Business opportunities in Seattle, WA

By Suresh Setty, April 25, 2020.

Introduction

Background

Seattle is a seaport city on the West Coast of the United States. According to census data released in 2018, the Seattle metropolitan area population stands at 3.98 million, and ranks as the 15th largest in the US. The Seattle area developed into a technology center from the 1980s onwards with companies like Microsoft becoming established in the region; internet retailer Amazon was also founded in Seattle, and a major airline Alaska Airlines is based in Seattle as well. Additionally, a lot of new software and biotechnology companies have been moving into the area, contributing to the area's vast economic and population growth.

Business Environment and Opportunities

The objective of this study is to identify various Seattle neighborhoods with good potential for new business opportunities. This is done by investigating the popular venues in each neighborhood to gain insights into the current business environment there. These insights in conjunction with house prices in the neighborhood could throw light at potential business opportunities.

Data

Data Acquisition

We will be using the following sources for acquiring data needed for this study.

- There are a lot of sources on the internet identifying and classifying various Seattle neighborhoods. After some research, the data from Zillow.com seemed to be more comprehensive. So, the data from this Zillow page was used - https://www.zillow.com/east-queen-anne-seattle-wa/home-values/ .

- Once we have the list of neighborhoods, we use geopy and Nominatim Geocoding service to obtain latitude and longitude values of these neighborhoods.

- We use developer access to Foursqaure (https://foursquare.com) to explore the venues in all the neighborhoods. All venues in Foursquare are categorized into 10 main categories - Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport. For each neighborhood, we will get the number of venues in all categories.

- Once we have the venue data for all neighborhoods, we can use k-means clustering algorithm to segment neighborhoods. We can then analyze the segments for any patterns and look for potential business opportunities.

Data Wrangling

- The neighborhood data from Zillow.com has a lot more information than what's needed. So, only the relevant information was kept.
- It was realized that some ZRI (rental index) values were missing from the Zillow.com data, so those missing values were replaced with the mean of all other ZRI values.
- Nominatim Geocoding service couldn't provide latitude and longitude values for a couple of neighborhoods, so those were entered manually after a google search.
- After obtaining latitude and longitude values for the remaining neighborhoods (using Nominatim), it was realized that there were some duplicates in the coordinate values. In Zillow data, the neighborhoods with the duplicate coordinate values had considerable differences in their ZHVI (home value index) values, so instead of deleting or merging these neighborhoods, the choice was made to replace the duplicate coordinate values manually with data from google search just for these neighborhoods.

Final Data

After acquiring and cleaning up the data like mentioned above, the resulting data looks like the table below.

	Neighborhood	Latitude	Longitude	Current ZHVI	Current ZRI	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Servic
0	Downtown	47.604872	-122.333458	815000	2656.000000	73	45	6	155	100	126	184	66	17
1	Adams	47.565271	-122.279546	723500	2575.000000	4	2	1	29	6	20	48	1	6
2	Admiral	47.581195	-122.386546	834900	2840.000000	4	2	0	26	9	21	71	10	4
3	Alki	47.576209	-122.409851	880000	2842.000000	5	0	1	25	5	18	6	6	1
4	Arbor Heights	47.512899	-122.381359	624000	2350.000000	5	3	0	4	1	4	13	1	
5	Atlantic	47.590493	-122.324313	714200	2692.701299	21	6	2	105	31	50	84	7	12
6	Beacon Hill	47.552600	-122.300900	641100	2453.000000	2	2	0	15	1	5	10	2	1
7	Belltown	47.613231	-122.345361	572000	2302.000000	97	49	7	213	105	135	145	88	17
8	Bitter Lake	47.726236	-122.348764	607800	2266.000000	9	2	0	57	5	16	51	10	7
9	Brighton	47.546210	-122.275679	577500	2380.000000	7	4	0	29	4	9	45	4	3

Methodology

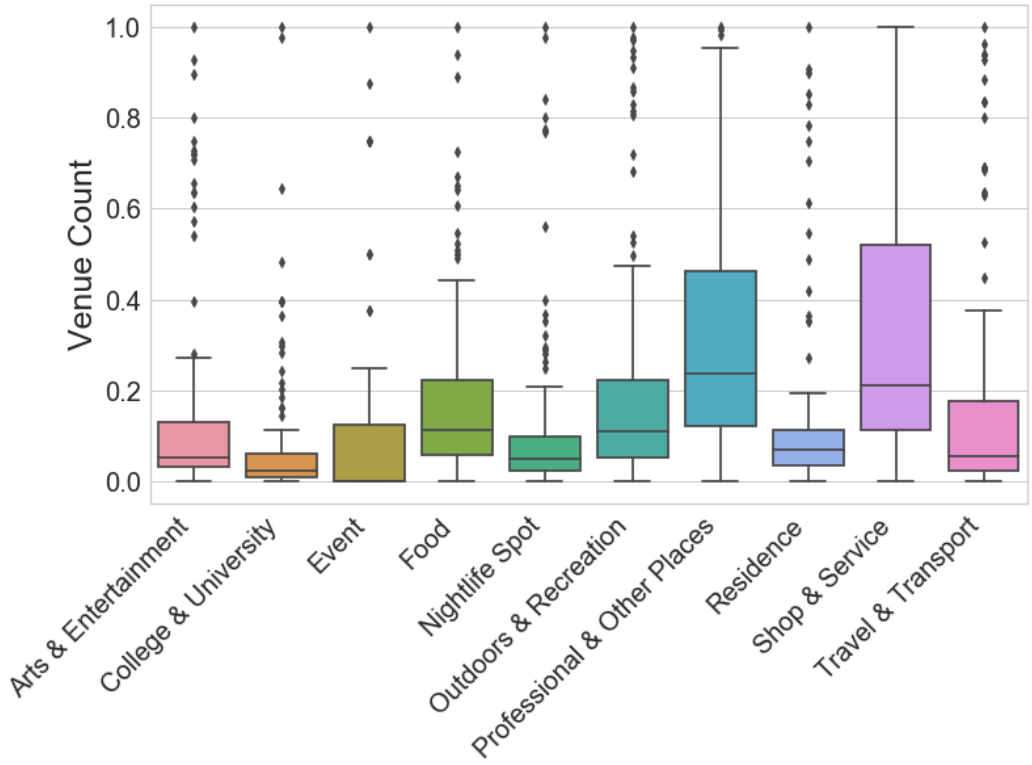
The methodology will include

- Data acquisition (as mentioned above), data exploration and wrangling.
- Using K-means clustering algorithm, perform neighborhood segmentation.
- Analyze the clusters using various statistical methods/tools.
- Understand the current business environment in the clusters and look for clusters that present growth opportunities.

Analysis

Scaling

The data is first normalized using MinMaxScaler and the category venue data is plotted below using boxplot. From the below box plot, we can see that some venue categories have a lot of outliers. We can also see that there are a lot more professional and shop & service venues across neighborhoods.



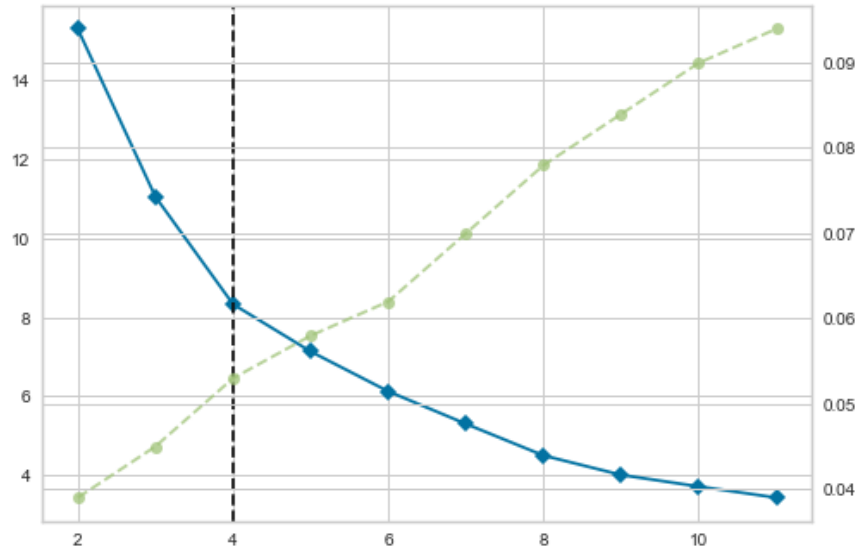
After looking at the mean of all venue categories, the 'Event' category is dropped from the data as the average number of 'event' venues is less than zero. The scaled data looks like below

	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	0.750000	0.362903	0.726415	0.800	0.911111	0.952632	0.750000	0.911917	1.000000
1	0.031250	0.016129	0.132075	0.048	0.125926	0.236842	0.011364	0.316062	0.030303
2	0.031250	0.016129	0.117925	0.072	0.133333	0.357895	0.113636	0.227979	0.012121
3	0.041667	0.000000	0.113208	0.040	0.111111	0.015789	0.068182	0.062176	0.006061
4	0.041667	0.024194	0.014151	0.008	0.007407	0.052632	0.011364	0.015544	0.018182

K-Means

Using the Elbow method, the optimum K is determined which turns out to be 4 in this case.

```
KElbowVisualizer(ax=<matplotlib.axes._subplots.AxesSubplot object at 0x0000021D2808DF08>,
                 k=None, locate_elbow=True, metric='distortion', model=None,
                 timings=True)
```

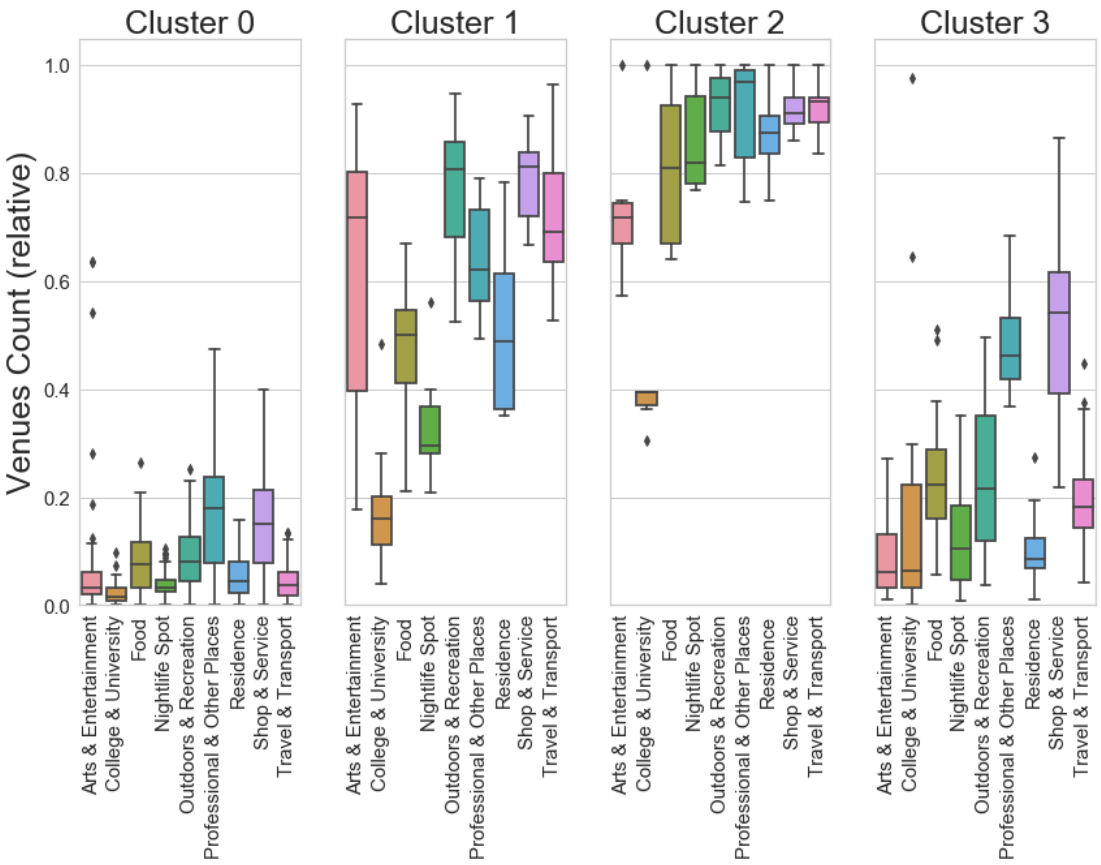


Results

Then, the K-means is performed on the scaled data and the cluster information is added back to the original data (see below).

Latitude	Longitude	Current ZHVI	Current ZRI	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport	Cluster
7.604872	-122.333458	815000	2656.000000	73	45	6	155	100	126	184	66	178	166	2
7.565271	-122.279546	723500	2575.000000	4	2	1	29	6	20	48	1	63	6	0
7.581195	-122.386546	834900	2840.000000	4	2	0	26	9	21	71	10	46	3	0
7.576209	-122.409851	880000	2842.000000	5	0	1	25	5	18	6	6	14	2	0
7.512899	-122.381359	624000	2350.000000	5	3	0	4	1	4	13	1	5	4	0
7.590493	-122.324313	714200	2692.701299	21	6	2	105	31	50	84	7	121	61	3
7.552600	-122.300900	641100	2453.000000	2	2	0	15	1	5	10	2	17	5	0
7.613231	-122.345361	572000	2302.000000	97	49	7	213	105	135	145	88	173	156	2
7.726236	-122.348764	607800	2266.000000	9	2	0	57	5	16	51	10	79	11	0
7.546210	-122.275679	577500	2380.000000	7	4	0	29	4	9	45	4	36	4	0

The venue data for each cluster is then plotted using a box plot as shown below. From the plot, we can see that cluster 2 has the most venues for all categories, followed by cluster 1, cluster 3 and cluster 0 in that order.



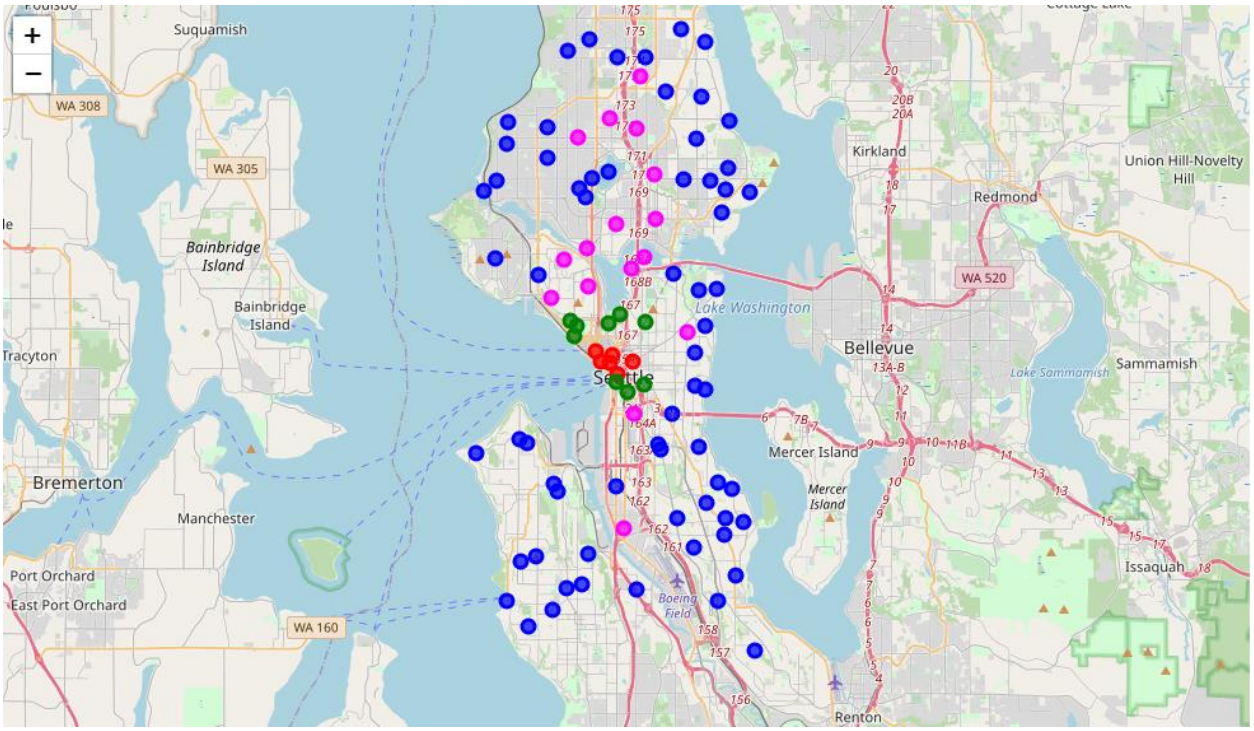
A table is created showing the number of neighborhoods, the average number of venues in each category, and average house/rent prices for each of the 4 clusters. Each cluster is also assigned a color for easy visualization on the map.

Cluster	Color	#Neighorhoods	Current ZHVI	Current ZRI	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	blue	65	856680.0	2699.7	8.1	2.4	0.2	18.1	5.0	14.9	36.7	4.5	31.4	8.4
1	green	9	674644.4	2604.8	59.3	23.4	2.1	101.8	41.4	105.7	125.7	45.2	154.9	119.4
2	red	6	685233.3	2509.1	71.7	59.0	5.8	172.3	107.5	127.7	176.2	76.8	179.5	153.0
3	magenta	16	838925.0	2782.6	10.0	23.1	0.2	53.6	15.6	34.8	94.8	9.1	101.5	35.1

For further analysis, we can also look at correlation between home/rent values and various categories as show below.

	Current ZHVI	Current ZRI	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
Current ZHVI	1.000000	0.859908	-0.979281	-0.695782	-0.788724	-0.869194	-0.783058	-0.975036	-0.854180	-0.906843	-0.909006	-0.966017
Current ZRI	0.859908	1.000000	-0.932089	-0.730055	-0.924550	-0.867740	-0.890665	-0.888682	-0.758437	-0.939943	-0.723032	-0.889972
Arts & Entertainment	-0.979281	-0.932089	1.000000	0.793641	0.895402	0.936042	0.887059	0.992015	0.896331	0.971403	0.909832	0.989482
College & University	-0.695782	-0.730055	0.793641	1.000000	0.928610	0.956960	0.958629	0.837229	0.953529	0.900585	0.871366	0.856890
Event	-0.788724	-0.924550	0.895402	0.928610	1.000000	0.964845	0.995737	0.892336	0.893559	0.973831	0.814117	0.905913
Food	-0.869194	-0.867740	0.936042	0.956960	0.964845	1.000000	0.978061	0.957214	0.978598	0.984507	0.937182	0.967405
Nightlife Spot	-0.783058	-0.890665	0.887059	0.958629	0.995737	0.978061	1.000000	0.895659	0.923628	0.971281	0.844563	0.910591
Outdoors & Recreation	-0.975036	-0.888682	0.992015	0.837229	0.892336	0.957214	0.895659	1.000000	0.940545	0.971102	0.954754	0.999294
Professional & Other Places	-0.854180	-0.758437	0.896331	0.953529	0.893559	0.978598	0.923628	0.940545	1.000000	0.935059	0.978550	0.950177
Residence	-0.906843	-0.939943	0.971403	0.900585	0.973831	0.984507	0.971281	0.971102	0.935059	1.000000	0.901050	0.977453
Shop & Service	-0.909006	-0.723032	0.909832	0.871366	0.814117	0.937182	0.844563	0.954754	0.978550	0.901050	1.000000	0.957289
Travel & Transport	-0.966017	-0.889972	0.989482	0.856890	0.905913	0.967405	0.910591	0.999294	0.950177	0.977453	0.957289	1.000000

All Seattle neighborhoods based on their cluster are shown on the map below – cluster 0 (blue), cluster 1 (green), cluster 2(red), and cluster 3(magenta).



Observations

- From the above correlation table too, we can see that cluster 2 (red) has the most venues for all categories, followed by cluster 1 (green) for almost all categories.
- Cluster 0 (blue) and cluster 3 (magenta) have the least number of venues. But these clusters have relatively higher house values compared to the other two clusters. This can also be observed from the correlation table as well - house prices are negatively correlated to most venue categories.
- From the above map, it is also interesting to see how neighborhoods in cluster 1 and cluster 2 (cluster with the most venues) are concentrated closer to each other.
- Based on all these observations, the clusters could be classified based on the number of venues and home values using 4 levels (least, slight, moderate, high) as follows
    - Cluster 0 (blue) – Least developed, Highly expensive neighborhoods.
    - Cluster 1 (green) – Moderately developed, least expensive neighborhoods.
    - Cluster 2 (red) – Highly developed, slightly expensive neighborhoods.
    - Cluster 3 (magenta) – Slightly developed, moderately expensive neighborhoods.

Note: The input for K-means algorithm (as shown in this notebook) was the data without home and rental values (just the category venue count). However, k-means was also performed on data that included home and rental values, but the result yielded (segmentation) was identical and hence that analysis wasn't discussed or included in this final notebook.

Conclusion

The neighborhoods in cluster 0 (blue) and cluster 3 (magenta) present most opportunities for anyone wanting to start a business as they are clearly undeveloped in terms of business venues. But, on the flipside, the neighborhoods in these two clusters might be expensive (as reflected by house values) to start a business. So, one should keep this in mind and do a thorough ROI (Return on Investment) analysis as location expenses and other operating costs could be higher in these neighborhoods.