Battle of the Midsize University Towns in US

Introduction

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 also favored by numerous investors. Because university towns usually have students forming a major fraction of their population, businesses that typically cater to students are in huge demand in such places.

In this project, I will explore the midsize university towns in US to study their similarities and dissimilarities. Here by "midsize", I meant to exclude the big cities with large populations. 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.

This study will be helpful to both investors (for example those who want to open restaurants or bookstores in one of the university towns) and students who are going to apply colleges in the near future to make their decision.

Data

The data used in this study include three main parts.

1. A list of university towns in the United States.
2. Some basic information about those towns, such as population, location (latitude, longitude coordinates), etc.
3. Once having the database of university towns and their locations, *Foursquare* API [1] is used to explore the nearby venues.

By analyzing the complete dataset, it is then possible to find the relation between the common venue types and certain features of the university towns such as locations, populations etc. and possibly other features (e.g. household income, international student percentage, etc.) if available.

After an internet search, I found a list of university towns around the world in Wikipedia [2]. The US part can be scrapped using the *BeautifulSoup* package [3]. As for the population (scrapped from [4]) and location data (downloaded from [5]), online databases are used. One note on the location data, I first tried to use *geopy* package [6] to extract their location coordinates using name address, but it took quite a few tries to get a complete list (due to limitations of their service) and the returned results are not always accurate. Instead, I decided to use the tabulated

online resources with all the US cities and their locations (later we'll see that there are also inaccurate location coordinates for some cities in this database).

Methodology

1. **Data collection and cleaning**

   The data used in this study were mostly scrapped from the web pages using *BeautifulSoup*. The location of US cities was downloaded online in the format of CSV file. *Pandas* package [7] was used to read in the data file and save them to DataFrame.

   1.1 List of university towns in US

   In Wikipedia, the list of university towns includes information from all over the world, while the states of US are listed as level 3 (h3) headings. All university towns in each state are stored using unordered lists (ul), so I first saved all the ul content in a python list (**us_UTown_lists**) and loop through them to create the DataFrame for university towns.

   After extracting the US states (50 states + DC) and all the unordered lists belonging to US, I noticed that the lengths of them do not match: there are 66 items in **us_UTown_lists**. The reason was found to be existence of sub-lists in some of the high college density areas, such as in the city of Los Angeles, Atlanta, Chicago, etc. To skip those sub-lists, I used a python dictionary to store how many lists to skip for each state containing sub-list(s). Note that this does not affect the final results, because the content in the sub-lists also appear in the main list.

   The DataFrame **df_utown** is created with 918 university towns.

   1.2 Populations of US cities

   Ref. 4 lists all the US cities in the order of population size based on 2017 data. The top 2000 cities are scrapped. The 2000^{th} city has a population of 17,828 (White Settlement, Texas), which is already rather small and unlikely to have national universities. As here we focus on midsize university towns, cutoff at 2000^{th} would be a reasonable choice. The data in the referred website are stored in a table, so scrapping is straightforward: simply go through each row in the table and save the needed content (City name, State name, Population).

   The data is saved to DataFrame **df_city_pop** with 2000 rows.

   1.3 Locations of US cities

   Locations of US cities were downloaded from Ref. 5 in the format of .CSV. We read in the file using pandas.read_csv and saved the 'City', 'State', 'lat', and 'lng' in the DataFrame **df_loc** (total row number = 37842).

   Finally, merging above three DataFrames with both the 'City' and 'State' columns to join on, I got a dataset with 520 rows to proceed, i.e. there are 520 university towns

with full information. Explanation for the reduction in total number: If we look into **df_utown** from Wikipedia, we can see some of them are either misnamed/informally-named or belong to neighborhoods of cities, which cannot be found in the **df_loc** or **df_city_pop** DataFrames. Therefore, the merged final dataframe has only 521 rows left. In addition, "Memphis, Tennessee" appeared twice in **df_utown** (i.e. Wikipedia), so we need to remove one of them. Thus, in the end there are 520 university towns left in the dataset.

I could spend more time correcting those town/region names and find their populations from other resources, but considering that most of them are in fact neighborhoods of large cities, and we only focus on midsize university towns, so for this study, I decided to proceed with the 520 towns. Using *folium* package [7], I plotted the university towns in the US map as shown in Figure 1.

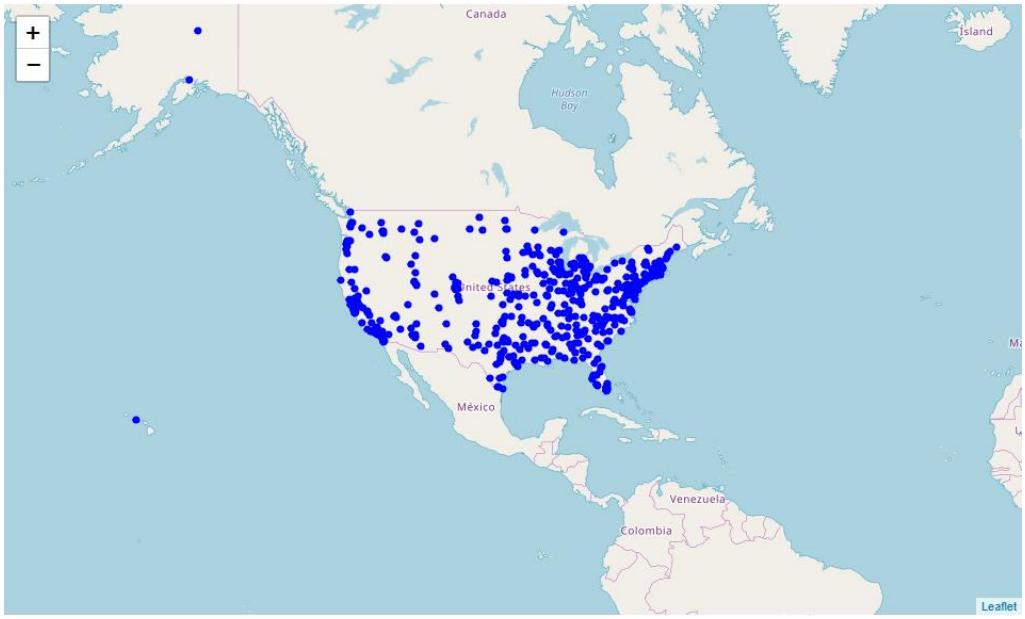


Figure 1. Map of the university towns in US obtained from internet with full information. There are 520 towns in total.

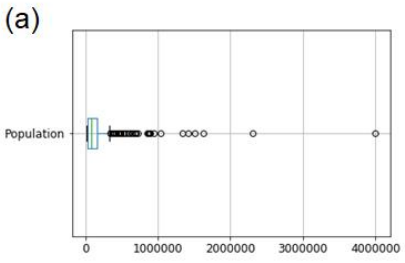
1.4 Venue data

Venue data are collected using *Foursquare* API [1]. Due to the limitation of their service, for each location, we can only get a maximum of 100 venues nearby. This may result in incomplete representation of the towns, but as a practice, the amount of data is enough for analysis and the result could still be informative.

2. Statistical analysis

After finishing the preparation of the dataset, I first examined the distribution of the populations. A box plot using the 'Population' column is plotted as the top panel of Figure 2(a). We can see that there are a lot of outliers identified. From the descriptive

statistics of the population data shown in Figure 2(a) lower panel, 25th percentile has a population of 36,855 and 75th percentile has a population of 160,000. After sorting the dataset in descending order of population (see Figure 2b), we can tell the outlier university towns are in fact some of the largest cities in US, which is not what we want to focus on.




```
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

Figure 2. (a) Boxplot of the population of the university towns. (b) A few examples of top cities sorted by population.

As argued in the introduction section, in the following analysis, 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. Their population distribution is plotted in Figure 3. Now the distribution of population is more compact and no outlier is detected. From Figure 3(b), we see that university towns tend to have small populations, and as the population size increases, the number of identified university towns decreases.

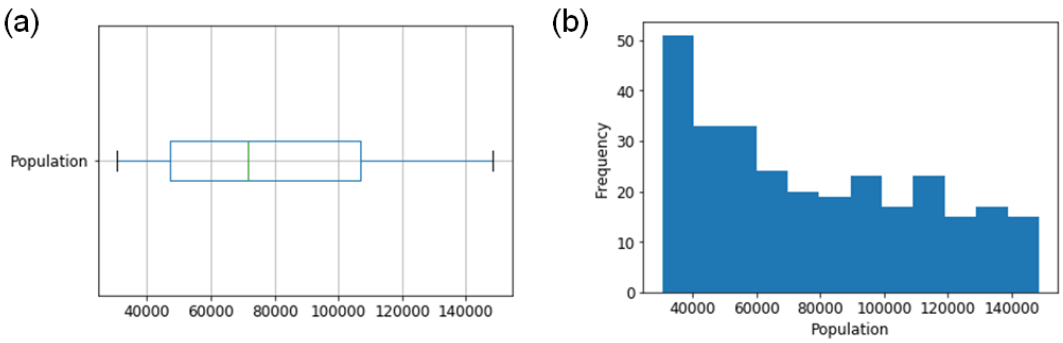


Figure 3. Population distribution of the final 290 university towns. (a) Boxplot; (b) Histogram.

3. Clustering method
   To study the similarities and dissimilarities between university towns, k-means clustering is used. K-means clustering is a method of vector quantization, originally from signal processing. k-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. In our case, we have 290 university towns to cluster and a cluster number of 4 was used in the analysis.
   scikit-learn package [8] is used to perform the machine learning tasks.

Results

There are 290 university towns in the final dataset. First let's plot them on the US map and draw the towns using different colors by dividing them into four population intervals, as shown in Figure 4. First, we note that the university towns in Hawaii and Alaska are removed due to small population sizes. From the map, we can see that the university towns with population between 30,000 and 60,000 are scattered among various states. But for towns with relative large populations, they mainly locate in the metropolitan area of large cities, such as Los Angeles, New York, Chicago, Detroit, etc.

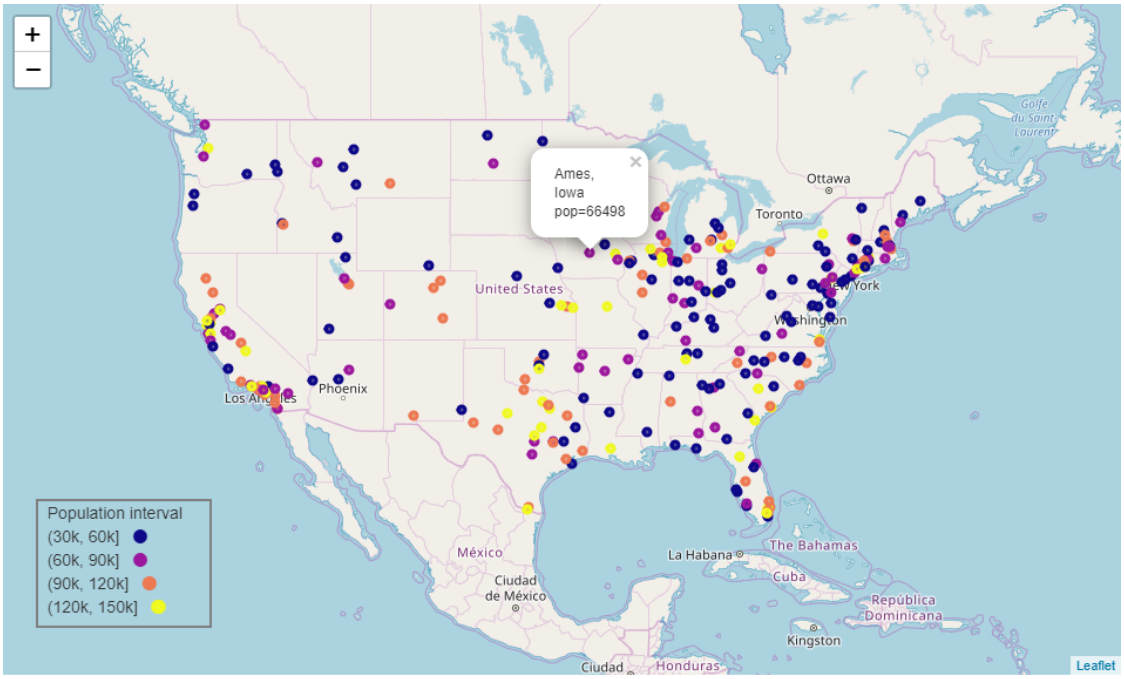


Figure 4. Map of the final university towns (total 290) selected for analysis. The color scheme is based on the different population intervals, as shown in the legend. By clicking each town, the pop out message shows the name of the town and its population, as indicated by the example "Ames, Iowa" which is the town for Iowa State University.

Next, 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. One note

on the gathering of venues: in the first try of exploring, one town (Norman, Oklahoma) only has 4 venues located. The reason was found to be its wrong location coordinates. After manually correcting for this town, the venue counter showed that the least number of venues found for any town was 38 (for Columbia, South Carolina). Although a venue count of 38 is still not ideal, especially for a town with population larger than 100,000, it might be statistically endurable. On the other hand, it can be expected that the location coordinates of other towns could also be inaccurate. This could have some influence on the following analysis.

By checking the returned results, 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. In addition, 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.

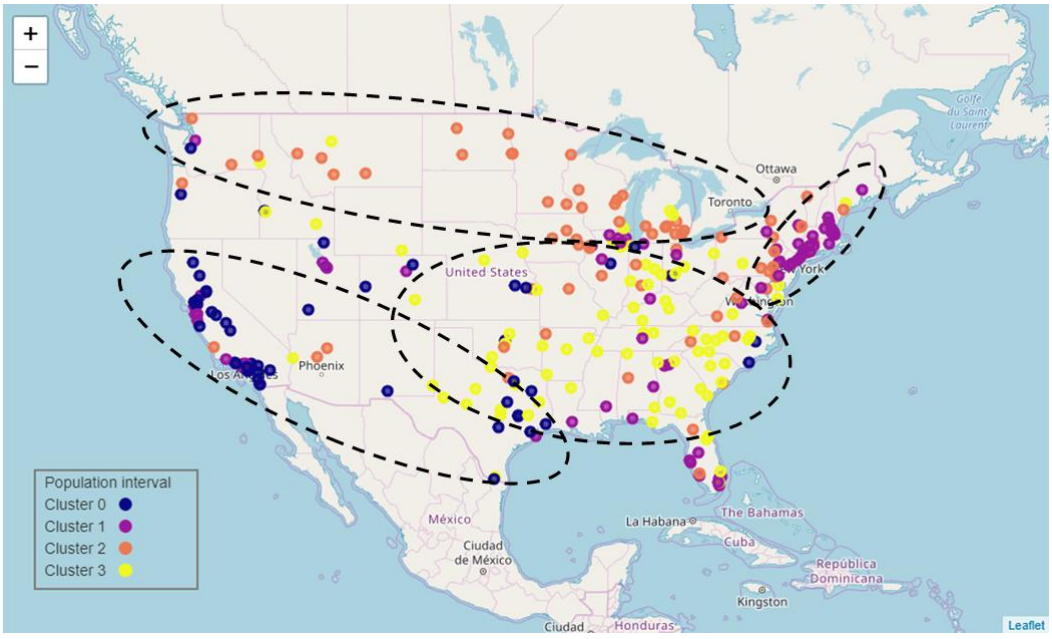


Figure 5. Map of the university towns (total 290) clustered according to their venue distributions. The color scheme is based on the different cluster labels, as shown in the legend. The dashed black ellipses are to guide the eyes to show the relation between cluster group and the town location.

Among the 27889 venues, there are in total 481 unique venue categories. Using them as the features, we can now cluster the 290 university towns into different groups using k-means method. The cluster number of 4 was used and the results are plotted in Figure 5. We can see that the clustering of the university towns has a clear dependency on the town's location. As guided by the dash black ellipses, the southeast, north, northeast, and southwest regions of US roughly correspond to the 4 clusters obtained based on venue analyses. It indicates that although

university towns usually have people gathered from the whole world, the life style of each town is clearly influenced by the regional culture.

Some examples of the university towns belonging to each cluster are shown in the following. In each table, the five examples (based on the alphabetical order of the town name) are listed with the top 10 most common venues. We can again see that the most common venues in different towns are food-related places. The difference between different clusters comes from the relative weights of different types of restaurants.

Table 1. 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]))]].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
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

Table 2. 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]))]].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
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

Table 3. 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]))]].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
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

Table 4. 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

If focusing on the 1st most common venue in each university town, we can create a set (i.e. no duplicate items) of venue categories for each cluster. The results are summarized in Figure 6, where we can tell that besides food-related venues, Coffee Shop, Convenience/Grocery Store, and Bar/Brewery are also very popular in all clusters.

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'}
```

Figure 6. Collections of the 1st most common venue categories for each cluster.

Discussion

First I'd like to make a few comments on the reliability of the data sources. During the data collection and analysis, clearly we can see the data can easily causing troubles due to inaccurate information.

1. Unreliable data sources from Wikipedia

Since some of Wikipedia pages are unprotected and can be edited by anyone. This causes the reliability of information from Wikipedia is not always high. Taking the list of university towns in the US as an example, not all towns are followed by the corresponding universities. If anyone wants to further analyze the subgroup of the population from the university, it will be difficult. In addition, duplicate items and misnamed towns are also troublesome.

2. 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. But as plotted in Figure 7, we can see that some of them are less than 100, even for towns with population larger than 100,000. I believe this is related to the inaccurate location coordinates of those towns. One obvious one (Norman, Oklahoma) that we manually corrected because there were only 4 venues returned is plotted in Figure 8 as an example. The wrong coordinates point to a place clearly outside Norman.
   If more time can be devoted to correcting those, the results should be more reliable. But again, maximum 100 venues can be return by *Foursquare* API, so the dataset must still be an incomplete representation of towns.

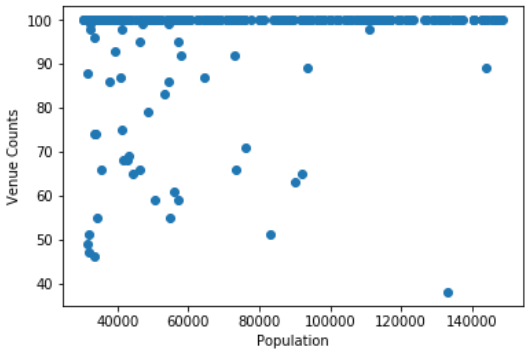


Figure 7. Venue counts as the function of the town population for the 290 university towns.

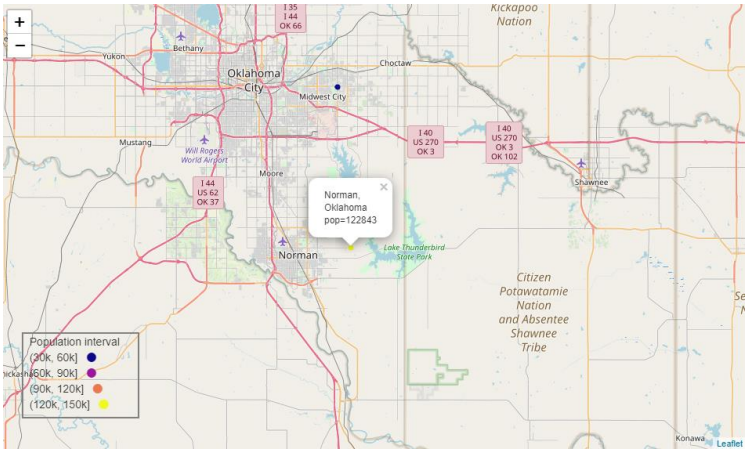


Figure 8. Map to show the wrong coordinates of Norman, Oklahoma.

Another thing I want to add is on finding more features to describe the similarities and dissimilarities of the university towns. Information about the universities/institutions in each town can be good features to classify the towns, which will be studied in the future.

Conclusion

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

a. Food-related venues are most popular in the university towns in US.
b. 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.
c. 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.

By devoting more time to this, there are a few ways to improve the analysis:

a. We could find ways to combine the 481 venue categories into more representative features, so that the clustering may give better results. Also, if more information about the town can be gathered, such as the university related features, household income of the town, etc., we can do better clustering on the university towns.
b. Another improvement can come from more accurate location data. Clearly some towns' location coordinates from the online resource are not very accurate, resulting in incomplete/wrong venue information.
c. If there is no limitation from *Foursquare* API (i.e. it is possible to obtain a more comprehensive exploration of the nearby venues), the clustering of the university towns can also be improved. Ways to collect more data using *Foursquare* need to be studied.

References

1. https://developer.foursquare.com/
2. University town list from Wikipeida, https://en.wikipedia.org/wiki/List_of_college_towns#College_towns_in_the_United_States
3. https://www.crummy.com/software/BeautifulSoup/
4. 2017 Population data for US cities, https://www.biggestuscities.com/
5. Location data for US cities, https://simplemaps.com/data/us-cities

6. https://geopy.readthedocs.io/en/stable/
7. https://pandas.pydata.org/
8. https://python-visualization.github.io/folium/
9. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.