Analysis of Budapest Neighborhoods

DATE : 8th Aug 2020

By
Sethumadhavan Aravindakshan

INTRODUCTION

Budapest is ranked as the top Eastern European ecosystem for scaleups due to the city's availability of capital. The tech infrastructure is exceptional and continues to develop further creating many opportunities for entrepreneurs and working professionals.

Aim of this project is to cluster the Budapest neighborhood according to their similarities into groups as :

- Fully developed neighborhood
- Moderately developed neighborhood
- Developing neighborhood
- Underdeveloped neighborhood

This will aid anyone who visits Budapest city to decide the place of stay based on what each neighborhood has to offer and their personal preferences.

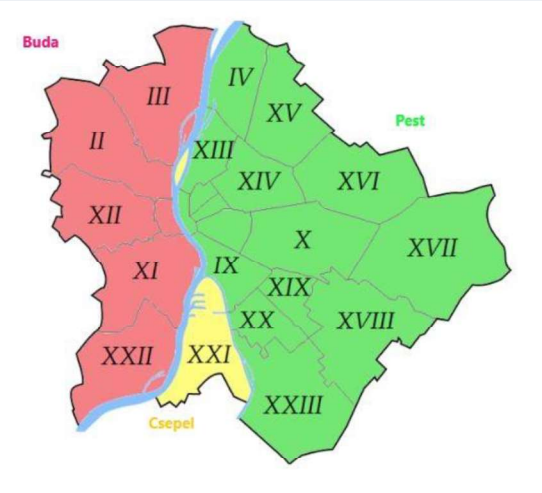
Data

In order to perform a meaningful analysis, we need data from reliable sources. To understand our problem and quantify results, the following data is used:

1. List of all districts of the city and postal codes are extracted from the below url: https://en.wikipedia.org/wiki/List_of_postal_codes_in_Hungary

	Postal Code	District names	Neighborhood
0	1010	Várkerület	I.
1	1020	2nd district	II.
2	1030	Óbuda-Békásmegyer	III.
3	1040	Újpest	IV.
4	1050	Belváros-Lipótváros	V.

2. Then the Borough names are added as per the below map:



	Postal Code	District names	Neighborhood	Borough
0	1010	Várkerület	I.	Buda
1	1020	2nd district	II.	Buda
2	1030	Óbuda-Békásmegyer	III.	Buda
3	1040	Újpest	IV.	Pest
4	1050	Belváros-Lipótváros	V.	Pest

3. Geographical co-ordinates of each neighborhood is extracted using the Geocoder from Geopy python library.

	Postal Code	District names	Neighborhood	Borough	Latitude	Longitude
0	1010	Várkerület	I.	Buda	47.499163	19.035143
1	1020	2nd district	II.	Buda	47.538887	18.982636
2	1030	Óbuda-Békásmegyer	III.	Buda	47.567611	19.036780
3	1040	Újpest	IV.	Pest	47.558687	19.079662
4	1050	Belváros-Lipótváros	V.	Pest	47.499945	19.050549

4. Foursquare developer access to venue data : https://foursquare.com/developers//
   Foursquare API is used to explore and extract the list of all venues and their frequencies of the neighborhood of the Budapest city. This data is used to further cluster the neighborhood based on their similarities.

Later a list of all venues for the top 10 categories are extracted using the Foursquare API and then frequency for all categories for each neighborhood is determined. Finally, Unsupervised Machine Learning Clustering techniques are applied to determine a similar pattern between the neighborhoods to group them to similar clusters for further analysis.

Methodology

One of the first steps we engaged in was to outline the sequence of steps that we will be following for our project. Each of these steps are elaborated below later:

- ✓ Data Preparation and Data Engineering
- ✓ Performing K-means clustering algorithm to segment neighborhoods

Data Preparation and EDA:

Foursquare API is used to explore types of venues in each area. Foursquare identified 10 top level categories available in the city and there are multiple sub categories which will not be used it for the time and they are as follows:

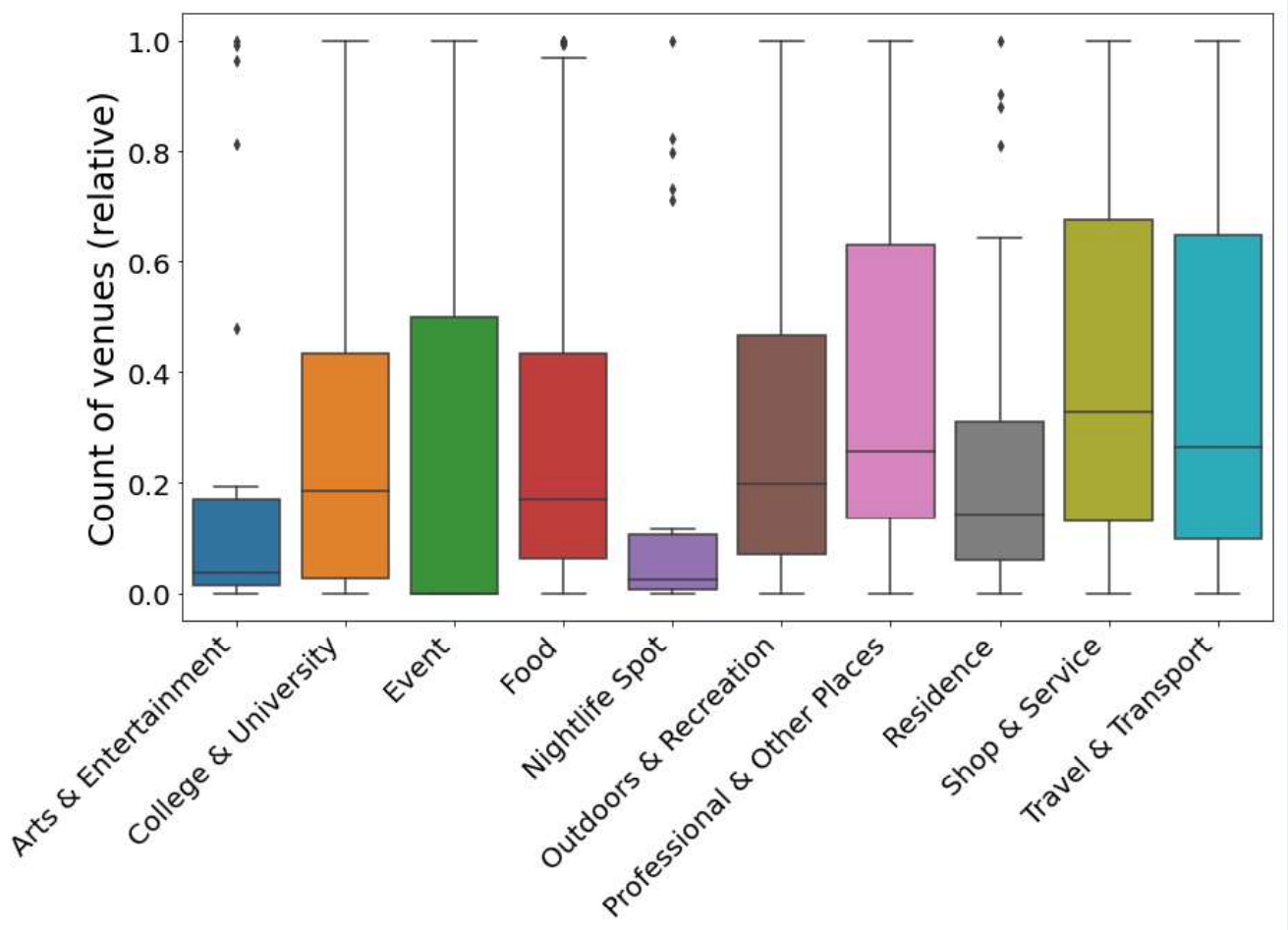
```
[('Arts & Entertainment', '4d4b7104d754a06370d81259'),
 ('College & University', '4d4b7105d754a06372d81259'),
 ('Event', '4d4b7105d754a06373d81259'),
 ('Food', '4d4b7105d754a06374d81259'),
 ('Nightlife Spot', '4d4b7105d754a06376d81259'),
 ('Outdoors & Recreation', '4d4b7105d754a06377d81259'),
 ('Professional & Other Places', '4d4b7105d754a06375d81259'),
 ('Residence', '4e67e38e036454776db1fb3a'),
 ('Shop & Service', '4d4b7105d754a06378d81259'),
 ('Travel & Transport', '4d4b7105d754a06379d81259')]
```

Then frequencies for each category is determined for all Neighborhoods of the city and stored as a data frame for further analysis.

	Neighborhood	Latitude	Longitude	Arts & Entertainment	College & University	...	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	I.	47.499163	19.035143	116	53	...	184	174	27	178	176
1	II.	47.538887	18.982636	6	2	...	39	36	1	33	30
2	III.	47.567611	19.036780	6	10	...	52	67	11	62	44
3	IV.	47.558687	19.079662	11	27	...	68	89	12	94	80
4	V.	47.499945	19.050549	142	92	...	199	175	37	192	193

Data Normalization:

As the frequencies of each categories may differ from each other, the above data is normalized using MinMax Scaler to bring the frequencies in the scale 0 to 1. After normalization, they can be visualized as below using Box plot for the entire dataset.

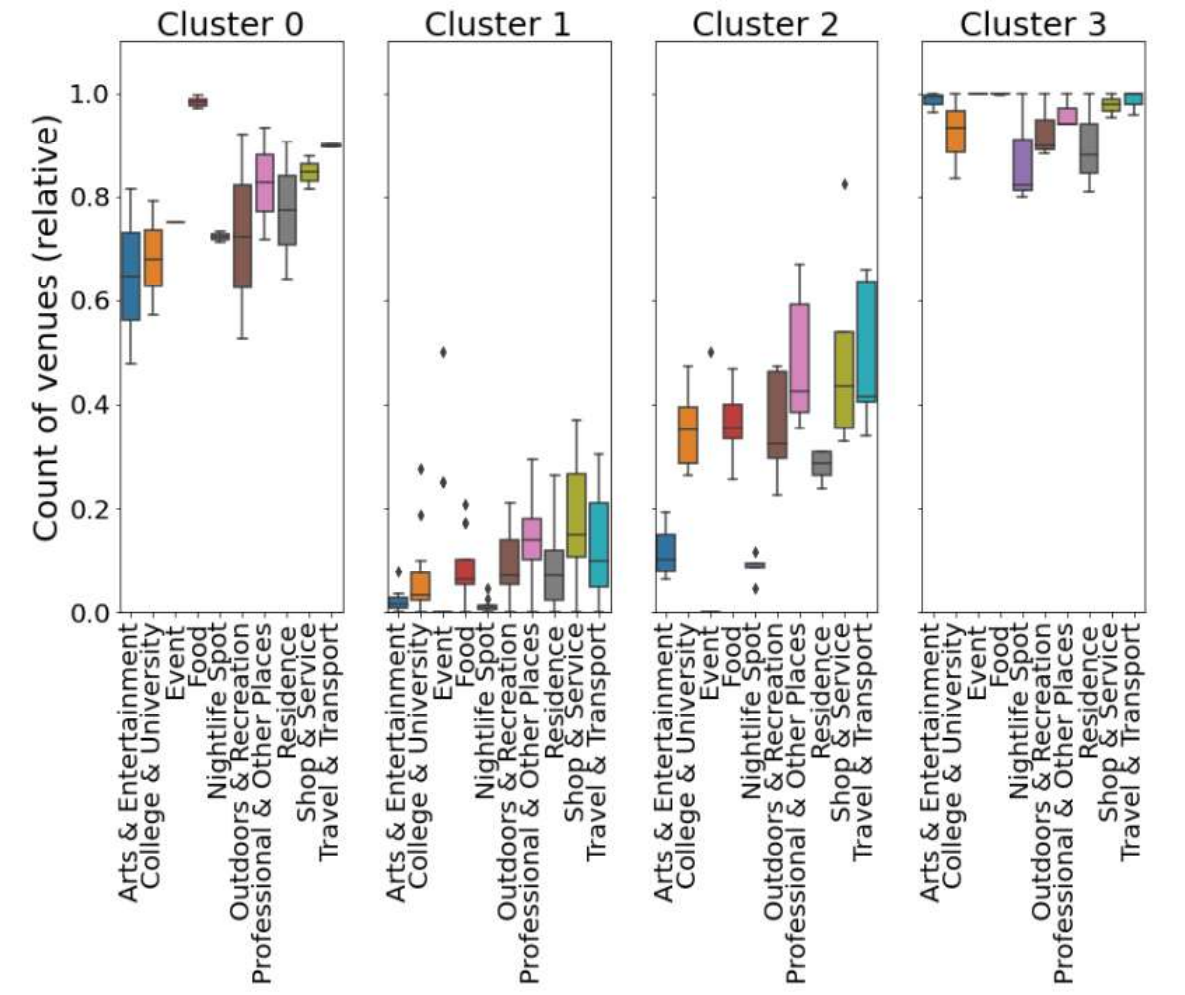


Descriptive statistics of each categories:

	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
count	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000
mean	0.222050	0.294792	0.260870	0.340761	0.201467	0.314399	0.383234	0.291925	0.407407	0.387999
std	0.354476	0.324360	0.380477	0.373558	0.334975	0.321622	0.329554	0.320181	0.332037	0.351402
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.014286	0.027473	0.000000	0.062500	0.006173	0.069892	0.134731	0.059524	0.129630	0.096491
50%	0.035714	0.186813	0.000000	0.170833	0.024691	0.198925	0.257485	0.142857	0.328042	0.263158
75%	0.171429	0.434066	0.500000	0.433333	0.104938	0.467742	0.631737	0.309524	0.677249	0.649123
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Performing K-means clustering algorithm to segment neighborhoods

Based on our problem statement, it would be ideal to choose a cluster value of 4 as it would cluster the neighborhoods better and make it more interpretable. Upon building the model and clustering the neighborhoods to 4 clusters we can visualize them as follows:

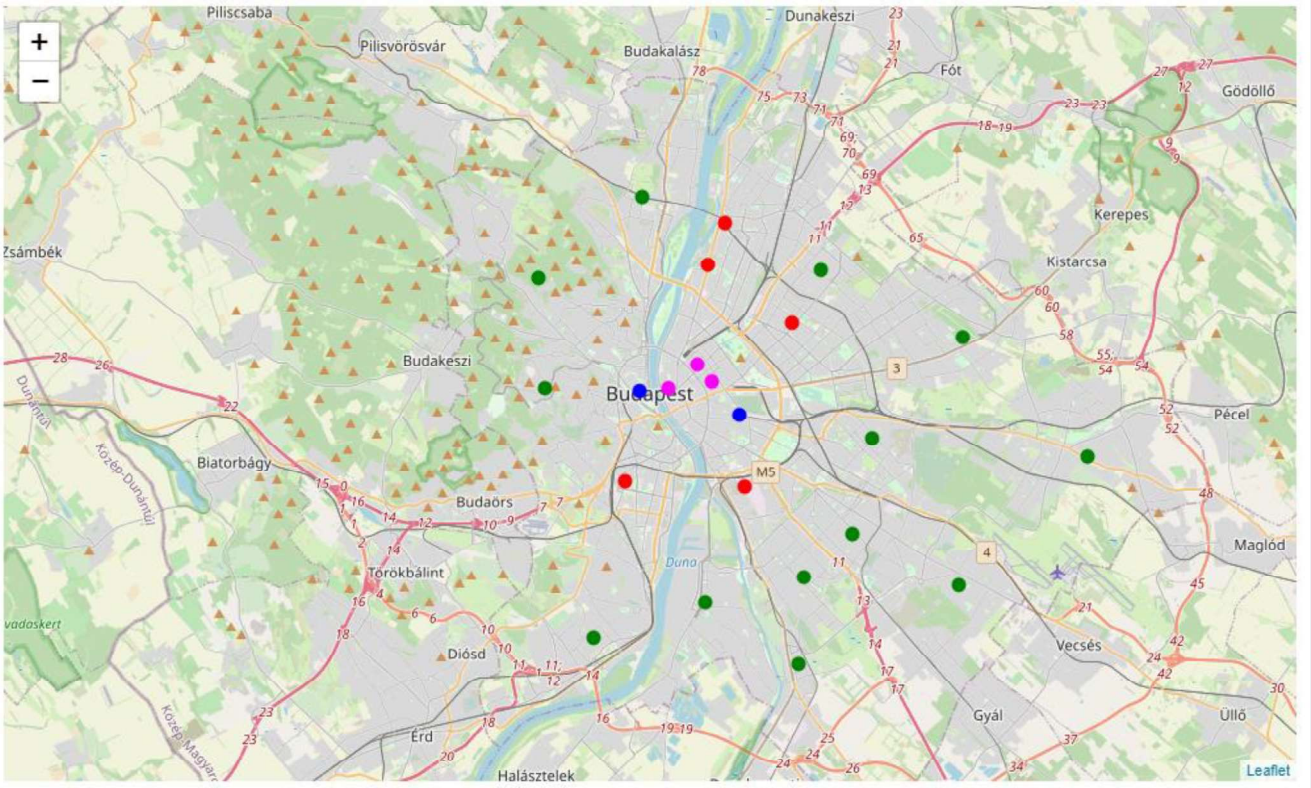


Results

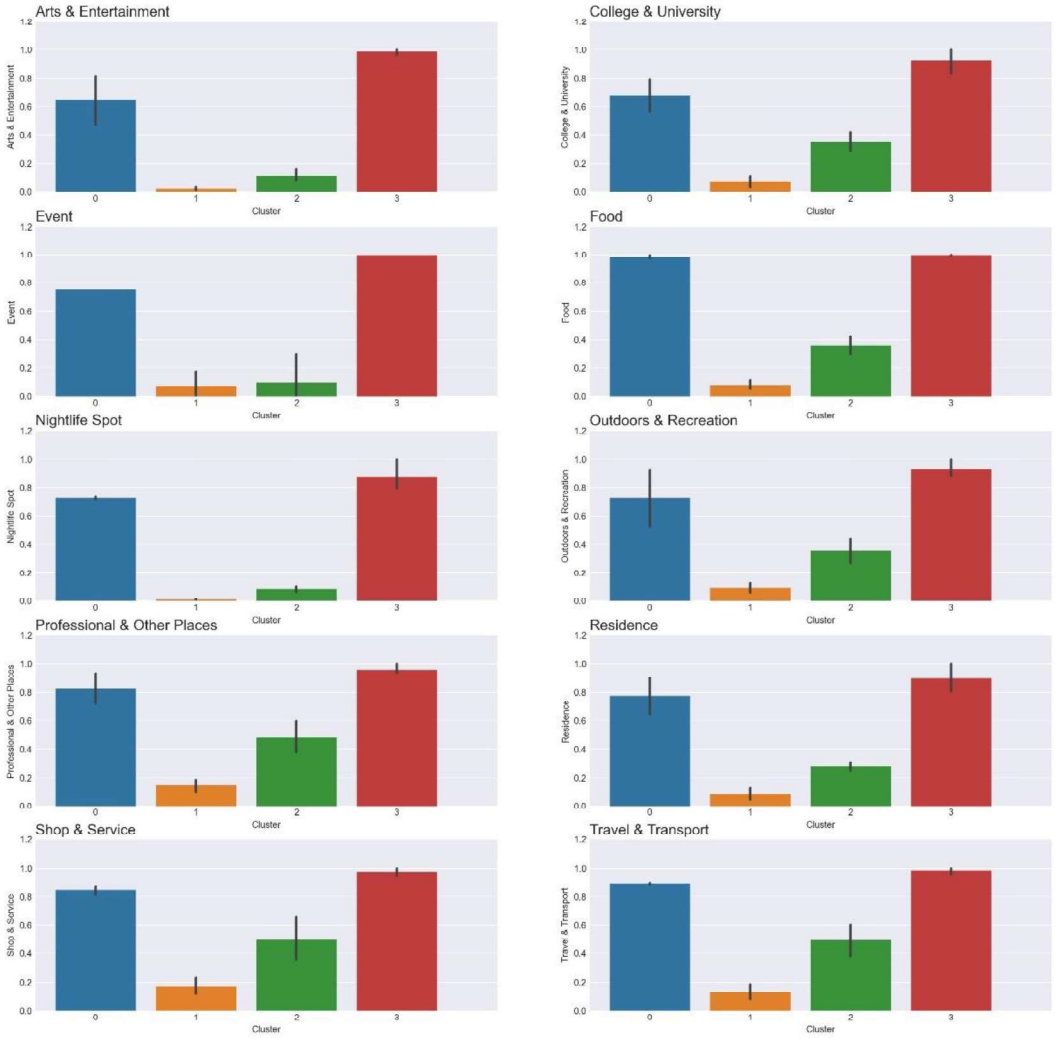
	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	...	Professional & Other Places	Residence	Shop & Service	Travel & Transport	Cluster
0	0.813793	0.602151	0.75	0.995833	0.744856	...	0.957055	0.630435	0.879581	0.905882	3
1	0.027586	0.010753	0.00	0.054167	0.000000	...	0.110429	0.021739	0.109948	0.047059	0
2	0.027586	0.096774	0.00	0.170833	0.024691	...	0.300613	0.239130	0.251309	0.129412	0
3	0.062069	0.279570	0.00	0.395833	0.094650	...	0.435583	0.260870	0.413613	0.341176	2
4	0.965517	1.000000	1.00	1.000000	1.000000	...	1.000000	0.913043	0.963351	0.976471	1
5	1.000000	0.763441	1.00	0.995833	0.794239	...	0.944785	0.847826	1.000000	0.947059	1

The obtained clusters can be interpreted and visualized on the map as below:

- ✓ Cluster 0 (Blue) - is the Moderately developed districts of the city.
- ✓ Cluster 1 (Green) - has low frequencies for all venue categories. They appear to be underdeveloped neighborhoods of the city.
- ✓ Cluster 2 (Red) - has average scores with more professional places and Transport services being the most popular hence falling under the developing neighborhoods bucket. These are mostly residential suburbs.
- ✓ Cluster 3 (Pink) - has consistently high frequencies for all venue categories. This is the most diversely developed part of city.



Cluster wise frequencies for each category:



Discussions

Additional Key Findings:

As per our analysis we can see that few neighborhoods of Budapest city can be classified as fully developed, moderately developed, developing and underdeveloped Neighborhoods based on the data from Foursquare API. But with Budapest city as one of the fast-growing tech hubs of Eastern Europe, we can expect the urban footprint of the Budapest city will expand as new growth areas keeps establishing every year.

Districts IV, IX, XI, XIII & XIV emerge as the major growing areas and might even see a transformation from Developing neighborhood to downtown/ uptown as its land usage and the number of tech companies in these districts is increasing in last few years.

Future scope:

It would be interesting to further study how we need to consider rental prices and the population data to Predict a growth pattern in these neighborhoods which will also lead us to identify early business and service opportunities in currently underdeveloped areas.

Conclusion :

Our aim is to aid anyone who visits Budapest city for either Employment, Business or Tourism to decide the place of stay based on what each neighborhood has to offer and their personal preferences and our analysis was able to successfully serve the purpose.

Thank You!!!