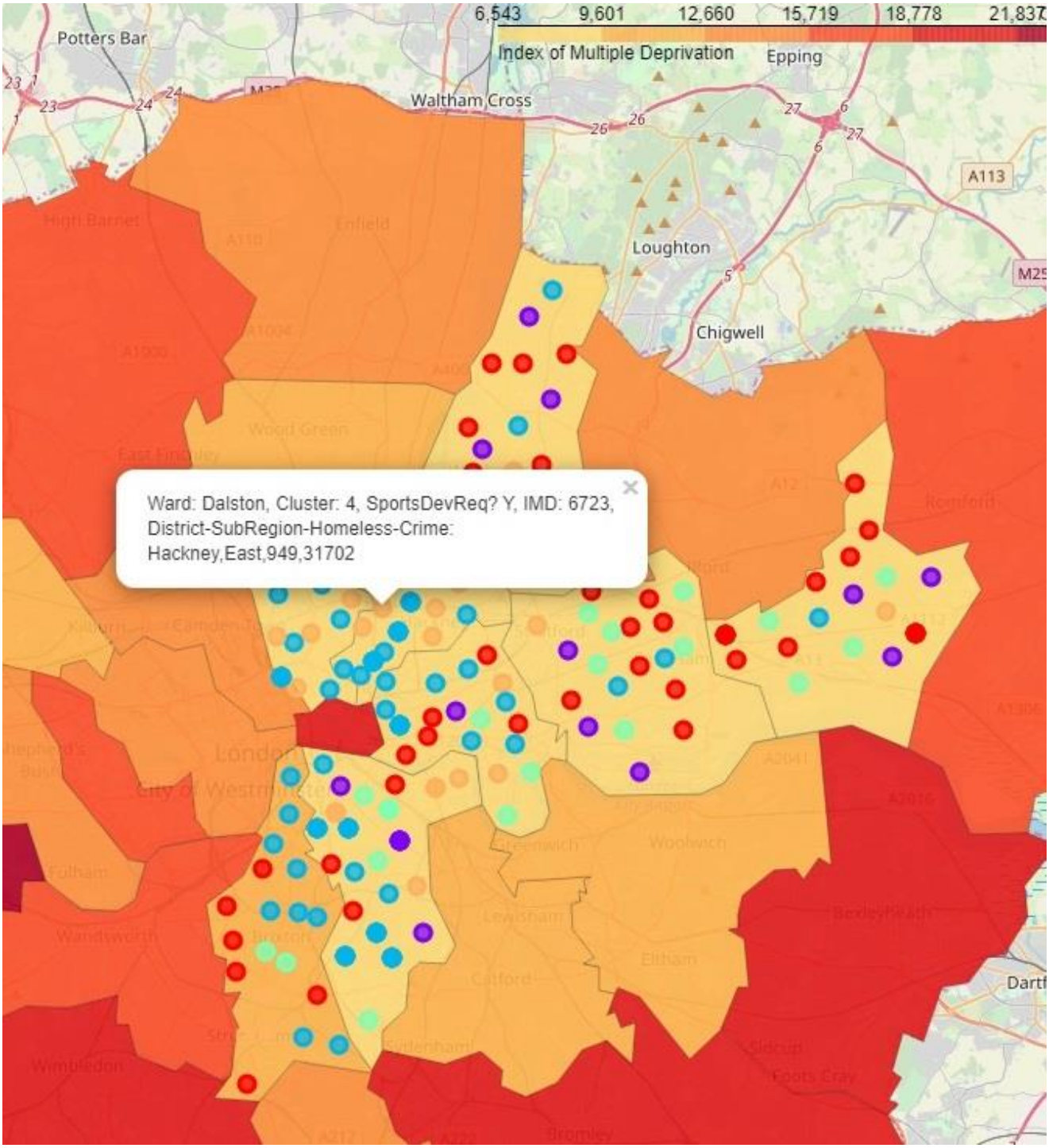
I assigned the 2 Wards from Cluster 4 to Cluster 1, and then assigned all Wards that require Sports development to Cluster 4. By doing do so, we will have a dedicated Cluster just for the wards that require Sports development, which can be easily identifiable in the map

4. Results

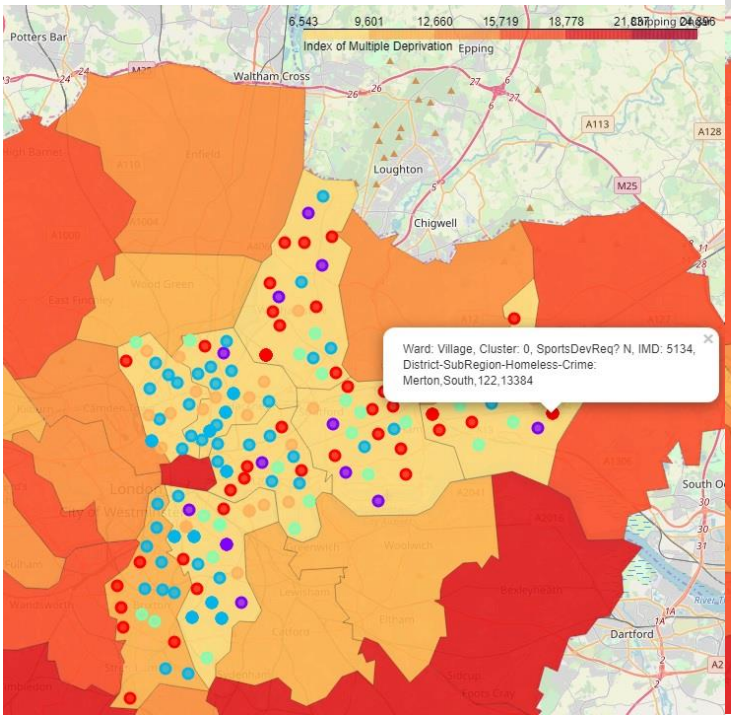
In final section, I created choropleth map which also has the below information for each Ward:

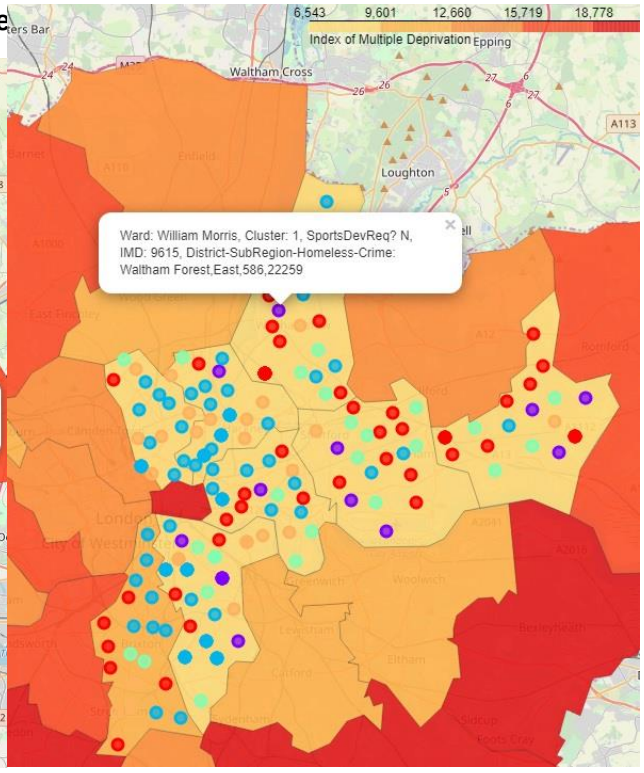
1. Ward name
2. Cluster Label
3. IMD value
4. District
5. Sub-Region
6. Homeless
7. Crime

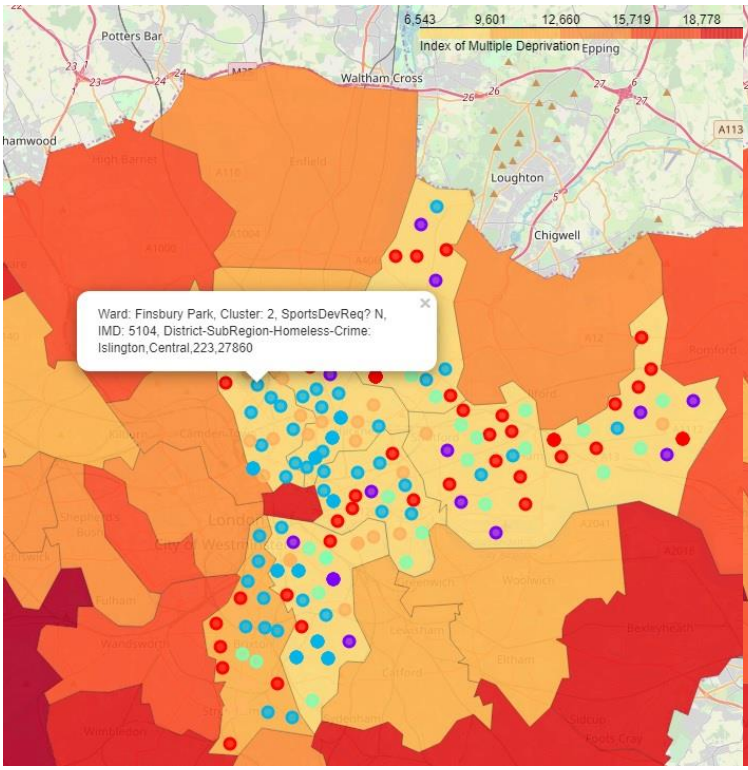
Cluster 4 - Wards with 'Pale brown colour' are the ones that require Sports development

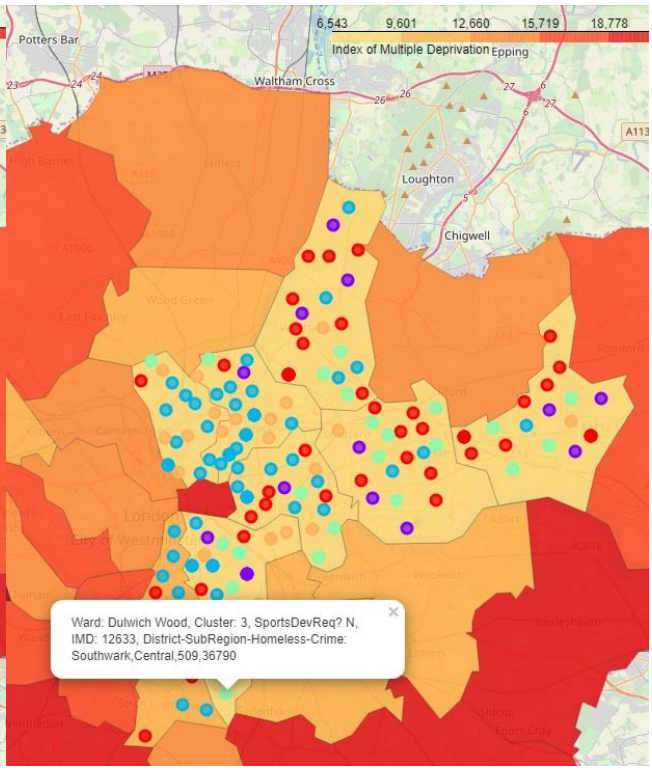


Other Clusters 0, 1, 2 and 3 do not require Sports deve









5. Discussion

I first found out the most deprived districts of London during Exploratory data analysis and created a choropleth map to display the London districts based on the Index of Multiple Deprivation values. The Darker areas are the prosperous districts while the light shaded ones are the deprived districts that we wanted to explore more to identify the most deprived neighbourhoods for development in London.

I chose Sport development as the development type because of the following reasons.

> Sports play a great role in family connections and social activities. It can help to motivate teenagers stay away from gang culture, increase employment opportunities, team work, and improve health. A lot of researches have shown that sport is a precious way to maintain physical health and has a close relationship with mental health and especially prevents mental disorders, anxiety, emotional and mental breakdowns and other disorders

By using Kmeans algorithm and Elbow method I found the optimal k value of 5 for the main data IMD, and by using Foursquare data I then identified the top 20 venues (mostly Sports category) for each Ward in the deprived London districts. I created a new feature called SportsDevRequired?Y/N and set a criteria for Sports development if there are no spots categories in the top 12 venues. After examining the clusters, I managed to assign wards that require Sports development to a dedicated Cluster (without affecting other Clusters). By doing do so, I have assigned a dedicated Cluster just for the wards that require Sports development, which can be easily identifiable in the map.

I created a choropleth map with clusters for deprived London districts with the below information for each Ward:

1. Ward name
2. Cluster Label
3. IMD value
4. District
5. Sub-Region
6. Homeless
7. Crime

Though IMD is our main factor for identifying the deprived areas for development, the additional information, namely, Homelessness and Crime are very useful when prioritising the development work. For example, there are Wards in deprived districts, Hackney and Islington where Homelessness and Crime are high as well so they need urgent attention

This study can be extended to include other development types (ex: Education) and a user-friendly Web portal can be built to assist the users

6. Conclusion

As stated in the business problem section, the results and recommendation of the projects will be useful for Charities, NGOs and local authorities to identify most deprived neighbourhoods and initiate, organise and carry out appropriate development projects for the affected areas that need urgent attention.

7. Appendix

7.1 Evidence of how sports can change lives for better.

Stephen Addison, a former London gang member who runs boxing classes to draw young people away from crime, has been recognised in the 2018 New Year Honours list. Since 2013 Stephen has helped more than 4,000 young people change their lives.

https://www.bbc.co.uk/news/education-46702775
https://www.ilfordrecorder.co.uk/news/crime-court/teens-encouraged-to-swap-knives-for-boxing-1-6093881