Locating the Ideal Location for New Pizzeria In Wake County, North Carolina

IBM Data Science Professional Certification
Capstone Project

By Mark Trimer
June 2020

Background
Wake County is located in central North Carolina. Its largest city is Raleigh, the state capital. Along with eleven other municipalities that surround Raleigh, Wake County has a population of 1.1M, making it the most populated county in North Carolina. It is also home to the largest school district in the state.

Within a half-day's drive from both the mountains and the ocean, and midway along the Atlantic seaboard, Wake County's location is ideal to access many points of interest. Ranking 3rd for the Best Counties to Live in America, 14th for Best Counties for Families in America, and combined with low unemployment, Wake County has experienced phenomenal population growth.

Recently, a friend has decided to move to Wake County and start a pizzeria. In order to help, an analysis will be created to pinpoint the top-5 locations in the county to open the shop.

Problem Statement
In order to find the best potential locations, one of the factors will be to find dense population centers which lack pizzerias. This approach should maximize clientele while minimizing competition.

This analysis will uncover a metric to calculate and locate where the top-5 potential locations should be, through analytics of data from multiple sources.

With that said, other factors should come into play, too, in order to find the best potential locations. However, these are out of scope for this analysis but are discussed briefly later in the document.

The ideal audience for this analysis would be any entrepreneur looking to open a pizzeria in Wake County. The approach, though, could then be easily modified to accommodate any cuisine or location.

Data Sources
Two data sources will be used.

1) Wake County Open Data (http://data-wake.opendata.arcgis.com/)

   This resource contains a wealth of data about Wake County, including properties, parks, education, public safety, and zoning data. Of interest, it contains demographic information. Specifically, one table includes 2010 census data by census tract. This table includes the area of the tract and population. This is ideal to calculate segmented population density throughout the county.
   (http://data-wake.opendata.arcgis.com/datasets/census-tracts-2010)

2) Foursquare (https://developer.foursquare.com/)
   A comprehensive resource of crowd-sourced data, Foursquare will be leveraged to find all the pizzeria locations in Wake County via its developer portal.

Approach

a) Using the **Wake County Census Tract Data**, obtain population and area (km^2), along with longitude and latitude of each tract. This is ideal for calculating relative population density throughout the county.

   Census Tract is an excellent a proxy for neighborhood data, especially considering customers will be coming various neighborhoods. Wake County does not have boroughs.

b) Using **Foursquare**, pull all the restaurants in Wake County that can be keyed on Italian cuisine or pizza, along with its location.

   This list will be further refined, as the entrepreneur believes his pizza product and its unique recipe stands alone. Therefore, he does NOT consider national chains to be direct competitors. The Foursquare data will be scrubbed to exclude the following: Papa John's, Domino's, Little Caesar's, and Pizza Hut.

c) Merge the two data sets on longitude and latitude. For the location of each Census Tract, count the number of pizzerias and Italian restaurants within a 5-kilometer radius. This is roughly the extent of current delivery radii and translates into a 10-minute drive.

d) Calculate the population per square kilometer and divide it by the number of competitor locations. This provides the maximum number of potential customers within close driving proximity to the location, relative to competing restaurants.

e) Sort this metric in descending order and document the location of the top-5 census tracts. A higher metric translates into higher potential customer-base with minimal competition.

The Data

The first source accessed was the **Wake County Census Tract Data**, in order to obtain population and area (km^2), along with longitude and latitude information. Data was downloaded as a .csv., where it was scrubbed to delete unnecessary columns, to add calculated column for Population Density, and to improve the column header names.

It's important to note that the data is from the 2010 census. It has not been updated with newer estimates, since then.

The .csv file was loaded as a pandas dataframe.

	OBJECTID	Population	Land_Area	Population_Density	Latitude	Longitude	NAMELSAD10	GEOID10
0	1	4567	4.682359	975.363060	35.808957	-78.915059	Census Tract 534.10	37183053410
1	2	4672	4.088491	1142.719893	35.784092	-78.762571	Census Tract 535.17	37183053517
2	3	7605	108.199695	70.286705	36.015825	-78.651009	Census Tract 539	37183053900
3	4	3610	4.115246	877.225809	35.756804	-78.759858	Census Tract 535.05	37183053505
4	5	3186	33.858770	94.096744	35.736014	-78.477742	Census Tract 541.09	37183054109

The dataframe has 187 Census Tracts, labeled 1-187 under OBJECTID column. Each tract has a population count, density calculation, land area (in square kilometers), latitude and longitude.

This data set is ideal to discover the densest areas within the county, and where they are located. Again, Census Tract data is an excellent a proxy for neighborhood data, especially considering customers will be coming various neighborhoods.

The second source accessed was **Foursquare**, in order to obtain all the venues in Wake County filtered on Italian cuisine or pizza, along with its location. Therefore, the analysis can pinpoint where the competition is located.

The initial data pull accesses all the venues within a 5km radius of each Census tract. Five kilometers is roughly a 10-minute drive, and the extent of a pizza delivery range.

A staggering 16,047 venues were pulled initially from Foursquare across the 187 Census tracts. Here is that pandas dataframe header:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	1	35.808957	-78.915059	American Tobacco Trail - New Hope Church Rd Ac...	35.816495	-78.927656	Trail
1	1	35.808957	-78.915059	Andia's Homemade Ice Cream	35.820146	-78.901185	Ice Cream Shop
2	1	35.808957	-78.915059	USA Baseball National Training Complex	35.797485	-78.892539	Baseball Stadium
3	1	35.808957	-78.915059	Los Tres Magueyes	35.819912	-78.900982	Mexican Restaurant
4	1	35.808957	-78.915059	Thomas E. Brooks Park	35.794564	-78.895054	Park

Obviously, a great deal of irrelevant information was included. The dataframe was pared down on the Venue Category column to only include pizzerias and Italian restaurants.

As mentioned earlier, too, national chains are out-of-scope, as they are not considered direct-competition, as the entrepreneur believes his unique recipe stands alone. The Foursquare data was scrubbed to exclude the following: Papa John's, Domino's, Little Caesar's, and Pizza Hut.

By scrubbing the Venue Category frame, the number of rows in the dataframe dropped to 938. Then, by eliminating the national chains, another 10 Little Caesar's, 25 Papa John's, 26 Domino's, and 25 Pizza Hut venues were dropped.

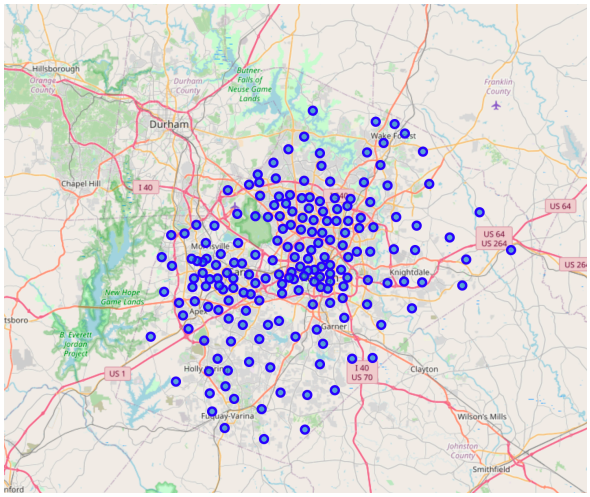
This leaves 852 venues throughout the county to consider. These are essentially the competitors. Here is a header of the data frame:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
10	1	35.808957	-78.915059	MOD Pizza	35.823656	-78.903501	Pizza Place
16	1	35.808957	-78.915059	Brixx Wood Fired Pizza	35.847511	-78.889840	Pizza Place
28	1	35.808957	-78.915059	Stellino's	35.848652	-78.889156	Italian Restaurant
33	1	35.808957	-78.915059	New York Pizza	35.785960	-78.872490	Pizza Place
72	1	35.808957	-78.915059	Cicis	35.784665	-78.870492	Pizza Place

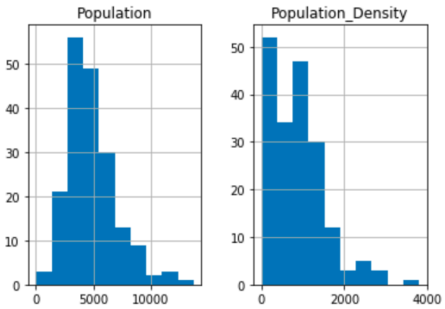
This is all the data needed to determine where to ideally locate the new pizzeria. The next step is to look at the data and pinpoint the best locations.

Data Exploration

The Wake County Census tracts are located as follows throughout the county:



The population and Population Density of each Census tract appears as follows in a histogram. Not much can be gleaned from these histograms. There are a few outliers, but nothing to insightful.



The Foursquare data must be grouped by neighborhood, so number of nearby competitors can be calculated more efficiently. The 852 competitors were grouped into each neighborhood. Column headings were updated. Here is a header of the revised dataframe:

	Competitors
OBJECTID	
1	5
2	2
4	1
7	5
8	1
9	2
10	4
11	9
12	10
13	2

Analysis and Methodology

The two data frames must be merged into one data frame.

The following steps were taken.

a) Merge the **Wake County Population by Census tract** data frame with the **Foursquare** pizza and Italian restaurant dataframe. Merge them on the OBJECTID column.
b) Add a column called Competition Density, which takes the **Population Density** and divides it by the **Competitors**.
c) The **Competition Density** column is sorted in descending order.

The following dataframe was created. The header is as follows:

	OBJECTID	Population	Land_Area	Population_Density	Latitude	Longitude	NAMELSAD10	GEOID10	Competitors	Competition_Density
125	134	4196	1.104549	3798.835543	35.777891	-78.684019	Census Tract 524.09	37183052409	2	1899.417772
126	135	6413	3.862827	1660.183073	35.778728	-78.733541	Census Tract 524.06	37183052406	1	1660.183073
138	147	3490	2.634857	1324.550061	35.771635	-78.751730	Census Tract 535.16	37183053516	1	1324.550061
130	139	4255	1.738790	2447.104021	35.781942	-78.665827	Census Tract 511.02	37183051102	2	1223.552011
24	28	4696	1.635126	2871.949929	35.788013	-78.688316	Census Tract 511.01	37183051101	3	957.316643

Of note, this dataframe only included 171 rows, not the original 187 rows. These 16 rows were dropped, because they had no competing venues. This will be investigated further to ensure there are no missed opportunities in these 16 locations.

Competition Density Metric Discussion

The analysis and conclusion hinge on the definition of the Competition Density metric. Essentially, the entrepreneur wants to locate his new business where

a) There is a high concentration of potential customers
b) The is minimal competing restaurants serving the same cuisine.

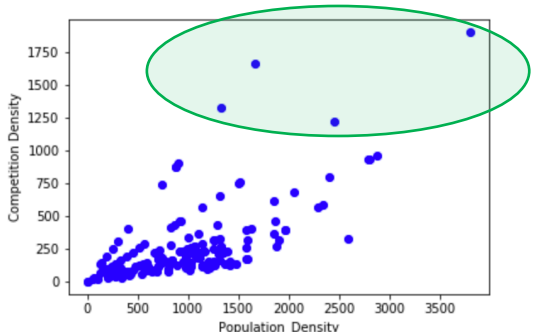
For example, two areas of the same size each has a population density of 1000 people per square kilometer.

a) Area 1 has 4 competing restaurants. Its competition density is 1000/4 = 250.
b) Area 2 has 8 competing restaurants. Its competition density is 1000/8 = 125.

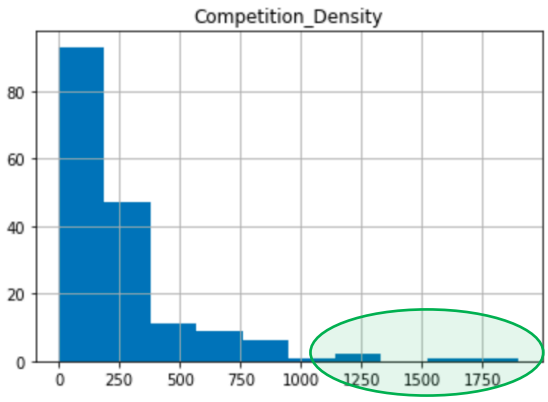
Ideally, one would want to locate the new restaurant in Area 1. Its Competition Density is higher, and therefore the better choice in this analysis.

Discussion of Wake County Competition Density Results

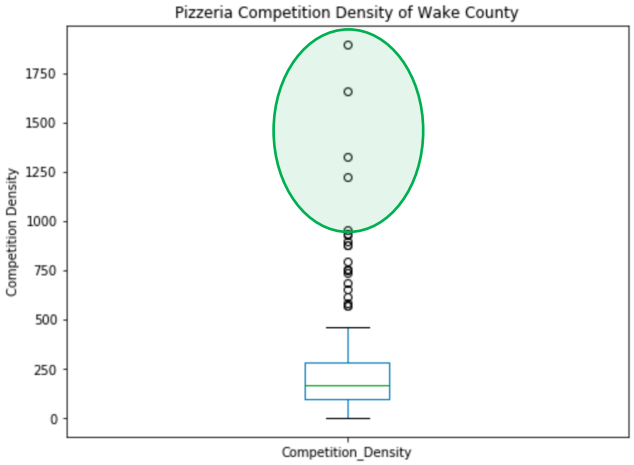
Generating a scatter plot, there is some clustering of the population density, however, the locations away from the cluster are more ideal.



Pulling the histogram, there are some outliers with high Competition Density numbers.



Looking at the Boxplot, this is more evident. Four (4) Census tracts are outliers. The conclusion will need to key on these locations.



Here are the details of these 4 locations. This is the answer to our problem statement:

	OBJECTID	Population	Land_Area	Population_Density	Latitude	Longitude	NAMELSAD10	GEOID10	Competitors	Competition_Density
125	134	4196	1.104549	3798.835543	35.777891	-78.684019	Census Tract 524.09	37183052409	2	1899.417772
126	135	6413	3.862827	1660.183073	35.778728	-78.733541	Census Tract 524.06	37183052406	1	1660.183073
138	147	3490	2.634857	1324.550061	35.771635	-78.751730	Census Tract 535.16	37183053516	1	1324.550061
130	139	4255	1.738790	2447.104021	35.781942	-78.665827	Census Tract 511.02	37183051102	2	1223.552011

The original goal to recommend 5 locations will be revised to 4, as there are clearly 4 outliers.

Recall there were 16 rows that dropped because they had no competing venues. Here is a dataframe of these 16. Looking at the population densities of each of these Census tracts, if an entrepreneur were to locate in any one of these locations, they are not dense enough, still, to override the 4 locations just identified.

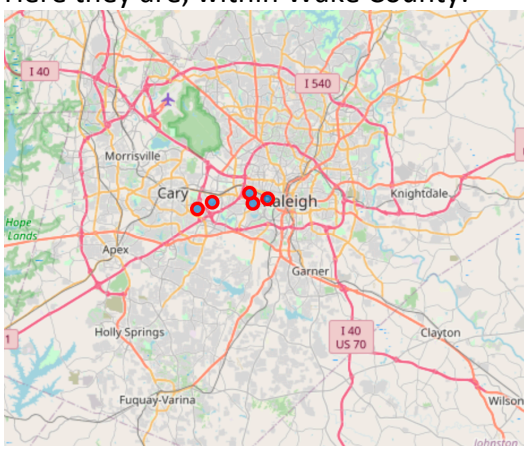
	OBJECTID	Population	Land_Area	Population_Density	Latitude	Longitude	NAMELSAD10	GEOID10
2	3	7605	108.199695	70.286705	36.015825	-78.651009	Census Tract 539	37183053900
4	5	3186	33.858770	94.096744	35.736014	-78.477742	Census Tract 541.09	37183054109
5	6	5648	93.506728	60.402071	35.871812	-78.360604	Census Tract 543.01	37183054301
41	42	3318	29.050411	114.215252	35.978970	-78.665727	Census Tract 538.04	37183053804
42	43	3737	11.885501	314.416700	35.960815	-78.693149	Census Tract 538.03	37183053803
50	51	2021	19.548598	103.383373	35.853275	-78.469168	Census Tract 541.11	37183054111
63	64	3886	33.645535	115.498238	35.836862	-78.411865	Census Tract 544.03	37183054403
85	86	2060	10.413494	197.820251	35.955480	-78.638863	Census Tract 538.05	37183053805
156	157	4388	42.237359	103.889071	35.713578	-78.531464	Census Tract 528.07	37183052807
159	160	13750	17.829267	771.203886	35.742080	-78.556768	Census Tract 528.06	37183052806
160	161	5312	16.320889	325.472467	35.652969	-78.725817	Census Tract 531.11	37183053111
161	162	5523	29.089704	189.860990	35.616998	-78.707297	Census Tract 531.10	37183053110
164	165	2871	21.190013	135.488355	35.564721	-78.664885	Census Tract 529.01	37183052901
165	166	7260	27.413833	264.829803	35.613935	-78.660517	Census Tract 529.04	37183052904
166	167	2914	12.707176	229.319244	35.657370	-78.638305	Census Tract 529.03	37183052903
177	178	4853	45.265980	107.210757	35.696650	-78.934694	Census Tract 534.13	37183053413

Insight and Conclusion

A recommendation is made to the pizzeria entrepreneur to consider four (4) locations in Wake County. These locations maximize nearby potential customers, while minimizing existing competition. The following locations are provided.

	OBJECTID	Population	Land_Area	Population_Density	Latitude	Longitude	NAMELSAD10	GEOID10	Competitors	Competition_Density
125	134	4196	1.104549	3798.835543	35.777891	-78.684019	Census Tract 524.09	37183052409	2	1899.417772
126	135	6413	3.862827	1660.183073	35.778728	-78.733541	Census Tract 524.06	37183052406	1	1660.183073
138	147	3490	2.634857	1324.550061	35.771635	-78.751730	Census Tract 535.16	37183053516	1	1324.550061
130	139	4255	1.738790	2447.104021	35.781942	-78.665827	Census Tract 511.02	37183051102	2	1223.552011

Here they are, within Wake County:



Other Considerations:

Obviously, location is not the only consideration. Other factors that can be researched and weighed by the entrepreneur include:

I. Demographics – ensuring pizza and Italian cuisine are favored within the area
II. Income – does the area have adequate disposable income to support the restaurant
III. Real Estate Costs – rent is a substantial overhead to manage
IV. Future Development Potential – is the area growing, or potentially ready to grow?
V. Density of Population – Is the Population density calculation skewed, as the Census tract is zoned commercial and industrial versus residential?
VI. Refreshing analysis when 2020 census results are complete – Population data is nearly 10 years old and may not reflect substantial changes over the last decade.

References:

https://www.niche.com/places-to-live/c/wake-county-nc/rankings/
https://en.wikipedia.org/wiki/Wake_County,_North_Carolina