Using k-Means Clustering to Identify Locations to Visit that are Like the Houston Heights

Introduction

My wife and I retired 5 years ago. We got rid of 99% of our stuff, sold our house, and bought a fifth wheel trailer to travel the United States and Canada with our two pugs, Pancho and Lefty. Over this time, we have visited over one hundred locations, generally staying between four to seven days. While it has been an amazing ride, we are starting to think about where we might want to settle down. By settle down, we mean to continue living in the 5th wheel but stay in a location from one to three months.

Earlier this year, I began taking classes on Coursera to qualify for the IBM Data Science Professional Certificate. The certification program consists of nine courses which cover a variety of data science topics including: open source tools and libraries, methodologies, Python, databases and SQL, data visualization, data analysis, and machine learning. The final course is a capstone project where the student applies the skills developed in the previous eight courses. For my capstone project, I decided to find locations Deb and I had not visited but that we might want to spend at least a month.

Before we retired, we lived in a neighborhood called the Houston Heights in Houston, Texas. We loved the area because of its convenient location (less than 5 miles from Downton), the friendly neighbors, and the neighborhood's eclectic and quirky shops and restaurants.

For the project, I chose to find areas in the United States with RV parks that are like the Houston Heights neighborhood we lived in. The criteria I used were areas (a) with RV parks (b) that are less than 60 miles from a Costco Warehouse and (c) that have similar venues within a 5-mile radius as our old neighborhood. We keep all our prescriptions with Costco pharmacy and like to visit Costco at least once a month to buy bulk supplies. The 5-mile radius allows for a short drive to local venues.

Data

I relied on three primary data sources to find RV parks within 60 miles of a Costco Warehouse and located within the contiguous 48 United States in areas like the Houston Heights.

- Costco_USA_Canada.csv, a list of Costco Warehouses located in the United States and Canada. I found this dataset on POI Factory
- GoodSam.csv, a list of campgrounds in the United States and Canada that offer discounts to members of Good Sam Club. I found this dataset on POI Factory
- Demographic data from the Census Bureau's American Community Survey 5-Year Data (2009-2018) ("ACS").

The datasets for the location of RV Parks and Costco Warehouses included Zip Codes. However, the dataset for Census data used ZCTA as a key. To insure all three datasets were based on the same key, I used the US Zip Codes Database provided by Simple Maps to cross reference the ZCTA included in the results of the demographic analysis to the related Zip Codes, and modify the results to present Zip Codes instead of ZCTAs.

To find which locations are most common with the Houston Heights, I compared the five most popular venues within a five-mile location of the center of each zip code. Data for each zip code was provided by Foursquare, a social location service that allows users to explore the world around them.

Methodology

- Good Sam RV Parks Dataset

My first step was to remove from the Good Sam RV Parks dataset any RV Parks outside of the 48 contiguous United States and any entries that did not have location data, such as latitude, longitude, city, and state. The

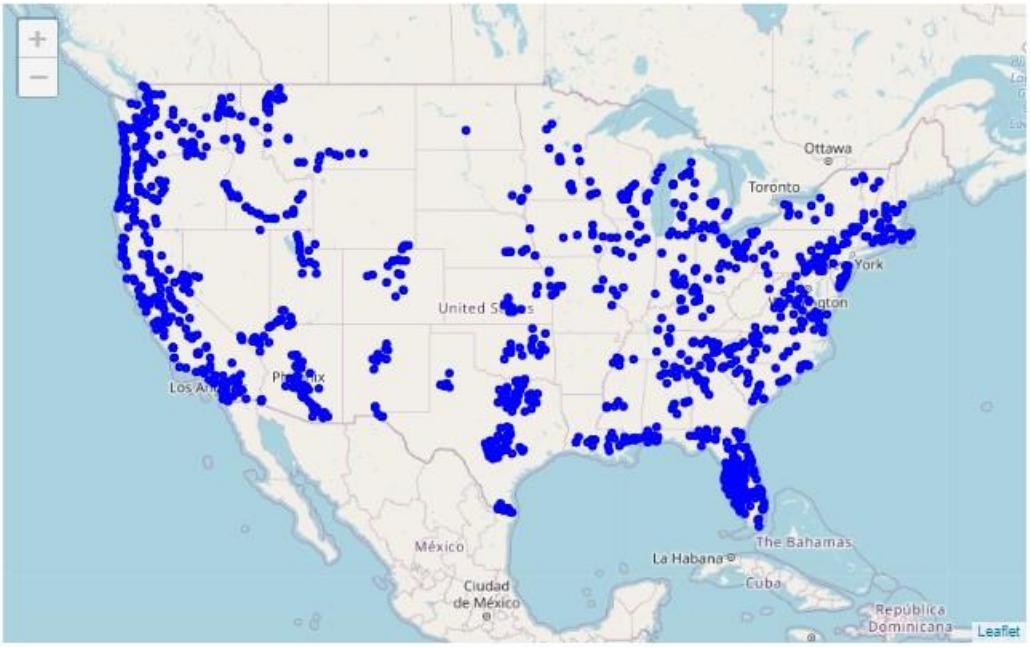
GoodSam.csv dataset had four columns – “Latitude”, “Longitude”, “Description”, and “Address”. To make the dataset usable for my analysis, I disaggregated the Description column and the Address column. I divided the Description column into two columns named “Park Name” and “Location”. I dropped the Location column as its data was partially redundant with data derived from the Address column. I divided the Address Column into five columns titled “Address”, “City”, “State, “Zip”, and “Phone Number”. After completing my data wrangling, the Good Sam database included 2,216 unique RV Parks in the 48 Contiguous United States.

- Costco Warehouse Dataset

I removed from the Costco Warehouse Dataset all Costco Warehouses located outside of the 48 contiguous United States and any entries that did not have location data, such as latitude, longitude, city, and state. Like the GoodSam.csv dataset, the Costco_USA_Canada.csv dataset consisted of four columns – “Latitude”, “Longitude”, “Description”, and “Address”. To make the dataset usable for my analysis, I disaggregated the Description column and the Address column. I divided the Description column into two columns named “Park Name” and “Location”, then dropped the Location column as its data was partially redundant with data derived from the Address column. I divided the Address Column into five columns titled “Address”, “City”, “State, “Zip”, and “Phone Number”. Upon completion of my data preparation, the Good Sam database included 543 Costco Stores located in the 48 Contiguous United States.

- . Find RV Parks within 60 miles of a Costco Warehouse

After cleaning and reformatting the Good Sam RV Park dataset and the Costco Warehouse dataset, I wanted to find the distance from each RV Park to the closest Costco Warehouse. To do so, I designed an algorithm which calculated for each RV Park the geodesic distance between it and each Costco Warehouse, saving the lowest distance calculated. The analysis found 1,473 RV Parks in 984 cities and 1,146 unique zip codes that are within 60 miles of a Costco.



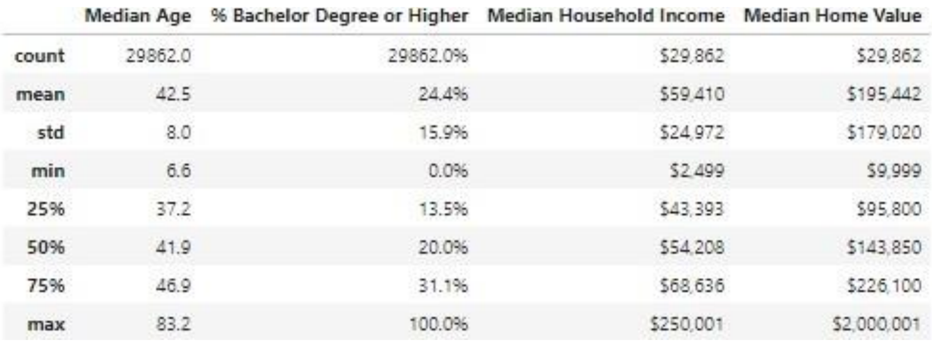
- Demographic Data

My wife and I discussed factors that we considered most important to us to find areas which we would like to spend more time. We reviewed the variables reported in the Census Bureau’s American Community Survey to select the specific demographic characteristics we wanted to consider. The ACS is an ongoing annual survey covering a broad range of topics about social, economic, demographic, and housing characteristics of the U.S. population. We decided to analyze data provided in the ACS 5-Year Data “Data Profiles”. Data Profiles is the smallest dataset in the ACS and includes over 1,000 variables covering a broad range of social, economic, housing, and demographic information presented as population counts and percentages.

The list of locations in the United States to consider will be Zip Code Tabulation Areas ("ZCTA"). Per the Census Bureau, ZCTA "are generalized areal representations of United States Postal Service (USPS) ZIP Code service areas." ZCTA is the smallest geographical area for which the Census Bureau provides demographic data. As such, I believe ZCTA and Zip Codes best represent neighborhoods within given locations.

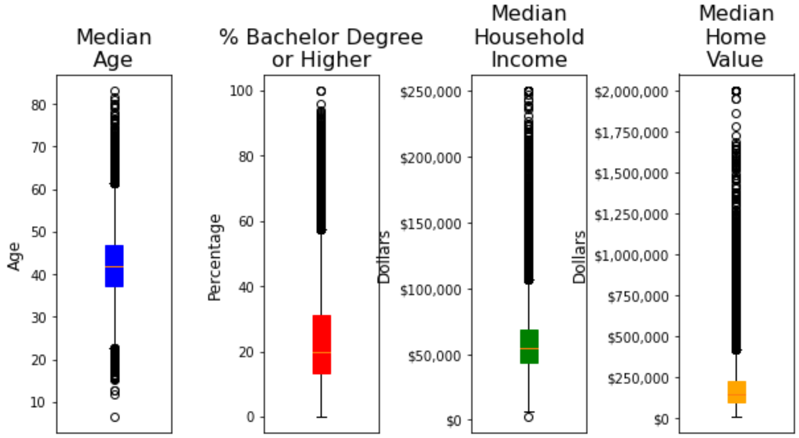
The four demographic variables we chose to consider were the estimated median age of the ZCTA population (DP05_0018E), estimated percentage of the ZCTA population over 25 with a Bachelor's degree or higher (DP02_0067PE), estimated median household income for each ZCTA (DP03_0062E), and the estimated median value of owner-occupied residences (DP04_0089E). We chose estimated median age because, although we are retired, we wanted to be in an area filled with a range of ages like the Houston Heights. We selected population with a bachelor's degree or higher because we both have graduate degrees and like being around people with whom we can discuss issues and new ideas. We chose estimated median household income and median home value to represent housing affordability.

Using the Census API, I retrieved the four variables for every ZCTA. For my next step, I cleaned the Census data by dropping all rows where one or more values were less than zero or blank. The table below presents a summary of basic statistical details of the Census data.

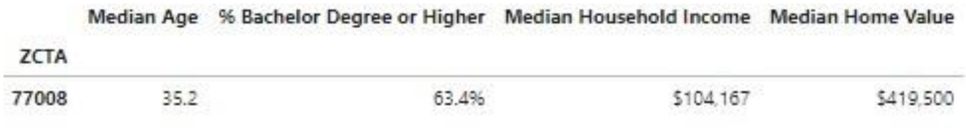


	Median Age	% Bachelor Degree or Higher	Median Household Income	Median Home Value
count	29862.0	29862.0%	$29,862	$29,862
mean	42.5	24.4%	$59,410	$195,442
std	8.0	15.9%	$24,972	$179,020
min	6.6	0.0%	$2,499	$9,999
25%	37.2	13.5%	$43,393	$95,800
50%	41.9	20.0%	$54,208	$143,850
75%	46.9	31.1%	$68,636	$226,100
max	83.2	100.0%	$250,001	$2,000,001

The chart below presents a box plot of the four demographic criteria considered.



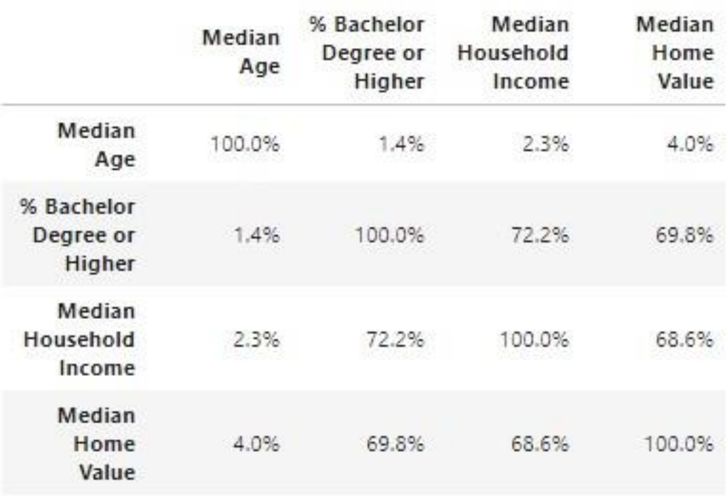
The table below presents the demographic data for the ZCTA in which our Houston Heights neighborhood was located.



	Median Age	% Bachelor Degree or Higher	Median Household Income	Median Home Value
ZCTA				
77008	35.2	63.4%	$104,167	$419,500

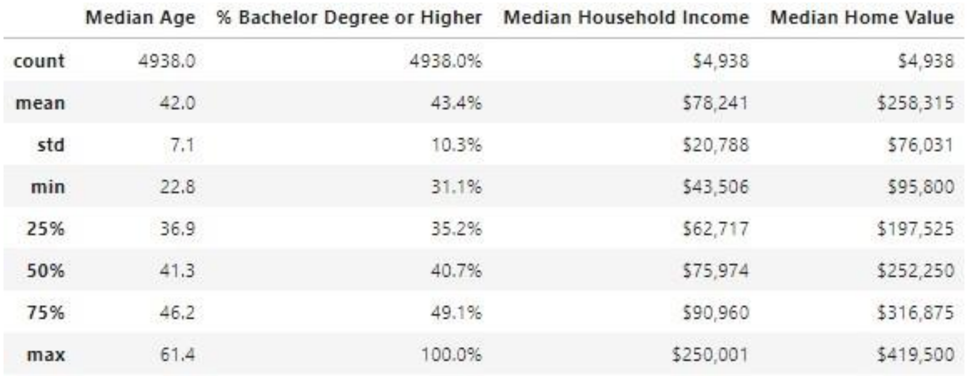
For three demographic criteria (Percentage of population with a bachelor's degree or higher, Median Household Income, and Median Home Value), the Houston Heights is in an outlier while the estimated Median Age is in the bottom quartile.

As the table below shows, Median Household Income, Median Home Value, and % Bachelor's Degree or higher are highly correlated.



	Median Age	% Bachelor Degree or Higher	Median Household Income	Median Home Value
Median Age	100.0%	1.4%	2.3%	4.0%
% Bachelor Degree or Higher	1.4%	100.0%	72.2%	69.8%
Median Household Income	2.3%	72.2%	100.0%	68.6%
Median Home Value	4.0%	69.8%	68.6%	100.0%

To filter the Census demographic data, I eliminated ZCTAs (a) which were outliers[1] of the Median Age, (b) where the percentage of the population over 25 with a bachelor's degree or higher were not in the fourth quartile, (c) where Median Household Income was in the first quartile, and (d) where the Median Home Value was greater than the Median Home Value in ZCTA 77008 or in the first quartile. The table below presents the summary statistics of the filtered Census demographic dataset.



	Median Age	% Bachelor Degree or Higher	Median Household Income	Median Home Value
count	4938.0	4938.0%	$4,938	$4,938
mean	42.0	43.4%	$78,241	$258,315
std	7.1	10.3%	$20,788	$76,031
min	22.8	31.1%	$43,506	$95,800
25%	36.9	35.2%	$62,717	$197,525
50%	41.3	40.7%	$75,974	$252,250
75%	46.2	49.1%	$90,960	$316,875
max	61.4	100.0%	$250,001	$419,500

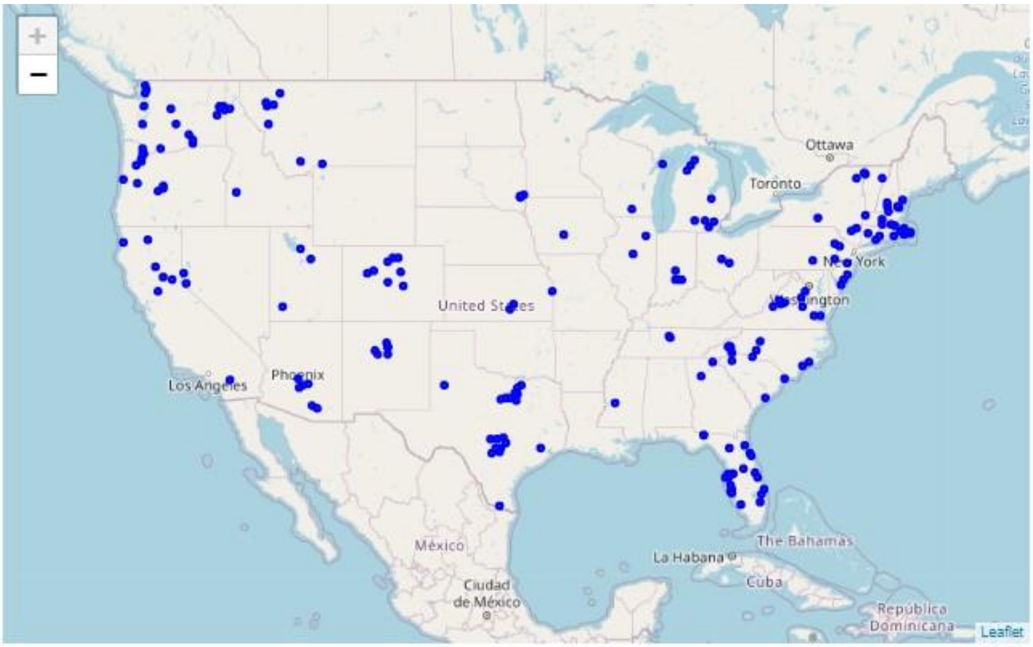
As discussed above, the dataset of RV Parks within 60 miles of a Costco Warehouse uses Zip Codes as a reference, while the dataset for Census data uses ZCTAs as a reference. I used an inner merge of the filtered Census

[1] Outliers are points outside of the range from the 1st quartile – 1.5*IQR to 3rd quartile +1.5 IQR, where IQR is the difference between the 3rd quartile and the 1st quartile.

demographic dataset and the Simple Maps US Zip Codes dataset based on ZCTA, then dropped the ZCTA column of the merged dataset to yield a Census demographic dataset with Zip Codes instead of ZCTA, and the latitude and longitude of each Zip Code.

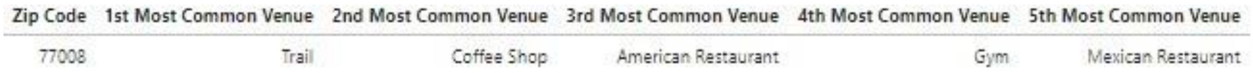
- Combine Filtered Demographic Data and the Combined RV Parks and Costco Dataset

Next, I inner merged the filtered Census demographic data with Zip Codes dataset and the dataset of RV Parks within 60 miles of a Costco Warehouse, which yielded a dataset (the "Final Dataset") of 198 unique zip codes in 186 cities. The map below presents those locations.



- Download Venue Data for Selected Zip Codes from Foursquare

To find the locations most common with the Houston Heights, I compared the most popular venues within a five-mile location of the center of each zip code in the Final Dataset. I collected such data for each zip code using the Foursquare API. From Foursquare, I requested the 100 most popular venues in each Zip Code within a five-mile radius. I then removed any Zip Code which had less than 50 venues. For each remaining Zip Code, I determined the 5 most popular venues. The screening found 371 unique venues over 110 unique Zip Codes with 50 venues or more. The table below presents the information for Zip Code 77008, the zip code for the Houston Heights neighborhood in which we lived.



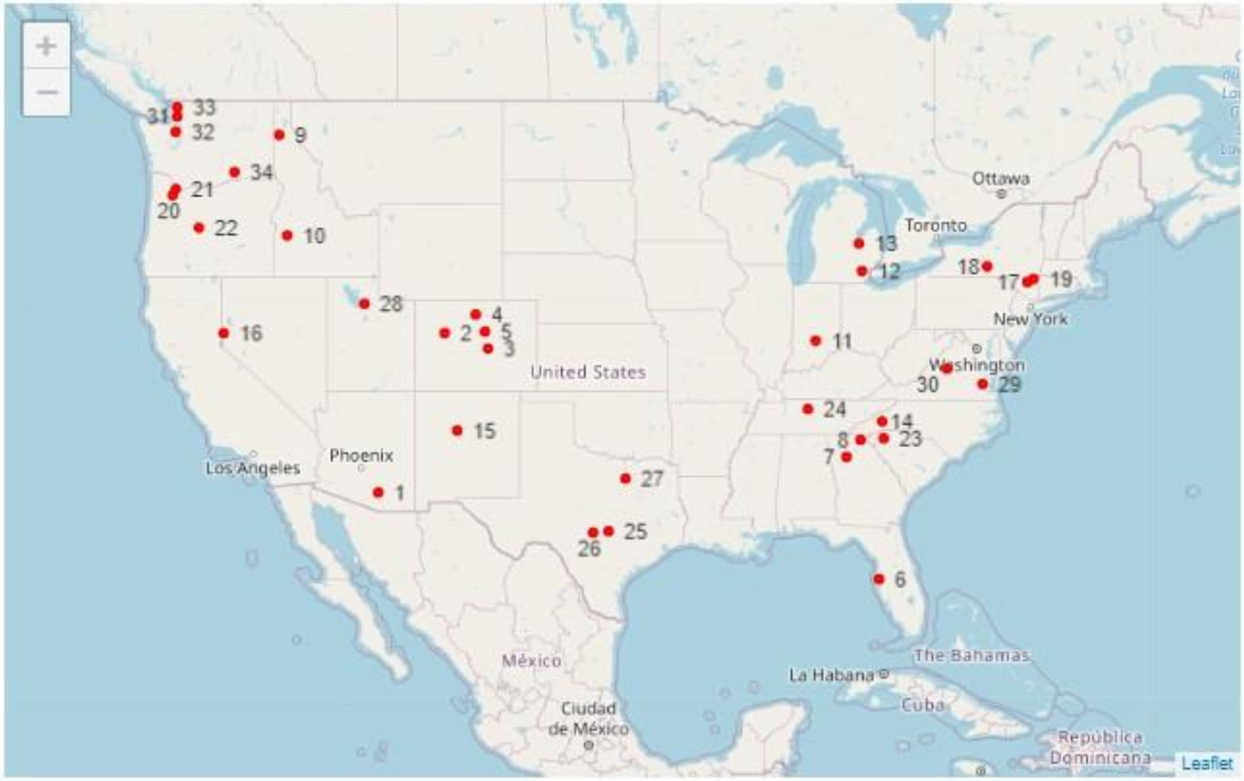
Zip Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
77008	Trail	Coffee Shop	American Restaurant	Gym	Mexican Restaurant

- Find Zip Codes Similar to the Houston Heights

To find zip codes most like the Houston Heights, I used the *k*-means clustering method by using. The *k*-means clustering algorithm identifies *k* number of centroids, then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest unsupervised machine learning algorithms and is highly suited for this project. For this analysis, I iterated the number of *k*-means to cluster the Zip Codes until the algorithm found 35 or fewer unique counties with Zip Codes with venues like the Houston Heights. The search found 34 unique locations.

Results

The map below shows the final 34 locations found by my analysis.



The table below is list of the locations sorted by state presented in the final map. The ID number corresponds to the number on the map.

ID	City	County	State	ID	City	County	State
1	Tucson	Pima County	AZ	18	Ithaca	Tompkins County	NY
2	Glenwood Springs	Garfield County	CO	19	Rhinebeck	Dutchess County	NY
3	Colorado Springs	El Paso County	CO	20	Wilsonville	Clackamas County	OR
4	Estes Park	Larimer County	CO	21	Portland	Multnomah County	OR
5	Englewood	Arapahoe County	CO	22	Bend	Deschutes County	OR
6	Tampa	Hillsborough County	FL	23	Greenville	Greenville County	SC
7	Marietta	Cobb County	GA	24	Nashville	Davidson County	TN
8	Sautee Nacoochee	White County	GA	25	Austin	Travis County	TX
9	Coeur D Alene	Kootenai County	ID	26	Fredericksburg	Gillespie County	TX
10	Garden City	Ada County	ID	27	Grapevine	Tarrant County	TX
11	Nashville	Brown County	IN	28	North Salt Lake	Davis County	UT
12	Ypsilanti	Washtenaw County	MI	29	Williamsburg	James City County	VA
13	Frankenmuth	Saginaw County	MI	30	Greenwood	Albemarle County	VA
14	Swannanoa	Buncombe County	NC	31	La Conner	Skagit County	WA
15	Albuquerque	Bernalillo County	NM	32	Poulsbo	Kitsap County	WA
16	Reno	Washoe County	NV	33	Bellingham	Whatcom County	WA
17	Accord	Ulster County	NY	34	Richland	Benton County	WA

Discussion

The *k*-means clustering algorithm found 34 unique locations in 18 states. Several locations we have visited and enjoyed like Tucson, Arizona, Tampa, Florida, Coeur D'Alene, Idaho, Swannanoa, North Carolina, Albuquerque New Mexico, Portland, Oregon, Bend, Oregon, Nashville, Tennessee, Austin, Texas, Fredericksburg, Texas, North Salt Lake, Utah, and the area around Seattle, Washington (La Conner, Poulsbo and Bellingham).

The clustering identified several areas that we had not visited and had not considered like Glenwood Springs, Colorado, the Denver, Colorado area (Colorado Springs, Estes Park, and Englewood), Ypsilanti, Michigan, Frankenmuth, Michigan, Reno, Nevada, Ulster County, New York, Dutchess County, New York, Ithaca, New York, Greenville, South Carolina, and Greenwood, Virginia. Since the clustering identified several places that we have enjoyed and have considered staying at for at least a month, we plan to include these places that are new to us in our travels to see how much we like them.

My sense is the final results would differ if the number of venues used to cluster areas was increased to 10, the maximum number of unique counties to end the *k*-clustering algorithm had been reduced or increased, the demographic data was included in clustering process. Results may have been different if I had been able to measure the drive time between the RV Parks and Costco Warehouses instead of the distance. It would be interesting to use a larger geographic area, like congressional districts or Standard Metropolitan/Micropolitan Statistical Area, to filter the demographic data and to identify the 100 most popular venues in the area to find places to visit.

I believe the analysis would have been faster if I had screened the demographic data first then merged that resulting data with the RV Park data. I expect that such a step would have reduced the number of RV Parks to find measure the distance to the closest Costco Warehouse.

We enjoy spending two to three months during the summer in Canada. I would like to prepare a similar analysis using Canadian demographic data, RV Parks, Costco Warehouses, and Foursquare data.

Conclusion

I used the *k*-means clustering technique to find areas like the Houston Heights neighborhood where we lived before we retired. I used location data for RV Parks and Costco Warehouses in the United States along with demographic data for ZCTAs from the Census Bureau to find 198 unique locations to consider. I then retrieved data the100 most popular venues within a 5-mile radius of the Zip Code's geolocation. The *k*-means clustering method found 34 unique areas with similar venues as the Houston Heights neighborhood. The clustering found several locations that we have visited, have enjoyed, and have considered spending at least one month in the future. As a result, I believe it is highly likely that we would enjoy the areas found which we have not visited found by the *k*-means clustering method.