Features of Toronto Neighbourhoods for Prospective Home Owners

IBM Applied Data Science Capstone

By: Jerry Tan

20/03/2019

Problem Introduction

- Choosing an area to live in is a big, but difficult decision
  - Lack good information regarding facilities in living areas
  - Infeasible to physically scour all prospective neighbourhoods

- Very important to make a good decision
  - A foodie will enjoy higher quality of life if he / she stays in a neighbourhood with many dining venues
  - An athletic person would want to stay near sport venues etc.

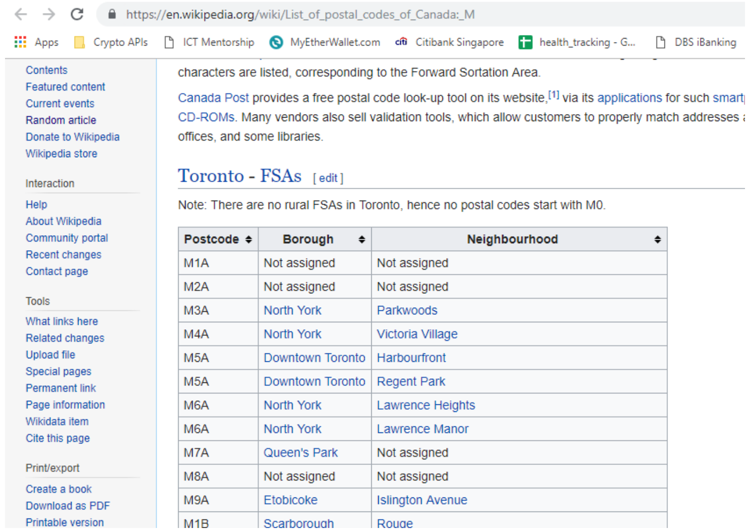
Solution

- Use Foursquare API to aggregate venue data, so as to make comparisons across different neighbourhoods based on latitude and longitude data

- Python API and statistical modules to be used to provide high-level insights regarding the concentration and types of facilities available within each neighbourhood in Toronto

Data

- Firstly, scrap list of neighbourhoods in Toronto from Wikipedia, convert data into a pandas dataframe



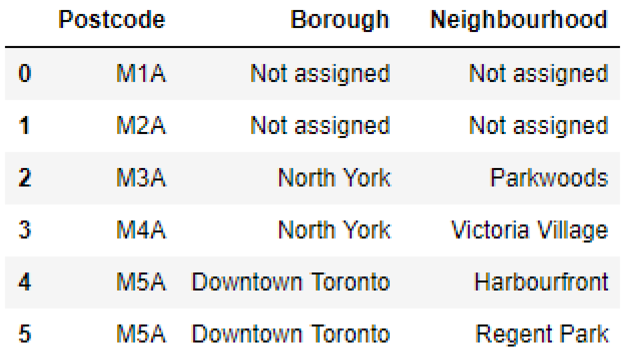
characters are listed, corresponding to the Forward Sortation Area.

Canada Post provides a free postal code look-up tool on its website,[1] via its applications for such smart
CD-ROMs. Many vendors also sell validation tools, which allow customers to properly match addresses
offices, and some libraries.

Toronto - FSAs [edit]

Note: There are no rural FSAs in Toronto, hence no postal codes start with M0.

Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M5A	Downtown Toronto	Regent Park
M6A	North York	Lawrence Heights
M6A	North York	Lawrence Manor
M7A	Queen's Park	Not assigned
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue
M1B	Scarborough	Rouge

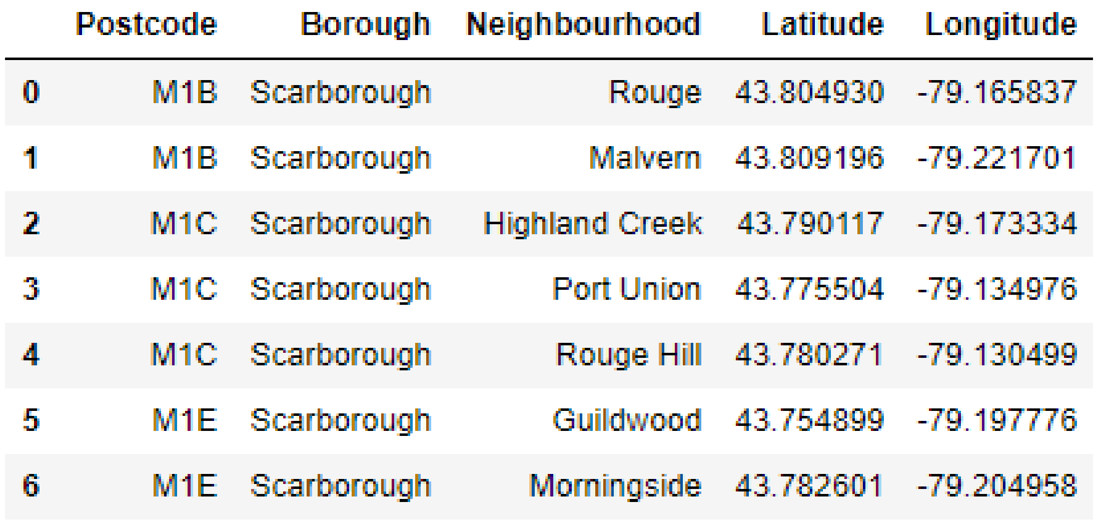


	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M5A	Downtown Toronto	Regent Park

- Data cleaning – remove rows where neighbourhood value is not assigned
- Initial 289 rows, cut to 211 rows after first round of cleaning

Data

- With list of neighbourhoods, use Geopy module in Python to generate latitudinal and longitudinal data

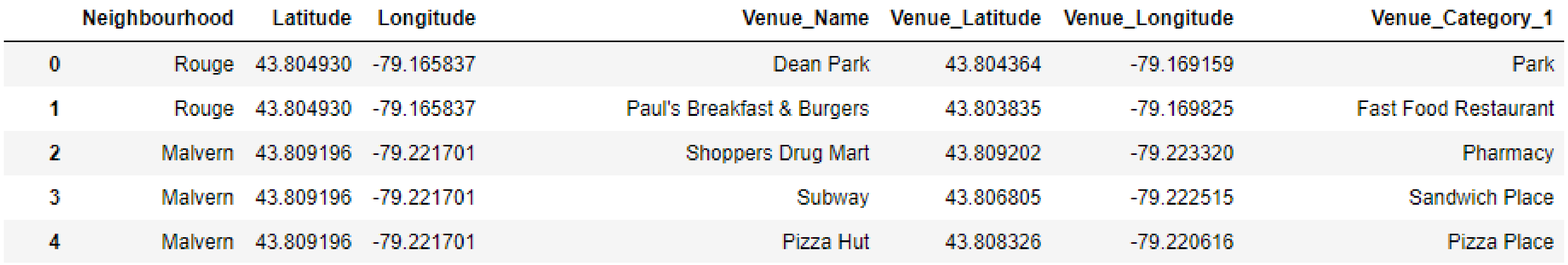


	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge	43.804930	-79.165837
1	M1B	Scarborough	Malvern	43.809196	-79.221701
2	M1C	Scarborough	Highland Creek	43.790117	-79.173334
3	M1C	Scarborough	Port Union	43.775504	-79.134976
4	M1C	Scarborough	Rouge Hill	43.780271	-79.130499
5	M1E	Scarborough	Guildwood	43.754899	-79.197776
6	M1E	Scarborough	Morningside	43.782601	-79.204958

- Data cleaning – remove neighbourhoods whereby latitude / longitude data is unavailable (reduced from 211 to 198 rows)

Data

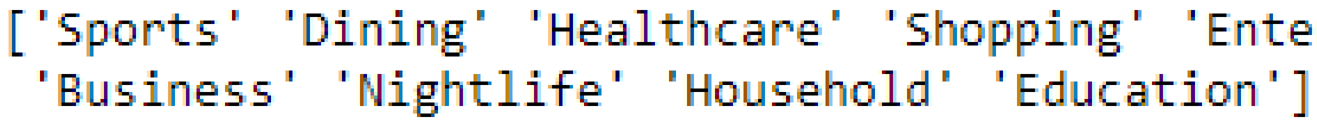
- With latitude and longitude data, we can now use Foursquare API to pull out venue data for each neighbourhood
- API call used the following parameters (500m from latitude / longitude point, 100 venues limit)



	Neighbourhood	Latitude	Longitude	Venue_Name	Venue_Latitude	Venue_Longitude	Venue_Category_1
0	Rouge	43.804930	-79.165837	Dean Park	43.804364	-79.169159	Park
1	Rouge	43.804930	-79.165837	Paul's Breakfast & Burgers	43.803835	-79.169825	Fast Food Restaurant
2	Malvern	43.809196	-79.221701	Shoppers Drug Mart	43.809202	-79.223320	Pharmacy
3	Malvern	43.809196	-79.221701	Subway	43.806805	-79.222515	Sandwich Place
4	Malvern	43.809196	-79.221701	Pizza Hut	43.808326	-79.220616	Pizza Place

Methodology

- As there were too many unique venue categories, to conduct more meaningful analysis, each unique venue category was bucketed into a higher-level venue category (Venue_Category_2), with values as listed below:

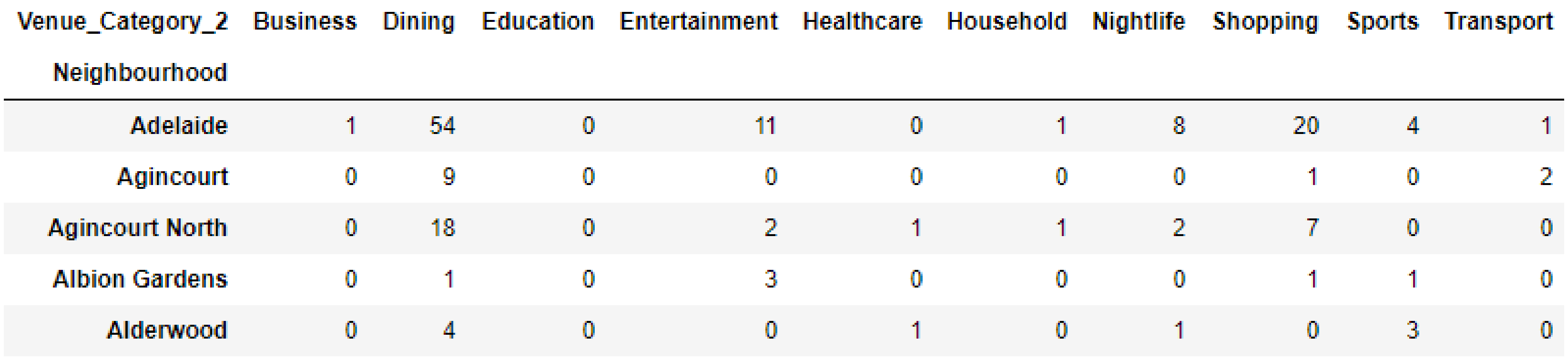


```
['Sports' 'Dining' 'Healthcare' 'Shopping' 'Entertainment' 'Transport'
 'Business' 'Nightlife' 'Household' 'Education']
```

- We then conducted statistic summaries to identify the most relevant locations for each category, as well as clustering analysis to identify neighbourhoods that were similar

Methodology

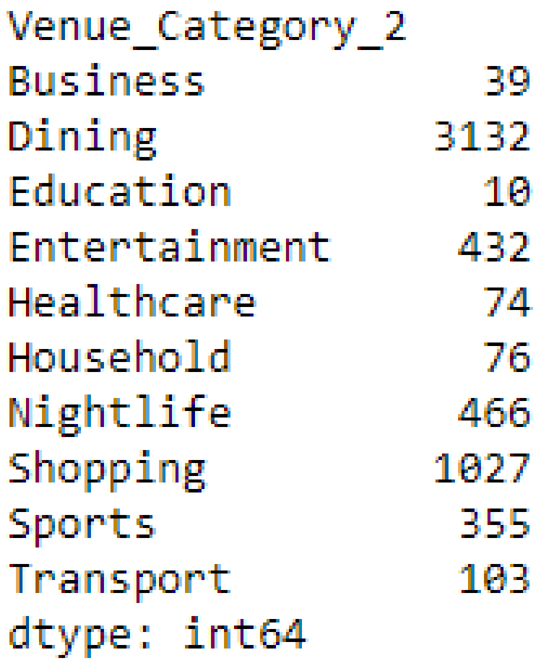
- Count of each higher-level category by neighbourhood was then found and saved in a dataframe as per below



Venue_Category_2	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
Neighbourhood										
Adelaide	1	54	0	11	0	1	8	20	4	1
Agincourt	0	9	0	0	0	0	0	1	0	2
Agincourt North	0	18	0	2	1	1	2	7	0	0
Albion Gardens	0	1	0	3	0	0	0	1	1	0
Alderwood	0	4	0	0	1	0	1	0	3	0

Results – Summary

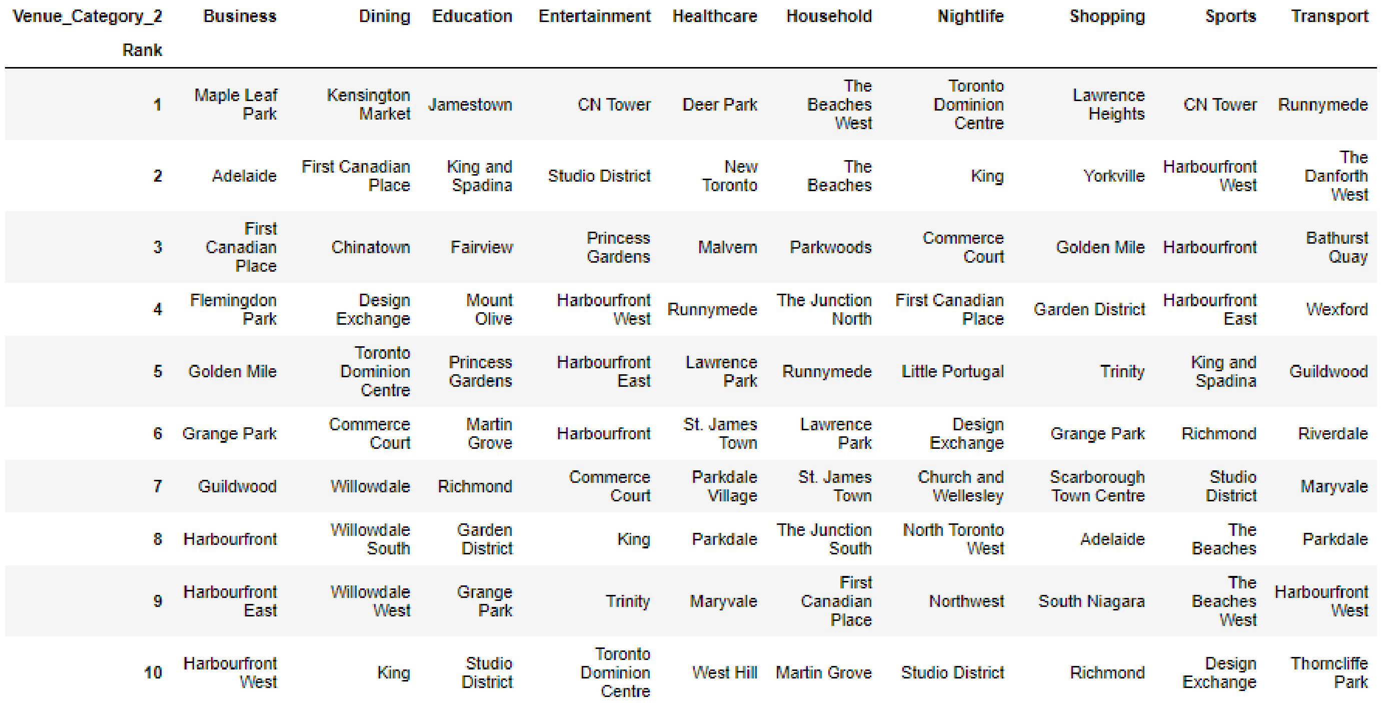
- The overall count of venues per category is summarized below:



```
Venue_Category_2
Business           39
Dining           3132
Education          10
Entertainment     432
Healthcare         74
Household          76
Nightlife         466
Shopping         1027
Sports            355
Transport         103
dtype: int64
```

Results – Summary

- Top 10 neighbourhoods for each venue category were found to be as follows:

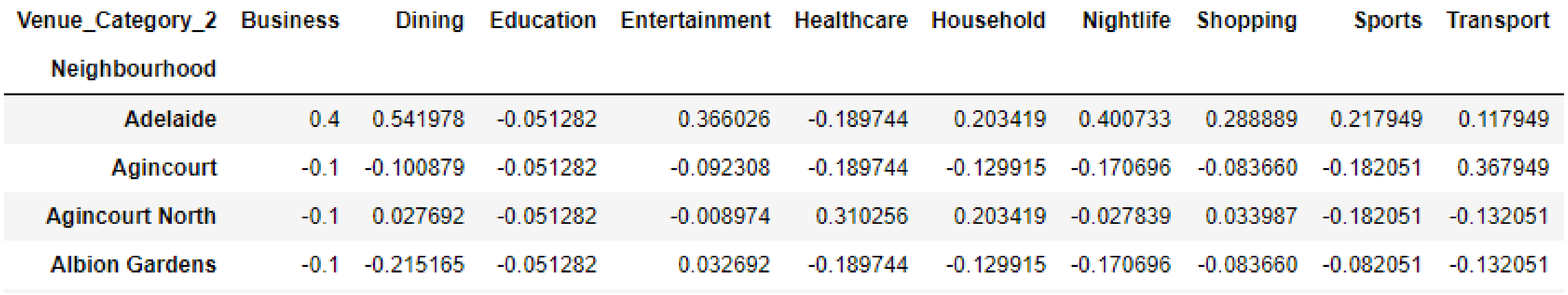


Venue_Category_2	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
Rank										
1	Maple Leaf Park	Kensington Market	Jamestown	CN Tower	Deer Park	The Beaches West	Toronto Dominion Centre	Lawrence Heights	CN Tower	Runnymede
2	Adelaide	First Canadian Place	King and Spadina	Studio District	New Toronto	The Beaches	King	Yorkville	Harbourfront West	The Danforth West
3	First Canadian Place	Chinatown	Fairview	Princess Gardens	Malvern	Parkwoods	Commerce Court	Golden Mile	Harbourfront	Bathurst Quay
4	Flemingdon Park	Design Exchange	Mount Olive	Harbourfront West	Runnymede	The Junction North	First Canadian Place	Garden District	Harbourfront East	Wexford
5	Golden Mile	Toronto Dominion Centre	Princess Gardens	Harbourfront East	Lawrence Park	Runnymede	Little Portugal	Trinity	King and Spadina	Guildwood
6	Grange Park	Commerce Court	Martin Grove	Harbourfront	St. James Town	Lawrence Park	Design Exchange	Grange Park	Richmond	Riverdale
7	Guildwood	Willowdale	Richmond	Commerce Court	Parkdale Village	St. James Town	Church and Wellesley	Scarborough Town Centre	Studio District	Maryvale
8	Harbourfront	Willowdale South	Garden District	King	Parkdale	The Junction South	North Toronto West	Adelaide	The Beaches	Parkdale
9	Harbourfront East	Willowdale West	Grange Park	Trinity	Maryvale	First Canadian Place	Northwest	South Niagara	The Beaches West	Harbourfront West
10	Harbourfront West	King	Studio District	Toronto Dominion Centre	West Hill	Martin Grove	Studio District	Richmond	Design Exchange	Thorncliffe Park

- Useful for people who are particularly interested in specific venue categories

Results – Clustering Analysis

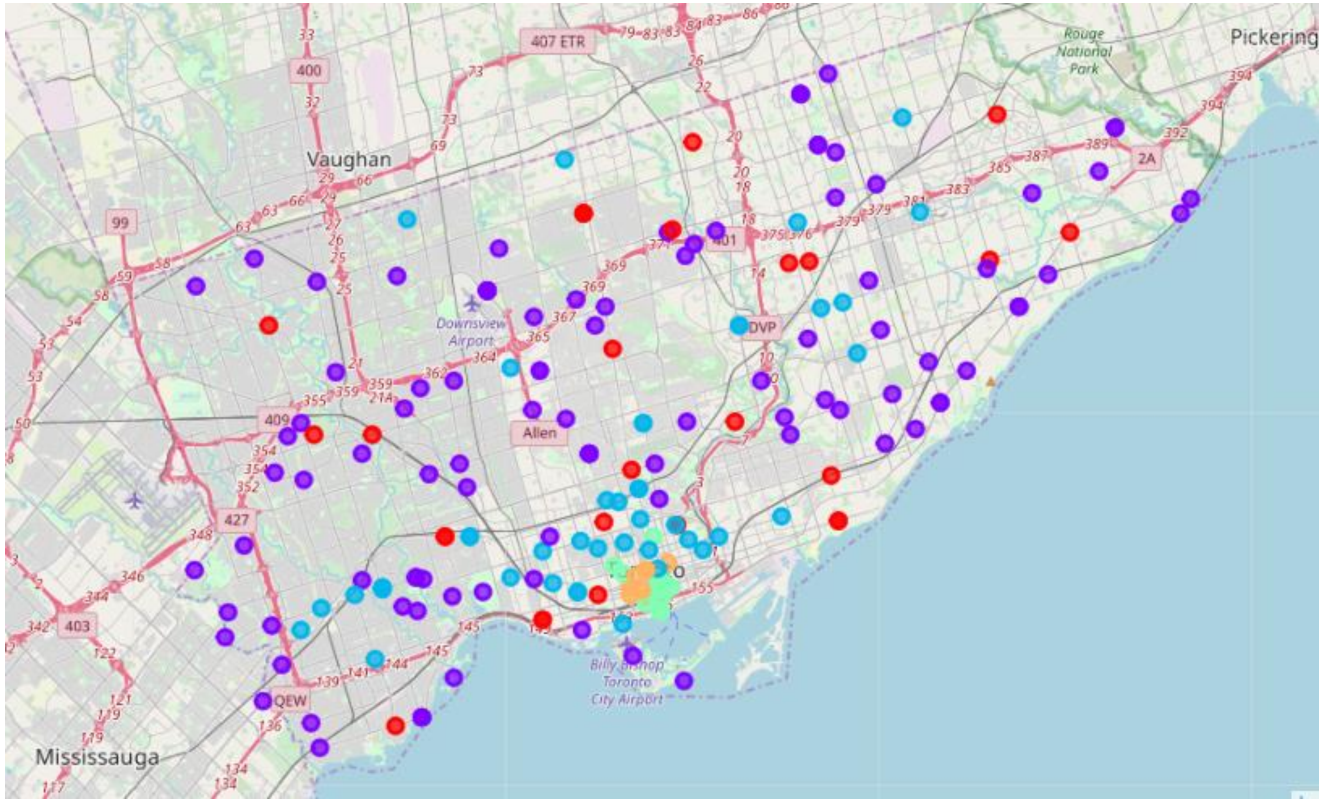
- The dataframe containing counts of each higher-level category by neighbourhood was normalized before conducting k-means clustering analysis, with k = 5



Venue_Category_2	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
Neighbourhood										
Adelaide	0.4	0.541978	-0.051282	0.366026	-0.189744	0.203419	0.400733	0.288889	0.217949	0.117949
Agincourt	-0.1	-0.100879	-0.051282	-0.092308	-0.189744	-0.129915	-0.170696	-0.083660	-0.182051	0.367949
Agincourt North	-0.1	0.027692	-0.051282	-0.008974	0.310256	0.203419	-0.027839	0.033987	-0.182051	-0.132051
Albion Gardens	-0.1	-0.215165	-0.051282	0.032692	-0.189744	-0.129915	-0.170696	-0.083660	-0.082051	-0.132051

Results – Clustering Analysis

- A map of Toronto was then created using Folium for easy visualization by the audience



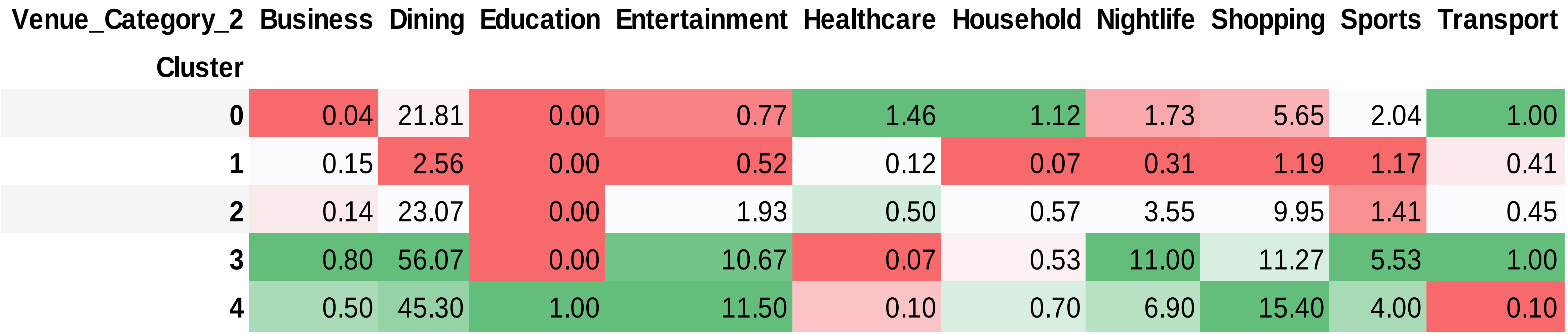
Legend

{Cluster: Color}

{0: Red, 1: Purple, 2: Blue, 3: Green, 4: Orange}

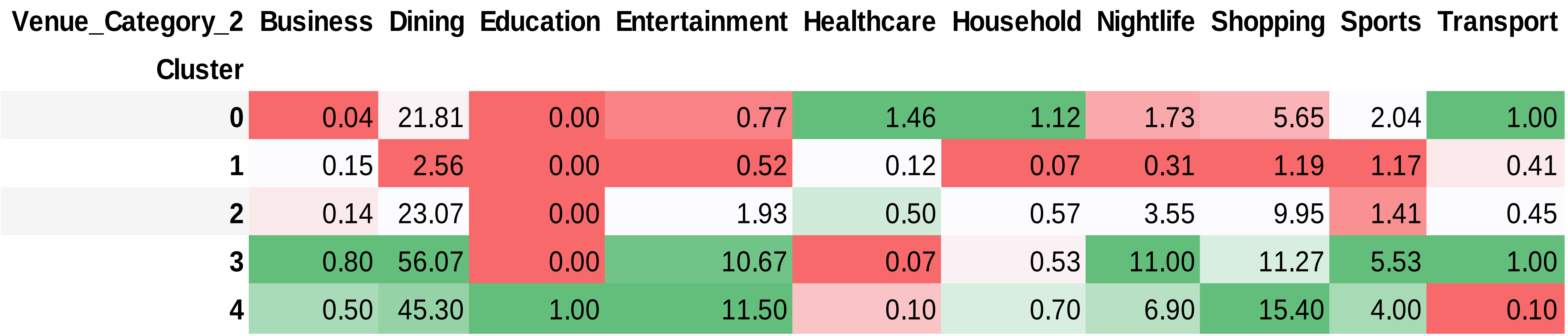
Results – Clustering Analysis

- The means of count for each category by cluster are summarized in the table below



Venue_Category_2	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
Cluster										
0	0.04	21.81	0.00	0.77	1.46	1.12	1.73	5.65	2.04	1.00
1	0.15	2.56	0.00	0.52	0.12	0.07	0.31	1.19	1.17	0.41
2	0.14	23.07	0.00	1.93	0.50	0.57	3.55	9.95	1.41	0.45
3	0.80	56.07	0.00	10.67	0.07	0.53	11.00	11.27	5.53	1.00
4	0.50	45.30	1.00	11.50	0.10	0.70	6.90	15.40	4.00	0.10

Discussion – Key Insights

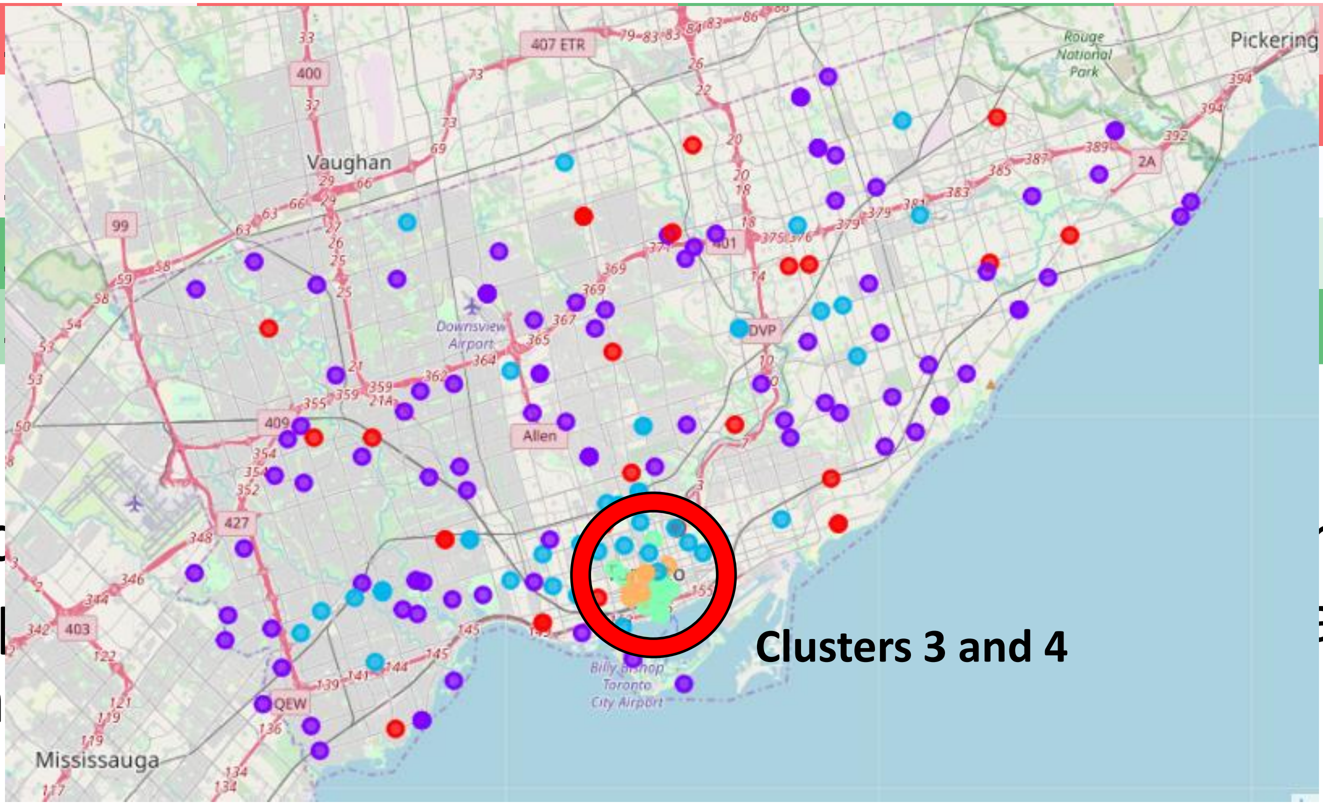


Venue_Category_2	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
Cluster										
0	0.04	21.81	0.00	0.77	1.46	1.12	1.73	5.65	2.04	1.00
1	0.15	2.56	0.00	0.52	0.12	0.07	0.31	1.19	1.17	0.41
2	0.14	23.07	0.00	1.93	0.50	0.57	3.55	9.95	1.41	0.45
3	0.80	56.07	0.00	10.67	0.07	0.53	11.00	11.27	5.53	1.00
4	0.50	45.30	1.00	11.50	0.10	0.70	6.90	15.40	4.00	0.10

- People who have kids may want to find housing in cluster 4 neighbourhoods, where there are educational facilities nearby

Discussion – Key Insights

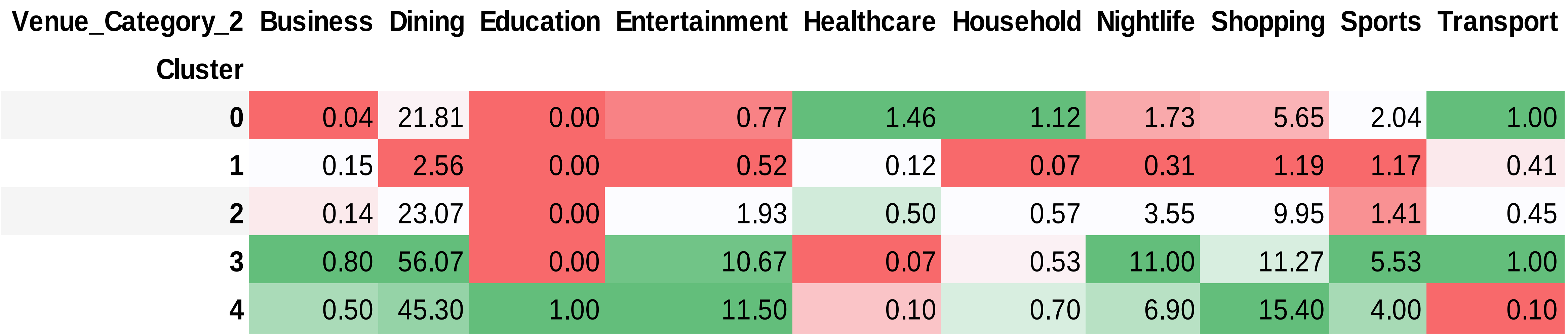
Venue_Category_2	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
Cluster										
0	0.							5.65	2.04	1.00
1	0.							1.19	1.17	0.41
2	0.							9.95	1.41	0.45
3	0.							11.27	5.53	1.00
4	0.							15.40	4.00	0.10



Clusters 3 and 4

- Cluster 3 anc ntration of dining, shop around city centre, migh

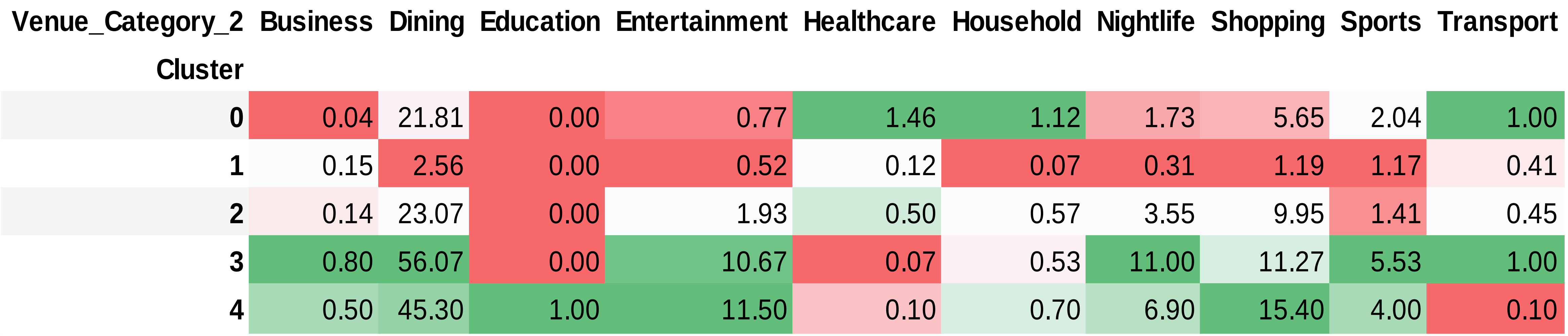
Discussion – Key Insights



Venue_Category_2 Cluster	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
0	0.04	21.81	0.00	0.77	1.46	1.12	1.73	5.65	2.04	1.00
1	0.15	2.56	0.00	0.52	0.12	0.07	0.31	1.19	1.17	0.41
2	0.14	23.07	0.00	1.93	0.50	0.57	3.55	9.95	1.41	0.45
3	0.80	56.07	0.00	10.67	0.07	0.53	11.00	11.27	5.53	1.00
4	0.50	45.30	1.00	11.50	0.10	0.70	6.90	15.40	4.00	0.10

- Avoid cluster 1 neighbourhoods which are seemingly lacking in all sorts of facilities.

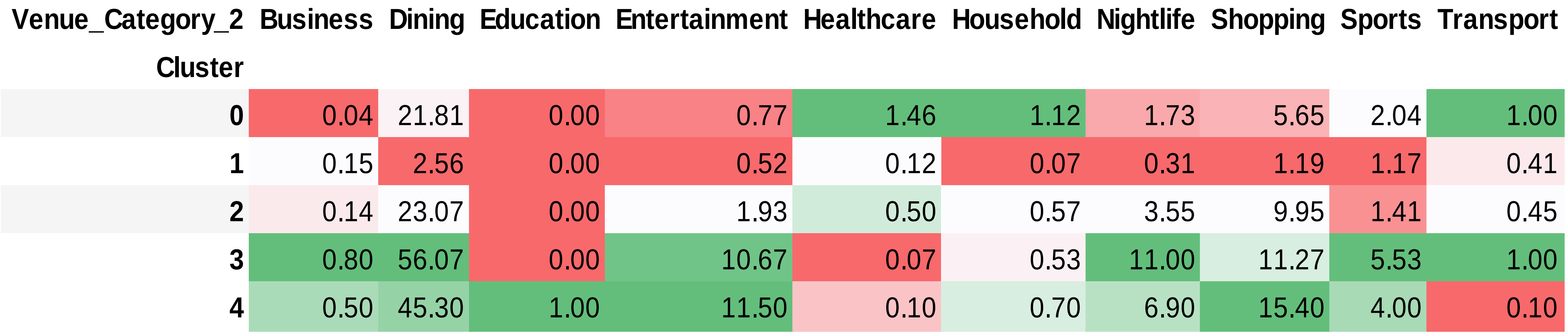
Discussion – Key Insights



Venue_Category_2	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
Cluster										
0	0.04	21.81	0.00	0.77	1.46	1.12	1.73	5.65	2.04	1.00
1	0.15	2.56	0.00	0.52	0.12	0.07	0.31	1.19	1.17	0.41
2	0.14	23.07	0.00	1.93	0.50	0.57	3.55	9.95	1.41	0.45
3	0.80	56.07	0.00	10.67	0.07	0.53	11.00	11.27	5.53	1.00
4	0.50	45.30	1.00	11.50	0.10	0.70	6.90	15.40	4.00	0.10

- Cluster 0 and cluster 2 are likely to be the cheaper alternatives for prospective home owners

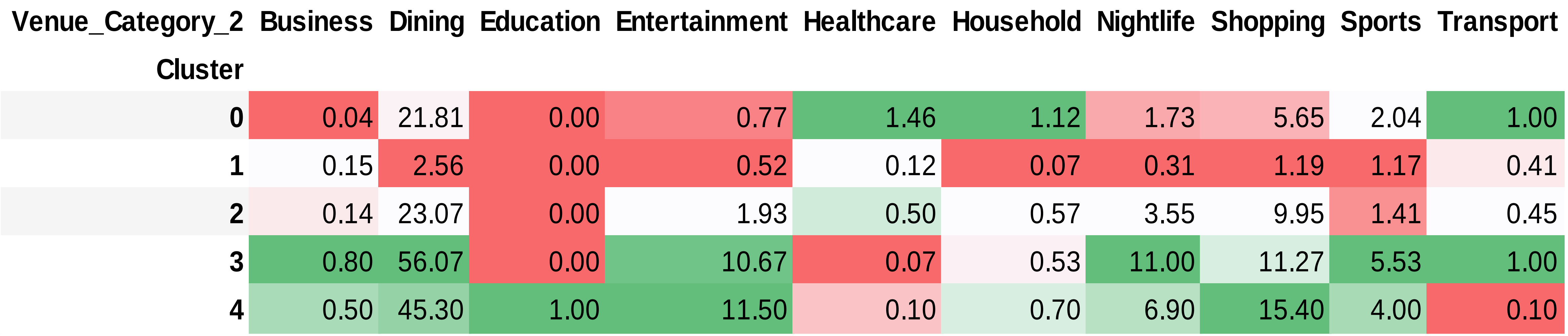
Discussion – Key Insights



Venue_Category_2	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
Cluster										
0	0.04	21.81	0.00	0.77	1.46	1.12	1.73	5.65	2.04	1.00
1	0.15	2.56	0.00	0.52	0.12	0.07	0.31	1.19	1.17	0.41
2	0.14	23.07	0.00	1.93	0.50	0.57	3.55	9.95	1.41	0.45
3	0.80	56.07	0.00	10.67	0.07	0.53	11.00	11.27	5.53	1.00
4	0.50	45.30	1.00	11.50	0.10	0.70	6.90	15.40	4.00	0.10

- Families may wish to live in cluster 0 rather than cluster 2 neighbourhoods given the higher concentration of healthcare, household and sports facilities

Discussion – Key Insights



Venue_Category_2	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
Cluster										
0	0.04	21.81	0.00	0.77	1.46	1.12	1.73	5.65	2.04	1.00
1	0.15	2.56	0.00	0.52	0.12	0.07	0.31	1.19	1.17	0.41
2	0.14	23.07	0.00	1.93	0.50	0.57	3.55	9.95	1.41	0.45
3	0.80	56.07	0.00	10.67	0.07	0.53	11.00	11.27	5.53	1.00
4	0.50	45.30	1.00	11.50	0.10	0.70	6.90	15.40	4.00	0.10

- Younger people may want to consider cluster 2 over cluster 0 neighbourhoods, given the higher availability of entertainment, nightlife and shopping venues

Conclusion

- This research is a simple but effective way of helping prospective home owners and renters get better knowledge of facilitates available around potential neighbourhoods
- Ways to improve on it:
  - Increase radius and limit when pulling Foursquare API, allow audience to visualize concentration of facilities within different ranges (500m, 1000m 2000m etc.)
  - Augment research with property and rental prices – price plays a huge role in the audiences’ decisions