Battle of the Neighborhoods

Best Place to Relocate in NC - A decision between Wake County and Mecklenburg County Neighborhoods.

IBM Data Science Capstone Project

Introduction

- When someone or a family is trying to find the best places to live, it's always a good idea to compare cities and if possible, to compare neighborhoods to see if its suites your taste.

- When thinking about the best place to live, lots of things are considered when trying to make a comparison between cities, towns, or neighborhoods. Some of these includes:
  - **Overall Comparison:** This is a comparison of the same factors for each city, resulting in having a general overview of the two cities.
  - **Crime Rates:** Here, the comparison is made to know the crime rates of two cities, then measures them both against the national statistics.
  - **Cost of Living and Salary Comparison:** This considers comparing salaries and cost of living within cities for a decision to be made
  - **Compare Schools:** This is helpful in finding the best school in a vicinity by doing a comparison between different places.
  - **Neighborhood Comparison:** This looks at neighborhood comparison and helps one choose the best place to live within any given city.

Introduction

- Objectives
  - To apply machine learning techniques (k-means) and data science principles to a US Cities dataset to help individuals and families make a decision on where to move or relocate to.
- Study Area
  - After an active exploration of the data, we chose to focus our efforts on neighborhoods in North Carolina, particularly on Wake County and Mecklenburg County should one decide to make a relocation choice between those two places.
  - Through a combination of Folium, Foursquare API and the K-means algorithm, we were able to obtain 4 cluster of neighborhoods with its associated venues.

Data

- The simplemaps website has maintained a large database of the United States Cities available at https://simplemaps.com/data/us-cities.
- From there, we can access information about the
  - State,
  - County,
  - Cities,
  - Latitude and Longitude,
  - Pop. Density,
  - time Zones as well as the
  - County and
  - State ID.
- With all this information we can proceed and see what insight can be extracted from the data.

Methods

- **Data Exploration**
  - First, US Cities data, which is in a .json format, can be easily imported into pandas , plotted and viewed using folium and geolocator.

```
df = pd.read_json(json_data, orient='columns')
```

```
df.head(10)
```

	City Name	County Code	County Name	Density	Id	Latitude	Longitude	Source	State Id	State Name	Timezone
0	Prairie Ridge	53053	Pierce	1349.8	1840037882	47.1443	-122.1408	polygon	WA	Washington	America/Los_Angeles
1	Edison	53057	Skagit	127.4	1840017314	48.5602	-122.4311	polygon	WA	Washington	America/Los_Angeles
2	Packwood	53041	Lewis	213.9	1840025265	46.6085	-121.6702	polygon	WA	Washington	America/Los_Angeles
3	Wautauga Beach	53035	Kitsap	261.7	1840037725	47.5862	-122.5482	point	WA	Washington	America/Los_Angeles
4	Harper	53035	Kitsap	342.1	1840037659	47.5207	-122.5196	point	WA	Washington	America/Los_Angeles
5	Telma	53007	Chelan	1.4	1840037825	47.8432	-120.8157	point	WA	Washington	America/Los_Angeles
6	Kahlotus	53021	Franklin	156.0	1840018478	46.6436	-118.5566	polygon	WA	Washington	America/Los_Angeles
7	Mondovi	53043	Lincoln	9.4	1840037576	47.6813	-118.0164	point	WA	Washington	America/Los_Angeles
8	Washtucna	53001	Adams	108.0	1840022312	46.7539	-118.3104	polygon	WA	Washington	America/Los_Angeles
9	Pleasant Hill	53033	King	25.8	1840037805	47.6154	-121.9096	point	WA	Washington	America/Los_Angeles

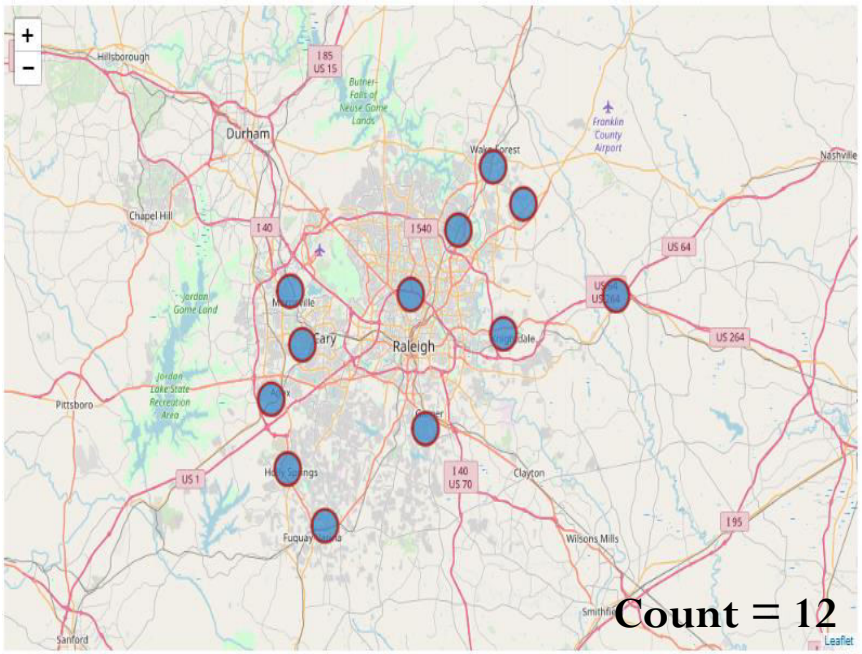
Methods

- Slicing and segmenting Wake and Mecklenburg Data
  - There are lots of columns. But for our purpose, we only need 'City Name', 'County Name', 'State Name', 'Density', 'Latitude,' and 'Longitude.'
  - After dropping unwanted columns, checking and cleaning up NaNs, the Pop. The data can be plotted and viewed as above.

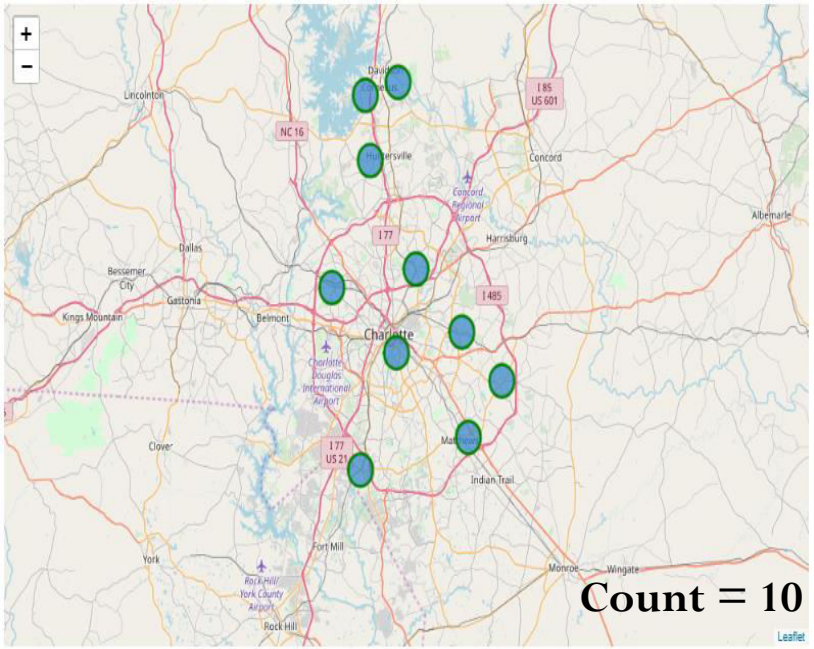
Methods

- Neighborhoods in NC Counties

The map did not give any information about the venues



Wake County



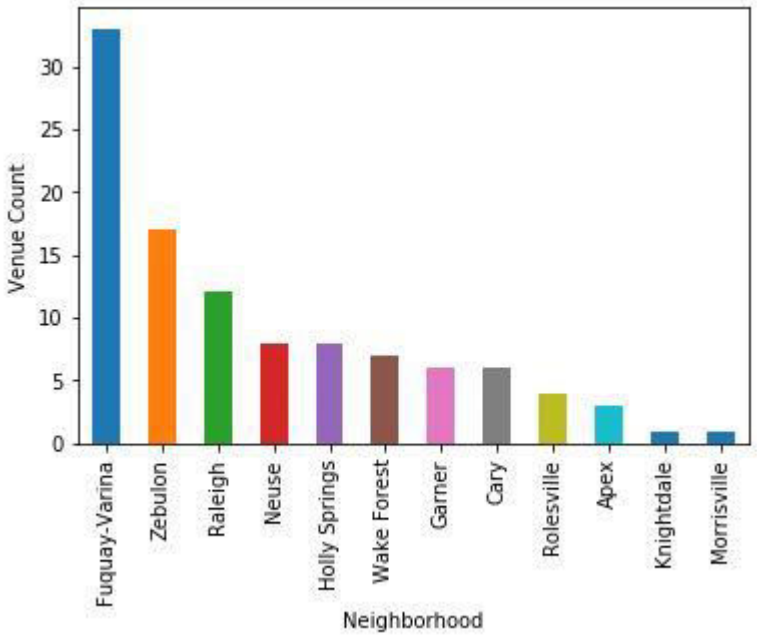
Mecklenburg County

Methods

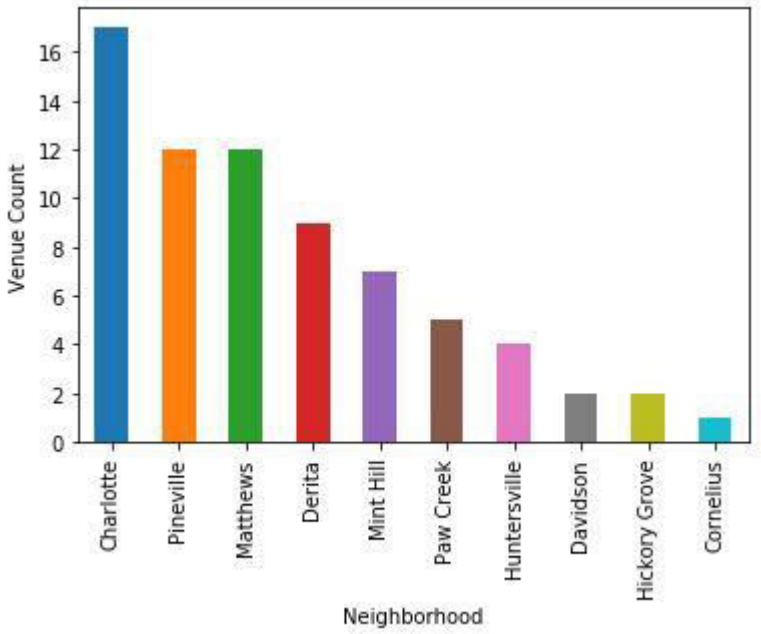
- Now that we have the neighborhood maps, next we are going to study these neighborhoods using Foursquare APIs.
  - For each neighborhood, the Foursquare search engine returns a list of top ten most common venues.
  - Based on the venue data, the neighborhoods can be clustered according to some similarity measures. Results from the Foursquare Explore API call are the following:

Methods

- Using the value_counts function, a bar chart of the total number of venues grouped by neighborhood is plotted as follows:



Wake County



Mecklenburg County

Methods

- While the above bar charts are useful when comparing different neighborhood venue categories, it would be nice if we could put these data on a map using their spatial location coordinates.

- This could be achieved through using folium package.

- We first form a Shapely geometry object by combining ‘Latitude’ and ‘Longitude’ columns and merge it with the original dataframe to form a GeoDataFrame:

Methods

- One-hot encoding
  - Prior to forming the GeoDataFrame, one-hot encoding was performed on the venues to convert the venues into categorical data grouped by neighborhood.
    - This is very important to aid in clustering of the neighborhood venues.

Methods

```
wake_grouped = wake_onehot.groupby('Neighborhood').mean().reset_index()
wake_grouped
```

	Neighborhood	American Restaurant	Antique Shop	Asian Restaurant	Athletics & Sports	BBQ Joint	Baseball Field	Basketball Court	Big Box Store	Boutique	Breakfast Spot	Brewery	Bus Station
0	Apex	0.000000	0.000	0.000000	0.000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000000
1	Cary	0.000000	0.000	0.000000	0.000	0.000000	0.166667	0.0	0.000000	0.000000	0.166667	0.000	0.166667
2	Fuquay-Varina	0.090909	0.000	0.030303	0.000	0.000000	0.000000	0.0	0.000000	0.030303	0.030303	0.000	0.000000
3	Garner	0.000000	0.000	0.000000	0.000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000000
4	Holly Springs	0.000000	0.000	0.000000	0.000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.125	0.000000
5	Knightdale	0.000000	0.000	0.000000	0.000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000000
6	Morrisville	0.000000	0.000	0.000000	0.000	0.000000	0.000000	1.0	0.000000	0.000000	0.000000	0.000	0.000000
7	Neuse	0.000000	0.125	0.000000	0.125	0.125000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000000
8	Raleigh	0.083333	0.000	0.000000	0.000	0.000000	0.000000	0.0	0.083333	0.000000	0.000000	0.000	0.000000
9	Rolesville	0.000000	0.000	0.000000	0.000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000000
10	Wake Forest	0.000000	0.000	0.000000	0.000	0.000000	0.000000	0.0	0.000000	0.000000	0.142857	0.000	0.000000
11	Zebulon	0.000000	0.000	0.000000	0.000	0.058824	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000000

```
mecklenburg_grouped = mecklenburg_onehot.groupby('Neighborhood').mean().reset_index()
mecklenburg_grouped
```

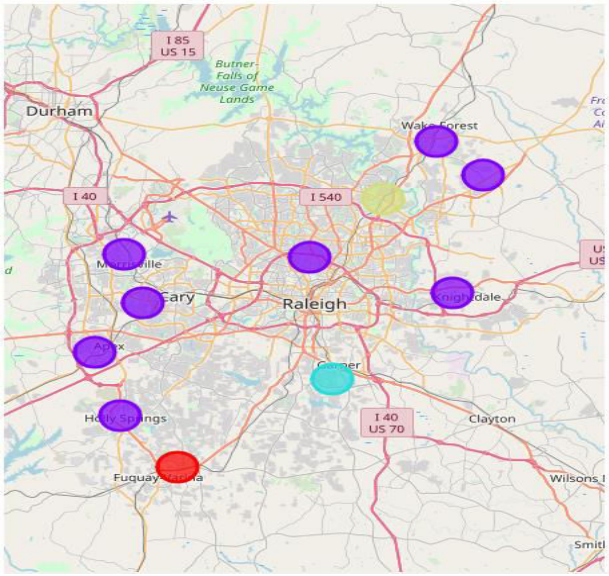
	Neighborhood	American Restaurant	Bakery	Bar	Baseball Field	Basketball Court	Beer Bar	Beer Garden	Big Box Store	Breakfast Spot	Brewery	Business Service	Café	
0	Charlotte	0.000000	0.117647	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00	0.058824	0.058824	0
1	Cornelius	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0
2	Davidson	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0
3	Derita	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0
4	Hickory Grove	0.000000	0.000000	0.000000	0.000000	0.5	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0
5	Huntersville	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.25	0.000000	0.000000	0
6	Matthews	0.083333	0.000000	0.000000	0.083333	0.0	0.000000	0.000000	0.083333	0.083333	0.00	0.000000	0.000000	0
7	Mint Hill	0.000000	0.000000	0.142857	0.000000	0.0	0.142857	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0
8	Paw Creek	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0
9	Pineville	0.000000	0.000000	0.000000	0.000000	0.0	0.083333	0.083333	0.000000	0.000000	0.00	0.000000	0.000000	0

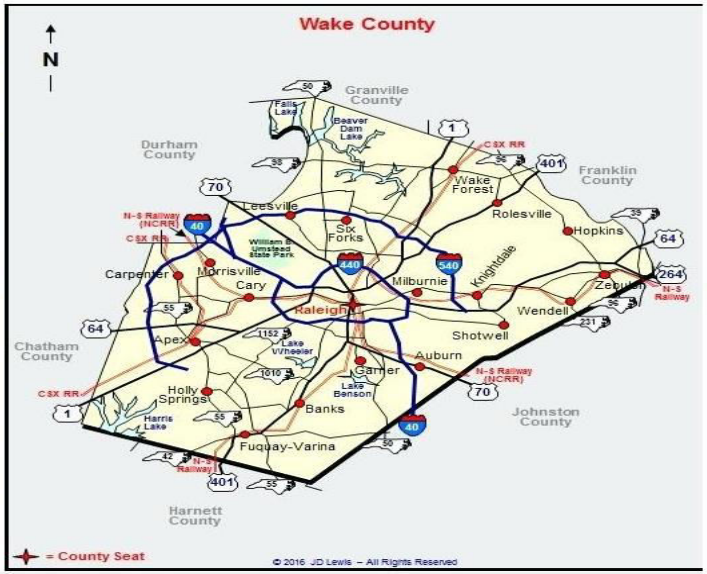
Method

- Cluster-kmeans
  - To cluster the neighborhoods, we simply apply the k-means algorithm to the one-hot encoded venue dataset.
  - This was done for both counties
  - Assuming there is 4 different clusters, the table and map looks like this

Cluster Results

	Neighborhood	County	Density	Latitude	Longitude	State	Cluster Labels	Cluster Colors
37225	Neuse	Wake	1314.1	35.8974	-78.5692	NC	3	Yellow
32748	Zebulon	Wake	472	35.8311	-78.3185	NC	1	Purple
32750	Rolesville	Wake	662	35.9249	-78.4654	NC	1	Purple
32751	Knightdale	Wake	893	35.7921	-78.4968	NC	1	Purple
32752	Morrisville	Wake	1135	35.8359	-78.8349	NC	1	Purple
32753	Fuquay-Varina	Wake	772	35.5956	-78.7801	NC	0	Red
32754	Garner	Wake	737	35.6949	-78.6212	NC	2	Blue
32755	Holly Springs	Wake	825	35.6544	-78.8392	NC	1	Purple
32756	Wake Forest	Wake	971	35.963	-78.5144	NC	1	Purple
32757	Apex	Wake	1057	35.7248	-78.866	NC	1	Purple
32758	Cary	Wake	1128	35.7815	-78.8162	NC	1	Purple
32759	Raleigh	Wake	1225	35.8323	-78.6441	NC	1	Purple

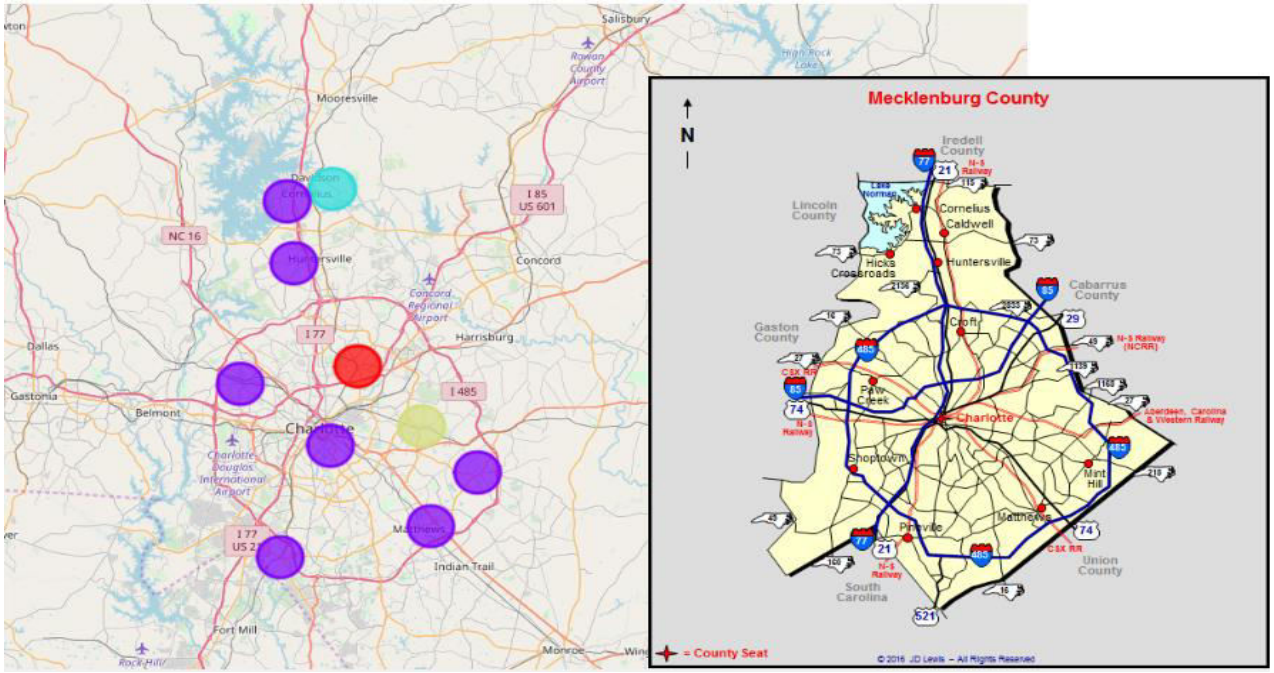




Map of clustered neighborhoods in Wake County. The results are obtained using k-means algorithm.

Cluster Results

	Neighborhood	County	Density	Latitude	Longitude	State	Cluster Labels	Cluster Colors
7182	Paw Creek	Mecklenburg	533	35.2749	-80.9384	NC	1	Purple
7183	Hickory Grove	Mecklenburg	992.4	35.2288	-80.7206	NC	3	Yellow
7184	Derita	Mecklenburg	1123.7	35.2938	-80.7976	NC	0	Red
32556	Pineville	Mecklenburg	500	35.0864	-80.8915	NC	1	Purple
32557	Davidson	Mecklenburg	835	35.4861	-80.8272	NC	2	Blue
32558	Mint Hill	Mecklenburg	424	35.1781	-80.6538	NC	1	Purple
32559	Cornelius	Mecklenburg	951	35.4733	-80.8833	NC	1	Purple
32560	Matthews	Mecklenburg	710	35.1196	-80.7101	NC	1	Purple
32561	Huntersville	Mecklenburg	530	35.4055	-80.8741	NC	1	Purple
32562	Charlotte	Mecklenburg	1065	35.208	-80.8308	NC	1	Purple



Map of clustered neighborhoods in Wake County. The results are obtained using k-means algorithm.

Cluster Results

- Top 10 venues
  - Again, we can print the top 10 venues in each neighborhood.
  - This can give us insight about the most common venues within each neighborhood within the various counties.

Cluster Result

Details of the purple clusters shown in the above maps.

- Wake County

	County	State	Cluster Labels	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 Com V
32748	Wake	NC	1	Fast Food Restaurant	Pizza Place	Video Store	Department Store	Pharmacy	Sandwich Place	Supermarket	Steakhouse	BBQ Joint	Optica Shop
32750	Wake	NC	1	Pizza Place	Sandwich Place	Restaurant	Gym Pool	Athletics & Sports	Cosmetics Shop	Gift Shop	Gas Station	Furniture / Home Store	Food Drink Shop
32751	Wake	NC	1	Park	Yoga Studio	Cosmetics Shop	Gift Shop	Gas Station	Furniture / Home Store	Food & Drink Shop	Fast Food Restaurant	Farmers Market	Farm
32752	Wake	NC	1	Basketball Court	Yoga Studio	Cosmetics Shop	Gym	Gift Shop	Gas Station	Furniture / Home Store	Food & Drink Shop	Fast Food Restaurant	Farme Marke
32755	Wake	NC	1	Pizza Place	Pharmacy	Ice Cream Shop	Brewery	Restaurant	Gym	Gas Station	Furniture / Home Store	Food & Drink Shop	Fast F Resta
32756	Wake	NC	1	Gas Station	Supermarket	Other Repair Shop	Breakfast Spot	Italian Restaurant	Pharmacy	Hobby Shop	Fast Food Restaurant	Farmers Market	Food Drink Shop
32757	Wake	NC	1	Health & Beauty Service	Farm	Playground	Asian Restaurant	Athletics & Sports	Gift Shop	Gas Station	Furniture / Home Store	Food & Drink Shop	Fast F Resta
32758	Wake	NC	1	Dance Studio	Breakfast Spot	Gym	Baseball Field	Chinese Restaurant	Bus Station	Discount Store	Dog Run	Donut Shop	Farm
32759	Wake	NC	1	American Restaurant	Spa	Gym	Hotel	Italian Restaurant	Coffee Shop	Big Box Store	Smoothie Shop	Pharmacy	Furnit Home Store

Cluster Result

Details of the purple clusters shown in the above maps.

- Mecklenburg County

	County	State	Cluster Labels	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 Com V
7182	Mecklenburg	NC	1	Chinese Restaurant	Discount Store	Pizza Place	Restaurant	Gas Station	Women's Store	Fried Chicken Joint	Dry Cleaner	Donut Shop	Diner
32556	Mecklenburg	NC	1	Beer Garden	Beer Bar	Pool Hall	Latin American Restaurant	Indian Restaurant	Deli / Bodega	Shoe Store	Pharmacy	Hobby Shop	Diner
32558	Mecklenburg	NC	1	Gym	Beer Bar	Restaurant	Sandwich Place	Shipping Store	Pizza Place	Bar	Discount Store	Diner	Came Store
32559	Mecklenburg	NC	1	Ice Cream Shop	Women's Store	Gas Station	Fried Chicken Joint	Dry Cleaner	Donut Shop	Discount Store	Diner	Deli / Bodega	Cosm Shop
32560	Mecklenburg	NC	1	Gym	Pharmacy	Baseball Field	Big Box Store	Breakfast Spot	Construction & Landscaping	Dry Cleaner	Video Store	Martial Arts Dojo	Ameri Resta
32561	Mecklenburg	NC	1	Pool	Kids Store	Brewery	Women's Store	Chinese Restaurant	Fried Chicken Joint	Dry Cleaner	Donut Shop	Discount Store	Diner
32562	Mecklenburg	NC	1	Bakery	Women's Store	Residential Building (Apartment / Condo)	Convenience Store	Ice Cream Shop	Indian Restaurant	Camera Store	Men's Store	Mobile Phone Shop	Café

Cluster Result Discussion

- Apparently, a lot of the neighborhoods are in the purple cluster for both Wake and Mecklenburg County. When we look at the purple cluster for Wake County, it becomes clear that the first two most common venues in the neighborhoods contain a lot of mixed amenities: Fast Food Restaurant, Pizza Place, Park, Basketball Court, Gas Station, Health & Beauty Service, Dance Studio, American Restaurant, Pharmacy, and SPA.

- Again, looking at that of Mecklenburg County, we have: Chinese Restaurant, Beer Garden, Gym, Ice Cream Shop, Pool, Bakery, Discount Store, Women's Store, Pharmacy, Kids store, Brewery, and Pool Hall. So, the question is where should someone considering relocating move to a new neighborhood given the choice between Wake and Mecklenburg County? Well, by looking at the two neighborhood maps, it appears that the anyone not a fun of beer and does not want to expose his or her children to alcohol would prefer moving to Wake County since there are lots on beer and brewery venues in the neighborhoods in Mecklenburg County.

- However, decision is left to the individual looking at relocating to make. But in general, though all these analyses are useful, there is nothing like visiting the actual city, seeing the neighborhoods, and speaking with residents. If it's possible, an in-person visit is highly recommended before making a big move.

Conclusion

- My analysis has shown that using folium- python library that assists in building a quick interactive data visualization and Foursquare API for neighborhood data collection, it is feasible to cluster neighborhoods cities data based on known and accepted machine learning techniques –K-Means Algorithm.
- These results must be considered bounded in scope to the dataset used, since there is no information available as to its provenance. Such results will be of interest to people or citizens whose aim to compare different neighborhoods when thinking about relocation or vacationing in a different environment, considering the ease of accessing numerous venues within a clustered setting.
- There certainly is lot of room for improvement. For example, obtaining more than the current neighborhood locations to analyze and cluster a wide expanse of geographical setting.
- We may also use and analyze crime data – which is publicly available for his two counties - to help to provide enough room for decision making with regards to choosing a location to relocate.
- This information may be extremely useful because we certainly don't want to live in a crime infested neighborhood. Though the approach used here may not be vigorous enough, it nevertheless showcases the usefulness of neighborhood data analysis.