The Pacific Northwest of United States Cities Analysis

INTRODUCTION

Background

The Pacific Northwest of the United States is one of the most quickly developed regions in this country. Besides the unparalleled natural landscape, PNW is also the hub of many well-known international companies, such as Amazon, Microsoft, Boeing, Starbucks, Nike, etc. Given all the things that PNW has to offer, lots of people from outside of this region are looking forward to moving here. Entrepreneurs/business owners are exploring opportunities to start or extend their businesses in this region. Meanwhile, some current residents/businesses of this region would like to see if they can move to nearby cities in order to reduce daily expenses.

Problem/Solution

In all of these situations above, it would be beneficial to conduct an analysis of different cities and better understand what each place has to offer. One way of doing this analysis is to identify similarities among different cities and assign them to different clusters by comparing existing ventures. In this way, people/businesses can narrow down the scope of potential places that they are interested in and make more informed decisions.

Audience

The targeted audience of this analysis are people who are interested in comparing the characteristics of different cities in the northwest corner of the US, based on various venues that are offered in these places. These people include but not limited to potential property buyers, business owners, government officials, etc.

DATA

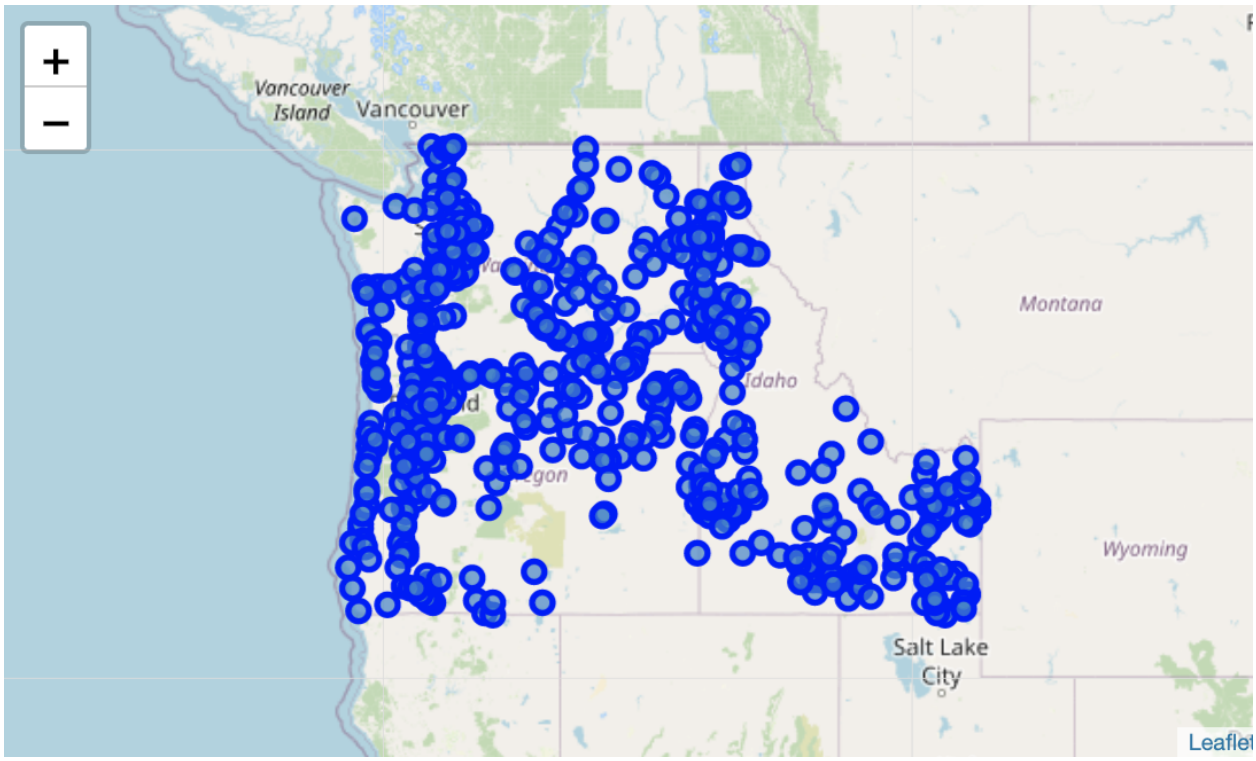
To conduct this analysis about analyzing and clustering cities in the Northwestern United States, I will use data from three different sources as follows:

- Wikipedia
  Obtain a list of cities in Washington, Oregon and Idaho.
  https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_Washington
  https://en.wikipedia.org/wiki/List_of_cities_in_Oregon
  https://en.wikipedia.org/wiki/List_of_cities_in_Idaho
- Foursquare venue recommendation API
  Obtain a list of most popular venues in cities of three states mentioned above
  https://developer.foursquare.com/docs/api/venues/explore
- Mapquest geocoding API
  Obtain latitude and longitude data associated with each location.
  https://developer.mapquest.com/documentation/geocoding-api/

Exploration

By using the Pandas library, I was able to import the tables on Wikipedia websites into the notebook as dataframes. After filtering and clearing the raw table, simplified dataframes were generated with four columns: city, state, longitude and latitude. Then city lists of three states were consolidated into one single dataframe named df_pnw.

Through connecting to the Mapquest API, I was able to look up for coordinates information of each city and import cities' longitude and latitude into the notebook dataframe. Then after geocoding all the cities, I used the Folium mapping library to visualize all the cities in the three states on a map.



Then, by connecting to the Foursquare API, I was able to query the top venues in each city based on the coordinates of cities and venue ratings of Foursquare users. While examining the Foursquare results, I noticed that setting a limit of at least 10 venues entries per city can help generate more meaningful clustering results. Given that, cities below this threshold were dropped from the dataframe. As a result, the data list was truncated from 593 rows to 180 rows, which can drastically reduce the required computing power and conduct more effective analysis.

Methodology

By using the unsupervised machine learning method, k-means clustering, I further analyzed the data and revealed some underlying patterns among different datapoints.

The top 5 venue categories were used to uncover similarities in various cities. The average frequency for each venue category was calculated. Then I converted each venue category into boolean variables using Pandas one-hot encoding method, so I was able to verify that the new dataframe's column count equaled the number of distinct venue categories. Moreover, all rows in the dataframe are grouped by city mean of frequency for each category. As a result, the top 5 most common venues in each city can be populated.

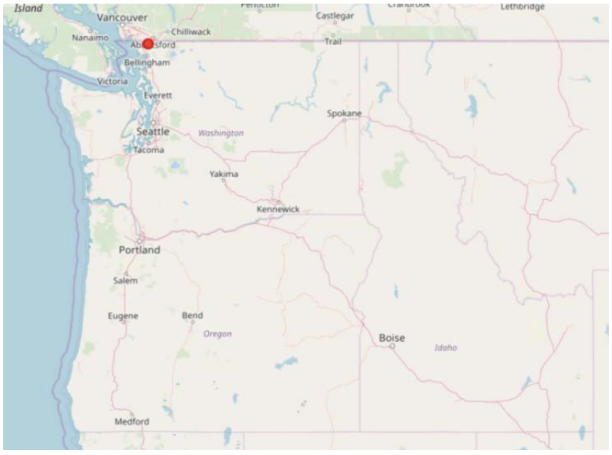
In addition, after several trials, when k=6, the k-means algorithm seems to generate the most meaningful results. The output of executing k-means clustering is an array of cluster assignments for each row of the dataframe. Therefore, I was able to group data by on cluster assignments.

Results and Findings

The Folium library was used to render clusters generated above. The results look reasonable, and the clusters are generally dispersed geographically.

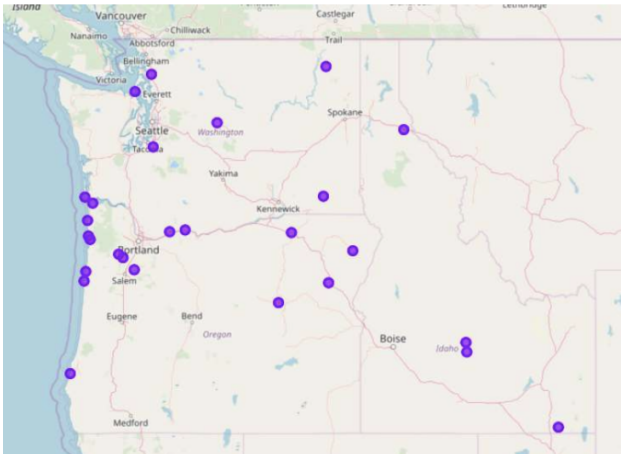
Cluster 0: Outlier City

The one city that was assigned to this cluster is Lynden, WA. Although individually top common venues appear to be normal, but the unique combination of offerings makes this city to be an outlier and different from all other cities in the dataset.

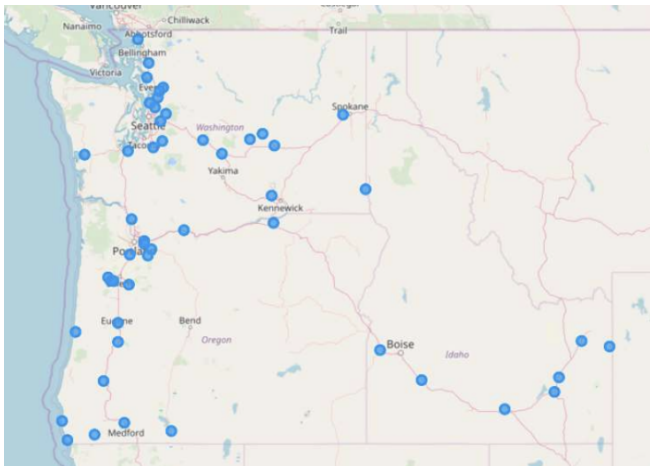


Cluster 1: Vacation Destinations

With lots of accommodations, restaurants and nightlife venues, cities in this cluster are ideal to be treated as vacation destinations. Individuals who are looking for weekend gateway locations can consider these places. Also, businesses that are relating to travel and leisure can see these cities as markets with great potential.

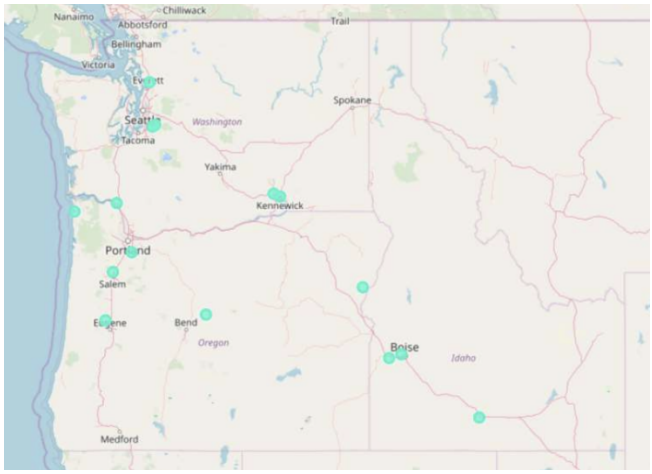


Cluster 2: Restaurant Cities



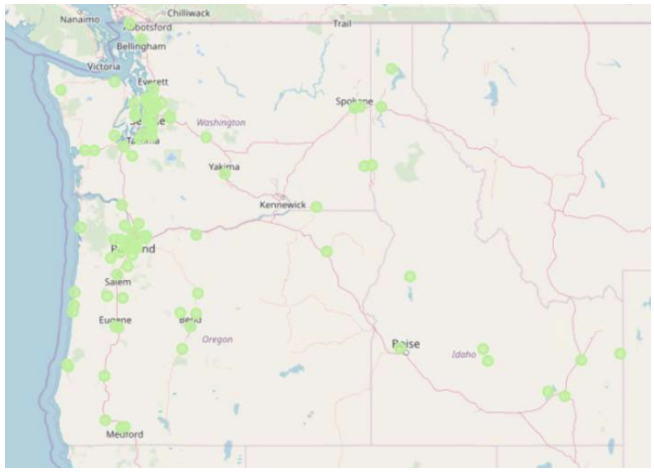
These cities have a high concentration of different restaurants including pizza places, Mexican restaurants, Chinese restaurants, etc. They are generally larger than simple bedroom communities and have some small business centers. But these cities can't be compared with major urban centers in terms of population and city sizes. People who prefer convenient but not too chaotic urban life can consider moving to these places. Businesses who have family spending as major revenue sources can see these spots as profitable markets.

Cluster 3: Fast Food Cities



The popularity of fast food spots is the major characteristic of these cities. These cities are bedroom communities that tend to be located between larger cities and are built for people who enjoy quiet suburban lives. These cities tend to have low living costs and real estate prices. Therefore, small businesses can consider these cities as great options.

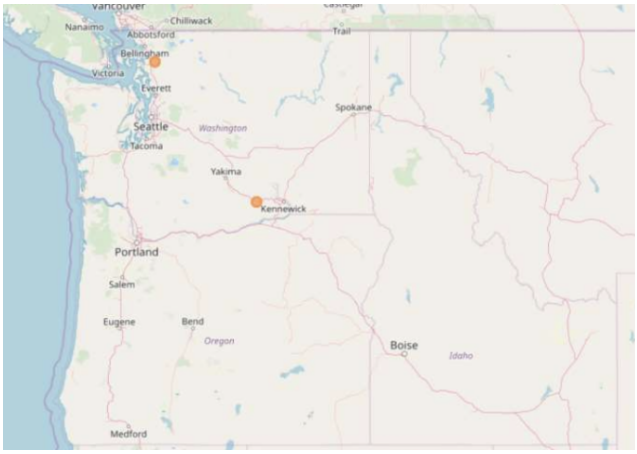
Cluster 4: Coffee Shop Cities



These are clusters of cities that have a great number of coffee shops, which can usually be seen at larger urban centers that also come with a wide diversity of other amenities. People who enjoy busy urban lives should definitely consider these places and evaluate based on their ability to bear high living costs. Businesses can target a very diverse group of customers here and have great chances to collaborate with other businesses or expand its own in a relatively short amount of time.

Cluster 5: Additional outliers

The final cluster suggests two more outlier cities within the State of Washington. They seem to have pretty unique venue categories, like the Bowling Alley in Sedro Woodley and the Gastropub in Prosser. People who are big fans of those venues can consider these cities.



Conclusion

By taking cities in Washington, Oregon and Idaho states into consideration, I conducted an analysis about their similarities based on the most common venue types that are available in different cities. Explaining the meaning behind common patterns within a cluster of cities, I was able to provide suggestions for individuals and businesses that are interested in the Pacific Northwest region of the United States.

Admittedly, due to the lack of data availability, the insights from this analysis can be partial and need to be further improved. However, the result of this report can still give its target audience a general sense of city vibes that are existing in the beautiful PNW.