The Battle of Neighborhoods

A Capstone Project for IBM Applied Data Science Capstone Course

Zachary Switaj

**Introduction/Business Problem**

I am an analyst for FooBar Cycles, a small chain of indoor cycling studios looking to expand to the beautiful Seacoast New Hampshire area (note: this is a made up scenario). The company’s senior management has tasked me with finding the ideal area to put our new studio. They want to find a location which has a relatively high population compared to surrounding areas, but mostly one with high income earners who are willing to pay for our premium indoor cycling services. This new location will need to be home to fitness minded folk, those who care about getting and staying fit, and not afraid to pay to help get them there!

**Data**

I have chosen to use the Foursquare API and web scraping to tackle this task. Using the Foursquare API I can obtain data about companies, businesses, and points of interest (called *venues* going forward) in the areas of interest, and compare these to economic data scraped from the web to make an informed, data-driven decision about where to put in our new location.

The Foursquare API is used to extract venue data in a .json format (see example below) within a set radius of the location provided. This can be done to get venue data for all surrounding venues up to a set limit, and the category of each venue is pulled from the .json snippet and used to categorize each venue appropriately.



Figure .json snippet showing venue data

Senior management would like the new location to be 30 miles of Rollinsford, NH so towns that fit this criteria were identified using zip-codes.com (see figure 2 below for top 10 towns). As previously mentioned they would also like the new location to be in an area with high income (relative to surrounding areas), and in towns with reasonable populations. This data was taken from webatlas.com for both New Hampshire and Maine (towns in both state are within 30 miles of Rollinsford, NH). Figure 3 below shows the data for the top 10 towns in New Hampshire.

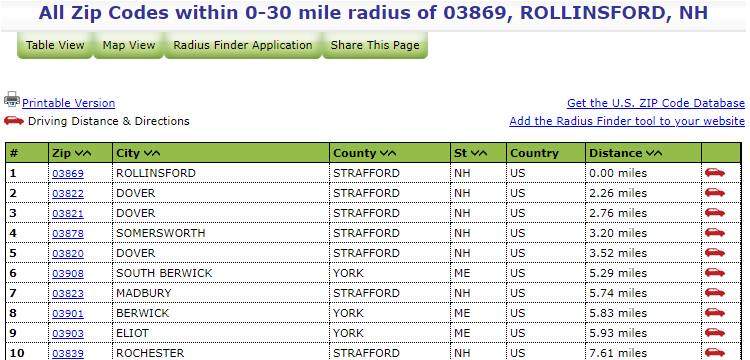


Figure zip-codes.com page showing towns within 30mi of 03869

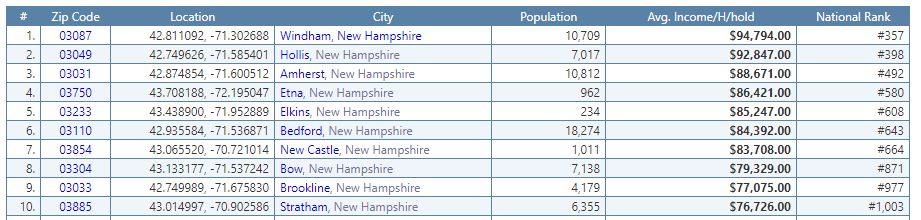


Figure zipatlas.com data showing population and household income

**Methodology**

To find a suitable location for the new FooBar Cycles location given the parameters laid out by senior management the above data was analyzed using a python Jupyter Notebook. The BeautifulSoup python package was used to scrape location, population, and income data for the towns within 30 miles of Rollinsford, NH and this data was cleaned and combined into a pandas DataFrame object. Pandas’ DataFrame object allows for further analysis and represents a very powerful tool for data science (see figure 4 below).

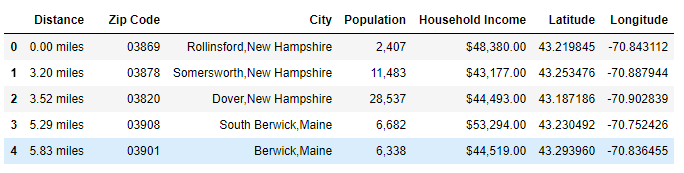


Figure DataFrame object with city, population, income, and location data

Using the folium python package the towns identified above were plotted to get an idea of the range we are looking at for locations, as well as the density of towns.

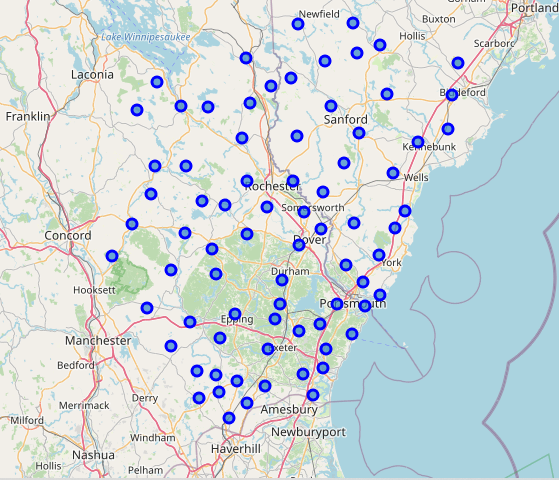


Figure Map of possible cities for new location

The Foursquare API was then used to retrieve venue data for each of the above towns. Given the relatively low density of towns in this area, a ‘radius’ of 5km was used to identify up to the top 500 venues within 5km of the town center. This search resulted in 2520 venues across the 103 towns identified. The category of each venue was extracted and resulted in 250 unique venue categories. One-hot encoding was used to indicate whether or not each of the 250 categories was present in each town, and a group mean was taken in each town to show the proportion of each venue category that is present in a given town.

Following this step each of the categories was placed into a group to reduce the number of variables present in the following classification algorithms. Identified groups which were made into list of categories include: shopping, travel, dining, fitness, entertainment, outdoor, hospitality, and other. The above encoded data was then summed for each town to indicate what proportion of each new category class was present, and this data was then merged with the population and income data for further analysis (see figure 6).

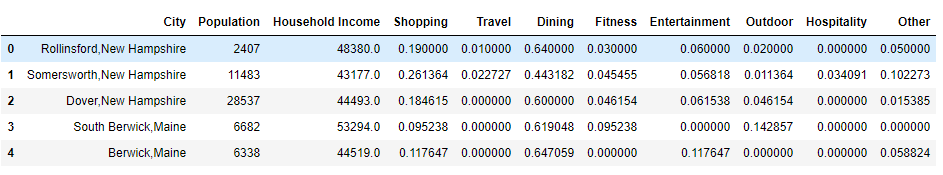


Figure Town data summarized by new category classes

**Results**

Data was preprocessed for classification using a MinMaxScaler function, ensuring each values was scaled from 0-1 so that features such as population and household income would not have any larger effect on classification than the classified category data. A k-means clustering algorithm with k = 4 was run on the data to separate towns into 4 clusters (k = 4 picked after some experimentation by author). Cluster labels were added and this new DataFrame was joined with the location data for a visual map of the clusters by color (see figure 7).

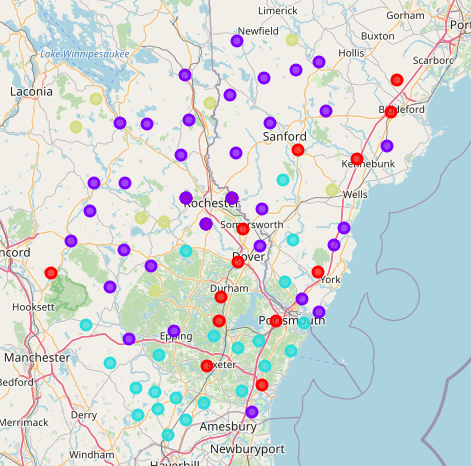


Figure Town clusters shown by color

The clusters were then further analyzed by individual cluster to identify cluster parameters which can be used for our selection of an appropriate location. From the summarized data shown in figure 8 below the below parameters of each cluster were identified:

* Cluster #0: High population, medium income, high density of dining venues, low density of outdoor venues
* Cluster #1: Medium population and income, low density of fitness venues
* Cluster #2: Medium population, high income, high density of shopping and fitness venues, low density of hospitality venues
* Cluster #3: Low population, medium-high income, low density of shopping and dining venues, very high density of outdoor venues

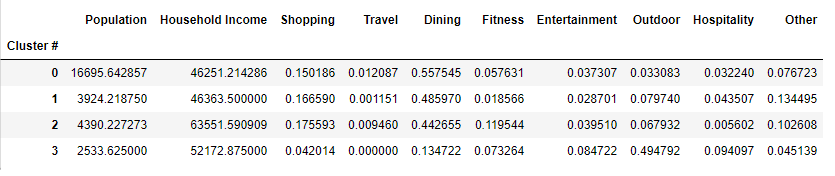


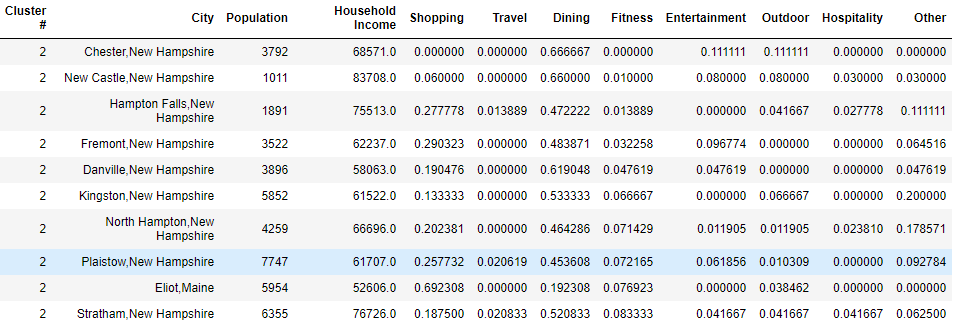
Figure Summarized cluster data for the identified clusters

**Discussion**

The above data shows that cluster #2 is likely the best location for our new FooBar Cycles location. This cluster represents a decent relative population, with the highest average income of the clusters and also the highest proportion of fitness locations. This indicates that venues in the ‘fitness’ category are currently present in this cluster, which suggests that the towns in the cluster are willing to support and are interested in these venues. If these towns were not fond of fitness venues than they would not support these venues and they would go out-of-business or be required to move to a location where they can get support. The high income earners also represent a potential customer for the premium indoor cycling services offered by FooBar Cycles.

However, senior management does not want to pick a town with a lot of competition, i.e. a lot of other fitness venues present. Figure 9 below shows towns with low density of fitness venues. These towns are in the same cluster as other towns where fitness venues have been deemed successful, and are therefore identified as being similar with respect to the features provided.

Looking at the table in figure 9 as well as the map in figure 10 it appears there is a group of towns in cluster #2 which are located at the south-western portion of our towns of interest. Looking at the towns present in this group shows that both Fremont and Danville are included. Figure 9 shows that these two towns have moderate populations, relatively high household income, and a low density of fitness venues. For these reasons, further analysis into these towns should be done to determine if their town ordinances, taxes, etc. would be a good fit for a new FooBar Cycles location.



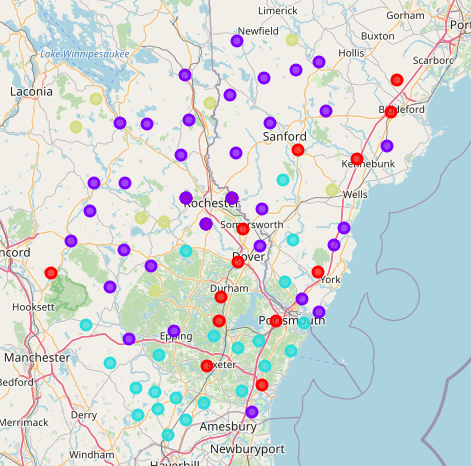


Figure 10 Identified area for new FooBar Cycles Location

Figure Venues in cluster #2 with low density of fitness venues

**Conclusion**

The above analysis identified a group of towns where a new FooBar Cycles location may be successful given the parameters laid out by senior management. Further analysis should be done into the town ordinances and taxes on this type of business to see if any of these towns would be a good fit for our new location. A similar analysis could be done using additional demographic data if deemed important by senior management. Additionally, if other successful FooBar Cycles locations were identified, important parameters of the locations of each could be identified to provide further features to add to our classification model. Given the data presented above and our domain knowledge of the subject, either Danville or Fremont appear to be good locations to add a new FooBar Cycles location.