Retirement Possibilities in the State of West Virginia

Introduction

Many of the capstone projects for the Applied Data Science specialization focus on beginnings, such as deciding where to open a new business. This project will focus instead on endings: can neighborhoods be classified in such a way as to reveal their potential as good retirement areas? The criteria can be very different than those used to decide where to live for younger adults. No longer concerned about job prospects, average income and the quality of schools, retirees instead focus on where their money will go furthest. Access to medical care, recreational opportunities, a mild climate, and ease of access to necessities can top the list of what is important in making the decision.

Among the Mid-Atlantic states, the dollar goes furthest in West Virginia*. This makes it good candidate to consider for retirement as people from more expensive states like Maryland, New Jersey, Pennsylvania and Virginia would not have to relocate very far. Using machine learning and big data, we can find comparable zip codes in West Virginia with desirable characteristics, highlighting possibilities to those not familiar with the state. This information would be of value to those that offer professional services to retirees, such as real estate agents and financial planners. West Virginia could use this information to market itself as a place to retire, or to identify new areas to promote.

* Numbers compiled from the Bureau of Labor Statistics by 24/7 Wall Street, a branch of USAToday. https://www.usatoday.com/story/money/2019/05/25/us-dollar-how-much-its-worth-value-in-every-state/39501091/

The Data

West Virginia will be divided by zip code, and those areas will be grouped by similarities in the following areas:

- Access to food: supermarkets, restaurants, fast food
- Access to medical care: distance to hospital, number of doctors' offices
- Access to entertainment: distance to nearest university/college, number of entertainment venues
- Number of real estate agencies (measure of housing demand)

There are 703 unique zip codes in West Virginia. Roughly a quarter of those are for Post Office Boxes only, and so do not represent potential retirement areas. These will have to be removed from the list. The complete list can be found at https://www.unitedstateszipcodes.org/wv/#zips-list.

The information for food, medical care, entertainment, and real estate can be gotten through the Foursquare API. Note that in some cases simply the number of venues will be sufficient (number of doctors' offices, number of supermarkets), and in other cases the relevant distance will need to be determined (hospitals, university/college).

A radius of approximately 6 miles (10,000 meters) will be used when searching for the number of venues, as this should be comparable to a 15-20 minute drive on local roads, a reasonable short drive. The venue categories used in this analysis are:

- Supermarkets (52f2ab2ebcbc57f1066b8b46)
- Restaurants, or food (4d4b7105d754a06374d81259)
- Medical centers (4bf58dd8d48988d104941735)
- Spiritual centers (4bf58dd8d48988d131941735)
- Arts and entertainment venues (4d4b7104d754a06370d81259)

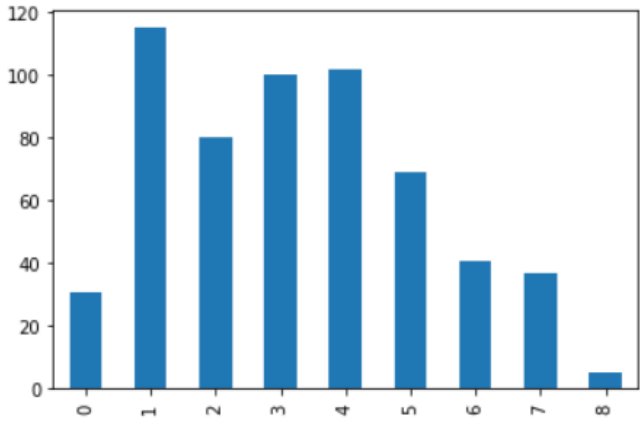
- Outdoors and recreations(4d4b7105d754a06377d81259)
- Real estate offices (5032885091d4c4b30a586d66)
- Bus stops(52f2ab2ebcbc57f1066b8b4f)

For the hospitals and universities, the maximum radius of 100,000 meters (approximately 60 miles) will be used. The program will then calculate the minimum distance to one of these venues as the appropriate variable to be used. The Foursquare venue categories use are:

- Hospital (4bf58dd8d48988d196941735)
- University/College (4d4b7105d754a06372d81259)

Methodology

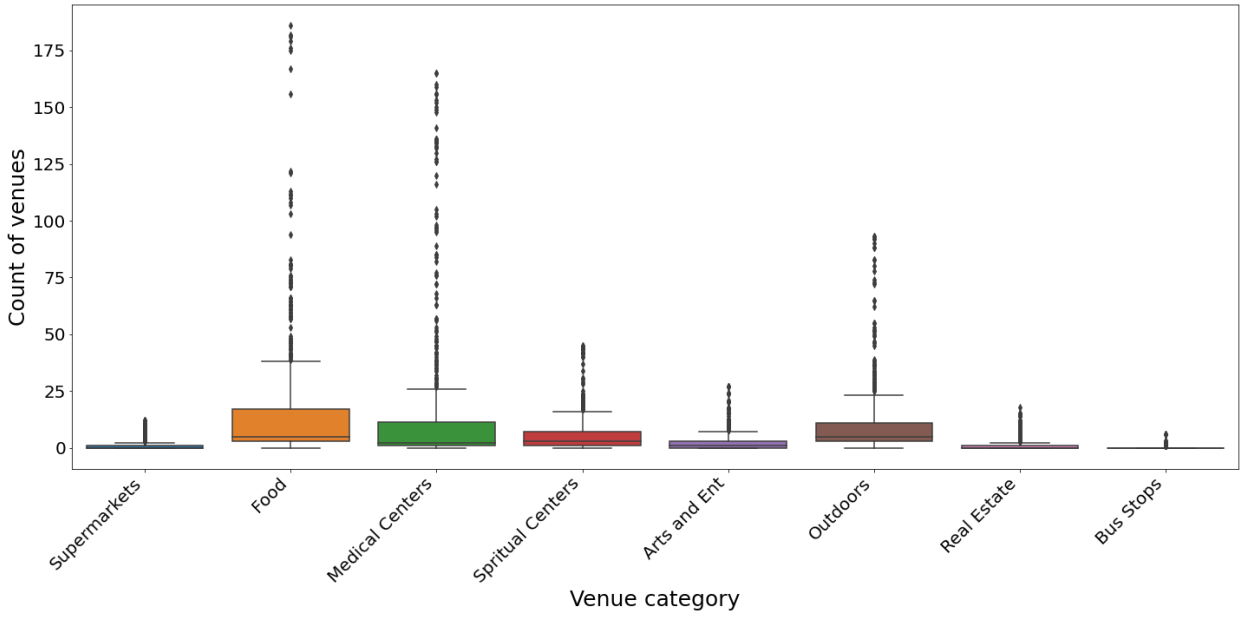
The first concern is that in a state as poor as West Virginia that there might be zip codes that have no services at all. A large number of rows of zeroes will not allow for much sorting to be done. A quick check of this data using a bar graph:



Number of venues categories that are zero

Only 5 ZIP codes have none of the venue types available and all venue categories are available in 31 ZIP codes. The most commonly missed type of venue was bus stops. These categories should be helpful in sorting the ZIP codes.

The next step in the analysis was to visualize the variability in the outcomes of the venue categories using boxplots.

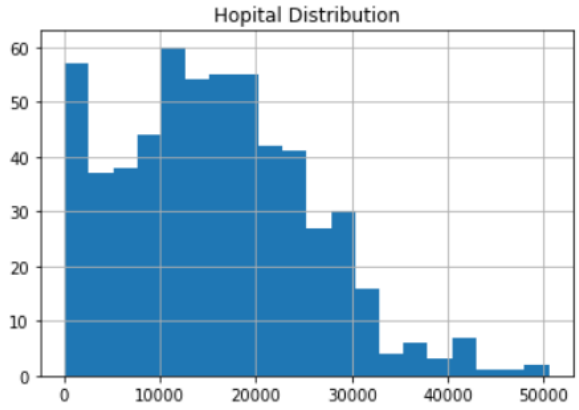


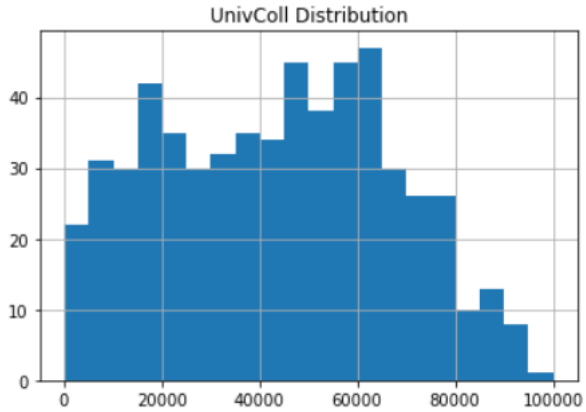
Most of the values reported are low, with the upper fence being 25 or less in every category except food. Looking at the statistics underlying this graph can give us some insights.

	Supermarkets	Food	Medical Centers	Spritual Centers	Arts and Ent	Outdoors	Real Estate	Bus Stops
count	580.000000	580.000000	580.000000	580.000000	580.000000	580.000000	580.000000	580.000000
mean	1.120690	16.537931	16.987931	5.817241	2.644828	10.948276	1.312069	0.131034
std	2.192605	28.652590	35.165508	8.670183	4.665094	15.850361	2.828585	0.630324
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	2.750000	1.000000	1.000000	0.000000	3.000000	0.000000	0.000000
50%	0.000000	5.000000	2.000000	3.000000	1.000000	5.000000	0.000000	0.000000
75%	1.000000	17.000000	11.250000	7.000000	3.000000	11.000000	1.000000	0.000000
max	12.000000	186.000000	165.000000	45.000000	27.000000	93.000000	18.000000	6.000000

Note that over half of the ZIPcodes don't have any supermarkets, arts and entertainment venues, or real estate agents nearby. Three-quarters of the ZIPcodes have no bus stops, indicating that mass transit is not widely available. This may make these areas undesirable for retirees. The 75th percentile indicates that about a quarter of the zip codes are more metropolitan, having plenty of food places, medical centers and spiritual centers to choose from. These would be more likely areas of interest for retirees.

The distances to hospitals and universities or colleges follow a very different distribution.





The distances to the nearest hospital show a distribution similar to a truncated normal distribution. (Distances cannot go below zero, so that part of the distribution is compressed at the low end. The distances to the nearest university or college show closer to a uniform distribution over the distances between 0 and 80,000 meters with a few ZIPcodes as far away as 100,000 meters from any venue of higher education.

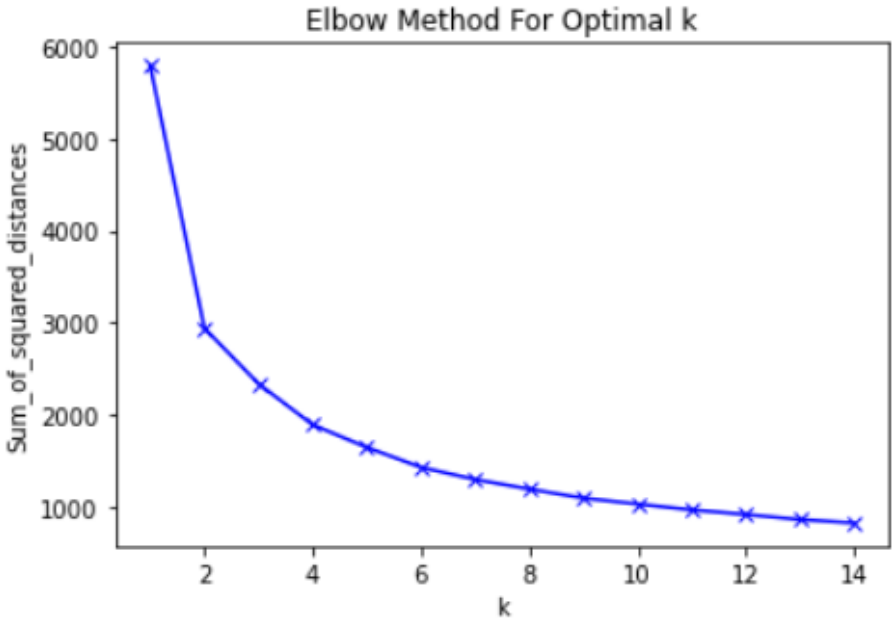
Correlation matrix

The presence of these types of venues seem to be highly correlated. ZIPcodes that have more of one type of venue are likely to have more of the other venues as well, as is shown by the high positive correlations between many of the venue categories. As expected, distances to Hospital and Universities or Colleges is negatively correlated with the number of other venues. ZIPcodes that have many places to eat, recreate, etc., are closer to Hospitals and Universities and colleges.

	Super-markets	Food	Med Centers	Spiritual Centers	Arts and Ent	Out-doors	Real Estate	Bus Stops	Hosp	Univ/ Coll
Supermarkets	1.000									
Food	0.918	1.000								
Medical Centers	0.911	0.877	1.000							
Spiritual Centers	0.848	0.903	0.850	1.000						
Arts and Ent	0.780	0.838	0.769	0.785	1.000					
Outdoors	0.841	0.904	0.829	0.870	0.844	1.000				
Real Estate	0.827	0.875	0.802	0.798	0.767	0.829	1.000			
Bus Stops	0.522	0.425	0.493	0.468	0.547	0.474	0.407	1.000		
Hospital	-0.533	-0.533	-0.522	-0.543	-0.500	-0.484	-0.472	-0.230	1.000	
UnivColl	-0.315	-0.321	-0.296	-0.303	-0.190	-0.257	-0.254	-0.007	0.342	1.000

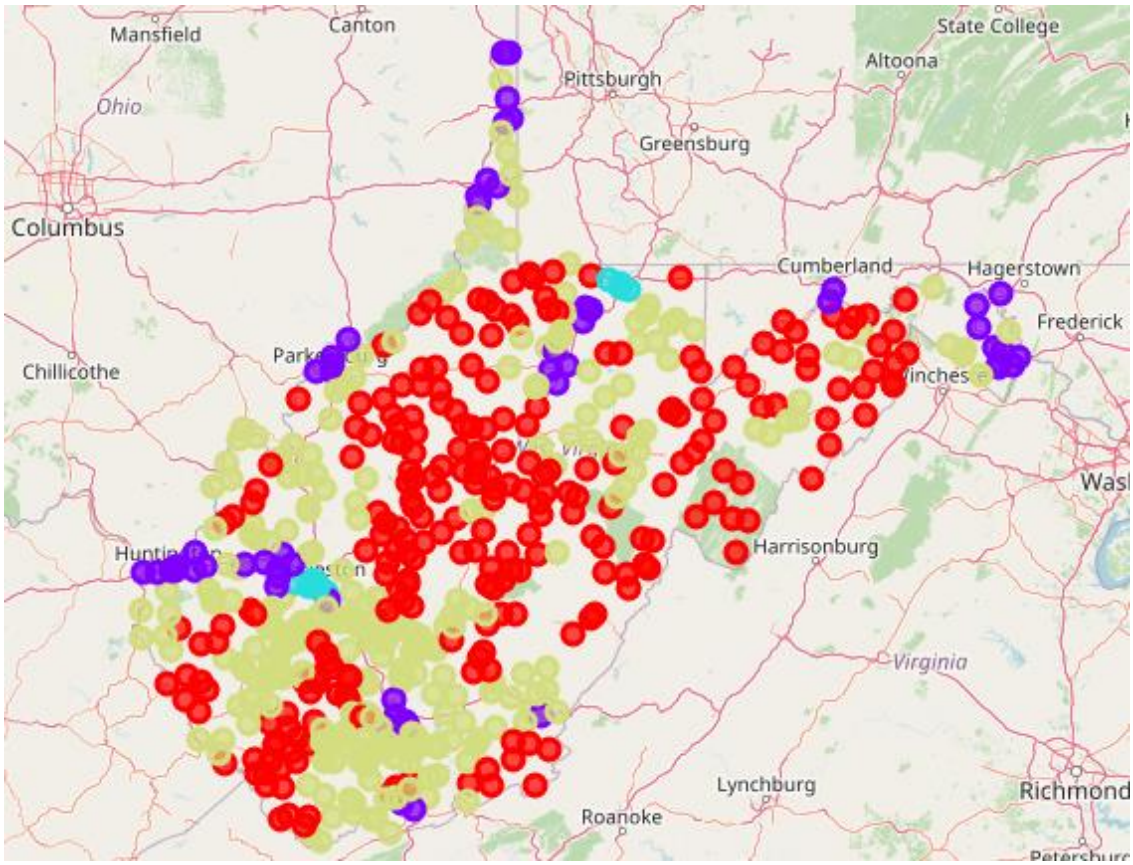
The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. This is an appropriate method when there are not many groups. This paper is only interested in finding areas that retirees might be interested in, not in finding fine gradations of differences between areas, so K-means should be appropriate for our needs. The main disadvantage of K-means is determining the value of K to use in a particular problem.

We will run the K-means algorithm multiple times, starting with K = 1 up to K = 15, and calculate the sum of squared distances from the centroids in each solution. This sum of squares will always decrease with an increase of K, but graphing the results we can see where these improvements level off, giving up the most likely value of K.



We can see that the curve flattens at K = 2 and again at K = 4. This is the value we will use to sort the ZIPcodes into groups. There is a slight flattening at K = 6, but using 6 clusters did not give a good fit, as the fifth and sixth cluster only had one or two points in them.

Results



The first cluster (red dots on the map) – **The Isolated/Rural ZIPcodes**

	Supermarkets	Food	Medical Centers	Spritual Centers	Arts and Ent	Outdoors	Real Estate	Bus Stops	Hopital	UnivColl
count	248.000000	248.000000	248.000000	248.000000	248.000000	248.000000	248.000000	248.000000	248.000000	248.000000
mean	0.141129	3.625000	1.850806	1.661290	0.637097	4.641129	0.185484	0.012097	23691.971774	60124.354839
std	0.371344	3.868342	3.529792	2.410796	1.222824	4.738959	0.560031	0.109539	8039.585298	15990.381749
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	479.000000	24688.000000
25%	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	18562.250000	48583.500000
50%	0.000000	3.000000	1.000000	1.000000	0.000000	4.000000	0.000000	0.000000	23757.500000	60219.500000
75%	0.000000	5.000000	2.000000	2.000000	1.000000	6.000000	0.000000	0.000000	28244.750000	72266.250000
max	2.000000	24.000000	38.000000	19.000000	10.000000	31.000000	4.000000	1.000000	50510.000000	94705.000000

These are the ZIPcodes that have the fewest venues of any type nearby:

- 43% of ZIPcodes, a very common result.
- 75% of them have no supermarkets within 10,000 meters.
- They are furthest on average from hospitals and universities or colleges.
- Almost no Arts and Entertainment venues.
- There is an average of only 1-2 medical centers and spiritual centers.
- Suitable only for retirees that value isolation more than access to basic needs.

The second cluster (yellow dots on the map) – **The Outer Suburban ZIPcodes**

	Supermarkets	Food	Medical Centers	Spritual Centers	Arts and Ent	Outdoors	Real Estate	Bus Stops	Hopital	UnivColl
count	261.000000	261.000000	261.000000	261.000000	261.000000	261.000000	261.000000	261.000000	261.000000	261.000000
mean	0.747126	11.739464	9.065134	4.827586	2.206897	8.068966	0.720307	0.057471	11586.636015	31163.609195
std	0.951336	10.634900	12.284746	4.154350	2.776846	7.091913	1.203492	0.362353	5750.319625	17312.505496
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	320.000000	861.000000
25%	0.000000	3.000000	1.000000	2.000000	0.000000	4.000000	0.000000	0.000000	8186.000000	17604.000000
50%	0.000000	7.000000	5.000000	4.000000	1.000000	6.000000	0.000000	0.000000	11634.000000	27060.000000
75%	1.000000	19.000000	12.000000	7.000000	4.000000	10.000000	1.000000	0.000000	15868.000000	42971.000000
max	4.000000	45.000000	77.000000	19.000000	16.000000	39.000000	7.000000	3.000000	24453.000000	80079.000000

These ZIPcodes also have little access to many amenities, but do have more variety in the offerings:

- The largest cluster – the most common outcome.
- Half have no supermarket nearby, but have a median of 7 eateries within 10,000 meters.
- The mean distance to a hospital is less than half that of the first cluster.
- Are roughly the same distance to a university or college as the last two clusters.
- Still very few Arts and Entertainment venues.
- Have a median of about 5 medical and spiritual centers.
- Good locations for retirees that like less dense areas, but are not too far from necessities and a variety of entertainment.

The third cluster (purple dots on the map) – **Inner Suburban/Population Center ZIPcodes**

	Supermarkets	Food	Medical Centers	Spritual Centers	Arts and Ent	Outdoors	Real Estate	Bus Stops	Hopital	UnivColl
count	58.000000	58.000000	58.000000	58.000000	58.000000	58.000000	58.000000	58.000000	58.000000	58.000000
mean	4.931034	64.637931	89.982759	19.913793	8.362069	34.241379	6.241379	0.431034	4181.155172	33150.931034
std	2.183367	20.825674	40.460720	8.698224	4.478793	13.981580	3.315621	0.818901	2835.180354	29056.526668
min	1.000000	38.000000	32.000000	5.000000	1.000000	15.000000	1.000000	0.000000	90.000000	141.000000
25%	3.250000	47.250000	52.250000	15.000000	4.250000	22.250000	4.000000	0.000000	1598.000000	8090.250000
50%	5.000000	61.500000	85.000000	18.500000	8.000000	31.500000	6.000000	0.000000	4528.000000	17618.500000
75%	6.750000	73.750000	131.500000	23.000000	11.750000	43.250000	8.000000	1.000000	6237.000000	63839.500000
max	9.000000	122.000000	156.000000	42.000000	20.000000	72.000000	18.000000	3.000000	14163.000000	88170.000000

Much smaller than the first two clusters of ZIPcodes, with a lot more access to venues of all varieties.

- Areas in cluster very close to the larger cities in West Virginia.
- Average of 5 grocery stores and over 60 eateries to choose from within a 20 minute drive.
- At least 32 medical centers within a 20 minute drive, and less than 3 miles from a hospital on average.
- Still not many bus stops, so retirees would still need to plan on driving.
- Average of 8 Arts and Entertainment venues, so more opportunities for going out.
- Not much closer to university or college than the second cluster.
- Three times as many spiritual centers than the second cluster, allowing for more diversity.
- More possible ZIPcodes for retirees here than in the Urban cluster 4.

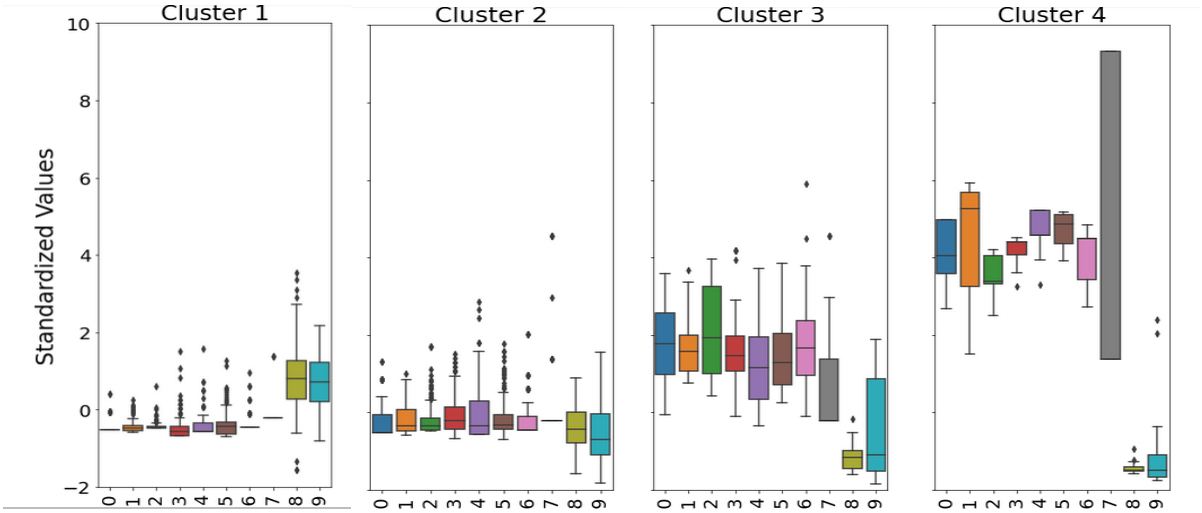
The fourth cluster (light blue dots on the map) – **Walkable Urban ZIPcodes**

	Supermarkets	Food	Medical Centers	Spritual Centers	Arts and Ent	Outdoors	Real Estate	Bus Stops	Hopital	UnivColl
count	13.000000	13.000000	13.000000	13.000000	13.000000	13.000000	13.000000	13.000000	13.000000	13.000000
mean	10.307692	144.615385	139.153846	42.076923	24.230769	85.153846	12.692308	2.538462	1944.000000	23318.384615
std	1.548366	43.509651	18.497055	3.328201	2.586949	7.162939	2.097006	2.401922	1743.658415	33328.978934
min	7.000000	60.000000	105.000000	34.000000	18.000000	73.000000	9.000000	1.000000	382.000000	2356.000000
25%	9.000000	110.000000	134.000000	41.000000	24.000000	80.000000	11.000000	1.000000	1010.000000	4048.000000
50%	10.000000	167.000000	136.000000	44.000000	24.000000	88.000000	14.000000	1.000000	1281.000000	8371.000000
75%	12.000000	179.000000	159.000000	44.000000	27.000000	92.000000	14.000000	6.000000	2055.000000	18006.000000
max	12.000000	186.000000	165.000000	45.000000	27.000000	93.000000	15.000000	6.000000	6513.000000	99997.000000

The smallest cluster of ZIPcodes (only 13), with the most venues within reach.

- Areas within the cluster are where the largest cities are.
- Has the most bus stops, so may appeal to retirees that don't want to drive.
- Less than a mile to the nearest hospital, closest of all clusters to a university or college.
- Average of 24 Arts and Entertainment venues, many more than any other cluster.
- The least number of supermarkets within a 20 minute drive is 7, the average is 10.
- Well over 100 eateries in most ZIPcodes.
- Would appeal to retirees that want diverse venues within near reach, and many venues reachable by walking.

Some of the differences in clusters can be more easily visualized by looking at these boxplots. It is easy to see the increases in the number of venues available as we move from one cluster to another. What is not easy to see here is the difference in cluster sizes. Many more ZIPcodes (88%) fall in clusters 1 and 2 than in 3 or 4.



Discussion

The four clusters identified here do correspond to different residential environments that would appeal to different retirees. The use of the number of easily available venues can give a real taste of what is available in each ZIPcode, allowing retirees to focus their search on those ZIPcodes with desirable attributes.

There are improvements that can be made to this project. It was originally intended to include socio-demographic information as well as venues to allow for further differentiation between ZIPcodes. Such variables as housing prices, population density and real estate taxes are expected to have a great impact on the desirability of an area. Unfortunately, most of this information was only available by county, and it was decided that including information in that form would too heavily favor creating a cluster for each county, which would add no real information to the retiree.

The other change that should be made is that instead of including all outdoor venues, only local/state/national parks should have been used, as these are mainly low level of activity venues suitable for retirees, as opposed to a venue like rock climbing!

Conclusion

Adding Foursquare data to other information about ZIPcodes can benefit retirees trying to decide where to retire and should influence the location of those service venues that cater to that demographic.