Battle of the Neighborhoods

Which neighbourhood should I move to in the number 1 ranked county in New York, Tompkins?

Introduction

- When comparing neighbourhoods, people often compare cost of living, crime rates, quality of schools
- Those statistics are relatively easy to compare, but what about the differences in venues around each neighbourhood?
- I will be using the foursquare API to pull location data of venues surround each neighbourhood to compare neighborhoods.

Introduction

- - We aim to apply k-means-clustering and basic data analysis to the dataset of Tompkins neighborhoods to determine which neighbourhood suits you best.
- Using the Foursquare API and the K-means algorithm, we are able to obtain 4 clusters of neighborhoods with its associated venues.

Data

- The data has been obtained from https://simplemaps.com/data/us-cities.
- The columns given in the data set are :
  - State
  - County
  - Cities
  - Latitude, Longitude
  - Population Density
  - Timezones
  - County and State ID

Methodology

```
pd.read_csv("uscities.csv")
```

```
ead(10)
```

city	city_ascii	state_id	state_name	county_fips	county_name	county_fips_all	county_name_all	lat	lng	population	density
South Creek	South Creek	WA	Washington	53053	Pierce	53053	Pierce	46.9994	-122.3921	2500.0	125.0
Roslyn	Roslyn	WA	Washington	53037	Kittitas	53037	Kittitas	47.2507	-121.0989	947.0	84.0
Sprague	Sprague	WA	Washington	53043	Lincoln	53043	Lincoln	47.3048	-117.9713	441.0	163.0
Gig Harbor	Gig Harbor	WA	Washington	53053	Pierce	53053	Pierce	47.3352	-122.5968	9507.0	622.0
Lake Cassidy	Lake Cassidy	WA	Washington	53061	Snohomish	53061	Snohomish	48.0639	-122.0920	3591.0	131.0
Tenino	Tenino	WA	Washington	53067	Thurston	53067	Thurston	46.8537	-122.8607	1830.0	491.0
amestown	Jamestown	WA	Washington	53009	Clallam	53009	Clallam	48.1229	-123.0911	289.0	191.0
Three Lakes	Three Lakes	WA	Washington	53061	Snohomish	53061	Snohomish	47.9420	-121.9924	3390.0	112.0
Curlew Lake	Curlew Lake	WA	Washington	53019	Ferry	53019	Ferry	48.7311	-118.6663	573.0	50.0
hain Lake	Chain Lake	WA	Washington	53061	Snohomish	53061	Snohomish	47.9038	-121.9861	4280.0	156.0

- Data Exploration
  - We first take a look at the original dataset
  - There are way too many columns that we will not be using, so we will remove them and the segment the data to our Tompkins data.

Methodology

- Now this is a lot better.

Segmenting Tompkins County

In [49]:
```
tompkins_data = NY_data[NY_data.County == 'Tompkins']
tompkins_data.reset_index()
tompkins_data
```

Out[49]:

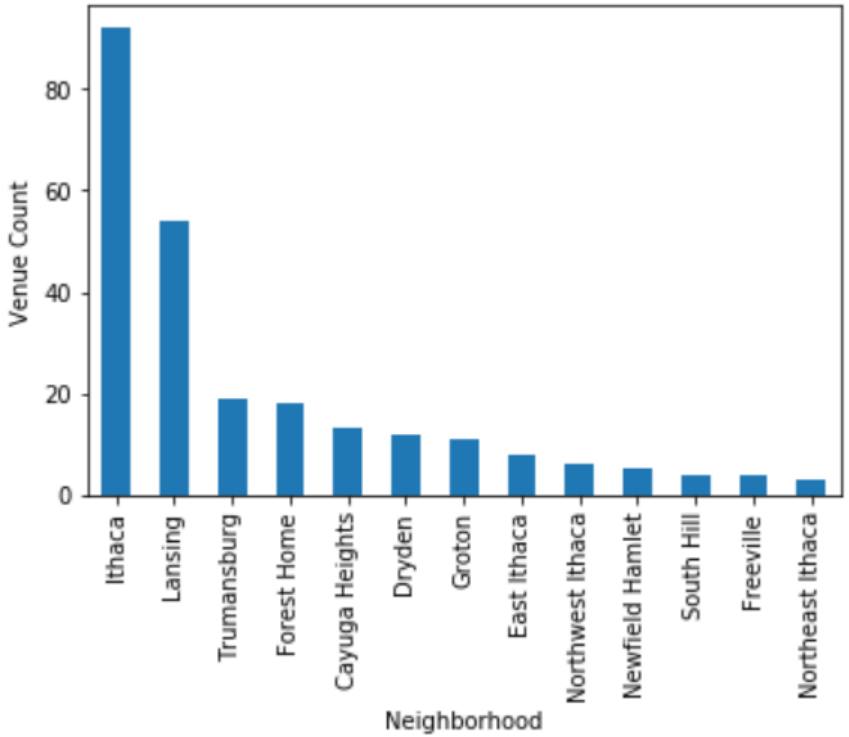
	Neighborhood	State	County	lat	lng	density
8860	South Hill	NY	Tompkins	42.4113	-76.4883	438.0
8981	Dryden	NY	Tompkins	42.4912	-76.2996	474.0
9012	Cayuga Heights	NY	Tompkins	42.4680	-76.4874	819.0
9048	Northeast Ithaca	NY	Tompkins	42.4703	-76.4623	837.0
9229	Lansing	NY	Tompkins	42.4901	-76.4856	299.0
9238	Trumansburg	NY	Tompkins	42.5410	-76.6618	505.0
9274	Forest Home	NY	Tompkins	42.4530	-76.4702	917.0
9390	East Ithaca	NY	Tompkins	42.4263	-76.4627	499.0
9470	Groton	NY	Tompkins	42.5875	-76.3630	524.0
9546	Newfield Hamlet	NY	Tompkins	42.3582	-76.5923	222.0
9759	Northwest Ithaca	NY	Tompkins	42.4706	-76.5414	156.0
9774	Freeville	NY	Tompkins	42.5114	-76.3457	187.0
9860	Ithaca	NY	Tompkins	42.4442	-76.5032	2221.0

Methodology

- Now that we have our neighbourhood data, we will use Foursquare API to return a list of the top 3 most common venues in a 1km radius of each neighbourhood.
- Based the venue data, we will classify each neighbourhood according to some similarity measures.

Methodology

- Using the value_counts function, we plot a bar chart of the total number of venues grouped by neighbourhood.



Methodology

- We then use one-hot encoding to classify the frequency of each type of venue for each neighbourhood.

```
kins_onehot.groupby('Neighborhood').mean().reset_index()
```

ries tore	American Restaurant	Arts & Crafts Store	Asian Restaurant	Automotive Shop	Bagel Shop	Bakery	Bar	Bed & Breakfast	...	Trail	Turkish Restauran
000	0.076923	0.000000	0.00000	0.00000	0.00000	0.076923	0.000000	0.000000	...	0.00000	0.0000
000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.000000	...	0.00000	0.0000
000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.000000	...	0.37500	0.0000
000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.000000	...	0.00000	0.0000
000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.000000	...	0.00000	0.0000
000	0.000000	0.000000	0.00000	0.00000	0.00000	0.090909	0.090909	0.000000	...	0.00000	0.0000
000	0.021739	0.010870	0.01087	0.01087	0.01087	0.010870	0.010870	0.021739	...	0.01087	0.0108
037	0.018519	0.018519	0.00000	0.00000	0.00000	0.000000	0.000000	0.000000	...	0.00000	0.0000
000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.200000	0.000000	...	0.00000	0.0000
000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.000000	...	0.00000	0.0000
000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.000000	...	0.00000	0.0000
000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.000000	...	0.00000	0.0000
000	0.105263	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.000000	...	0.00000	0.0000

Methodology

- We then use our k-means algorithm to cluster the neighbourhoods to 4 different clusters.

Neighborhood	State	County	lat	lng	density	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
South Hill	NY	Tompkins	42.4113	-76.4883	438.0	1	Concert Hall	Student Center	Park
Dryden	NY	Tompkins	42.4912	-76.2996	474.0	1	Pharmacy	Grocery Store	Gas Station
Cayuga Heights	NY	Tompkins	42.4680	-76.4874	819.0	3	Shopping Mall	Café	Gym / Fitness Center
Northeast Ithaca	NY	Tompkins	42.4703	-76.4623	837.0	1	Playground	Park	Shopping Mall
Lansing	NY	Tompkins	42.4901	-76.4856	299.0	0	Clothing Store	Shoe Store	Japanese Restaurant
Trumansburg	NY	Tompkins	42.5410	-76.6618	505.0	1	American Restaurant	Brewery	Golf Course
Forest Home	NY	Tompkins	42.4530	-76.4702	917.0	1	Food Truck	College Lab	Theater
East Ithaca	NY	Tompkins	42.4263	-76.4627	499.0	1	Trail	Cosmetics Shop	Park
Groton	NY	Tompkins	42.5875	-76.3630	524.0	1	Pizza Place	Gas Station	Gift Shop

Methodology

- We can then group the clusters and attach our labels to each cluster to determine the type of neighbourhood that each cluster contains.

Cluster 0 is by itself, mainly with fashion stores.

	State	density	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
8860	NY	438.0	1	Concert Hall	Student Center	Park
8981	NY	474.0	1	Pharmacy	Grocery Store	Gas Station
9048	NY	837.0	1	Playground	Park	Shopping Mall
9238	NY	505.0	1	American Restaurant	Brewery	Golf Course
9274	NY	917.0	1	Food Truck	College Lab	Theater
9390	NY	499.0	1	Trail	Cosmetics Shop	Park
9470	NY	524.0	1	Pizza Place	Gas Station	Gift Shop
9759	NY	156.0	1	Farm	Gift Shop	Museum
9860	NY	2221.0	1	Coffee Shop	Thai Restaurant	Bookstore

Cluster 1 mainly consists of entertainment and leisure venues.

	State	density	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
9546	NY	222.0	2	Yoga Studio	Deli / Bodega	Diner
9774	NY	187.0	2	Breakfast Spot	Ice Cream Shop	Sports Club

Methodology

- For the 2 remaining clusters.

	State	density	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
9546	NY	222.0	2	Yoga Studio	Deli / Bodega	Diner
9774	NY	187.0	2	Breakfast Spot	Ice Cream Shop	Sports Club

Cluster 2 mainly consists of eateries.

	State	density	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
9012	NY	819.0	3	Shopping Mall	Café	Gym / Fitness Center

Cluster 3 is once again by its lonesome, with a shopping mall, café, and gyms.

Cluster Result Discussion

- From our analysis, it is clear that it is not easy to come to an easy conclusion about which neighbourhood is best.
- Coming from Singapore myself, although my area does not have much good food and eateries which is important to me, I still enjoy living here, and do not mind commuting to other areas for food.
- Similarly, many people who go to prestigious schools do not live near the area, and come from other neighbourhoods to attend this particular school, because their neighbourhood is more affordable to live in.

Cluster Result Discussion

- However, this analysis is a useful beginner level analysis of neighbourhood data, and shows its potential. In the future, I will be able to use these skills of using the foursquare API to conduct better analyses with geographical data.