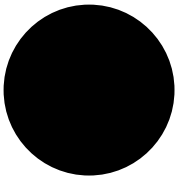
Finding alternative cities to live in the Northwest US

IBM Applied Data Science Capstone Project



Introduction

Background

Over the past decade, Seattle has led other metropolitan cities as the fastest growing city in the United States[1]. Such rapid growth, however, has caused a general sense of crowding and displacement for existing residents[2], particularly those who were first drawn to the area for its "small city" character and affordable housing[3]. This real estate boom has not been limited to the greater Seattle area. Housing prices across the entire state of Washington have undergone unprecedented growth[4].

Problem

The steep rise in the cost of living is starting to compel a number of Washington residents to seek alternative places to live, particularly those who are able to telecommute and/or whose job prospects aren't tied to a particular location and retirees. The question for this subset of people is *how to even get started browsing prospective places to move*, as Washington State alone has 211 cities[5].

To answer this question, we'll start with the assumption that greater Seattle residents looking to move are still interested in in living in the Northwest US and seek to find alternative cities with similar amenities as their current one. Given this scope, its possible to sample the superset of cities in Washington and adjacent states (Oregon and Idaho) to create a kind of "fingerprint" of popular venues (such as certain types of restaurants, stores and natural areas) for each city, and then use this to identify potential similarities with other cities. The findings of this exercise could then be used as a recommendation guide for further, in-person real estate research.

Audience

The primary audience of this study might include realtors and potential home buyers/renters in the Northwest (Washington/Idaho/Oregon) region. The findings could also be used by Northwest entrepreneurs looking to open new businesses or even a way of fostering outreach and partnerships among Northwest municipal chambers of commerce.

Data

To obtain a list of Northwest cities, I scraped Wikipedia for a list of all the cities in Washington[5], Oregon[6] and Idaho[7]. Next I used the Foursquare venue recommendation API[8] to obtain a list of the most popular venues for each city and queried location data (latitude/longitude) using the Mapquest

Geocoding API[9] and the Folium[10] mapping library in order to visualize the cities and how they cluster.

Preparation

With the lists of cities from the source Wikipedia pages structured as tables, it was easy to use the Pandas library to read in the HTML tables and convert them to dataframes. Using the columns of city names from those dataframes, I then created new dataframes to store city, state, longitude and latitude values and combine the Washington, Oregon, and Idaho dataframes into a single dataframe representing all the Northwest US cities.

I then ran the master city dataframe through the Mapquest API to look up the location (in terms of longitude and latitude) of each city. This is called *geocoding*. Using the Folium mapping library, I was then able to render the full set of cities in the Northwest US.

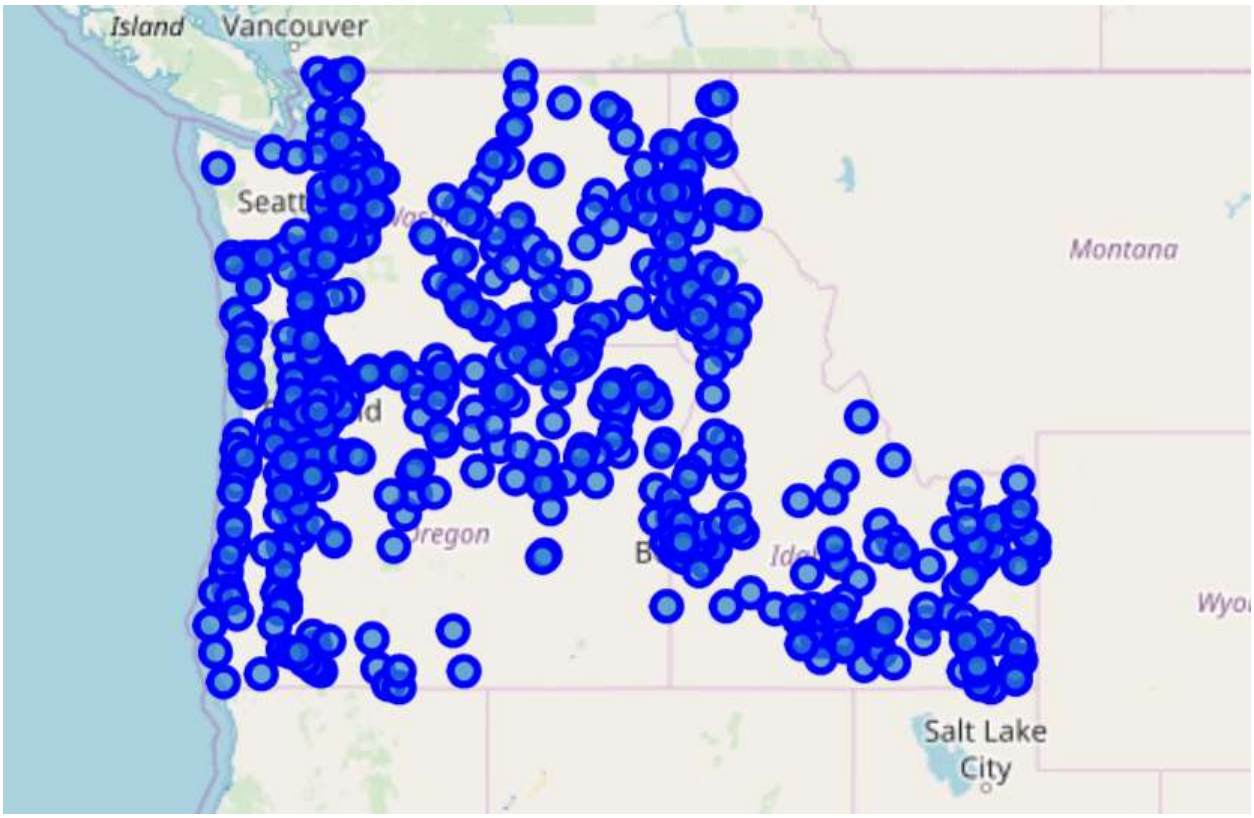


Figure 1: Map of all cities in Washington, Oregon, and Idaho state

The next step was to run the list of Northwest cities through the Foursquare API to query the top venues in each city (according to the ratings of Foursquare users). I abstracted the individual venues by filtering the returned data based on the general category of each venue.

Next I counted the number of venues for each city. Some cities (for example, bedroom communities) have very few venue entries on Foursquare. After testing different limits, I found that a city requires at least about 10 venue entries in order to have an adequate venue "profile" for meaningful clustering results with other cities. Given that, I dropped cities with fewer than 10 venues for the remainder of the study. This was a necessary step for further analysis, however it drastically truncated the list of 593 Northwest cities down to only 180 cities. Even so, the 180 remaining cities represented 108 unique venue categories, which seemed suitable for expressing interesting clustering patterns, as I later discovered.

Methodology

To analyze the data, I used a popular unsupervised machine learning algorithm called k-means clustering to partition observations into a specified number of clusters in order to discover underlying patterns. Specifically, I used the top 5 venue categories for each city (based on occurrences in the dataset) as each city's vector profile for finding similarities with other cities.

The first step was to calculate the average frequency for each venue category across each city. Using a Pandas dataframe I converted each venue category into a boolean (yes/no) column using the One-hot encoding method, verifying that new dataframe's column count equaled the number of unique venue categories (108) I identified during data preparation. Next I grouped rows by city mean of frequency for each category, and used that to find the five most common venues for each city.

With that I were ready to apply the K-means clustering algorithm. After trying out different k values (where k= *number of clusters*), I found the clusters to be most meaningful and interesting with around k=6. The output of the K-means algorithm is an array of cluster assignments for each row in the dataframe. With that I then stitched the cluster labels back into the dataframe and combined city location data in order to print out and visualize the results.

Results and discussions

I used the Python Folium library to render the clusters, using a distinct color for each. At first glance, the results look promising in terms of holding some patterns about the dataset. The clusters seem generally dispersed geographically and balanced in terms of member count.

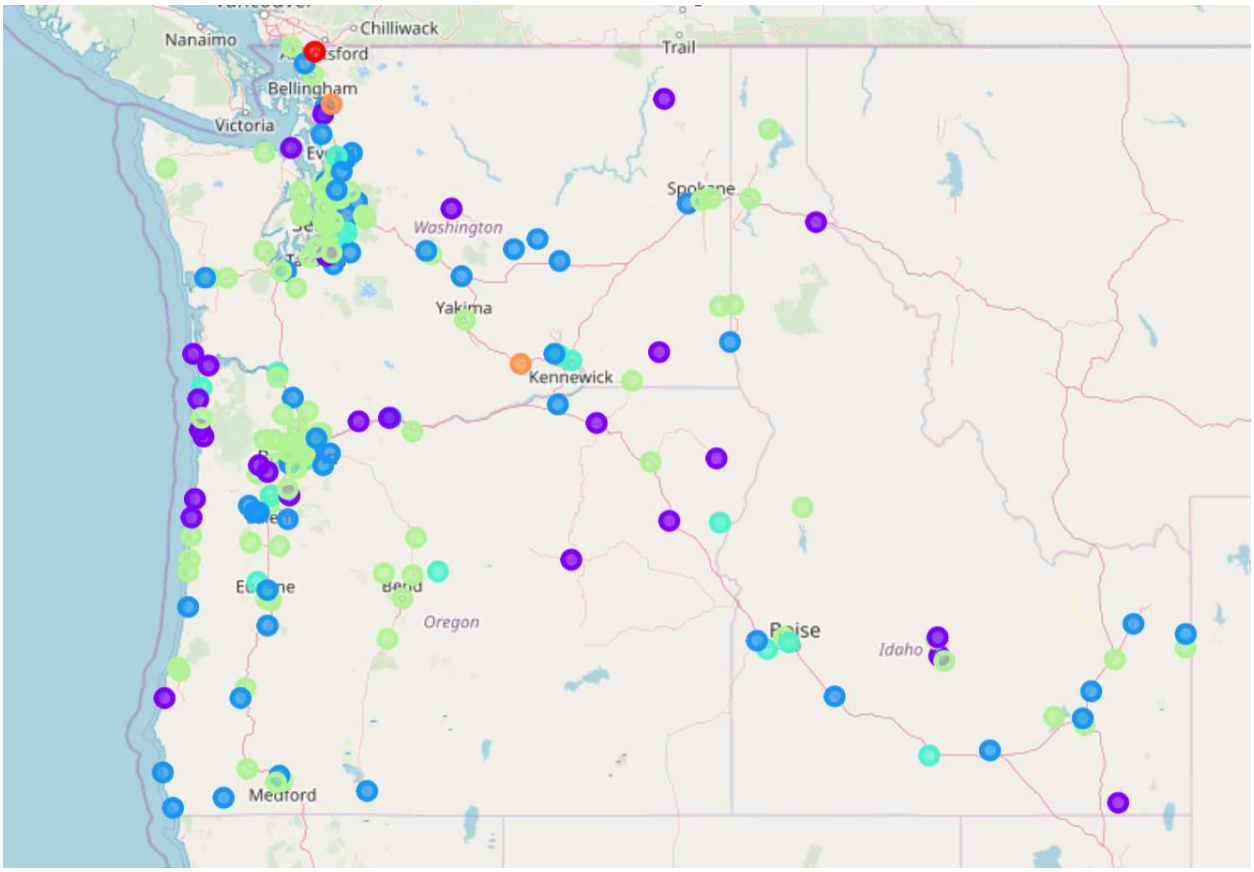
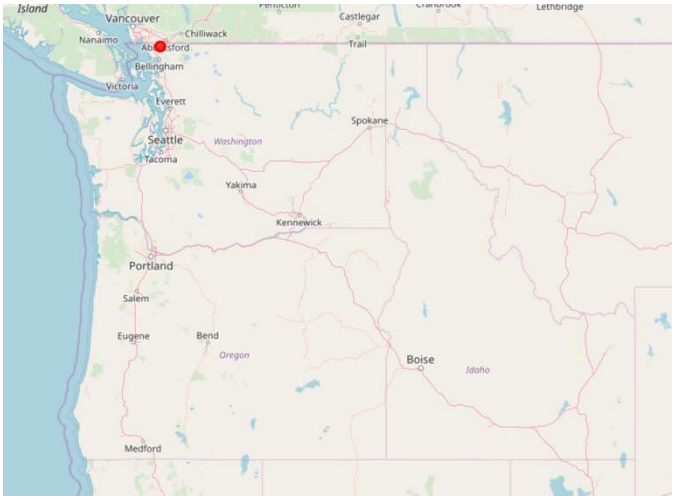


Figure 2: Clusters of Northwest cities

Here's how the clusters broke out and some further observations.

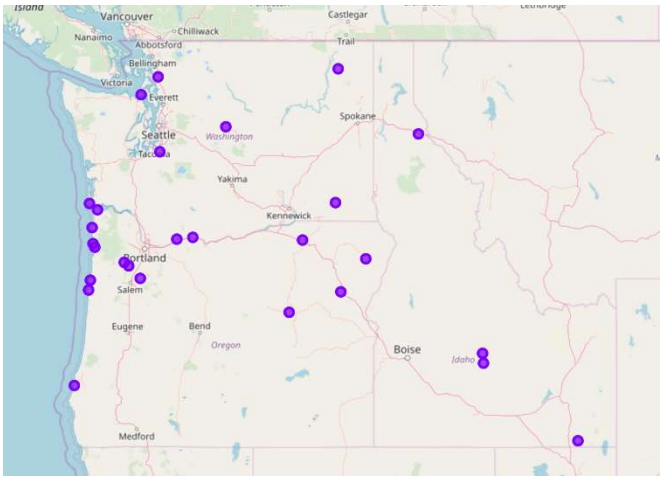
Cluster 0: Outlier city

This cluster consists of a single outlier city, Lynden WA. Although none of its top common venues seems uncommon in itself, perhaps it was the combination of all 5 that proved particularly unique among the dataset.



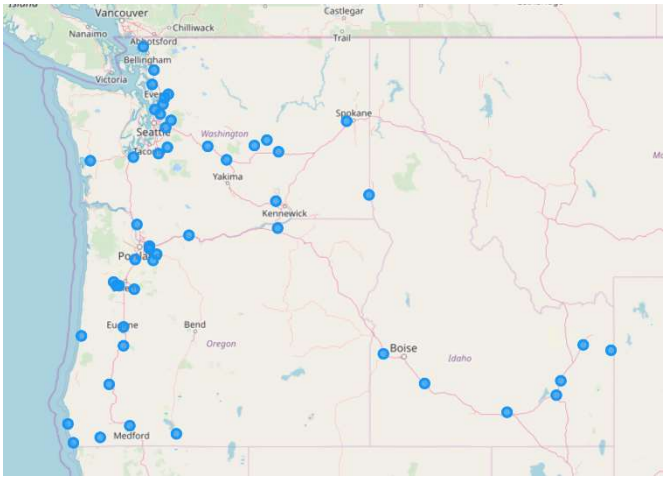
Cluster 1: Vacation destinations

Cluster 1 is characterized by hotels/resorts, restaurants and nightlife (bars, breweries, etc). Many of the cities on this list are vacation destinations and/or popular weekend getaway spots.



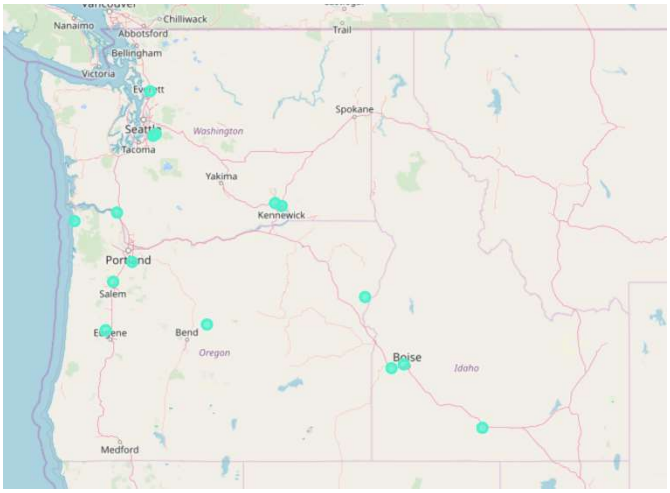
Cluster 2: Restaurant cities

Pizza place is common among nearly all of the cities in this cluster. It looks like *Pizza Place*, *Mexican Restaurant*, *Chinese Restaurant*, *American Restaurant*, and *Bar* are all grouping together here. The cities on this list tend to be larger than bedroom communities, but somewhat smaller than major uban centers.



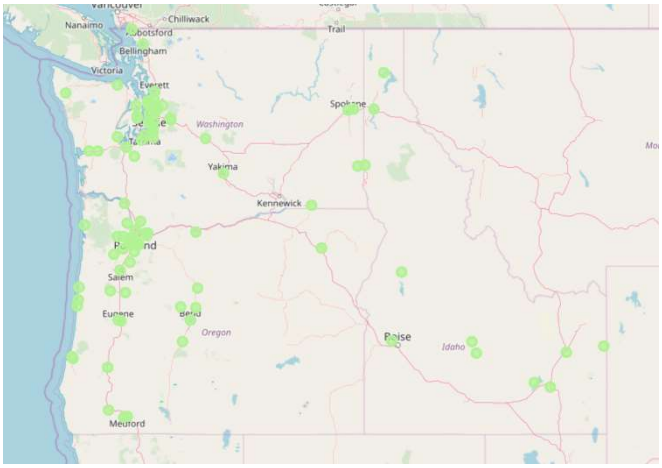
Cluster 3: Fast food cities

The unifying characteristic among these cities is that *Fast Food Restaurant* is the prominent venue type. These are smaller cities and bedroom communities that tend to be located between larger cities with more ammentiies. At first glance, a "Fast food city" might not seem particularly attractive, but this cluster deserves further exploration for prospective home buyers looking for more seclusion and lower real estate prices.



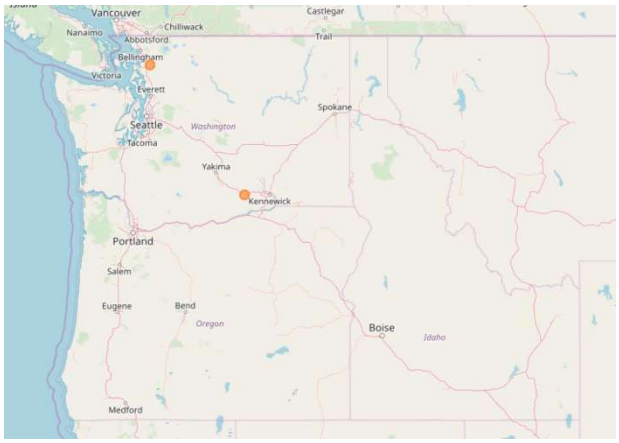
Cluster 4: Coffee shop cities

In addition to coffee shops being the prevalent venue type, the cities in this cluster are charaterized by a diverse set of ammenities, indicative of larger urban centers.



Cluster 5: More outliers

The yield for the final cluster was a couple more outlier cities, also in Washington state. These cities both have a category in their top 5 venue types that is unique among our dataset: the *Gastropub* in Prosser, and the *Bowling Alley* in Sedro-Woolley.



Conclusion

Starting from a list of 653 total cities across the states of Washington, Oregon, and Idaho, I found 593 cities with Foursquare venue data. A Foursquare query of venues in those cities yielded 5935 venues, however it was necessary to filter out cities with fewer than 10 venues, as their data profile later proved insufficient for meaningful clustering. After filtering out those cities, only 180 cities remained—less than 30% of the original group of cities.

The 180 cities used in the final analysis represented 4272 venues and 309 unique venue types. I used the k-means clustering algorithm to group them into six distinct clusters, however only four of those clusters were truly meaningful in terms of revealing insights among the dataset that I could use to answer the original question of the business problem: *how can Northwest residents identify similar cities as prospective places to move?* The results of the analysis certainly provide one answer to the question, with the caveat that there remains a host of many other Northwest cities (not to mention towns) that I weren't able to include in the study for lack of adequate data.

Throughout the process of this study I uncovered limitations in comprehensively addressing the business problem at hand. Nevertheless, I did find some interesting patterns among the refined dataset of larger Northwest cities with an adequate amount of Foursquare venue data. Next steps in the process might be to supplement the data used to cluster cities with additional sources, such as the average home price and population size. With additional data like this, it might be possible to retain and cluster the full list of Northwest cities, while still providing finer-grained grouping patterns for cities with ample Foursqure venue data.

Footnotes

[1] https://www.seattletimes.com/seattle-news/data/114000-more-people-seattle-now-this-decades-fastest-growing-big-city-in-all-of-united-states/

[2] https://www.seattletimes.com/pacific-nw-magazine/surviving-seattles-sidewalks-pedestrian-rage-rises-as-the-population-grows/

[3] For example, https://www.seattletimes.com/opinion/after-14-years-ive-had-it-im-leaving-seattle/

[4] https://www.seattletimes.com/business/home-prices-rising-faster-in-washington-than-in-any-other-state/

[5] https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_Washington

[6] https://en.wikipedia.org/wiki/List_of_cities_in_Oregon

[7] https://en.wikipedia.org/wiki/List_of_cities_in_Idaho

[8] https://developer.foursquare.com/docs/api/venues/explore

[9] https://developer.mapquest.com/documentation/geocoding-api/

[10] https://python-visualization.github.io/folium/