Clustering of the world most visited cities

How to recommend the next best travel destination to clients

The aim of this paper is to help travel agencies in grouping typical destinations based on the most common types of venues found in each location. This will allow a more informed recommendation for their clients seeking advice on where to go next.

Fabrice PETITFRERE

4/22/2020

Contents

PART 1 – INTRODUCTION

Introduction

Do you know what it feels to be thrown in a foreign city and be miles away from home, losing all sense of familiarity? That's what is so particular with our planet is that there is such a variety of cultures and environments. Sometimes however, even far from home, you feel a sense of familiarity with the place you're staying and/or visiting and can't yet tell why this is. There are obviously similarities among all the other differences between towns across all the continents. These similarities are sometimes noticeable enough to make you feel in a familiar place even though environment, culture or language is different. Whether you like a place or not is sometimes linked to these little things, hard to notice. What if we are able to cluster the world cities according to the number and the categories of venues found in each location?

Business problem

Travel agencies role is (more often than we can think) to advise their customers on best holiday locations. Not all people decide of their next holiday plans before they walked into a travel agency. Selecting a destination for their customers required to understand their preferences and presumably know of places they visited and which they liked. Destinations are sometimes quite hard to compare and it would be a good idea to have at least the most visited cities arranged in a few similar groups. It would enable travel agents to say: "Oh you've been to Milan last year, I would suggest you try Lisbon if you haven't been as it has some similarities you may like!" (obviously if we can group Milan and Lisbon together but the analysis will tell us!).

Business audience

This work can be helpful to travel agents in need of information to advise their customers.

PART 2 – DATA

Data sources

To undertake this work, we will first need a pertinent list of destinations to investigate, their accurate locations with GPS coordinates, some information about each location including data about different types of venues that can be found there.

List of destinations

Since we are investigating destinations for tourists, we will use a list of the most visited cities in the world. This list of the top 100 most visited cities is ranked by the number of international visitors, including all international arrivals by land, sea or plane, for tourist of business purposes. More precisely, we will the Euromonitor count which counts a visitor as any person visiting a city another country for at least 24 hours, for a period not exceeding 12 months, and staying in paid or unpaid, collective or private accommodation. Each arrival is counted separately and includes people travelling more than once a year and people visiting several cities during one trip. Useful fields to retrieve contains: city name, country and number of arrivals in 2018.

Source: https://en.wikipedia.org/wiki/List_of_cities_by_international_visitors

World city population

In order to conduct some exploratory analysis and look at the data by inhabitant, we also need to import the population of each city. This information can be found on the link below.

Source: https://worldpopulationreview.com/world-cities/.

GPS coordinates

To find the GPS coordinates of each city location, we will use Geopy, a popular python client for geocoding web services. This will return latitude and longitude for each location.

Source: https://geopy.readthedocs.io/en/stable/#

Venues data

To explore each location, we will then use the Foursquare location data and explore each location for the most popular venue categories. This will be described in more details further on but we will only use the main venue categories which consist of 10 distinct categories.

Source : https://developer.foursquare.com/docs/build-with-foursquare/categories

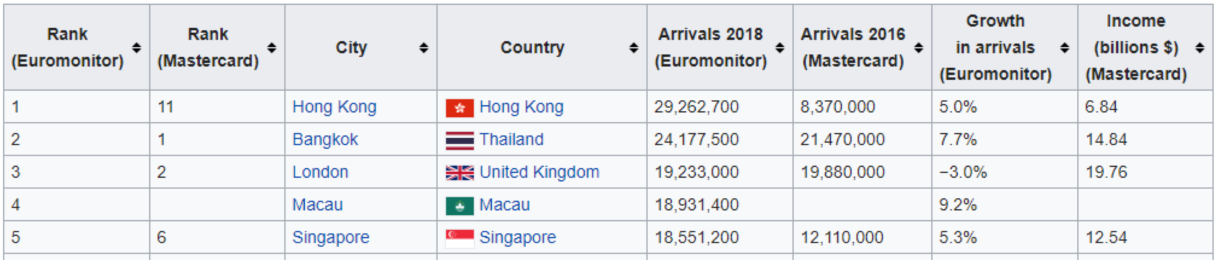
PART 3 – METHODOLOGY

Data fetching and cleaning

The first step of the work consisted in fetching and cleaning all the necessary data to undertake the analysis.

List of destinations

As described in the previous section, a list of 100 cities was used in the analysis and the original table can be found on Wikipedia (link above). The display of the table is shown on Figure 1 below.

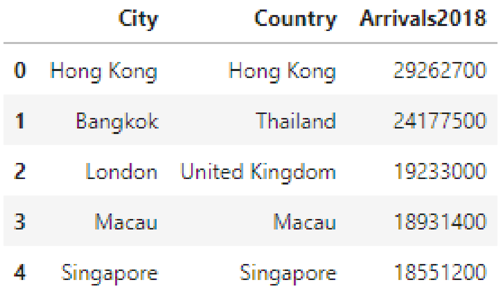


Rank (Euromonitor)	Rank (Mastercard)	City	Country	Arrivals 2018 (Euromonitor)	Arrivals 2016 (Mastercard)	Growth in arrivals (Euromonitor)	Income (billions $) (Mastercard)
1	11	Hong Kong	Hong Kong	29,262,700	8,370,000	5.0%	6.84
2	1	Bangkok	Thailand	24,177,500	21,470,000	7.7%	14.84
3	2	London	United Kingdom	19,233,000	19,880,000	−3.0%	19.76
4		Macau	Macau	18,931,400		9.2%	
5	6	Singapore	Singapore	18,551,200	12,110,000	5.3%	12.54

Figure 1 - Most visited cities in the world – Source: Wikipedia

We are only interested in retrieve the following fields: ‘City’, ‘Country’ and ‘Arrivals 2018 (Euromonitor). The table includes more than 100 cities but beyond the 100th city, the arrivals are based on the Mastercard Index. In order to keep the arrivals number to one source only, we select the cities ranked from 1 to 100 in the ‘Rank (Euromonitor)’ column.

The BeautifulSoup Python library was used to scrape the table data from the webpage. The scraping was made easier as there is only one table of class ‘wikitable’ on the webpage. We can create lists for the fields ‘City’, ‘Country’ and ‘Arrivals2018’. After some text cleaning and the conversion of the arrival numbers in integer, we can create a dataframe with the 3 column series.



	City	Country	Arrivals2018
0	Hong Kong	Hong Kong	29262700
1	Bangkok	Thailand	24177500
2	London	United Kingdom	19233000
3	Macau	Macau	18931400
4	Singapore	Singapore	18551200

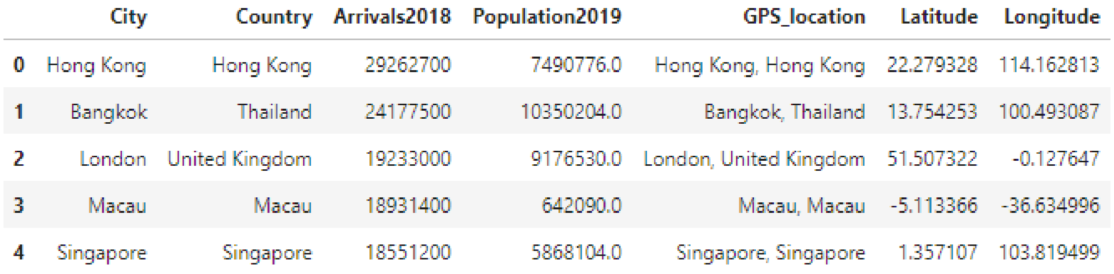
Figure 2 – Top100 cities dataframe (first 5 rows)

World city population

City population was extracted from the Worldpopulationreview website (link above). City populations could not be found for 9 of them. Considering that this feature was not crucial in the analysis, we will not consider these 9 cities in the analysis including populations.

GPS coordinates

Geopy was used to find the city latitudes and longitudes. As the API does not always return answers to requests, a 'while' loop was used to make sure all latitudes and longitudes were returned for the 100 cities in our dataframe.



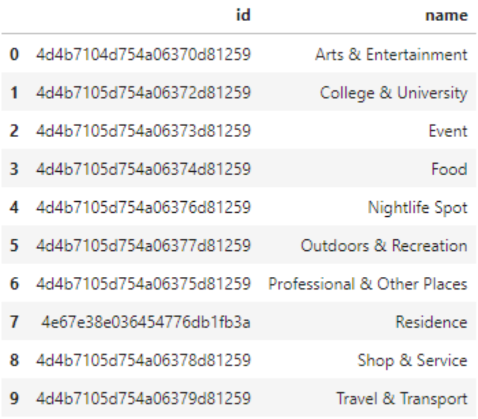
	City	Country	Arrivals2018	Population2019	GPS_location	Latitude	Longitude
0	Hong Kong	Hong Kong	29262700	7490776.0	Hong Kong, Hong Kong	22.279328	114.162813
1	Bangkok	Thailand	24177500	10350204.0	Bangkok, Thailand	13.754253	100.493087
2	London	United Kingdom	19233000	9176530.0	London, United Kingdom	51.507322	-0.127647
3	Macau	Macau	18931400	642090.0	Macau, Macau	-5.113366	-36.634996
4	Singapore	Singapore	18551200	5868104.0	Singapore, Singapore	1.357107	103.819499

Figure 3 – Top 100 cities with selection of features (first 5 rows)

Venues data

The data about venues in each city is obtained from Foursquare location data. Foursquare classifies each venue into a category tree with 3 levels. The higher level of this tree consists of 10 categories. The analysis was conducted at this level in order to reduce the number of parameters and making sure we had venues categories that were relatively consistent in all cities. Foursquare contains some particular sub-categories that may clearly not be found in all destinations (i.e. Chinese Aristocrat Restaurant, West-Ukrainian Restaurant, Pachinko Parlor or Nudist Beach).

A request can be sent to Foursquare to retrieve only the first level of the category tree. The results are shown below in Figure 4. Each category is attributed with a specific id value.



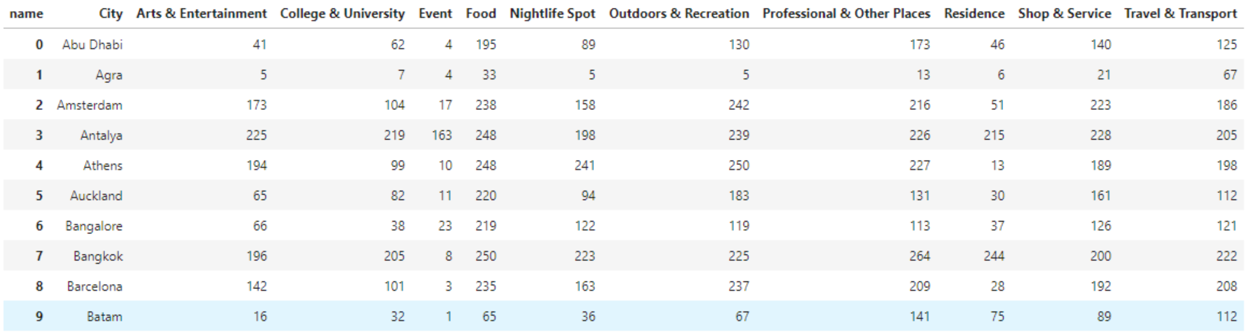
	id	name
0	4d4b7104d754a06370d81259	Arts & Entertainment
1	4d4b7105d754a06372d81259	College & University
2	4d4b7105d754a06373d81259	Event
3	4d4b7105d754a06374d81259	Food
4	4d4b7105d754a06376d81259	Nightlife Spot
5	4d4b7105d754a06377d81259	Outdoors & Recreation
6	4d4b7105d754a06375d81259	Professional & Other Places
7	4e67e38e036454776db1fb3a	Residence
8	4d4b7105d754a06378d81259	Shop & Service
9	4d4b7105d754a06379d81259	Travel & Transport

Figure 4 – Top level Foursquare categories

Requests were sent to the Foursquare venues API to retrieve the number of venues per category and for each city. In order to explore cities which may be very different in terms of land expansion, we set the radius to 10km to cover large metropolitan areas and smaller cities. The limit number of venues per city per category was set at 500.

Feature selection

The final dataset comprises of 100 rows for 10 columns. Each row represents one of the top100 most visited city. Each column represents the number of venues per Foursquare categories returned for each city as shown below in Figure 5.



name	City	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Abu Dhabi	41	62	4	195	89	130	173	46	140	125
1	Agra	5	7	4	33	5	5	13	6	21	67
2	Amsterdam	173	104	17	238	158	242	216	51	223	186
3	Antalya	225	219	163	248	198	239	226	215	228	205
4	Athens	194	99	10	248	241	250	227	13	189	198
5	Auckland	65	82	11	220	94	183	131	30	161	112
6	Bangalore	66	38	23	219	122	119	113	37	126	121
7	Bangkok	196	205	8	250	223	225	264	244	200	222
8	Barcelona	142	101	3	235	163	237	209	28	192	208
9	Batam	16	32	1	65	36	67	141	75	89	112

Figure 5 – Final dataset (first 10 rows)

Exploratory data analysis

City locations

To start, it is worth plotting the city locations on a world map (see below in Figure 6). This is done with a bubble marker sized in proportion to the arrivals in 2018. Clearly we can see the Asian hubs that are Hong Kong, Bangkok, Macau and Singapore which are all part of the top 5 most visited cities (the last one being London) in 2018. Also Europe and Asia seem to concentrate most of the top100 cities. On the other side, the African continent has only 2 cities in the top100.

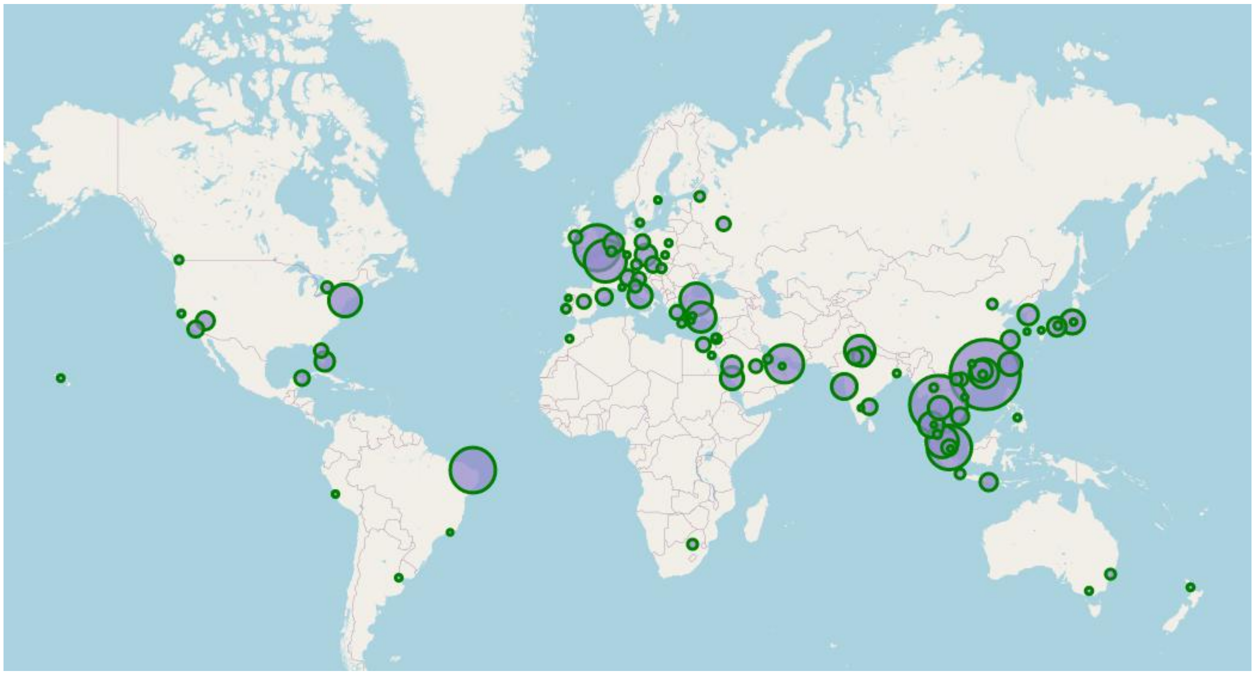


Figure 6 – Locations of the top100 most visited cities in the world

Statistics

Before attempting to cluster the cities, we conduct some statistical analysis to see patterns in our features for each category and city location.

After displaying the cities on a map, one could question the factors that contribute to the number of arrivals. Clearly air traffic must be a major contributor as Hong Kong, Singapore and London are some of the major airports in the world. Unfortunately we did not have any information about flights for the cities. One set of data that we had collected was the population in each city. As shown in Figure 7, when plotting the arrivals against the population in each city, there is no clear correlation between the 2 variables. Arrivals could be fairly independent to the city population and particularly for touristic destinations or regional travels hubs (Macau for example is one the most visited city but its population is relatively small compared to other cities).

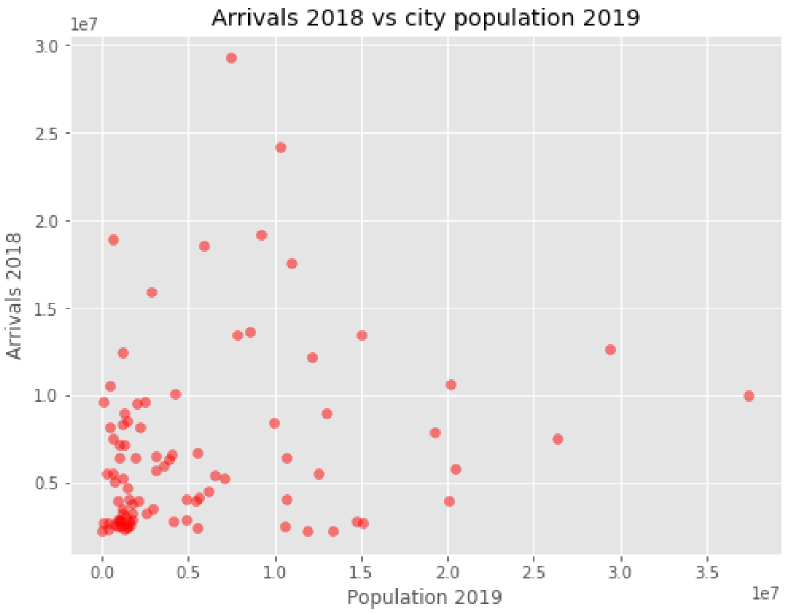


Figure 7 – Correlation between arrivals and population

Another way to explore the data is to look at the number of venues in comparison to the arrivals and to detail this analysis across the 10 Foursquare categories previously mentioned. Scatter plots can be drawn for each category with the number of category venues (as X) and the city arrivals (as Y). The data visualization is displayed in Figure 8. Again no clear patterns can be identified for any venue category. It is worth noticing however that the number of venues per category seem to be often limited to 250 which may be a limitation of the Foursquare data (even though the limit was set at 500 in the request). For example the number of food venues for a few cities is around 250.

The same analysis conducted on population rather than arrivals leads to the same conclusions. There are no clear correlations between number of venues per category and city population. The visualization is displayed in Figure 9 below.

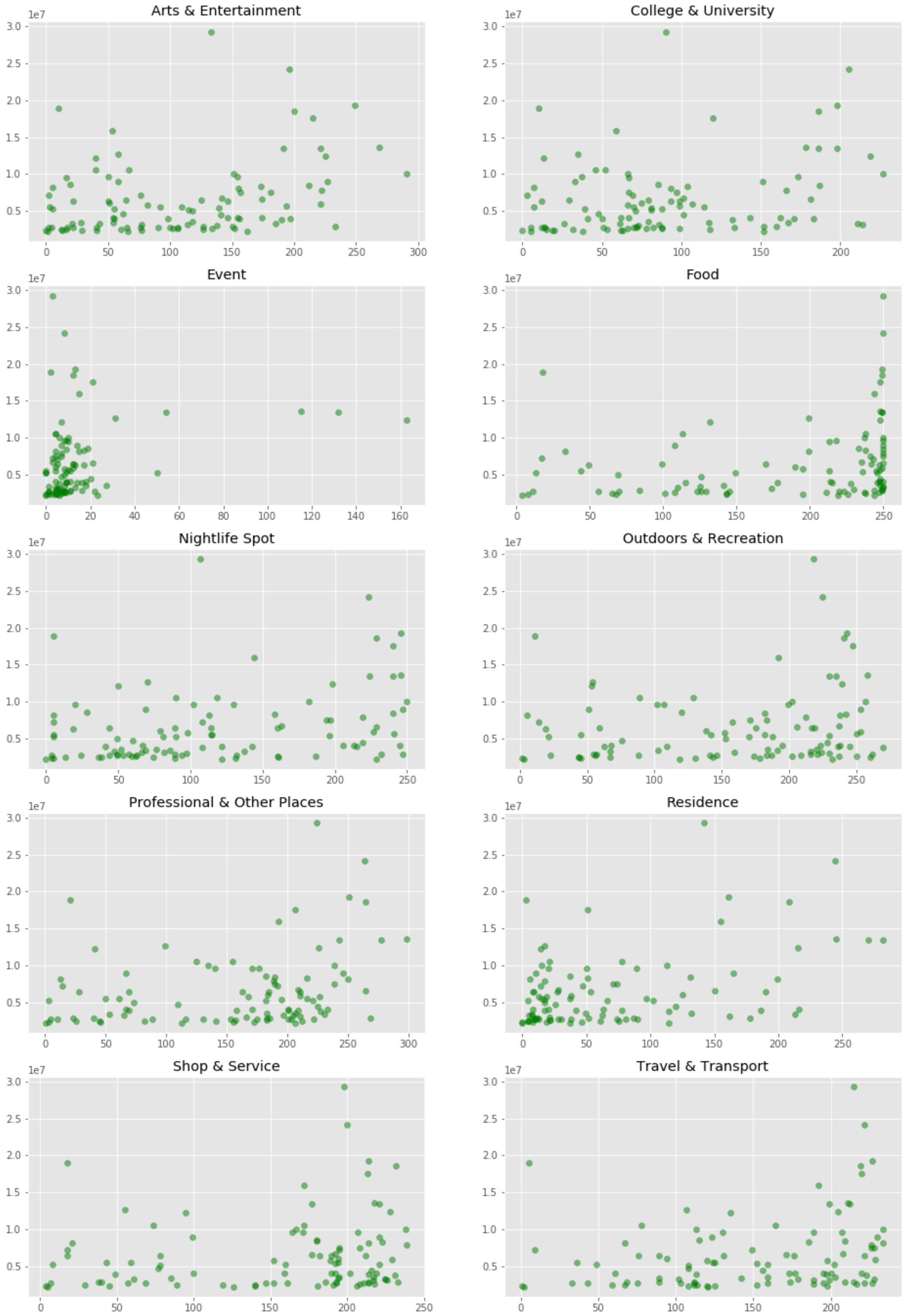
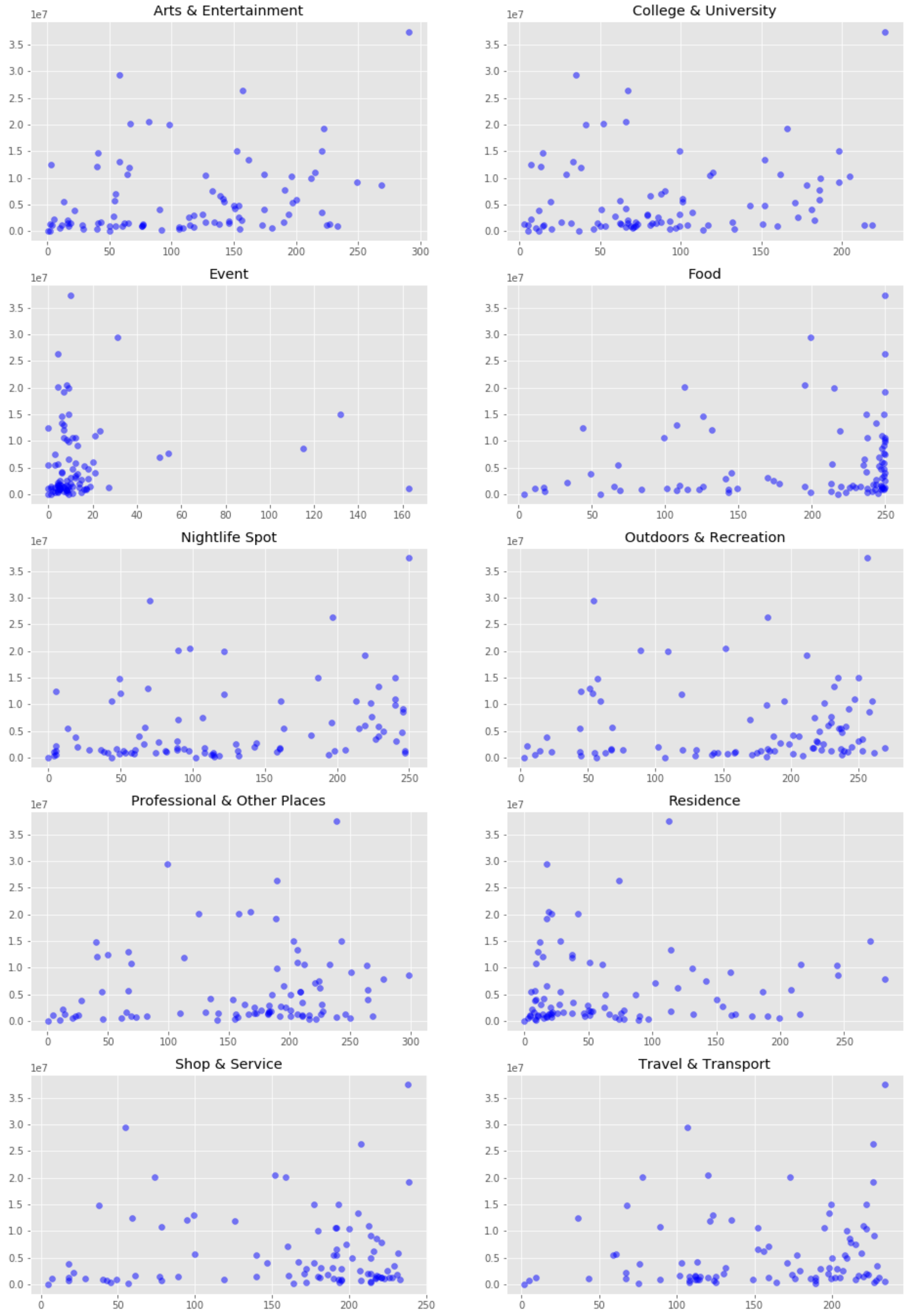


Figure 8 – Scatter plot of arrivals vs. venues by category



Arts & Entertainment

College & University

Event

Food

Nightlife Spot

Outdoors & Recreation

Professional & Other Places

Residence

Shop & Service

Travel & Transport

Figure 9 – Scatter plot of population vs. venues by category

To push this exploratory analysis a bit further, we examine the top10 cities for each Foursquare venue category in terms of venue number weighted by the city arrivals. This visualization is shown in Figure 10 below. It is interesting to note that Rio de Janeiro tops 4 categories and comes second on 4 others out of 10. Its relative low number of arrivals probably plays in its favor. San Francisco in the US is also present in a lot of categories. This may be due again to its number of arrivals less than other US cities like New York, Miami or Las Vegas where the venue offering might be fairly similar.

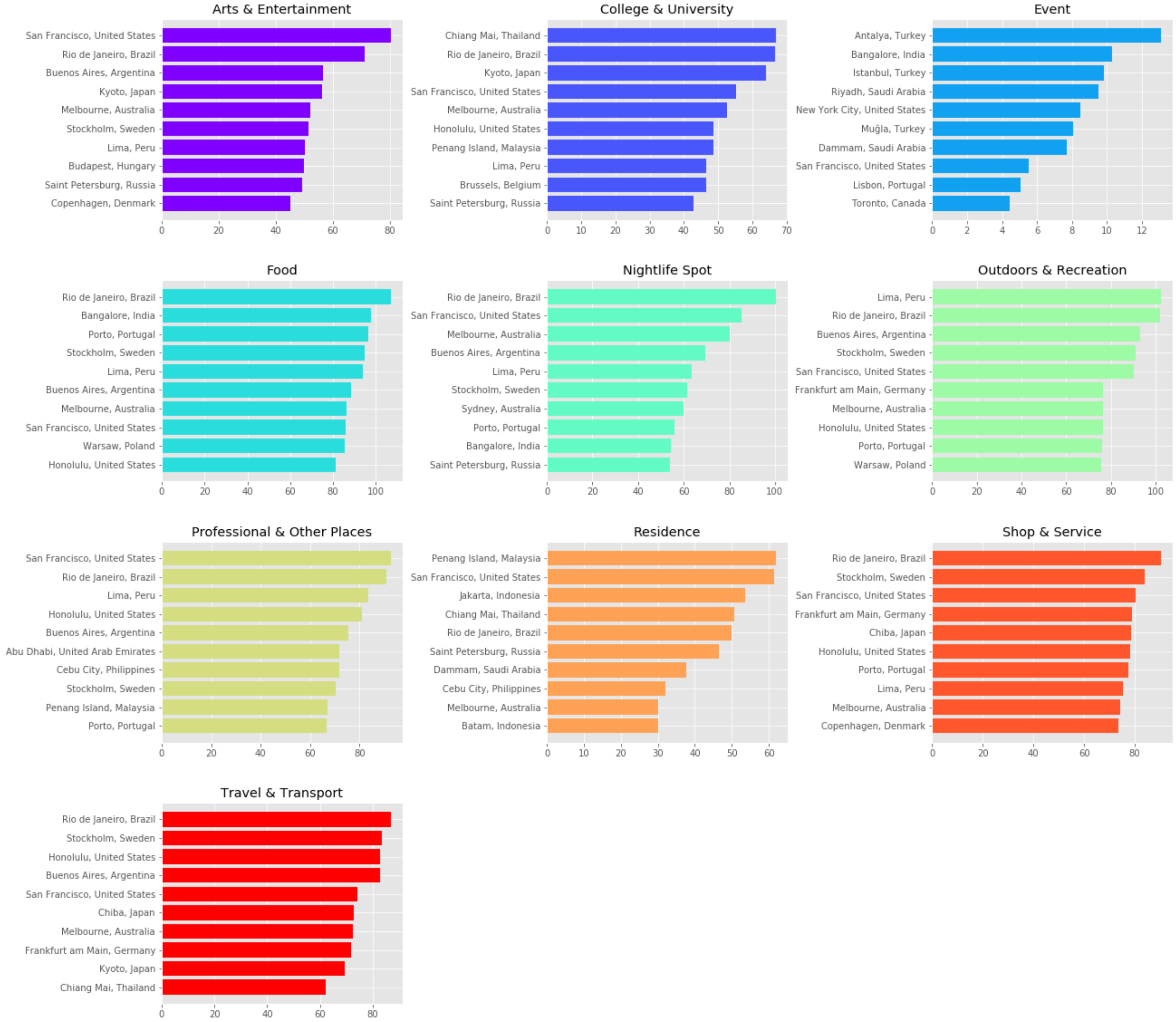


Figure 10 – Top10 cities on venue numbers weighted by city arrivals per category

Finally, as part of the exploratory data analysis, we look at the distribution of the venues numbers per category as shown in Figure 11, thanks to a boxplot diagram. The number of venues for events is questionable as most numbers are very low although we note some clear outliers. We can also clearly see the number of food venues is “skewed” towards the number of 250 which may resemble a limit in the Foursquare data returned.

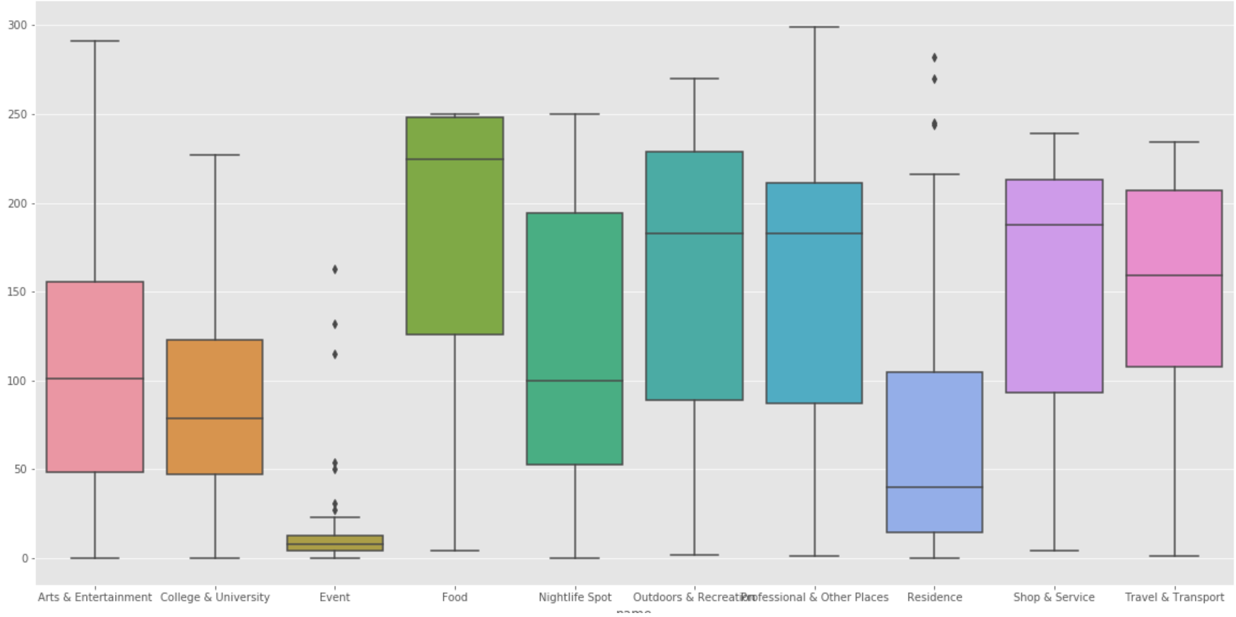


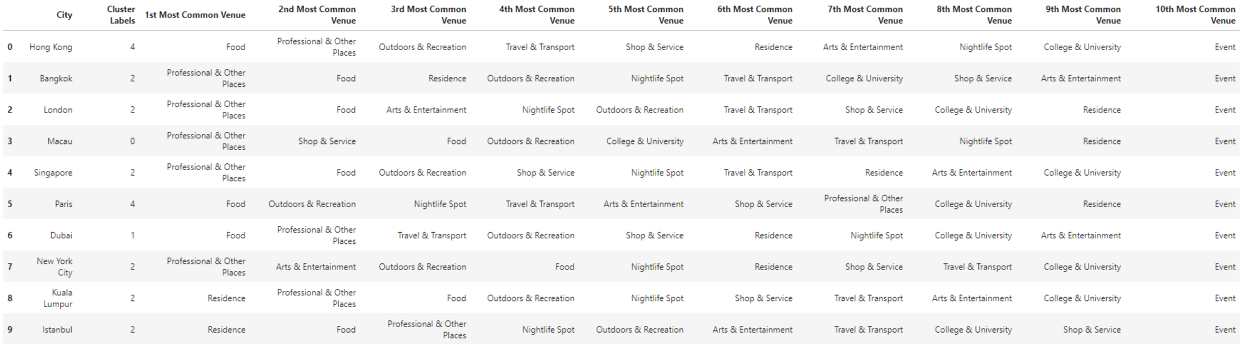
Figure 11 – Distribution of venue numbers per category

Clustering

Since the main objective of this analysis is to group cities which share similar features, we will use a K-means clustering algorithm. K-means clustering is a method of finding clusters and cluster centers in a set of unlabeled data. Our unlabeled data consist of the number of venues in each of the top 10 Foursquare category for each of the 100 most visited in the world.

A number of 5 k-clusters was selected to run the algorithm.

First the most common venue categories were displayed for each city as shown below in Figure 12. Not surprisingly the ‘food’ venue category seemed to appear in the first or second most common venue category for a lot of cities as this is the highest mean value of venues per category (as seen in previous Figure 11). On the other hand, the ‘events’ venue category is the last venue category for almost all cities as it has the lowest mean value of all categories. To explore further the ‘events’ category, we selected the city which tops the category on Figure 10 previously, that is Antalya in Turkey with 100 venues. Unfortunately names of venues in Turkish did not shine any new light for us!



	City	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Hong Kong	4	Food	Professional & Other Places	Outdoors & Recreation	Travel & Transport	Shop & Service	Residence	Arts & Entertainment	Nightlife Spot	College & University	Event
1	Bangkok	2	Professional & Other Places	Food	Residence	Outdoors & Recreation	Nightlife Spot	Travel & Transport	College & University	Shop & Service	Arts & Entertainment	Event
2	London	2	Professional & Other Places	Food	Arts & Entertainment	Nightlife Spot	Outdoors & Recreation	Travel & Transport	Shop & Service	College & University	Residence	Event
3	Macau	0	Professional & Other Places	Shop & Service	Food	Outdoors & Recreation	College & University	Arts & Entertainment	Travel & Transport	Nightlife Spot	Residence	Event
4	Singapore	2	Professional & Other Places	Food	Outdoors & Recreation	Shop & Service	Nightlife Spot	Travel & Transport	Residence	Arts & Entertainment	College & University	Event
5	Paris	4	Food	Outdoors & Recreation	Nightlife Spot	Travel & Transport	Arts & Entertainment	Shop & Service	Professional & Other Places	College & University	Residence	Event
6	Dubai	1	Food	Professional & Other Places	Travel & Transport	Outdoors & Recreation	Shop & Service	Residence	Nightlife Spot	College & University	Arts & Entertainment	Event
7	New York City	2	Professional & Other Places	Arts & Entertainment	Outdoors & Recreation	Food	Nightlife Spot	Residence	Shop & Service	Travel & Transport	College & University	Event
8	Kuala Lumpur	2	Residence	Professional & Other Places	Food	Outdoors & Recreation	Nightlife Spot	Shop & Service	Travel & Transport	Arts & Entertainment	College & University	Event
9	Istanbul	2	Residence	Food	Professional & Other Places	Nightlife Spot	Outdoors & Recreation	Arts & Entertainment	Travel & Transport	College & University	Shop & Service	Event

Figure 12 – Most common venue categories and cluster classification for a selection of 10 cities

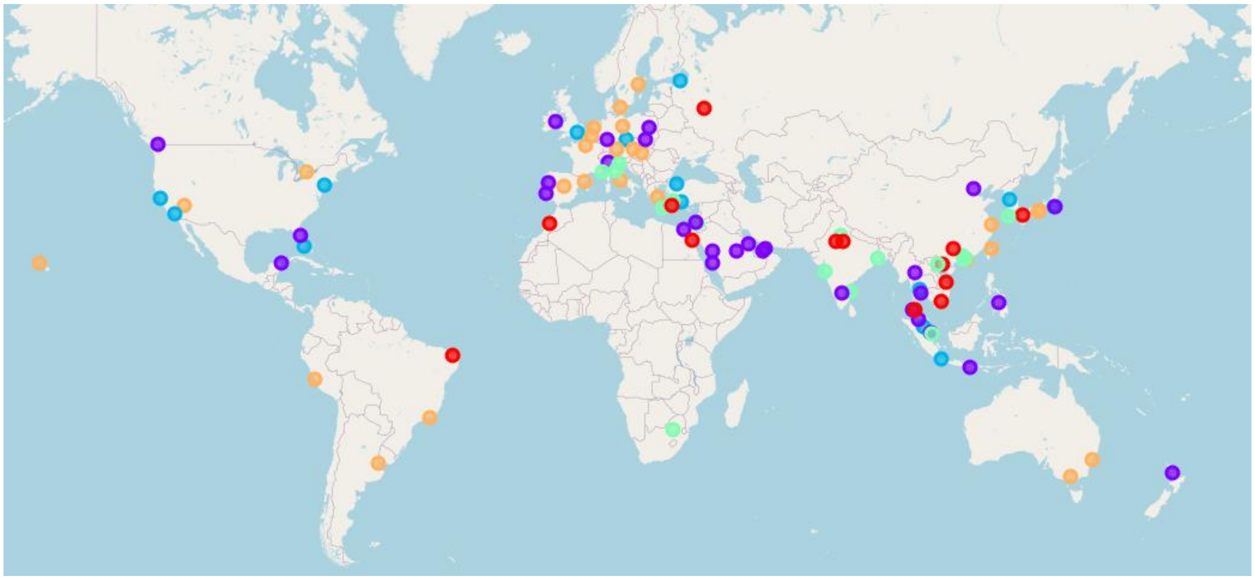


Figure 13 – Cities cluster classification

In order to understand further the cluster classification, for each city and each category we calculate the standard deviation of the city number of venues from each category mean. This will allow some normalization across categories rather than comparing absolute number of venues per category. This will also allow to visualize the patterns in each of the 5 clusters of cities (shown on Figure 14).

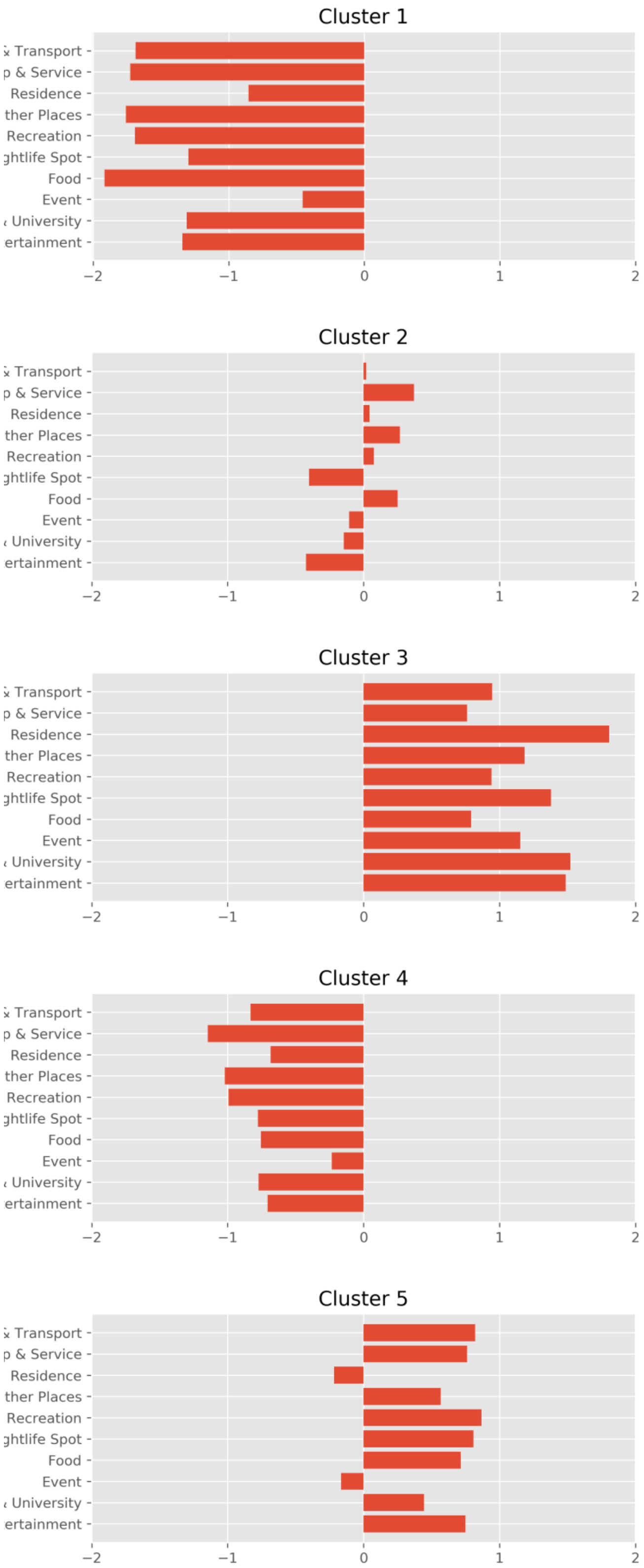
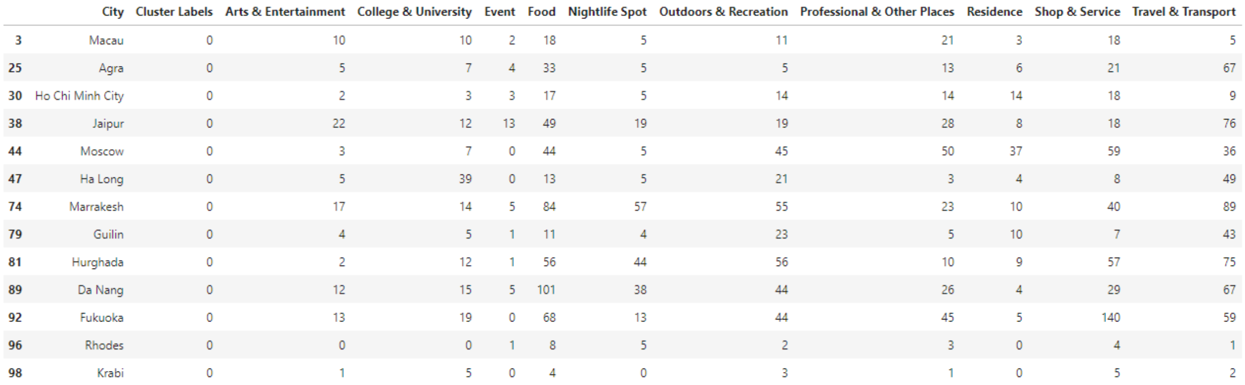


Figure 14 – Cluster category means standard deviations

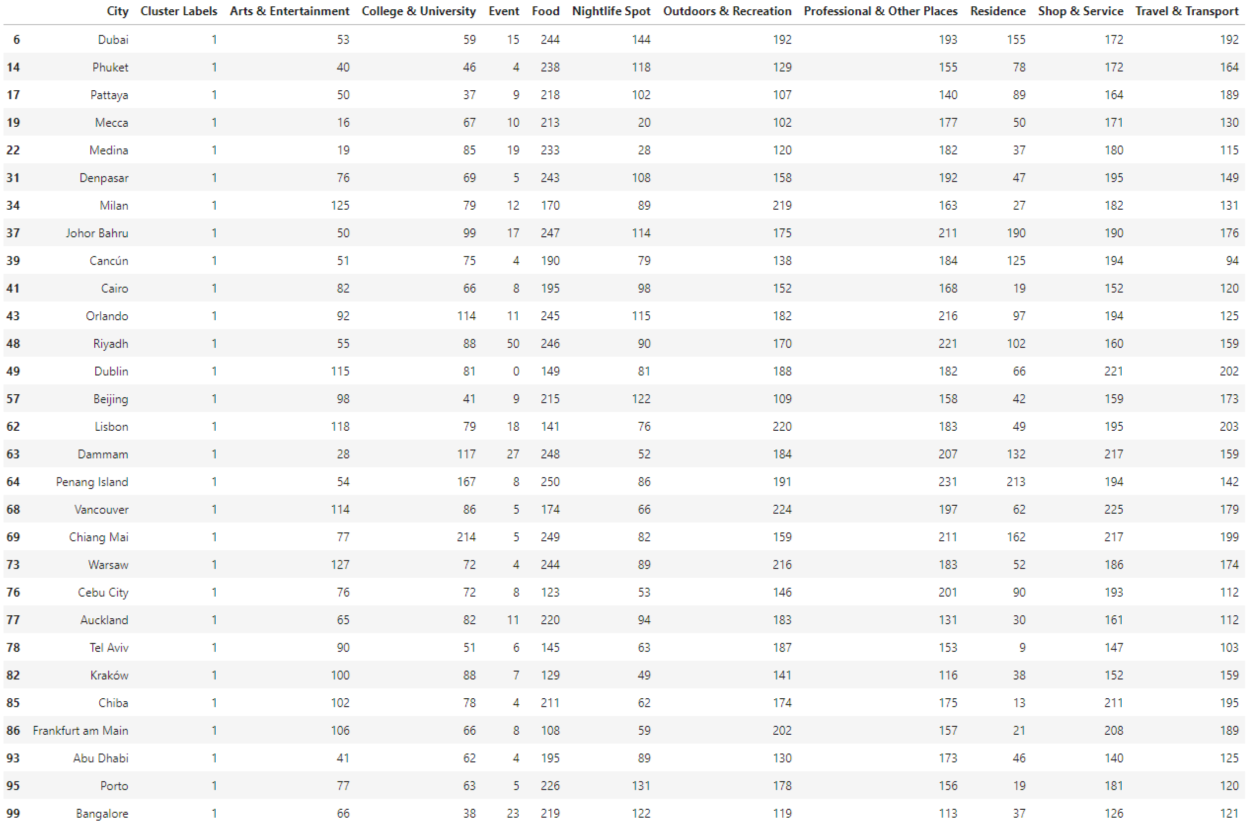
We can note the following:

- **Cluster 1** seems to include all cities where venue numbers for all categories are in the lower end of the scale. We find cities like Macau, Agra, Ho Chi Minh or Fukuoka (Japan).



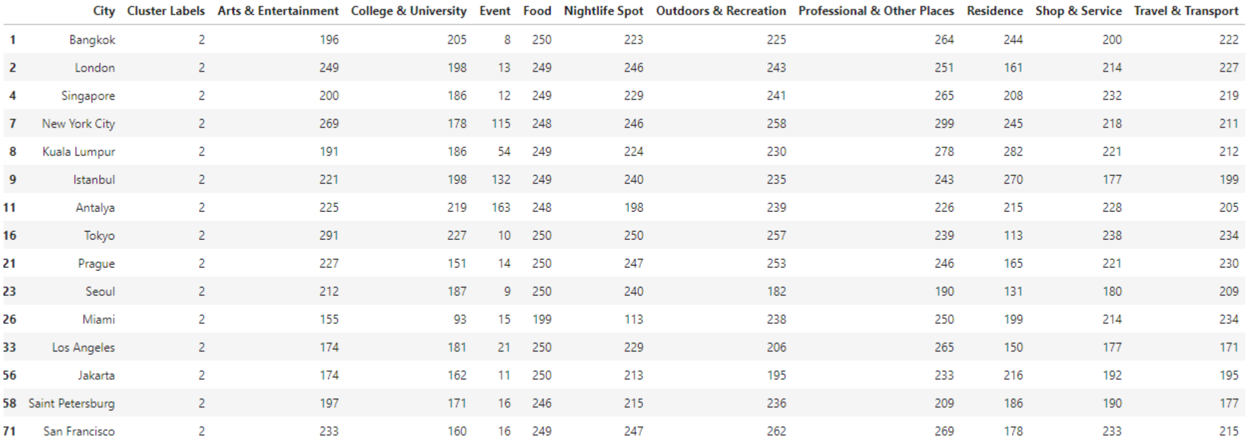
	City	Cluster Labels	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
3	Macau	0	10	10	2	18	5	11	21	3	18	5
25	Agra	0	5	7	4	33	5	5	13	6	21	67
30	Ho Chi Minh City	0	2	3	3	17	5	14	14	14	18	9
38	Jaipur	0	22	12	13	49	19	19	28	8	18	76
44	Moscow	0	3	7	0	44	5	45	50	37	59	36
47	Ha Long	0	5	39	0	13	5	21	3	4	8	49
74	Marrakesh	0	17	14	5	84	57	55	23	10	40	89
79	Guilin	0	4	5	1	11	4	23	5	10	7	43
81	Hurghada	0	2	12	1	56	44	56	10	9	57	75
89	Da Nang	0	12	15	5	101	38	44	26	4	29	67
92	Fukuoka	0	13	19	0	68	13	44	45	5	140	59
96	Rhodes	0	0	0	1	8	5	2	3	0	4	1
98	Krabi	0	1	5	0	4	0	3	1	0	5	2

- **Cluster 2** seems to include all cities where venue numbers for all categories are near the mean. We find cities like Dubai, Phuket, Milan, Lisbon or Dublin for example



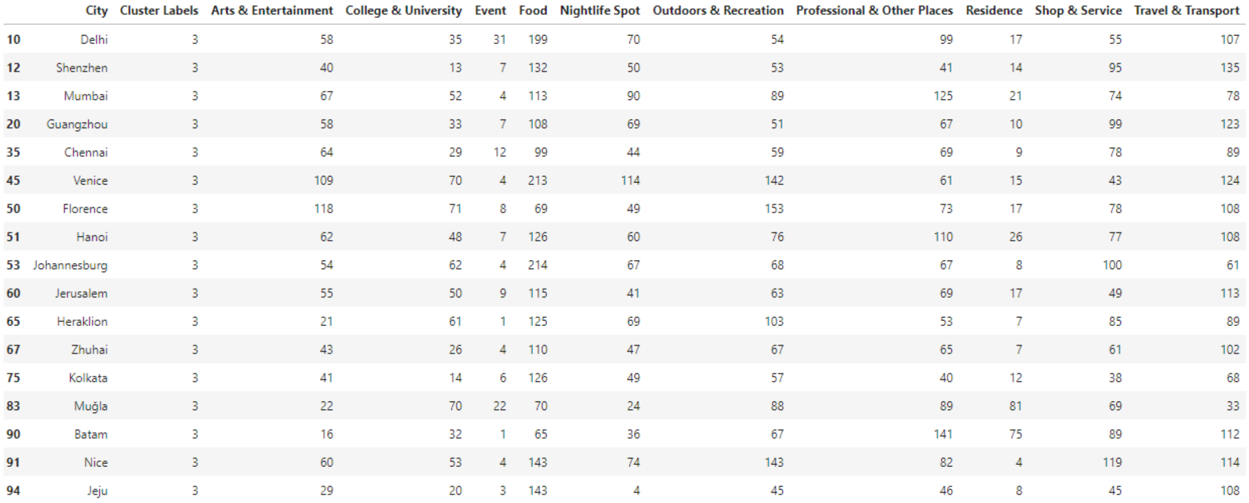
	City	Cluster Labels	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
6	Dubai	1	53	59	15	244	144	192	193	155	172	192
14	Phuket	1	40	46	4	238	118	129	155	78	172	164
17	Pattaya	1	50	37	9	218	102	107	140	89	164	189
19	Mecca	1	16	67	10	213	20	102	177	50	171	130
22	Medina	1	19	85	19	233	28	120	182	37	180	115
31	Denpasar	1	76	69	5	243	108	158	192	47	195	149
34	Milan	1	125	79	12	170	89	219	163	27	182	131
37	Johor Bahru	1	50	99	17	247	114	175	211	190	190	176
39	Cancún	1	51	75	4	190	79	138	184	125	194	94
41	Cairo	1	82	66	8	195	98	152	168	19	152	120
43	Orlando	1	92	114	11	245	115	182	216	97	194	125
48	Riyadh	1	55	88	50	246	90	170	221	102	160	159
49	Dublin	1	115	81	0	149	81	188	182	66	221	202
57	Beijing	1	98	41	9	215	122	109	158	42	159	173
62	Lisbon	1	118	79	18	141	76	220	183	49	195	203
63	Dammam	1	28	117	27	248	52	184	207	132	217	159
64	Penang Island	1	54	167	8	250	86	191	231	213	194	142
68	Vancouver	1	114	86	5	174	66	224	197	62	225	179
69	Chiang Mai	1	77	214	5	249	82	159	211	162	217	199
73	Warsaw	1	127	72	4	244	89	216	183	52	186	174
76	Cebu City	1	76	72	8	123	53	146	201	90	193	112
77	Auckland	1	65	82	11	220	94	183	131	30	161	112
78	Tel Aviv	1	90	51	6	145	63	187	153	9	147	103
82	Kraków	1	100	88	7	129	49	141	116	38	152	159
85	Chiba	1	102	78	4	211	62	174	175	13	211	195
86	Frankfurt am Main	1	106	66	8	108	59	202	157	21	208	189
93	Abu Dhabi	1	41	62	4	195	89	130	173	46	140	125
95	Porto	1	77	63	5	226	131	178	156	19	181	120
99	Bangalore	1	66	38	23	219	122	119	113	37	126	121

- **Cluster 3** seems to include all cities where venue numbers for all categories are in the higher end of the scale. We find cities like Bangkok, London, Singapore, New York, Tokyo for example.



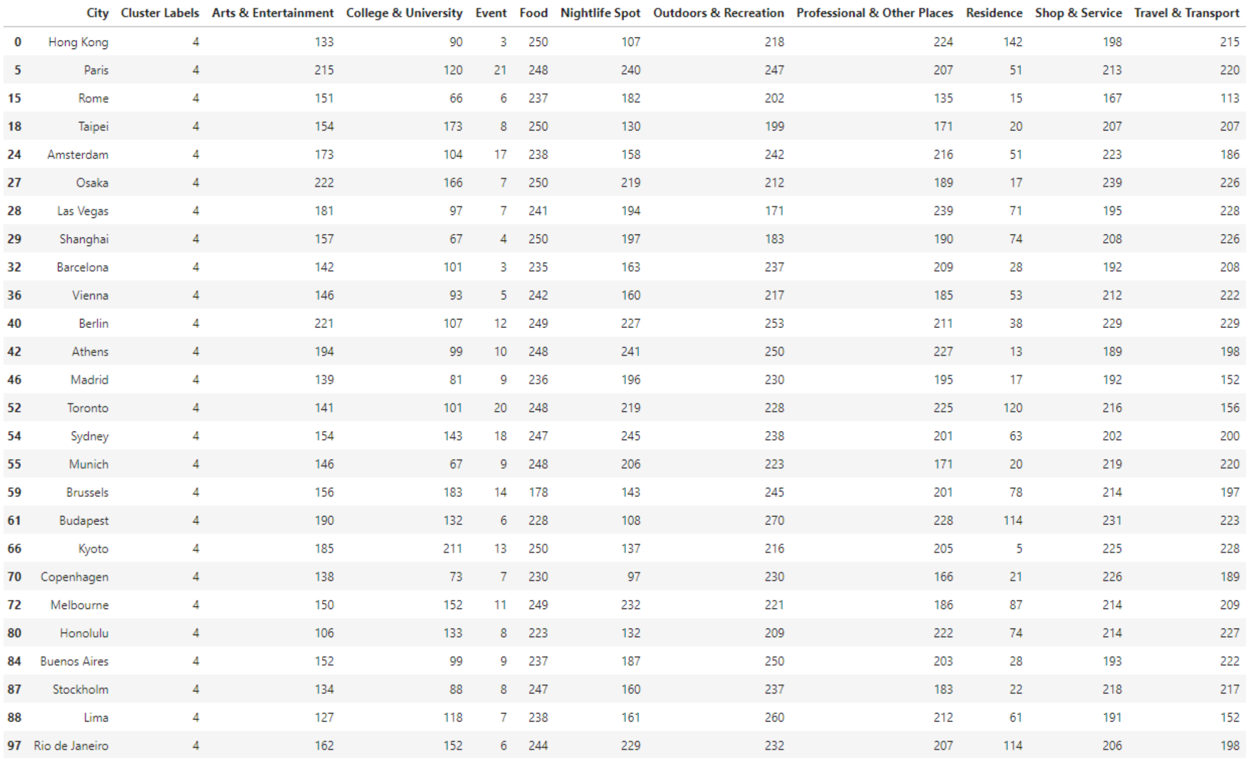
	City	Cluster Labels	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
1	Bangkok	2	196	205	8	250	223	225	264	244	200	222
2	London	2	249	198	13	249	246	243	251	161	214	227
4	Singapore	2	200	186	12	249	229	241	265	208	232	219
7	New York City	2	269	178	115	248	246	258	299	245	218	211
8	Kuala Lumpur	2	191	186	54	249	224	230	278	282	221	212
9	Istanbul	2	221	198	132	249	240	235	243	270	177	199
11	Antalya	2	225	219	163	248	198	239	226	215	228	205
16	Tokyo	2	291	227	10	250	250	257	239	113	238	234
21	Prague	2	227	151	14	250	247	253	246	165	221	230
23	Seoul	2	212	187	9	250	240	182	190	131	180	209
26	Miami	2	155	93	15	199	113	238	250	199	214	234
33	Los Angeles	2	174	181	21	250	229	206	265	150	177	171
56	Jakarta	2	174	162	11	250	213	195	233	216	192	195
58	Saint Petersburg	2	197	171	16	246	215	236	209	186	190	177
71	San Francisco	2	233	160	16	249	247	262	269	178	233	215

- **Cluster 4** seems to include all cities where venue numbers for all categories are somewhat lower than average but better than in the cluster 1. We find cities like Delhi, Venice, Hanoi and Nice.



	City	Cluster Labels	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
10	Delhi	3	58	35	31	199	70	54	99	17	55	107
12	Shenzhen	3	40	13	7	132	50	53	41	14	95	135
13	Mumbai	3	67	52	4	113	90	89	125	21	74	78
20	Guangzhou	3	58	33	7	108	69	51	67	10	99	123
35	Chennai	3	64	29	12	99	44	59	69	9	78	89
45	Venice	3	109	70	4	213	114	142	61	15	43	124
50	Florence	3	118	71	8	69	49	153	73	17	78	108
51	Hanoi	3	62	48	7	126	60	76	110	26	77	108
53	Johannesburg	3	54	62	4	214	67	68	67	8	100	61
60	Jerusalem	3	55	50	9	115	41	63	69	17	49	113
65	Heraklion	3	21	61	1	125	69	103	53	7	85	89
67	Zhuhai	3	43	26	4	110	47	67	65	7	61	102
75	Kolkata	3	41	14	6	126	49	57	40	12	38	68
83	Muğla	3	22	70	22	70	24	88	89	81	69	33
90	Batam	3	16	32	1	65	36	67	141	75	89	112
91	Nice	3	60	53	4	143	74	143	82	4	119	114
94	Jeju	3	29	20	3	143	4	45	46	8	45	108

- **Cluster 5** seems to include all cities where venue numbers for all categories are somewhat higher than average but not as good as the cluster 3. We find cities like Hong Kong, Paris, Rome, Las Vegas or Berlin.



	City	Cluster Labels	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Hong Kong	4	133	90	3	250	107	218	224	142	198	215
5	Paris	4	215	120	21	248	240	247	207	51	213	220
15	Rome	4	151	66	6	237	182	202	135	15	167	113
18	Taipei	4	154	173	8	250	130	199	171	20	207	207
24	Amsterdam	4	173	104	17	238	158	242	216	51	223	186
27	Osaka	4	222	166	7	250	219	212	189	17	239	226
28	Las Vegas	4	181	97	7	241	194	171	239	71	195	228
29	Shanghai	4	157	67	4	250	197	183	190	74	208	226
32	Barcelona	4	142	101	3	235	163	237	209	28	192	208
36	Vienna	4	146	93	5	242	160	217	185	53	212	222
40	Berlin	4	221	107	12	249	227	253	211	38	229	229
42	Athens	4	194	99	10	248	241	250	227	13	189	198
46	Madrid	4	139	81	9	236	196	230	195	17	192	152
52	Toronto	4	141	101	20	248	219	228	225	120	216	156
54	Sydney	4	154	143	18	247	245	238	201	63	202	200
55	Munich	4	146	67	9	248	206	223	171	20	219	220
59	Brussels	4	156	183	14	178	143	245	201	78	214	197
61	Budapest	4	190	132	6	228	108	270	228	114	231	223
66	Kyoto	4	185	211	13	250	137	216	205	5	225	228
70	Copenhagen	4	138	73	7	230	97	230	166	21	226	189
72	Melbourne	4	150	152	11	249	232	221	186	87	214	209
80	Honolulu	4	106	133	8	223	132	209	222	74	214	227
84	Buenos Aires	4	152	99	9	237	187	250	203	28	193	222
87	Stockholm	4	134	88	8	247	160	237	183	22	218	217
88	Lima	4	127	118	7	238	161	260	212	61	191	152
97	Rio de Janeiro	4	162	152	6	244	229	232	207	114	206	198

As we can see from some of the examples of cities clustered together, there is not a clear pattern of classification that we can highlight. It seems the algorithm has not perceived the differences in the categories as much as one would have expected. The classification is influenced by the number of venues consistently across all categories. We did not find a cluster that could have been the result of particular cities where venue numbers in a particular category would have been higher.

The results are discussed further in the next chapter.

PART 4 – RESULTS

There is no correlation between number of arrivals and local venues

From the attempts to link the number of arrivals in each of the most visited city with either the local population numbers or the venues numbers, there is no clear correlation between these variables. The number of arrivals in a location seems disconnected to the location itself. This behavior is partially explained by the fact that some of the cities have become a regional hub for travel and particularly by plane. It is not surprising to see some of the biggest airports in the top 10 most visited cities in the world, like Hong Kong, Singapore, Dubai or Macau.

Total venue numbers have a high influence on cluster classification

When trying to cluster cities across the world, we note lots of similarities in the number of venues per category. This might be due to the fact that past a certain size, we expect cities to have a relatively close number of venues for each category. This flattens a bit the cluster classification and does not enable to distinguish cities better by certain categories. In other words, the cluster differences highlight differences in the global number of venues across all categories (cf. Figure 14 – Cluster category means standard deviations).

PART 5 – DISCUSSION

Understanding data feed is crucial

The use of the Foursquare data location is prone to a lot of questions. The ‘explore’ function of the Foursquare API returns recommended venues only. This could be an issue if our analysis would benefit more from an actual count of venues per category. For our most visited cities, Foursquare was able to return a number of venues in each location fairly similar. In other words, the distribution of number of venues does not show great dispersion and flattens the differences that way.

Moreover the date and time the request is made will change the results. As we are submitting requests at the same time for world locations, one would question if less venues would be returned for cities where it is night time than for cities where it is day time.

Changing the radius (which was set at 10km) of the explore request did not seem to always have the expected results either. Sometimes we would find fewer venues with a greater radius. Again increasing the maximum limit of returned venues did not yield the expected results. Even though this limit was set at 500, the results never went beyond 300 per category.

Improving the quality of data and adding features

On top on fixing some of the issues with the location data from Foursquare, the analysis would benefit from additional features to be more relevant. This would require to find other data sources. One can think to add other features like the following:

- City land area
- Air quality
- Traffic
- Weather
- Population
- Cost of living

Testing on a smaller scale

As shown in examples of analyses of neighborhoods, the scale of this analysis may have been too large to yield good results. Focusing the analysis to a country or a region may be more pertinent. The issues with time of day will be lifted in that case and the Foursquare API may be able to return more pertinent results.

Would you recommend destinations based on this analysis?

This analysis provides a first cluster classification of the most visited cities in the world. It can be used to give travel agency clients a sense of each destination in terms of their number of venues, whether the destination is a rather ‘active’ place or not. This would be not be appropriate at this stage to make recommendations of destinations based on destinations that fall in the cluster. For example, Hong Kong and Paris fall in the same cluster but they are rather very different cities. One would feel clearly that other features will need to be considered to offer a proper recommendation system.

PART 6 – CONCLUSION

This analysis has created 5 clusters of the 100 most visited cities in the world. The classification has however largely used the global venue numbers per city location to create the clusters. On both ends, cities with high numbers of venues in most categories tend to be in one cluster together when cities with low numbers of venues in most categories tend to be in another cluster.

Using machine learning to distinguish world cities require a good selection of features. It appears that limiting the analysis to the number of recommended venues per category does not highlight clear differences between cities. The use of sources other than Foursquare may be necessary to improve the classification as Foursquare returns only recommended venues when we would rather need to count venues in each location for example.

Table of figures