Discovering Similar Neighborhoods in New Cities

Aaron Oldre

April 26, 2020

1 Introduction

1.1 Description & Discussion of the Background

Seattle is a wonderful city, but what if I had to relocate another city? I am offered the option of going to Austin, Denver, Phoenix, or Portland. I like Seattle and the neighborhood I live in with the restaurants, breweries, coffee shops, and bars all within walking distance of my house. Can I find neighborhoods in these new cities with these same types of establishments and similar housing costs and area demographics?

I will use my data science powers to generate a few most promising neighborhoods based on these criteria. Advantages of each area will then be clearly expressed so that best possible final location that will feel like home.

1.2 Data Description

Based on definition of the problem, factors that will influence the decision are:

- similarity of venues across neighborhoods
- population density
- average housing cost

I will be using a polygon shapefile includes neighborhood boundaries supplied by Zillow.com, the Los Angeles Times and the City of Denver together with a number of demographic variables from the 2010 U.S. census [1]

Following data sources will be needed to extract/generate the required information:

- centers of neighborhoods are defined in the shapefile
- neighborhood boundaries are defined in the shapefile
- neighborhood demographics are defined in the shapefile
- city centers with be obtained using **Geopy Geocoders**

- number of venues and their type and location in every neighborhood will be obtained using **Foursquare API**

The geo_json is a list of dictionaries. There are also more data fields than are needed for the data modeling. Only the following fields are needed. The remaining fields can be discarded.

- state: State
- city: City
- name: Neighborhood Name
- x: Neighborhood Centroid Latitude
- y: Neighborhood Centroid Longitude
- avg_hval: Average Home Value
- pct_own: Percentage of Owned Homes
- pct_rent: Percentage of Rented Homes
- medage_cy: Median Age of Residents
- popdensity: Population Density
- geometry: Neighborhood Boundary Box

2 Methodology

I utilized the Foursquare API to explore the neighborhoods. The API limits the results at 100 venues, and I set the radius of the search to and the radius 800 meters (~.5 mile) from the defined centroid of each neighborhood. Here is a head of the list venues name, category, latitude and longitude information from Foursquare API along with the city and neighborhood they are from.

	Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category	City
0	Capitol Hill	Hudson Hill	39.736960	-104.979407	Cocktail Bar	Denver
1	Capitol Hill	Wax Trax Records	39.736810	-104.979230	Record Shop	Denver
2	Capitol Hill	Jelly Cafe	39.736791	-104.979730	Breakfast Spot	Denver
3	Capitol Hill	Snarf's Sandwiches	39.733891	-104.975320	Sandwich Place	Denver
4	Capitol Hill	Wokano Asian Bistro	39.733414	-104.974988	Asian Restaurant	Denver

Table 1. Foursquare API call results

I used python **folium** library to visualize geographic details of Ballard in Seattle, the neighborhood we are trying to match in other cities. First a created a map of the neighborhood based on the centroid defined in the dataset and marked all the venues found in the **Foursquare API.**

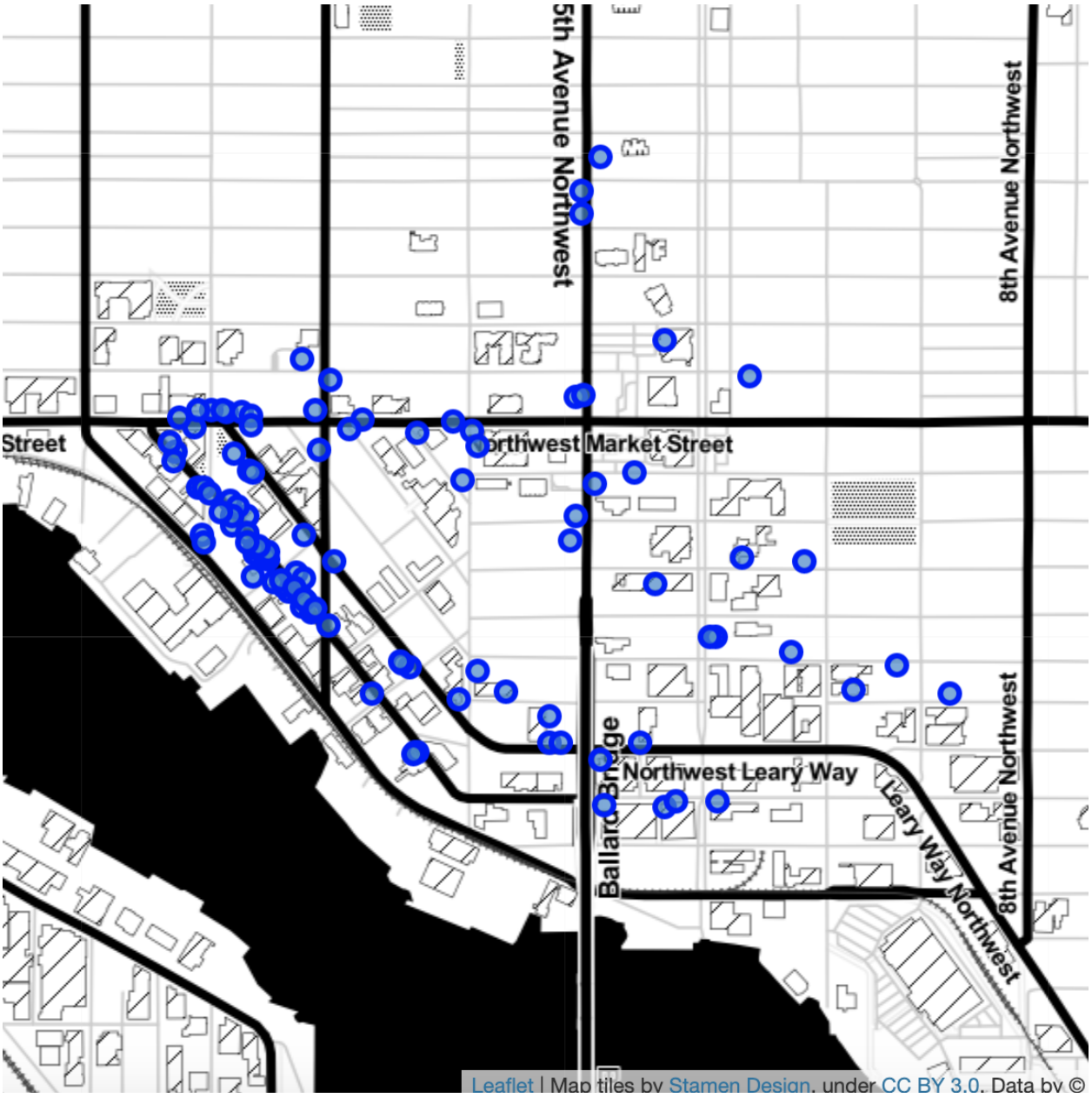

Figure 1. Venues overlaid on Ballard map.

All of the venues where converted from categorical data to numerical data using Pandas get dummies and group by methods. This created a dataframe with the percent of each venue in each neighborhood. This was merged with census data to be used in clustering.

	Neighborhood	ATM	Accessories Store	Adult Boutique	Advertising Agency	African Restaurant	Airport	Airport Gate	Airport Lounge	Airport Service	...	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Wi
0	Admiral_Seattle	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.020408	0.0	
1	Ahwatukee Foothills_Phoenix	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	
2	Alahambra_Phoenix	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	
3	Alameda_Portland	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	
4	Alki_Seattle	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
289	Windsor_Denver	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	
290	Woodlawn_Portland	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	
291	Woodstock_Portland	0.018519	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.018519	0.0	
292	Wooten_Austin	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	
293	Zilker_Austin	0.010000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.020000	0.0	

Table 2. Venue results categorized by city and neighborhood.

In Woodstock, Portland we see 1.86% of venues found on Foursquare are ATMs and Ziker, Austin 2% are Wine Bars.

I used unsupervised learning K-means algorithm to cluster the neighborhoods. K-Means algorithm is one of the most common cluster methods of unsupervised learning. The Elbow Method was used to determine the optimal number of clusters into which the neighborhood data can be clustered. KElbowVisualizer looks for strong inflection points at in the curve of distortion scores to determine the optimal number of clusters to use to segregate the data.

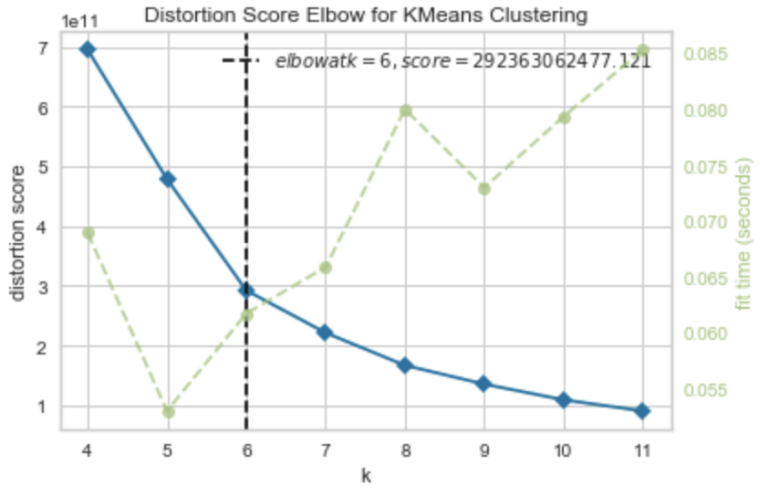

Figure 2. Distortion scores of k (blue) with computation times (green) in elbow score for optimum number of clusters (black).

Using the elbow method, we are seeing 6 different clusters for the neighborhood data. I used this value of k to find clusters of similar neighborhoods among the 5 cities.

3 Results

The clusters from each city are shown below. The colors identify the clusters labels and are continuous for each city.

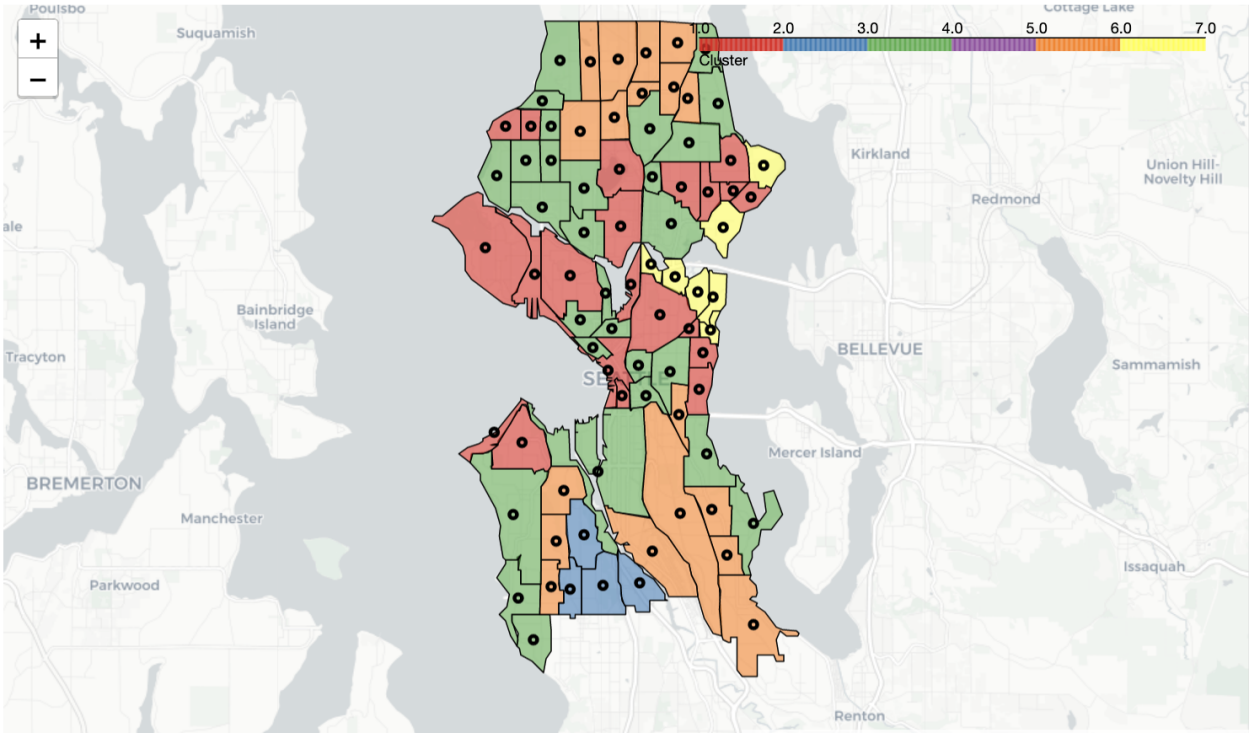

Figure 3. Seattle, WA (Ballard is green).

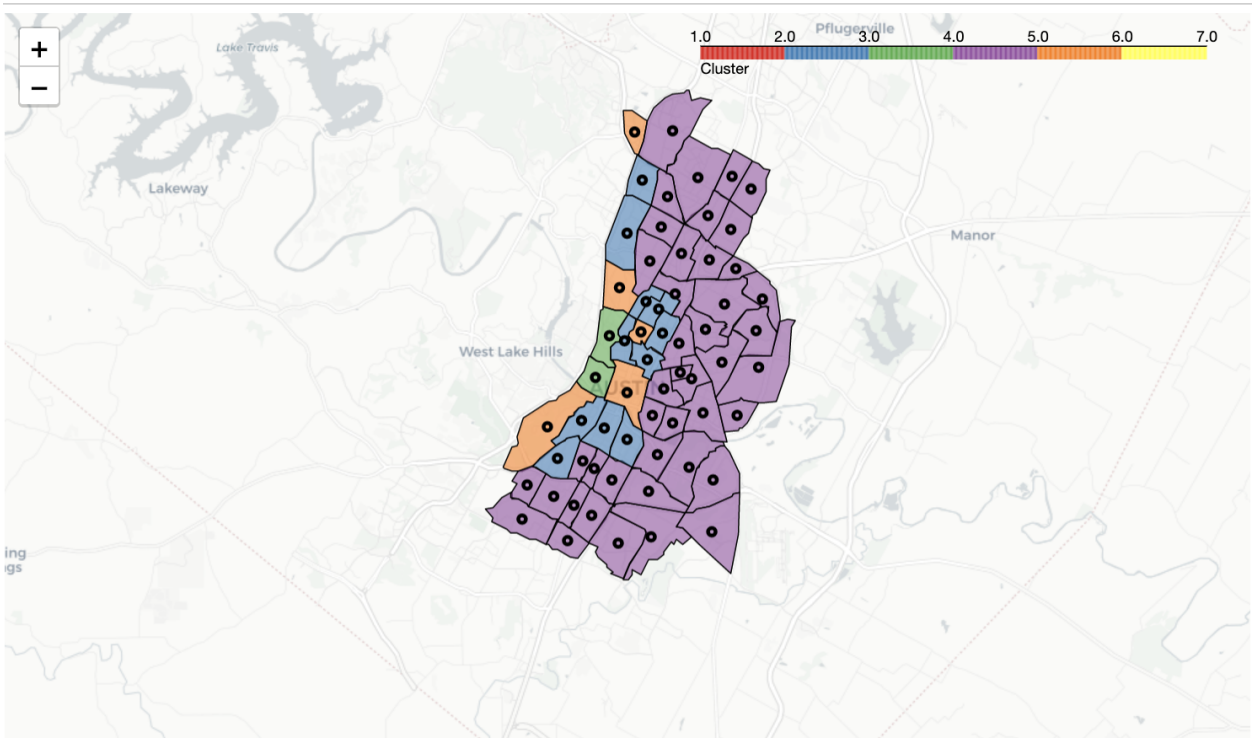

Figure 4. Austin, TX.

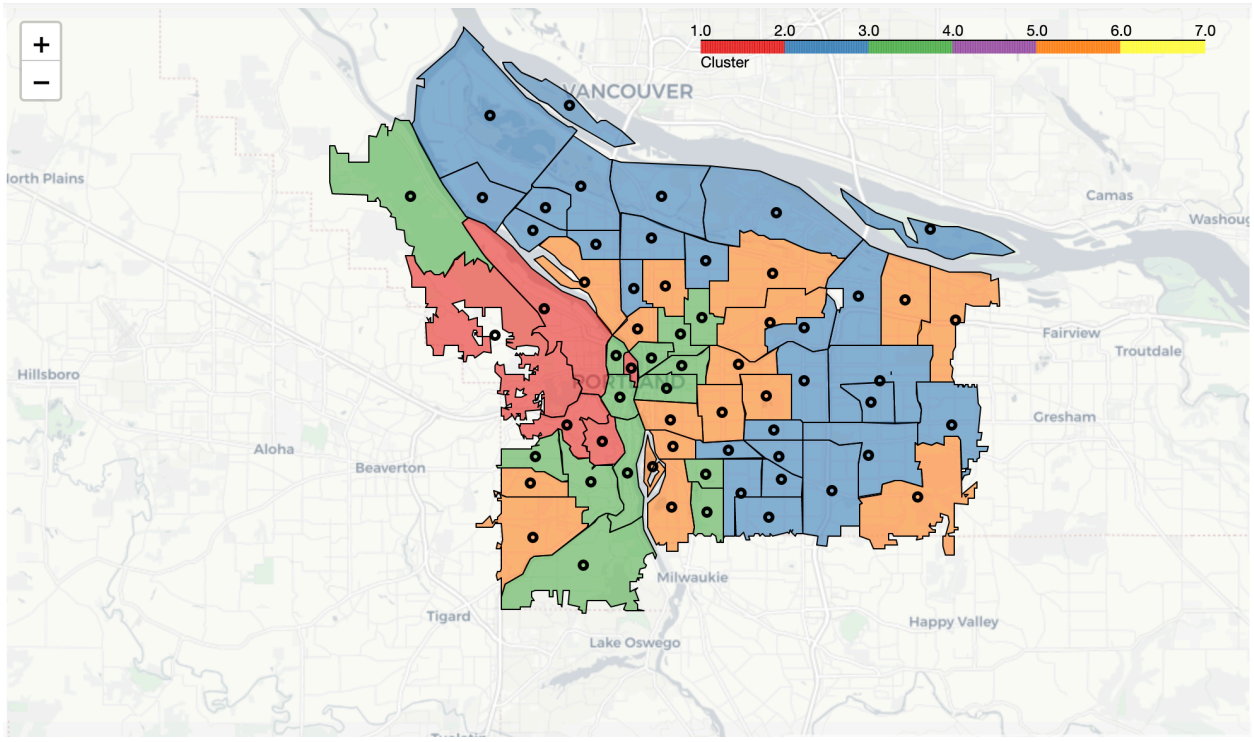

Figure 4. Portland, OR.

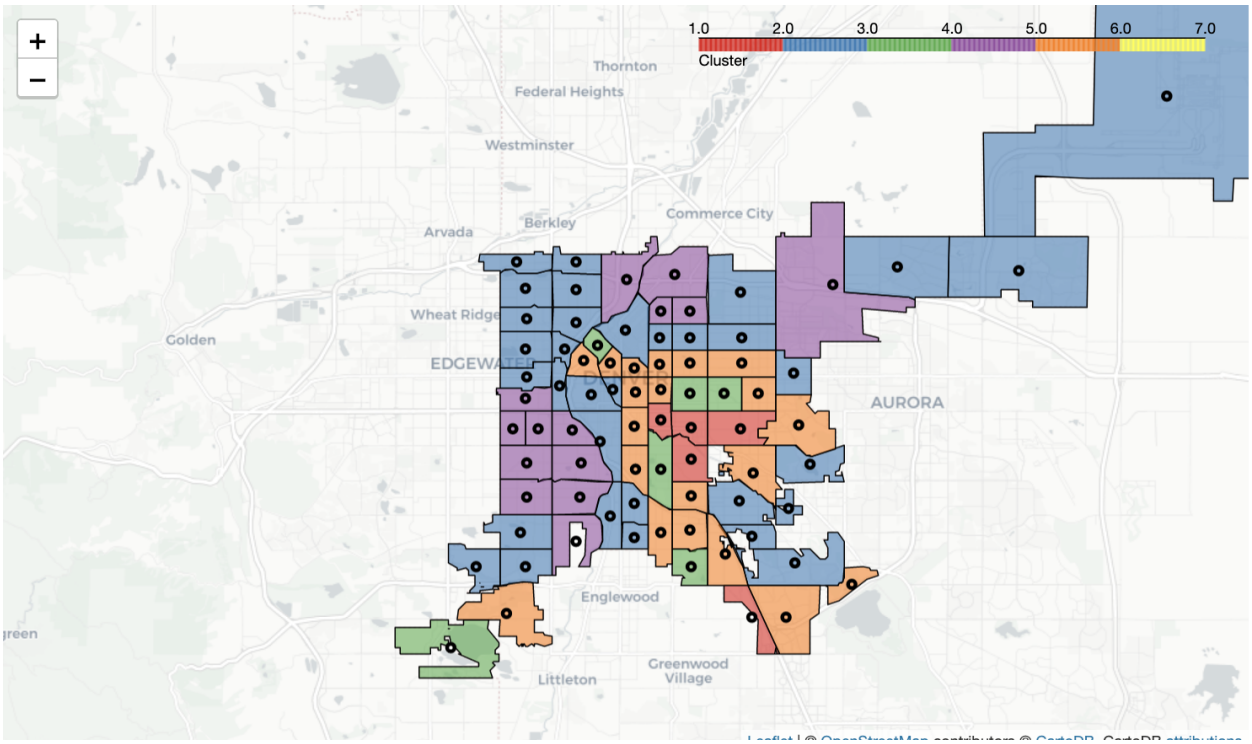

Figure 6. Denver, CO.

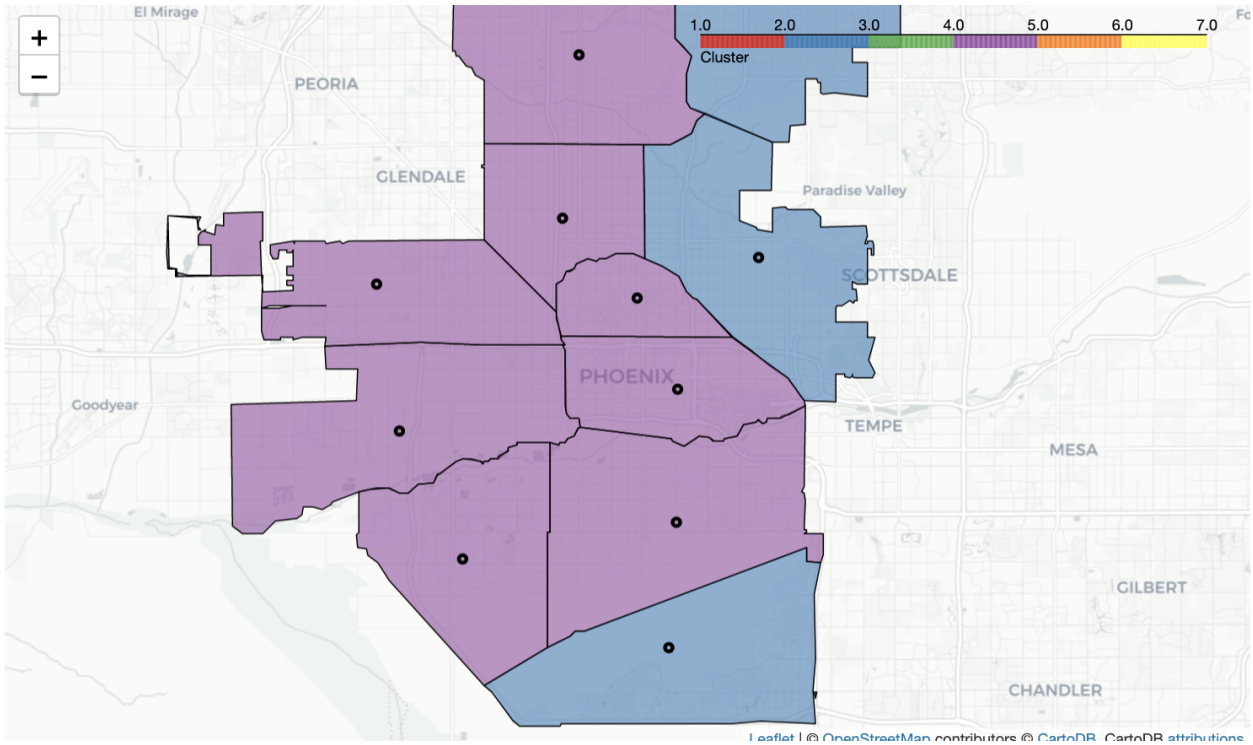

Figure 7. Phoenix, AZ.

4 Discussion

This was a fun exercise into data science. All 5 of these cities are vastly different and it is hard to capture the essence of a neighborhood. The census data used to describe the demographics of each neighborhood is from 2010 and housing prices have vastly changed in the last 10 years. There is better, more current and more detailed data available on housing prices through other sources, but they are not freely available like the NYU data used here. The Foursquare API is limited to the searching within the set radius from the centroid of each neighborhood. In some cases, like Phoenix and Denver there are large neighborhoods where 750 meters around the centroid includes only a small portion of the neighborhood and other neighborhoods in Seattle and Austin where the search radius includes portions of adjacent neighborhoods.

Looking into the top 10 venues of each cluster and the mean demographics of clusters we can start so see the breakdown of what defines the clusters. Cluster 6 looks like a nicer, residential neighborhood with the top venues including a park, a bus stop, a trail, soccer field and a playground with the highest housing prices and lowest rentership percentage and highest population age. Phoenix is dominated with clusters that don't match to Ballard with lower housing costs, lower population density, but similar venues.

	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	1	Park	Coffee Shop	Trail	Pizza Place	Gym	Café	Grocery Store	Playground	Italian Restaurant	Burger Joint
1	2	Park	Coffee Shop	Pizza Place	Convenience Store	Bar	Mexican Restaurant	Sandwich Place	Trail	Brewery	Grocery Store
2	3	Coffee Shop	Park	Trail	Pizza Place	Mexican Restaurant	Bar	Grocery Store	American Restaurant	Playground	Food Truck
3	4	Mexican Restaurant	Convenience Store	Park	Discount Store	Food Truck	Coffee Shop	Pizza Place	Gas Station	Liquor Store	Fast Food Restaurant
4	5	Coffee Shop	Park	Pizza Place	Bar	Mexican Restaurant	Sandwich Place	Grocery Store	Convenience Store	Food Truck	Café
5	6	Park	Bus Stop	Café	Coffee Shop	Trail	Gym	Pharmacy	Soccer Field	Playground	Pizza Place

Table 3. Most Common Venues in each cluster

	Cluster Labels	Home Ownership	Home Rentership	Home Value	Median Age	Population Denisty
0	1	0.541638	0.387761	562483.677419	41.761062	8452.580796
1	2	0.505959	0.411856	223708.462500	36.130668	6561.008287
2	3	0.487702	0.439953	431347.980000	39.892811	9082.096665
3	4	0.413438	0.516924	129842.121212	31.399679	6340.309719
4	5	0.478288	0.448951	311301.366667	38.268215	7800.960587
5	6	0.680126	0.259367	769085.571429	43.702007	5902.288526

Table 4. Mean Census Demographics from each cluster.

We can measure the Euclidean distance between the centroid of cluster 3 to the centroid of all other clusters. Clusters 1 and 5 are the closest to cluster 3 while 4 and 6 are the furthest away. There are some similarities in neighborhoods in Seattle that fall into 1 and 3.

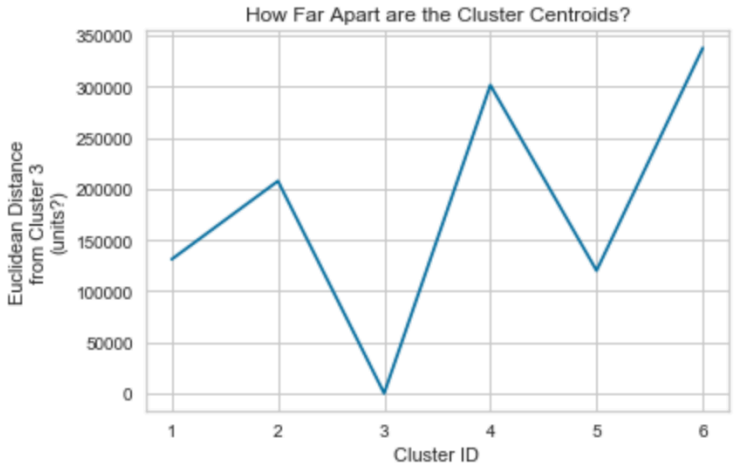

Figure 8. Euclidean distances of cluster centroids based on cluster 3.

5 Conclusion

To find a new neighborhood to move across the country, Portland or Austin are the better choices with the most options of neighborhoods similar to Ballard, Seattle. They cluster similar to my current home in housing costs, and similar types of venues.

6 References

- [1] NYU Spatial Data Repository
- [2] Foursquare API