Comparison of neighbourhoods in major cities of United States and Canada

Yogesh Kuvelkar

28 Feb 2020

Contents

1. Introduction

Background

United States of America is a great place for the people across the world to dream visiting or getting a job or starting their own business and eventually settling down. I think there are many resources available that differentiate the various major cities in USA / Canada in terms of the weather, population, demographics etc. Such readily available resources help aspirants planning to move to USA to decide their cities of preference based on weather, demographics, etc. I think what can help is a readily available resource that also describes how the major cities are similar or dissimilar comparing the neighbourhoods in each city.

Problem

Such a resource will require access to data of various neighbourhoods and the venues in each neighbourhood. Once the data is accessible, it requires analysis and co-relation of various venues in various cities. The project aims to study the neighbourhoods of major cities in United States and Canada and the venues around these cities and identify the pattern or similarity or dissimilarity of venues that are observed in cities.

Interest

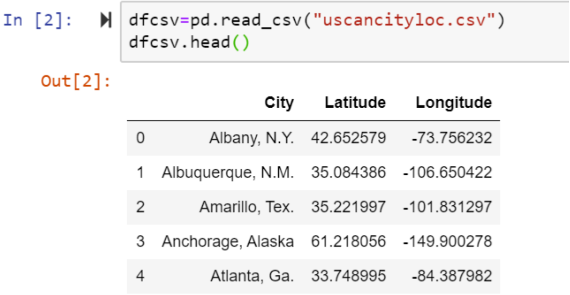
Such a resource will find interest from visitors, entrepreneurs, job-seekers and immigrants who are either new or are planning to move to United States/Canada. For individuals planning to start a restaurant business, this data will provide insights on in which city their cuisine of interest will flourish. For individuals looking for a job, this data will provide insights on neighbourhoods where they can reside and grow their family.

2. Data

Data Sources

There are two data sources that have been utilised for this project :

a) I myself prepared a list of major cities in United States/Canada and their location coordinates. I did use https://gpscoordinates.info/ for the latitude and longitude details.



In [2]:

```
dfcsv=pd.read_csv("uscancityloc.csv")
dfcsv.head()
```

Out[2]:

	City	Latitude	Longitude
0	Albany, N.Y.	42.652579	-73.756232
1	Albuquerque, N.M.	35.084386	-106.650422
2	Amarillo, Tex.	35.221997	-101.831297
3	Anchorage, Alaska	61.218056	-149.900278
4	Atlanta, Ga.	33.748995	-84.387982

b) The Venues data from FourSquare
For this I have registered with FourSquare in their developer environment and utilised their public APIs to access the Venue data. This API is based on the location latitude and longitude passed to it.
API used: https://api.foursquare.com/v2/venues/explore
Documentation: https://developer.foursquare.com/docs/api/venues/explore

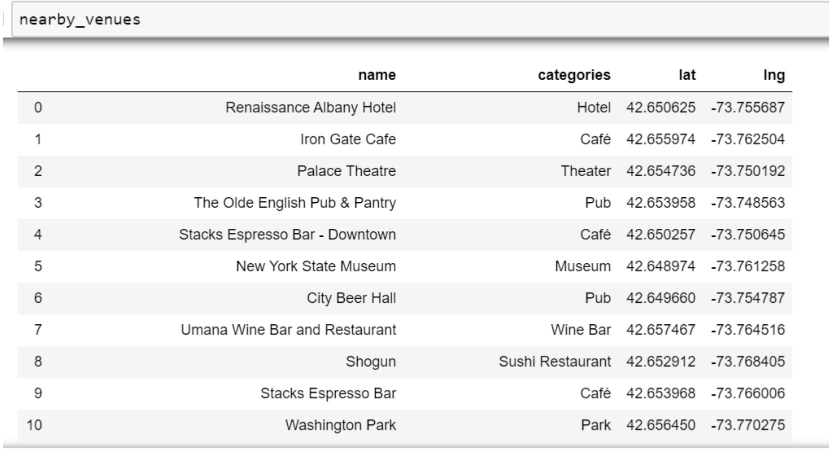
3. Methodology

Exploratory Data Analysis

I plotted the various cities of the map of the USA/ Canada using the folium package. This provided a good visualization of the cities.

United
México

Then the FourSquare API call was tested for 1 city in the dataset. Because we are looking at venues of the entire city, we need a rather large radius. To illustrate, New York City’s Manhattan alone can stretch for 20km. Thus, a rather large radius of 15km was selected. The json response was then cleansed to get the city neighbourhood details in the below format



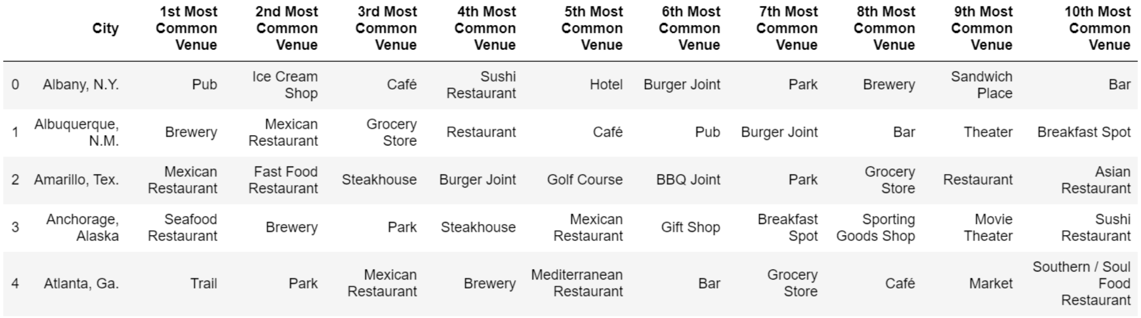
```
nearby_venues
```

	name	categories	lat	lng
0	Renaissance Albany Hotel	Hotel	42.650625	-73.755687
1	Iron Gate Cafe	Café	42.655974	-73.762504
2	Palace Theatre	Theater	42.654736	-73.750192
3	The Olde English Pub & Pantry	Pub	42.653958	-73.748563
4	Stacks Espresso Bar - Downtown	Café	42.650257	-73.750645
5	New York State Museum	Museum	42.648974	-73.761258
6	City Beer Hall	Pub	42.649660	-73.754787
7	Umana Wine Bar and Restaurant	Wine Bar	42.657467	-73.764516
8	Shogun	Sushi Restaurant	42.652912	-73.768405
9	Stacks Espresso Bar	Café	42.653968	-73.766006
10	Washington Park	Park	42.656450	-73.770275

Then a loop structure was used to retrieve the neighborhood details of each city in the dataset. A combined. The most common venues can then be determined which will form the basis for clustering.

Note: it was observed that some venues are quite common across all cities, Coffee Shops, American Restaurants and Pizza Place in particular. These outliers were removed from the dataset.

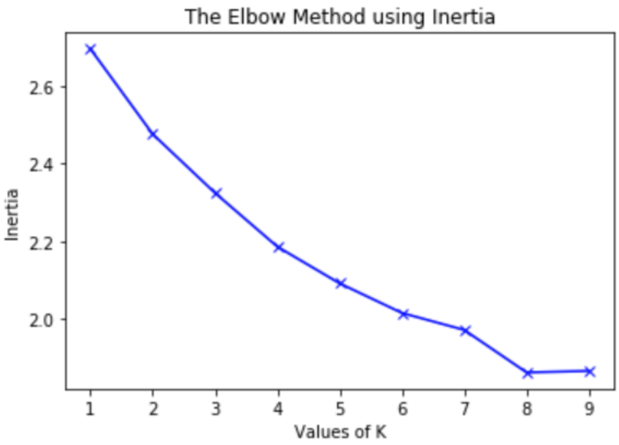
The following table excerpt was produced to give a sense of the types of venues are common in each city.



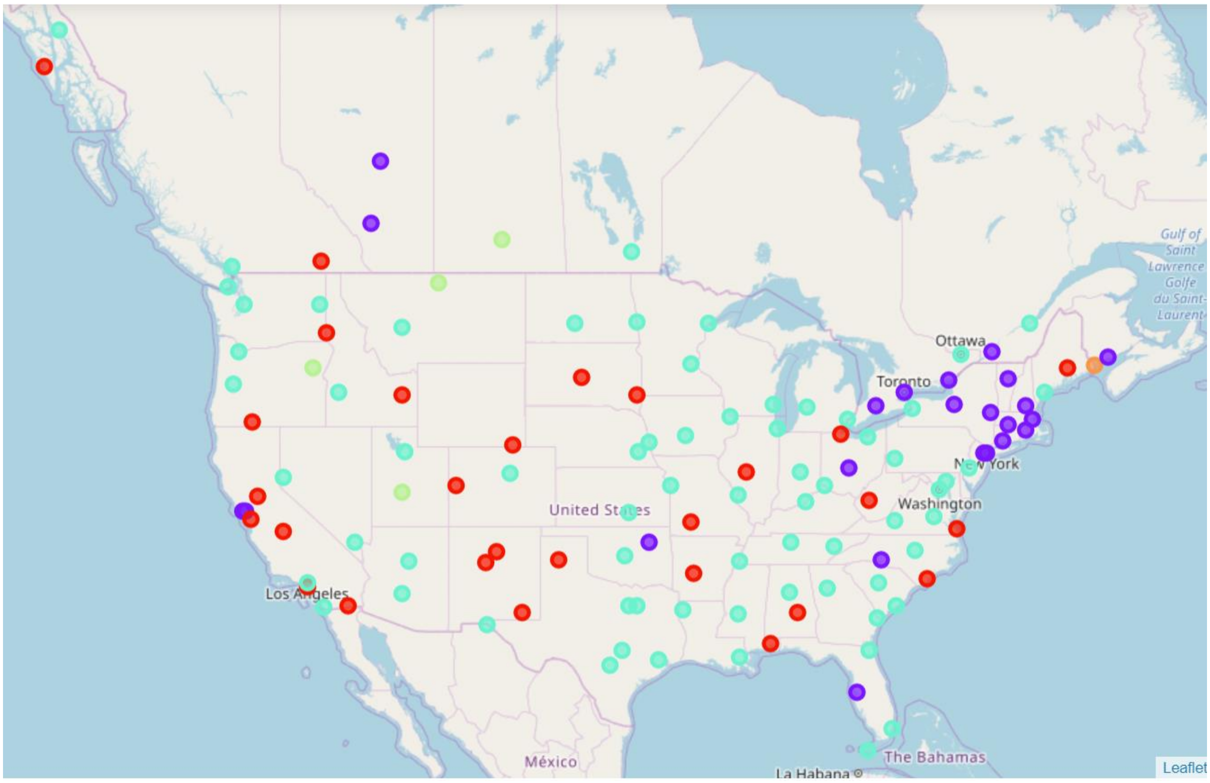
	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Albany, N.Y.	Pub	Ice Cream Shop	Café	Sushi Restaurant	Hotel	Burger Joint	Park	Brewery	Sandwich Place	Bar
1	Albuquerque, N.M.	Brewery	Mexican Restaurant	Grocery Store	Restaurant	Café	Pub	Burger Joint	Bar	Theater	Breakfast Spot
2	Amarillo, Tex.	Mexican Restaurant	Fast Food Restaurant	Steakhouse	Burger Joint	Golf Course	BBQ Joint	Park	Grocery Store	Restaurant	Asian Restaurant
3	Anchorage, Alaska	Seafood Restaurant	Brewery	Park	Steakhouse	Mexican Restaurant	Gift Shop	Breakfast Spot	Sporting Goods Shop	Movie Theater	Sushi Restaurant
4	Atlanta, Ga.	Trail	Park	Mexican Restaurant	Brewery	Mediterranean Restaurant	Bar	Grocery Store	Café	Market	Southern / Soul Food Restaurant

Modelling

I used the clustering machine learning technique to determine which cities are similar or dissimilar to each other. I used the K-means elbow method to determine the optimal number of clusters, as below.



It seems a toss-up here between k = 5 and k = 6. Because I would like to have some nuance in the clustering, I would prefer to use k = 6. After running the clustering with this k value, the clustering looks rather interesting.



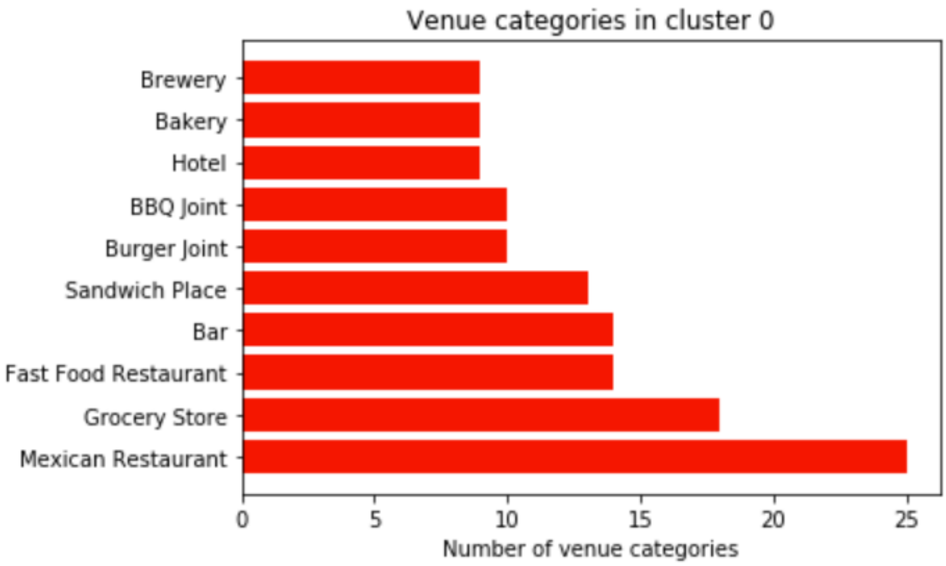
The different coloured dots represent these clusters:

● Cluster 0
● Cluster 1
● Cluster 2
● Cluster 3
● Cluster 4
● Cluster 5

Then I explored each cluster to determine the common venue categories that define each cluster and listed the cities that form the cluster.

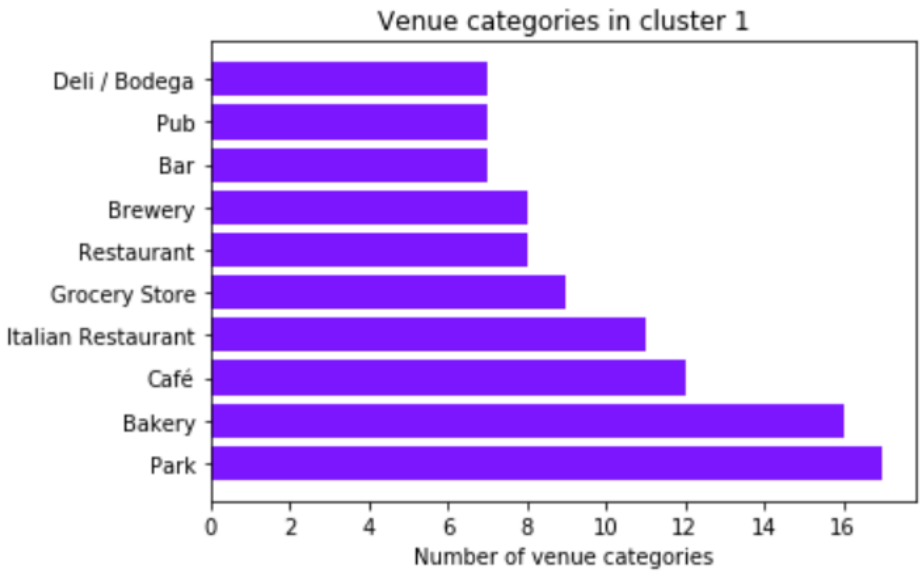
4. Results

Cluster 0



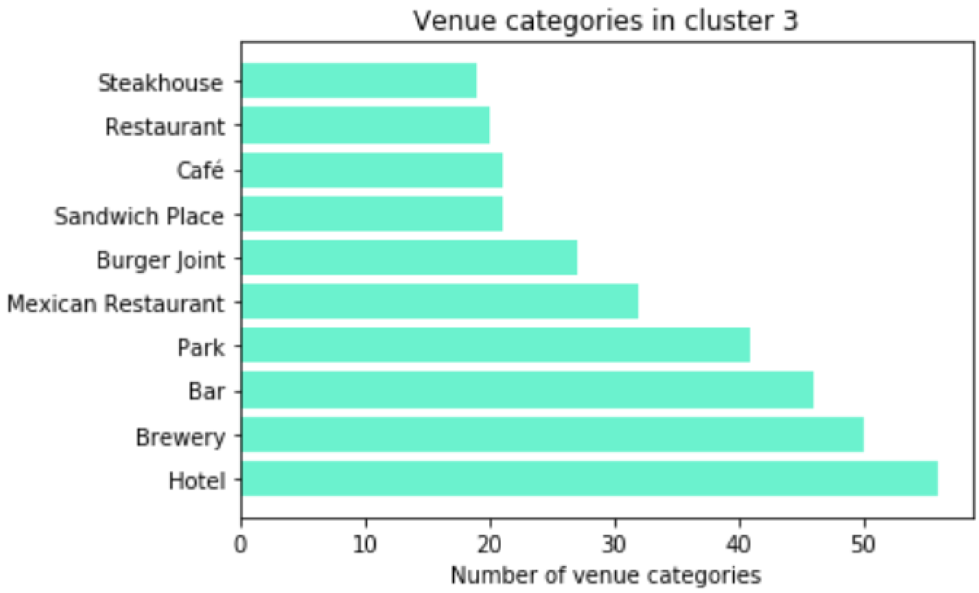
```
['Albuquerque, N.M.',
 'Amarillo, Tex.',
 'Bangor, Maine',
 'Carlsbad, N.M.',
 'Charleston, W. Va.',
 'Cheyenne, Wyo.',
 'El Centro, Calif.',
 'Fresno, Calif.',
 'Grand Junction, Colo.',
 'Hot Springs, Ark.',
 'Idaho Falls, Idaho',
 'Klamath Falls, Ore.',
 'Lewiston, Idaho',
 'Long Beach, Calif.',
 'Mobile, Ala.',
 'Montgomery, Ala.',
 'Nelson, B.C., Can.',
 'Pierre, S.D.',
 'Sacramento, Calif.',
 'San Jose, Calif.',
 'Santa Fe, N.M.',
 'Sioux Falls, S.D.',
 'Sitka, Alaska',
 'Springfield, Ill.',
 'Springfield, Mo.',
 'Toledo, Ohio',
 'Virginia Beach, Va.',
 'Wilmington, N.C.']
```

Cluster 1



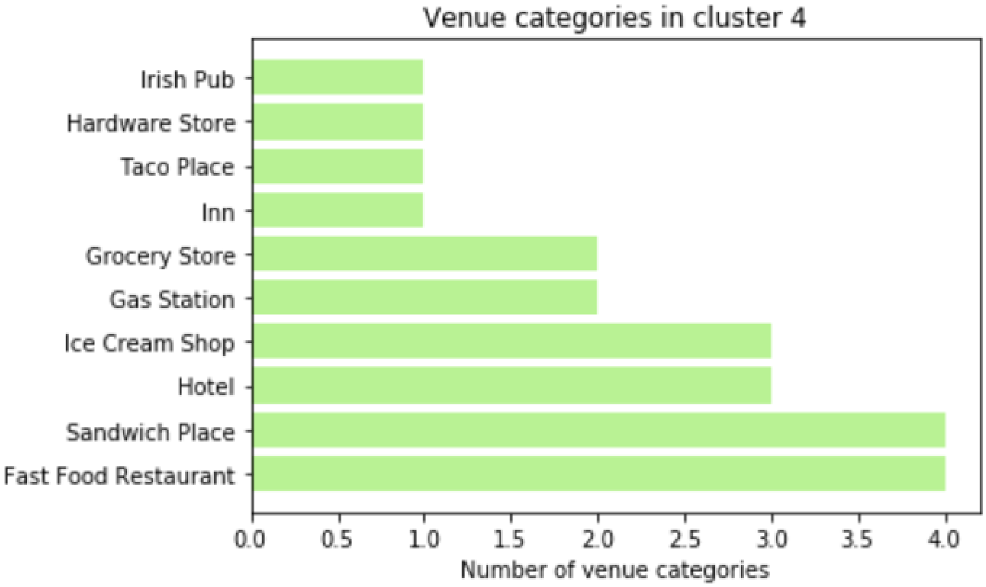
```
['Albany, N.Y.',
 'Boston, Mass.',
 'Calgary, Alba., Can.',
 'Charlotte, N.C.',
 'Columbus, Ohio',
 'Edmonton, Alb., Can.',
 'Kingston, Ont., Can.',
 'London, Ont., Can.',
 'Manchester, N.H.',
 'Montpelier, Vt.',
 'Montreal, Que., Can.',
 'New Haven, Conn.',
 'New York, N.Y.',
 'Newark, N.J.',
 'Oakland, Calif.',
 'Providence, R.I.',
 'San Francisco, Calif.',
 'Springfield, Mass.',
 'St. John, N.B., Can.',
 'Syracuse, N.Y.',
 'Tampa, Fla.',
 'Toronto, Ont., Can.',
 'Tulsa, Okla.']
```

Cluster 3



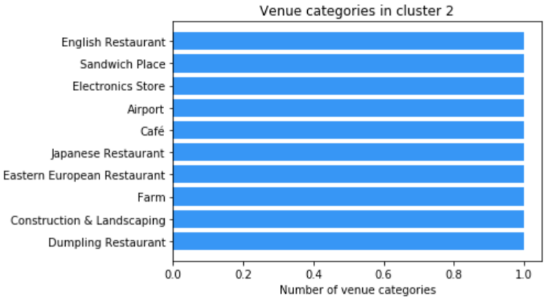
```
['Anchorage, Alaska',
 'Atlanta, Ga.',
 'Austin, Tex.',
 'Baltimore, Md.',
 'Birmingham, Ala.',
 'Bismarck, N.D.',
 'Boise, Idaho',
 'Buffalo, N.Y.',
 'Charleston, S.C.',
 'Chicago, Ill.',
 'Cincinnati, Ohio',
 'Cleveland, Ohio',
 'Columbia, S.C.',
 'Dallas, Tex.',
 'Denver, Colo.',
 'Des Moines, Iowa',
 'Detroit, Mich.',
 'Dubuque, Iowa',
 'Duluth, Minn.',
 'El Paso, Tex.',
 'Eugene, Ore.',
 'Fargo, N.D.',
 'Flagstaff, Ariz.',
 'Fort Worth, Tex.',
 'Grand Rapids, Mich.',
 'Helena, Mont.',
 'Honolulu, Hawaii',
 'Houston, Tex.',
 'Indianapolis, Ind.',
 'Jackson, Miss.',
 'Jacksonville, Fla.',
 'Juneau, Alaska',
 'Kansas City, Mo.',
 'Key West, Fla.',
 'Knoxville, Tenn.',
 'Las Vegas, Nev.',
 'Lincoln, Neb.',
 'Los Angeles, Calif.',
 'Louisville, Ky.',
 'Memphis, Tenn.',
 'Miami, Fla.',
 'Milwaukee, Wis.',
 'Minneapolis, Minn.',
 'Nashville, Tenn.',
 'New Orleans, La.',
 'Oklahoma City, Okla.',
 'Omaha, Neb.',
 'Ottawa, Ont., Can.',
 'Philadelphia, Pa.',
 'Phoenix, Ariz.',
 'Pittsburgh, Pa.',
 'Portland, Maine',
 'Portland, Ore.',
 'Quebec, Que., Can.',
 'Raleigh, N.C.',
 'Reno, Nev.',
 'Richmond, Va.',
 'Roanoke, Va.',
 'Salt Lake City, Utah',
 'San Antonio, Tex.',
 'San Diego, Calif.',
 'San Juan, P.R.',
 'Savannah, Ga.',
 'Seattle, Wash.',
 'Shreveport, La.',
 'Spokane, Wash.',
 'St. Louis, Mo.',
 'Vancouver, B.C., Can.',
 'Victoria, B.C., Can.',
 'Washington, D.C.',
 'Wichita, Kan.',
 'Winnipeg, Man., Can.']
```

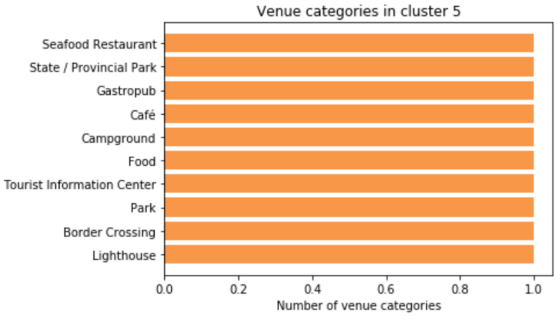
Cluster 4



```
['Baker, Ore.',
 'Havre, Mont.',
 'Moose Jaw, Sask., Can.',
 'Richfield, Utah']
```

Outlier Clusters - Cluster 2 (Nome, Alaska) and Cluster 5 (Eastport, Maine)





5. Discussion

Based on the above, the aspirants should get a good sense of similarities and dissimilarities of the major cities in USA / Canada. Below are some guidelines that the aspirant could use.

Residential: less shops and restaurants than elsewhere

Residential with services: well connected with shops and restaurants

Cultural & going out places: well connected with lots of cultural places and clubs

Activity centres: large activity and large number of events

Shopping areas: more shops and less cultural/going out places

Universities: well connected with public services and university facilities

Randall's Island: alone within its cluster of parkings and entertainement

6. Conclusion

Above insights combined with other readily available resources such as weather, population, demographics can help aspirants planning to come to United States make a good informed decision.

7. Acknowledgment

A big thanks to Coursera and IBM for offering the IBM Data Science Professional Certificate.