Introduction

Suppose a venue in Albany, NY wanted to open a second location elsewhere in New York State. The business owner would like to make the second location be similar to the current location. The owner wants to balance the population of the cities with the percentage of current similar venues that exist. This is to help ensure that there is enough population to support the business, as well as minimizing competition from similar venues.

Data

The data is primarily from two sources:

- Wikipedia list of cities in NY with population from 2010 census
  -https://en.wikipedia.org/wiki/List_of_cities_in_New_York_(state)
- Foursquare API with data about venues in the selected cities.

From wikipedia the list of cities in new york along with longitude, longitude and population for each was scraped. Then repeated calls to the Foursquare API were made to build the data set of all cities, their populations, and the venues within 3 km of the city center. 62 cities were generated spreading across NY state with blue pins at each city center. 4430 venues were generated across 327 unique categories to provide the data set.

```
Venue                  4430
There are 327 uniques categories.
```

Methodology

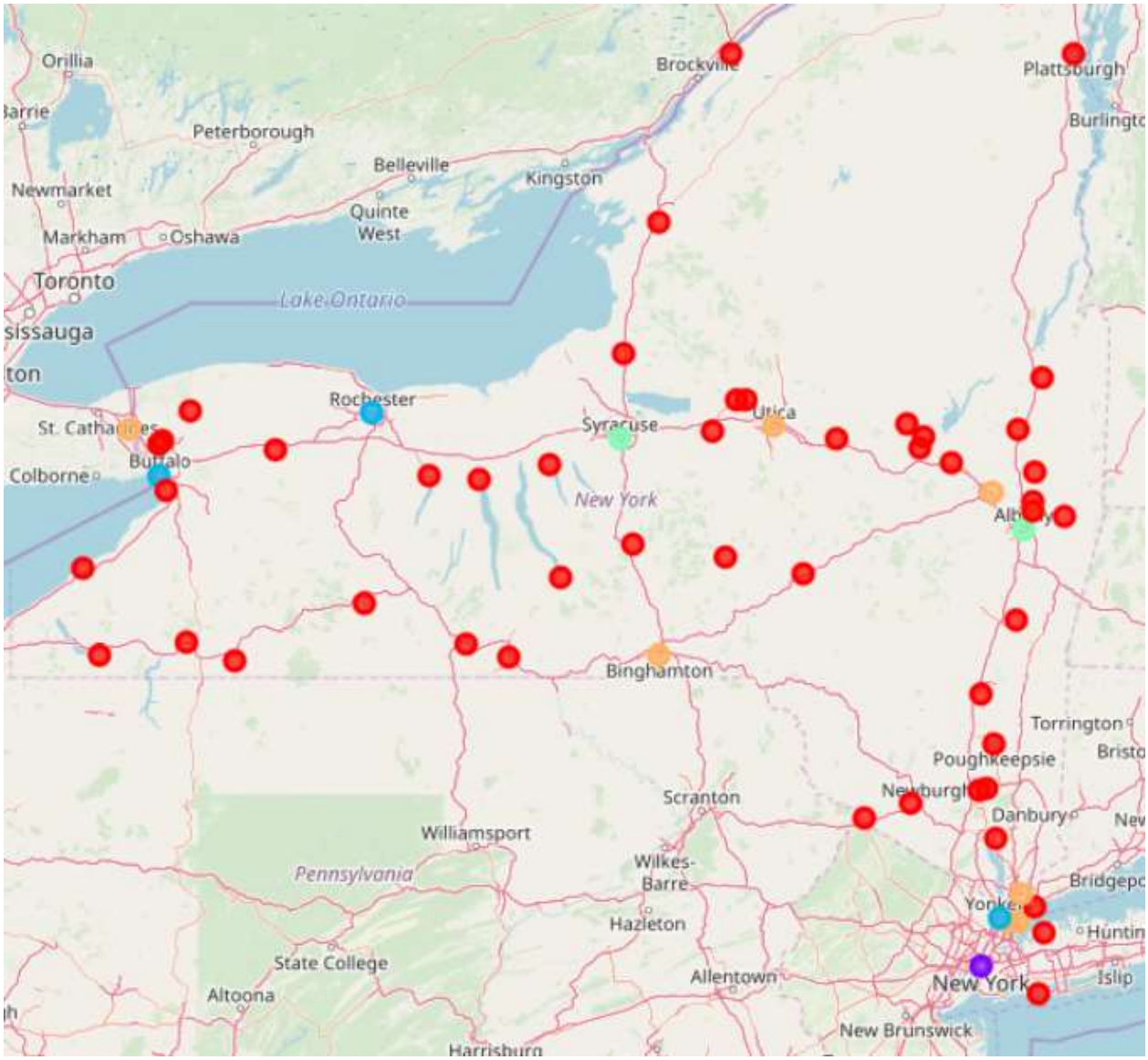
Now that the data was collected, a Onehot algorithm was used and organized by city and mean counts to group the venues. The shape of the dataframe was (62, 328). The top 10 venues in each city were sorted and grouped. An example of Albany, NY, indicates that no single venue really dominates the landscape.

```
----Albany----
                      venue  freq
0                       Café  0.07
1      American Restaurant  0.06
2                       Pub  0.05
3        Convenience Store  0.04
4                       Bar  0.03
5                      Park  0.03
6                   Theater  0.03
7                   Brewery  0.03
8            Deli / Bodega  0.03
9  New American Restaurant  0.02
```

Once the data was all collected and cleaned a k-means algorithm was used to split the cities into groups based on venues and populations within 3km of the city center. A k size of 5 was used as it provided the best results. K selection after 5 yielded little change in outcome.

Results

In the end it appears that most of the cities are pretty similar to each other. With clusters as such:



Cluster Labels	
0	48
1	1
2	3
3	2
4	8

Discussion

At first it appeared that the groupings were incorrect in some way, so I did a manual comparison of some of the 48 cities in cluster 0. And they do seem pretty similar to each other.

----Little Falls----

	venue	freq
0	Convenience Store	0.15
1	Beer Bar	0.08
2	Bakery	0.08
3	Grocery Store	0.08
4	Pizza Place	0.08
5	Bowling Alley	0.08
6	Sandwich Place	0.08
7	Fast Food Restaurant	0.08
8	Hotel	0.08
9	Discount Store	0.08

----Glens Falls----

	venue	freq
0	Convenience Store	0.10
1	Pharmacy	0.07
2	American Restaurant	0.06
3	Sandwich Place	0.06
4	Brewery	0.05
5	Discount Store	0.04
6	Supermarket	0.04
7	Pizza Place	0.03
8	Sushi Restaurant	0.03
9	Ice Cream Shop	0.03

----Saratoga Springs----

	venue	freq
0	Hotel	0.07
1	Italian Restaurant	0.06
2	American Restaurant	0.06
3	Coffee Shop	0.05
4	Pizza Place	0.04
5	Ice Cream Shop	0.04
6	Food & Drink Shop	0.03
7	Bar	0.03
8	Mexican Restaurant	0.03
9	Park	0.03

When compared to group 3 it becomes more obvious. That Syracuse is the best choice given the conditions of the tests.

----Syracuse----

	venue	freq
0	Coffee Shop	0.07
1	Italian Restaurant	0.06
2	Bakery	0.06
3	Pizza Place	0.05
4	Mexican Restaurant	0.04
5	American Restaurant	0.04
6	Hotel	0.04
7	Pub	0.03
8	Clothing Store	0.03
9	Bar	0.03

----Albany----

	venue	freq
0	Café	0.07
1	American Restaurant	0.06
2	Pub	0.05
3	Convenience Store	0.04
4	Bar	0.03
5	Park	0.03
6	Theater	0.03
7	Brewery	0.03
8	Deli / Bodega	0.03
9	New American Restaurant	0.02

Conclusion

Using data science methods alone isn't enough to make a fully informed business decision about opening a new branch of your current venue. It does however provide a good starting point for continued market and feasibility research. By narrowing the field

of cities to go to down to a few, it is easier to spend resources in a more efficient manner looking for the right location within the selected city.

Without these methods it would take a long time, guessing, and hunting and pecking around the map checking out different trends and venues.