Comparing Neighborhoods in Pittsburgh, PA to Austin, TX

Taylor Vactor

November 3, 2020

1. Introduction

My family and I have been considering moving back to Pittsburgh, PA from Austin, TX, but we really love our current neighborhood in Austin. We love that we are technically in the suburbs and have space in our house and neighborhood, parks and green space, good schools and crime is remarkably low. What we really enjoy most though is the fact that we are close to downtown (~8 miles) and south Austin (~8 miles) and have easy access to the multitude of amenities that these parts of town offer. These include great restaurants, music venues, eclectic shops, coffee houses, etc. My family and I would love to find a neighborhood in Pittsburgh that has similar proximal amenities to our current neighborhood, so that we could move closer to our extended families and really enjoy where we live.

2. Data Acquisition and Cleaning

Pittsburgh neighborhood data from the Western Pennsylvania Regional Data Center (https://data.wprdc.org/dataset/neighborhoods-with-snap-data/resource/cdea4e5c-646d-4924-84b1-afc3a7206eb9) and Foursquare's location data/API (https://foursquare.com/city-guide) were the primary sources of data used to solve this problem. The Pittsburgh neighborhood data provided the geographical coordinates of all the Pittsburgh neighborhoods and then the Geopy library was used to determine the latitude and longitude of our neighborhood in Austin. Foursquare's API will be used to see what type of venues are nearby. It will consider such venues as parks, restaurants, coffee shops, stores, music venues, etc.

The data was preliminarily reviewed, and it was determined that a significant number of items could be removed from the Pittsburgh Neighborhood dataset. Being that all that was required for this exercise was the neighborhood name and latitude and longitude, features such as Neighborhood ID, Perimeter, Acres, Population Change, etc were removed from the dataset.

3. Methodology

The Pittsburgh dataset JSON file was initially brought into the notebook using wget and then was converted into a data frame for easier manipulation. Using the Geopy library, the coordinates for the neighborhood in Austin were retrieved and appended to the existing neighborhoods data frame. Those

coordinates were plotting using the Folium library and the data was reviewed for geographical accuracy (Fig 1a & 1b).

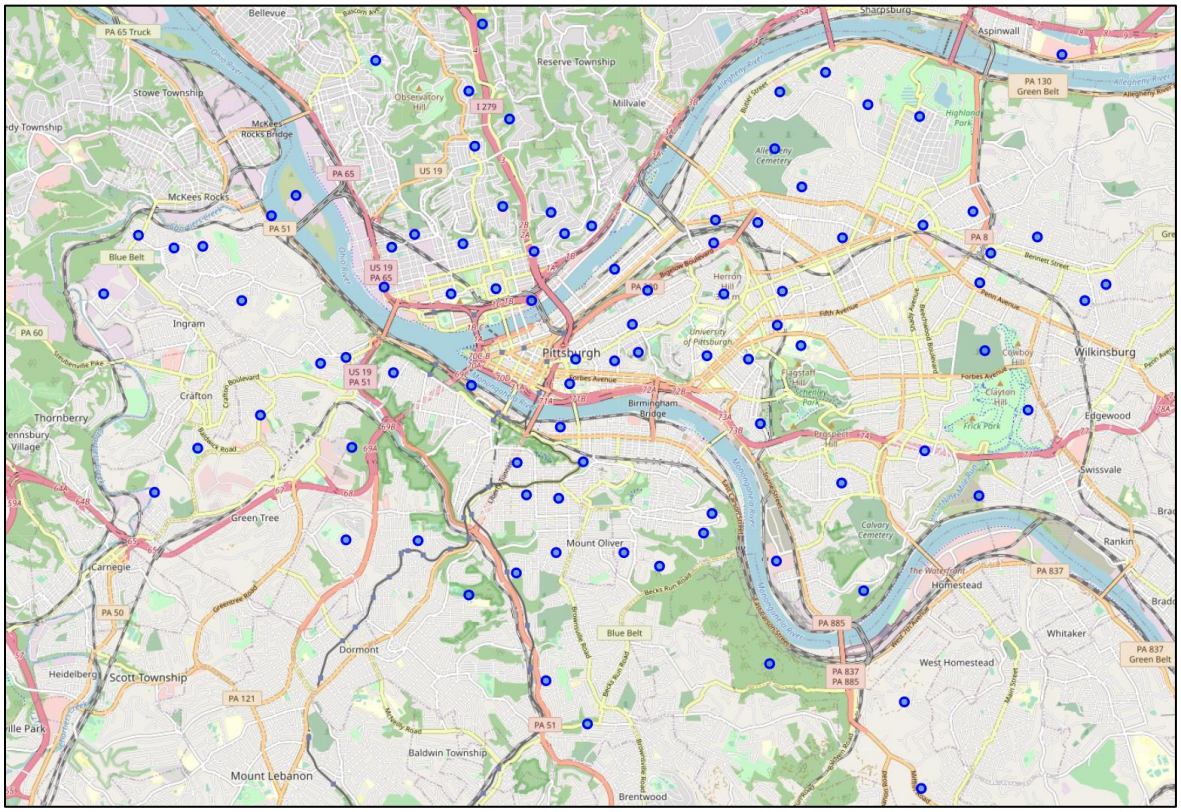

Figure 1a. Pittsburgh Neighborhoods

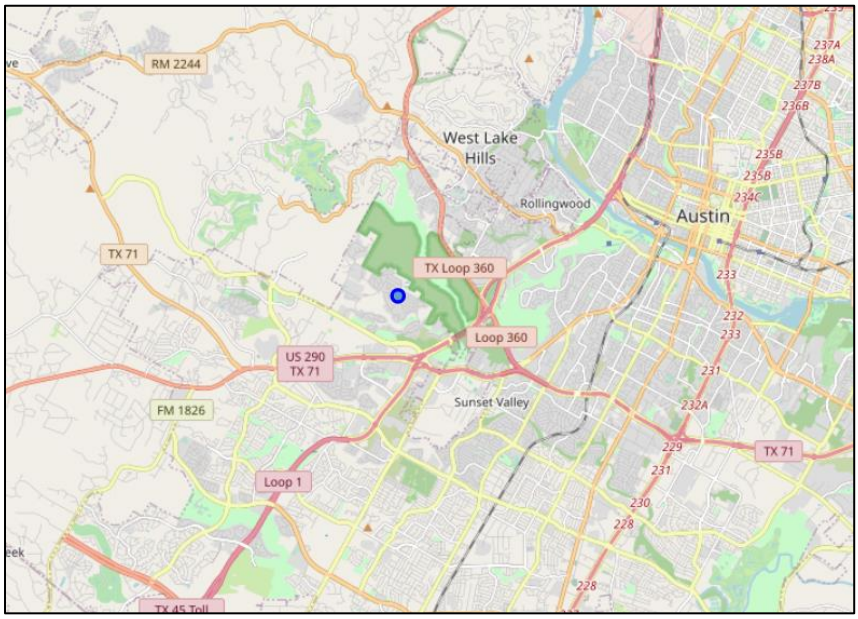

Figure 1b. Austin Neighborhood

The Foursquare API was then called to retrieve all venues within 6 Km of each neighborhood's lat/longs. These venues were then grouped by neighborhood and category. The frequency of each category per neighborhood was then calculated to normalize the data. The 10 most common venues of each neighborhood were populated into a chart for visual comparison. An example of this table is given in Figure 2 where the top 5 venues are shown for the first 5 neighborhoods by alphabetical order.

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Allegheny Center	Hotel	Sandwich Place	Bar	American Restaurant	Scenic Lookout
Allegheny West	American Restaurant	Hotel	Bar	Scenic Lookout	Coffee Shop
Allentown	Hotel	American Restaurant	Bar	Ice Cream Shop	Scenic Lookout
Arlington	Hotel	American Restaurant	Ice Cream Shop	Bar	Pizza Place
Arlington Heights	American Restaurant	Hotel	Ice Cream Shop	Bakery	Coffee Shop

Figure 2. Most Common Venues by Neighborhood

It was determined that KMeans Constrained Clustering would be the most effective machine learning algorithm to solve this problem. Again, the goal is to find neighborhoods in Pittsburgh that are most similar to the one in Austin. This can be achieved through a clustering algorithm. KMeans was initially chosen but was not working as the Austin neighborhood was being put into a cluster all by itself. Kmeans Constrained was then chosen because it allows the use of a minimum number of neighborhoods required in each cluster.

By using the "elbow method" it was determined that the appropriate number of clusters for the KMeans Constrained algorithm was 4. This was selected by referring to the graph in Figure 3 and choosing the inflection point in the data.

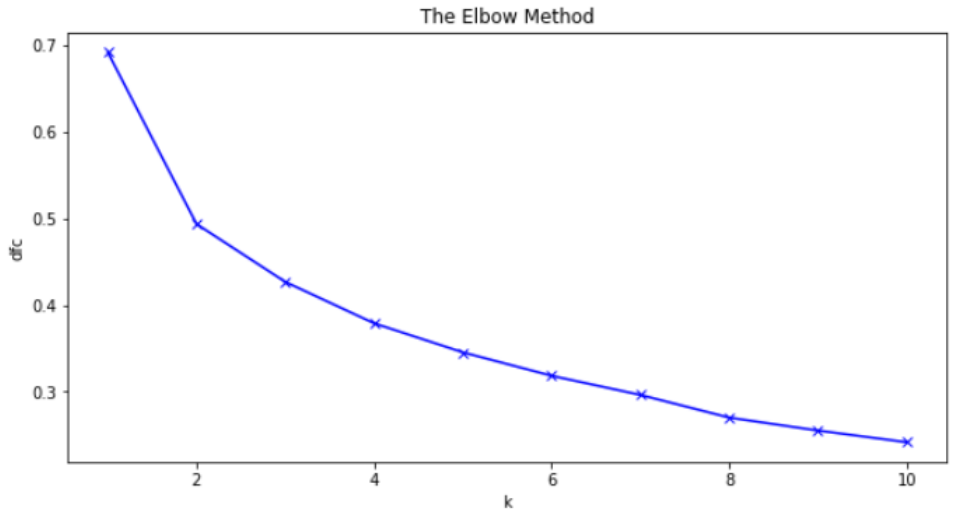

Figure 3. "Elbow Method" Graph

The KMeans Constrained algorithm was then run and cluster numbers were assigned to individual neighborhoods. Finally, this information was mapped to spatially visualize which neighborhoods were most similar.

4. Results & Discussion

From the mapping of the resulting clustering you can see that all the Pittsburgh neighborhoods that were grouped with the Austin Neighborhood (red) are on the eastern side of Pittsburgh (Figure 4a & b).

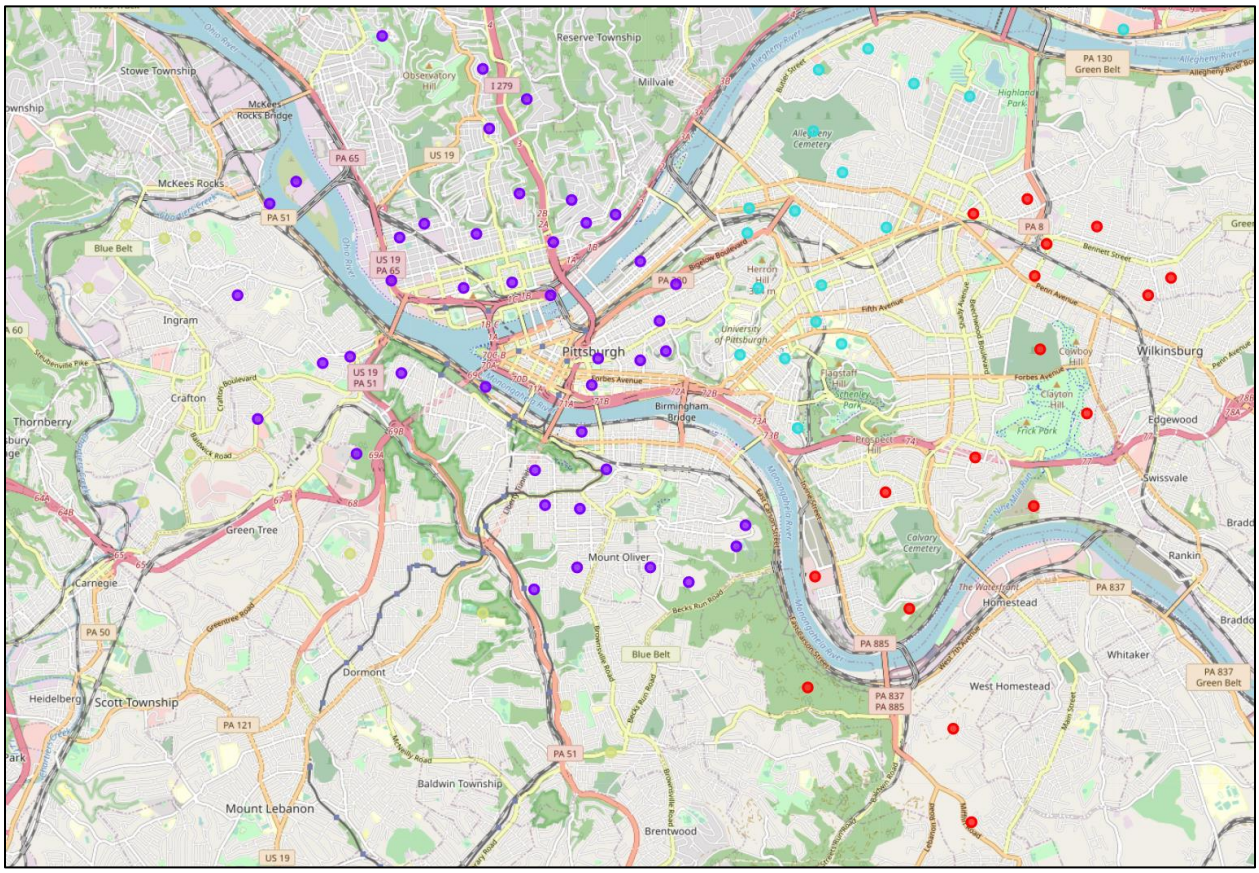

Figure 4a. Pittsburgh Clustering

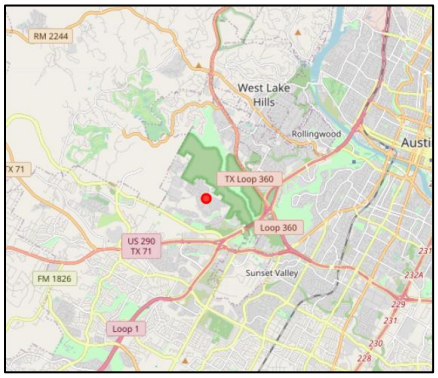

Figure 4b. Austin Clustering

These neighborhoods are New Homestead, East Liberty, Homewood West, Larimer, Homewood South, Lincoln Place, Point Breeze North, Hays, Regent Square, Swisshelm Park, Point Breeze, Squirrel Hill South, Greenfield, East Hills, Hazelwood, Homewood North and Glen Hazel.

In being a former Pittsburgher and current Austinite, the red grouping makes sense to me. The eastern side of Pittsburgh is home to some of the city's biggest parks and greenspace. It also has an eclectic mix of ethnic restaurants, coffee shops and bars. This is very similar to where we currently live in Austin. Furthermore, during the last 5 years we lived in Pittsburgh, we lived in two different neighborhoods in the red cluster, Regent Square and Squirrel Hill. That is certainly not a coincidence and was driven by the fact that we were drawn to the amenities that that area had to offer. Therefore, the red cluster appears like they are suitable neighborhoods for my family and I to consider moving to.

5. Conclusions

In this study, I analyzed similarities in neighborhoods in Pittsburgh versus my family's current neighborhood in Austin to try and determine which Pittsburgh neighborhoods we might like to move to. This was done by comparing what venues were in each neighborhood's general vicinity such as restaurants, stores, coffee shops, music venues and more. I used a clustering algorithm to solve this problem and came up with a list of potential neighborhoods to move to based on this analysis.

In the future, I believe that this work could further be refined by incorporating crime statistics and housing prices. I noticed that a few of the neighborhoods that were in the red cluster are in high crime areas even though they border locations with great amenities. I would like to trim these neighborhoods out based perhaps on a crime rate threshold or just by incorporating it into the clustering. Housing prices should also be included because we can obviously only purchase a home within our budget. Once those things are incorporated, a wholistic approach to determining the best neighborhoods for us should be achievable.