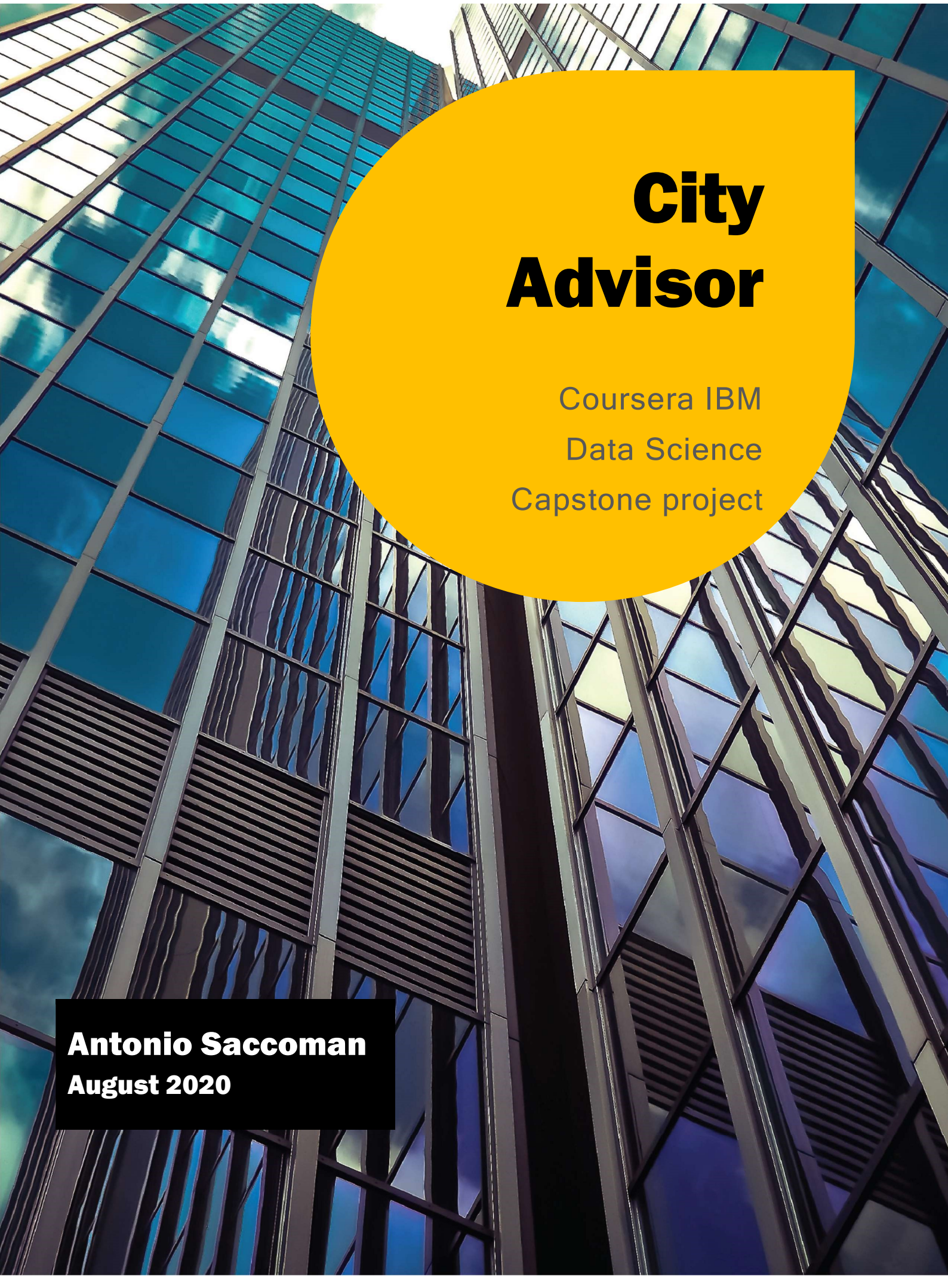


City Advisor

Coursera IBM
Data Science
Capstone project

Antonio Saccoman
August 2020

City Advisor

by Antonio Saccoman
August 2020

1. Introduction. Description of the problem and discussion of the background.

The business case for the analysis being developed here consists of advisory value for a wealthy client, who is looking for consultation over the potential purchase of a second home, in one of the largest metropolitan areas around the world.

Client is aiming at a purchase in central areas, also to maximize the number of available venues within short distance. He/she also does not express preference for a specific city, but wishes to extend the search to multiple cities in different countries. For more details about client's assumptions, please see Methodology chapter.

Data Science tools will be applied to 1) select most central areas of a given list of cities; 2) retrieve venues info for all the candidate areas; 3) analyse venues in terms of category frequency per area; 4) narrow the area search progressively via clustering and ranking functions; 5) provide final suggestions to the client in the form of best areas where to look for a property.

2. Data description and sources

Analysis as introduced above will require multiple types of data; see below.

1. Initial boroughs list for each city will be retrieved ('scraped') from the relevant city's Wikipedia page info (e.g. initial boroughs list); via BeautifulSoup.
   See example of query output below. Also, further below example of additional data such as Quality of Life index from Numbeo website.

https://it.wikipedia.org/wiki/Municipi_di_Milano#Schema_delle_zone_di_Milano

...

	#	Denominazione	Superficie(km²)	Abitanti(31.12.2018)	Densità(ab/km²)
0	Municipio 1	Centro storico	967	98 531	10 189
1	Municipio 2	Stazione Centrale, Gorla, Turro, Greco, Cresce...	1258	162 090	12 884
2	Municipio 3	Città Studi, Lambrate, Venezia	1423	144 110	10 127
3	Municipio 4	Vittoria, Forlanini	2095	161 551	7 711
4	Municipio 5	Vigentino, Chiaravalle, Gratosoglio	2987	126 089	4 221

https://www.numbeo.com/quality-of-life/rankings_current.jsp

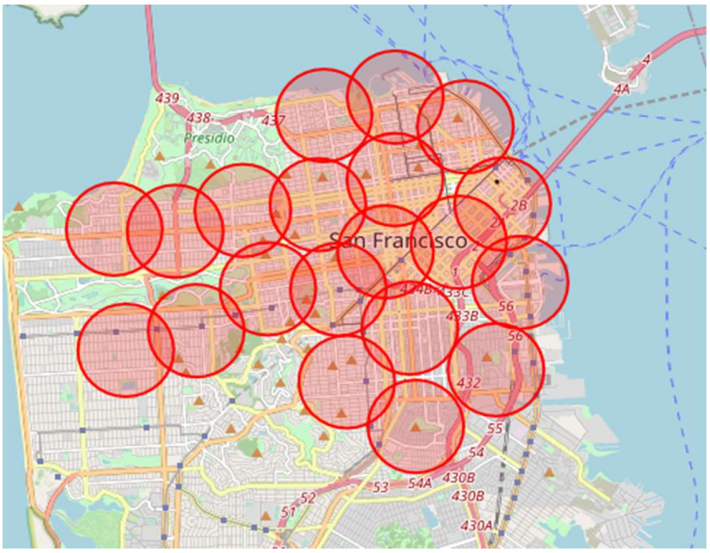
...

	Rank	City	Quality of Life Index	Purchasing Power Index	Safety Index	Health Care Index	Cost of Living Index	Property Price to Income Ratio	Traffic Commute Time Index	Pollution Index	Climate Index
0	NaN	Adelaide, Australia	207.83	108.87	71.37	80.91	71.20	4.27	24.13	17.94	94.96
1	NaN	Canberra, Australia	206.65	103.63	78.93	81.44	78.38	5.21	24.05	14.01	82.72
2	NaN	Raleigh, NC, United States	202.53	129.80	66.17	75.62	70.25	2.82	32.41	21.95	83.88
3	NaN	Wellington, New Zealand	200.21	97.75	70.07	74.90	71.02	6.36	27.74	13.66	97.68
4	NaN	Madison, WI, United States	197.34	119.50	68.89	78.89	67.15	3.43	23.73	17.51	51.64

2. City areas geographical coordinates; via Nominatim.
   See example of query output below.

	Borough	City+	Latitude	Longitude
62	Hammersmith	London, United Kingdom	51.492038	-0.223640
63	Bond street	London, United Kingdom	51.514299	-0.149002
64	Kennington	London, United Kingdom	51.488286	-0.105883
65	Battersea	London, United Kingdom	51.470793	-0.172214
66	Chelsea	London, United Kingdom	51.487542	-0.168220
67	Holborn	London, United Kingdom	51.517934	-0.119528
68	Shoreditch	London, United Kingdom	51.526669	-0.079893

3. Maps of areas in a given city; via Folium.
   See example of query output below.



4. Venues info for each area, such as geo coordinates and category; via Foursquare API. See example of query output below; also mapping via Folium further below.

	Borough	City+	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
4250	6th Arrondissement	Paris, France	48.850433	2.332951	Place Saint-Sulpice	48.850823	2.333323	Plaza
4251	6th Arrondissement	Paris, France	48.850433	2.332951	Pierre Hermé	48.851532	2.332766	Pastry Shop
4252	6th Arrondissement	Paris, France	48.850433	2.332951	Evi Evane	48.851886	2.334288	Greek Restaurant
4253	6th Arrondissement	Paris, France	48.850433	2.332951	L'Avant-Comptoir du Marché	48.851781	2.335429	Bistro
4254	6th Arrondissement	Paris, France	48.850433	2.332951	Muji	48.851272	2.335605	Miscellaneous Shop

3. Methodology

3.1 Overview, limitations and assumptions

As per introduction above, the purpose of this project is to provide consultation about best city areas to a *wealthy* client (i.e. the budget level is high, hence there are no financial limitations in the choice of property areas). The key question then is: *based on available venues in the areas, which one best fits the client's personal preferences with regards to venue categories?*

Since large cities tend to be very diverse within their boundaries, the analysis here proposed must be conducted at neighbourhood (area) level, in order to provide valuable insight to the client. Thus, the analysis looks into the central areas of every city; these are selected to be the **10-to-20** circular areas covering the city centre and the closest surrounding neighbourhoods. Radius of each area is set to **1Km**; note that 1) actual number of areas per city is not fixed, as it depends on city size, and 2) areas selection is performed to minimize overlapping, hence areas often do *not* match the administrative boroughs or districts of a given city.

For practical reasons, area search is limited here to 10 cities only. The total number of city areas is finalized as 150. The total number of venues fetched via Foursquare API is more than 14 *thousands.*

Note that the query is limited to fetch a maximum of 100 venues per area; and this is just one of the Foursquare's (queries) limitations affecting the overall results which will be discussed later on.

Every city area defined as per above is then analysed with regards to the venues located within the 1Km circular area. As an overview the areas analysis will:

1. apply clustering via machine learning, based on the areas' venues category frequency, to observe any similarity across different areas;
2. apply ranking and scoring functions, in order to identify the most desirable areas. The scoring algorithm will consist of a highly customizable function based on the assumed client's personal preferences with regards to venue categories.

Ultimately, the selection of most desirable areas to be suggested to client will be based not only on the mere *number* of venues in an area, but, more importantly, on the *categories* of the local venues. More specifically, ranking and scoring functions will be defined so that, for example, in any given neighbourhood, a park or an Italian/French restaurant will be preferred (i.e. will contribute to a higher score) compared to an auto repair shop or an airport.

Note that the assumptions over client's personal preferences are multiple: for the purpose of this project in fact the client is considered to be neutral to features such as *language, currency, tax regime, local weather, geographical location,* etc. The client is also assumed to prefer a *central* area, with as many venues as possible; other assumptions will be made about the personal preferences over venue categories (as anticipated just above).

Final product of the areas analysis will be suggestions to the client, in the form of best 10 areas given their venues contents. The results discussion will also take into consideration other parameters such as Numbeo's indices levels (e.g. quality of life) for the relevant cities; see summary table further below.

3.2 Cities and areas selection

As anticipated, cities have been selected among the largest metropolitan areas in the word, with a view to diversify as much as possible in terms of country, language, weather, etc. The total number of cities has been finalized to 10 only, for practical reasons. See below table summarizing key comparison indices (source: Numbeo website)

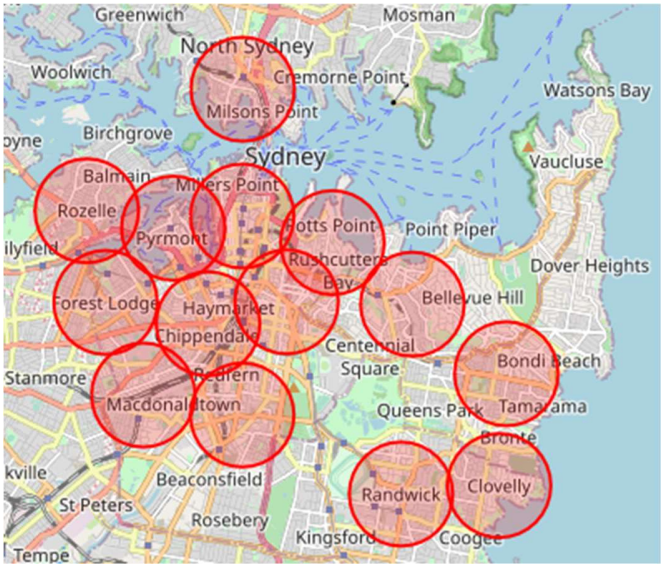
	City	Country	Price per sqm - Apartment in City Centre	Quality of Life Index	Cost of Living Index	Groceries Index	Restaurant Price Index	Health Care Index	Crime Index	Pollution Index	Climate (Weather) Index	Gross Rental Yield City Centre	Affordability Index
1	Dubai	UAE	$3,653	154.03	68.83	53.87	65.63	67.34	16.70	50.65	50.27	10.10	2.62
2	London	UK	$14,284	128.23	81.10	59.12	82.82	70.28	52.71	58.57	88.25	3.34	0.93
3	Milan	Italy	$10,214	117.27	82.37	71.36	79.45	71.57	43.19	66.06	88.12	3.16	0.87
4	New York	United States	$15,685	139.69	100.00	100.00	100.00	62.96	45.14	57.00	79.66	5.10	1.39
5	Paris	France	$14,105	117.69	91.01	89.96	77.63	78.58	51.85	64.23	88.39	2.36	0.80
6	San Francisco	United States	$13,489	161.12	92.68	91.85	87.01	66.02	55.30	46.95	97.26	5.70	1.89
7	Singapore	Singapore	$20,086	140.47	82.51	71.56	57.91	71.07	32.10	33.26	57.45	2.35	0.71
8	Sydney	Australia	$10,136	175.79	84.63	77.99	71.35	77.75	33.32	26.82	97.07	4.07	1.39
9	Tokyo	Japan	$11,274	153.26	91.75	94.46	49.48	80.13	23.40	42.25	85.26	2.65	1.05
10	Zurich	Switzerland	$13,639	196.06	136.91	139.13	121.60	74.58	16.56	17.17	81.48	3.43	2.22

Gross Rental Yield: to be considered in a potential buy-to-let scenario. *Affordability Index*: to be considered in a potential permanent residence transfer scenario. For more information about above indices description and their computation please visit www.numbeo.com.

Quality of Life index per city mapped below. Note that the colour shades (green for highest index levels, passing by yellow and down to red for lowest levels) are *relative* to the chosen 10 cities only.

As anticipated, a variable number of circular areas have been selected in each city; to cover the city centre and its immediate surroundings. Geographical coordinates for each area have been fine-tuned to 1) minimize overlapping among areas, and 2) avoid pinpointing any area *centre* into a large park location (i.e. restrict area selection to urban areas). Then venues information data have been downloaded via Foursquare API for each of the areas. See examples below from Sydney queries; areas summary table and map.

Borough	City+	Latitude	Longitude
City {*}	Sydney, Australia	-33.865423	151.207317
Woollahra - Double Bay	Sydney, Australia	-33.880873	151.243412
Bondi	Sydney, Australia	-33.893056	151.263333
Surry Hills - Darlinghurst	Sydney, Australia	-33.880456	151.216719
Elizabeth Bay House	Sydney, Australia	-33.870085	151.226444
Randwick	Sydney, Australia	-33.914121	151.241005
Rozelle	Sydney, Australia	-33.864500	151.174354
Waterloo	Sydney, Australia	-33.900276	151.207314
Chippendale - Ultimo	Sydney, Australia	-33.884304	151.200078
Clovelly	Sydney, Australia	-33.912639	151.261790
Forest Lodge	Sydney, Australia	-33.880556	151.178333
Lavender Bay	Sydney, Australia	-33.843200	151.207415
Pyrmont	Sydney, Australia	-33.867555	151.192691
Macdonaldtown	Sydney, Australia	-33.896783	151.186337



The total number of selected areas is 150, for an *average* of 15 per city. See below areas distribution.

City	Dubai	London	Milan	New York	Paris	San Francisco	Singapore	Sydney	Tokyo	Zurich	Total Areas
#Areas	16	16	15	16	15	19	12	14	16	11	150

3.3 Venues info preliminary analysis

A first overview of venues data by Foursquare: see below the 10 most frequent venue categories for each city. Note on most top spots the local cuisine restaurants as an immediate city characterization.

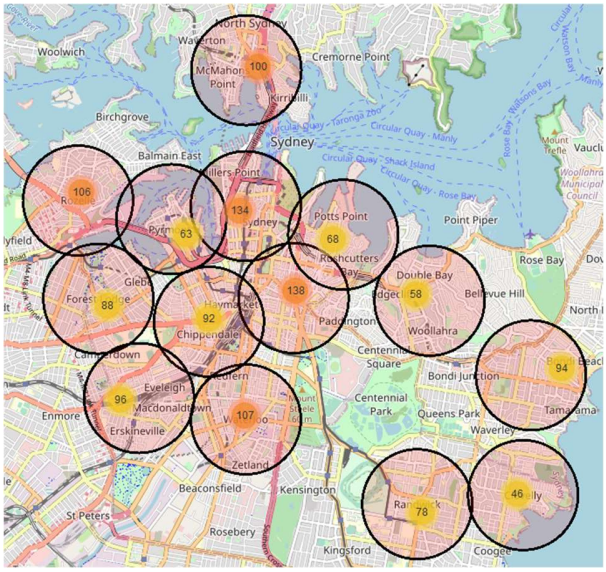
Most frequent venue categories per city

	Dubai	London	Milan	New York	Paris	San Francisco	Singapore	Sydney	Tokyo	Zurich
1	Hotel	Pub	Italian Restaurant	Coffee Shop	French Restaurant	Coffee Shop	Hotel	Café	Japanese Restaurant	Italian Restaurant
2	Café	Coffee Shop	Pizza Place	Italian Restaurant	Hotel	Park	Chinese Restaurant	Bar	Ramen Restaurant	Swiss Restaurant
3	Coffee Shop	Café	Café	Park	Italian Restaurant	Bakery	Coffee Shop	Park	Café	Café
4	Middle Eastern Restaurant	Hotel	Ice Cream Shop	American Restaurant	Bakery	Café	Café	Thai Restaurant	Coffee Shop	Hotel
5	Indian Restaurant	Italian Restaurant	Hotel	Bakery	Coffee Shop	Pizza Place	Japanese Restaurant	Pub	Sake Bar	Restaurant
6	Restaurant	Bakery	Plaza	Café	Plaza	Chinese Restaurant	Food Court	Coffee Shop	BBQ Joint	Bar
7	Asian Restaurant	Park	Japanese Restaurant	Mexican Restaurant	Bar	Sushi Restaurant	Bakery	Italian Restaurant	Italian Restaurant	Supermarket
8	Burger Joint	Gym / Fitness Center	Cocktail Bar	Gym	Japanese Restaurant	Wine Bar	Noodle House	Pizza Place	Chinese Restaurant	Bakery
9	Fast Food Restaurant	Theater	Restaurant	Pizza Place	Vietnamese Restaurant	Mexican Restaurant	Indian Restaurant	Bakery	Soba Restaurant	Tram Station
10	Gym / Fitness Center	French Restaurant	Seafood Restaurant	Grocery Store	Café	Vietnamese Restaurant	Italian Restaurant	Japanese Restaurant	Sushi Restaurant	Park

Going into more details, again for Sydney as example: see below summary table and venues cluster map. Note that some of the areas (in Sydney, and in most of the other cities too) do *not* reach the

max limit of 100; a sign that these areas are perhaps not *central* enough for the client's purposes, and are *predominantly* residential areas. Areas ranking will naturally account for this, by assigning a lower total score to the area. Hence, a fair expectation about the final areas ranking would be that all the top areas display a venue count of 100, or very close to 100.

		Venue
Borough	City+	
Bondi	Sydney, Australia	94
Chippendale - Ultimo	Sydney, Australia	100
City (*)	Sydney, Australia	100
Clovelly	Sydney, Australia	48
Elizabeth Bay House	Sydney, Australia	100
Forest Lodge	Sydney, Australia	100
Lavender Bay	Sydney, Australia	100
Macdonaldtown	Sydney, Australia	100
Pyrmont	Sydney, Australia	100
Randwick	Sydney, Australia	76
Rozelle	Sydney, Australia	92
Surry Hills - Darlinghurst	Sydney, Australia	100
Waterloo	Sydney, Australia	100
Woollahra - Double Bay	Sydney, Australia	58



A practical preliminary way to examine venue distributions in the cities is to define venues *macro*-categories. They will be called here **Generic Venue Category** (**GVC**); and will gather the 5 most common venue categories (broadly defined). Venue exploration will be greatly facilitated, considering that the unique venue categories count will drop from the original 500+ to just 5. See GVC list below, and their most frequent components.

1. *Arts / Culture;*
2. *Bars / Pubs / Cafés / Coffee Shops;*
3. *Parks / Gardens / Outdoors;*
4. *Restaurants / Food places;*
5. *Shops (excluding Coffee Shop) / Stores.*

Most frequent venue categories per GVC

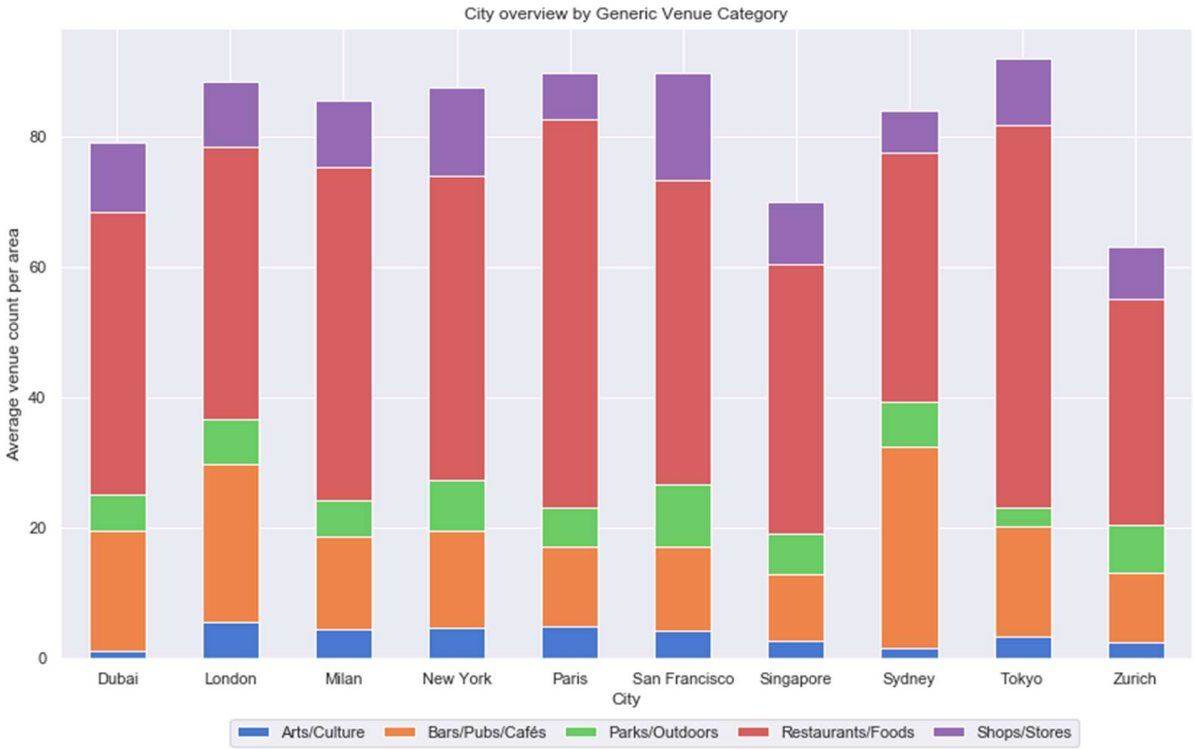
	Arts/Culture	Bars/Pubs/Cafés	Parks/Outdoors	Restaurants/Foods	Shops/Stores
1	Theater	Café	Park	Italian Restaurant	Supermarket
2	Art Gallery	Coffee Shop	Plaza	French Restaurant	Grocery Store
3	Art Museum	Bar	Garden	Bakery	Bookstore
4	Museum	Pub	Playground	Japanese Restaurant	Clothing Store
5	Movie Theater	Cocktail Bar	Pool	Pizza Place	Spa
6	Historic Site	Wine Bar	Scenic Lookout	Chinese Restaurant	Boutique
7	Performing Arts Venue	Juice Bar	Trail	Restaurant	Convenience Store
8	Concert Hall	Sake Bar	Beach	Ice Cream Shop	Cosmetics Shop
9	Music Venue	Beer Bar	Beer Garden	Indian Restaurant	Shopping Mall
10	Monument / Landmark	Hotel Bar	Dog Run	Thai Restaurant	Gift Shop

PLEASE NOTE: GVC total venues figure per area may not add up to total venues for the area; as GVC is by construction a *subset* of the Foursquare's venue categories. Total venues count under GVC is 12,614, i.e. GVC is close to a 90% coverage of the 14,241 total venues.

See below summary table per city with GVC count

Generic Venue Category City	Arts/Culture	Bars/Pubs/Cafés	Parks/Outdoors	Restaurants/Foods	Shops/Stores	Total Venues
Dubai	20	295	88	693	173	1269
London	89	387	110	667	161	1414
Milan	67	212	86	766	154	1285
New York	75	239	123	748	217	1402
Paris	73	185	89	892	110	1349
San Francisco	80	246	181	890	313	1710
Singapore	32	123	76	496	114	841
Sydney	23	431	96	537	89	1176
Tokyo	54	268	45	942	164	1473
Zurich	28	118	79	382	88	695
Total Venues	541	2504	973	7013	1583	12614

A better view of the same data is found below, after *normalizatio*n: since number of areas per city is not constant, the below histogram shows the *average* venue count per area under GVC.



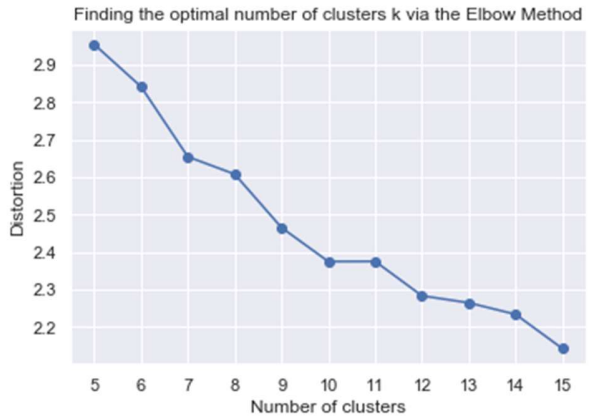
A few quick observations:

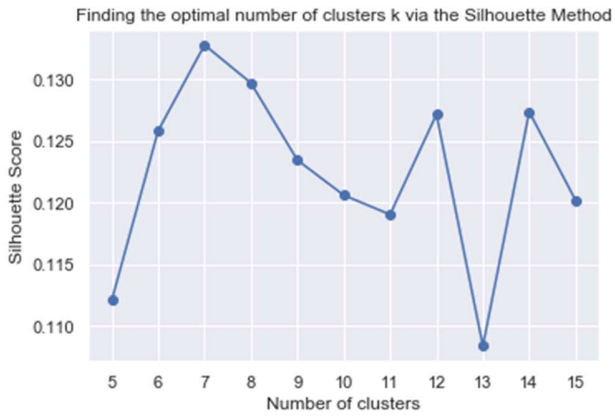
- ✓ **Sydney** dominates from the point of view of **Bars** et simil venues, closely followed by **London**;
- ✓ **Paris** and **Tokyo** lead for **Restaurants** et simil, with **Milan** just behind;
- ✓ **San Francisco** tops the other cities for both **Parks** and **Shops** GVCs;
- ✓ **London** shows the highest **Arts** et simil average per area, then **Paris** and **New York**;
- ✓ **Singapore** and **Zurich** do not seem to be at the same level as other cities in terms of overall number of venues per area.

3.4 Areas clustering

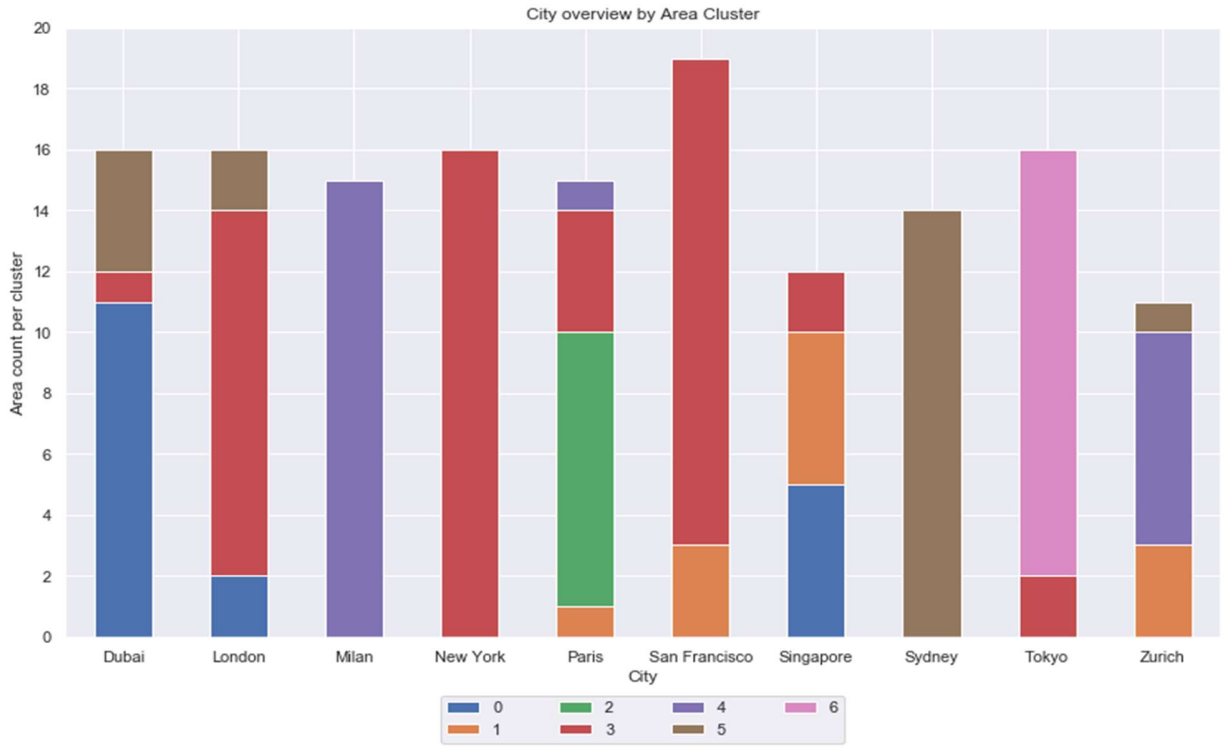
K-Means clustering is performed to help identifying common traits across the 150 different areas.

Firstly, an attempt at finding optimal number of clusters: Elbow method does not indicate a clear figure, while Silhouette method looks more helpful (graphs below).





Optimal number of clusters is set at **7**. See clustering results and observations below.



Examination of clustering results yields the following learning points:

- **Cluster 0** (blue)**:** composed prevalently of hotels, cafés, coffee shops, Middle Eastern and Indian restaurants. **Dubai** is mostly represented by this cluster; a good part of **Singapore** too. Total areas: 18.

- **Cluster 1** (orange): displays multiple types of Asian restaurants (Chinese, Thai, Vietnamese, Korean, etc); some cafés, some bakeries. The remaining part of **Singapore** is mostly represented by this cluster; some parts of **San Francisco** and **Zurich** are to be mentioned too. Total areas: 12.

- **Cluster 2** (green)**:** mainly shows French restaurants, hotels, Italian restaurants, bakeries, bars. **Paris** (most of its areas) is the only city allocated to this cluster. Total areas: 9.

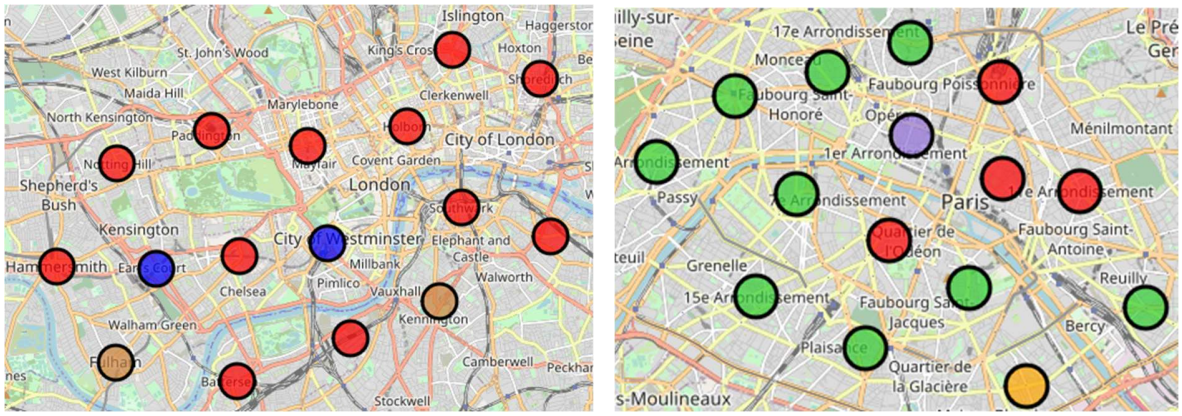
- **Cluster 3** (red)**:** this seems to be the "*anything else*" cluster, as compared to the others; there are no clear similarities among areas. This cluster displays a wide variety of restaurants of different cuisines, then cafes, coffee shops, bars, pubs, bakeries, parks and gardens. **New York** (in its entirety), then **San Francisco** and **London** are the main representatives of this "*diversity*" cluster. Total areas: 53.

- **Cluster 4** (purple)**:** most frequent venues are Italian restaurants and pizza places, hotels, cafes, bars, plazas, ice creams. **Milan** (in its entirety), then **Zurich** are the main representatives of this cluster. Total areas: 23.

- **Cluster 5** (brown)**:** predominantly includes cafés, coffee shops, bars, pubs, then some Asian and Italian restaurants, parks. **Sydney** (in its entirety) is the main representative of this cluster. Total areas: 21.

- **Cluster 6** (pink)**:** characterized mostly by the presence of Japanese restaurants (various local types), sake bars, BBQ joints, and some Italian restaurants. **Tokyo** (almost entirely) is the main representative of this cluster. Total areas: 14.

See below a couple of clusters city maps for reference; London and Paris.



3.5 Areas preliminary ranking

Each of the 500+ unique venue categories from Foursquare API is assigned with a **Tier** representing the client's degree of appeal towards the given venue category: Tier-3 is assigned to any venue category client is *indifferent* to; Tier-2 is assigned to venue category that the client is assumed to consider *good-to-have* in the neighbourhood; and finally, Tier-1 is assigned to client's assumed most preferred venue categories, i.e. *must-have* in the area.

PLEASE NOTE: **Tiers** are the features under which areas search and ranking are *customised* to each client, based on his/her personal preferences; as opposed to **GVC**, which is *generic* venue flag (macro-category) *independent* from client preferences.

For the purpose of this project, Tiers are customized to a *generic client*. E.g. the client is assumed to favour restaurants and shops (Tier-2), but assigns even higher value to parks and outdoors in general, plus cultural spots like museums and theatres (Tier-1). As *direct* result of these assumptions, see below table of most frequent Tier-1 / 2 / 3 categories.

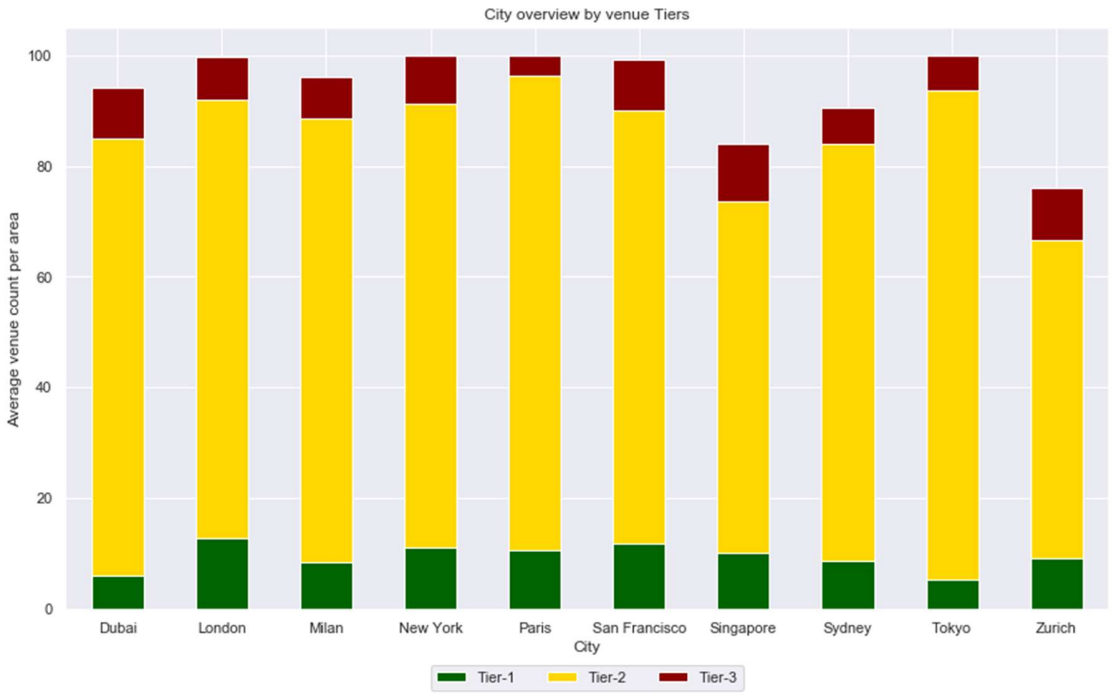
Most frequent venue categories per Tier

	Tier-1	Tier-2	Tier-3
1	Park	Café	Sandwich Place
2	Plaza	Coffee Shop	Burger Joint
3	Theater	Italian Restaurant	Vegetarian / Vegan Restaurant
4	Garden	Hotel	Yoga Studio
5	Deli / Bodega	French Restaurant	Fast Food Restaurant
6	Art Museum	Bakery	Nightclub
7	Food Court	Japanese Restaurant	Hostel
8	Museum	Pizza Place	Tram Station
9	Farmers Market	Bar	Salad Place
10	Pool	Chinese Restaurant	Bus Station

It looks clear now that Tiers may be - to some degree - proxied to GVCs: Tier-1 (i.e. most desirable venue categories) is a mix of most Parks/Outdoors venues and most Art/Culture venues; while Tier-2 is a mix of most Restaurants/Foods and most Bars/Cafes venues.

See below overview of Tiers per city: firstly, a table with absolute venues numbers, then histogram of Tiers distribution per city as *average* venue count per area.

Venue Tier	Tier-1	Tier-2	Tier-3	Total Venues
City				
Dubai	94	1264	149	1507
London	205	1266	124	1595
Milan	128	1202	113	1443
New York	180	1282	138	1600
Paris	159	1285	56	1500
San Francisco	224	1485	176	1885
Singapore	123	761	124	1008
Sydney	122	1055	91	1268
Tokyo	85	1415	100	1600
Zurich	101	630	104	835
Total Venues	1421	11645	1175	14241



A preliminary ranking starts to appear, at least from a **city** point of view:

- ✓ **Tier-1** venues are on average most frequent in **London**, followed by **San Francisco**;
- ✓ Aggregate of **Tier-1 & Tier-2** venues is most frequent in **Paris**, then in **Tokyo**.

For a preliminary ranking of **areas** see below table instead. Areas are ranked by highest Tier-1 venues count; then by Tier-2 count, then Tier-3.

Position	Area	City	Tier-1	Tier-2	Tier-3	Total Venues
1	Southwark	London	27	65	8	100
2	Marina South	Singapore	25	54	13	92
3	Buena Vista	San Francisco	24	62	14	100
4	East Harlem - 104th St	New York	23	72	5	100
5	Invalides - La Tour-Maubourg	Paris	22	76	2	100
6	Anderson Bridge	Singapore	22	65	13	100
7	Riesbach	Zurich	21	70	6	97
8	Place Monge	Paris	20	80	0	100
9	Lincoln Square - 66 St	New York	20	75	5	100
10	National Library (Richelieu)	Paris	19	79	2	100

Southwark in London is the top ranked area thanks to highest number of Tier-1 venues; looking at the area venues, its high Tier-1 count is mostly due to the high concentration of Arts/Culture venues (14/27) - mostly due to Theatres (7/27) and Museums (4/27); plus Outdoors (7/27), and Food places other than regular restaurants (6/27).

Southwark , Tier-1 venues composition

Generic Venue Category	Venue Category	Venue
Arts/Culture	Art Museum	3
	Concert Hall	1
	History Museum	1
	Performing Arts Venue	2
	Theater	7
Parks/Outdoors	Park	2
	Pedestrian Plaza	1
	Plaza	1
	Scenic Lookout	3
Restaurants/Foods	Deli / Bodega	1
	Farmers Market	2
	Street Food Gathering	3
Total Venues		27

3.6 Areas final scoring

Final scoring is based on *points* assigned for each Tier. After a brief set of simulations, total computation follows a linear method, but points to venues are allocated in a *non-linear* way: a Tier-2 venue is valued 3 times as much as a Tier-3 venue, while a Tier-1 venue is valued *more than twice* a Tier-2 venue. Hence the below points system.

Venue Tiers	Points per Venue
Tier-3	**1**
Tier-2	**3**
Tier-1	**7**

See below table for resulting scores.

Position	Area	City	Tier-1	Tier-2	Tier-3	Total Venues	Total Score
1	Southwark	London	27	65	8	100	392
2	Invalides - La Tour-Maubourg	Paris	22	76	2	100	384
3	East Harlem - 104th St	New York	23	72	5	100	382
4	Place Monge	Paris	20	80	0	100	380
5	National Library (Richelieu)	Paris	19	79	2	100	372
6	Lincoln Square - 66 St	New York	20	75	5	100	370
7	Buena Vista	San Francisco	24	62	14	100	368
8	West Bermondsey	London	16	84	0	100	364
9	Riesbach	Zurich	21	70	6	97	363
10	Anderson Bridge	Singapore	22	65	13	100	362

4. Results and Discussion

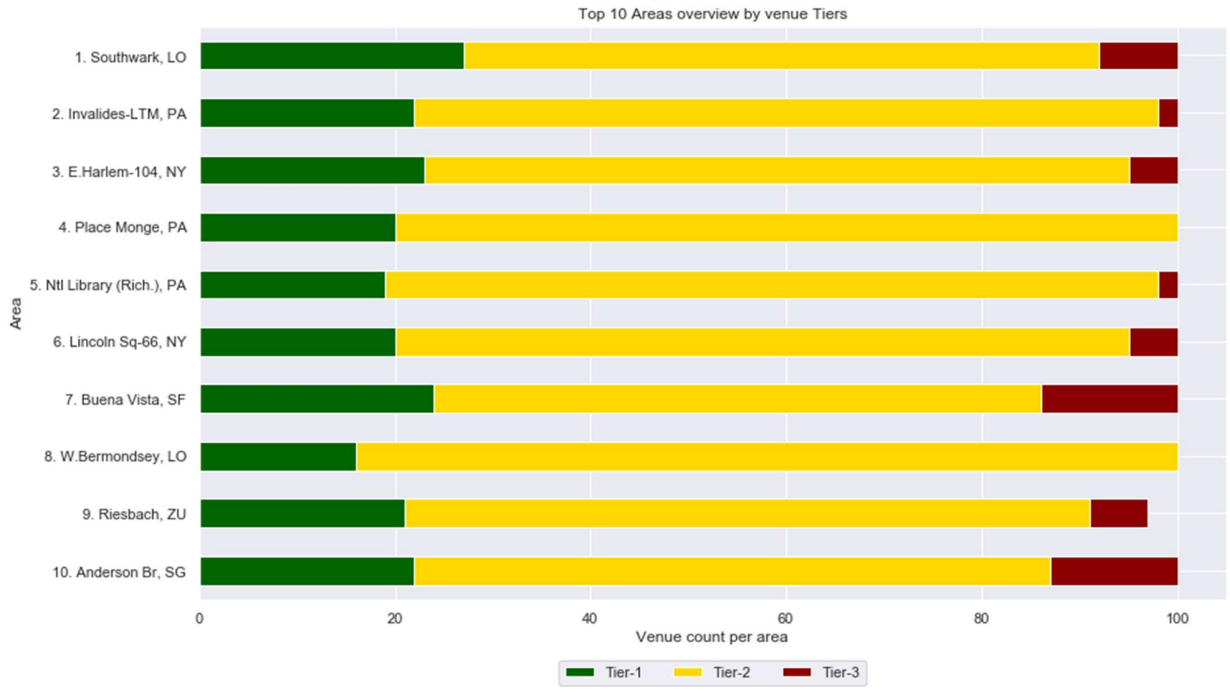
Southwark in London is confirmed as best area out of 150 selected world-wide, according to the described methodology; **Invalides** area in Paris and **East Harlem** in New York are the second and third best areas, respectively.

Before commenting further on the results, please find below relevant summary tables and graphs.

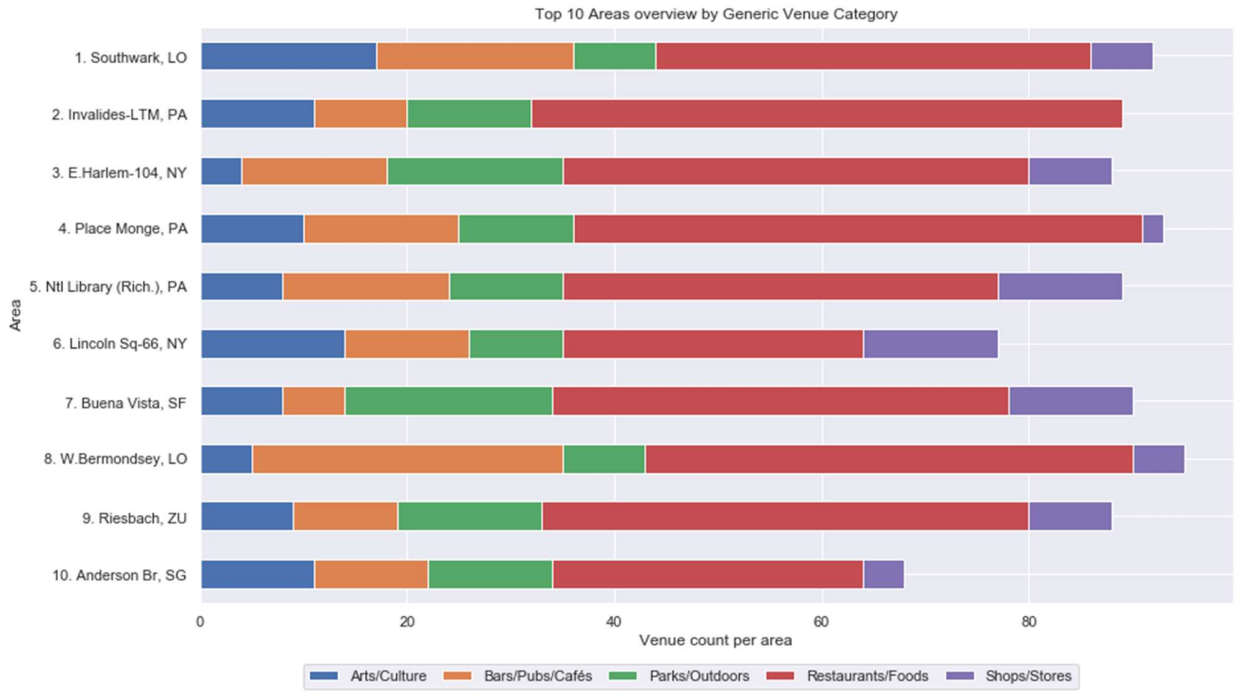
Table of top 10 areas inclusive of both **Tier** counts and **GVC** counts.

Area	City	Tier-1	Tier-2	Tier-3	Total Venues	Total Score	Arts/Culture	Bars/Pubs/Cafés	Parks/Outdoors	Restaurants/Foods	Shops/Stores
Southwark	London	27	65	8	100	392	17	19	8	42	6
Invalides - La Tour-Maubourg	Paris	22	76	2	100	384	11	9	12	57	0
East Harlem - 104th St	New York	23	72	5	100	382	4	14	17	45	8
Place Monge	Paris	20	80	0	100	380	10	15	11	55	2
National Library (Richelieu)	Paris	19	79	2	100	372	8	16	11	42	12
Lincoln Square - 66 St	New York	20	75	5	100	370	14	12	9	29	13
Buena Vista	San Francisco	24	62	14	100	368	8	6	20	44	12
West Bermondsey	London	16	84	0	100	364	5	30	8	47	5
Riesbach	Zurich	21	70	6	97	363	9	10	14	47	8
Anderson Bridge	Singapore	22	65	13	100	362	11	11	12	30	4

Top 10 areas view by **Tiers**.



Top 10 areas view by **GVC**.



Top 10 areas **most frequent venue categories**; with K-Means cluster.

Area	City	Cluster#	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Southwark	London	3	Pub	Theater	Coffee Shop	Hotel	Italian Restaurant
Invalides - La Tour-Maubourg	Paris	2	French Restaurant	Hotel	Italian Restaurant	Plaza	Historic Site
East Harlem - 104th St	New York	3	Mexican Restaurant	Park	Gym	Coffee Shop	Deli / Bodega
Place Monge	Paris	2	French Restaurant	Coffee Shop	Italian Restaurant	Café	Hotel
National Library (Richelieu)	Paris	4	Hotel	Plaza	Japanese Restaurant	Historic Site	French Restaurant
Lincoln Square - 66 St	New York	3	Gym / Fitness Center	Italian Restaurant	Jazz Club	Plaza	Gym
Buena Vista	San Francisco	3	Seafood Restaurant	Park	Chocolate Shop	Historic Site	Bike Rental / Bike Share
West Bermondsey	London	3	Pub	Coffee Shop	Brewery	Italian Restaurant	Park
Riesbach	Zurich	4	Italian Restaurant	Swiss Restaurant	Café	Restaurant	Hotel
Anderson Bridge	Singapore	0	Hotel	Gym / Fitness Center	Waterfront	Event Space	Cocktail Bar

Some observations on final results:

1. Final top 10 areas list is just **30** points wide in terms of score; meaning that just above 4 Tier-1 venues - or just 10 Tier-2 venues – make the difference between the 1st and the 10th area.
2. **Paris** is the most frequent city in the top 10 areas, with 3 areas. Then both London and New York have 2 areas each.
3. Only **6** cities out of 10 are represented in the top 10 area ranking. The excluded cities are: Dubai, Milan, Sydney and Tokyo.
4. The most frequent K-Means **cluster** (5 areas out of 10) is the number **3**. Diversity of venue categories seems to play some positive role in areas searching. Although cluster 3 is the largest cluster, including 53 areas out of 150; for a fair expectation of 3-to-4 areas in the top 10.
5. Arts & Culture venues (14) are the driving force behind **Southwark's** 1st place. Yet questions must be asked to the client, as to determine whether (or how much more) he/she prefers going to theatre or to a concert (assumed as a weekly-to-monthly activity) as opposed to walks in parks/gardens et simil (an activity assumed to be daily or close to it). Southwark might be the best London area for theatres et simil, yet there exist other areas with much higher presence of parks and gardens. A hint here for any future methodology development: increase number/granularity of Tiers. To complete the area Tier-1 composition: Parks & Outdoors venues (7), Restaurants & Foods venues (6).
6. In addition to Arts & Culture venues (10) and Parks & Outdoors venues (12) (both mostly due to proximity to Les Invalides complex), the actual Restaurants (45, under Tier-2) seem to be one of the reasons for **Invalides** 2nd place (see GVC graph above). Yet the lack of *cuisine diversification* must be highlighted, as compared to other areas such as London's, New York's, and San Francisco's (remember K-Means cluster 3): most of actual restaurants under Invalides Tier-2 are in fact French et simil (78%), and only 6 other cuisine types can be found. If a client values cuisine diversification more than the mere number of restaurants (as one would expect) then Invalides

would be ranked much lower. A hint here for any future methodology development: add cuisine diversification index to scoring function.

7. Parks & Outdoors concentration helps **East Harlem** area to the 3rd place. There are 14 venues under this GVC; most of them due to proximity to one large park: Central Park. Plus, similarly to Southwark, the area includes Tier-1 food places like Deli/Bodega and Farmers Market (5 in total, vs Southwark's 6). Another positive feature of East Harlem is the cuisine diversification (as opposed to Invalides): various cuisines are available, summing up to 15 different ones, vs Southwark's 14 and Invalides' 7 only. To complete the area Tier-1 composition: Arts/Culture venues (4).
8. As per Tier graph above, highest count of aggregate (Tier-1+Tier-2) venues is reached by **Place Monge** in Paris (4th place) and by **West Bermondsey** in London (8th place).
9. As per GVC graph above, Parks & Outdoors are the most frequent in **Buena Vista** in San Francisco (20), thanks to combination of Fort Mason / Great Meadow park and coastal location; while Bars et simil are the most frequent in **West Bermondsey** in London (30).
10. Multiple parameters other than available venues would be considered by the client before property purchase (among which stamp duty tax). See for reference the Numbeo table of **city indices** in par. 3.2. Key index for client would obviously be the city's average property Price per sqm; other parameters too would be relevant, such as Quality of Life, Cost of Living, Weather, etc. Looking at final results, the 3 top cities (London, Paris, New York) are among the most expensive, but do not show extreme values in terms of **Price per sqm**; although New York's property is on average more expensive than London's or Paris'. Considering the **Quality of Life** index (higher is better) and the **Cost of Living** index (lower is better), London shows the best balance (highest ratio) between the 2 indices: ~128/81, vs Paris ~118/91 and New York ~140/100. Furthermore, depending on client's age and personal preferences, other indicators such as **Healthcare** index and **Weather** index may well be pondered upon.

Further observations on limitations *affecting* results.
All results should be seen as *indicative* only, due to Foursquare's venues data *limitations*:

a) **Randomness in venues selection** (i.e. 'sampling'). As the vast majority of areas reaches the max limit of 100 venues per query (Foursquare's default for the charge-free access?), most of them, if not all, must be assumed to have in reality *more* than 100 venues; the implication being that the 100 venues for the area are selected *randomly* by the query. Evidence that some venues - such as parks and historic sites for example (NB Tier-1) - are being left out is easily found by looking at Folium maps, and noticing that a venue, despite size and/or historical relevance, is not assigned with the expected marker.
   The higher the number of an area's venues (above 100), the higher the expected randomness in selecting only 100 venues ('sampling'). And the likelihood that such randomness preserves the area relative proportions of Tier-1, Tier-2, Tier-3 venues is fairly low, so selection randomness might well have a medium to high impact on shaping the final ranking. Hence one of the priorities in any future development is increasing the Foursquare's venues max limit; this may come with a cost (non-free access).
b) With particular reference to **Parks** (NB Tier-1 venues), the venue's **size** aspect is not considered in Foursquare's venue info. A client's standard preference would be assumed to be for *larger* parks as opposed to smaller ones. Thus, final ranking would likely be affected by this. Would it change much? Potentially yes; think of small parks - such as some in Southwark - compared to Hyde Park in London, Central Park in New York, Presidio or Golden Gate park in San Francisco.
c) **Gardens / trails / tracks** *within a park* (also Tier-1 venues): these are part of the "park" complex, then they - altogether with the park - should in theory count as *one* Tier-1 only. See for example

Invalides in Paris or areas close to Central Park in New York. Tentative manual adjustments would reduce Tier-1 venues for a few areas adjacent to parks → would this change the final top 3 areas? Yes, potentially lowering ranking for both Invalides and East Harlem. *However*, point b above may be seen as counter-balancing point c here: the more features within a park, the higher likelihood of a larger park, so, to some degree, observation c *offsets* observation b, as the number of *internal* gardens/trails may be seen as a *proxy* for a park size.

d) Again, on **parks' size**: a park is counted as venue belonging to an area *only if* a very specific point inside the park (usually the very centre of the park) falls into the area circle. Hence an area attached to a *large* park may *not* be assigned with the park as venue, simply because the area does not manage to contain the Foursquare's park centre coordinates. Though also *proximity* must be assumed to have value to the client! Ideally then, a park venue should be allocated to all areas which include an **entrance** to the park, not just areas including the *centre* of the park. Tentative manual adjustments would add park as Tier-1 venue to a number of areas (e.g. Outer Richmond in San Francisco, Bond Street in London). Would this change the final top 3 areas? Potentially yes, as these adjustments would be mostly beneficial to some areas in San Francisco, London and New York.

5. Conclusions

The purpose of this project was to identify and suggest best city areas where to purchase property, given a scoring function based on 1) numbers and categories of venues available in an area, and 2) the venues degrees of appeal to a generic property purchaser (the "client"). After selecting a range of cities (10) and central areas (150), and then collecting venues information for each of the areas, this project has performed 1) exploratory analysis, by means of 5 venue macro-categories, 2) K-Means clustering, to identify common traits across different areas, 3) areas ranking, by count of most appealing venue categories, and finally 4) areas scoring, by points system, to narrow down the search to the best scoring (i.e. most desirable) 10 areas.

In general, the top areas identified here are as expected among the most expensive in terms of average property Price per sqm, although not the most expensive (Singapore). Other indicators place them in a nice mid spot where Quality of Life is medium-to-high, while Cost of Living is not too high; while Weather extremes are rare.

Finally, it is worth reminding again that final areas ranking here, while reasonable, is *indicative* only, as affected by several (venue data) issues, and by generic assumptions on client's preferences. Thus, to fully pursue this project's recommendation purposes: 1) venue data issues would have to be addressed and adjusted for, and 2) assumptions on client's preferences would have to be fine-tuned to a specific client. Final property purchase decisions would then be up to the client, and would likely include many additional factors; hence the choice here to collect and display a summary of cities' economic and social indicators sourced from Numbeo website (see par. 3.2).

6. Further developments

A few suggestions for further developments:

- set a higher max limit on Foursquare's API queries (e.g. 200+, to minimize *randomness* in venues selection);
- add more cities added to the search;
- include more areas per city, and/or set smaller area radius (walking distance);

- account/adjust for Foursquare's venue data limitations (see end of par. 4);
- define more venue Tiers (5? 10?), to enhance customization;
- include *negative* weights for *undesired* venues if any ("negative Tiering");
- add *cuisine diversification* index to scoring;
- add Numbeo's indicators to scoring;
- experiment non-linear scoring functions (e.g. exponential), and/or alternative scoring point systems;
- consider actual distance from the city's most iconic places or landscapes; especially when such items are not in the top-ranking areas.