Capstone Project - Opening a new business venue in a different city.

Business Problem

An entrepreneur owns a shop in a specific area of a city. The shop has a specific theme, and the area the shop is located had been selected because of its character. The existing venues and amenities give to the surroundings a specific vibe and characterisation that the entrepreneur deems fit for the theme of the shop.

The entrepreneur wants to open a new location in a different city and would like to select an area which has the same characteristics of the one in the previous city.

For this specific project we will consider the following instance:

An entrepreneur owns an arty/concept coffee-shop in SoHo, Manhattan, New York and would like to open a new location in London. The entrepreneur wants to select an area which has the same characteristics and vibes of SoHo, an area known for its commercialization and eclectic mix of venues ranging from restaurants and coffee-shops, to shopping boutiques and art galleries.

Data

The following procedure is used to create the datasets used in the project:

- A list of all areas of London, containing area name and OS Grid Ref is scraped from Wikipedia (https://en.wikipedia.org/wiki/List_of_areas_of_London)
- The centre of each London areas is located by converting OS Grid Reference number in geodetic coordinates (latitude and longitude).
- SoHo, New York centre in located and added to the dataset
- A list of venues around a radius of 500mt around every area centre is retrieved using the Foursquare API. Venue categories hierarchies are scraped from the Foursquare documentation (https://developer.foursquare.com/docs/build-with-foursquare/categories/)
- For every venue retrieved using the Foursquare API, the number of category keywords available is expanded using the category hierarchies.

The datasets created:

- A list of areas in all the areas of London + SoHo, New York, with centre coordinates
- A list of venues in SoHo, New York, with venue coordinates and venue characteristics.
- A list of venues in all the areas of London, with venue coordinates and venue characteristics.

Areas of London (first 5)

	borough	area	arealD	latitude	longitude
0	Bexley	Abbey Wood	0	51.485964	0.110225
1	Ealing	Acton	1	51.510080	-0.263398
2	Croydon	Addington	2	51.362403	-0.024762
3	Croydon	Addiscombe	3	51.381096	-0.067074
4	Bexley	Albany Park	4	51.434403	0.126554

Venues of London and New York (first 5)

	venueID	venue	latitude	longitude	categoryID	arealD	Category	Category_lv1	Tokens
0	4bc11de1abf49521cf98c093	Dance With Me SoHo	40.722578	-74.001363	4bf58dd8d48988d134941735	-1	Dance Studio	Arts & Entertainment	dance studio,dance,studio,performing arts venu...
1	4b6705a3f964a5207e352be3	Sam Brocato Salon	40.722371	-74.002562	4bf58dd8d48988d110951735	-1	Salon / Barbershop	Shop & Service	barbershop,service,shop,salon,shop & service,s...
2	4b96c70ff964a520dfe334e3	Hair Toto Group	40.718629	-73.999593	4bf58dd8d48988d110951735	-1	Salon / Barbershop	Shop & Service	barbershop,service,shop,salon,shop & service,s...
3	45e98bacf964a52080431fe3	MarieBelle	40.723101	-74.002477	4bf58dd8d48988d1d0941735	-1	Dessert Shop	Food	dessert shop,dessert,shop,food
4	52eddc12498e40bb655e0d7a	Ladurée	40.724314	-74.002453	4bf58dd8d48988d1d0941735	-1	Dessert Shop	Food	dessert shop,dessert,shop,food

Methodology

The main objective is to find which area of London has the same characteristics and vibes of SoHo, New York.

Few assumptions are made:

- Characteristics of an area can be inferred by the characteristics of the venues present in said area.
- A theme is a specific mix of venue characteristics which are somehow related together. (e.g. the theme “Sport & Fitness” refers to all the venues related to sport and fitness, such as gyms, fitness centres, SPAs, golf courses, etc.).
- An area is defined by a combination of themes (e.g. an area could be 0.8 Sport & Fitness and 0.2 Residential)
- Similarity between two areas is calculated based on the proportion of the same themes they share.

In the available dataset, every venue is defined by a series of tokens, created using the Foursquare categories hierarchies. These tokens are labels expressing the business and main qualities of the venue (e.g. food, drinks, restaurant, Italian).

The tokens are used to calculate 10 distinct themes using Latent Semantic analysis. These themes are then used to calculate the similarity between Soho, New York and every Area of London.

Most similar areas are defined as the ones with a similarity higher than 0.9.

As some of the most similar areas are geographically neighbouring or even partially overlapping, they are clustered by proximity using DBSCAN algorithm. The resulting clusters are groups of areas close to each other and can be treated as a single larger area where a shop could potentially be opened.

As a cluster could be composed by areas with different similarities, 5 Nearest Neighbours algorithm in order to calculated which cluster is the most similar to the NewYork area.

Analysis

In the available dataset, every venue is defined by a series of tokens, created using the Foursquare categories hierarchies. These tokens are labels expressing the business and main qualities of the venue (e.g. food, drinks, restaurant, Italian).

A Tf-idf matrix is built using the tokens. Such matrix contains the occurrences of any token for every area (including SoHo New York). The 432 tokens are used to calculate 10 distinct themes using Latent Semantic analysis (LSA). LSA is used to reduce the numbers of features (tokens) that are used to calculated the similarity between areas. Instead, groups of these features are created (themes), reducing the feature space down to 10.

```
THEME 0: food / shop / restaurant / shop & service / service / nightlife / nightlife spot / bar / outdoors / recreation / outdoors & recreation / store / drink / food & drink shop / pub

THEME 1: outdoors / recreation / outdoors & recreation / athletics / athletics & sports / sports / park / transport / travel / travel & transport / nightlife / nightlife spot / gym / gym / fitness center / fitness

THEME 2: station / transport / travel / travel & transport / train / train station / service / shop & service / platform / store / shop / bus / stop / bus stop / convenience

THEME 3: transport / travel / travel & transport / restaurant / station / train / train station / food / nightlife / nightlife spot / bus / bar / platform / bus stop / stop

THEME 4: nightlife / nightlife spot / bar / pub / service / shop & service / store / shop / convenience / convenience store / construction / construction & landscaping / landscaping / arts & entertainment / entertainment

THEME 5: arts / entertainment / arts & entertainment / clothing / clothing store / hotel / bus / shop & service / service / bus stop / stop / shop / travel / transport / travel & transport

THEME 6: bus / bus stop / stop / grocery store / grocery / park / travel & transport / travel / transport / store / food & drink shop / drink / hotel / trail / metro station

THEME 7: park / landscaping / construction & landscaping / construction / outdoors / outdoors & recreation / recreation / shop / coffee shop / coffee / restaurant / clothing store / clothing / convenience / convenience store

THEME 8: construction & landscaping / construction / landscaping / athletics & sports / athletics / sports / course / golf course / golf / soccer field / field / soccer / shop / pub / fast food restaurant

THEME 9: landscaping / construction / construction & landscaping / indian / indian restaurant / grocery store / grocery / arts & entertainment / entertainment / arts / food & drink shop / drink / restaurant / fast / fast food restaurant
```

The groups created can be interpreted as:

0. Shops, Food & Drinks
1. Fitness, Outdoor & Travel
2. Commuting & Shops
3. Commuting & Food
4. Nightlife & Entertainment,
5. Shops, Art & Entertainment
6. Transport
7. Outdoor, Recreation & Shops
8. Residential & Outdoor sport
9. Residential & Ethnic food

Similarity between two areas is calculated based on the proportion of the same themes they share. Cosine similarity is used to calculate the similarity between Soho, New York and all the Areas of London.

London Areas and similarity (first 5)

	borough	area	arealD	latitude	longitude	similarity
0	Bexley	Abbey Wood	0	51.485964	0.110225	0.615897
1	Ealing	Acton	1	51.510080	-0.263398	0.622542
2	Croydon	Addington	2	51.362403	-0.024762	0.128970
3	Croydon	Addiscombe	3	51.381096	-0.067074	0.650411
4	Bexley	Albany Park	4	51.434403	0.126554	0.447080

Only the most similar areas (similarity >= 0.9) are considered

London most similar areas (similarity >= 0.9) (top)

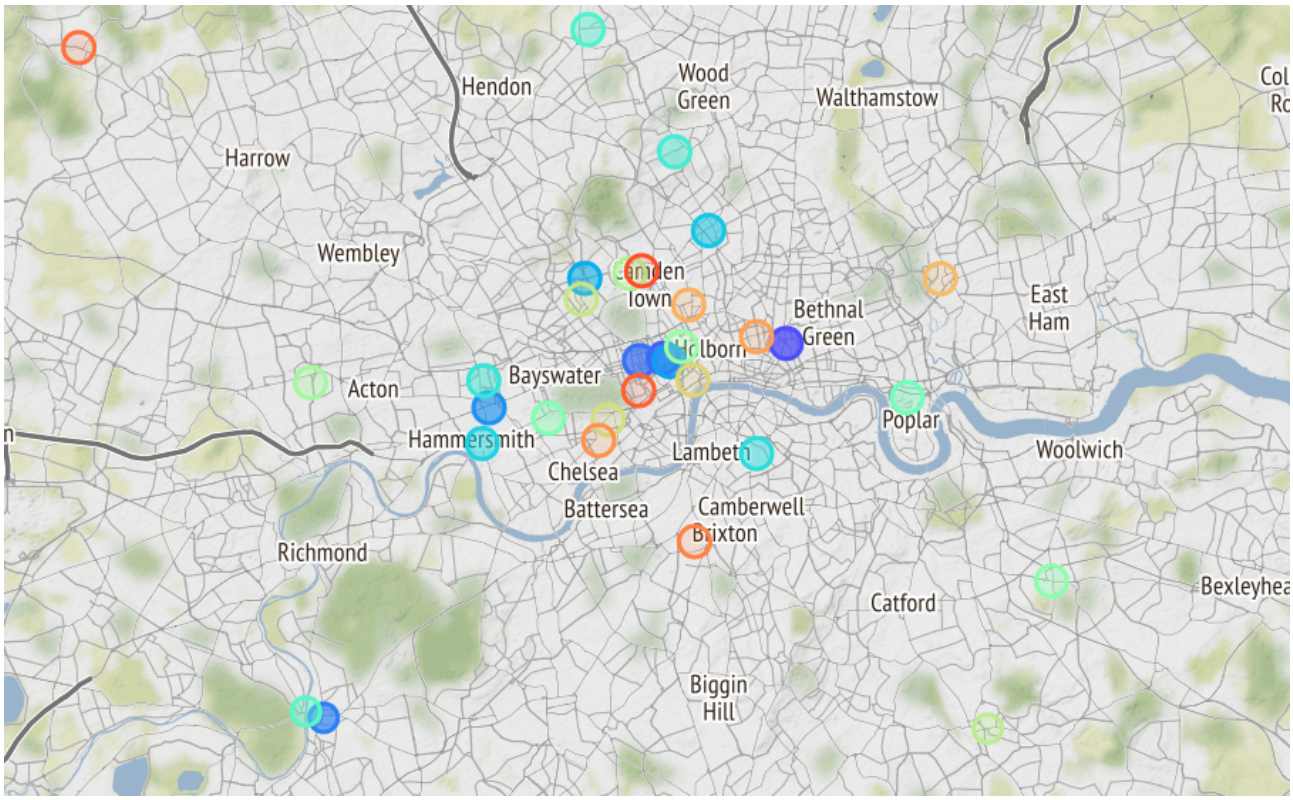
borough	area	similarity
Hackney	Shoreditch	0.973975
Camden	Fitzrovia	0.972802
Westminster	Marylebone (also St Marylebone)	0.965995
Kingston upon Thames	Kingston upon Thames	0.962092
Hammersmith and Fulham	Shepherd's Bush	0.960121
Westminster	Soho	0.956681
Camden	Swiss Cottage	0.952429
Bromley	Orpington	0.952374
Islington	Holloway	0.947718
Hammersmith and Fulham	Hammersmith	0.947667
Southwark	Walworth	0.942715

Many of the above areas are part of the same administrative zone (borough) and are located near each other:

Boroughs and Areas

borough	similarity_max	similarity_mean	nr_areas
Hackney	0.973975	0.973975	1
Camden	0.972802	0.930782	6
Westminster	0.965995	0.928376	6
Kingston upon Thames	0.962092	0.962092	1
Hammersmith and Fulham	0.960121	0.949979	3
Bromley	0.952374	0.937020	2
Islington	0.947718	0.927855	2
Southwark	0.942715	0.942715	1
Haringey	0.940209	0.940209	1
Barnet	0.936639	0.936639	1
Richmond upon Thames	0.935512	0.935512	1
Havering	0.934278	0.934278	1
Tower Hamlets	0.932519	0.932519	1
Kensington and Chelsea	0.932150	0.919616	2
Greenwich	0.926657	0.926657	1

Map of London Areas



Map of London: every circle is an area (radius 500m). Colour is unique per area, colour transparency identify the similarity (more solid the colour, higher the similarity)

As some of the most similar areas are geographically neighbouring or even partially overlapping, they are clustered by proximity using DBSCAN algorithm. The resulting clusters are groups of areas close to each other and can be treated as a single larger area where a shop could potentially be opened.

Clusters of London Areas

cluster	similarity_max	similarity_mean	nr_areas	nr_boroughs
0	0.973975	0.973975	1	1
1	0.972802	0.939781	6	2
2	0.962092	0.948802	2	2
3	0.960121	0.951135	2	1
4	0.952429	0.933862	2	2
5	0.952374	0.952374	1	1
6	0.947718	0.947718	1	1
7	0.947667	0.947667	1	1
8	0.942715	0.942715	1	1
9	0.940209	0.940209	1	1
10	0.936639	0.936639	1	1
11	0.934278	0.934278	1	1
12	0.932519	0.932519	1	1
13	0.932150	0.932150	1	1
14	0.926657	0.926657	1	1

The above table shows how clusters can be composed by several neighbouring areas belonging to different administrative zones (borough).

Example: details of cluster 1

	borough	area	areaID	latitude	longitude	similarity	cluster
172	Camden	Fitzrovia	172	51.518021	-0.136242	0.972802	1
299	Westminster	Marylebone (also St Marylebone)	299	51.517305	-0.147803	0.965995	1
406	Westminster	Soho	406	51.517076	-0.133397	0.956681	1
53	Camden	Bloomsbury	53	51.521478	-0.127451	0.924837	1
113	Westminster	Covent Garden	113	51.511500	-0.122095	0.912547	1
300	Westminster	Mayfair	300	51.508317	-0.148168	0.905823	1

As a cluster could be composed by areas with different similarities, 5 Nearest Neighbours (KNN) algorithm in order to calculated which cluster is the most similar to the NewYork area. KNN gives the probability for SoHo to belong to one of the top cluster identified

Best Clusters

	cluster	probability
1	1	0.424050
0	0	0.246223
2	2	0.169041
3	3	0.160686

Areas belonging to best clusters

areaID	cluster	area
172	1	Fitzrovia
299	1	Marylebone (also St Marylebone)
406	1	Soho
53	1	Bloomsbury
113	1	Covent Garden
300	1	Mayfair
400	0	Shoreditch
269	2	Kingston upon Thames
212	2	Hampton Wick
397	3	Shepherd's Bush
511	3	White City

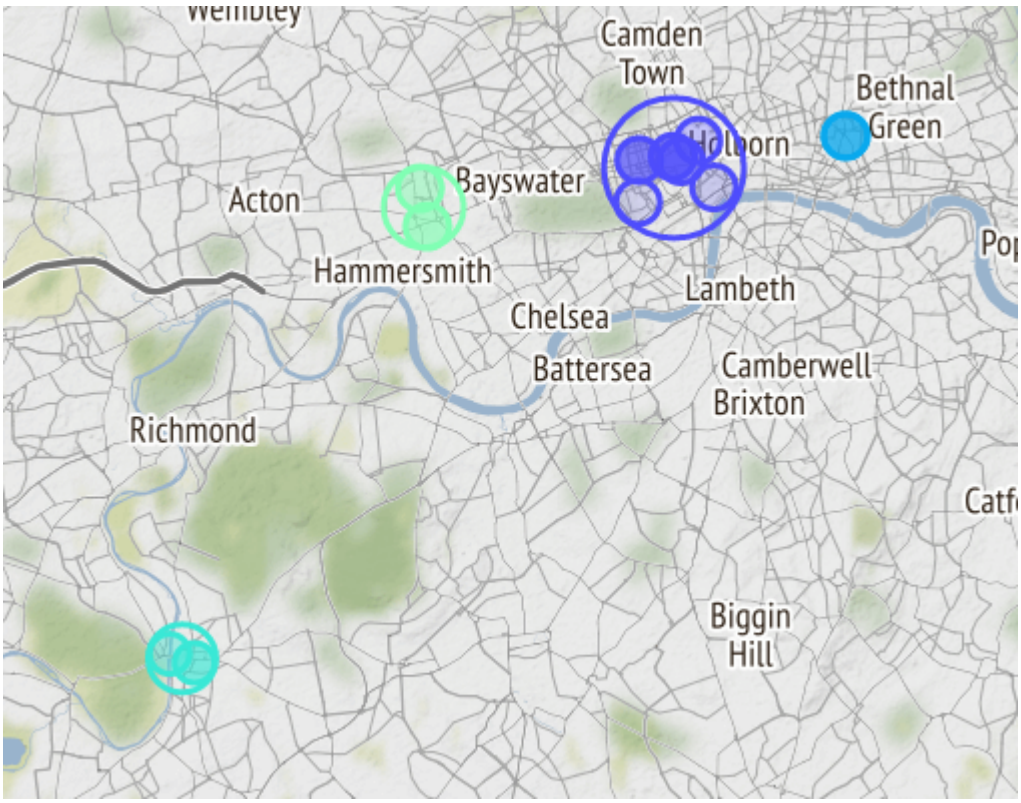
As the similarity calculation was based on the proportion of themes (as defined above) composing an area, it is possible to visualise the cluster theme composition and compare it to SoHo theme composition

Clusters' themes

	-1	1	0	2	3
Shops, Food & Drinks	0.791025	0.791164	0.824956	0.818866	0.777527
Shops, Art & Entertainment	0.286002	0.256987	0.171053	0.190239	0.295624
Residential & Outdoor sport	0.050814	-0.051587	0.046084	0.098623	-0.011683
Outdoor, Recreation & Shops	0.020506	0.078989	0.020517	0.156006	0.037745
Commuting & Shops	-0.078593	-0.101971	-0.123416	-0.078793	0.036479
Residential & Ethnic food	-0.086795	-0.065468	-0.095306	-0.173619	-0.106929
Transport	-0.097798	-0.068457	-0.104546	-0.046476	-0.055465
Nightlife & Entertainment	-0.112151	-0.033794	0.010567	-0.041894	-0.010793
Fitness, Outdoor & Travel	-0.129454	-0.169312	-0.064580	-0.297501	-0.143002
Commuting & Food	-0.159602	0.020927	-0.090477	-0.058833	-0.193290
similarity	1.000000	0.939781	0.973975	0.948802	0.951135
probability	NaN	0.424050	0.246223	0.169041	0.160686

In the table above SoHo, New york is identified with the column -1. every other column identifies a cluster. It is possible to notice how SoHo is characterised by being predominantly an area of Shops, Food & Drinks (0.79), and Shops, Art & Entertainment (0.286002), and not really a fitness location (-0.12) nor commuting and fast food (-0.15). Although the mean similarity for cluster 1 (0.939781) is lower than cluster 0 (0.973975), KNN(5) suggest that cluster 1 is a more denser area, with an higher probability for SoHo to belong to it (0.424050)

Map of London's clusters



Results and Discussion

4 main optimal zones are identified, by clustering the most similar areas of London and grouping them by proximity.

areaID	cluster	area
172	1	Fitzrovia
299	1	Marylebone (also St Marylebone)
406	1	Soho
53	1	Bloomsbury
113	1	Covent Garden
300	1	Mayfair
400	0	Shoreditch
269	2	Kingston upon Thames
212	2	Hampton Wick
397	3	Shepherd's Bush
511	3	White City

- **Cluster 0**, a cluster composed only by the areas of Shoreditch, a popular and fashionable part of London, particularly associated with the creative industries. Art galleries, bars, restaurants, media businesses are common in the area.
- **Cluster 1**, a large cluster composed by London areas belonging to the very central boroughs of Westminster and par of Camden. This cluster is dense (boundaries of the areas overlapping) and cover the biggest area in meters. This zone famous for the reputation as a major entertainment district of London . It is filled with galleries, bars, restaurants and major theatres.
- **Cluster 2**, a cluster composed by two areas belonging to the borough of Kingston Upon Thames. Although far from London's centre, Kingston is identified as a metropolitan area and is today a major retail centre, one of the biggest in the UK, receiving 18 million visitors a year.
- **Cluster 3**, a cluster composed by two areas hosting a major luxury retail centre and campus of universities

Several additional external factor could be taken in account to choose which cluster is optimal (population of the areas, percentage of tourists, proximity to tourist attraction, average cost of rental). Anyway, for this project only the previously defined internal metrics are used: mean similarity of the cluster and cluster density/areas coverage.

Clusters probability, similarity, and radius

cluster	probability	cluster		cluster	
1	0.424050	0	0.973975	0	500.000000
0	0.246223	1	0.972802	1	1676.553043
2	0.169041	2	0.962092	2	815.427496
3	0.160686	3	0.960121	3	960.810945
		Name: similarity,		Name: radius,	

Cluster 1 appears to be the optimal choice. Although it has an average lower similarity (0.93), it is composed by at least one areas with a similarity almost identical to the only area composing cluster 0 (0.97). Furthermore, as cluster 1 gives the client a more options on where to open a venue as it covers the biggest area, with a radius of 1676 meters compared to 500m radius area of cluster 0.

Conclusion

The purpose of the project was to identify the area of London that could best reproduce the characteristics and atmosphere of Soho New York.

To compare London areas to SoHo, a measure of similarity was defined.

Few assumptions were made:

- Characteristics of an area can be inferred by the characteristics of the venues present in said area.

- A theme is a specific mix of venue characteristics which are somehow related together. (e.g. the theme “Sport & Fitness” refers to all the venues related to sport and fitness, such as gyms, fitness centres, SPAs, golf courses, etc.).
- An area is defined by a combination of themes (e.g. an area could be 0.8 Sport & Fitness and 0.2 Residential)
- Similarity between two areas is calculated based on the proportion of the same themes they share.

Using geographical data and venues dataset from the Foursquare API, themes composition of all the London areas were derived, and similarities were calculated using cosine similarity.

Geographically closer areas were grouped into clusters.

A Cluster composed by the central areas of London (Fitzrovia, Marylebone, Soho, Bloomsbury, Covent Garden, Mayfair) was selected as the optimal cluster; having the optimal qualities of being composed by areas the most similar areas to SoHo and covering the biggest area in meters, giving gives the client a more options on where to open a venue.