Wake County ,North Carolina.

Battle of the neighborhoods

Vinil Beeravolu

Sr Data Engineer

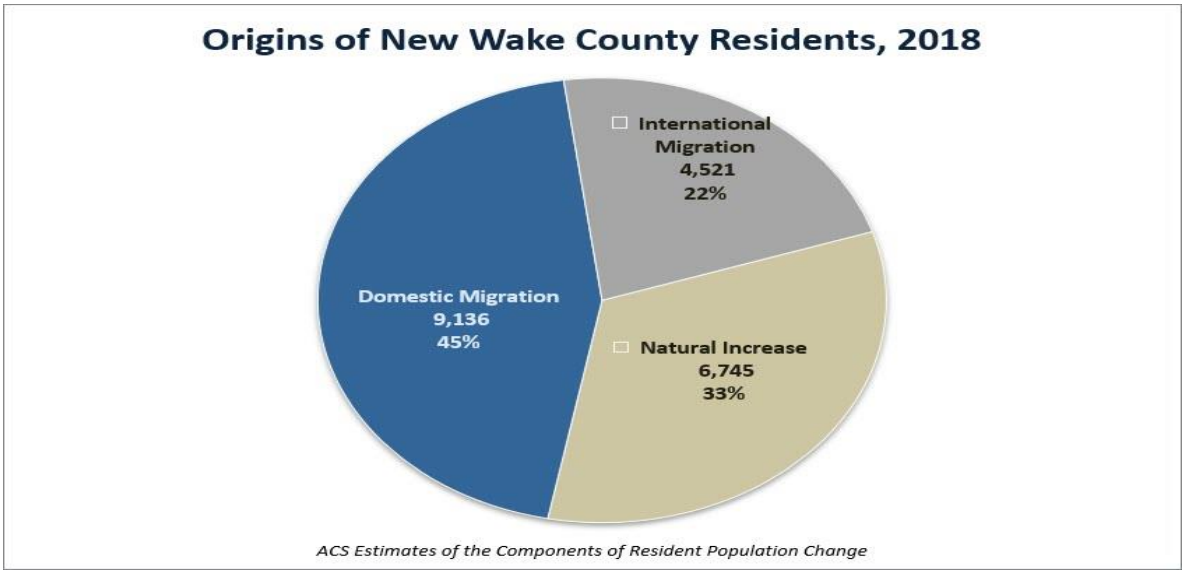
- **Introduction: Which neighborhoods are best for families with children who are migrating to Wake county?**

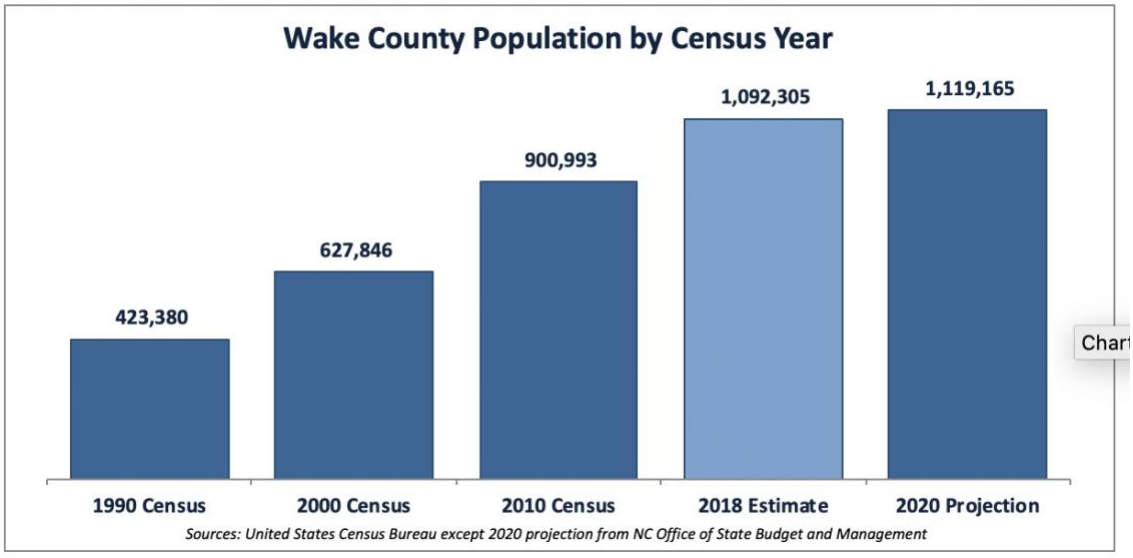
WAKE COUNTY

NORTH CAROLINA

Wake County is the most populous county in the state of North Carolina and it has some of the best rated cities/towns to live in, and it is one of the best schooling systems out there.

Because of low real estate prices compared to big cities , booming local economy and with a high standard of living , there has been a recent spike in domestic and international migration to this county. According to the recent estimates ,on average 70 people move to this county on a daily basis . Below is a pie chart form the county website showing the breakdown of the origins of new residents in 2018. I had a hard time picking the right place to rent/buy when I moved here with my family. This project is to help new comers to this county, especially families with school age children find the best neighborhoods suitable for their family. I will be bringing the Public Schools data into the mix to make this more oriented towards our goal to find the best spots for families moving with school going kids. This data can be very useful for real estate developers , investors , businesses and city planners to assess family friendliness of a neighborhood and invest .





- **Data**

a. Data Sources

Data for this project consists of Zoning data for the county, Public schools information, Population data for the neighborhoods and location data for the neighborhoods and finally Foursquare data for venues.

Wake County launched its “Open Data” program in January 2014 and provides various data sets for financial transactions, restaurant inspections, permitting, real estate, elections ,schools and property .

Zoning data and Public School data are available here at the Wake County Open Data Portal. The data came in as a Json file that contained different zip codes and city names associated with those zip codes. The data was unpacked using the read_json function from the Pandas library and looping through different levels of the data to gather the features necessary from this data set and stored it in a data frame. There is lot of other information that is not relevant to this project, so I was deleted. This But, what this data is lacking is the geo information (latitude and longitude) that is needed for the Foursquare app to gather venue data. To gather this missing geo data, I have used “geolocator.geocode “ from the python geopy library. This provides the latitude and longitude for each of the zip code in the zoning data. Here is a sample of the neighborhood data after gather the geo info.

	ZIPCODE	CITY	location	geo
0	27502	Apex	(Friendship, Apex, Wake County, North Carolina...	(35.7289032, -78.89319159574836)
1	27511	Cary	(Kildaire Farms, Raleigh, Wake County, North C...	(35.7564404, -78.781422)
2	27513	Cary	(Weston, Morrisville, Wake County, North Carol...	(35.812963749999994, -78.80742802541542)
3	27518	Cary	(Piney Plains, Raleigh, Wake County, North Car...	(35.75471245, -78.75532640688729)
4	27519	Cary	(Wake County, North Carolina, 27519, United St...	(35.797509370161215, -78.88121876337784)

The next step was to integrate this zip code data with the public schools information. This information was available in a CSV format. This has information about the schools , their type and address. For this project I just needed the zip code to associate it with the neighborhood data and the type of the school for clustering. Here is a sample of the school data set.

NAME	GRADELEVEL	PHONE	CALENDAR	ADDRESSNUM	ADDRESS	CITY	ZIPCODE	MAGNETPROGRAM	WEBSITE	DISTRICT	TYPE
Swift Creek Elementary	Elementary	919-233-4320	Traditional	5601	Tryon Rd	Raleigh	27606	N/A	http://www.wcpss.net/swiftcreekes	5	Elementary Schools
Briarcliff Elementary	Elementary	919-460-3443	Traditional	1220	Pond St	Cary	27511	N/A	http://www.wcpss.net/briarcliffes	9	Elementary Schools
Farmington Woods Elementary	Elementary	919-460-3469	Traditional	1413	Hampton Valley Rd	Cary	27511	International Baccalaureate Programme	http://www.wcpss.net /farmingtonwoodses	9	Magnet Elementary School
Cary High	High	919-460-3549	Traditional	638	Walnut St	Cary	27511	N/A	http://www.wcpss.net/caryhs	9	High School
Adams Elementary	Elementary	919-460-3431	Year-Round	805	Cary Towne Blvd	Cary	27511	N/A	http://www.wcpss.net/adamses	9	Elementary Schools

The next step was to gather the population data for each of these neighborhoods. For this I had to do some web scraping from a news website to get the 2018 population numbers. This data was saved to a csv and extracted in to a data frame. Here is the sample of the population data .

City	Population 2018
Aberdeen	7708
Ahoskie	4801
Alamance	1032
Albemarle	16106
Alliance	758

The last step in data gathering is to pull the venue data for all the neighborhoods from Foursquare through api calls. For this I have used my credentials and passed the latitude and longitude information to get the venues in 700 meters radius. I created a function to loop through all the lat and long data for all the zip codes and return venue information within 700 meters of the provided location. Here is the sample data from Foursquare.

	Postcode	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	27502	35.728903	-78.893192	Kidstown Playground / Kelly Road Park	35.732502	-78.894608	Playground
1	27502	35.728903	-78.893192	Kelly Road Softball Field	35.732697	-78.894114	Baseball Field
2	27502	35.728903	-78.893192	greenbrier pool	35.728583	-78.898426	Pool
3	27502	35.728903	-78.893192	Kleibers Kasa	35.723466	-78.893934	Australian Restaurant
4	27502	35.728903	-78.893192	Precision Renovations	35.730846	-78.900341	Construction & Landscaping

b. Cleanup:

The final step with organizing my data is cleaning up and merging all the components into a single data frame to work with . The population data needed some data cleansing as the names of the cities did not exactly match with the neighborhood data due to special characters. The final output is a data frame with foursquare venue data concatenated(Union) with the public school information and joined with population data along with geo information.

- # Methodology and Data Exploration:

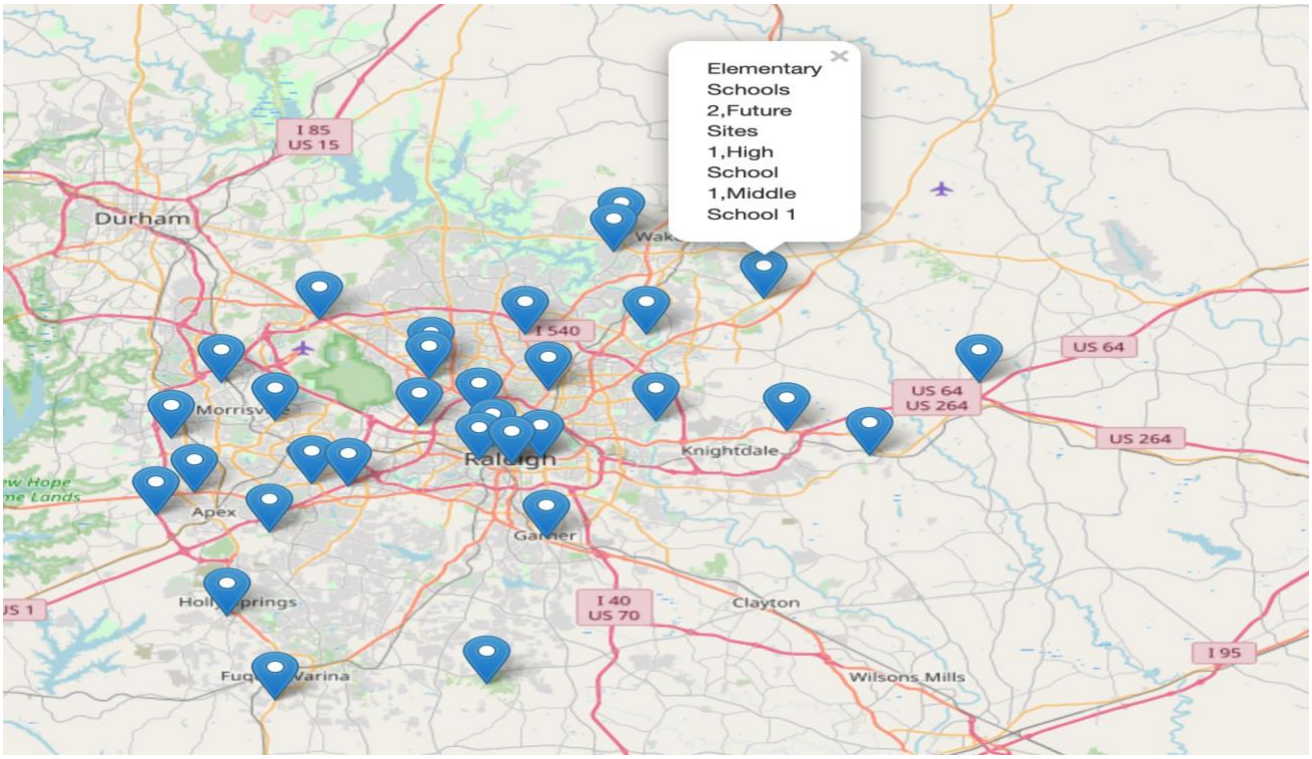
a. Explore:

Explore school data:

Converted the school data in to a pivot to see the counts of different types of public school available in different neighborhoods. Elementary schools seem to be dominating in numbers.

ZIPCODE	Academies	Administration	Elementary Schools	Future Sites	High School	Magnet Elementary School	Magnet High School	Magnet Middle School	Middle School	Special / Optional Schools
27502	0	0	4	1	2	0	0	0	2	0
27511	0	0	3	0	1	1	0	0	1	0
27513	0	0	3	0	0	1	0	1	1	0
27518	0	2	1	0	0	0	1	0	0	1
27519	0	0	8	0	3	0	0	0	3	0

Lets explore that data in a map to visually see where they are located using folium marker map.

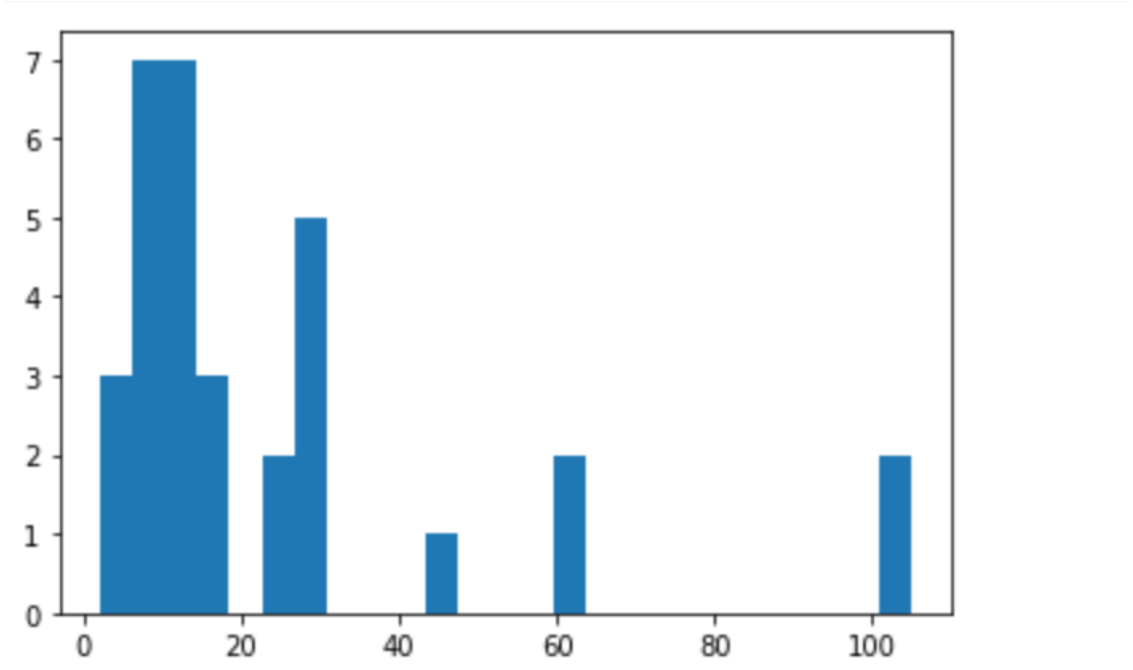


Now let's look at the final data set that contains Foursquare venues and the schools in the respective neighborhoods. The final table has 7 features and 772 samples.

Sample data:

Postcode	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
27502	35.728903	-78.893192	Kidstown Playground / Kelly Road Park	35.732502	-78.894608	Playground
27502	35.728903	-78.893192	Kelly Road Softball Field	35.732697	-78.894114	Baseball Field
27502	35.728903	-78.893192	greenbrier pool	35.728583	-78.898426	Pool
27502	35.728903	-78.893192	Kleibers Kasa	35.723466	-78.893934	Australian Restaurant
27502	35.728903	-78.893192	Precision Renovations	35.730846	-78.900341	Construction & Landscaping

Visualize the data to identify sparse data.



The above graph shows the histogram to get an understanding of how much data we have. In other words count of zip codes per venue. We can get rid of the neighborhoods that have sparse data , i.e. neighborhoods with less than 4 venues. Luckily wen only have one neighborhood with sparse data that was excluded.

b. Transform data:

Top 10 venues:

Transform data to pivot the venue data and create mean of occurrence for each venue type. Then pick the top 10 venues based on the mean. This is done by creating a data fame with the one hot method to pivot all the venue categories and then group the data to create the mean . Then we sort the data in descending order by mean to get the top 10 venues. Here is the sample before sorting.

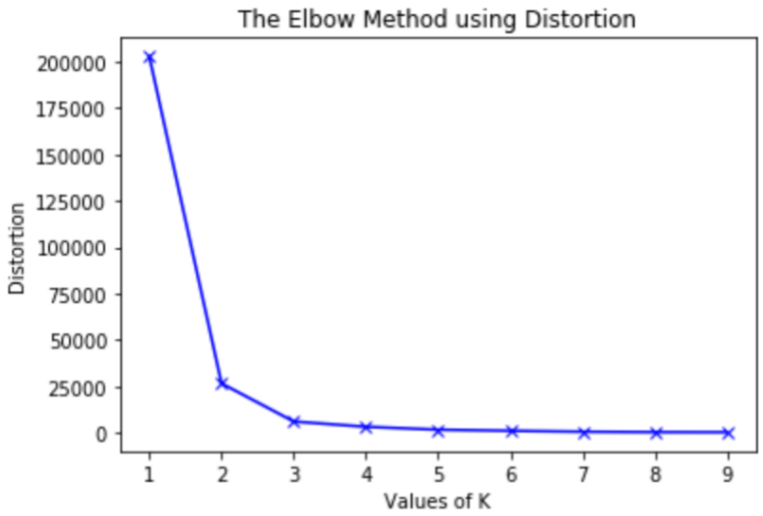
Academies	Administration	Airport	American Restaurant	Antique Shop	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Australian Restaurant	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Basketball Court	Beer Garden	Beer Store	Bookstore	Boutique
0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.071429	0.0	0.0	0.0	0.000000	0.0	0.0	0.071429	0.000000	0.0	0.0	0.0	0.0
0.0	0.000000	0.0	0.044444	0.0	0.0	0.0	0.0	0.0	0.0	0.022222	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0
0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.111111	0.0	0.0	0.000000	0.111111	0.0	0.0	0.0	0.0
0.0	0.166667	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0
0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0

c. Clustering:

K-Means clustering:

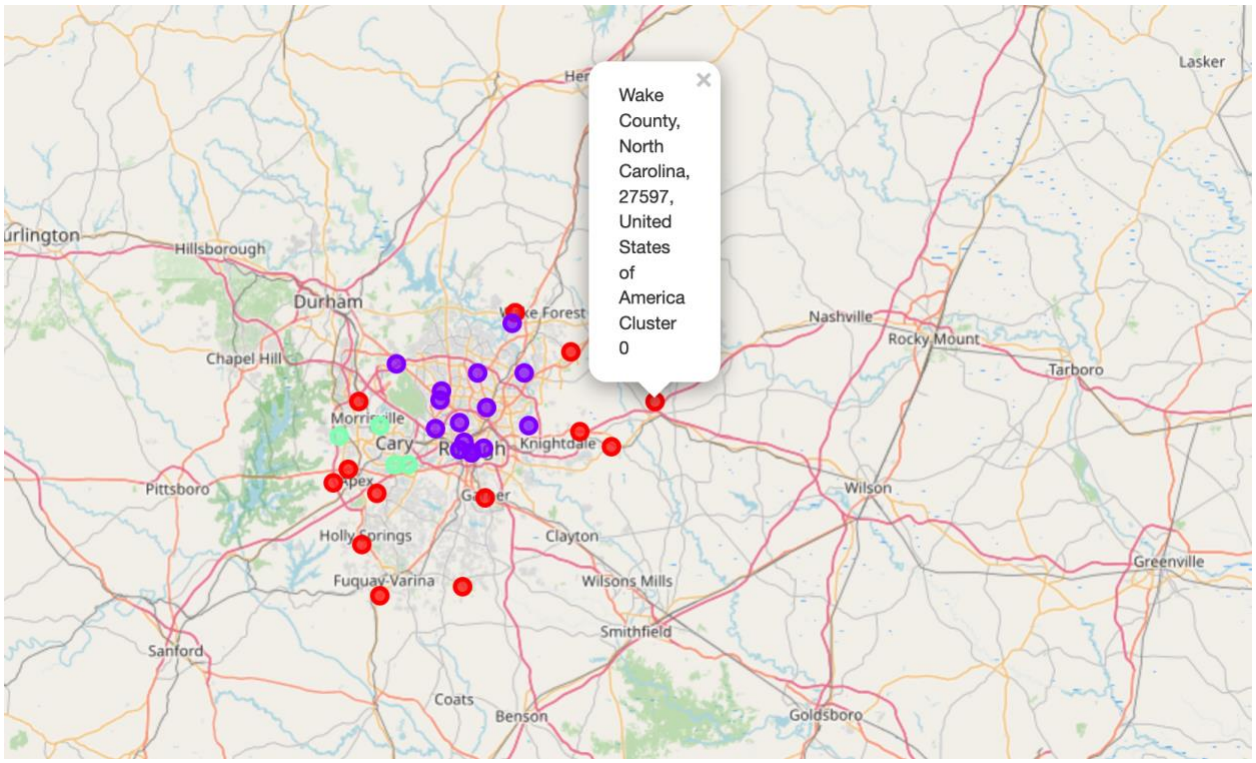
I picked K means method to cluster the data and analyze the determining differences between the clusters. This requires us to declare the K value i.e. the number of clusters the data should be clustered into.

I used the K-means elbow method to determine the optimal number of clusters, as below.



From the above graph, it is clear that K=3 would be the ideal number of clusters for this data.

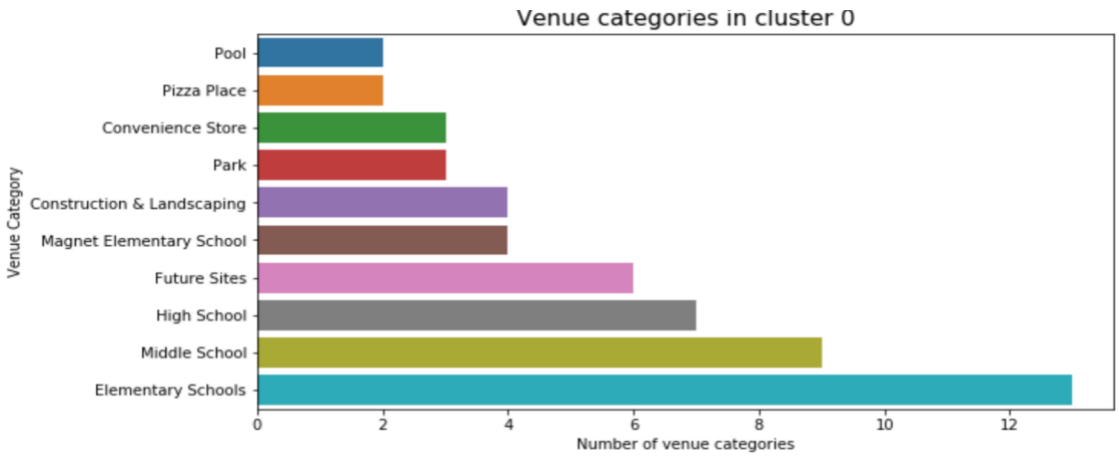
When we visualize the resulting clusters on a map for Wake County. Here is how it looks.

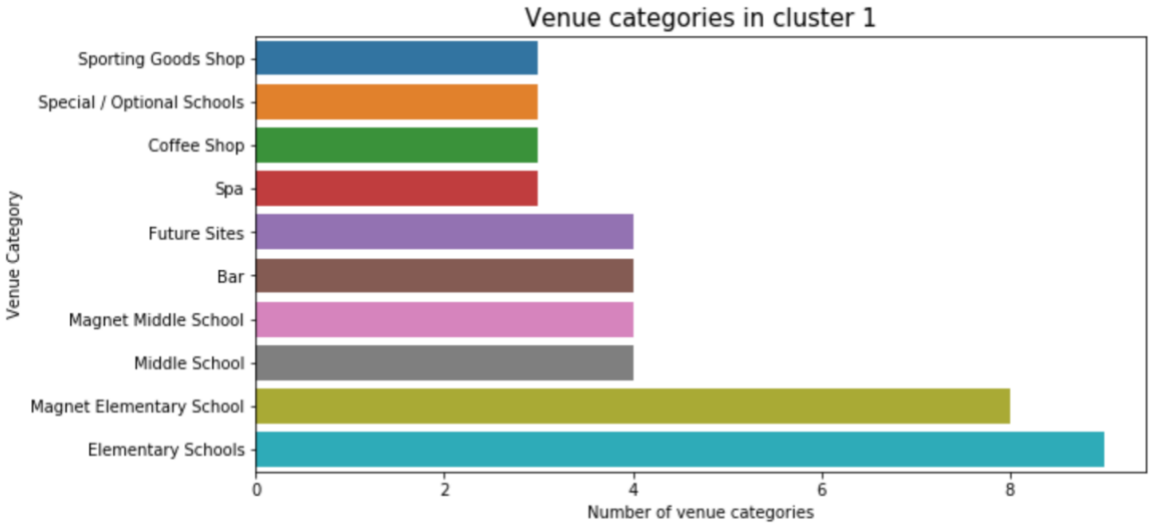


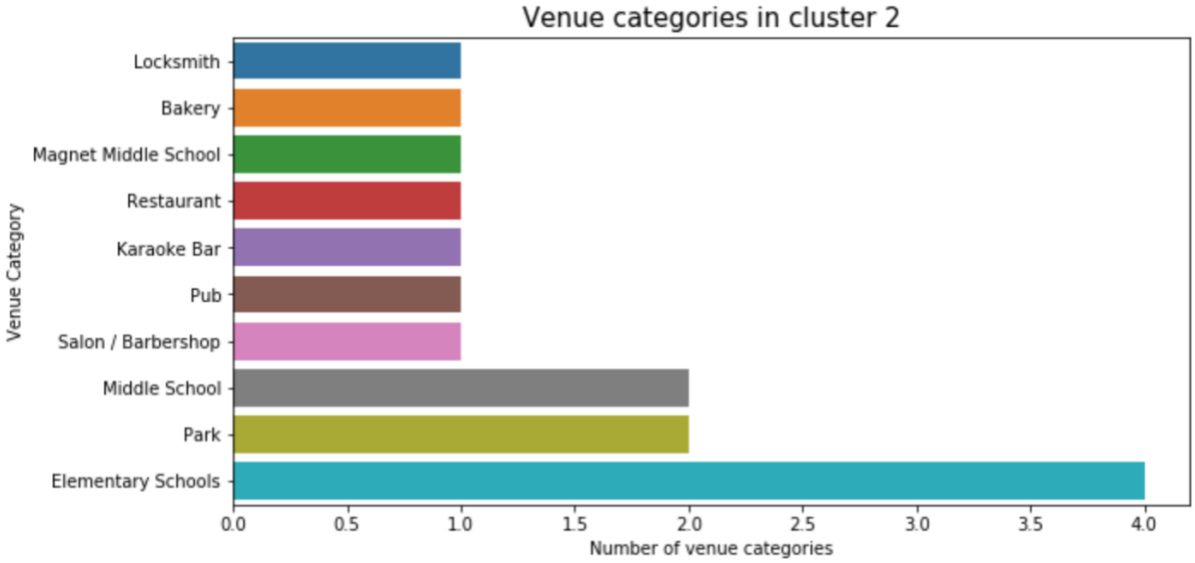
Data where you describe the data that will be used to solve the problem and the source of the data.

- # Results Section:

Here is how the clusters look visually:







As you can see the schools lead the results in all the clusters. But , that is what we want , since we are targeting the families with school going kids. The results differentiate the 3 clusters .

Elementary schools top the charts of most common venues in different clusters. Cluster 3 looks rather interesting. Let's discuss the differences between these clusters and what they mean.

• Discussion:

To get a better idea about where families with children typically tend to settle in the county, lets compare the different zip codes based on their average real estate process. The average real estate price for this county according to Zillow is **$310,628.**

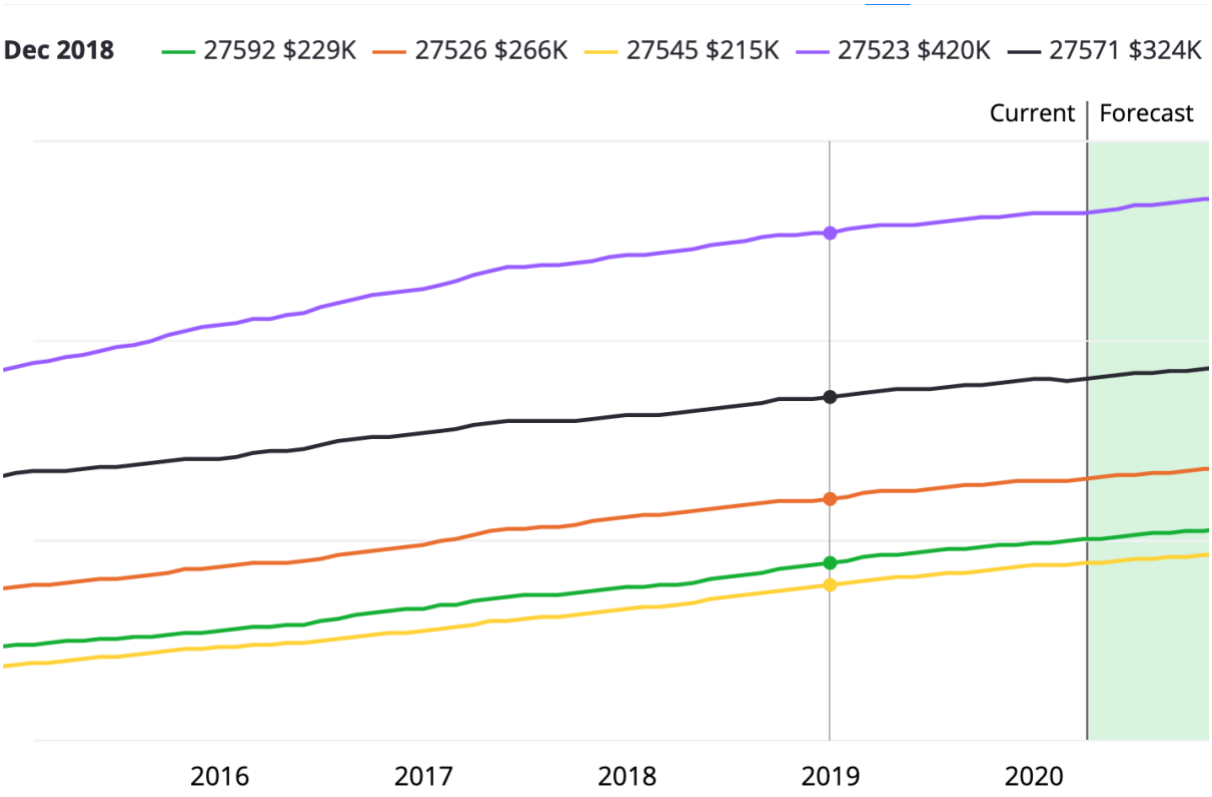
Cluster 0:

In Cluster 0 ,there are good number of schools in these areas. There aren't many restaurants in this area. There are Future sites, zoned for future use by the county. This tells that these areas are yet to reach their full potential.

My analysis is that, these zip codes are located away from busy areas like downtown, and the real estate price is typically at or below average when compared to the county average.

Schools Availability: High
Real Estate Price range Estimate: Low. This cluster is not crowded and is spread out away from the center.
Food and Entertainment index: Medium. Not many restaurants.
Investment Return potential: High

As you can find in the below chart gathered from Zillow about the average price of real estate for some of the zip codes in this cluster. You can see that the average price of this cluster is at or below the Wake County average real estate price.



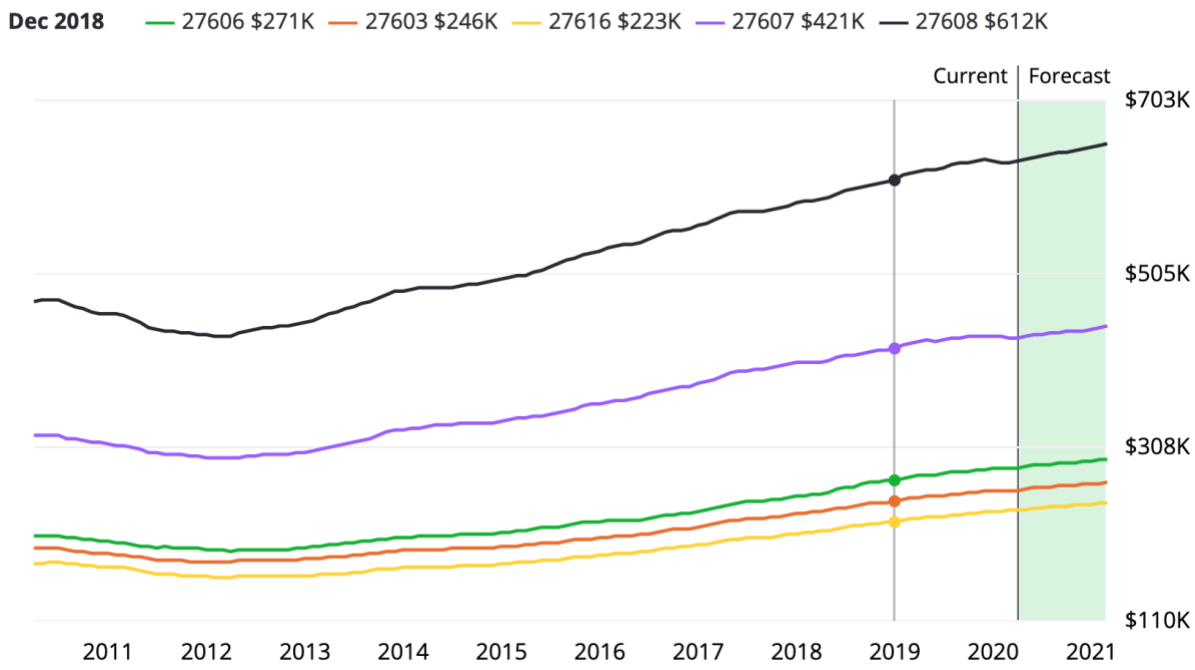
Cluster 1:

In cluster 1 there are very high number of schools in this area. There are good number of restaurants and entertainment places in this area. There are Future sites, zoned for future use by the county. This tells that the area is still yet to reach its full potential, but is growing fast.

My analysis is that, these zip codes are closer to busy areas like downtown, and the real estate price expected to be above average when compared to the county average.

Schools Availability: High (Not much choice for Middle Schools)
Real Estate Price range Estimate: High. This cluster is crowded and is closer to the center. The number of schools indicate a high population density in that area. But there are good number of future sites which indicates there is still some scope for development.
Food and Entertainment index: Medium.
Investment Return potential: High

As you can see in the Zillow chart below, the average price is above the Wake County's average price for this cluster as predicted.



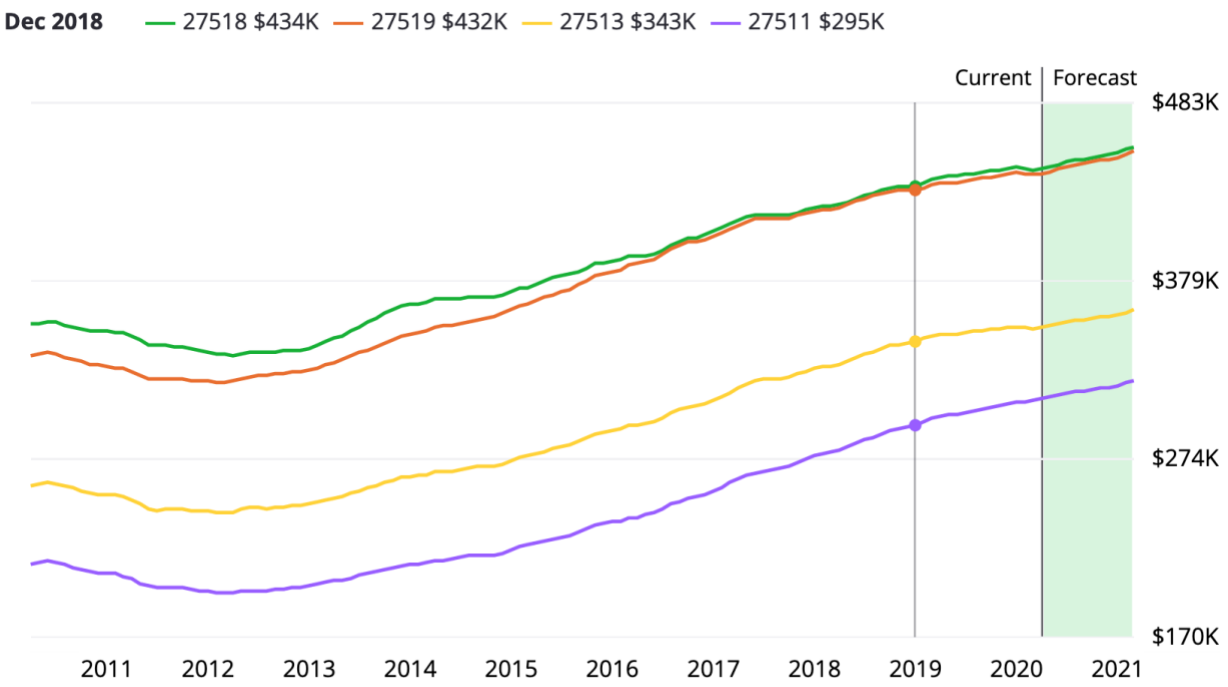
Cluster 2:

There are not many schools in this cluster compared to others. The number of restaurants and entertainment places in this area are high compared to others. There are no Future sites, zoned for future use by the county. This tells that the area is most likely well established and there is very little space for future development.

My analysis is that, these zip codes are in the downtown areas, and the real estate price expected to be high when compared to the county average.

Schools Availability: Medium (Not many Schools)
Real Estate Price range Estimate: High. This cluster is crowded and is closer to the center. No future sites which indicates the area is well established.
Food and Entertainment index: High.
Investment Return potential: High. Since the areas are well established the real estate is like expensive. But, the return of investment can go higher with the population increase.

As you can see in the Zillow chart below, the average price is above the Wake County's average price for this cluster as predicted.



For this particular problem, the analysis would have been even better if there was additional data available like Real estate data , Hospitals, universities, private schools information etc.

- **Conclusion:**

The purpose of this project was to help families with children migrating to Wake County to identify a good neighborhood to settle down . The analysis gives a good direction for the intended audience by giving a peep in to the different neighborhoods and how they are similar. This can be definitely improved on with additional data. Working with limited data was a challenge. Including more dimensions like Crime data, Real estate data , Hospitals, universities, School ratings, drug problems and hotspots would have made the analyses comprehensive.