Portland Neighborhood Search

Coursera IBM Data Science Capstone Project

Maribeth Todd

May 2020

Introduction

Choosing the right neighborhood to buy a house in can be difficult, with an overwhelming array of features and locations to choose from within an urban area. To narrow it down, we can use Foursquare and other data sources to locate the amenities we care about and find neighborhoods with the right mix of services for our personal preferences. This analysis would be of interest to anyone looking to choose a neighborhood in which to buy a house, as well as real estate professionals trying to help their clients choose the right neighborhood.

When I was looking to buy a house in Portland, Oregon, I was new to the city and didn't know much about the neighborhoods. I looked at data from the Census Bureau, local schools, and other sources to choose a few neighborhoods, and then went to Google maps to figure out if those neighborhoods had the amenities that were important to me (e.g. grocery stores, etc.) This research process was time consuming and somewhat arbitrary, so this project will create a more systematic process for narrowing the search for the right house in the right neighborhood, using Portland as an example.

Data

The data sources for this project include business location data from Foursquare, Census data, and the Mapquest geocoding API.

I use Census data to define the neighborhoods in the Portland region and obtain some basic information like population density and average housing values. The neighborhoods are based on census tracts in order to take advantage of the wide variety of data available in the Census American Community Survey. Since the Foursqaure API takes lat/long points as input, neighborhoods are defined using the mean center of population for every census tract in the Portland region. This concept is more representative of the location of population within each tract than a geographic center would be. The Census Bureau provides latitude and longitude data for the mean center of population for various geographic levels here.

Other attributes are obtained from the Census American Community Survey API and TIGER/Line files. The Census ACS includes data on a wide variety of aspects that may be of interest to a potential homebuyer, including average house values, and demographic information like age and the number of

households with kids. The TIGER/line shapefile provides the area of the tracts for calculating population densities.

Using the Foursqaure API, I find the locations of the types of businesses that I would like to have within walking distance in my neighborhood. My preferences for a neighborhood include some essential services like grocery stores, a library, a hardware store, and some restaurants. Someone else might prefer different amenities, like bars and bike shops, in their ideal neighborhood so this approach could be customized for different preferences.

The goal is to use the Census and Foursquare data to cluster neighborhoods into similar groups. This will help narrow the search for the right location and provide a manageable list of neighborhoods for further exploration. The Mapquest API is used to assign each census tract center to a zip code, to facilitate linking the results of the cluster analysis out to real estate listings.

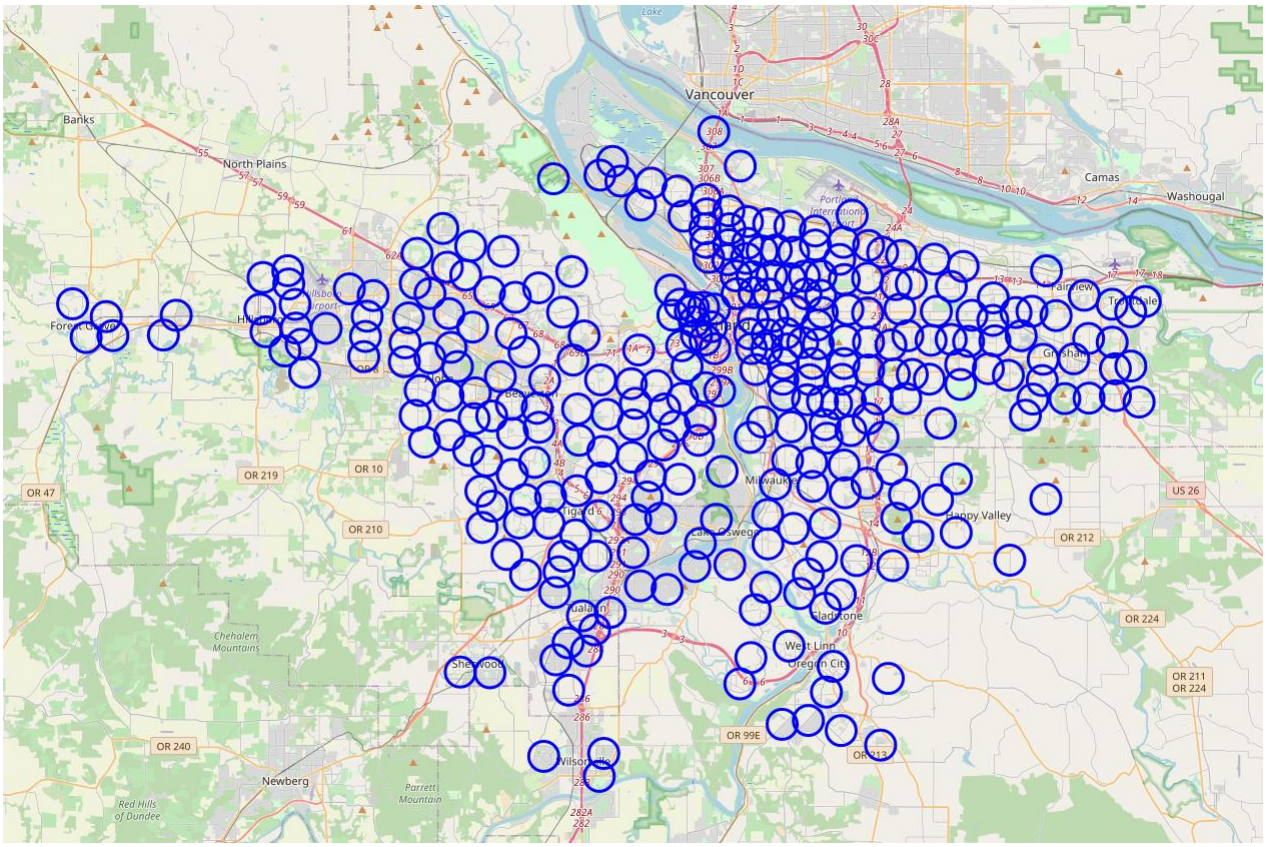
Methodology

The first step is importing the necessary Census data, including the neighborhood points and ACS data to link to those points. The Portland region includes three counties, and only tracts within the urban portions of those counties are included in the analysis.

The population center points are imported as a csv file directly from the Census website. I use a tract shapefile from the Census TIGER/line program to obtain the land area of each tract and add a flag identifying which tracts are in the urban area.

Next, download some ACS data from the Census API. The total population comes from ACS table B01003 and can be used along with the land area to calculate population density. Median house value in 2018 inflation-adjusted dollars comes from table B25107.

Now the lat/long data from the Census tract points can be used to collect information from Foursquare on the businesses located in each neighborhood. The map below shows the radius of each neighborhood, set to 800 meters or about a half mile, that will be sent to the Foursquare API. Many of the neighborhoods overlap to some extent, which means many businesses could be counted multiple times in the analysis, but that's ok. The point is to find the businesses that are within a reasonable walking distance of the neighborhood center.



I define a function that searches the Foursquare API by category, and input the appropriate venue category ID codes. I am using venues of interest to me in this analysis, but these could be changed if, for example, a real estate agent was assisting a client with a neighborhood search. The venue types and Foursquare category IDs included in this search are:

- **Libraries**: '4bf58dd8d48988d12f941735'
- **Grocery stores**: '4bf58dd8d48988d118951735'
- **Food** (i.e. restaurants and prepared food): '4d4b7105d754a06374d81259'
- **Hardware stores**: '4bf58dd8d48988d112951735'

The Foursquare search returns some venues that don't really fit the categories of interest, so an additional filtering step on the category descriptions is applied to ensure the venues are relevant. Each venue type is summarized at the tract level, so the result is the count of each venue type in each neighborhood.

Now the Census and Foursquare data can be collated and formatted for clustering analysis. One additional cleaning step is required to drop records with no house value data (where value < 0) because these are nonresidential areas.

Some summary statistics on the dataset (below) indicate that most neighborhoods do not have a library or hardware store. A little more than a third of neighborhoods have no grocery store, and all

neighborhoods have at least one restaurant, coffee shop, or other prepared food venue. This suggests that libraries and hardware stores may be limiting factors more than grocery stores and restaurants in which neighborhoods are most suitable.

	Number of neighborhoods with venues			
Number of venues	**library**	**hardware**	**grocery**	**restaurant**
0	206	153	115	0
1	78	66	71	32
2	12	45	58	13
3	6	14	32	8
4	6	14	12	10
5	1	5	9	7
6	1	4	3	7
7	1	3	5	9
8	0	1	2	7
9	0	1	3	6
10 or more	0	5	1	212

The clustering analysis includes four venue types and population density as prices will be used for filtering the neighborhoods after the clustering step. The data are scaled and then the neighborhoods are clustered using the k-means algorithm.

To check for the best number of clusters to use, we need to evaluate the output of the clustering algorithm using different values for k. There are several metrics for k-means including the silhouette coefficient, Davies-Bouldin index and Calinski-Harabsz index. Higher values of the silhouette and Calinski-Harabasz scores indicate better defined clusters, while a lower value of the Davies-Boulding score indicates a model with better separation between the clusters.

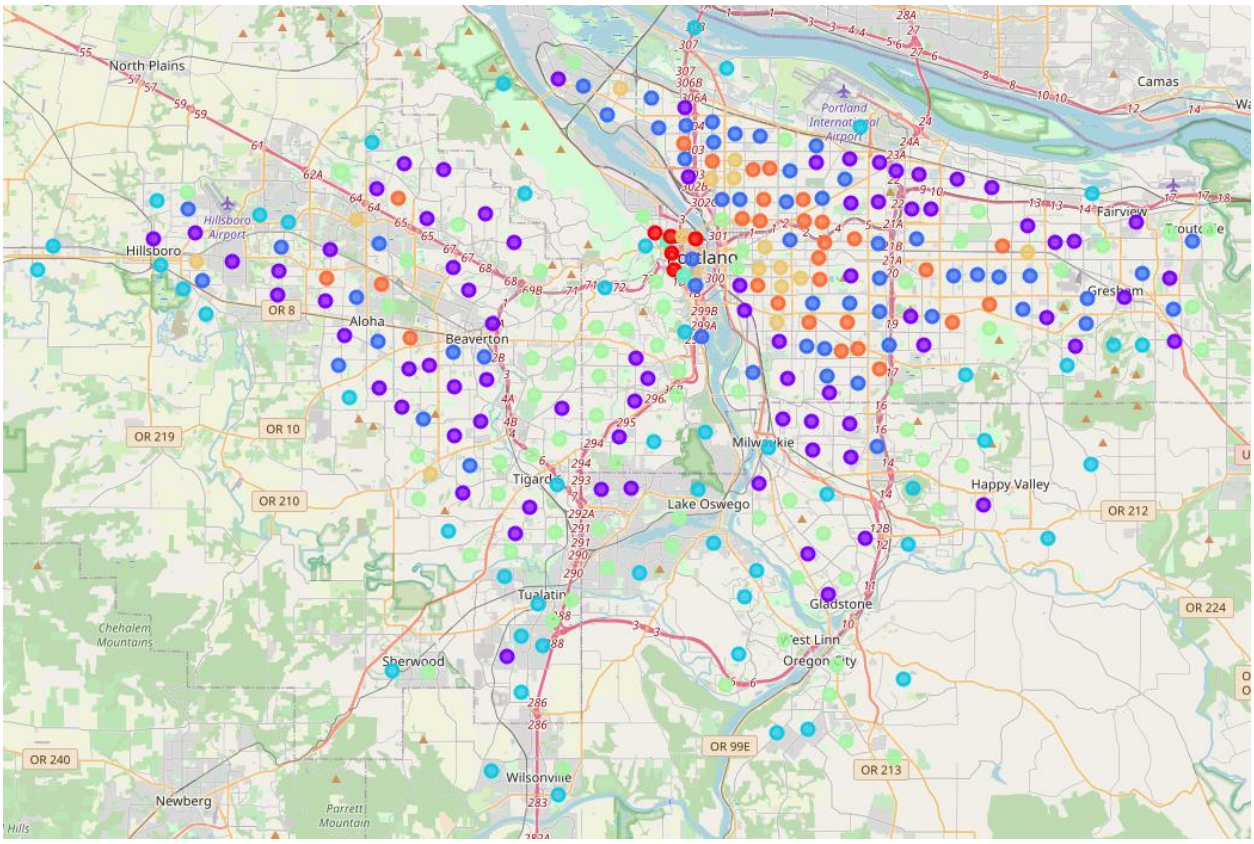
k-means cluster metrics	**6 clusters**	**7 clusters**	**8 clusters**	**9 clusters**	**10 clusters**
Average silhouette score	0.534	0.536	0.561	0.557	0.548
Davies-Bouldin score	0.522	0.511	0.490	0.398	0.412
Calinski-Harabasz score	868	1054	1213	1363	1622

These scores suggest that a model with 8 or 9 clusters presents the best fit, as 8 clusters produces the best silhouette score and 9 clusters produces the best Davies-Bouldin score. The Calinski-Harabasz score continues to improve as the number of clusters increases but at the expense of the other two measures. The resulting clusters from using k=8 are shown in the results section below.

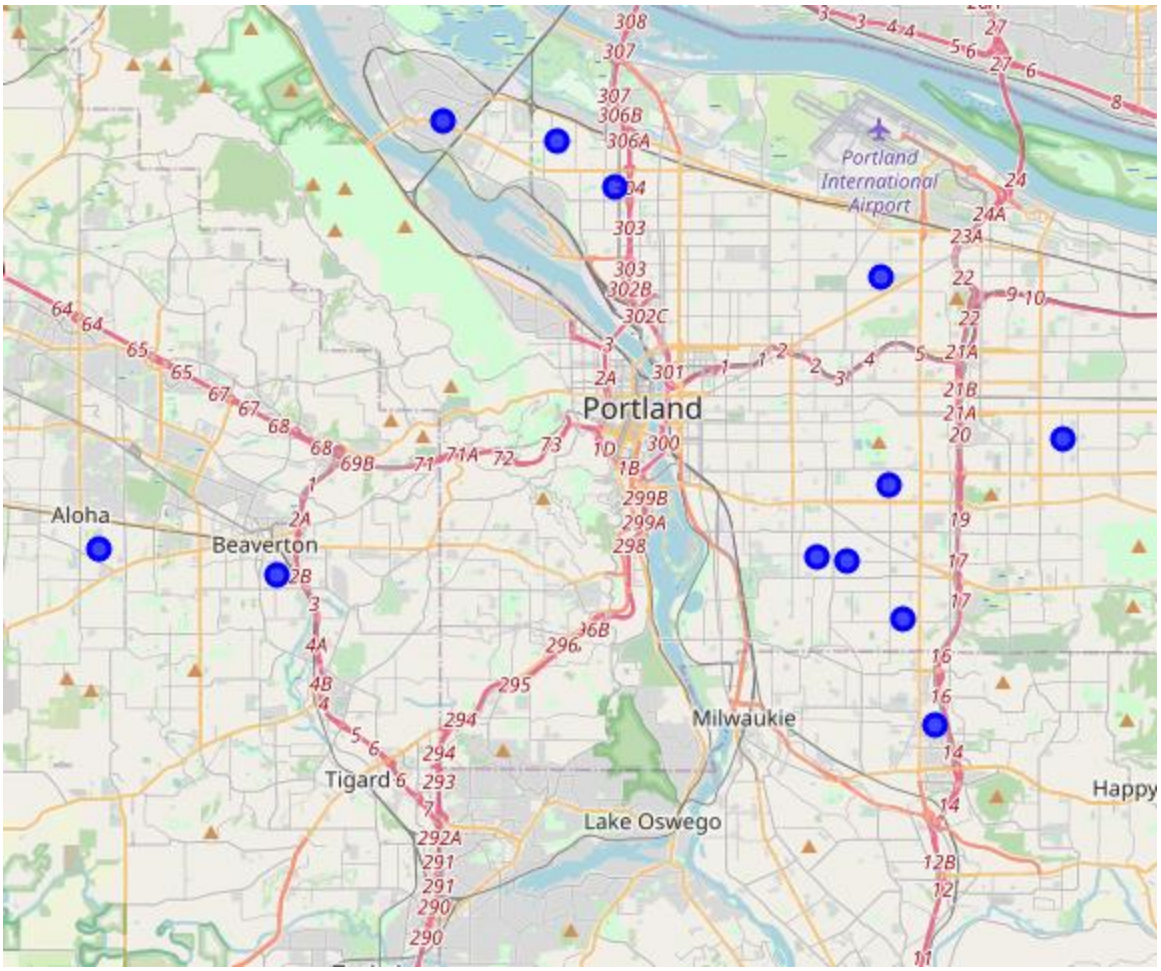
Results

Some summary statistics on the resulting clusters are shown below. Most neighborhoods fall within clusters 1, 2, 3, 5, and 7. The smaller clusters represent neighborhoods with more rare features such as the very high densities of population and services in the central city in clusters 0 and 4.From these summary statistics cluster 2 looks promising. It has a moderate population density, low to moderate price level, and most neighborhoods in this cluster have access to at least one of each of the venue types. Cluster 2 is shown in dark blue in the map below.

		Average cluster values					
cluster	**neighborhood count**	**food**	**grocery**	**library**	**hardware**	**population density**	**house value**
0	5	30	6.4	3.4	5.4	7,435	562,500
1	78	15.9	1.1	0.4	1	2,238	344,868
2	56	21.8	2.2	0.6	1.4	2,996	344,246
3	54	10.4	0.6	0.3	0.7	628	397,498
4	2	30	6	2.5	4	11,852	353,250
5	72	13.9	1	0.2	1.4	1,508	414,321
6	15	29.1	4.5	1.2	2.2	4,763	397,587
7	29	23.6	2.2	0.9	1.7	3,755	412,283



To better visualize the results, the next map shows just the neighborhoods in cluster 2. To narrow the list down further, we'll look at neighborhoods with a median house value of less than $400,000 and where there is access within a half mile of at least one library. This results in 12 neighborhoods for further exploration. Now that we have a reasonable number of neighborhoods to explore, we can look at some real estate listings. In the full analysis notebook, the points in this filtered map also include a link to homes that are currently for sale in that neighborhood's zip code on Zillow to facilitate further research on these neighborhoods. From here, choosing the best neighborhood depends on the type of houses available, the commute to work, and other variables that haven't been included so far.



Discussion

These cluster analysis results are a jumping off point for further exploration of the neighborhoods that meet the criteria laid out above. The direct links to real estate listings in the notebook make it easy to see the types of houses that are on the market in these neighborhoods and determine if any of them might be a good fit. If none of these neighborhoods seem quite right, we may need to explore a different cluster (cluster 7 seems to be most similar to cluster 2) or we may need to adjust our filtering criteria (price and library access) or redo the cluster analysis using different variables or parameters. Requiring easy library access eliminates about two thirds of the neighborhoods in the Portland region from consideration, so we may decide that library access can be omitted from the analysis or replaced with another venue type.

Conclusion

Choosing the right neighborhood to live in is a very subjective process and everyone will have their own set of preferences and priorities. In this analysis I have focused on easy walking access to grocery stores, hardware stores, libraries, and restaurants within a half mile. The inclusion of population density in the cluster analysis is a good proxy for the type of built environment and walkability in a neighborhood. The Foursquare data indicate that restaurants and other food venues like coffee shops are readily available in most neighborhoods in the Portland region, and grocery stores are fairly well distributed as well. Hardware stores are less common, and easy access to a library is a much more limiting factor in choosing a neighborhood.

The ultimate decision of which neighborhood to purchase a house in must include additional factors such as budget, the size and type of houses that are on the market in various neighborhoods, and commutes to work or school, but this is a good way to start narrowing the search. In addition to confirming that the neighborhood I chose to live in still meets all of my needs, this analysis has highlighted some other neighborhoods that I should consider if we decide to move in the future. This type of systematic neighborhood evaluation could be useful for other individuals looking for a new house or real estate professionals assisting clients with their search for a place to live, and can easily be customized to different priorities.

Reference and data sources

Census Centers of Population

- https://www.census.gov/geographies/reference-files/2010/geo/2010-centers-population.html

Census ACS data

- https://www.census.gov/data/developers/data-sets/acs-5year.html

Foursquare venue categories

- https://developer.foursquare.com/docs/build-with-foursquare/categories/

Scikit-learn k-means clustering metrics

- https://scikit-learn.org/stable/modules/clustering.html

Mapquest geocoding API (for zip codes)

- https://developer.mapquest.com/documentation/geocoding-api/