IBM Applied Data Science Specialization

Capstone Project

Part I b

Data

I will be using 3 data sets to analyze the relative attractiveness of the towns under consideration.

1. The Foursquare data to look at the number and variety of different venues in each town.
2. Quality of Life data which I will gather primarily from City-data.com, a website that contains a plethora of data on every city and town in the U.S. Additionally I will gather data on relative quality of the local schools from a USnews.com study of American high schools.
3. Statistics on local skiing mountains will be amalgamated from each mountain's own website.

Data from the various websites will be collected and then manually entered into a data frame in my code. Foursquare data will be downloaded directly from Foursquare and initially placed in its own data frame.

Foursquare Data

After downloading the information on the venues for each town the data will be added to single data frame containing all the foursquare data. After taking a look at the ten most frequently occurring venue types in each town I will consolidate the data into five broad categories by general venue type for easier analyses.

The five board categories that the venues will be divided into are:

1. Restaurants
2. Shopping
3. Recreation

4. Entertainment
5. Other

Each category will be grouped into its own data frame, then the towns will be ranked by category according to the number of venues in each town of each type. The towns will then by ranked in each category according to the number of venues of each type after normalizing each category from 0 to 1. The Restaurant and shopping categories will be double weighted giving a total of 7 points possible from the Foursquare venue data. I am giving double weighting to those two categories both because they are the most important categories in which a wide variety of choices is important. Furthermore, a large number of restaurants and shops should indicate a vibrant economy.

Quality of Life Data

The quality of Life data taken from city-data.com will come from five categories.

1. Average January Temperature (I prefer not to be too ridiculously cold!)
2. Air Quality Index
3. Average Sunny Days per Year
4. Cost of Living Index
5. Quality of Schools
6. Crime Rate

All the above statistics except for the school quality data are sourced from city-data.com. the School data comes from USnews.com/education/best-high-schools a study that attempts to rank every high school in America from 1 – 17,245. I assume that the quality of education in the primary schools is going to be roughly analogous to the high school. To calculate a quantitative value for education quality I take the ranking of the local high school according to U.S. News for each town and express that ranking as a percentile of the entire data base. For example, Aspen High School in Aspen Colorado is ranked # 1,430 in the U.S. In my data I express this as 1-(1,430/17,245) = .917.

Like the Foursquare data, the raw numbers in each category are normalized from 0-1. Air Quality Index, Cost of Living Index and Crime rate are indices in which lower numbers indicate a better situation, so I need to invert those statistics before normalizing them. I will double weight the crime data as that, to me, seems the most important category. I will then sum the scores for each town to arrive at a total Quality of Life score for each town. Like the Fourscore venue data, the maximum score for a town from this data set is 7.

Skiing Data

Data on available skiable terrain will be amalgamated from each mountain's website.

To evaluate the quality of the skiing experience at each mountain I will explore 6 variables.

1. Total Skiable Acres - (this includes any ski area within a 45-minute drive (according to Google) – the maximum distance I would be willing to sometimes commute for the sake of variety.)
2. Ski In/Out Acres - The number of skiable acres immediately accessible from the town.
3. Vertical – the largest vertical distance from the top to the bottom of any of the commutable skiing areas.
4. Lifts – the total number of lifts in all commutable ski areas combined.
5. Average Annual Snowfall – the more the better for skiing.
6. Lifts per Total Skiable Acres – derived from 1. And 4. Above, should provide a general indication of the length of lift lines.

I will follow the same general methodology as in the first two data sets. I will normalize each category among the different towns. The normalized value of Total Skiable Acres will be given double weight within this group giving a total of a maximum of 7 points for this data set, the same as the previous two groups.

Final Score for Each Town

Finally, the total column from each of the above 3 data frames will be placed in a final data frame and points summed to create a total score for each town to find which town is objectively most appealing based on the selected data and methodology. The maximum possible score will be 3 x 7 = 21, but I don't expect any town to come close to this. I will also then look at median home values (Also from city-data.com) to find which town represents the best real estate value.