Battle of the University Towns in US

For IBM Applied Data Science Capstone Project

Xin Zhao

5/7/2019

Introduction: What are university towns?

- A university town or college town is a community that is dominated by its **university population**.
- In US, it is often a separate town or city, but in some cases it can also be a city neighborhood or district.

- University towns are usually considered to be some of the best places to live in because of their **low cost of living**, **rich cultural activities**, **educational opportunities**, **fun and sports**, etc.

With the stable consumption power coming from students, faculties, staff, and retirees, university towns are favored by numerous investors.

Introduction: purpose of this study

In this project, I will explore the midsize university towns in US to study their similarities and dissimilarities.

- By “midsize”, the big cities with large populations are excluded.
- The reason is that cultures and business/life styles in large cities are influenced by many more factors compared with “university towns” where the university/institution plays important roles throughout the community.

Targeted audience

- Investors
  - e.g. Open restaurants or bookstores in one of the university towns
- Students
  - Apply colleges in the near future.

Data

The data used in this study include three main parts:

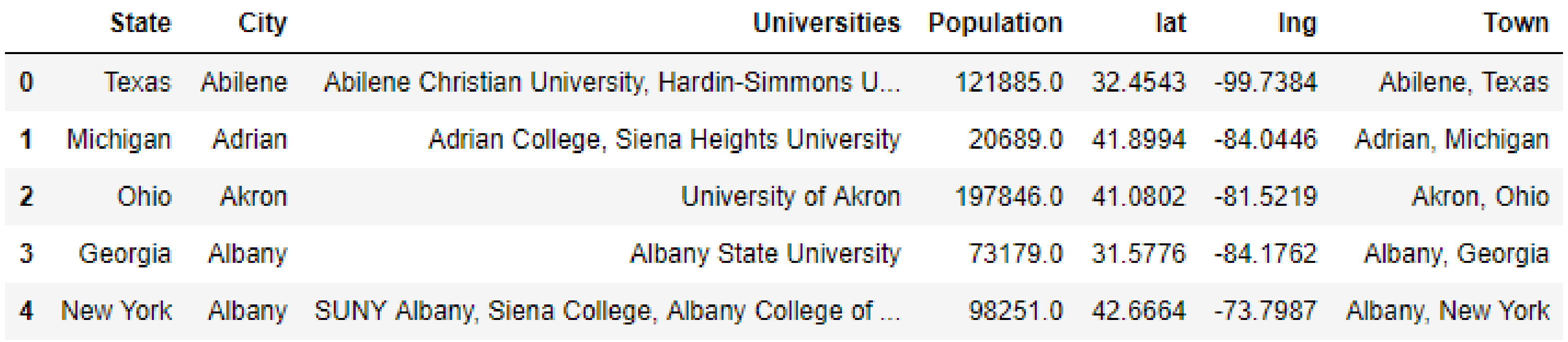
1. A list of university towns in the United States
   - Scrapped From Wikipedia using *BeatifulSoup*: https://en.wikipedia.org/wiki/List_of_college_towns#College_towns_in_the_United_States

2. Basic information about those towns, such as population, location (latitude, longitude coordinates), etc
   - 2017 Population data: https://www.biggestuscities.com/
   - Location data: https://simplemaps.com/data/us-cities

3. Nearby venues obtained using *Foursquare* API

Methodology

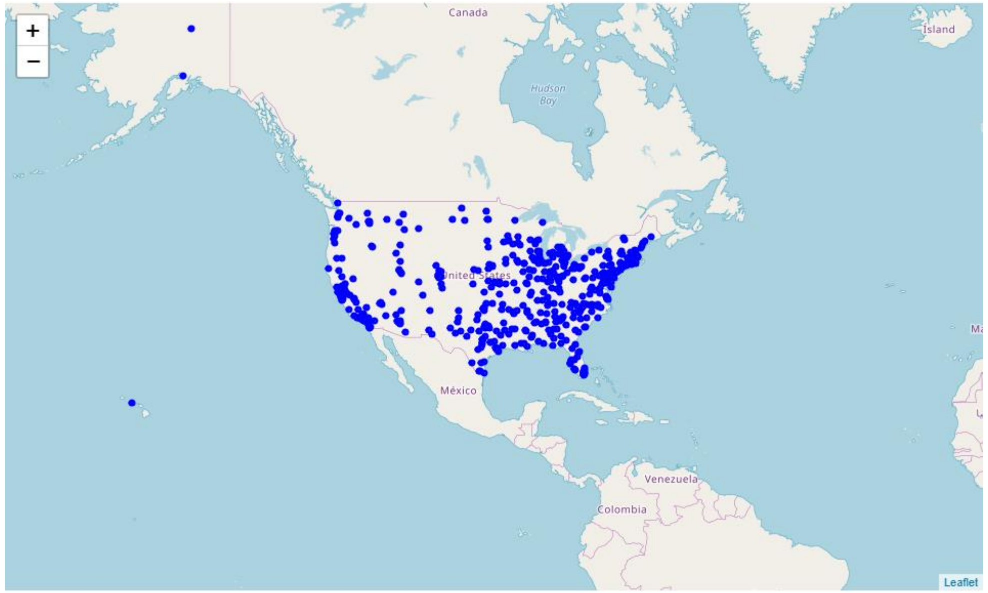
1. Data collection and cleaning
    - Website data were scrapped using *BeatifulSoup*:
    - Data are organized using *pandas* DataFrame
    - In the end, we have a dataframe of 520 university towns with complete information, e.g.



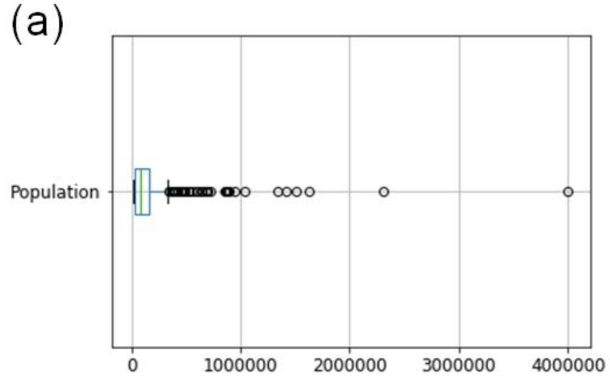
	State	City	Universities	Population	lat	lng	Town
0	Texas	Abilene	Abilene Christian University, Hardin-Simmons U...	121885.0	32.4543	-99.7384	Abilene, Texas
1	Michigan	Adrian	Adrian College, Siena Heights University	20689.0	41.8994	-84.0446	Adrian, Michigan
2	Ohio	Akron	University of Akron	197846.0	41.0802	-81.5219	Akron, Ohio
3	Georgia	Albany	Albany State University	73179.0	31.5776	-84.1762	Albany, Georgia
4	New York	Albany	SUNY Albany, Siena College, Albany College of ...	98251.0	42.6664	-73.7987	Albany, New York

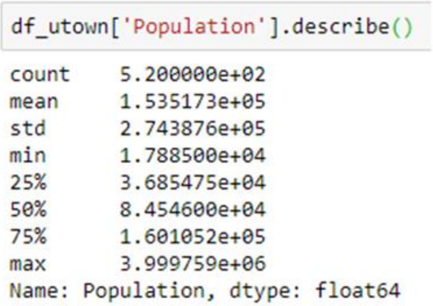
2. Statistical analysis
    - Basic statistical analysis was performed use Pandas describe function
    - Boxplot, histogram plot
3. Clustering method
    - K-means usng scikit-learn

Narrowing down the dataset using population size



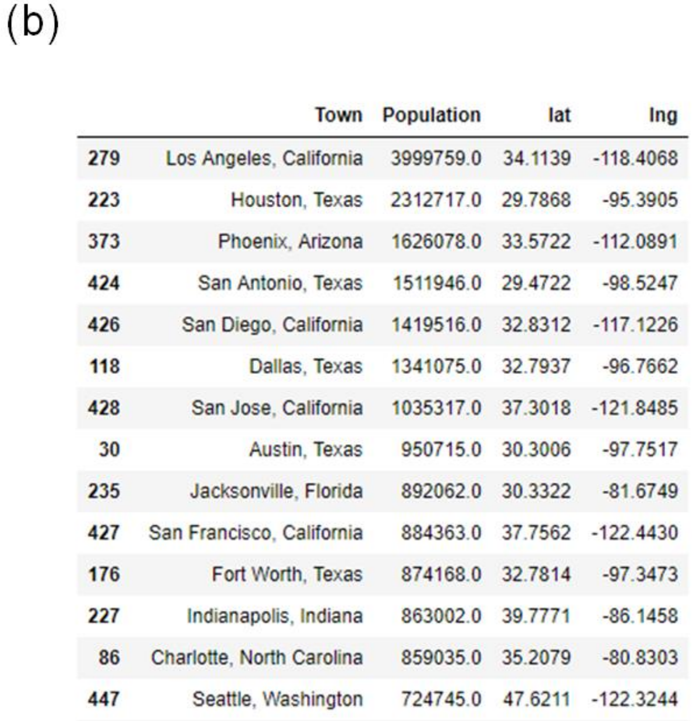
Map of the university towns in US obtained from internet with full information. There are 520 towns in total.





```
df_utown['Population'].describe()

count    5.200000e+02
mean     1.535173e+05
std      2.743876e+05
min      1.788500e+04
25%      3.685475e+04
50%      8.454600e+04
75%      1.601052e+05
max      3.999759e+06
Name: Population, dtype: float64
```



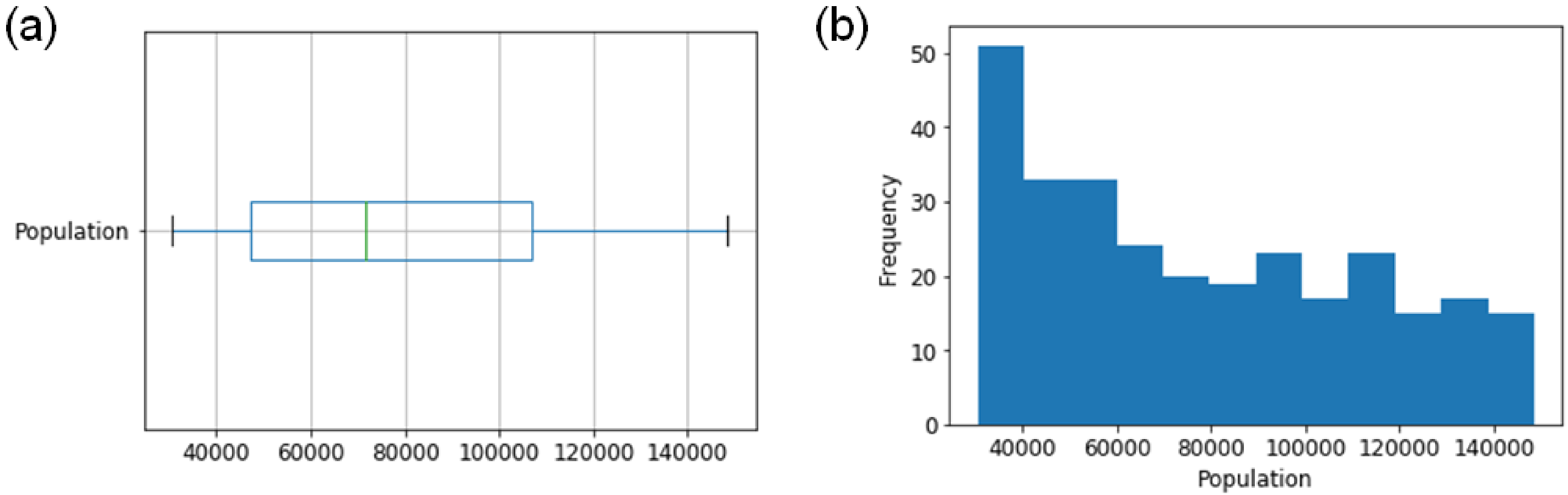
(b)

	Town	Population	lat	lng
279	Los Angeles, California	3999759.0	34.1139	-118.4068
223	Houston, Texas	2312717.0	29.7868	-95.3905
373	Phoenix, Arizona	1626078.0	33.5722	-112.0891
424	San Antonio, Texas	1511946.0	29.4722	-98.5247
426	San Diego, California	1419516.0	32.8312	-117.1226
118	Dallas, Texas	1341075.0	32.7937	-96.7662
428	San Jose, California	1035317.0	37.3018	-121.8485
30	Austin, Texas	950715.0	30.3006	-97.7517
235	Jacksonville, Florida	892062.0	30.3322	-81.6749
427	San Francisco, California	884363.0	37.7562	-122.4430
176	Fort Worth, Texas	874168.0	32.7814	-97.3473
227	Indianapolis, Indiana	863002.0	39.7771	-86.1458
86	Charlotte, North Carolina	859035.0	35.2079	-80.8303
447	Seattle, Washington	724745.0	47.6211	-122.3244

(a) Boxplot of the population of the university towns.
(b) A few examples of top cities sorted by population.

- Since here we are focusing on midsize university towns, we will remove the large cities and only keep those towns roughly within 25% ~ 75% of the population distribution.
- The range of population to be considered in this study is set to [30000, 150000] and as a result, there are 290 towns left in the dataset.

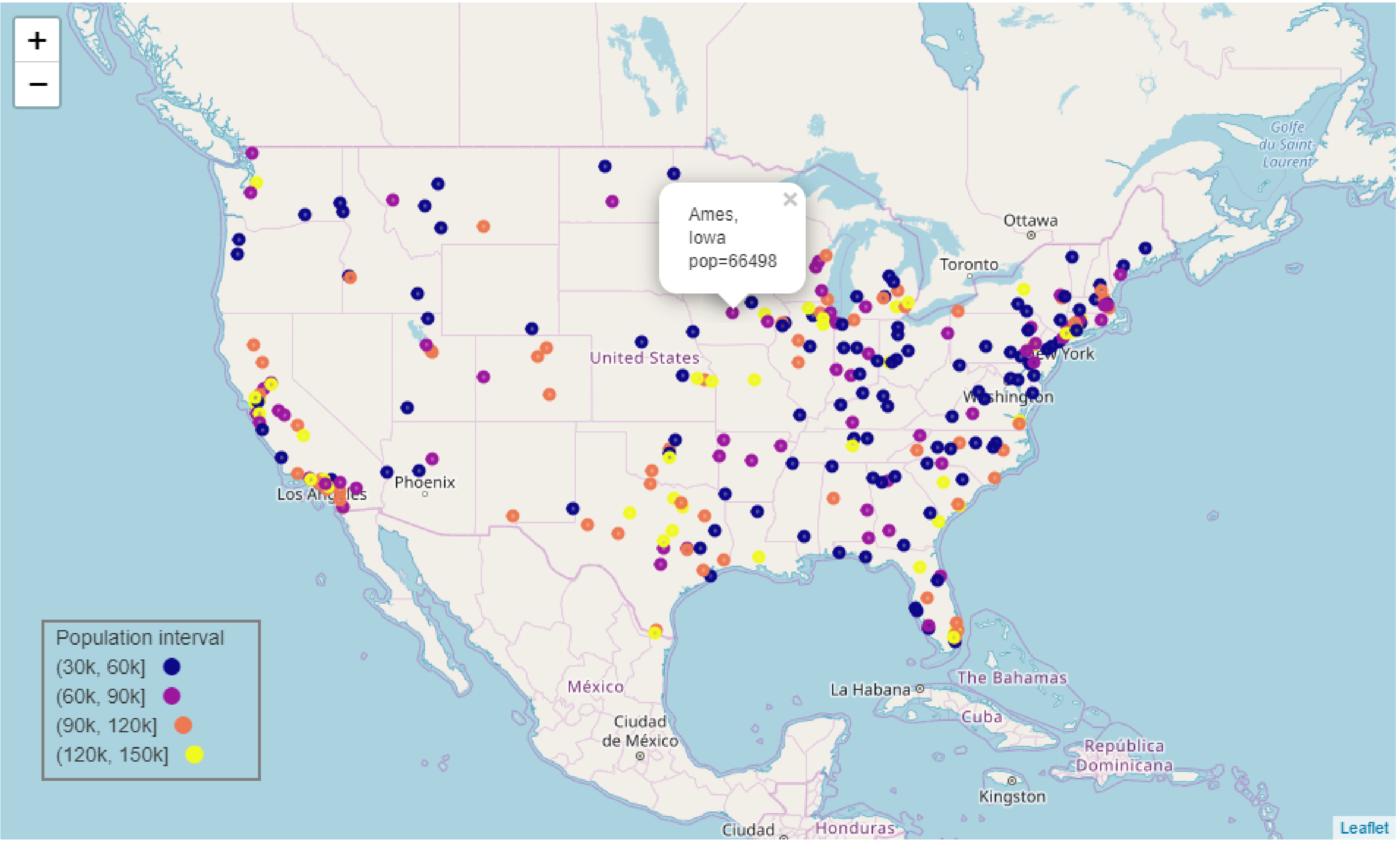
Population distribution of the final 290 university towns:



(a) Boxplot; (b) Histogram.

- We see that university towns tend to have small populations, and as the population size increases, the number of identified university towns decreases.

Mapping the final university towns (total 290) selected for analysis.



- The color scheme is based on the different population intervals, as shown in the legend.
- By clicking each town, the pop out message shows the name of the town and its population, as indicated by the example “Ames, Iowa” which is the town for Iowa State University.

Results

Summary of the venues returned by Foursquare:

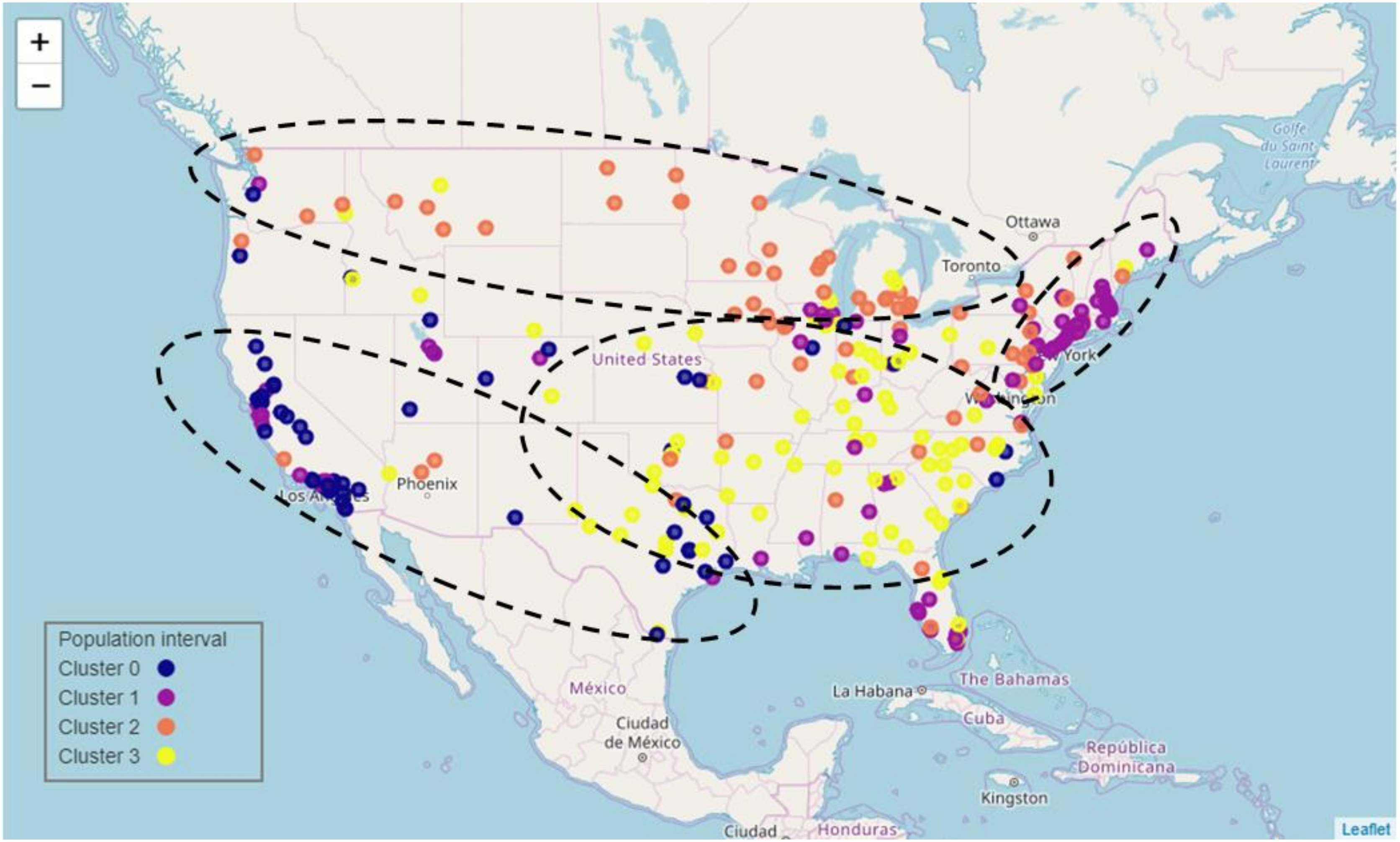
- Using Foursquare API, we can get a maximum of 100 venues near each university town.

- For 290 towns in total, 27889 venue places are returned, averaging 96 venues per town.

- 7161 out of the 27889 venues contain the word “Restaurant” in their venue category, i.e. nearly 1/3 of the venues found in the university towns are restaurants. Considering there are also other venue categories like “Pizza Place”, “Sandwich Place”, “Burger Joint”, etc., food related business are without any doubt the most popular one in university towns.

- Among the 27889 venues, there are in total 481 unique venue categories.

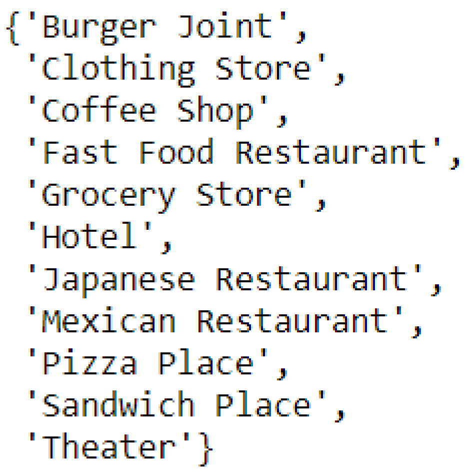
Results: Clustering of the university towns based on venues



- The color scheme is based on the different cluster labels, as shown in the legend.
- The dashed black ellipses are to guide the eyes to show the relation between cluster group and the town location.

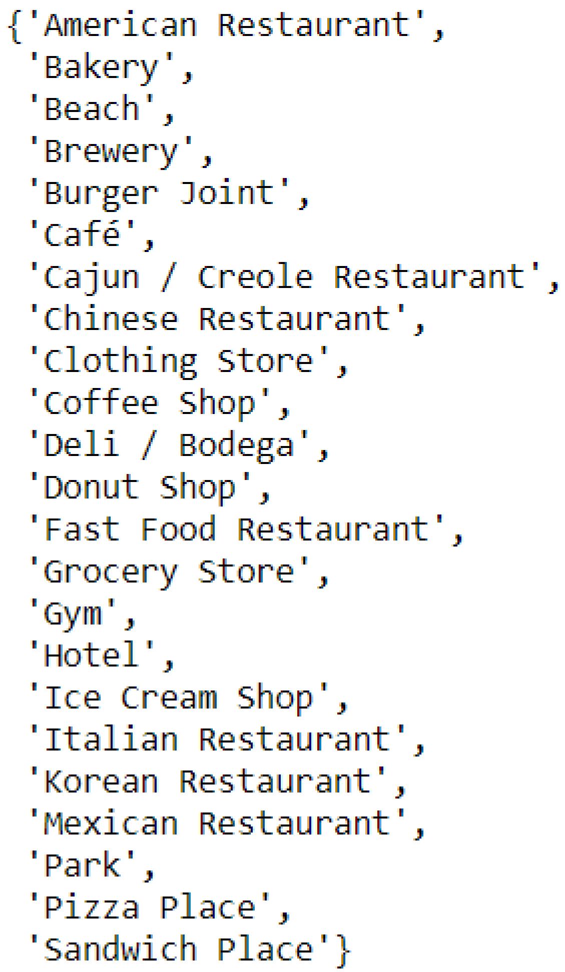
Most common venue categories for each cluster (4 cluster groups)

Cluster 0



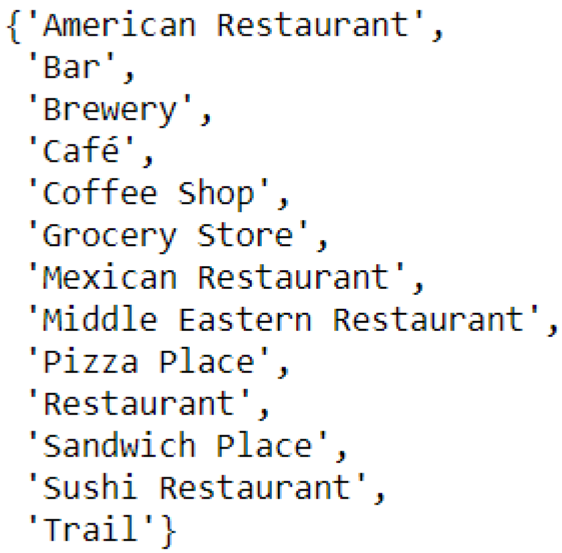
```
{'Burger Joint',
 'Clothing Store',
 'Coffee Shop',
 'Fast Food Restaurant',
 'Grocery Store',
 'Hotel',
 'Japanese Restaurant',
 'Mexican Restaurant',
 'Pizza Place',
 'Sandwich Place',
 'Theater'}
```

Cluster 1



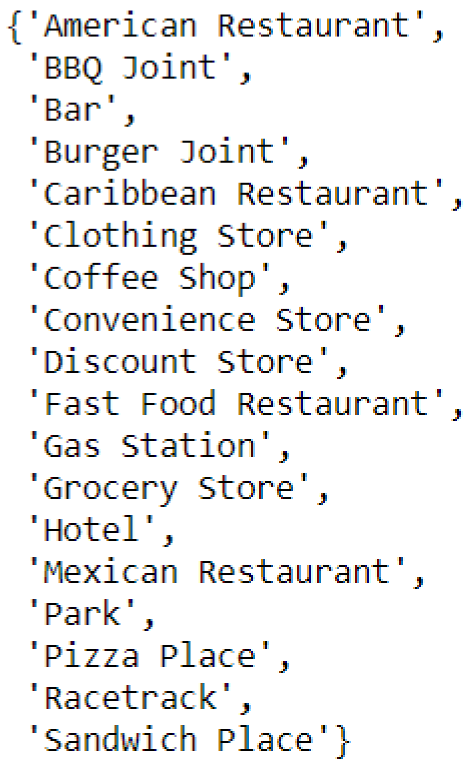
```
{'American Restaurant',
 'Bakery',
 'Beach',
 'Brewery',
 'Burger Joint',
 'Café',
 'Cajun / Creole Restaurant',
 'Chinese Restaurant',
 'Clothing Store',
 'Coffee Shop',
 'Deli / Bodega',
 'Donut Shop',
 'Fast Food Restaurant',
 'Grocery Store',
 'Gym',
 'Hotel',
 'Ice Cream Shop',
 'Italian Restaurant',
 'Korean Restaurant',
 'Mexican Restaurant',
 'Park',
 'Pizza Place',
 'Sandwich Place'}
```

Cluster 2



```
{'American Restaurant',
 'Bar',
 'Brewery',
 'Café',
 'Coffee Shop',
 'Grocery Store',
 'Mexican Restaurant',
 'Middle Eastern Restaurant',
 'Pizza Place',
 'Restaurant',
 'Sandwich Place',
 'Sushi Restaurant',
 'Trail'}
```

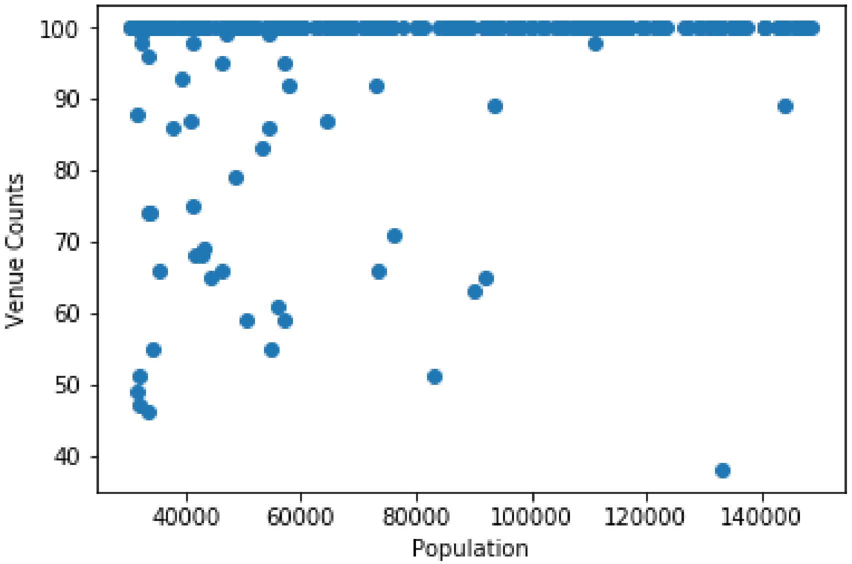
Cluster 3



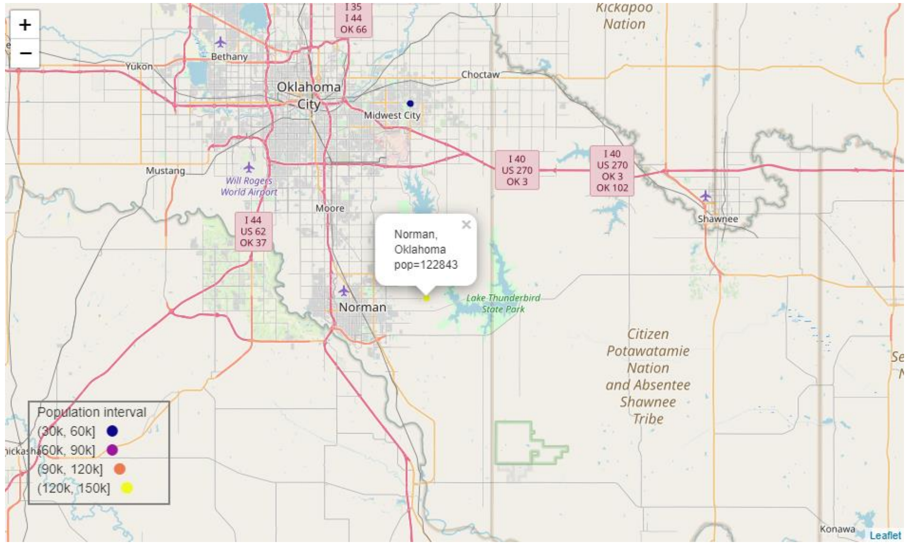
```
{'American Restaurant',
 'BBQ Joint',
 'Bar',
 'Burger Joint',
 'Caribbean Restaurant',
 'Clothing Store',
 'Coffee Shop',
 'Convenience Store',
 'Discount Store',
 'Fast Food Restaurant',
 'Gas Station',
 'Grocery Store',
 'Hotel',
 'Mexican Restaurant',
 'Park',
 'Pizza Place',
 'Racetrack',
 'Sandwich Place'}
```

Discussion: errors in the data sources

- Unreliable data sources from Wikipedia
  - Duplicate items
  - Misnamed towns
- Inaccurate location coordinates
  - Ideally, there should be no problem for each university town to return 100 venues as we set the lower limit of the population to be 30,000.



Venue counts as the function of the town population for the 290 university towns.



Map to show the wrong coordinates of Norman, Oklahoma (one example of errors in the location data)

Discussion: Feature designing

It will be helpful to find more features to describe the similarities and dissimilarities of the university towns.

- Combine the 481 venue categories into more representative features
- Information about the universities/institutions in each town
- Other information about the town, such as household income of the town, etc.

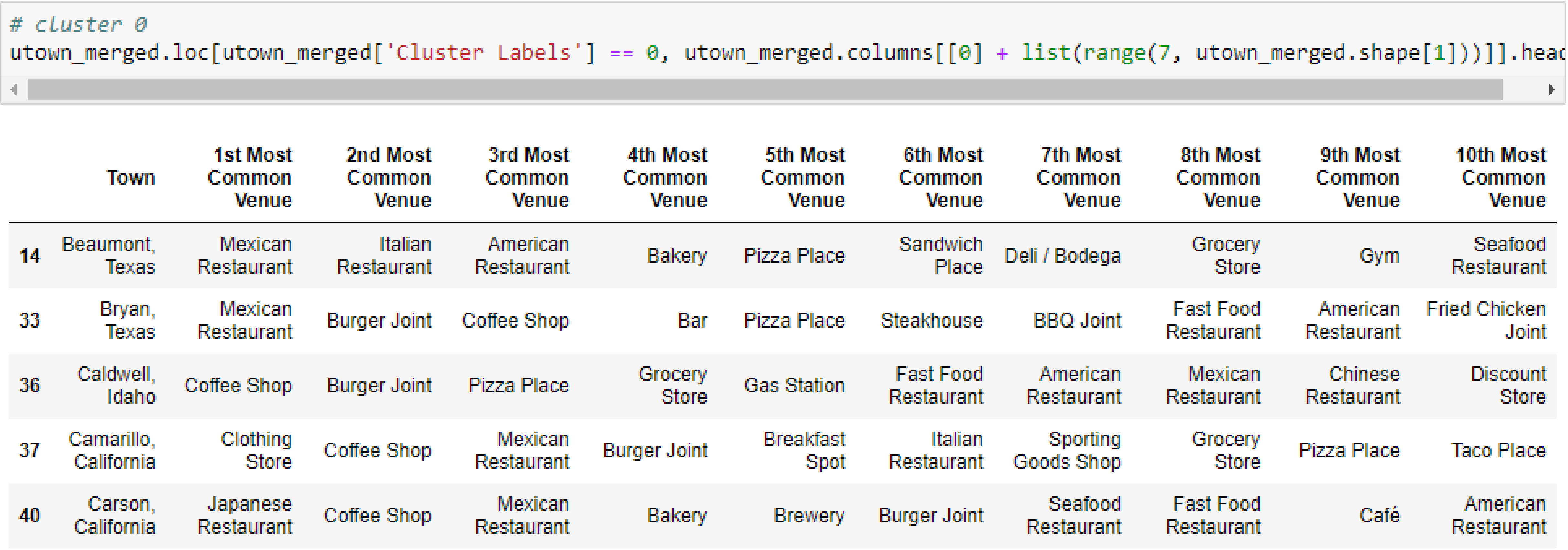
Conclusions

Based on the analysis on 290 university towns with population size among (30000, 150000), we can conclude that,

- Food-related venues are most popular in the university towns in US.
- Considering the most common venue category, “Restaurant”, “Coffee Shop”, “Convenience/Grocery Store”, and “Bar/Brewery” are most frequent in all the 4 groups we clustered into.
- The clustering of the university towns based on venue distributions shows clear dependency on the geometrical location of the town, indicating that the life style of each town is clearly influenced by the regional culture although people in the universities usually comes from all over the US/world.

Backup slides

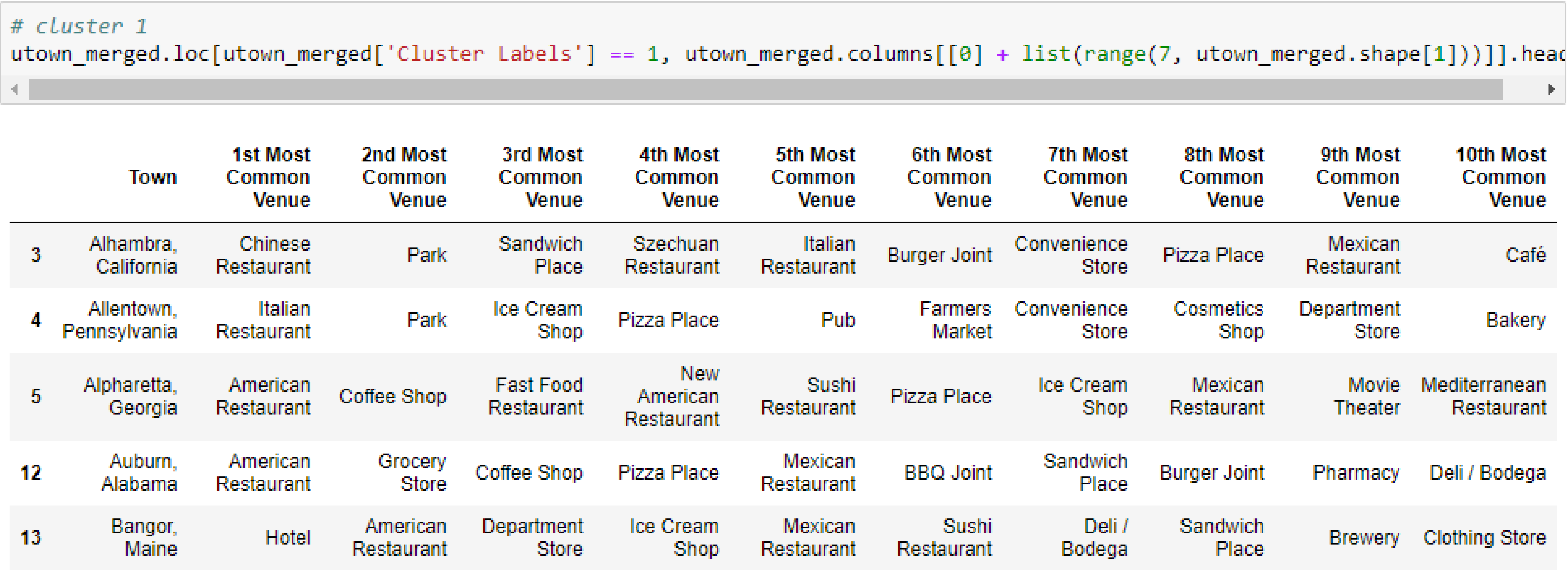
Some example university towns in cluster 0:



```
# cluster 0
utown_merged.loc[utown_merged['Cluster Labels'] == 0, utown_merged.columns[[0] + list(range(7, utown_merged.shape[1]))]].heac
```

	Town	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
14	Beaumont, Texas	Mexican Restaurant	Italian Restaurant	American Restaurant	Bakery	Pizza Place	Sandwich Place	Deli / Bodega	Grocery Store	Gym	Seafood Restaurant
33	Bryan, Texas	Mexican Restaurant	Burger Joint	Coffee Shop	Bar	Pizza Place	Steakhouse	BBQ Joint	Fast Food Restaurant	American Restaurant	Fried Chicken Joint
36	Caldwell, Idaho	Coffee Shop	Burger Joint	Pizza Place	Grocery Store	Gas Station	Fast Food Restaurant	American Restaurant	Mexican Restaurant	Chinese Restaurant	Discount Store
37	Camarillo, California	Clothing Store	Coffee Shop	Mexican Restaurant	Burger Joint	Breakfast Spot	Italian Restaurant	Sporting Goods Shop	Grocery Store	Pizza Place	Taco Place
40	Carson, California	Japanese Restaurant	Coffee Shop	Mexican Restaurant	Bakery	Brewery	Burger Joint	Seafood Restaurant	Fast Food Restaurant	Café	American Restaurant

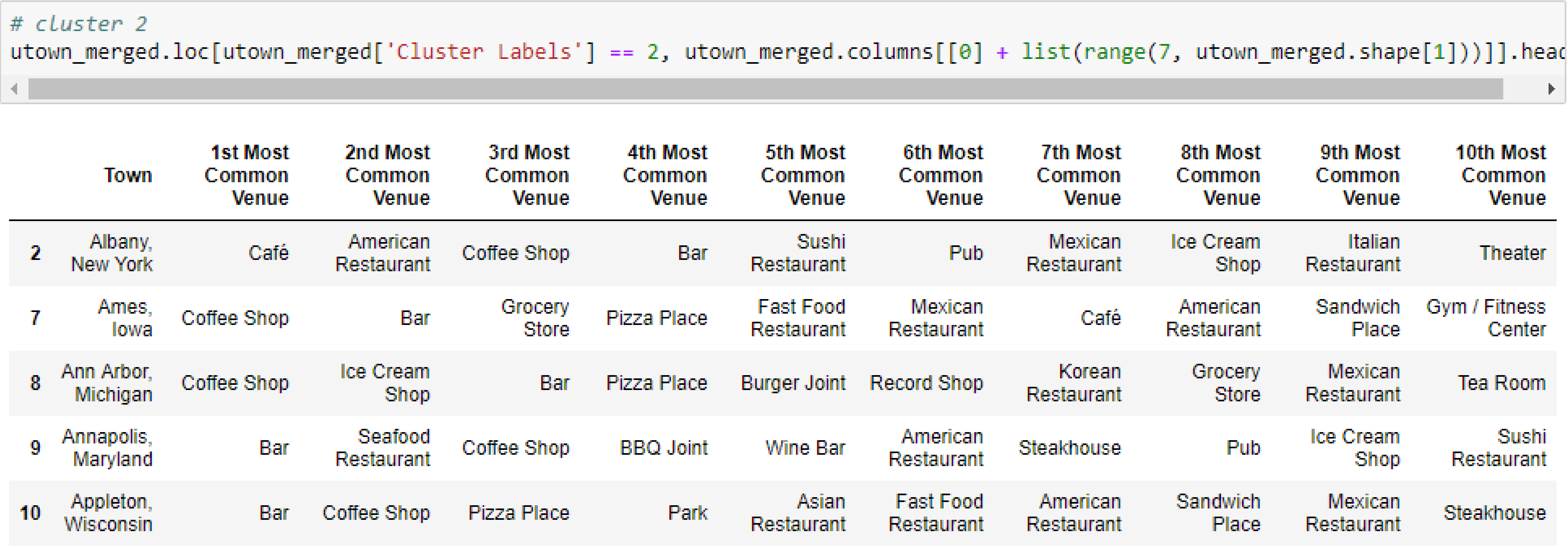
Some example university towns in cluster 1:



```
# cluster 1
utown_merged.loc[utown_merged['Cluster Labels'] == 1, utown_merged.columns[[0] + list(range(7, utown_merged.shape[1]))]].hea
```

	Town	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	Alhambra, California	Chinese Restaurant	Park	Sandwich Place	Szechuan Restaurant	Italian Restaurant	Burger Joint	Convenience Store	Pizza Place	Mexican Restaurant	Café
4	Allentown, Pennsylvania	Italian Restaurant	Park	Ice Cream Shop	Pizza Place	Pub	Farmers Market	Convenience Store	Cosmetics Shop	Department Store	Bakery
5	Alpharetta, Georgia	American Restaurant	Coffee Shop	Fast Food Restaurant	New American Restaurant	Sushi Restaurant	Pizza Place	Ice Cream Shop	Mexican Restaurant	Movie Theater	Mediterranean Restaurant
12	Auburn, Alabama	American Restaurant	Grocery Store	Coffee Shop	Pizza Place	Mexican Restaurant	BBQ Joint	Sandwich Place	Burger Joint	Pharmacy	Deli / Bodega
13	Bangor, Maine	Hotel	American Restaurant	Department Store	Ice Cream Shop	Mexican Restaurant	Sushi Restaurant	Deli / Bodega	Sandwich Place	Brewery	Clothing Store

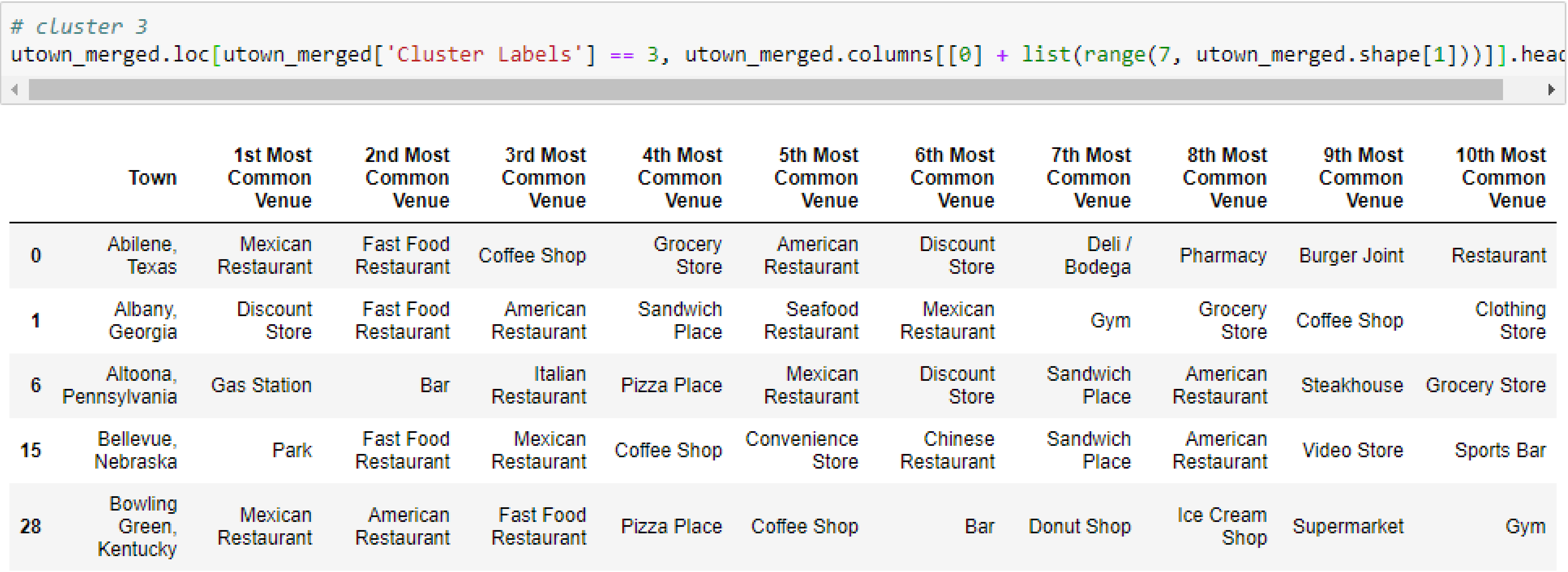
Some example university towns in cluster 2:



```
# cluster 2
utown_merged.loc[utown_merged['Cluster Labels'] == 2, utown_merged.columns[[0] + list(range(7, utown_merged.shape[1]))]].hea
```

	Town	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Albany, New York	Café	American Restaurant	Coffee Shop	Bar	Sushi Restaurant	Pub	Mexican Restaurant	Ice Cream Shop	Italian Restaurant	Theater
7	Ames, Iowa	Coffee Shop	Bar	Grocery Store	Pizza Place	Fast Food Restaurant	Mexican Restaurant	Café	American Restaurant	Sandwich Place	Gym / Fitness Center
8	Ann Arbor, Michigan	Coffee Shop	Ice Cream Shop	Bar	Pizza Place	Burger Joint	Record Shop	Korean Restaurant	Grocery Store	Mexican Restaurant	Tea Room
9	Annapolis, Maryland	Bar	Seafood Restaurant	Coffee Shop	BBQ Joint	Wine Bar	American Restaurant	Steakhouse	Pub	Ice Cream Shop	Sushi Restaurant
10	Appleton, Wisconsin	Bar	Coffee Shop	Pizza Place	Park	Asian Restaurant	Fast Food Restaurant	American Restaurant	Sandwich Place	Mexican Restaurant	Steakhouse

Some example university towns in cluster 3:



```
# cluster 3
utown_merged.loc[utown_merged['Cluster Labels'] == 3, utown_merged.columns[[0] + list(range(7, utown_merged.shape[1]))]].head
```

	Town	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abilene, Texas	Mexican Restaurant	Fast Food Restaurant	Coffee Shop	Grocery Store	American Restaurant	Discount Store	Deli / Bodega	Pharmacy	Burger Joint	Restaurant
1	Albany, Georgia	Discount Store	Fast Food Restaurant	American Restaurant	Sandwich Place	Seafood Restaurant	Mexican Restaurant	Gym	Grocery Store	Coffee Shop	Clothing Store
6	Altoona, Pennsylvania	Gas Station	Bar	Italian Restaurant	Pizza Place	Mexican Restaurant	Discount Store	Sandwich Place	American Restaurant	Steakhouse	Grocery Store
15	Bellevue, Nebraska	Park	Fast Food Restaurant	Mexican Restaurant	Coffee Shop	Convenience Store	Chinese Restaurant	Sandwich Place	American Restaurant	Video Store	Sports Bar
28	Bowling Green, Kentucky	Mexican Restaurant	American Restaurant	Fast Food Restaurant	Pizza Place	Coffee Shop	Bar	Donut Shop	Ice Cream Shop	Supermarket	Gym