Determining the Optimal Location for Retirement

IBM Applied Data Science Capstone Project

Isaac Shvartsman

February 3, 2021

Abstract

Whether you're young or old, the thought of retirement has most likely come across your mind at some point in your life. However, as many people have come to realize, this task is not as easy as one might first anticipate. Whether you're indecisive about a location, or your wallet just isn't fit to make the move, there's no denying it is extremely expensive to retire, and the amount of money that people have been putting aside has been drastically increasing over the years. However, as this report will try to address, the location you decide to retire can also be a have a significant financial impact on your decision. So, what is the best way to pick a place for retirement?

I. Introduction

In this report, I attempt to list five cities in the United States that are the most economically advantageous when considering a location for retirement. To accomplish this task, it is necessary to identify key features with financial implications that distinguish cities from one another, as well as various characteristics that would be beneficial for retirees. By selecting retirement locations based on these criteria, such as state income tax rates in addition to available health care, I was able to refine an initial list of over 300 uncategorized cities into one that shows five that are the least financially impactful. The remainder of this paper lists the features necessary for this analysis, explains the significance they hold for an individual who is looking to retire, and once the data is structured, it becomes possible to conduct exploratory analysis on these locations. Of the five cities chosen through the initial criteria, further research is conducted to gather information on the medical amenities that are available through the use of the foursquare API.

II. Data Section

In order to determine the five most optimal locations for retirement, there are several key characteristics that affect an individual's decision-making when picking where they will spend the remaining years of their life living in leisure. We need to look at variables that are financially impactful, as well as variables that are typically screened through when deciding retirement locations, such as population density and available health facilities per capita.

a. Income Tax

There are several important pieces of variables that can be financially impactful, depending on the location that is selected for retirement. This report will strictly be focusing on state income tax rate, where each state will take out differing amounts of money depending on which state you decide to live in, so it is important to filter for states that have the lowest income tax rate. This information was readily available on Wikipedia.

b. Population size

Population size is a necessary variable in this analysis because I needed to calculate the quantity of health facilities per 100,000 residents, a factor which is highly significant when determining a location to retire. I was able to gather this information on Wikipedia, where the page emphasizes that the list only contains cities with populations no less than 100,000 individuals. This fact is favorable for retirees as regions that have miniscule populations will most likely lack the necessary services for a comfortable retirement.

c. Population Density

i. Population density is important because
ii. It is significant because
iii. Wikipedia

d. Latitude / Longitude Coordinates

It was necessary to compile the latitude and longitude coordinates for all the cities that were in our initial data collection in order to properly use the Foursquare API to collect information relative to various health services that are available in each respective location. This information was available on the same Wikipedia page where I collected the population size and population density of each city.

e. Average House Prices

I collected average house prices in each location so as to determine the respective costs of living using a csv dataset offered by Zillow. The company offers data on housing availability and pricing, in addition to allowing for a compilation of their data sets into csv file format. This allowed me to determine the financial impact of living in each city that I've filtered for retirement.

f. Health Amenities in the area

Health amenities is an important variable to consider because "the growing prevalence of seniors is changing the American economy, increasing the need for medical professionals who can help people stay healthy as they age. Americans who can afford to retire comfortably look for communities that can meet their changing lifestyles and needs, such as quality medical care." (Suneson & Stebbins) To gather this information, I utilized the latitude and longitude coordinates of

the cities that were filtered and used the Foursquare API to gather lists of health facilities in each area. This list had to be refined because it also included either specialized medical facilities, or simply offices that do not provide medical attention.

III. Methodology section

a. State Income Tax

The state income tax data accessible on Wikipedia contains all 50 states and their income tax rates at each respective bracket, which is explicitly defined by each state. The dataset is also split into two columns, where the first represents the state income tax rate for those who file as individuals, and the second for those who file taxes jointly due to marriage. To filter for states with the lowest income tax rate, we restrict our search by keeping in the lowest income tax bracket for each state. After I filtered out each tax bracket that isn't the lowest, I sorted the states by their income tax rates ascending from lowest to highest. From this I was able to get a list of 5 states with the lowest income tax rate, being 0% in Florida, Alaska, Washington, Nevada, and Texas. Using these five states as the basis for my data analysis, I am able to continue in collecting additional significant variables.

	State	Single Filer Rates > Brackets	Married Filing Jointly Rates > Brackets
0	Ala.	2.00%\t>\t$0	2.00%\t>\t$0
1	NaN	4.00%\t>\t$500	4.00%\t>\t$1,000
2	NaN	5.00%\t>\t$3,000	5.00%\t>\t$6,000
3	Alaska	none	none
4	Ariz.	2.59%\t>\t$0	2.59%\t>\t$0

State	[S] Income Tax Rate (%)	[S] Tax Bracket ($)	[M] Income Tax Rate (%)	[M] Tax Bracket ($)
Fla.	0.0	0	0	0
Alaska	0.0	0	0	0
Wash.	0.0	0	0	0
Nev.	0.0	0	0	0
Tex.	0.0	0	0	0

b. Population Size, Population Density, Latitude / Longitude Coordinates

I was able to find a page on Wikipedia that had several important variables relative to my search for the most optimal locations for retirement. I was able to web-scrape a dataset that consists of population size, population density, locational coordinates, as well as other features that were not necessarily required for my analysis, however, the data was very messy and required significant cleaning. Below is a screenshot of the data before I had begun any cleaning of any sort.

	2019rank	City	State[c]	2019estimate	2010Census	Change	2016 land area	2016 land area.1	2016 population density	2016 population density.1	Location
0	1	New York City[d]	New York	8336817	8175133	+1.98%	301.5 sq mi	780.9 km2	28,317/sq mi	10,933/km2	.mw-parser-output .geo-default,.mw-parser-outp...
1	2	Los Angeles	California	3979576	3792621	+4.93%	468.7 sq mi	1,213.9 km2	8,484/sq mi	3,276/km2	34°01′10″N 118°24′39″W / 34.0194°N 118.4108°W
2	3	Chicago	Illinois	2693976	2695598	−0.06%	227.3 sq mi	588.7 km2	11,900/sq mi	4,600/km2	41°50′15″N 87°40′54″W / 41.8376°N 87.6818°W
3	4	Houston[3]	Texas	2320268	2100263	+10.48%	637.5 sq mi	1,651.1 km2	3,613/sq mi	1,395/km2	29°47′12″N 95°23′27″W / 29.7866°N 95.3909°W
4	5	Phoenix	Arizona	1680992	1445632	+16.28%	517.6 sq mi	1,340.6 km2	3,120/sq mi	1,200/km2	33°34′20″N 112°05′24″W / 33.5722°N 112.0901°W

i. Population Size

The population size was already structured in a clean way so there wasn't anything really needed with respect to data-cleaning, however, these values are important for when calculating the amount of health care facilities per capita at each location. So, I created a separate list from which I will be able to compile a final dataframe of when doing my analysis.

ii. Population Density

The population density column needed a little bit of work, where I had to split the value from the '/sq mi' and also get rid of the value separating commas so that I may convert the values into a usable integer format. Python does not allow for integers to contain string formatted characters, so there was definitely some cleaning that had to be done there. After cleaning, we sorted the dataframe by its population density from lowest to highest.

iii. Latitude / Longitude Coordinates

The latitude and longitude coordinates needed a significant amount of work in order to be able to use them in the Foursquare API. It was necessary to convert the geospatial coordinates from north and south degrees into negative and positive float values. I split the column into two separate columns, one for latitude values and the other for longitude values. Then I had to eliminate any string characters from the cells similar to how I did with the population density. Then, because the United States is located in the western hemisphere, I had to convert all the longitude values negative so that when the coordinates are inputted into the Foursquare API, we are given the proper location. Below is a screenshot of the dataframe after cleaning the features and subsequently concatenating them.

	City	State	Population	Population Density	Latitude	Longitude
0	Anchorage	Alaska	288000	175	61.1743	-149.2843
1	Abilene	Texas	123420	1146	32.4545	-99.7381
2	Jacksonville	Florida	911507	1178	30.3369	-81.6616
3	Reno	Nevada	255601	2286	39.5491	-119.8499
4	Spokane Valley	Washington	101060	2681	47.6733	-117.2394

c. Zillow Data (Average House Prices)

I imported housing data from Zillow into a csv file that shows the average house price of various cities in the United States. Because this is an analysis to determine whether a city is an adequate place for retirement, I limited the housing data to family style housing units only, under the assumption that retirees will likely have family living either in the house or maybe in the area, as elderly individuals may require assistance from an additional helping hand. When reading the csv file into my notebook, I only needed the most recent average housing prices and not the historical data that was available as well. I dropped any unneeded columns and filtered based on the cities that we filtered using population density in the states with the lowest income tax rate. This information is important when comparing costs of living between each location.

	RegionID	SizeRank	RegionName	RegionType	StateName	State	Metro	CountyName	1996-01-31	1996-02-29	...	2020-02-29	2020-03-31	2020-04-30	2020-05-31
0	6181	0	New York	City	NY	NY	New York-Newark-Jersey City	Queens County	207583.0	207009.0	...	665691.0	665820.0	666836.0	667437.0
1	12447	1	Los Angeles	City	CA	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	190758.0	190802.0	...	767433.0	774529.0	780670.0	783079.0
2	39051	2	Houston	City	TX	TX	Houston-The Woodlands-Sugar Land	Harris County	95357.0	95457.0	...	194			
3	17426	3	Chicago	City	IL	IL	Chicago-Naperville-Elgin	Cook County	136370.0	136234.0	...	244			
4	6915	4	San Antonio	City	TX	TX	San Antonio-New Braunfels	Bexar County	96098.0	96063.0	...	190			

City	Typical_Home_Value
Abilene	145682.0
Anchorage	375448.0
Jacksonville	214595.0
Reno	442817.0
Spokane Valley	309564.0

d. Health-Care per Capita (Foursquare API)

By inputting the geospatial coordinates of each city in my filtered dataframe into the Foursquare API, I was able to gather a list of medical related venues at each specific location in a given radius. I wrote a function that extracted the category type from each venue, however, this list included additional medical facilities that are not necessarily relevant to retirement, such as pet hospitals, food venues located within hospitals, as well as children's hospitals. So, I filtered any rows that did not pertain to hospitals, doctors' offices, emergency rooms, or medical centers, and got the length of the list so as to determine the quantity of health facilities in the area. After doing this for each location on my retirement list, I divided the quantity of health care amenities by the total population of each given city, which allows me to calculate the number of health-care facilities per-capita in each location. Below are the dataframes before and after restructuring/cleaning.

categories	name	id
Hospital	Renown Health - Rehabilitation Hospital	4d190cbb81cea35dce20f2ec
Hospital	West Hills Hospital	4e692f96ae6037916d8d317b
NaN	St. Mary's Hospital, Reno NV	53f9502c498e9dac1e32f4cc
Hospital	PAM Specialty Hospital of Sparks	5dc5eb9eb76cb700084bc537
Hospital	Saint Mary's Regional Medical Center	4b663767f964a52062192be3

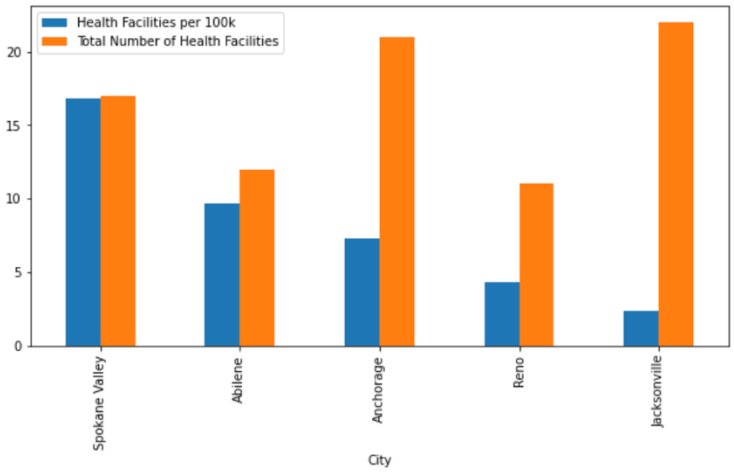
	Total Number of Health Facilities
0	21
1	12
2	22
3	11
4	17

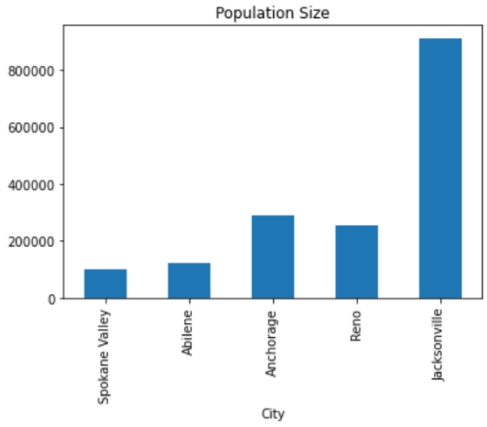
	Health Facilities per 100k
0	7.291667
1	9.722897
2	2.413585
3	4.303583
4	16.821690

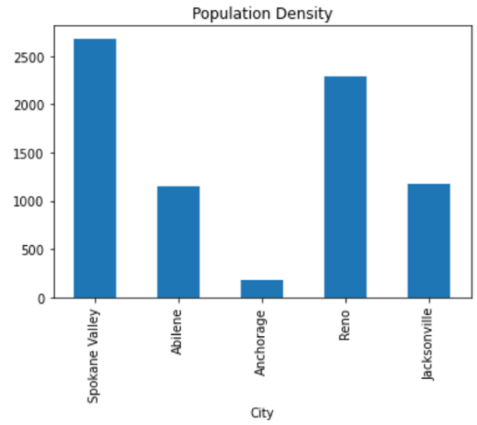
IV. Result section

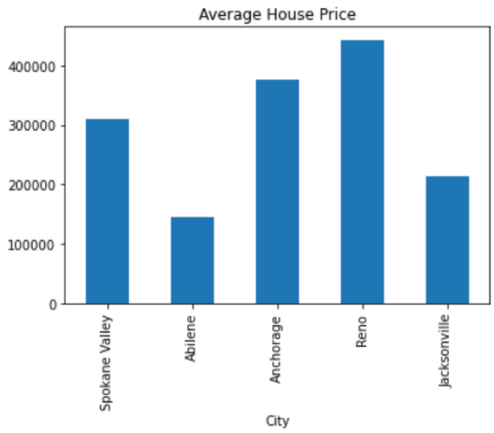
After filtering through cities in the United States based on the features selected above, the resulting list of optimal cities for retirement consists of five locations in the five states with the lowest income tax rate. Each city has relatively low population densities, as well as roughly the same number of total health facilities. The difference in locations occurs when observing the population size relative the quantity of facilities in each area. Spokane Valley, Washington, is going to have the most health facilities per capita on this list with a moderately low population, therefore it is going to take the number one spot for the best retirement location based on the criteria stated above. Although its average housing price is not the lowest, it provides the best opportunities to get medical care, which is extremely important for retirees. Abilene, Texas, takes second since its healthcare per capita is almost half as much as the number one spot, although the average house price is significantly lower. Anchorage, Alaska, has the lowest population density with moderately high housing prices, with a health care per capita of only 7.29 it is going to be the third retirement location on this list. Reno, Nevada, is the fourth city on this list, where, it has the most expensive home values and significantly lower health facilities per capita than the other cities on this list. Jacksonville, Florida, is the number 5 spot in this list for optimal retirement locations in the United States. It offers the lowest number of health amenities relative to its population, seeing as there are almost a million inhabitant in the city. Below is the final dataset that was used to make this analysis, as well as a couple bar graphs representing the feature differences between cities.

City	State	Total Number of Health Facilities	Population	Population Density	Health Facilities per 100k	Typical_Home_Value
Spokane Valley	Washington	17	101060	2681	16.821690	309564.0
Abilene	Texas	12	123420	1146	9.722897	145682.0
Anchorage	Alaska	21	288000	175	7.291667	375448.0
Reno	Nevada	11	255601	2286	4.303583	442817.0
Jacksonville	Florida	22	911507	1178	2.413585	214595.0









V. Discussion Section

While going through my analysis of cities in the United States to determine optimal retirement locations, it became apparent to me that there are several factors that need to be accounted for in order to create a more accurate list of cities. Inclusion of these various features would have been out of the scope of this report; thus, I chose to omit them from my analysis.

a. Shortcomings

I realize that there are many different variables to take into account when determining a retirement location, so much so that the scope of this project would be too large. I did not account for external geographical factors such as weather-patterns or susceptibility to natural disasters. I also did not fully encapsulate the total features that would be financially impactful towards retirees, such as property taxes and other cost-of-living factors like the price of goods. This possibly hindered my analysis as there may not have been enough features to make a reasonable decision, but, when cross-referencing the locations on my list with well-known retirement blogs, each location came up on other lists as prominent cities for retirement.

b. Future Work

In any future analysis of cities in the United States that I perform, there will surely be an inclusion of the various features that I had listed in the shortcomings section of my report. I am hoping to address these factors in the next report in such a way where the following iteration will be more precise and provide better locations for retirement than in this report.

VI. Conclusion

Using geospatial data in addition to other distinguishing characteristics, I was able to create a list of locations that are optimal for retirement. Given that my feature selection was not as inclusive as it could have been, there could definitely be improvements in the model. However, using these features which I have deemed most important relative to the decision-making of retirees, I was able to successfully put together a list of five cities that are most advantageous for individuals in retirement.

VII. References

https://medium.com/@yrnigam/how-to-write-a-data-science-report-181bd49d8f4d

https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

https://en.wikipedia.org/wiki/State_income_tax/

https://www.zillow.com/research/data/

https://developer.foursquare.com/docs/

https://www.kiplinger.com/retirement/happy-retirement/601003/find-a-great-place-to-retire

https://career-resource-center.udacity.com/portfolio/data-science-reports

https://www.usatoday.com/story/money/2019/04/15/for-retirement-30-best-cities-for-older-americans/39331431/