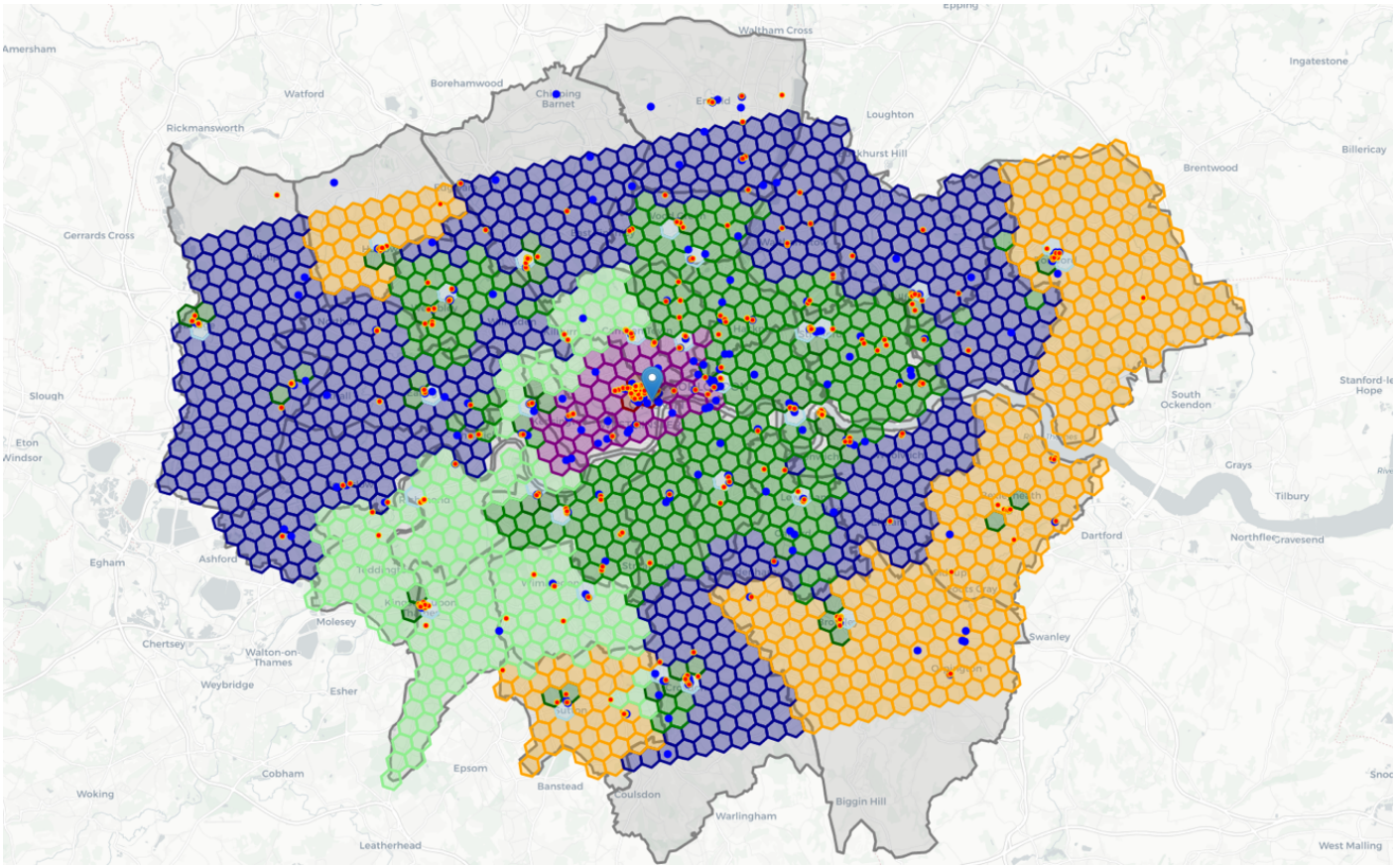

Figure 1 - Map showing clusters of areas to open a clothing store

The Battle of the Neighbourhoods – Opening a Clothing Store in London, UK

IBM Applied Data Science - Capstone Project

By: Sakib Chughtai

Date: 2/12/2020

1 Introduction

1.1 Background

With over 8.5 million inhabitants, London is one of the largest and most influential cities in the world. Often regarded as one of the world's most connected cities, London has a diverse range of cultures with over 250 languages spoken in the region. Each year, the city attracts over 21 million international visitors as well as 28 million domestic tourists.

London is home to over 40,000 shops and 26 major street markets. Oxford Street is known as London's busiest shopping street. Located in the heart of the city's West End, Oxford Street offers a wide variety of clothing stores from high end luxury brands such as Louis Vuitton, Burberry and Gucci to high street fashion brands such as Zara, Tommy Hilfiger and Polo Ralph Lauren.
There are many other types of clothing stores located in London to suit the different needs of customers, such as The North Face and Patagonia which sell mountain and winter clothing. There are also cheaper options for people looking to buy clothes on a budget, such as Primark and H&M. All in all, there is a clothing store for anyone and everyone.

1.2 Business Understanding

The goal of this project is to predict the optimal location to open a clothing store in London, UK. This report will provide stakeholders with an insight into which areas are the most promising based on a multitude of factors. The key stakeholder in this project are investors and owners looking to start or expand their clothing company into London. The reason why this project would be important to them is because they want to find the best area to open a store, and the analysis will allow them to make an informed decision backed by real data. Other stakeholders could be data enthusiasts and engineers looking to add more features to the model or perform the same analysis on another city such as Paris or New York.

Tackling this problem head on requires breaking down the end goal into objectives that are in support of the goal. By breaking down the objectives, structured discussions can take place where priorities can be identified in a way that can lead to organizing and planning how to tackle the problem at hand.

There are thousands of different clothing stores in London. The first objective is to locate all of these stores using the Foursquare API. We will also look at all shopping malls because this may have an impact on where the ideal location to open a store will be. The next step consists of using demographic data such as population density, average age and income to cluster similar areas in London. The optimal location to open a store will be in an area where there is a high concentration of clothing stores AND shopping malls. Ideally, this area should also have a high population density and average income.

1.3 Analytics Approach

Selecting the right analytic approach is key to ensuring we answer the question being asked correctly and accurately. Given that we are trying to show relationships between different features within the model, we will use a descriptive model. We will use K-means clustering which is a simple and popular unsupervised machine learning algorithm to answer our question. A cluster is defined as a collection of data points that are aggregated together because of certain similarities. This will be the best technique to use for our use case.

We will use several data science tools to achieve the end goal. First, We will use K-means clustering to answer our question. We will then use machine learning and visualization tools to generate the most promising areas based upon criteria we defined earlier.

Additionally, we will evaluate the quality of the model by through a diagnostic measures phase and statistical significance testing. These evaluation techniques ensure the model is working as intended and that the data is being properly handed and interpreted. Lastly, we will present our findings to the stakeholders so they can make an informed decision on where they would like to open their clothing store

2 Data

2.1 Data Sources

If the problem that needs to be solved is a ‘recipe’, the data is an ‘ingredient’, so to speak. As a data scientist, there are several aspects of the data that need to be addressed before moving onto the data preparation stage.

We need to identify what datasets are needed, how to source or collect them, how to understand or work with them and how to prepare the data to meet the desired outcome. Based upon the criteria we defined earlier, there are several factors that will impact the final decision of which area to open a clothing store. These are:

- The number of clothing stores within the neighbourhood
- The number of shopping malls within the neighbourhood
- London census information – population, density, average age, average income

We will define each neighbourhood using a grid format around the whole city. This grid will consist of many hexagons in a honeycomb layout. This will be the basis for defining different boroughs and neighbourhoods.
We will collect the data from the following sources:

- London borough boundaries shapefile which is defined using longitude and latitude values. Will be in geojson format to be easily integrated into analysis. Can be obtained from the London Datastore (publicly available).
- Coordinates which define the centre of London, this will be obtained using the Google Geocoding API.
- Clothing stores and shopping mall data in every borough will be obtained using Foursquare API.
- The centre of every borough will be defined using hexagon cells in a honeycomb layout, these will be calculated algorithmically. The addresses for those centres will be fetched by using the Google Geocoding API.
- London census data for the year 2016 will be obtained from the London Datastore (publicly available).

2.2 Data Preparation

Before we make use of the Foursquare API and census data for London, we need to implement a process to uniquely identify each neighbourhood on the map. For this job we will use a honeycomb hexagon grid which will span across the whole London map. This grid format is a popular choice when performing any geospatial analysis. The reason we are using hexagons instead of circles is because there is no spacing amongst hexagons when they are joined together in a grid. Each cell on the grid represents a candidate neighbourhood. The length of each side is $1000/\sqrt{3}$ meters, and covers an area of approximately 0.87 square kilometres. As we are only performing an analysis on London, we will only use the honeycomb grid on areas within London. Before we instantiate any honeycomb hexagon grid, we need to define the coordinates of central London as a starting point for our exploration. We will find the latitude and longitude of central London using the Google Geocoding API. The API provides geocoding and reverse geocoding of

addresses. Geocoding is the process of converting addresses into latitude and longitude coordinates, we can use these coordinates to place markers on the map or explore specific areas. Below we can see the output of geocoding and reverse geocoding the center of London. We perform reverse geocoding to make sure everything is working as it should.

```
Coordinate of Trafalgar Square, Charing Cross, London, England: [51.508039, -0.128069]

Reverse geocoding check
-----------------------
Address of [51.508039, -0.128069] is: 5 Trafalgar Square, Charing Cross, London WC2N 5NJ, UK
```

The next step involves defining the boundary of London as we are only focusing on analysis all areas within London. We will fetch the London boundary shape file which is publicly available from the London Datastore. The file, in geojson format, defines the boundaries of each borough. When we create the hexagon cell grid spanning across London, we will limit the grid within the boundaries of London. To make sure the shape file contains all the necessary data, we will perform some checks. These include checking the boundaries of London, and whether the coordinate we defined earlier is in London.

```
The boundaries of London: (-0.510375069478728, 51.2867601558474, 0.33401556346767, 51.6918
74109516)

Found in borough:  {'id': 25, 'name': 'Westminster', 'code': 'E09000033', 'area_hectares':
2203.005, 'inner_statistical': True}
London center (-0.128069, 51.508039) in geojson file: True
```

As we can see, all the necessary data is present. However, we notice that latitude and longitude coordinates are in a different format than what we need. The geojson file uses the WGS84 spherical coordinate system but we need to convert the coordinates into a common metric unit, which is the UTM Cartesian coordinate system (X/Y coordinates in meters). We will define several methods to convert these latitude and longitude values into meters and vice versa.
The next step involves creating the honeycomb grid of cells with hexagons. The centers of each hexagon are equally distant in one row, and every other row will be offset. Additionally, the center of each cell is the same distance from all surrounding cell centers. This ensures that we cover all areas in London without having to worry about spacing between each hexagon cell.

As mentioned earlier in the report, each side of a hexagon cell will be of $1000/\sqrt{3}$ meters in length. We can calculate the distance between the centers of two hexagons within the same row, which is 1000 meters. When we create the hexagon grid, we limit where the grid extends by referring the boundary shape file we imported earlier. As a general rule, if the hexagon cell falls within the boundary of a borough, we can create the cell. We found 1596 candidate hexagon cells and visualised this on a map with the boundary data. The boundaries are defined in grey, whilst the hexagon cells are in blue.

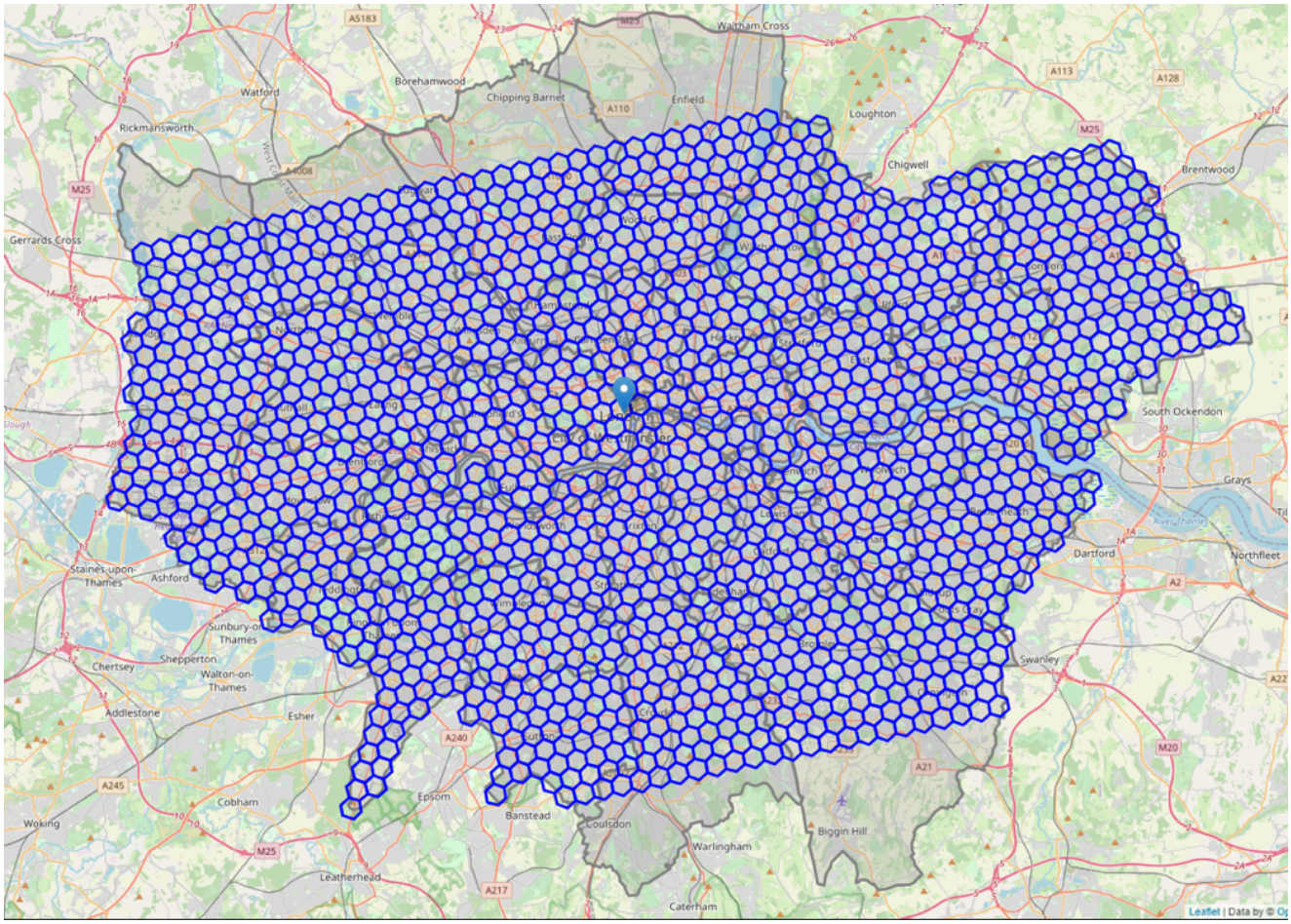

Figure 2 - Map showing hexagon cell grid covering London

Before we created the honeycomb grid covering the majority London, we had to define the coordinates of the center of each hexagon cell. The last important step of the data preparation stage is geocoding these coordinates and putting the data for all the generated locations into a Pandas Dataframe. We will retrieve the addresses of the coordinates we defined earlier by using reverse geocoding through the Google Geocoding API. The table below displays the first 10 rows of the locations Dataframe.

Table 1 - Addresses of the center of each hexagon cell

	Address	Latitude	Longitude	X	Y	Distance from central
0	Rose Cottage, Carshalton Rd, Banstead SM7 3JA	51.332255	-0.171414	-554116.768333	5.796889e+06	20048.954645
1	73 Lower Pillory Down, Little Woodcote, Little...	51.334091	-0.157560	-553116.768333	5.796889e+06	19717.417166
2	6 Verulam Ave, Purley CR8 3NQ	51.335926	-0.143705	-552116.768333	5.796889e+06	19431.739424
3	2 Silver Ln, Purley CR8 3HG	51.337760	-0.129848	-551116.768333	5.796889e+06	19193.969219
4	926 Brighton Rd, Purley CR8 2LN	51.339592	-0.115990	-550116.768333	5.796889e+06	19005.904655
5	5 Lexington Ct, Purley CR8 1JA	51.341422	-0.102130	-549116.768333	5.796889e+06	18869.032012
6	Old Court, Hook Hill, South Croydon CR2 0LA	51.343251	-0.088269	-548116.768333	5.796889e+06	18784.470352
7	149 Ridge Langley, South Croydon CR2 0AQ	51.345078	-0.074406	-547116.768333	5.796889e+06	18752.927339
8	47 Foxearth Rd, South Croydon CR2 8EL	51.346903	-0.060542	-546116.768333	5.796889e+06	18774.670200
9	33 Heathfield Vale, South Croydon CR2 8AG	51.348727	-0.046676	-545116.768333	5.796889e+06	18849.514541

3 Exploratory Data Analysis

3.1 Foursquare API

Now that we have generated all the candidate neighbourhoods and the honeycomb hexagon grid, we can now retrieve venue information using the Foursquare API. The Foursquare API allows application developers to interact with the Foursquare platform. We will be using the Places API in particular, which returns information about places using HTTP requests. We can look up information for many different types of venues. To make the search easier for users they are grouped into categories such as Arts & Entertainment, Food, Nightlife Spot, Outdoors & Recreation, Shop & Service and others.

We will be using the Foursquare API to find all Clothing Stores within London. This information is found under the Shops venue category. The corresponding ID for clothing stores is: **4bf58dd8d48988d103951735.**

We will use folium to visualize all clothing stores within London on a map.

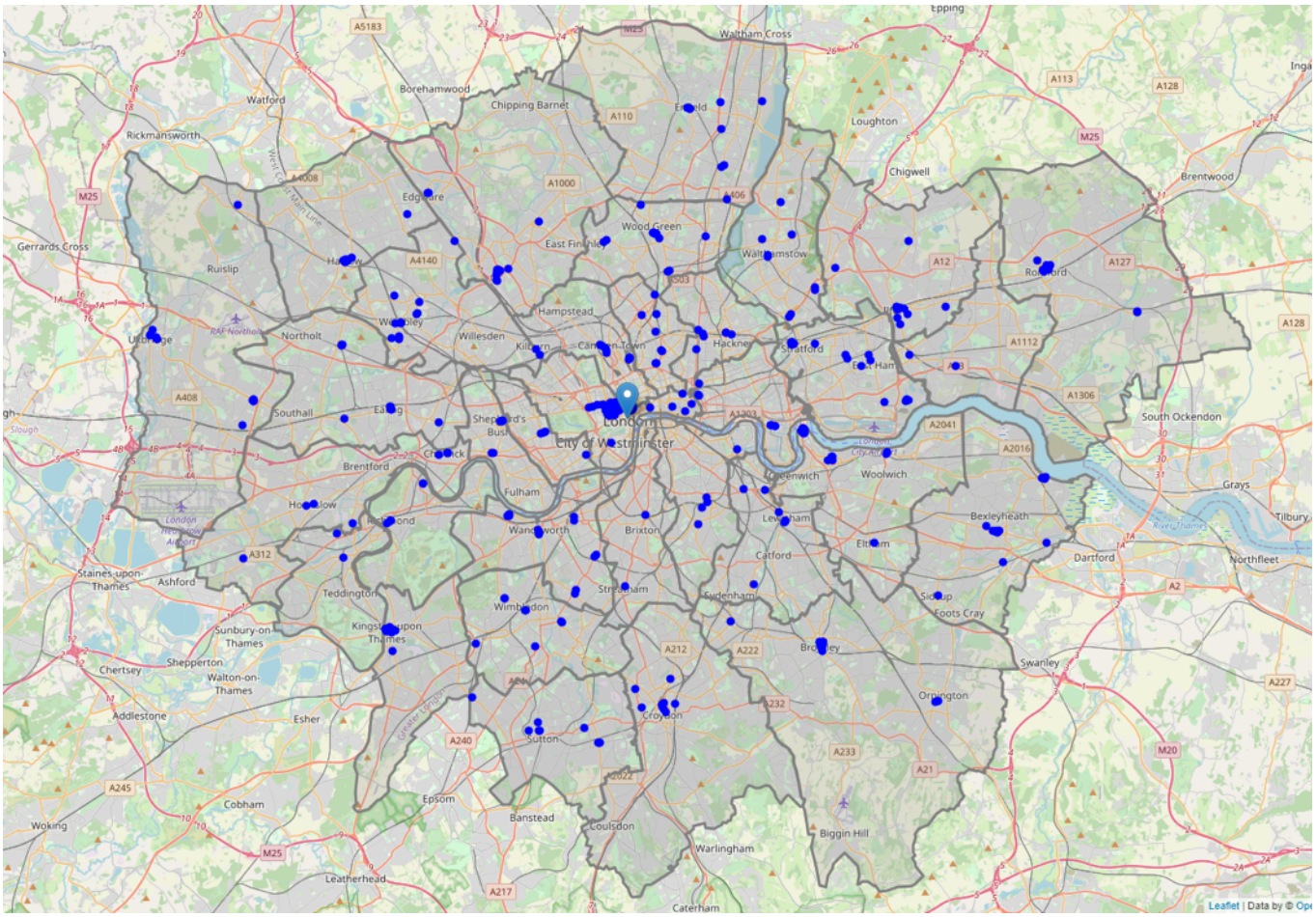

Figure 3 - Map showing all clothing stores in London(blue)

We now know where all clothing stores within London are located, but we don't know how many there are within a specific area. We will generate a heatmap to display which areas of London have a lower or higher concentration of clothing stores.

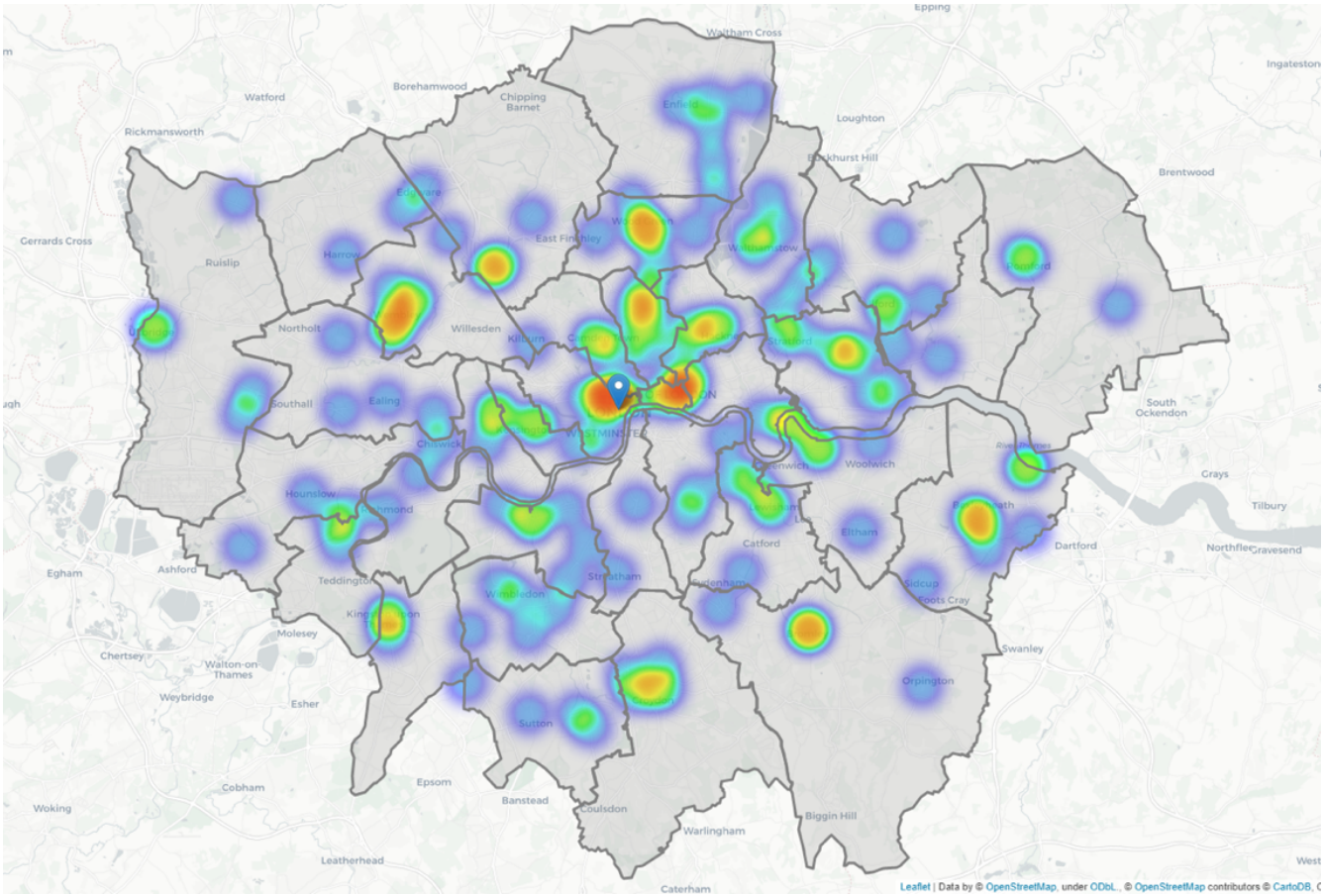

Figure 4 - Heatmap showing where clothing stores are concentrated within the city

Upon first glance, the heatmap shows that clothing stores are mostly spread out across the city, with the highest concentration in the city centre. The second highest concentration of clothing stores is located north of central London, in areas such as Camden and Islington.

The presence of shopping malls could explain why there are clusters of shops in certain neighbourhoods located far away from the city centre. We will explore all the shopping mall data in London using the Foursquare API. There are also several categories that are related to clothing stores that will be explored, such as Vintage and Skate shops.

Relevant Categories and ID's, respectively:

- **Shopping Mall:** 4bf58dd8d48988d1fd941735
- **Shopping Plaza:** 5744ccdfe4b0c0459246b4dc
- **Outlet Mall:** 5744ccdfe4b0c0459246b4df
- **Outlet Store:** 52f2ab2ebcbc57f1066b8b35

Let’s visualize all shopping malls on one map using folium.

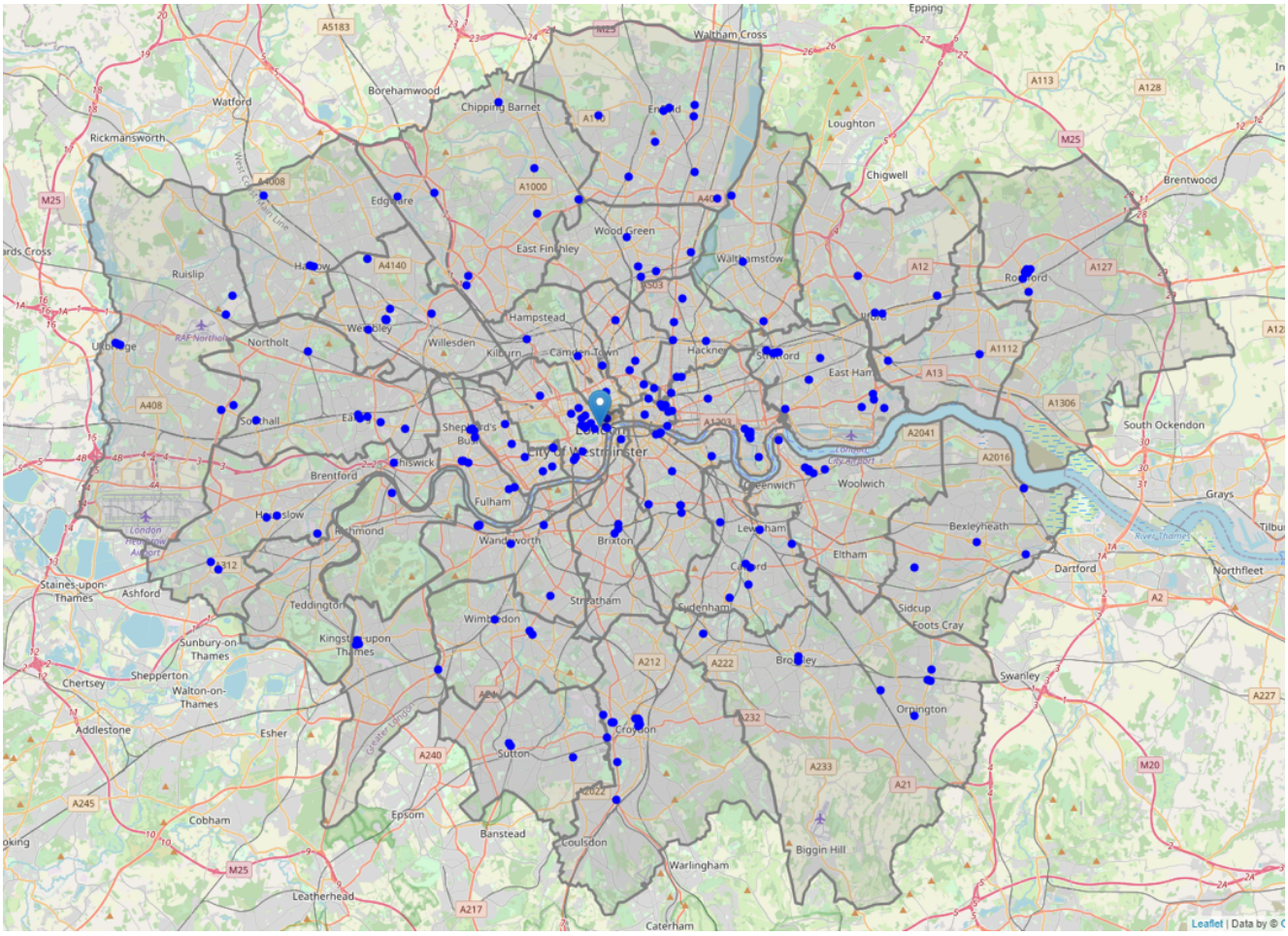

Figure 5 - Map showing where all shopping malls are located in London

The next step in the exploration of London is to visualize shopping malls and clothing stores in one map. Where shopping malls are marked in blue, and clothing stores in red.

From the map above, we can see that clothing stores in the center of London are mostly located in or near shopping malls. There is a higher concentration of shopping malls in the suburbs compared to clothing stores, which is expected. We can also see that there are several clusters of shops in central London. This shows that many clothing store owners prefer to open stores in areas that are highly built up, thus leading to higher customer counts.

In an ideal situation, opening a clothing store near a shopping mall and other clothing stores is preferred. This will maximize the potential customer visits to the physical store, leading to higher chances of clothes being purchased, and thus yields higher profits over time.

3.2 London Census Data

The next major step in predicting the ideal area to open a clothing store is looking at the census information for each borough in London. The census data will be retrieved from the London Datastore, which is a free and open data-sharing portal where anyone can access data relating to the capital. (Can be found at:https://data.london.gov.uk/). The file is in csv format, and is therefore easily integrated into the notebook using Pandas. We read the csv into a Pandas Dataframe and remove all unnecessary columns. We are only focusing on several basic census variables: Population, Area, Density, Average Age, and Average income.

Table 2 - Census information (uncleaned)

	GLA Population Estimate 2016	Inland Area (Hectares)	Population density (per hectare) 2016	Average Age, 2016	Modelled Household median income estimates 2012/13
Area name					
City of London	8,548	290.4	28.9	42.9	£99,390
Barking and Dagenham	205,773	3,610.80	57.3	32.9	£34,080
Barnet	385,108	8,674.80	44.5	37.2	£54,530
Bexley	243,303	6,058.10	39.9	38.9	£44,430
Brent	328,568	4,323.30	76.1	35.5	£39,630

Before we use this dataset in our analysis, we need to change the column names into more simple terms. Additionally, we need to remove all commas and currency symbols as machine learning algorithms only work with numbers. The last step requires casting all numbers as a float, as they are of type object in their initial state. The census dataset is cleaned and ready for analysis.

Table 3 - Census information (cleaned)

Borough	Population	Area	Density	Average Age	Average Income
City of London	8548	290.4	28.9	42.9	99390
Barking and Dagenham	205773	3610.8	57.3	32.9	34080
Barnet	385108	8674.8	44.5	37.2	54530
Bexley	243303	6058.1	39.9	38.9	44430
Brent	328568	4323.3	76.1	35.5	39630
Bromley	326560	15013.5	21.7	40.1	55140
Camden	240595	2178.9	109.8	36.2	67990
Croydon	383408	8650.4	44.4	36.9	45120

We will now explore the census data to gain a better understanding about every borough and study the relationships each variable may have with the amount of clothing stores located within a specific area or borough.

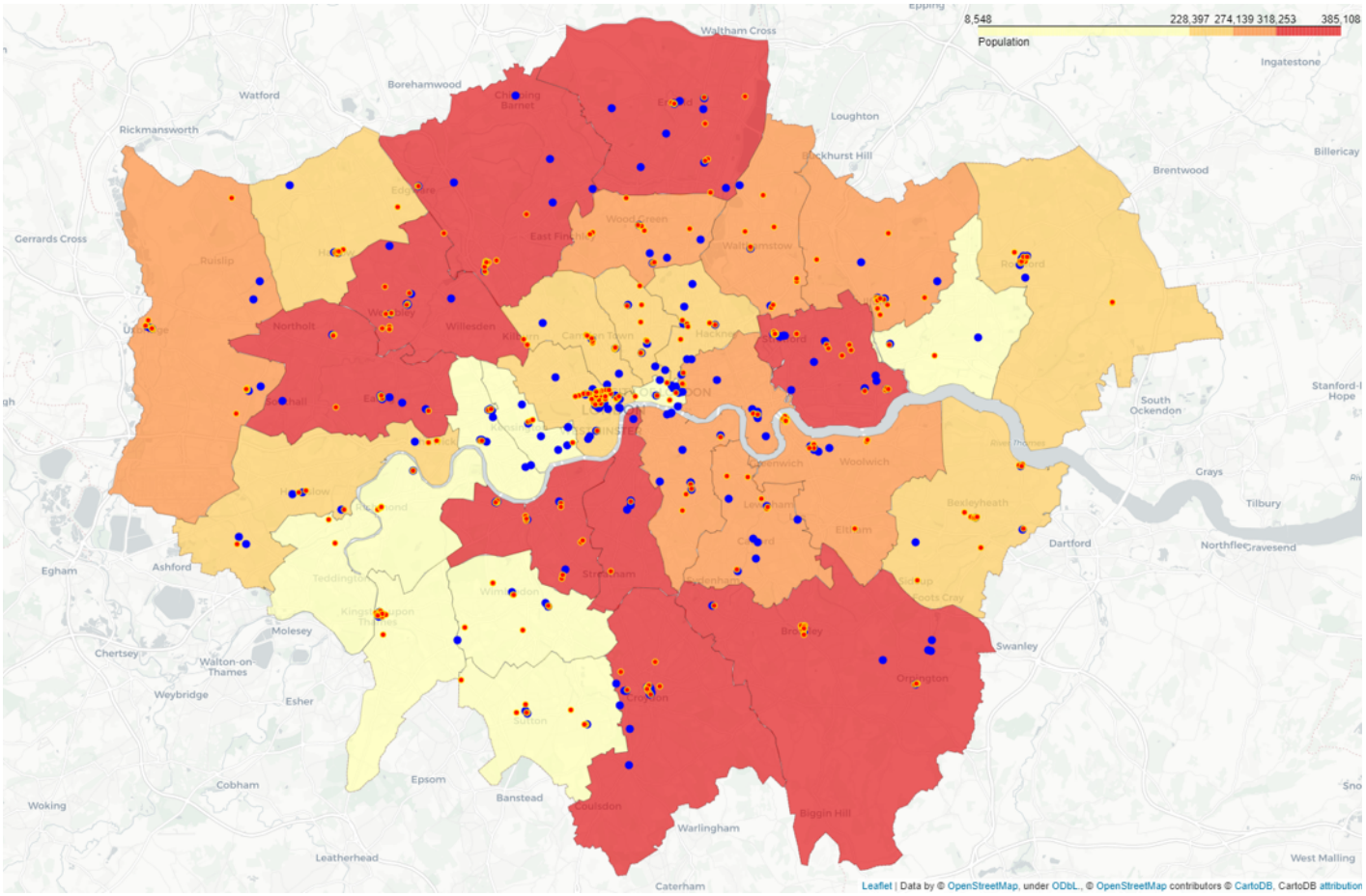

Figure 6 - Map showing population distribution of London

The choropleth map above highlights which boroughs of London have the highest and lowest population. Boroughs such as Ealing, Barnet, Brent, Croydon and Enfield all have a population above 318,000, and are marked in red.
Meanwhile on the lower end of the spectrum, boroughs Kensington and Chelsea, Hammersmith and Fulham, Kingston upon Thames and Richmond upon Thames all have populations below 228,000. Upon first glance, there doesn't seem to be a correlation between population and number of clothing stores or shopping malls within a borough.

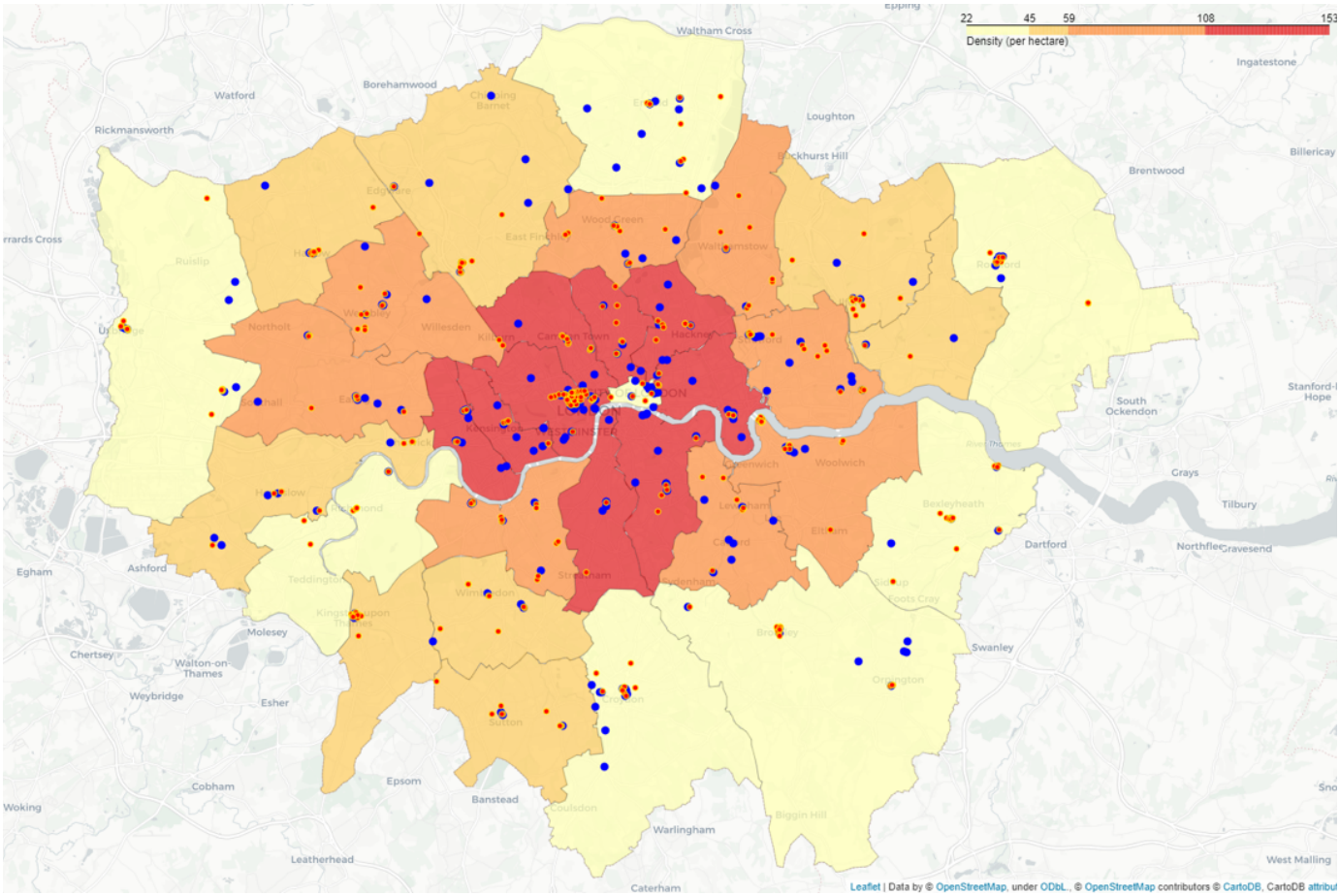

Figure 7 - Map showing density distribution

This one is interesting. We can clearly see that boroughs surrounding the center of London have the highest population density. In particular, boroughs such as Westminster, Camden, Hackney, Kensington and Islington all have a density above 108 people (per hectare). Meanwhile boroughs such as Hillingdon, Richmond, Havering and Bromley have a density below 45 people (per hectare).

More importantly, we can see that there is a correlation between density and the number of clothing stores and shopping malls. There is a higher population density in areas where there are a high concentration of clothing stores and malls. This is especially true for boroughs such as Westminster and Camden.

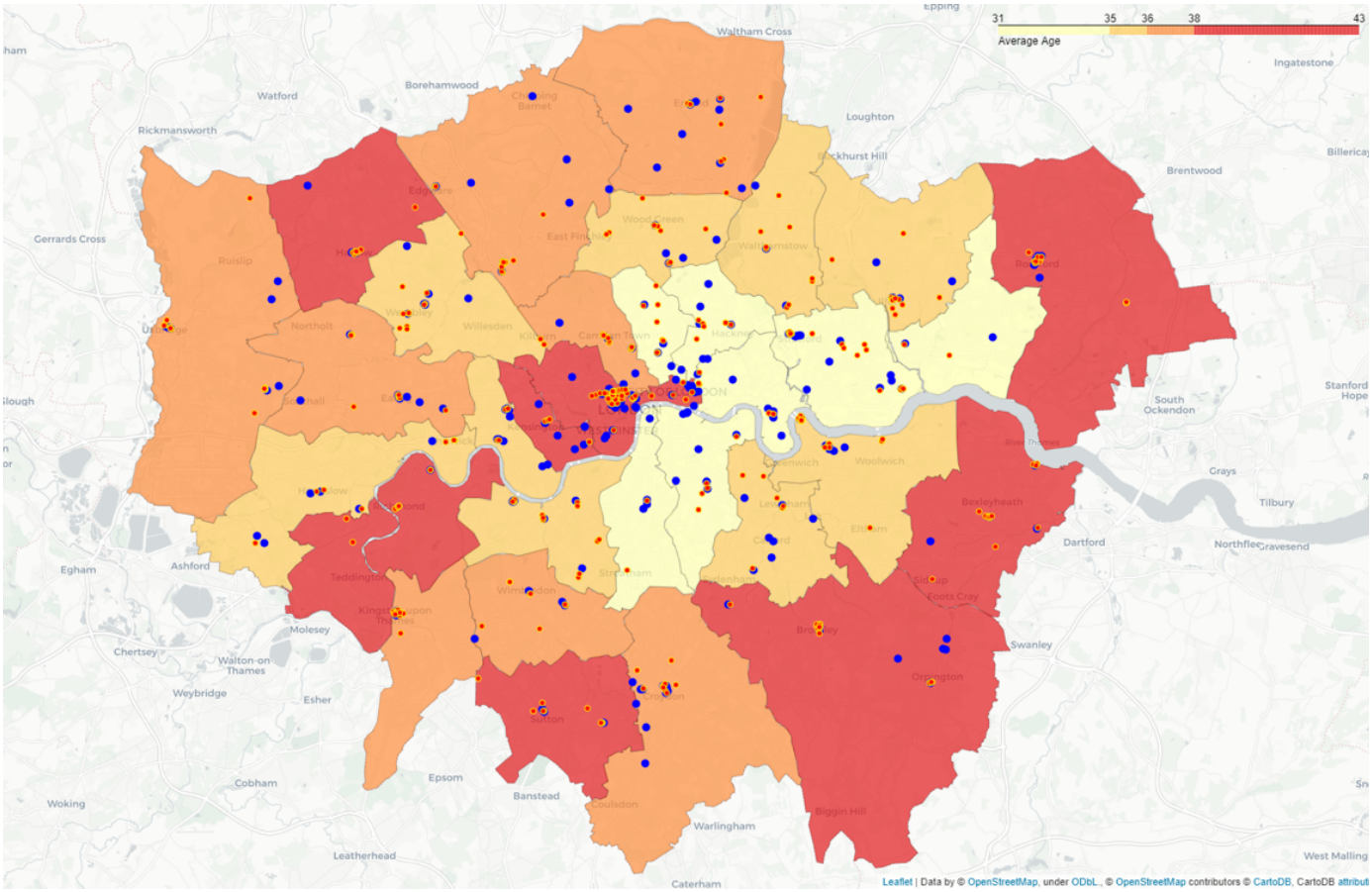

Figure 8 - Map showing age distribution

The map above shows the average age distribution in London. Upon first glance, we can see that boroughs in the east tend to have a lower average age (below 35). Meanwhile boroughs in the eastern suburbs have a higher average age (above 38). Many boroughs that lie in the suburbs tend to have a higher average age than those near central London. This is the case for boroughs such as Barnet, Enfield, Sutton and Bromley. In boroughs Westminster and Kensington (located in central London), we can see the average age is above 38. There doesn't seem to be a correlation between average age and the number of clothing stores/shopping malls in the borough.

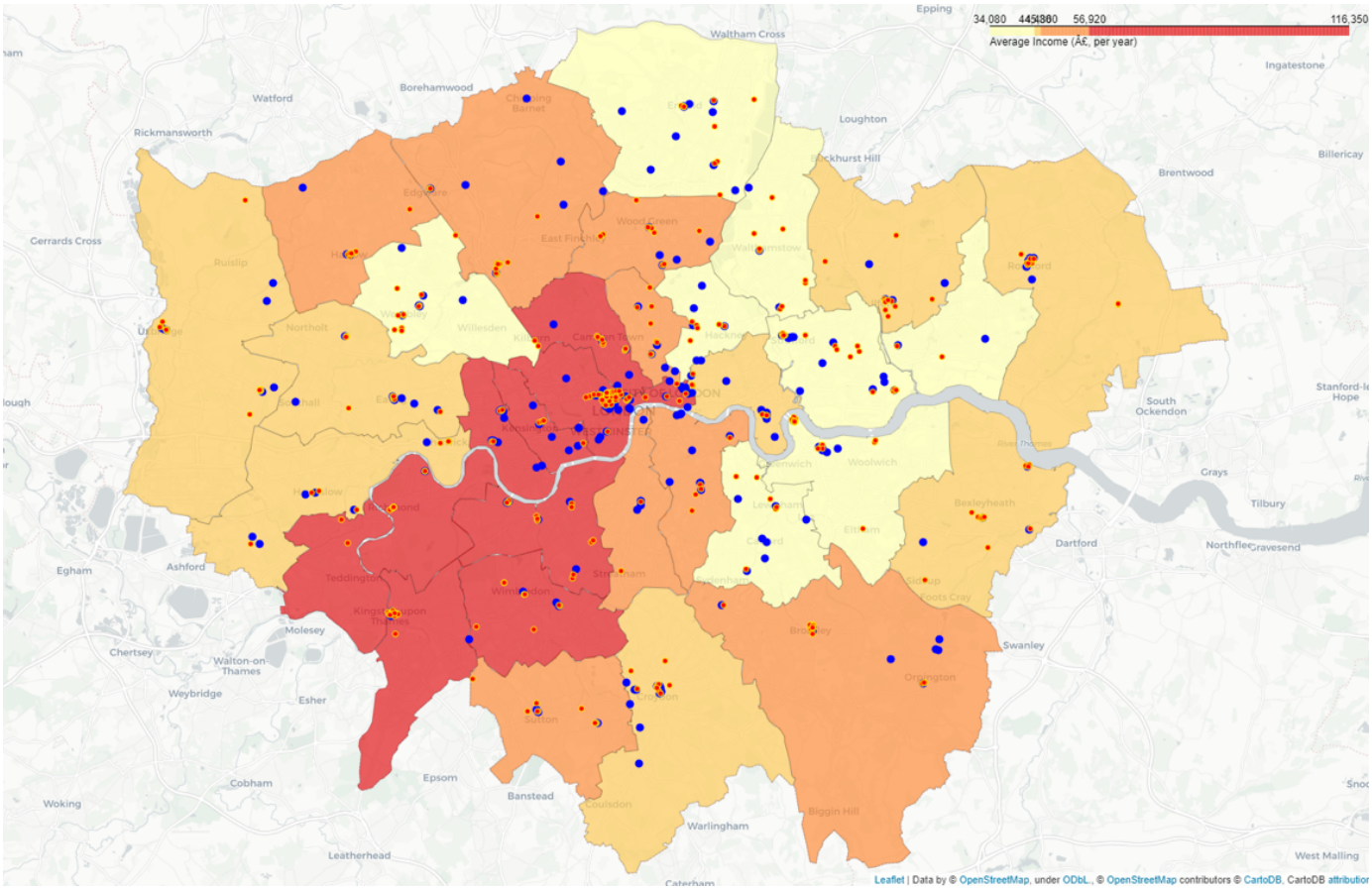

Figure 9 - Map showing income distribution

The choropleth map above shows the distribution of income in London. We can see that boroughs in central London and south-west London have the highest income on average. Westminster, Camden, Kensington and Hammersmith all have an average income above £56,000.

Boroughs in the east such as Lewisham, Hackney, Greenwich and Newham are amongst the lowest in terms of average income, with people earning less than £44,000. As for the rest of the city, most boroughs fall between £44,000 and £56,000. Boroughs located in the suburbs mostly have a moderate income compared to those living in the city center and west London.

4 Predictive Modelling

The goal of this project is to predict the optimal area to open a clothing store in London.

We have gathered, wrangled and cleaned the following data:

- All Clothing Stores data in London
- All Shopping Malls and Centres data in London
- London census data retrieved for every borough in the year 2016. We only picked several attributes, which were: Population, Area, Density, Age and Income.
- Boundary data for each borough in London, which was in geoJson format.

We have also created a honeycomb hexagon grid with each side being $1000/\sqrt{3}$ meters in length. This grid covers majority of the London land area. By processing all of the data above, we will create several new features that will be useful in the modelling stage of the project. These features will consist of census information for each candidate area (cell), and the number of clothing stores and shopping malls in the nearby area (within 1/2/3 km). Lastly, we will recommend the area with the most shops AND shopping malls nearby. This is because we want to maximize the potential amount of customers visiting the store.

4.1 Generating new features

There are multiple variables that impact our final prediction of which area looks most promising for stakeholders. These are:

- Number of clothing stores within the neighbourhood
- Number of shopping malls within the neighbourhood
- Census information – Population, density, average age, average income

Now that we have the census information of every London borough, we have to calculate the census data for each candidate hexagon cell (area) accordingly. We will do this by looking at which hexagon cells intersect with the corresponding borough. Once this is completed, we can easily generate census data for each individual hexagon cell and perform further analysis to derive deeper insights. As a general rule, if multiple hexagon cells are located entirely within a certain borough, the cells will all contain the same census information. However, if a hexagon cell is located between two boroughs, the census data will be calculated as an average between the two boroughs, respectively. We can now generate the census data for all candidate hexagon cells.

After this has been completed, we will merge the Dataframe we are currently working on with the previous Dataframe containing all location data. The new Dataframe contains information about each hexagon cell. Let us take a look at it.

Table 4 - Census information for each hexagon cell

Address	Latitude	Longitude	X	Y	Distance from central	Population	Density	Average Age	Average Income	Boundary
39 Albury Cl, Hampton TW12 3BB	51.423656	-0.369217	-565616.768333	5.809880e+06	19464.521435	196602.000000	34.300000	38.700000	76610.000000	POLYGON ((-0.3709380329785318 51.4286619235410...
The Pavilion Bushy Park, Cricket Ln, Hampton H...	51.425519	-0.355349	-564616.768333	5.809880e+06	18511.713914	196602.000000	34.300000	38.700000	76610.000000	POLYGON ((-0.3570681880841556 51.4305246464424...
Constitutional House, 5A Stanley Rd, Teddingto...	51.427380	-0.341479	-563616.768333	5.809880e+06	17564.154104	196602.000000	34.300000	38.700000	76610.000000	POLYGON ((-0.3431968561874282 51.4323857756264...
31 Manor Rd, Teddington TW11 8AA	51.429239	-0.327607	-562616.768333	5.809880e+06	16622.739447	196602.000000	34.300000	38.700000	76610.000000	POLYGON ((-0.3293240383506804 51.4342453106815...
7 Vancouver Rd, Richmond TW10 7YA	51.431096	-0.313734	-561616.768333	5.809880e+06	15688.576228	191305.275243	37.094000	38.304183	72025.512310	POLYGON ((-0.3154497356371143 51.4361032511968...

Now that the census data for each hexagon cell has been calculated, we can now move onto calculating the number of shops and shopping malls within each cell. In particular, we will generate 3 new features:

- The number of clothing stores and shopping malls within the selected hexagon cell.
- The number of clothing stores and shopping malls within 1 km of the center of the hexagon cell.
- The number of clothing stores and shopping malls within 3 km of the center of the hexagon cell

After this has been calculated and saved into a Pandas Dataframe, we can finally start using machine learning techniques to group similar areas into clusters.

4.2 Machine Learning – K-Means Clustering

K-means clustering is one of the most popular unsupervised machine learning algorithms. The algorithm groups data points together based on similar characteristics. These groups are referred to as clusters. Each and every data point is allocated to a cluster. At the start of the process the algorithm uses randomly selected centroids. The positions of each centroid are optimized by performing iterative calculations.

Table 5 - Clothing store/mall information for each hexagon cell

	Population	Density	Average Age	Average Income	Shops in cell	Shops within 1km	Shops within 3km	Malls in cell	Malls within 1km	Malls within 3km
500	196602.000000	34.300000	38.700000	76610.000000	0	0	0	0	0	0
501	196602.000000	34.300000	38.700000	76610.000000	0	0	1	0	0	0
502	196602.000000	34.300000	38.700000	76610.000000	0	0	1	0	0	0
503	196602.000000	34.300000	38.700000	76610.000000	0	0	14	0	0	4
504	191305.275243	37.094000	38.304183	72025.512310	0	0	14	0	0	4
505	193279.294637	36.052713	38.451699	73734.090351	0	0	13	0	0	4
506	196602.000000	34.300000	38.700000	76610.000000	0	0	7	0	0	1
507	196602.000000	34.300000	38.700000	76610.000000	0	0	0	0	0	0
508	200406.295264	39.095266	38.275280	73540.046943	0	0	0	0	0	0
509	275719.195601	72.996543	36.228624	69852.738142	0	0	2	0	0	0

We will be performing k-means clustering for each and every hexagon cell we defined earlier. The table above shows the inputs we will be using for machine learning. The process of selecting the best value of K is done by running an evaluation step. K is defined as the number of categories that the data can be split into.

The Sum of Squared Distance will be used to evaluate how accurate the data points and relevant clusters are. It measures the error between different data points and their allocated cluster centroids. We want to have the smallest value possible, which indicates the least error rate.

The Silhouette Score focuses on minimizing the sum of squared distance inside the cluster. It also maximizes the distance between its neighbours. The bigger value of the score, the better.

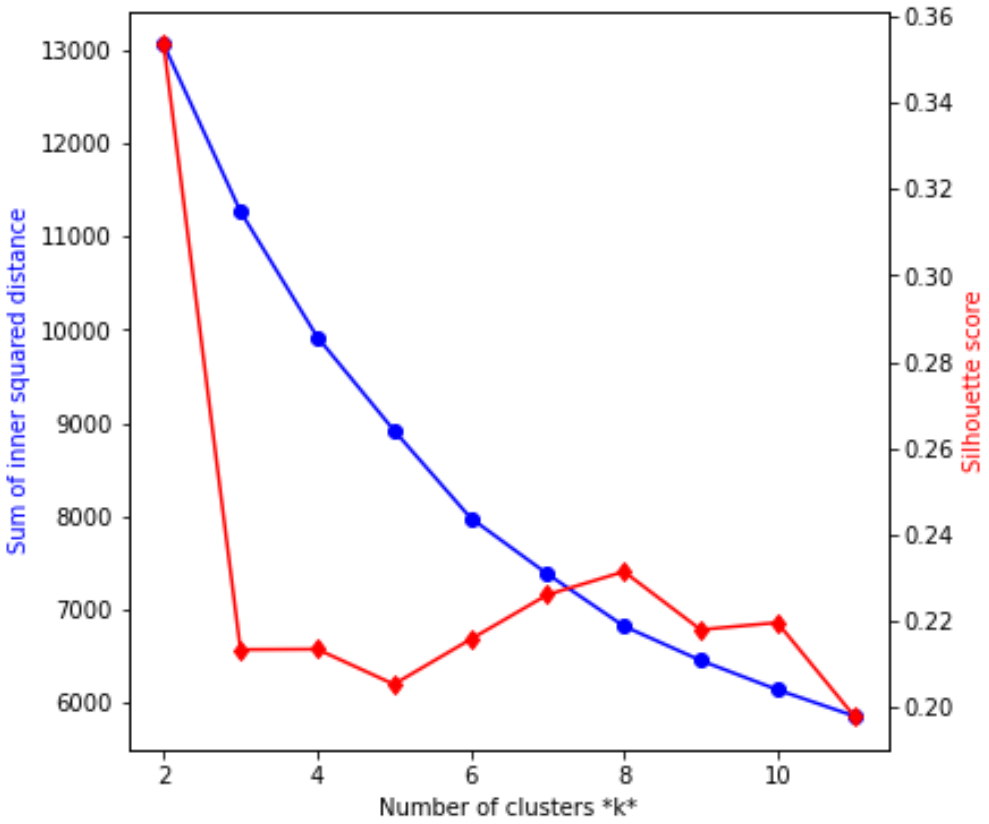

Figure 10 - Graph showing Sum of Squared Distance and Silhouette score as K increases

From the graph above, we can see that the Sum of Squared Distance (SSD) starts at a high value (13000) at K=2 and gradually decreases as K approaches 10.

The case for the Silhouette score is different, at K=2 the score starts off high (0.36) but drastically decreases to (0.21) at K=3. We then see a gradual increase in the score as K increases from 5. It then eventually peaks at K=8 with a score of approximately 0.24.

We want to choose the **lowest** sum of squared distance and the **highest** silhouette score. We will choose K=8 because it reflects a balanced number for both values.

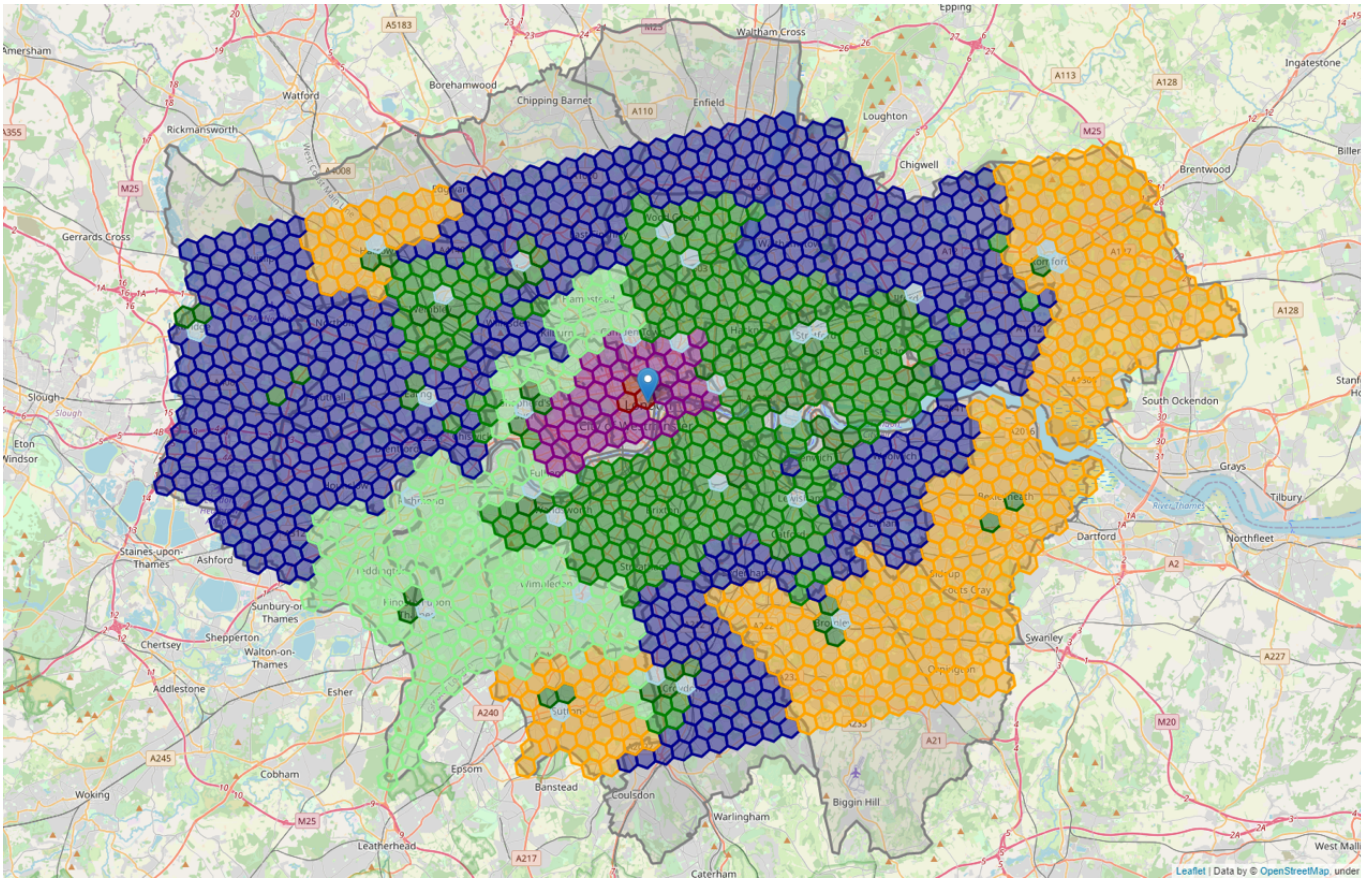

Figure 11 - Map showing clusters of areas to open a clothing store

Above we can see the final result of clustering each area of London. Boroughs located in the suburbs tend to be similar to their neighbours and this can be seen if we look at the boroughs in yellow and dark blue. These areas contain the least clothing stores and shopping malls and also have a lower average income. Moving onto the dark green cluster, this group of cells represent areas that have a moderate amount of clothing stores and shopping malls.
The clusters in red and purple represent the best possible areas to open a store. In particular, the red cluster containing 3 hexagon cells is the best possible area to open a store as there are an abundance of clothing stores and shopping malls and the population density and average income is very high. This is very promising for stakeholders looking for an area that returns the highest profits.

We will visualize all attributes on one map. These include shopping malls, marked as a blue point, and shops marked with a red point.

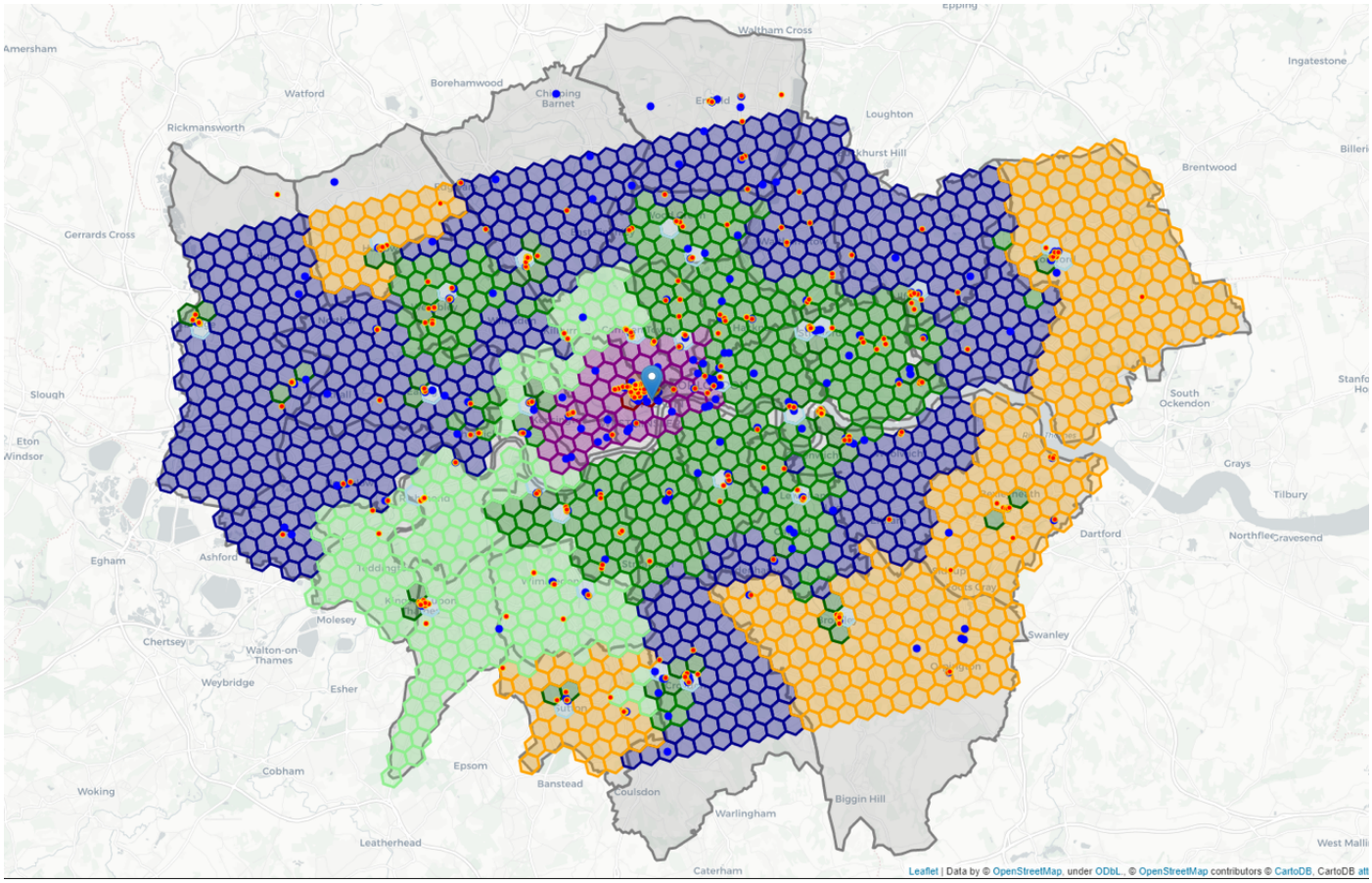

Figure 12 - Map showing clusters of areas to open a clothing store, where shops = red dot, malls = blue dot

From the cluster plot above including all of the clusters. We can see that there is a single cluster coloured in purple in central London. And there is a smaller cluster consisting of 3 hexagons in red, which is right in the center of the city. There are many clothing stores and shopping malls within both of these clusters, with the higher concentration of both within the red cluster.

As we move towards the outer areas of London, inevitably the amount of clothing stores and malls decrease. This is reflected by the green, blue and orange clusters which consist of many hexagons. This shows that these areas are mostly the same in nature.

4.3 Finding the best cluster

We will now combine all features related to clothing stores into one, and the same process for shopping malls as well. This allows us to more easily sort attributes. The clothing store score and shopping mall score will both be weighted, and then we will generate a final score which consists of both scores within one.

The higher the score, the larger amount of clothing stores and shopping malls within the area.

Table 6 - Scores for each cluster based upon stores/malls

Population	Density	Average Age	Average Income	Distance from central	Malls in cell	Malls within 1km	Malls within 3km	Shops in cell	Shops within 1km	Shops within 3km	Mall Score	Shop Score
239905.668501	110.364255	37.516595	79999.226792	917.894687	3.666667	7.000000	14.333333	10.000000	18.666667	15.000000	53.666667	121.000000
280924.065139	73.215516	36.067935	51153.733308	12025.755465	2.535714	0.714286	3.892857	4.821429	1.321429	5.392857	18.714286	33.464286
293783.432433	65.833071	36.190538	50194.282714	13176.414471	0.294118	2.901961	2.921569	0.450980	5.764706	4.549020	13.098039	24.098039
207824.025824	112.869691	37.323963	80442.703830	3209.241863	0.541667	1.562500	18.625000	0.395833	1.750000	28.291667	26.020833	35.520833
203956.511711	57.049161	37.348280	65556.229921	13400.482483	0.020101	0.120603	2.809045	0.075377	0.311558	6.356784	3.271357	7.668342
302211.166270	100.914130	34.105389	46286.704085	9032.289359	0.121107	0.384083	7.000000	0.159170	0.536332	9.937716	8.757785	12.342561
267222.139313	30.597509	39.548482	48272.560882	20896.037138	0.031579	0.094737	1.623684	0.063158	0.152632	3.592105	2.065789	4.365789
308998.590938	50.955520	35.854950	44636.905254	16313.188472	0.051839	0.122074	2.239130	0.068562	0.178930	3.707358	2.864548	4.586957

The table above shows the calculated shop and mall score based upon each feature (census and venue data). The clusters are ordered by highest score to lowest score.
The final score is calculated by subtracting the mall score from the shop score.

Table 7 - Final scores for each cluster

Cluster	Final Score
4	67.333333
7	14.750000
6	11.000000
5	9.500000
0	4.396985
1	3.584775
3	2.300000
2	1.722408

As shown above, we can clearly see that cluster 4 is the most promising cluster among the other 7. With a score of 67.3, it has the most clothing stores and shopping malls in the area. It also has one of the highest average income amongst all other clusters.
The second highest scoring cluster has a value of 14.7, which is significantly lower than its higher scoring counterpart. Several clusters have a score lower than 10, we will not be recommending areas in these clusters to stakeholders as they are not the optimal areas to open a store.
Let's have a closer look at our highest scoring cluster, number 4.

Table 8 - Information about target cluster (4)

	Distance from central	Population	Density	Average Age	Average Income	Shops in cell	Shops within 1km	Shops within 3km	Malls in cell	Malls within 1km	Malls within 3km	Cluster
count	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000	3.0	3.000000	3.0
mean	917.894687	239905.668501	110.364255	37.516595	79999.226792	10.000000	18.666667	15.000000	3.666667	7.0	14.333333	4.0
std	530.559682	44.548550	0.036465	0.085086	776.105034	2.645751	2.516611	3.605551	1.527525	2.0	2.081666	0.0
min	313.659591	239862.000000	110.327110	37.429922	79208.648625	8.000000	16.000000	12.000000	2.000000	5.0	12.000000	4.0
25%	723.050820	239882.978846	110.346382	37.474892	79618.840188	8.500000	17.500000	13.000000	3.000000	6.0	13.500000	4.0
50%	1132.442050	239903.957692	110.365655	37.519863	80029.031750	9.000000	19.000000	14.000000	4.000000	7.0	15.000000	4.0
75%	1220.012235	239927.502752	110.382828	37.559931	80394.515875	11.000000	20.000000	16.500000	4.500000	8.0	15.500000	4.0
max	1307.582420	239951.047812	110.400000	37.600000	80760.000000	13.000000	21.000000	19.000000	5.000000	9.0	16.000000	4.0

Table 9 - Score information for target cluster

	Mall Score	Shop Score	Score
count	3.000000	3.000000	3.000000
Mean	53.666667	121.000000	67.333333
std	6.110101	11.532563	15.947832
min	47.000000	109.000000	54.000000
25%	51.000000	115.500000	58.500000
50%	55.000000	122.000000	63.000000
75%	57.000000	127.000000	74.000000
max	59.000000	132.000000	85.000000

Shown in the tables above, we can see that there are 3 hexagons within Cluster 4. There is an average of 10 clothing stores within each cell whilst there is an average of 3.6 shopping malls. Additionally, there is an average of 18.6 clothing stores within 1km of the centre of the cluster center, whilst there is an average 14.3 shopping malls.
We will plot every cluster using matplotlib so we can compare each cluster from one another for each feature.

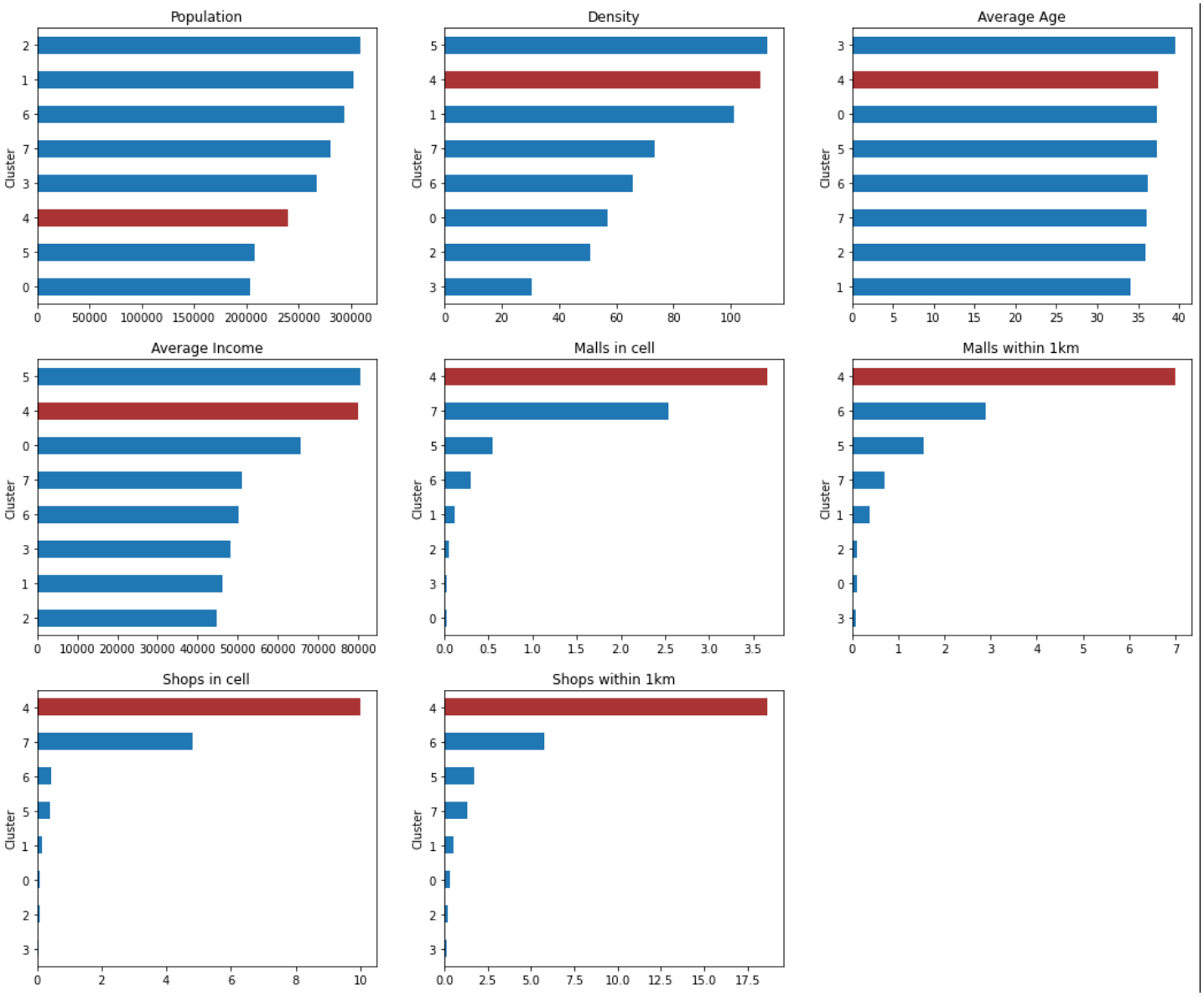

Figure 13 – Bar charts showing frequency of shops and malls, and census data

From the bar charts above, we can see that Cluster 4 has one of the lowest population counts, however it has one of the highest density counts. We can also see that Cluster 4 has the second highest average income, which is very promising for stakeholders who want to open a clothing store in that area. Lastly, we can see that there is a large number of shops and malls within cluster 4, compared to other cells.

Normally, we would pick the first 5 hexagons within the cluster to recommend opening a clothing store. However as the cluster only contains 3 hexagon cells, we can recommend all three of them as they will all reflect the most promising area to open a store.

We will still list these hexagons in descending order just to see which cells are the most promising.

Table 10 - Scores for each hexagon cell within the target cluster

	Population	Density	Average Age	Average Income	Malls in cell	Malls within 1km	Shops in cell	Shops within 1km	Mall Score	Shop Score	Score
914	239951.047812	110.327110	37.429922	79208.648625	2	7	13	16	47	132	85
857	239903.957692	110.365655	37.519863	80029.031750	4	9	9	21	59	122	63
856	239862.000000	110.400000	37.600000	80760.000000	5	5	8	19	55	109	54

The last task of this analysis is to plot the target cluster (4).

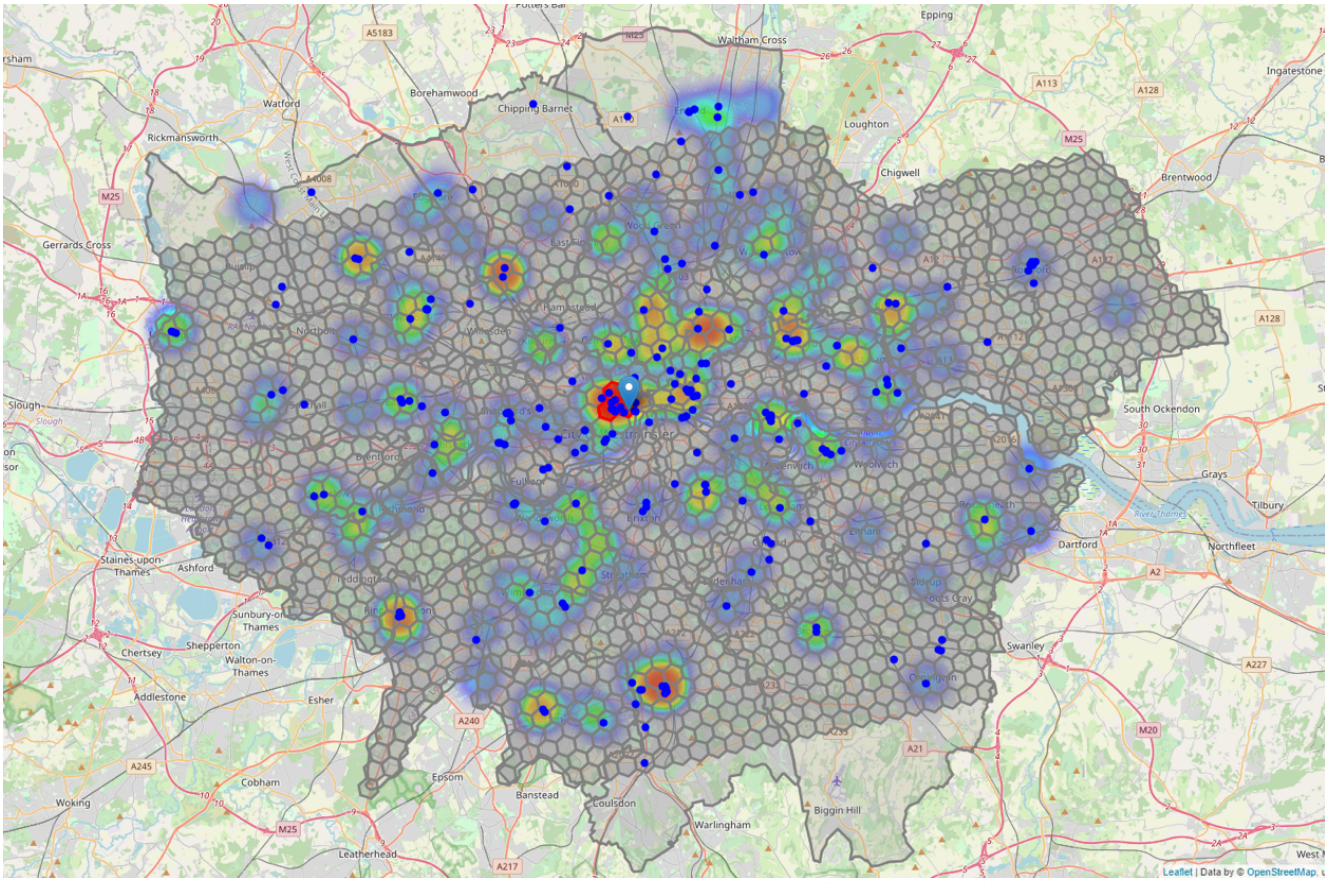

Figure 14 - Map showing target cluster and heatmap of clothing stores/malls

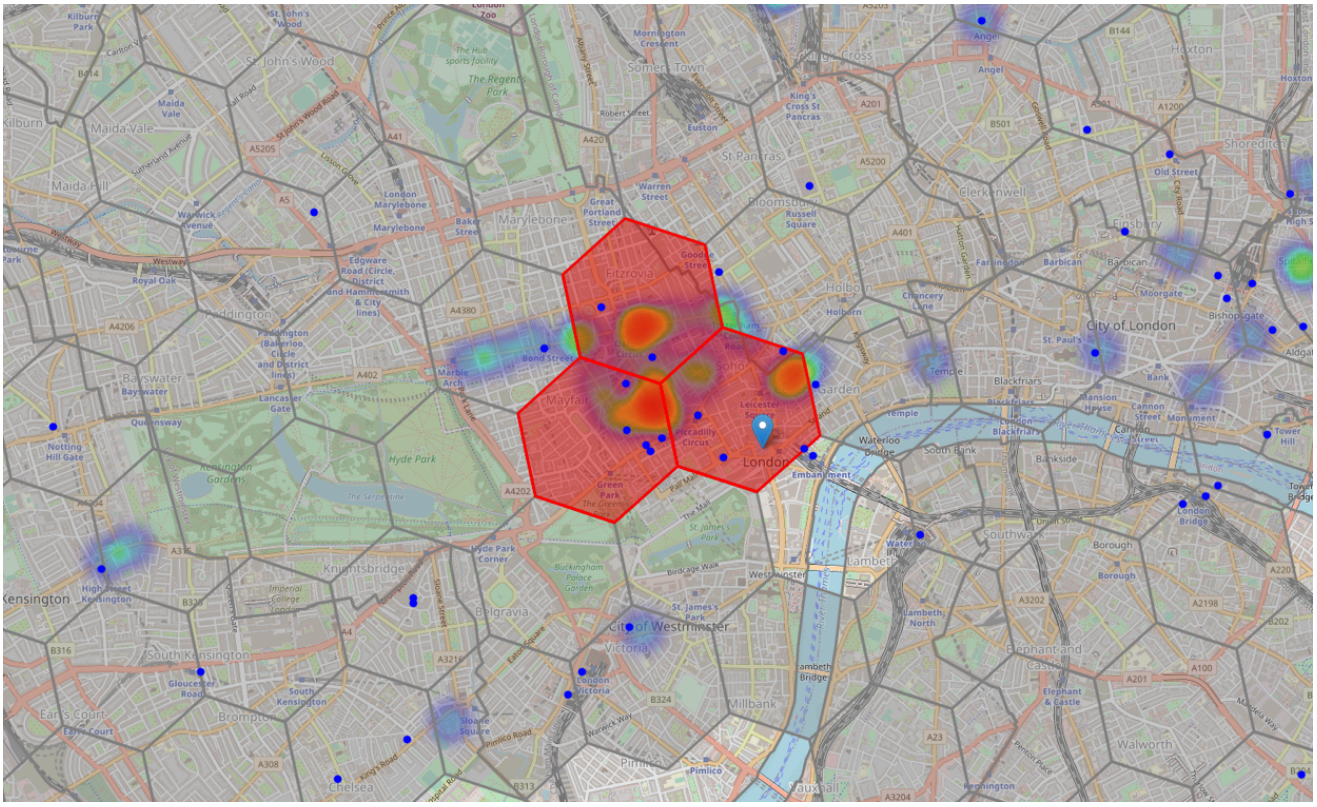

Figure 15 - Map showing target cluster in detail

This concludes our analysis. We have used K-means to segment and cluster the neighbourhoods of London to find the most promising areas open a clothing store. These areas contain the highest concentration of clothing stores and shopping mall, which satisfy the criteria we defined at the start of the project.

Each neighbourhood and borough are represented using a hexagon shape in a grid format, this is a popular method of viewing maps. The areas in the cluster highlighted in the middle also contain the highest average income, population and population density.

5 Results and Evaluation

After segmenting and clustering every single neighbourhood within London, we were able to accurately determine which areas are the most promising to open a clothing store.

We generated a hexagon cell grid that covers approximately 95% of the land area in London. We then grouped these hexagon cells into 8 clusters based upon features such as income, density, population, and age. We also considered existing clothing stores and shopping malls in our analysis to ultimately decide which area provides the most value to the primary stakeholder. From the analysis carried out, we can see that clothing stores are mostly located near shopping malls. Thus, we wanted to find the area which has a large number of both clothing stores and shopping malls.

The process of choosing how many clusters we needed was through statistical testing. In this project we used the sum of squared difference and silhouette score to determine how many clusters would produce the most accurate results. In the end we chose K=8 clusters. After determining the optimal amount of K neighbours to use, we clustered every hexagon cell within London. We found a hexagon cluster consisting of 3 cells which was the most promising cluster of all 8. Located right by the center of London (Trafalgar Square), this cluster contains the highest concentration of clothing stores and shopping mall. Additionally, the census data in this cluster is very promising. The cluster has the highest population density and second highest average income. A very high population density indicates that there are a lot of people present in the area, and most likely visiting different shops. A high average income is also beneficial because consumers are more likely to have a higher disposable income, which can be spent on clothing. This leads to more sales and ultimately higher profits, a vital factor for owners looking for ideal places to open a business.

The chosen cluster only contains 3 hexagon cells, therefore we can recommend all areas within this cluster, as they all represent the optimal location to open a clothing store. These recommended areas are a good starting point for further analysis and adding more features. There are a number of factors which could impact which area is optimal to open a store. for example, revenue for all clothing stores, rent/real estate prices, crime and traffic could all have an influence on which area is the best. Computing with more features and data, ultimately, will lead to more accurate results and higher confidence that the answer is correct.

6 Conclusion

The goal of this project was to find the best area to open a clothing store in London. There were several important objectives we defined to achieve this goal. First we divided London into hexagon cells in a honeycomb grid format. Ultimately, this enabled us to uniquely identify promising areas with high accuracy. Secondly, we used the Foursquare API to fetch clothing store and shopping centre venue data. Additionally, we obtained census data from the London Datastore to gain a further understanding of the city and the people living in it. Thirdly, we grouped similar areas together by using the K-means clustering algorithm. We determined the 3 most promising areas which represent a mixture of factors such as high average income, population density and most importantly, a lot of clothing stores nearby.

All in all, we were able to achieve the goal we set out at the beginning. Although we collected enough and analysed data to answer the question, adding more features to the model would allow stakeholders to gain a more sophisticated understanding of London. For example, if we added a feature dedicated to commercial real estate/rent prices, we could advise stakeholders different areas based upon their budget. That said, whilst we could collect more data, it would not have a significant impact upon the analysis and the predictions made.

Now equipped with insights and factual data, stakeholders can now ultimately decide and explore which area they think has the most potential. Ultimately, they have to make a decision and our job was to help inform that decision with analytics. Their decision, however, could be impacted by other factors that we have not covered in this analysis. For example, revenue for existing clothing stores, rent prices and traffic could all impact the stakeholder's final decision.

7 References

- *Open a Movie Theatre in Montreal (accessed at: https://towardsdatascience.com/the-battle-of-the-neighborhoods-open-a-movie-theater-in-montreal-355cf5c679b8)*
- *Census Data for London (accessed at: https://data.london.gov.uk/census/)*
- *London boundaries data (accessed at: https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london)*
- *Foursquare Places API Documentation (accessed at: https://developer.foursquare.com/docs/places-api/)*
- *Folium documentation (accessed at: https://python-visualization.github.io/folium/)*
- *Numpy documentation (accessed at : https://numpy.org/doc/)*
- *Pandas documentation (accessed at: https://pandas.pydata.org/docs/)*
- *K- means clustering documentation (accessed at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html)*
- *Google geocoding API documentation (accessed at: https://developers.google.com/maps/documentation/geocoding/overview)*