Applied Data Science Capstone

Introduction

Ever since the UK government introduced new planning policy guidelines aimed at revitalising city centres and control out-of-town developments, planners and developers have seen cinemas as key amenities to make urban masterplans and their associated schemes more attractive. This interest has grown since 2008, despite the retail-led regeneration having little effect in averting the decline of many regional city centres. Tough economic conditions combined with the growth in online shopping meant that retailers needed fewer physical outlets to establish their brand, and recession-hit shoppers knew that there were often better bargains to be had online. Consumers now need more compelling reasons to come into town, as well as a variety of ways to spend their time once they have arrived. Cinemas, as part of a mix of restaurants, cafes and other leisure facilities, can help address this, increasing consumer dwell time and helping to create a sustainable night-time economy that gives city centres a life beyond the usual retail opening hours.

As major retailers come under pressure to attract shoppers with leisure-related activities, a growing number of cinema construction projects are being built across the country to help ensure that new developments succeed. Developers are responding to the growing demand for ‘experiences’ at leisure destinations and a revival in the popularity of the big screen. Cinema attendance in the UK exceeded 177 million in 2018, the highest level since 1970. According to agents CBRE there were 33 new multiplex cinemas in the three-year development pipeline in the UK at the end of 2018. Half of the schemes are set to be in shopping centres, rather than out-of-town leisure parks. A recent report on the future of leisure commissioned by Legal & General Investment Management, envisages the spread of so-called ‘fusion cinemas’ which use new technology and act as anchor tenants for schemes. By the mid-2020s, it sees UK cinemas acting as cultural community spaces with film screenings running alongside co-working, education, and dining events.

The focus of this project will be establishing which Boroughs in Greater London are currently underserved by major Cinema chains alongside smaller Independent Cinemas, then viewing this data through the lens of population information using the most recent census data to gain an understanding of which boroughs would represent ideal investment opportunities for developers and chains seeking to build additional cinemas to serve the communities within which they reside.

Data Sources

There will be three primary data sources supporting this study. ONS population data will be used to establish the population levels within each of the 32 London boroughs. Further to this, the Wikipedia entry for the London Boroughs will be utilised to gain an oversight of the names of both Boroughs and Local Authorities that control those Boroughs. Finally, we will be using the Foursquare API to get an understanding of the number of cinemas already present within each of those locations.

Methodology

To produce a list of each of the London Boroughs, in addition to further information about their political control was scraped using from a table on Wikipedia using Urllib. Once the data was scraped and assigned to a variable, we used Beautiful Soup to extract and work with the data in it. Beautiful Soup is a Python library for pulling data out of HTML and XML files. In this case we used it alongside 'html5lib' as the parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. From this searching of the Parse tree it was possible to establish that the class ID of the table in questions as 'Wikitable sortable' so a further function pulled out only tables with this ID. Once these were established it was then possible to loop through the rows of the table using the tr (row) and td (cell) fields to populate the new Dataframe with the data in question. This newly created table was then cleaned to remove unnecessary columns from the point of view of future analysis.

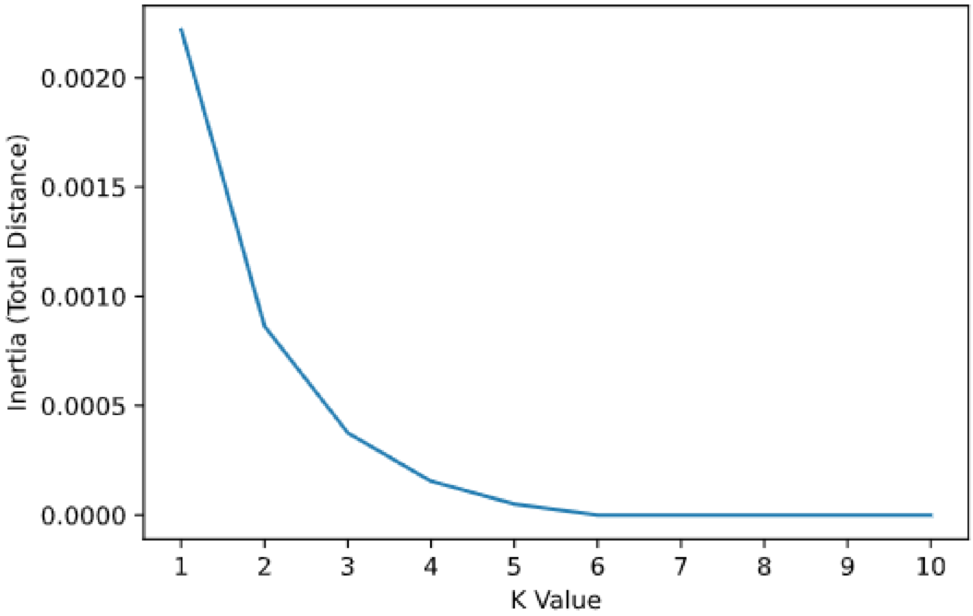
Once a table had been created and stored the data for each of the London Boroughs Geocoder was used to find the Latitude and Longitude of each of the geographic centres of the Boroughs. While the source table did include Lat/Long information, this was for the political headquarters of each Borough, rather than the centre, the latter being most suitable for the analysis. This was saved as a separate Dataframe.

latter being most suitable for the analysis. This was saved as a separate Dataframe.

These geographic coordinates were then passed through the Foursquare API to look for nearby venues within a 6 Kilometre range from the centre of each borough. A onehot was then produced to establish the mean number of each venue category produced by the resultant search and this Dataframe was then queried for unique values within 'Venue Category' which established 'Movie Theatre' and 'Indie Movie Theatre' as the categories of interest for this study. Onehots for these categories were extracted for each borough and saved as a new Dataframe.

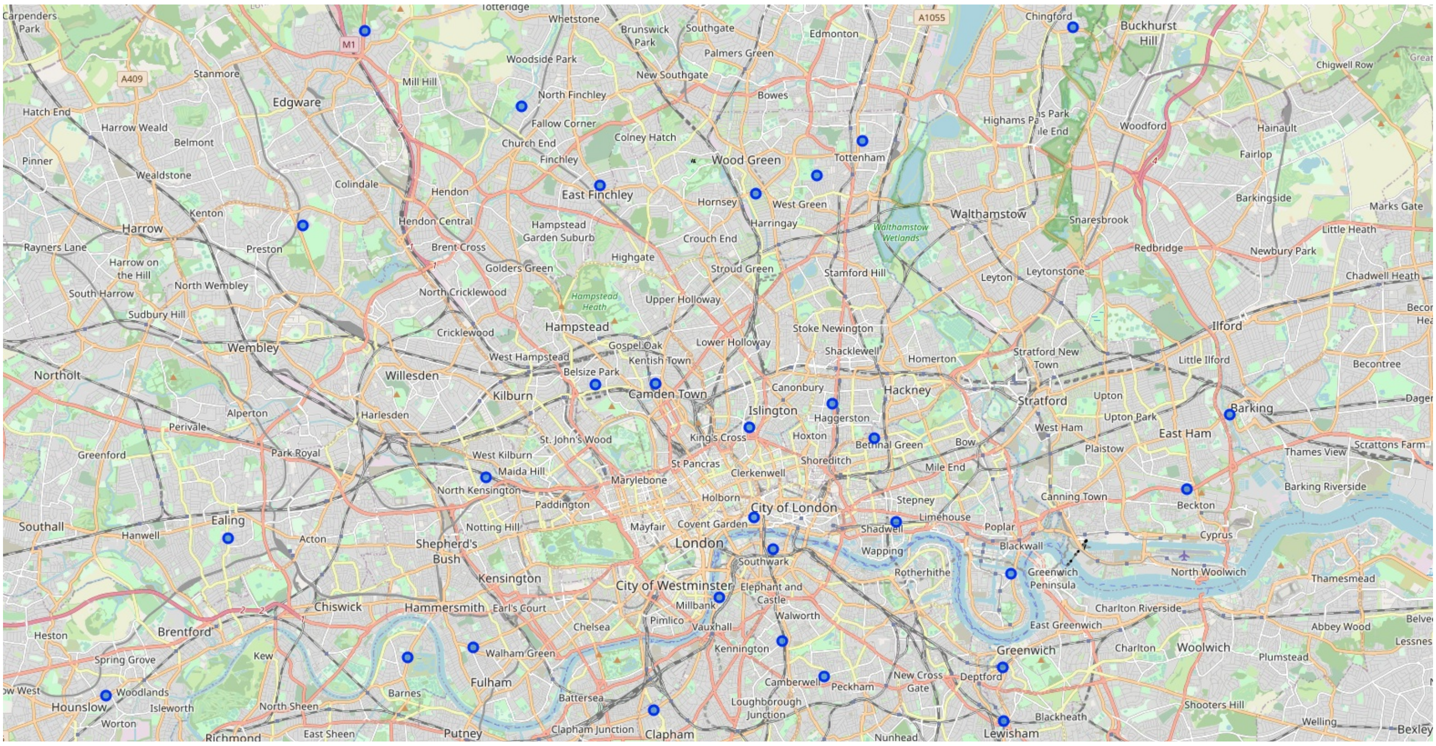
K-means Clustering was then utilised to cluster boroughs based on the onehot output around the venue categories previously established. Prior to a full run a cluster_variance function was defined to establish inertia looking at total Euclidean distance between data values for K=1 through K=10. This search produced the below plot, Figure 1, using the elbow method it was possible to identify a K value of 2 or 3 as being ideal for this study. A K value of 3 was ultimately used as this produced three clear clusters, one with presence of either a Movie Theatre, an Independent Movie Theatre or neither.

Figure 1



Results

Figure 2: The latitude and longitude of each London Borough mapped using Folium, using coordinates extracted using Geocoder.



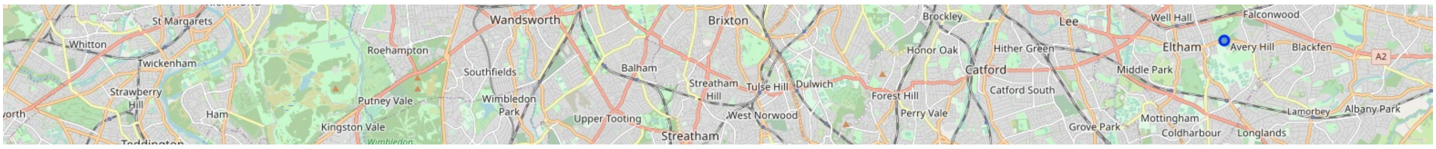
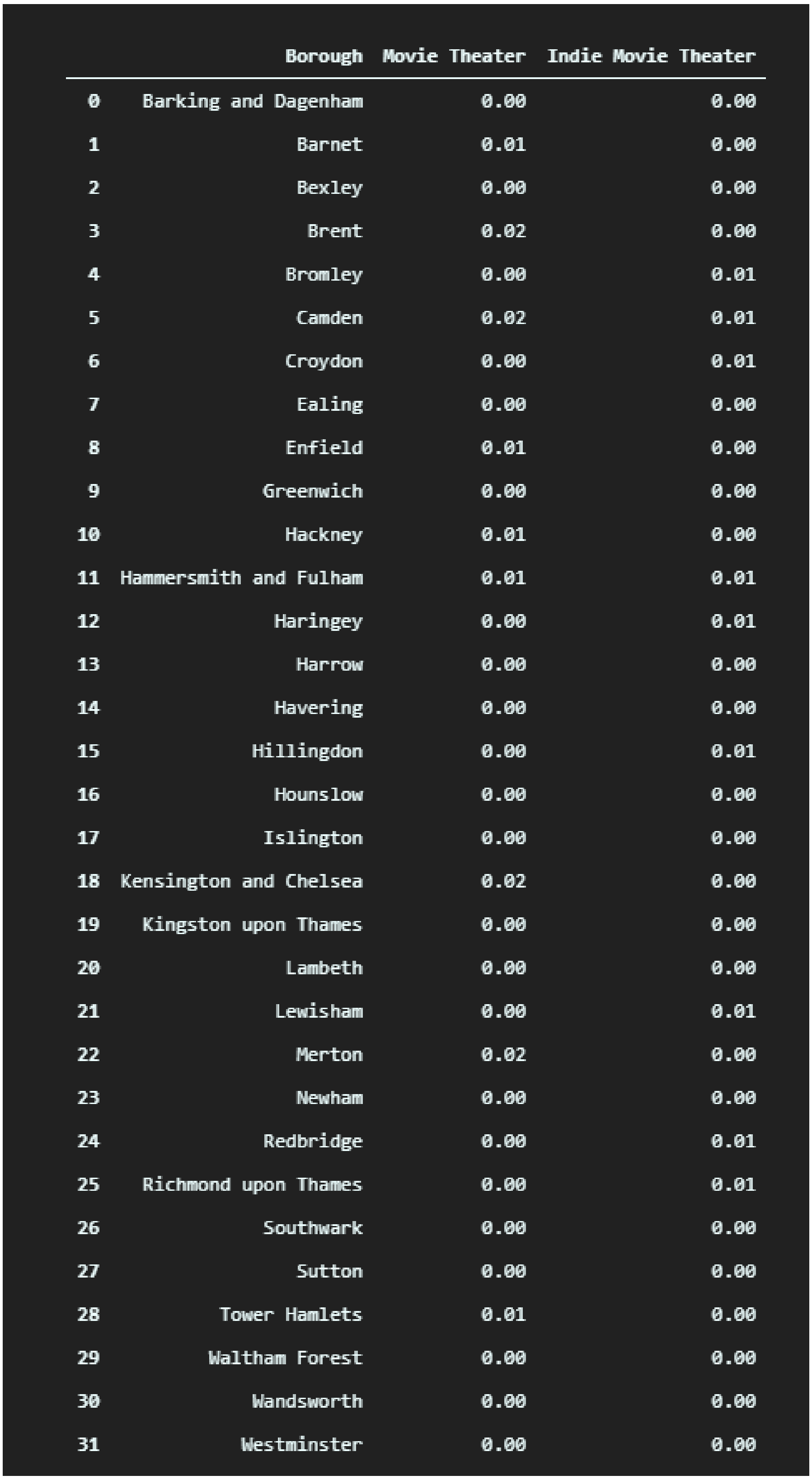
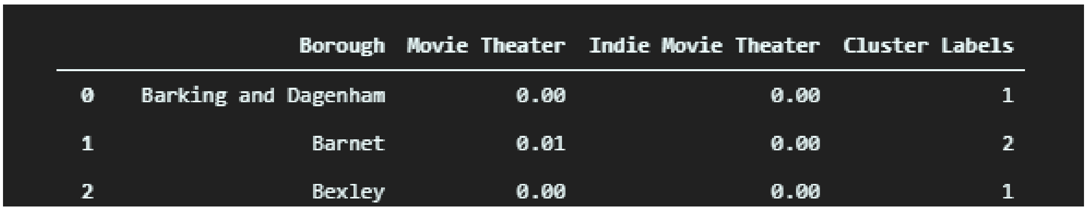


Figure 3: **Dataframe produced after querying the Foursquare API for either locations with a Venue Category of either Movie Theatre or Independent Movie Theatre, then passed through a onehot for each London Borough.**

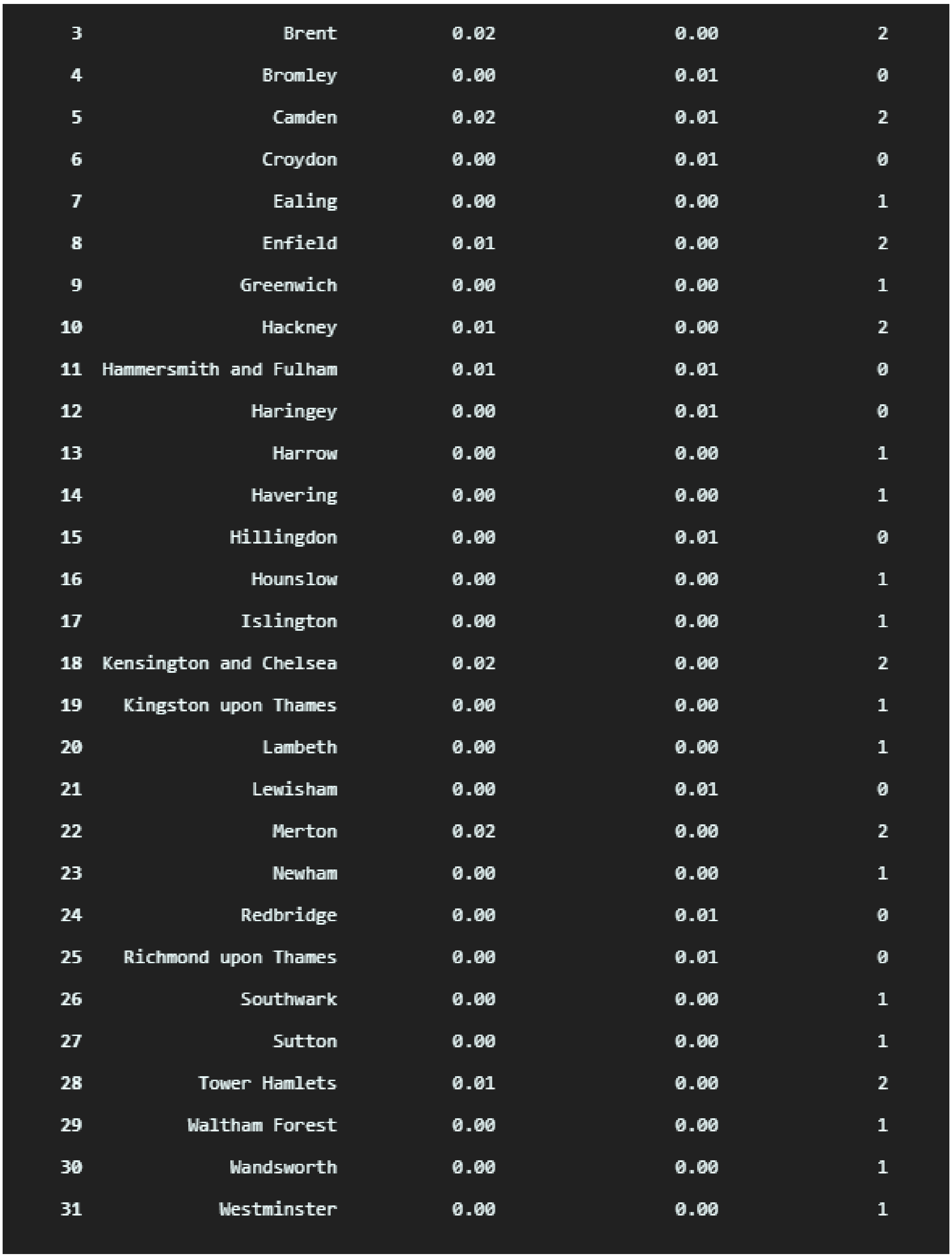


	Borough	Movie Theater	Indie Movie Theater
0	Barking and Dagenham	0.00	0.00
1	Barnet	0.01	0.00
2	Bexley	0.00	0.00
3	Brent	0.02	0.00
4	Bromley	0.00	0.01
5	Camden	0.02	0.01
6	Croydon	0.00	0.01
7	Ealing	0.00	0.00
8	Enfield	0.01	0.00
9	Greenwich	0.00	0.00
10	Hackney	0.01	0.00
11	Hammersmith and Fulham	0.01	0.01
12	Haringey	0.00	0.01
13	Harrow	0.00	0.00
14	Havering	0.00	0.00
15	Hillingdon	0.00	0.01
16	Hounslow	0.00	0.00
17	Islington	0.00	0.00
18	Kensington and Chelsea	0.02	0.00
19	Kingston upon Thames	0.00	0.00
20	Lambeth	0.00	0.00
21	Lewisham	0.00	0.01
22	Merton	0.02	0.00
23	Newham	0.00	0.00
24	Redbridge	0.00	0.01
25	Richmond upon Thames	0.00	0.01
26	Southwark	0.00	0.00
27	Sutton	0.00	0.00
28	Tower Hamlets	0.01	0.00
29	Waltham Forest	0.00	0.00
30	Wandsworth	0.00	0.00
31	Westminster	0.00	0.00

Figure 4: **Dataframe produced by using K-means Clustering on the dataset with a K of 3, including cluster labels for each location.**

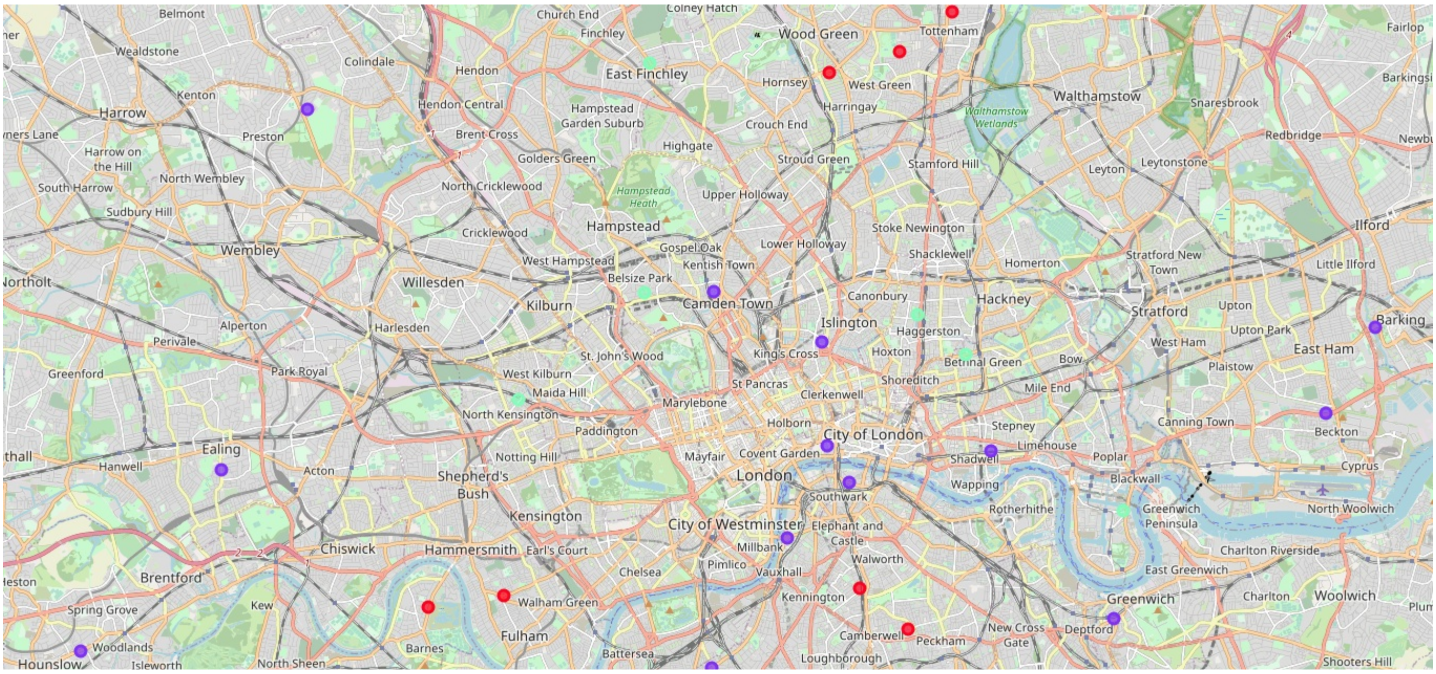


	Borough	Movie Theater	Indie Movie Theater	Cluster Labels
0	Barking and Dagenham	0.00	0.00	1
1	Barnet	0.01	0.00	2
2	Bexley	0.00	0.00	1



```
 3                   Brent       0.02        0.00      2
 4                 Bromley       0.00        0.01      0
 5                  Camden       0.02        0.01      2
 6                 Croydon       0.00        0.01      0
 7                  Ealing       0.00        0.00      1
 8                 Enfield       0.01        0.00      2
 9               Greenwich       0.00        0.00      1
10                 Hackney       0.01        0.00      2
11  Hammersmith and Fulham       0.01        0.01      0
12                Haringey       0.00        0.01      0
13                  Harrow       0.00        0.00      1
14                Havering       0.00        0.00      1
15              Hillingdon       0.00        0.01      0
16                Hounslow       0.00        0.00      1
17               Islington       0.00        0.00      1
18  Kensington and Chelsea       0.02        0.00      2
19    Kingston upon Thames       0.00        0.00      1
20                 Lambeth       0.00        0.00      1
21                Lewisham       0.00        0.01      0
22                  Merton       0.02        0.00      2
23                  Newham       0.00        0.00      1
24               Redbridge       0.00        0.01      0
25    Richmond upon Thames       0.00        0.01      0
26               Southwark       0.00        0.00      1
27                  Sutton       0.00        0.00      1
28           Tower Hamlets       0.01        0.00      2
29          Waltham Forest       0.00        0.00      1
30              Wandsworth       0.00        0.00      1
31             Westminster       0.00        0.00      1
```

Figure 5: **The below figure is the data from the previous Dataframe mapped spatially using Folium and the earlier defined Latitudes and Longitudes**



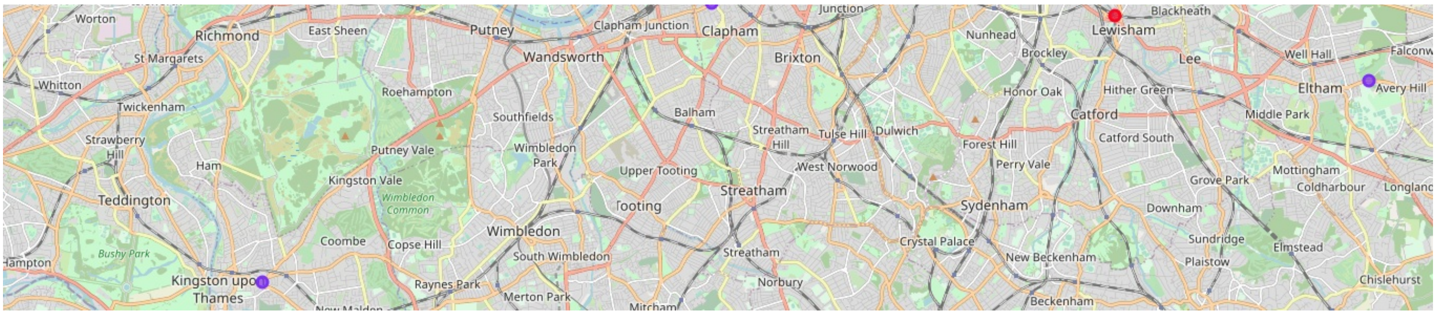
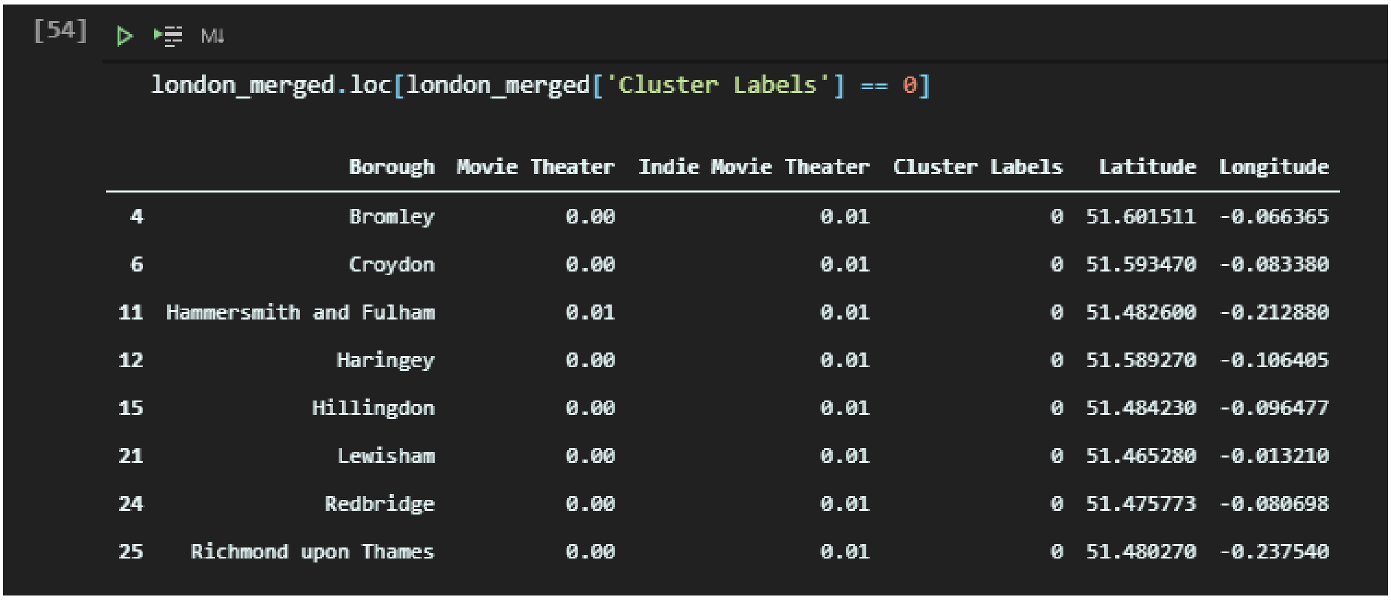


Figure 6: **Dataframe showing cluster '0'**

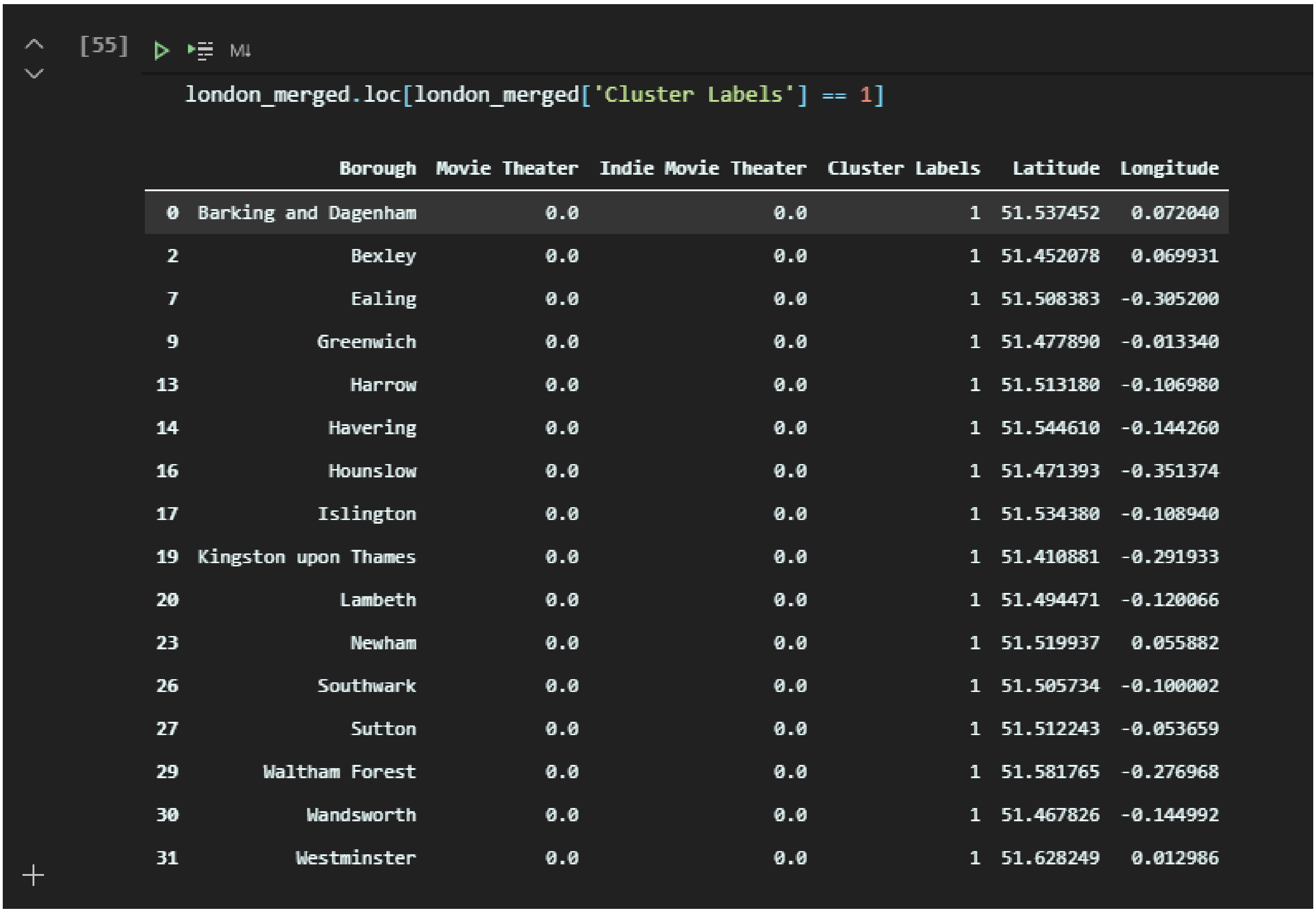


[54]

```
london_merged.loc[london_merged['Cluster Labels'] == 0]
```

	Borough	Movie Theater	Indie Movie Theater	Cluster Labels	Latitude	Longitude
4	Bromley	0.00	0.01	0	51.601511	-0.066365
6	Croydon	0.00	0.01	0	51.593470	-0.083380
11	Hammersmith and Fulham	0.01	0.01	0	51.482600	-0.212880
12	Haringey	0.00	0.01	0	51.589270	-0.106405
15	Hillingdon	0.00	0.01	0	51.484230	-0.096477
21	Lewisham	0.00	0.01	0	51.465280	-0.013210
24	Redbridge	0.00	0.01	0	51.475773	-0.080698
25	Richmond upon Thames	0.00	0.01	0	51.480270	-0.237540

Figure 7: **Dataframe showing cluster '1'**

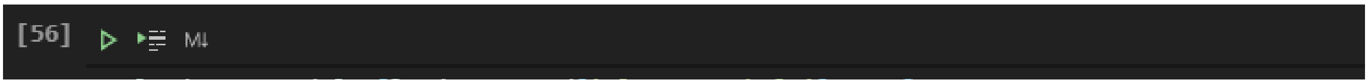


[55]

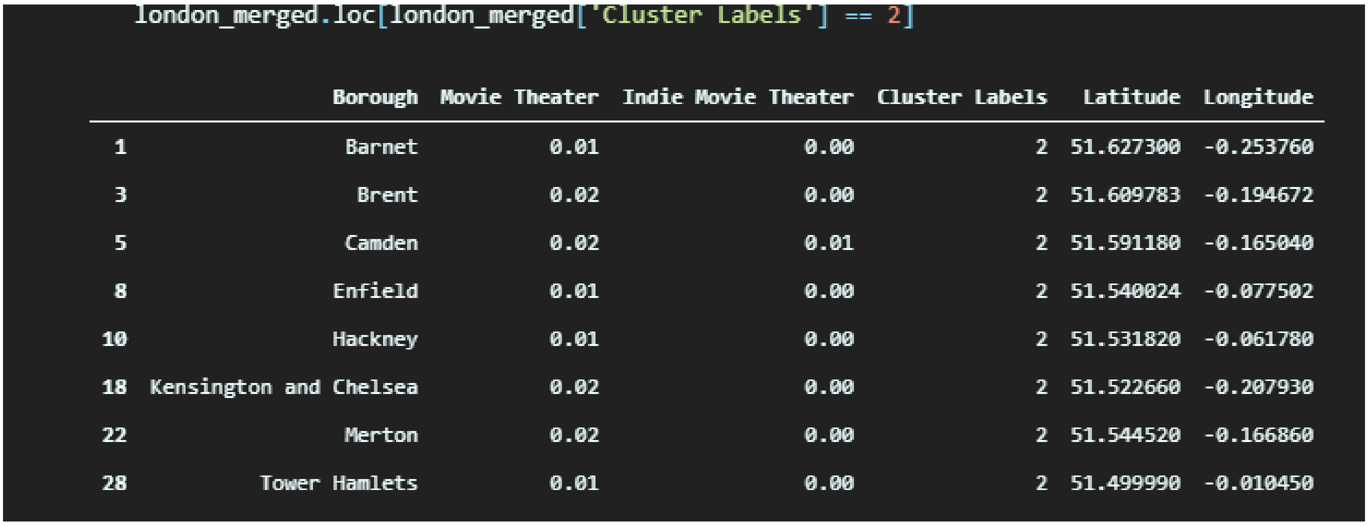
```
london_merged.loc[london_merged['Cluster Labels'] == 1]
```

	Borough	Movie Theater	Indie Movie Theater	Cluster Labels	Latitude	Longitude
0	Barking and Dagenham	0.0	0.0	1	51.537452	0.072040
2	Bexley	0.0	0.0	1	51.452078	0.069931
7	Ealing	0.0	0.0	1	51.508383	-0.305200
9	Greenwich	0.0	0.0	1	51.477890	-0.013340
13	Harrow	0.0	0.0	1	51.513180	-0.106980
14	Havering	0.0	0.0	1	51.544610	-0.144260
16	Hounslow	0.0	0.0	1	51.471393	-0.351374
17	Islington	0.0	0.0	1	51.534380	-0.108940
19	Kingston upon Thames	0.0	0.0	1	51.410881	-0.291933
20	Lambeth	0.0	0.0	1	51.494471	-0.120066
23	Newham	0.0	0.0	1	51.519937	0.055882
26	Southwark	0.0	0.0	1	51.505734	-0.100002
27	Sutton	0.0	0.0	1	51.512243	-0.053659
29	Waltham Forest	0.0	0.0	1	51.581765	-0.276968
30	Wandsworth	0.0	0.0	1	51.467826	-0.144992
31	Westminster	0.0	0.0	1	51.628249	0.012986

Figure 8: **Dataframe showing cluster '2'**



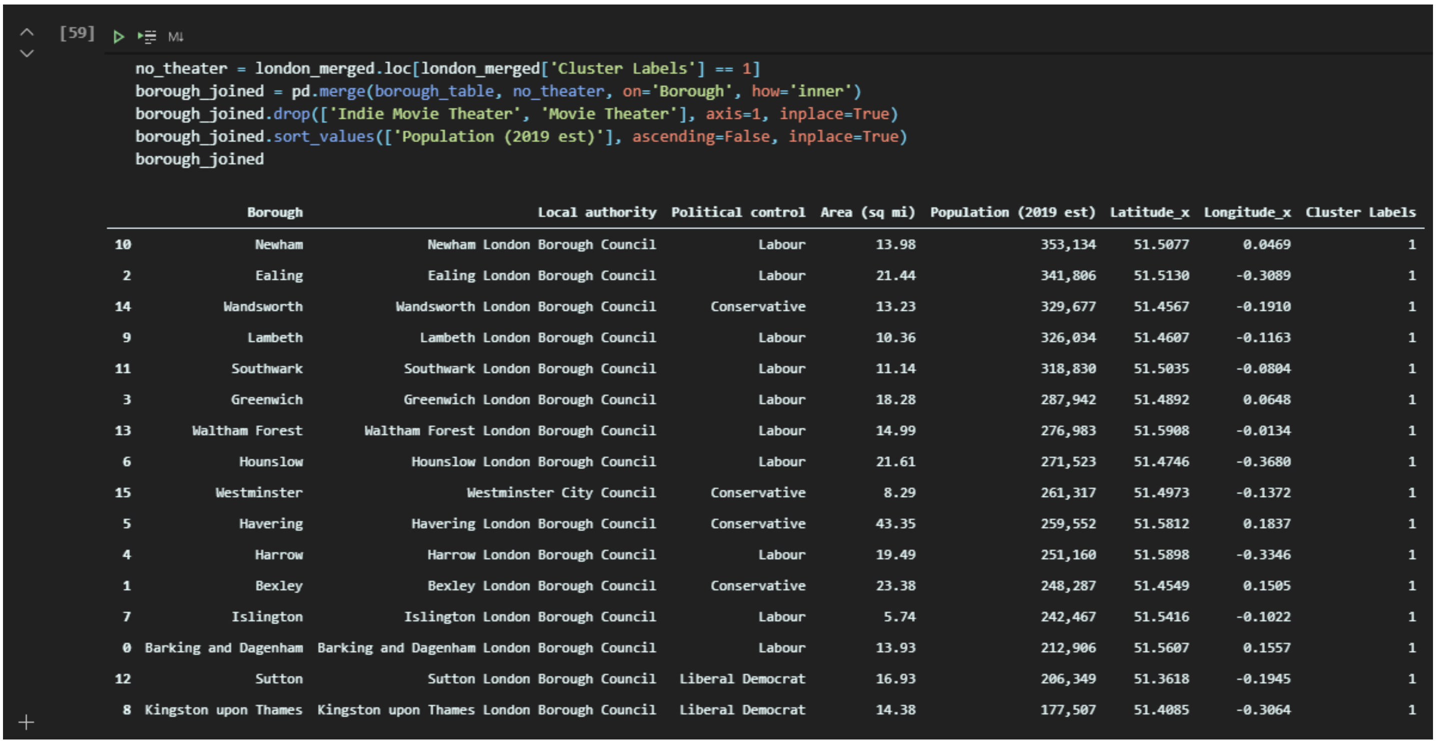
[56]



```python
london_merged.loc[london_merged['Cluster Labels'] == 2]
```

	Borough	Movie Theater	Indie Movie Theater	Cluster Labels	Latitude	Longitude
1	Barnet	0.01	0.00	2	51.627300	-0.253760
3	Brent	0.02	0.00	2	51.609783	-0.194672
5	Camden	0.02	0.01	2	51.591180	-0.165040
8	Enfield	0.01	0.00	2	51.540024	-0.077502
10	Hackney	0.01	0.00	2	51.531820	-0.061780
18	Kensington and Chelsea	0.02	0.00	2	51.522660	-0.207930
22	Merton	0.02	0.00	2	51.544520	-0.166860
28	Tower Hamlets	0.01	0.00	2	51.499990	-0.010450

Figure 9: **Final Dataframe including information from above, but having used pd.Merge to merge frame with previously defined Dataframe with London Borough data. Sorted by the values in the 'Population (2019 est.)' column.**



[59]

```python
no_theater = london_merged.loc[london_merged['Cluster Labels'] == 1]
borough_joined = pd.merge(borough_table, no_theater, on='Borough', how='inner')
borough_joined.drop(['Indie Movie Theater', 'Movie Theater'], axis=1, inplace=True)
borough_joined.sort_values(['Population (2019 est)'], ascending=False, inplace=True)
borough_joined
```

	Borough	Local authority	Political control	Area (sq mi)	Population (2019 est)	Latitude_x	Longitude_x	Cluster Labels
10	Newham	Newham London Borough Council	Labour	13.98	353,134	51.5077	0.0469	1
2	Ealing	Ealing London Borough Council	Labour	21.44	341,806	51.5130	-0.3089	1
14	Wandsworth	Wandsworth London Borough Council	Conservative	13.23	329,677	51.4567	-0.1910	1
9	Lambeth	Lambeth London Borough Council	Labour	10.36	326,034	51.4607	-0.1163	1
11	Southwark	Southwark London Borough Council	Labour	11.14	318,830	51.5035	-0.0804	1
3	Greenwich	Greenwich London Borough Council	Labour	18.28	287,942	51.4892	0.0648	1
13	Waltham Forest	Waltham Forest London Borough Council	Labour	14.99	276,983	51.5908	-0.0134	1
6	Hounslow	Hounslow London Borough Council	Labour	21.61	271,523	51.4746	-0.3680	1
15	Westminster	Westminster City Council	Conservative	8.29	261,317	51.4973	-0.1372	1
5	Havering	Havering London Borough Council	Conservative	43.35	259,552	51.5812	0.1837	1
4	Harrow	Harrow London Borough Council	Labour	19.49	251,160	51.5898	-0.3346	1
1	Bexley	Bexley London Borough Council	Conservative	23.38	248,287	51.4549	0.1505	1
7	Islington	Islington London Borough Council	Labour	5.74	242,467	51.5416	-0.1022	1
0	Barking and Dagenham	Barking and Dagenham London Borough Council	Labour	13.93	212,906	51.5607	0.1557	1
12	Sutton	Sutton London Borough Council	Liberal Democrat	16.93	206,349	51.3618	-0.1945	1
8	Kingston upon Thames	Kingston upon Thames London Borough Council	Liberal Democrat	14.38	177,507	51.4085	-0.3064	1

Discussion

The project's main goal is to determine the best location for opening a Movie Theatre business in London. Discussing what locations can be considered "the best" may vary but we can equate it as the most conducive ones by considering the following criteria:

1. **Presence of Competitor Sites**

Figure 5 helps to fully illustrate a significant clustering of current Movie Theatres in the areas South of the River Thames, firmly within central London. This also extends to the area within the North of London. Again, all these areas clustered within the major commercial zones. With the coverage of these locations decreasing as you head towards the suburbs around the M25. If referring to the clusters of data produced in Figures 6 it's possible to identify those Boroughs well served by independent cinemas without any real coverage of mainstream chains and the influence brought by size of location, both in terms of number of screens and maximum footfall. By comparison however, Figure 8 helps to illustrate Boroughs with good coverage of mainstream cinemas, without a great deal of coverage of independent locations, and the ability for those locations to show the mixture of blockbuster and arthouse films that make them so well loved by their clientele.

Figure 7 however, illustrates locations that are completely unserved by either type of location and might possibly present good investment opportunities by developers or chains looking to open new locations. This however leads on to the next analysis.

1. Population in Boroughs

Figure 9 displays those locations from figure 7, with neither mainstream nor independent cinemas, overlaid with the population data extracted earlier. It stands to reason that a significant factor in deciding where to open a new cinema might be the potential number of customers that would help any newly opened location reach the critical mass of business required to transition into an established site. The ONS population figures indicate that locations such as Newham or Ealing are best placed as areas of further exploration. This however leads us onto a few of the limiting factors of this study.

Areas for development

As completed to date, there are two obvious flaws in the methodology of this work. Firstly, having searches cantered around a signal geographical point for each of the boroughs with a set radius for exploration leave open the possibility that a particular location, geographically proximate to two of the centres of each borough could well have been identified twice and assigned as a location for each borough. Similarly, it is entirely possible that a location was outside of the 6-kilometre radius that represented the search zone of venues around those Borough centroids. With that in mind it would be worth expanding the search radius, but then completing extra work to remove duplicates from the study, assigning a location identified multiple times to the centroid that it is closest to, rather than either counting multiple times, or not at all.

Furthermore, based on the analysis displayed in figure 9 it would be reasonable to suggest that any of the top 5 or 10 locations would make suitable candidates for a new cinema location. With that in mind it would therefore be useful to conduct further analysis on average travel times to and from each location to establish the ease of access to any new site using existing public transport links.

Conclusion

Although the service sector in London has certainly been significantly impacted by the volatility brought to the nation as a result of the ongoing pandemic, it stands to reason that markets are poised to return to some semblance of normality in a post-Covid world. None the less, businesses are certainly planning their exit strategy which may well involve expanding their footprint to include additional sites in underserved communities. It is with this fact in mind that this report focused on the integration of data science and machine learning technics to assist decision makers in making informed and efficient choices.

While this study was narrow in scope and certainly would require a significant expansion to provide any results that would be actionable in a real-world scenario. It could be reasonable to state that from the results provided, we could potentially recommend potential locations for future research into their theoretical profitability as venue locations.