By HARSHAVARDHAN,

Battle of the Neighbourhoods, North Carolina – Looking for the Best Place to Relocate.

1. 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 neighbourhoods to see if its suites your taste. After all, when you go to buy a car or a house or any big-ticket item, you usually try out a few models or visit a few homes before you decide. The same tactic applies to finding the best places to live. It is always advisable to do it before you start planning your move or to help narrow down your choices.

When thinking about the best place to live, lots of things are considered when trying to make a comparison between cities, towns, or neighbourhoods. 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. Some of the popular factors include population, cost of living, average rent, crime rate, tax rates, and air quality.

- **Neighbourhood Comparison:** This looks at neighbourhood comparison and helps one choose the best place to live within any given city. These sites allow you to see some interesting facts about the various communities.

- **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. Some factors for this comparison include statistics on food, housing, utilities, transportation and more. This is a useful way to find out if your salary will measure up in the new city.

- **Compare Schools:** This is helpful in finding the best school in a vicinity by doing a comparison between different places. It mostly takes into consideration test scores and teacher and student ratios, including the teacher's experience of the least schools in the city of your choice.

The data set includes the coordinates of the cities/neighbourhoods in the USA. If we had the venue information, we could easily find out more information about the neighbourhoods. For example, how many restaurants are there, are there parks or cinemas? What about banks and glossary stores? If all this information is known, we could better understand or make educated decision about where to move or relocate to. Hence, the purpose of this project is to, algorithmically, find a way to use the location coordinates and tag each data point into a neighbourhood in two Counties in North Carolina-Wake County and Mecklenburg County. The algorithm used is k-means clustering. The main idea is to determine neighbourhood with venues clustered around each other so that one can decide on the right neighbourhood to choose based on the proximity of amenities and venues to each other.

2. DATA

2.1. Background of the data

The dataset for this project consists of information regarding the cities in the USA obtained from https://simplemaps.com/data/us-cities. Specifically, the data contain: City Name, County Code, County Name, Density, Id, Latitude, Longitude, Source, State Id, State Name, and Time zone. Though this data came with the coordinates, Tableau was used for geocoding the data to obtain the correct coordinates. The data was then exported and converted into a excel csv file, read into a pandas data frame and sliced into Wake and Mecklenburg data for use in the project. Besides this data, the Foursquare API was be used to collect venues near the neighbourhoods for cluster analysis to be performed on the data.

3. METHODOLOGY

3.1. Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the downloaded data. This was used to:

a. Segment the data into Cities in Wake and Mecklenburg Counties in North Carolina

b. Identify missing values, verify the quality of the data

c. Determine likely approaches to modelling, which might best yield to good clustering.

3.2. Segmenting and Slicing, and visualizing the data

An important part of cluster modelling is the careful selection of the variables of available data. A prerequisite of the study is that the foursquare API is used to collect the venue information. Hence it is very important that the dataset for this work includes the coordinates of the cities to be studied. Segmentation and slicing of the data resulted in Table 1. The subjects included in the data for analysis includes: Neighbourhood Name, County Name, Density, Latitude, Longitude, and State Name.

Table 1. Segmented and sliced data

Name of dataset	Subjects included	No. of Rows	No. of Columns
Wake County	0,2,3,5,6,8	6	16
Mecklenburg County	0,2,3,5,6,8	6	10

To view the sliced data for both counties, folium was used. One may ask what folium is. Folium is a powerful python library that builds on the data wrangling strengths of the python ecosystem and the mapping strengths of the Leaflet.js library. Generally, data is manipulating in Python, and then visualize it in on a Leaflet map via Folium. Hence to visualize the data in folium, the coordinate of a location in Wake County was obtained and then looped through the rest of the neighborhoods and plotted to view the location on a map. This was also done for the Mecklenburg County data.

3.3. Neighborhood Exploration and Cluster – Wake and Mecklenburg County

For the neighbourhood exploration, the Foursquare API was used. The get request was deployed on the Foursquare URL to get the category types of venues limiting the number of venues to 100 within a 500 radius.Because the aim of the project is to determine the cluster of venues in the neighborhoods, one-hot encoding was performed on the venue categories to get dummies for each venue. That is to say, the venues were coded into 0s and 1s. The result was then grouped by neighborhood by taking the mean of the frequency of occurrence of each category

3.4. Cluster of Neighborhoods in Wake County.

For the clustering of venues categories in the neighborhoods, the k-means cluster was employed. to cluster the neighborhood into 4 clusters. The k-means clustering machine learning algorithm is an unsupervised clustering technique searches for a pre-determined number of clusters within an unlabeled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like:

- The "cluster center" is the arithmetic mean of all the points belonging to the cluster.
- Each point is closer to its own cluster center than to other cluster centers in the dataset.

The two assumptions above are presumably the basis of the k-means model. To be able to produce the clusters and visualize it on a map, the sliced wake and Mecklenburg county data were merged with the grouped venue data. This was done so that the coordinates form the sliced data can aid in visualizing the clusters on a map.

4. RESULT

The source .json data contained a total of 36,651rows and 11 columns. The sliced data for the two counties came out with 12 rows and 6 columns for Wake County, and 10 rows and 6 columns for Mecklenburg County as shown in Table 2 and 3 below. This makes it easy for the data to be easily analyzed.

Table 2. Wake County data

	Neighborhood	County	Density	Latitude	Longitude	State
7225	Neuse	Wake	1314.1	35.8974	-78.5692	NC
32748	Zebulon	Wake	472	35.8311	-78.3185	NC
32750	Rolesville	Wake	662	35.9249	-78.4654	NC
32751	Knightdale	Wake	893	35.7921	-78.4968	NC
32752	Morrisville	Wake	1135	35.8359	-78.8349	NC
32753	Fuquay-Varina	Wake	772	35.5956	-78.7801	NC
32754	Garner	Wake	737	35.6949	-78.6212	NC
32755	Holly Springs	Wake	825	35.6544	-78.8392	NC
32756	Wake Forest	Wake	971	35.963	-78.5144	NC
32757	Apex	Wake	1057	35.7248	-78.866	NC
32758	Cary	Wake	1128	35.7815	-78.8162	NC
32759	Raleigh	Wake	1225	35.8323	-78.6441	NC

Table 3. Mecklenburg County data

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

Using the Foursquare API, the venues within the neighborhoods in both wake and Mecklenburg counties resulted in a vast number of outcomes. The radius defined for the venue returned venues with 106 rows and 7 columns for Wake County and 71 rows and 7 columns for Mecklenburg County. The one-hot encoding produced a total number of 106, 63 and 71, 53 rows and columns for wake

and Mecklenburg counties respectively. Table 4 and 5 shows the results of the top 3 venues in each neighborhood venues for both wake and Mecklenburg County was grouped by neighborhood.

Table 4. Top 3 venues in each neighborhood in Wake County

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Apex	Health & Beauty Service	Farm	Playground
1	Cary	Dance Studio	Breakfast Spot	Gym
2	Fuquay-Varina	American Restaurant	Mexican Restaurant	Sandwich Place
3	Garner	Park	Yoga Studio	Coffee Shop
4	Holly Springs	Pizza Place	Pharmacy	Ice Cream Shop
5	Knightdale	Park	Yoga Studio	Cosmetics Shop
6	Morrisville	Basketball Court	Yoga Studio	Cosmetics Shop
7	Neuse	Dance Studio	Farmers Market	Antique Shop
8	Raleigh	American Restaurant	Spa	Gym
9	Rolesville	Pizza Place	Sandwich Place	Restaurant
10	Wake Forest	Gas Station	Supermarket	Other Repair Shop
11	Zebulon	Fast Food Restaurant	Pizza Place	Video Store

Table 5. Top 3 venues in each neighborhood in Mecklenburg County

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Charlotte	Bakery	Women's Store	Residential Building (Apartment / Condo)
1	Cornelius	Ice Cream Shop	Women's Store	Gas Station
2	Davidson	Cosmetics Shop	Construction & Landscaping	Women's Store
3	Derita	Pharmacy	Donut Shop	Video Store
4	Hickory Grove	Track	Basketball Court	Women's Store
5	Huntersville	Pool	Kids Store	Brewery
6	Matthews	Gym	Pharmacy	Baseball Field
7	Mint Hill	Gym	Beer Bar	Restaurant
8	Paw Creek	Chinese Restaurant	Discount Store	Pizza Place
9	Pineville	Beer Garden	Beer Bar	Pool Hall

Since the aim of the project is to cluster the neighborhoods, the k-means algorithm is applied to the onehot encoded venue dataset, assuming there are 4 different clusters. The tables below show the neighborhood and the cluster labels assigned to it after the k-means algorithm was applied. Cluster label '0' represents the 1st cluster and '3' the 4th cluster. This series of plots shows the data for each pair of variables with different clusters shown with different cluster plotting symbols on the maps in Figure 1 and 2.

Table 6. Neighborhood Clusters for Wake County

	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

Table 7. Neighborhood Clusters for MecklenburgCounty

	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

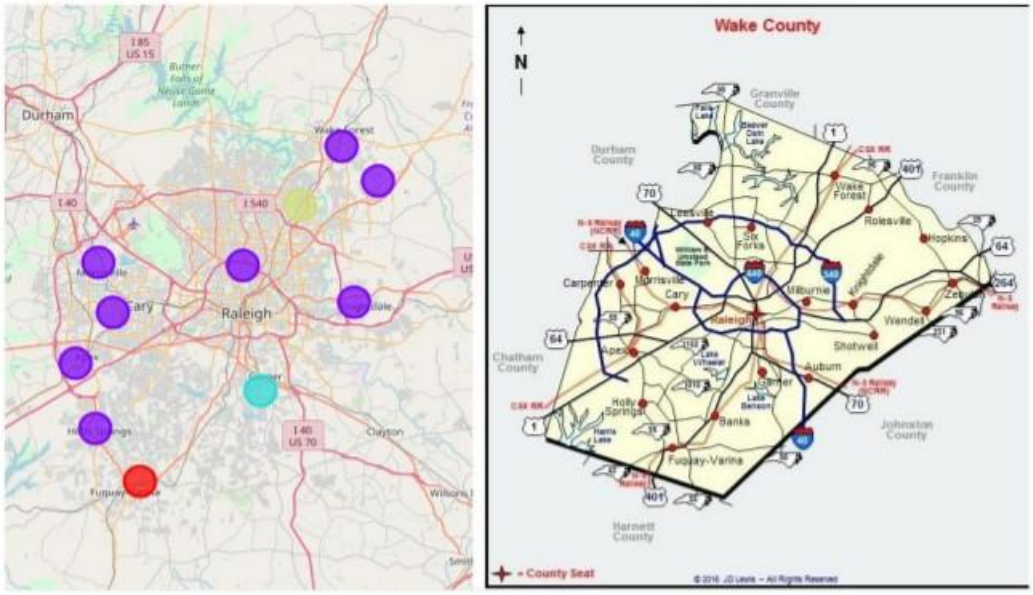

Figure 1. Neighborhood Clusters for Wake County

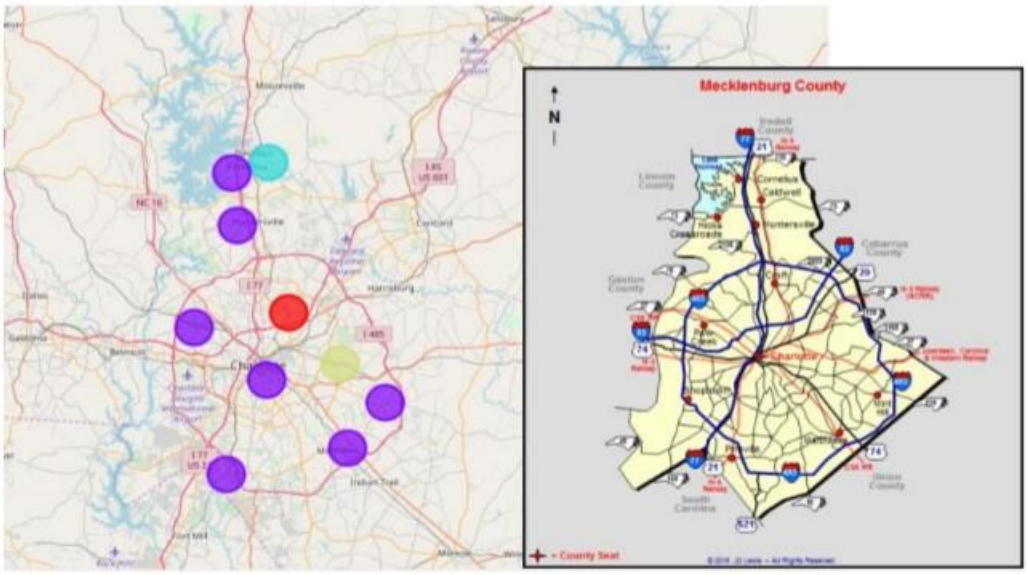

Figure 2. Neighborhood Clusters for Mecklenburg County

Examining each cluster for the various neighborhoods in the analyzed counties, it was determined that some discriminating venue categories were distinguished each cluster. Based on the defining categories, the following names were assigned to each cluster. Since 10 common venues were defined in this work, the assigned names were based only on the 2 st common venues for ease of name assignment.

5. 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 neighbourhood 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.

6. CONCLUSION

The aim of this work is to provide the necessary amenities to help people decide on the best to live or relocate to should they think about that. Using public datasets obtained from the web, iwas able to, the address a few factors by analyzing the neighborhoods within two major Counties in North Carolina, Wake and Mecklenburg by based on the spatial distribution of venues in the chosen neighborhoods. 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, obtainingmore 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.