Towns in Connecticut with Healthy Lifestyle

Isha Sodhi

14th Sep 2020

1. INTRODUCTION

1.1 Description of the problem

There are always ups and downs in the real estate market. However, following the occurrence of the pandemic, a new trend is being observed among the people looking for their new home- 'Healthy Lifestyle and healthy neighborhood'. Home hunters are now looking for a home in a town/city where they can enjoy the nature as well as stay fit by involving themselves in different activities. They want to buy a home in a neighborhood which gives plenty of different options to have an active lifestyle. As social distancing has become the need of the hour, people are preferring more open ground activities, rather than those inside closed rooms.

Real estate agents in Connecticut need the help of data science to present the neighborhoods to their customer which have different healthy activities around, like state parks for hiking, trails, farmers market for fresh food, beaches for recreation.

1.2 Solution

In this project, I'll be using my knowledge of data science to analyze which towns/cities in Connecticut state have more number of venues that the real estate agents are currently interested in, viz. state parks, playgrounds, beaches, farmers markets, yoga studios etc.

To start with, I'll need the town/city data of Connecticut, analyze the data and find clusters of towns/cities in Connecticut having more number of venues of interest. Clustering method will help me achieve the required clusters.

2. DATA

2.1 Source

The data of list of Towns & Cities will be downloaded from Wikipedia: https://en.wikipedia.org/wiki/List_of_towns_in_Connecticut .

2.2 Description

The list on Wikipedia gives the table along with list of towns, population of towns, their respective counties, land area, designation, Council of governments. For this project, many of the columns will be dropped which will not contribute to the solution. Wrangling of data will be done in order to filter out the unnecessary data. The main column that we are focused on is the 'Town', which will be used to obtain the geo-coordinates.

Table 1: Raw data from Wikipedia

	Number	Town	Designation	Dateestablished	Land area(square miles)	Population(in 2010)	Form ofgovernment	County	Council of Governments	Native Americanname
0	1.0	Andover	Town	1848	15.46	3303	Town meeting	Tolland County	Capitol Region	NaN
1	2.0	Ansonia	City	1889	6.03	19249	Mayor-council	New Haven County	Naugatuck Valley	NaN
2	3.0	Ashford	Town	1714	38.79	4100	Town meeting	Windham County	Northeast CT	NaN
3	4.0	Avon	Town	1830	23.12	18098	Council-manager	Hartford County	Capitol Region	NaN
4	5.0	Barkhamsted	Town	1779	36.22	3620	Town meeting	Litchfield County	Northwest Hills	NaN

The unwanted columns were dropped, and a new table was created using the wrangling functions available in python.

Table 2: Cleaned Data

	Town	Designation	Land area(square miles)	Population	Form ofgovernment	County
0	Andover	Town	15.46	3303	Town meeting	Tolland County
1	Ansonia	City	6.03	19249	Mayor-council	New Haven County
2	Ashford	Town	38.79	4100	Town meeting	Windham County
3	Avon	Town	23.12	18098	Council-manager	Hartford County
4	Barkhamsted	Town	36.22	3620	Town meeting	Litchfield County

Coordinates of the towns will be collected using the geocoder package and these coordinates will be used in the foursquare application to find out the venues of interest.

Table 3: Coordinates obtained using geocoder

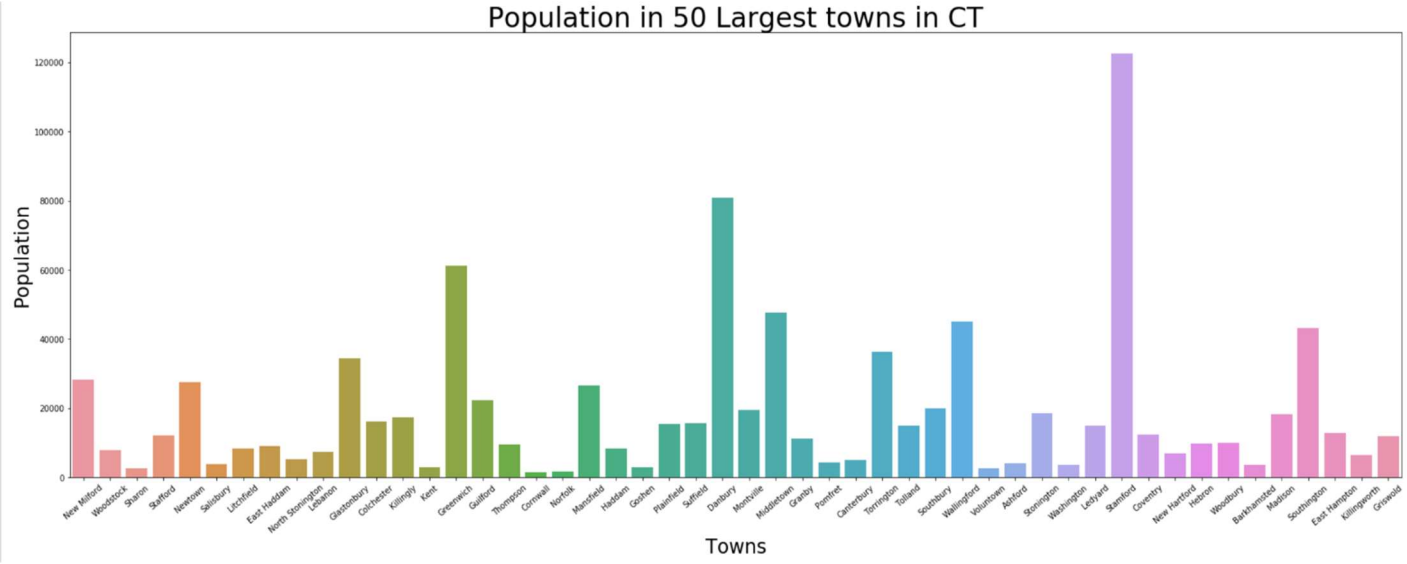
	Latitude	Longitude
0	41.74159	-72.38047
1	41.34317	-73.07875
2	41.86293	-72.18840
3	41.81088	-72.82965
4	41.91212	-72.98829

The two above tables were merged together to get a better idea of which coordinates corresponds to which Towns.

The data obtained from foursquare will provide the information of all different types of venues in all the towns of Connecticut. Our aim is to filter the venues of interest i.e parks, farmer's markets, beaches, yoga studios, gyms. All these venues contribute towards a healthy lifestyle.

3. METHODOLOGY

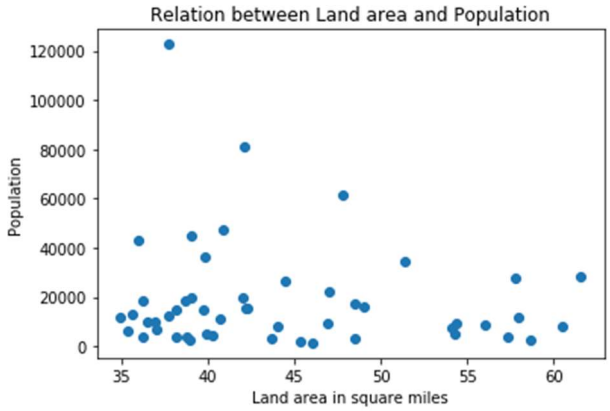
3.1 Analysis of data



With the help of plots, data is analyzed to find relation between different factors that can contribute to our solution. I started with observing the population of different towns in Connecticut and plotted 50 most populous towns. The aim is to find if there is any relation between land area, population and the venues that we are considering.

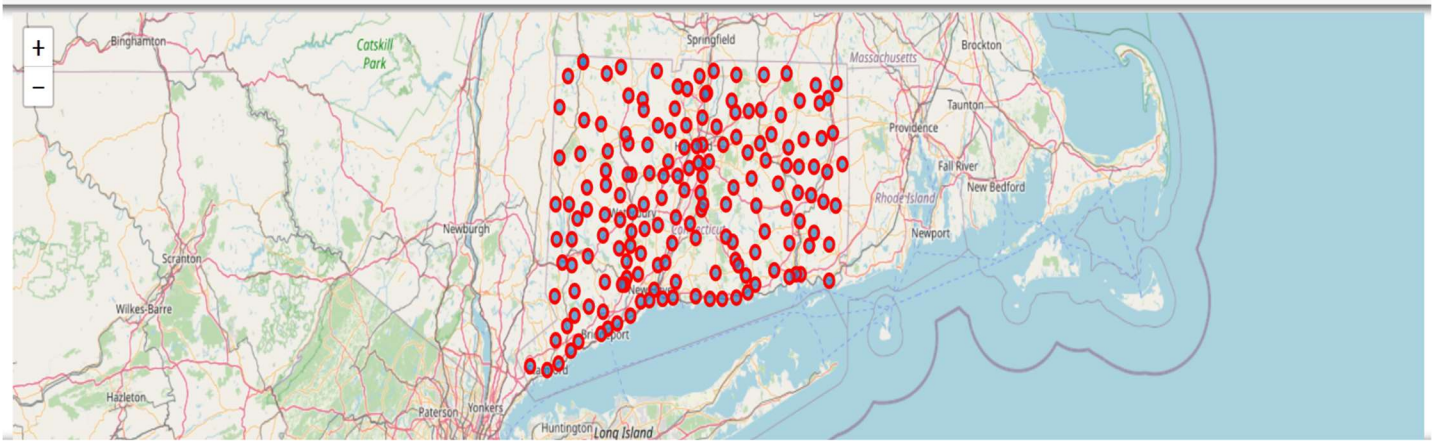
It was observed that Stamford was the most populated town of Connecticut, followed by Danbury, Greenwich and Southington.

Further, I plotted a scatter plot of population verses the land area. One would think that larger land area should correspond to more population; however, no specific relation was observed between these two factors. This would give us an idea that there would be non-residential land in towns with larger land area; the reason could be the town has more state parks, playgrounds, theme parks or it would have more industries, malls etc.



3.2 Visualization of Coordinates

Folium is a powerful python library that makes it easy to visualize the data in the form of an interactive map. In this project I used folium.Map function to get a map of Connecticut with marked coordinates. We can zoom in and out as per our need and when the mouse pointer is hovered over the marked coordinate, it shows the name of the town.



Map of Connecticut with marked Coordinates

3.3 Use of Foursquare

As we are interested in specific type of venues viz. state parks, beaches, gyms, outdoor activity venues we will have to use an API which can provide us with the data of the interested venues in Connecticut. I've taken the advantage of having an account with FourSquare API to get the required data.

Different types of calls can be made on FourSquare but as the numbers of calls per day are limited, I've set the 'Limit' to 100.

There is another variable which is very important to raise a call with FourSquare, to get proper data: Radius. It decides the range in which the API will search for the interested Venues and will provide the venues within that range. Generally, the center point is the downtown of that particular town. Initially, I had set the radius as 500, which did not give me enough data to work on as I'm focused on a limited number of venue categories. With trials I found the appropriate radius to be 5000.

Raw data obtained from FourSquare had a total of 361 unique categories. After scanning the different categories, the following were chosen for this project:

- Athletics & Sports
- Campground
- Cycle Studio
- Dance Studio
- Fair

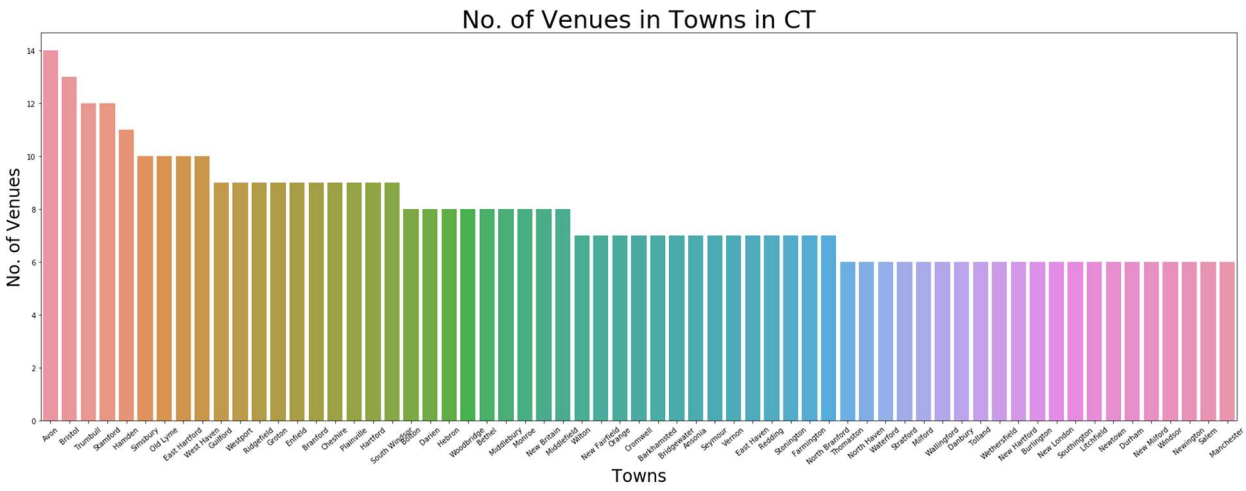
- Farmers Market
- Gym
- Gym/Fitness Center
- Gym Pool
- Gymnastics Gym
- Harbor/Marina
- Lake
- Monument/Landmak
- Paintball field
- Park
- Playground
- River
- State/Provincial Park
- Theme Park Ride / Attraction
- Trail
- Yoga Studio

3.4 Analysis of FourSquare Data

Since the data returned by FourSquare was huge, some filtering was needed. Only the towns which had more than 5 venues of interest were extracted and a table was formed. The top 10 towns with most venues of interest is shown in table below:

Table 3: Coordinates obtained using geocoder

	Town	Population	Latitude	Longitude	Land area(square miles)	Venue Category
0	Avon	18098	41.81088	-72.82965	23.12	14.0
1	Bristol	60477	41.67548	-72.94652	26.51	13.0
2	Trumbull	36018	41.23495	-73.21967	23.29	12.0
3	Stamford	122643	41.05195	-73.54222	37.75	12.0
4	Hamden	60960	41.38423	-72.90161	32.78	11.0
5	Simsbury	23511	41.88046	-72.80000	33.88	10.0
6	Old Lyme	7603	41.32391	-72.33397	23.10	10.0
7	East Hartford	51252	41.76897	-72.64401	18.02	10.0
8	West Haven	55564	41.27228	-72.94998	10.84	10.0
14	Guilford	22375	41.28424	-72.68142	47.05	9.0



3.5 One hot encoding

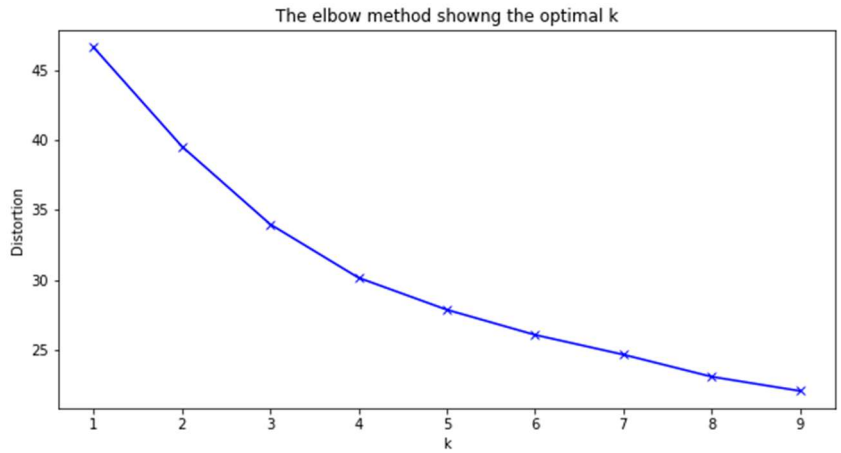
Out[49]:

	Town	Athletics & Sports	Boxing Gym	Campground	College Gym	Cycle Studio	Dance Studio	Fair	Farmers Market	Gym	Gym / Fitness Center	Gym Pool	Gymnastics Gym	Harbor / Marina	Lake	Monument / Landmark	National Park	Outdoor Gym	Paintball Field	Park	Parking	Playground	RV Park	River	Skate Park	S Provi
0	Andover	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.25	0.000000	0.000000	0.0	0.0	0.0	0.250000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.00
1	Ansonia	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.142857	0.142857	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.142857	0.0	0.0	0.0	0.0	0.0	0.14
2	Ashford	0.000000	0.0	1.0	0.0	0.0	0.0	0.0	0.00	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.00
3	Avon	0.071429	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.071429	0.142857	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.214286	0.0	0.0	0.0	0.0	0.0	0.07
4	Barkhamsted	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.000000	0.000000	0.0	0.0	0.0	0.285714	0.0	0.0	0.0	0.0	0.285714	0.0	0.0	0.0	0.0	0.0	0.00

3.6 Clustering

Finding suitable value of 'k': Elbow Method

Clustering method requires to define the value of ‘k’ which decides how many clusters will be formed. For selecting a proper ‘k’ Elbow method is used. The value that I found to be appropriate was 5.

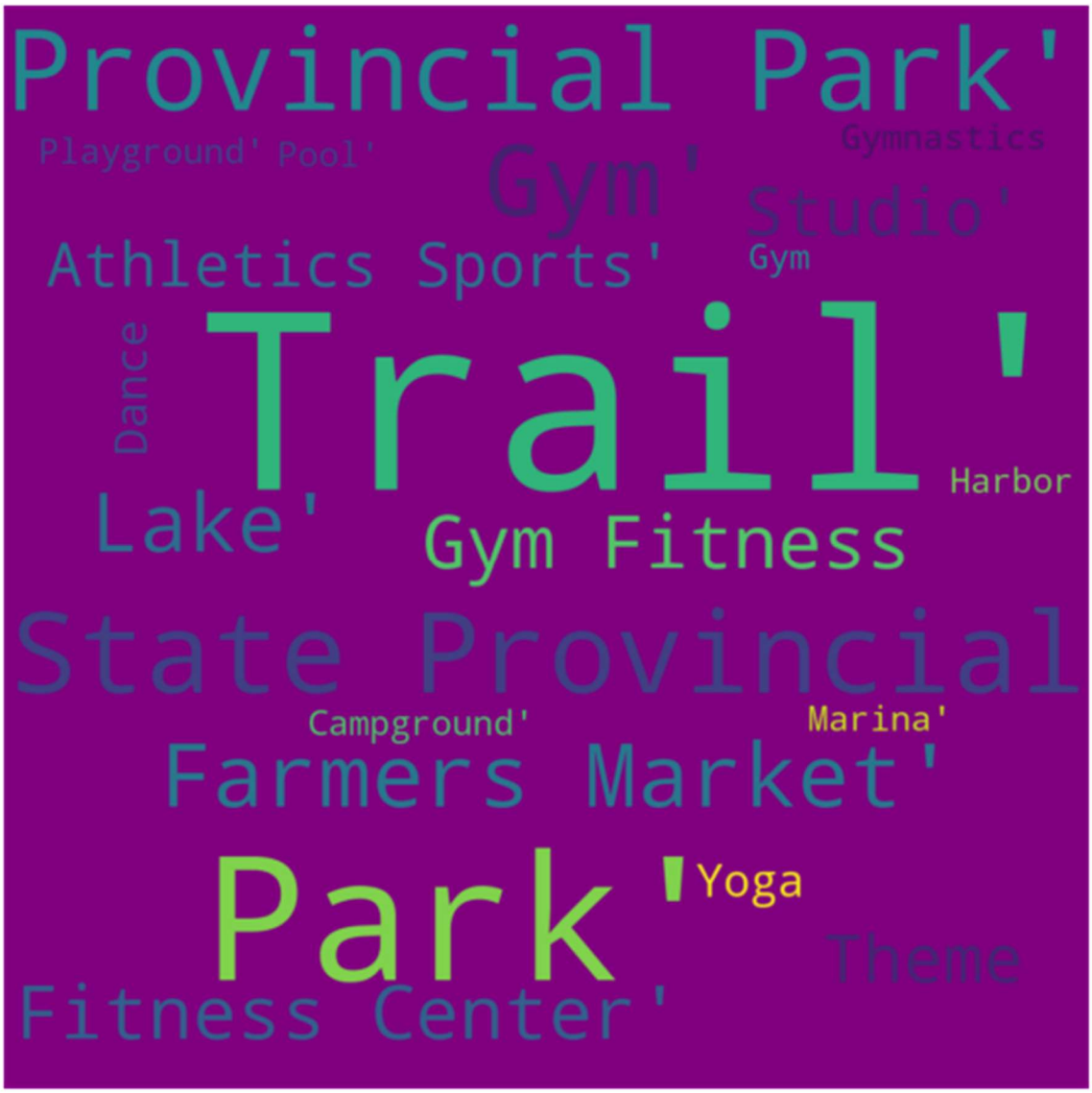


4. RESULTS

Total 5 clusters were obtained as k-clustering method was used.

Cluster0: This cluster has total of 25 Towns and the most common venues include Gym and fitness centers. The venue categories had different names for these types of venues as listed above. The Majority of the venues were gym and fitness center and a few of trails and parks.

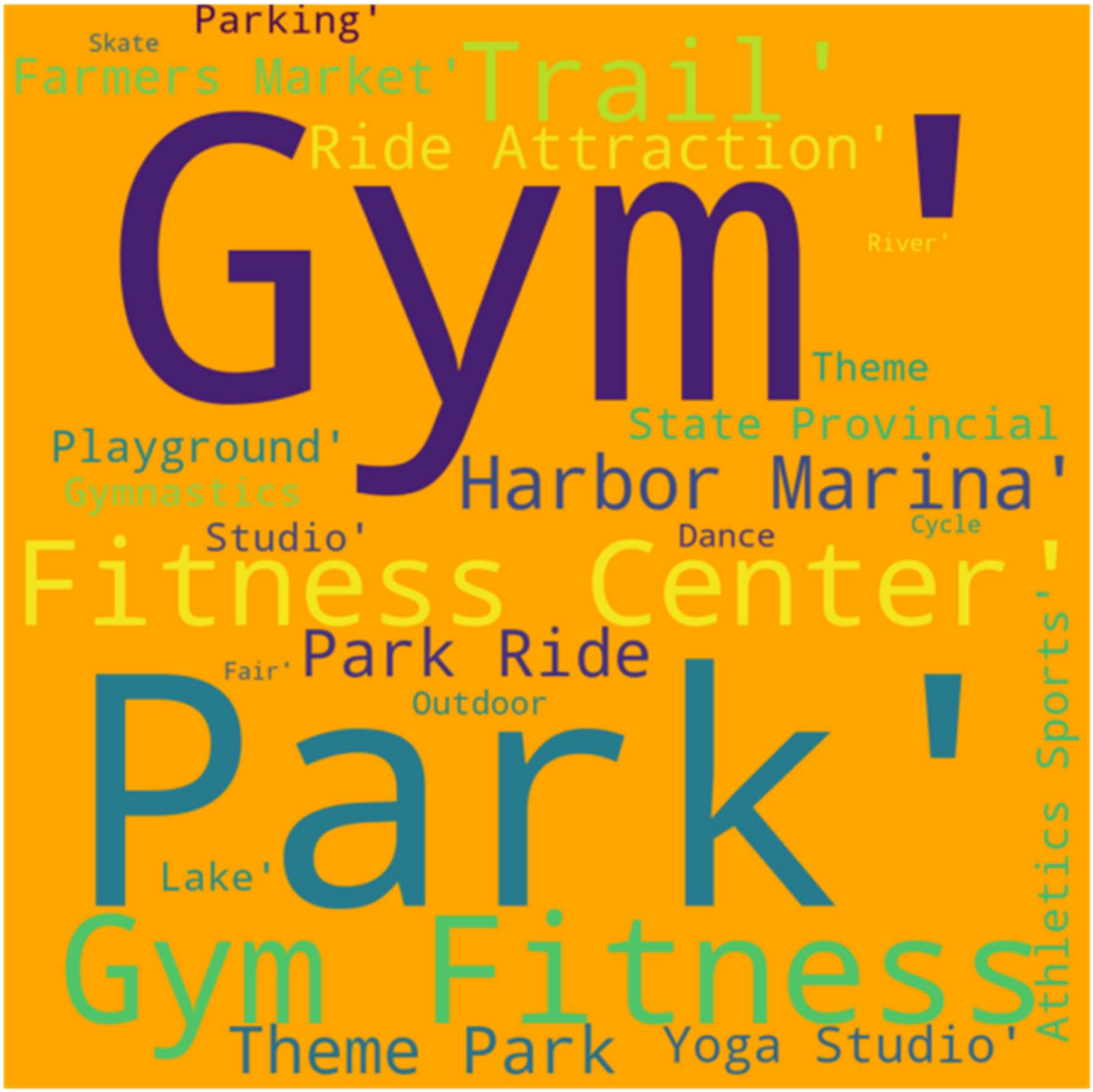
Cluster1: This cluster has total of 31 Towns and the most common venues include Trails, Parks and Lakes. The venue categories had different names for these types of venues as listed above. The Majority of the venues in this cluster were open ground recreational venues and a few of gyms and studios.



Cluster2: This cluster has total of 33 Towns and the most common venues include Campground, Harbor marina and Lakes. The venue categories had different names for these types of venues as listed above. The Majority of the venues in this cluster were open ground/water recreational venues and a few of fitness centers.

Cluster3: This cluster has total of 37 Towns and the most common venues include Farmers Market, Parks and Theme parks. The venue categories had different names for these types of venues as listed above. The Majority of the venues in this cluster were open ground recreational venues and a few of gyms and studios.

Cluster4: This cluster is a combination of Gym, Parks and Fitness centers as the most common venues followed by yoga studios, playgrounds, and farmers markets.



5. DISCUSSION

From the above results I've observed the clusters are formed which has similar venues:

Cluster0- Gym & fitness centers

Cluser1- Open ground activity venues

Cluster2- Water venues

Cluster3- Farmers' Market and theme parks

Cluster4- Combination of Gyms & Parks

The below table shows the towns in each cluster:

	Towns_C0	Towns_C1	Towns_C2	Towns_C3	Towns_C4
0	Beacon Falls	Ansonia	Andover	Bethany	Berlin
1	Bethel	Avon	Ashford	Bridgeport	Bristol
2	Branford	Barkhamsted	Bozrah	Bridgewater	Brookfield
3	Cromwell	Bethlehem	Canaan	Cheshire	Canton
4	Derby	Bloomfield	Chaplin	Colchester	Clinton
5	Farmington	Bolton	Chester	Danbury	Darien
6	Granby	Burlington	Columbia	Deep River	Durham
7	Griswold	Colebrook	Coventry	East Haven	East Hartford
8	Manchester	Cornwall	East Granby	Ellington	East Windsor
9	Meriden	Easton	East Haddam	Groton	Enfield
10	Newington	Hartland	East Hampton	Haddam	Essex
11	Newtown	Kent	East Lyme	Hebron	Fairfield
12	North Branford	Lebanon	Eastford	Ledyard	Franklin
13	Norwich	Litchfield	Goshen	Lisbon	Glastonbury
14	Orange	Lyme	Killingworth	Madison	Greenwich
15	Plainville	Mansfield	Marlborough	Middletown	Guilford
16	Putnam	Morris	New Fairfield	Milford	Hamden
17	Rocky Hill	Norfolk	New Hartford	Montville	Hartford
18	Shelton	North Haven	North Canaan	Naugatuck	Harwinton
19	Stratford	North Stonington	Norwalk	New Canaan	Killingly
20	Tolland	Pomfret	Old Lyme	New Haven	Middlebury
21	Vernon	Preston	Old Saybrook	New London	Middlefield
22	Westport	Prospect	Plymouth	New Milford	Monroe
23	Wethersfield	Redding	Ridgefield	Oxford	New Britain
24	Wolcott	Roxbury	Salem	Portland	South Windsor
25	NaN	Seymour	Salisbury	Scotland	Stonington
26	NaN	Simsbury	Southbury	Sherman	Suffield
27	NaN	Washington	Thomaston	Somers	Torrington
28	NaN	Weston	Thompson	Southington	Wallingford
29	NaN	Woodbridge	Trumbull	Sprague	Watertown
30	NaN	Woodbury	Union	Stamford	Westbrook
31	NaN	NaN	Willington	Sterling	Wilton

	Towns_C0	Towns_C1	Towns_C2	Towns_C3	Towns_C4
32	NaN	NaN	Winchester	Voluntown	Windham
33	NaN	NaN	NaN	Waterbury	Windsor
34	NaN	NaN	NaN	Waterford	Windsor Locks
35	NaN	NaN	NaN	West Hartford	Woodstock
36	NaN	NaN	NaN	West Haven	NaN

As the size of each cluster varies, the empty cells are filled with NaN.

6. CONCLUSION

Based on the above results and discussions, I would recommend a home finder to choose a town from cluster4. This cluster offers mostly all the venues of interest. However, if someone is interested in gym and fitness centers as compared to other venues, they can select cluster0.

The above results and discussions will be very helpful to the real estate agents to provide suggestions to their clients and support their suggestions with the facts and figures of this report.