Applied Data Science Capstone Project – The Battle of Neighborhoods

Business Problem

The city of Charlotte, North Carolina has seen tremendous economic and population growth over the past 20 years. Between 2004 and 2014, Charlotte was ranked as the country's fastest-growing metro area, with 888,000 new residents. According to U.S. Census data, from 2005 to 2015, Charlotte topped the U.S. in millennial population growth. It is the second-largest city in the southeastern United States, just behind Jacksonville, Florida. Additionally, it is currently the third-fastest-growing major city in the United States, per the Seattle Times.

A group of investors is considering adding a quick-service empanada franchise to the Charlotte region and needs help determining which of Charlotte's 199 bustling neighborhoods is the best fit. The ideal location would provide a consistent mix of traffic to the restaurant from the lunch hour business crowd, working parents on the go in the early evening, as well as weekend traffic from the arts and entertainment crowd. Due to previous lack luster results from investing in a different southeastern city, this investment group is keen on a successful launch.

Data

This analysis will include location data gathered via the Foursquare API to determine the ideal neighborhood in which to launch the empanada restaurant. The neighborhoods will be clustered and subsequently evaluated based on their proximity to corporate office buildings, family style attractions, gyms, other Latin American-style restaurants as well as non-Latin American-style restaurants. Because the data available in Foursquare does not include neighborhood specific details, it will be analyzed based on venue categories by postal code first, followed by finding the 10 most common venues per postal code to find the best location.

Methodology

Pandas will be deployed to analyze the data pulled from the Foursquare API. The python folium library will be integral in creating several maps to display all of the venues in and around Charlotte as well as the different categories in comparison to Latin American-themed restaurants. Clustering will be used to make a cleaner map of all of the venues and the k-means algorithm will be used to determine the number of clusters to use in order to find the ideal neighborhood where this restaurant could succeed.

When pulling data using any API or website, it is necessary to narrow down the dataframe to include only relevant information. It also requires checking and, where necessary, changing data formats in order to make the best use of the data analysis tools available, such as pandas and folium. Below are the top five rows of the dataframe created from the downloaded .json file from Foursquare.

	name	categories	address	crossStreet	Y	X	labeledLatLngs	distance	postalCode	cc	city	state	country	formattedAddress	neighborhood	id
0	Blumenthal Performing Arts Center	Performing Arts Venue	130 N Tryon St	at 5th St	35.227930	-80.841951	[{'label': 'display', 'lat': 35.22792953956913...	130	28202	US	Charlotte	NC	United States	[130 N Tryon St (at 5th St), Charlotte, NC 282...	NaN	4bad5829f964a52071483be3
1	Romare Bearden Park	Park	300 S Church St	at W 3rd St	35.226927	-80.847685	[{'label': 'display', 'lat': 35.22692655213674...	419	28202	US	Charlotte	NC	United States	[300 S Church St (at W 3rd St), Charlotte, NC ...	NaN	4e60dc33483bd9a9739e0b07
2	The Capital Grille	American Restaurant	201 N Tryon St	at E 5th St	35.228216	-80.841974	[{'label': 'display', 'lat': 35.228215965114 68...	150	28202	US	Charlotte	NC	United States	[201 N Tryon St (at E 5th St), Charlotte, NC 2...	NaN	4b05863ef964a52045922e3
3	Knight Theater	Theater	430 S Tryon St	btw 1st & 2nd	35.224415	-80.847743	[{'label': 'display', 'lat': 35.22441516682533...	525	28202	US	Charlotte	NC	United States	[430 S Tryon St (btw 1st & 2nd), Charlotte, NC...	NaN	4b1ab236f964a5209bef23e3
4	Belk Theater	Concert Hall	130 N Tryon St	NaN	35.227711	-80.841663	[{'label': 'display', 'lat': 35.2277106677568,...	140	28202	US	Charlotte	NC	United States	[130 N Tryon St, Charlotte, NC 28202, United S...	NaN	4b058640f964a520495a22e3

It is critical to take a look at the data in order to identify any potential gaps in information and to ensure that enough information is available to properly conduct the necessary analysis.

```
<class 'pandas.core.frame.DataFrame'>
click to scroll output; double click to hide
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   name              100 non-null    object
 1   categories        100 non-null    object
 2   address           100 non-null    object
 3   crossStreet       43 non-null     object
 4   Y                 100 non-null    float64
 5   X                 100 non-null    float64
 6   labeledLatLngs    100 non-null    object
 7   distance          100 non-null    int64
 8   postalCode        98 non-null     string
 9   cc                100 non-null    object
 10  city              100 non-null    object
 11  state             100 non-null    object
 12  country           100 non-null    object
 13  formattedAddress  100 non-null    object
 14  neighborhood      7 non-null      object
 15  id                100 non-null    object
dtypes: float64(2), int64(1), object(12), string(1)
memory usage: 12.6+ KB
```

*Note that the neighborhood classification was available for only seven of the one hundred venues, therefore postal codes will be used instead of neighborhood names for clustering purposes. The crossstreet column is not relevant for this analysis because latitude (Y) and longitude (X) will be used for mapping purposes. The remaining information appears to be complete.

With 58 individual categories, it would be best to combine similar ones for the purposes of this analysis.

Brewery	13
Park	5
American Restaurant	5
Pizza Place	4
Deli / Bodega	3
Grocery Store	3
Bakery	3
Italian Restaurant	3
Mexican Restaurant	3
BBQ Joint	2
Theater	2
Gift Shop	2
Southern / Soul Food Restaurant	2
Bar	2
Coffee Shop	2
Peruvian Restaurant	2
Gym	2
Farmers Market	2
Café	2

Restaurants were consolidated into four main types: American, Latin American, European and Asian. These could have been combined in another fashion, such as by price range or quick service versus sit-down style, however complete restaurant details were outside of the available information provided via the Foursquare API.

Other combinations of note include the consolidation of 'Science Museum', 'General Entertainment', 'Park', 'Theater', 'Baseball Stadium', 'Art Museum', 'Art Gallery', 'Basketball Stadium', 'Rock Club', 'Performing Arts Venue', 'Concert Hall', 'Football Stadium' into the Arts and Entertainment category; 'Grocery Store', 'Farmers Market', 'Market', 'Convenience Store', 'Deli / Bodega', 'Beer Store', 'Wine Shop' into the Food Market category; and 'Gift Shop', 'Smoke Shop', 'Comic Shop', 'Antique Shop', 'Bookstore', 'Liquor Store' into the Specialty Shop category.

American Restaurant	21
Brewery / Bar	18
Arts and Entertainment	16
Food Market	12
Latin American Restaurant	8
Café	7
Specialty Shop	5
European Restaurant	5
Asian Restaurant	4
Gym	3
Hotel	1

Name: categories, dtype: int64

Now there are only 11 categories.

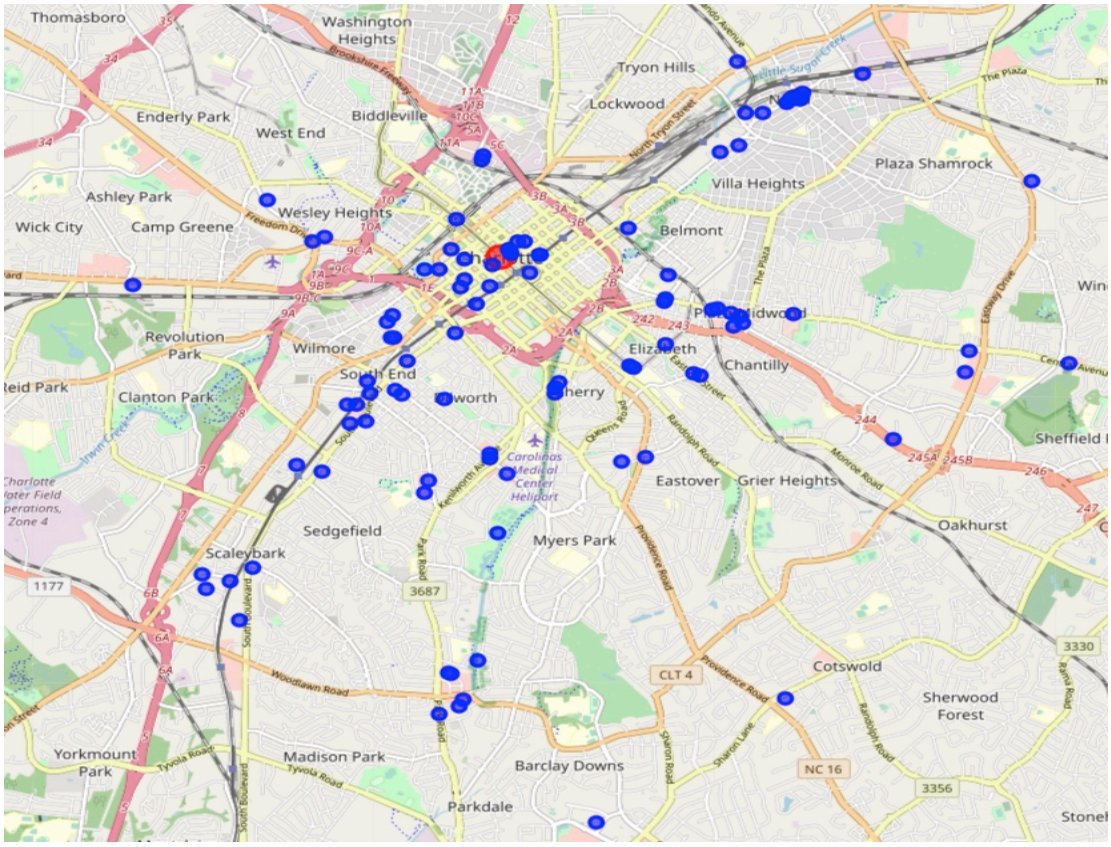
Anecdotally, there are some clear gaps with what was provided in the Foursquare .json file. There is more than one hotel and more than three gyms in Charlotte. However, at first glance the restaurant as well as the arts and entertainment categories appear to be more accurate than the hotel and gym information.

	name	categories	address	crossStreet	Y	X
22	Sabor Latin Street Grill	Latin American Restaurant	415 Hawthorne Ln	Hawthorne and 7th St	35.216258	-80.821734
24	Viva Chicken Elizabeth Avenue	Latin American Restaurant	1617 Elizabeth Ave	NaN	35.213407	-80.825966
30	Superica	Latin American Restaurant	101 W Worthington Ave	Suite 100	35.211736	-80.860303
38	Bakersfield	Latin American Restaurant	1301 East Blvd	NaN	35.202040	-80.844430
73	Cabo Fish Taco	Latin American Restaurant	3201 N Davidson St	at E 35th St	35.247173	-80.805700
78	Sabor Latin Street Grill	Latin American Restaurant	3205 N Davidson St	NaN	35.247129	-80.805511
92	Viva Chicken Park Road	Latin American Restaurant	4500 Park Rd, Suite 100	Montford	35.169891	-80.851038
99	Paco's Tacos & Tequila	Latin American Restaurant	6401 Morrison Blvd Ste 8A	NaN	35.156275	-80.830696

Per the Foursquare information, there are only eight Latin American restaurants in the greater Charlotte metro area.

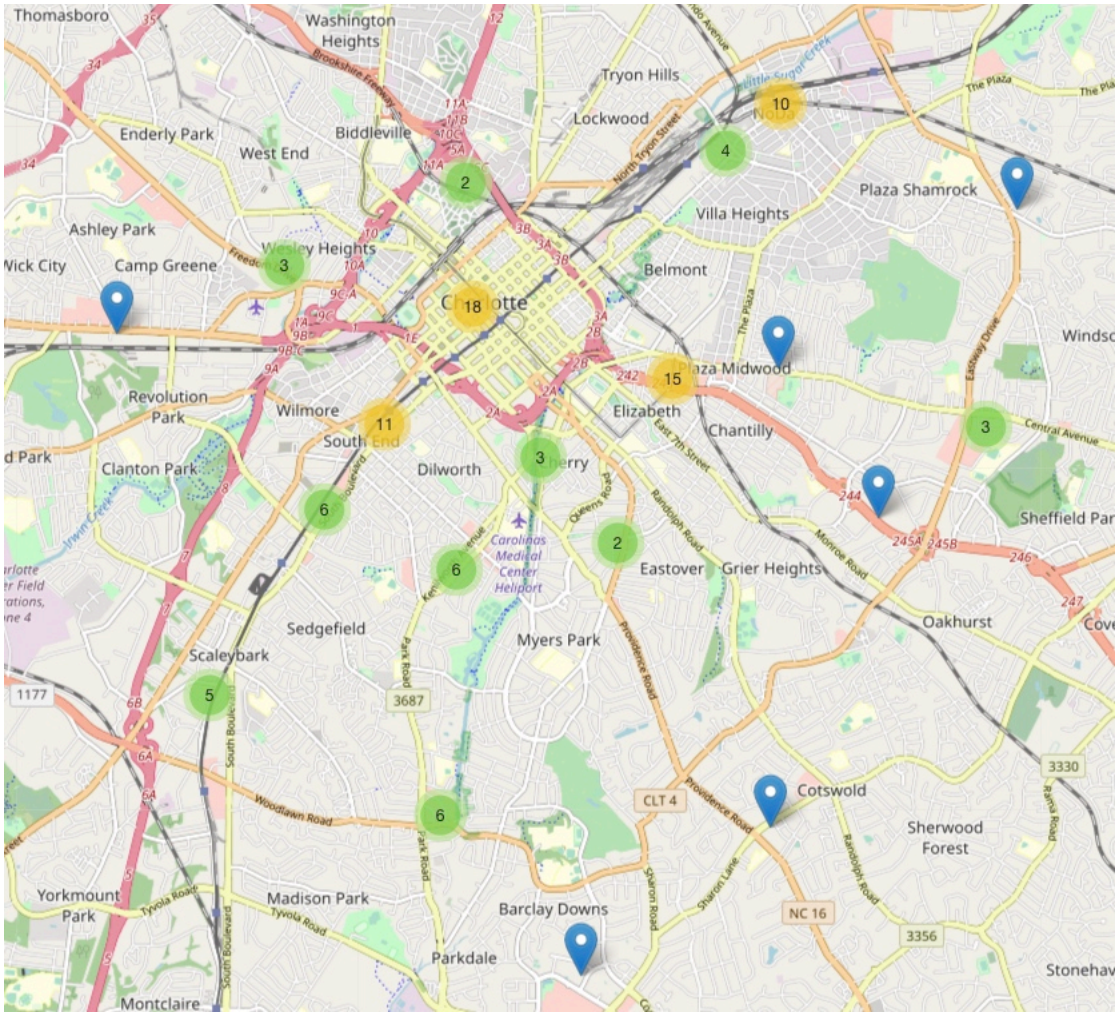
Now that it has been cleaned, classified and filtered, it is time to visualize the data set. First up, mapping all of the Foursquare venues (in blue) around the center of Charlotte (red dot).

Charlotte Venues Map



With so many venues in a concentrated area, it helps to cluster the data points to get a better picture of the areas of saturation.

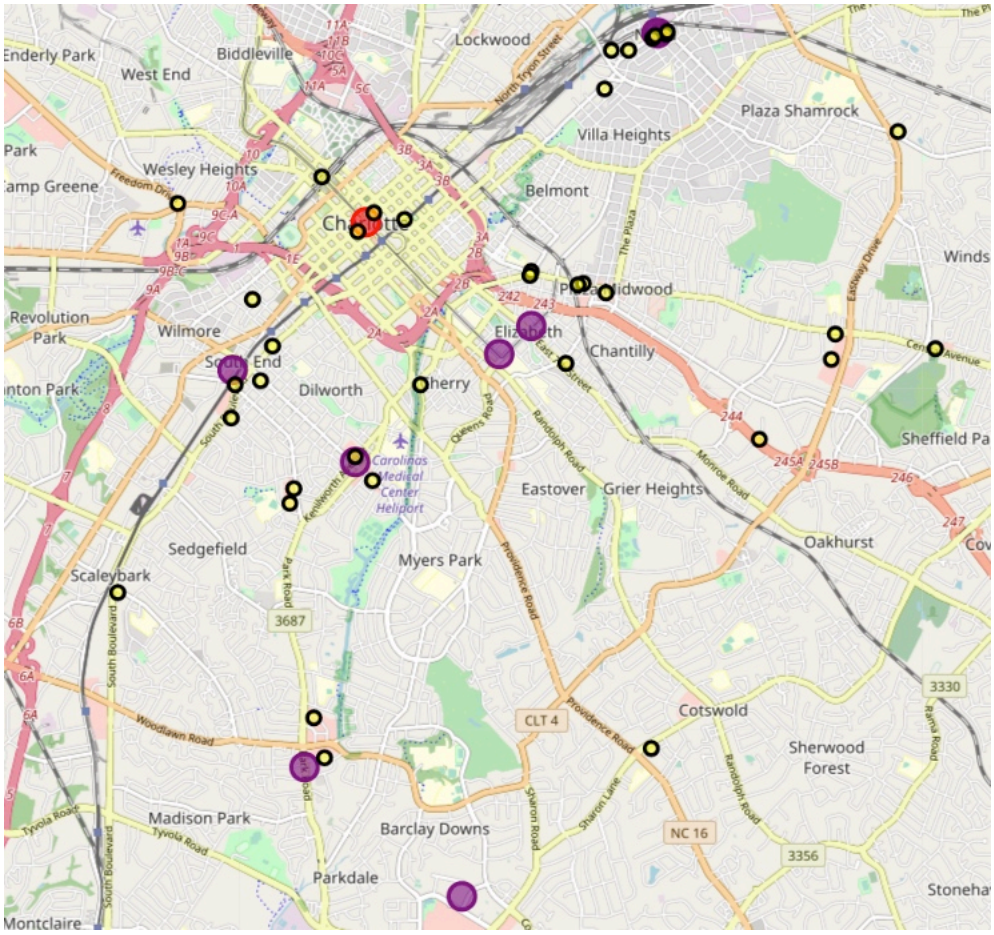
Venue Clustering Map of Charlotte



The heaviest concentrated areas (in yellow) are in Uptown, Elizabeth, South End and NODA. This map displays all 11 of the unique venue categories, so further drilling into specific categories is necessary.

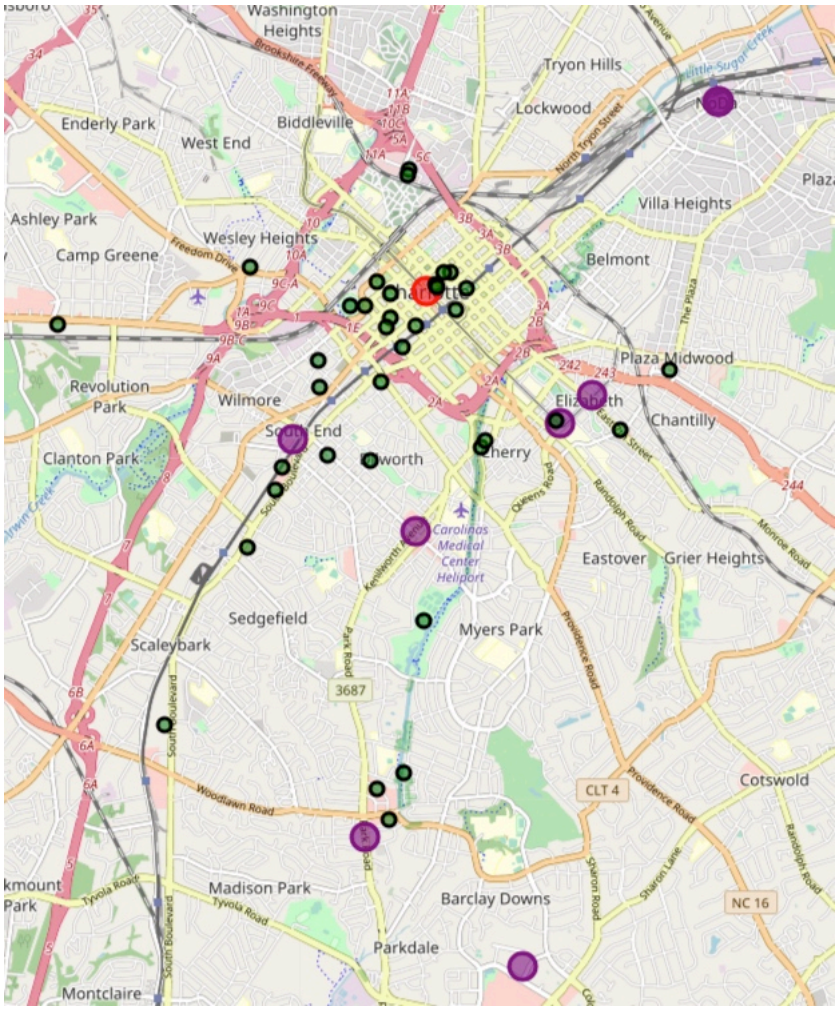
In order to adequately solve the business problem, Latin American-themed restaurants will first be compared against all other restaurants in the Charlotte metro area, followed by a comparison of where the cultural and entertainment venues are compared to the existing Latin American restaurants.

Comparison map of the Latin American restaurants around Charlotte (in purple) and the other restaurants (in yellow)



The eight Latin American restaurants are all located outside of the center of Charlotte. This map exposes yet another gap in the reliability of the Foursquare data. There are definitely more than four restaurants in Uptown Charlotte, for example. Because the project instructions explicitly state to use the Foursquare API, this information will be treated as the best source of information for the purposes of this analysis and therefore considered the most accurate available.

Comparison map of the Latin American restaurants around Charlotte (in purple) and Cultural Attractions, Shopping and Gyms (in green)



Uptown Charlotte contains the heaviest concentration of the cultural, arts, entertainment, gym and shopping venues.

The K-Nearest Neighbors algorithm will be using the labeled points to assign each venue to similar clusters based on their features. First up is to check how many venues were returned for each postal code.

postalCode	name	categories	a
28202	19	19	
28203	20	20	
28204	11	11	
28205	24	24	
28206	4	4	
28207	2	2	
28208	4	4	
28209	8	8	
28211	2	2	
28215	1	1	
28217	3	3	

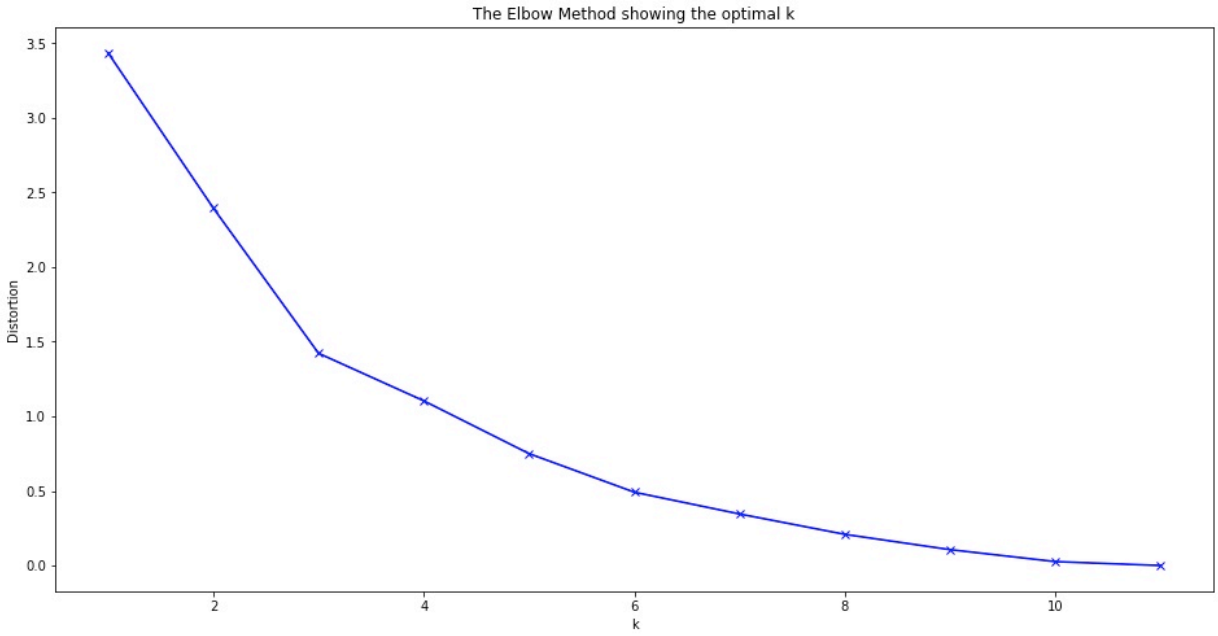
Next, checking the concentration of each venue category per postal code by taking the average of the frequency of occurrence of each type of venue.

postalCode	American Restaurant	Arts and Entertainment	Asian Restaurant	Brewery / Bar	Café	European Restaurant	Food Market	Gym	Hotel	Latin American Restaurant	Specialty Shop
28202	0.052632	0.526316	0.000000	0.000000	0.052632	0.105263	0.105263	0.105263	0.0	0.000000	0.052632
28203	0.200000	0.100000	0.100000	0.150000	0.100000	0.000000	0.150000	0.050000	0.0	0.100000	0.050000
28204	0.090909	0.090909	0.000000	0.090909	0.181818	0.090909	0.181818	0.000000	0.0	0.181818	0.090909
28205	0.375000	0.000000	0.041667	0.291667	0.083333	0.083333	0.041667	0.000000	0.0	0.083333	0.000000
28206	0.000000	0.500000	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000
28207	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.5	0.000000	0.000000
28208	0.250000	0.000000	0.000000	0.250000	0.000000	0.000000	0.500000	0.000000	0.0	0.000000	0.000000
28209	0.375000	0.000000	0.000000	0.000000	0.000000	0.000000	0.250000	0.000000	0.0	0.125000	0.250000
28211	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.500000	0.000000
28215	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000
28217	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000

	postalCode	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 Most Common Venue
0	28202	Arts and Entertainment	Gym	Food Market	European Restaurant	Specialty Shop	Café	Latin American Restaurant	Hotel	Brewery / Bar	Asian Restaurant
1	28203	Food Market	Brewery / Bar	Latin American Restaurant	Café	Asian Restaurant	Arts and Entertainment	Specialty Shop	Gym	Hotel	European Restaurant
2	28204	Latin American Restaurant	Food Market	Café	Specialty Shop	European Restaurant	Brewery / Bar	Arts and Entertainment	Hotel	Gym	Asian Restaurant
3	28205	Brewery / Bar	Latin American Restaurant	European Restaurant	Café	Food Market	Asian Restaurant	Specialty Shop	Hotel	Gym	Arts and Entertainment
4	28206	Brewery / Bar	Arts and Entertainment	Specialty Shop	Latin American Restaurant	Hotel	Gym	Food Market	European Restaurant	Café	Asian Restaurant
5	28207	Hotel	Brewery / Bar	Specialty Shop	Latin American Restaurant	Gym	Food Market	European Restaurant	Café	Asian Restaurant	Arts and Entertainment
6	28208	Food Market	Brewery / Bar	Specialty Shop	Latin American Restaurant	Hotel	Gym	European Restaurant	Café	Asian Restaurant	Arts and Entertainment
7	28209	Specialty Shop	Food Market	Latin American Restaurant	Hotel	Gym	European Restaurant	Café	Brewery / Bar	Asian Restaurant	Arts and Entertainment
8	28211	Latin American Restaurant	Specialty Shop	Hotel	Gym	Food Market	European Restaurant	Café	Brewery / Bar	Asian Restaurant	Arts and Entertainment
9	28215	Asian Restaurant	Specialty Shop	Latin American Restaurant	Hotel	Gym	Food Market	European Restaurant	Café	Brewery / Bar	Arts and Entertainment
10	28217	Brewery / Bar	Specialty Shop	Latin American Restaurant	Hotel	Gym	Food Market	European Restaurant	Café	Asian Restaurant	Arts and Entertainment

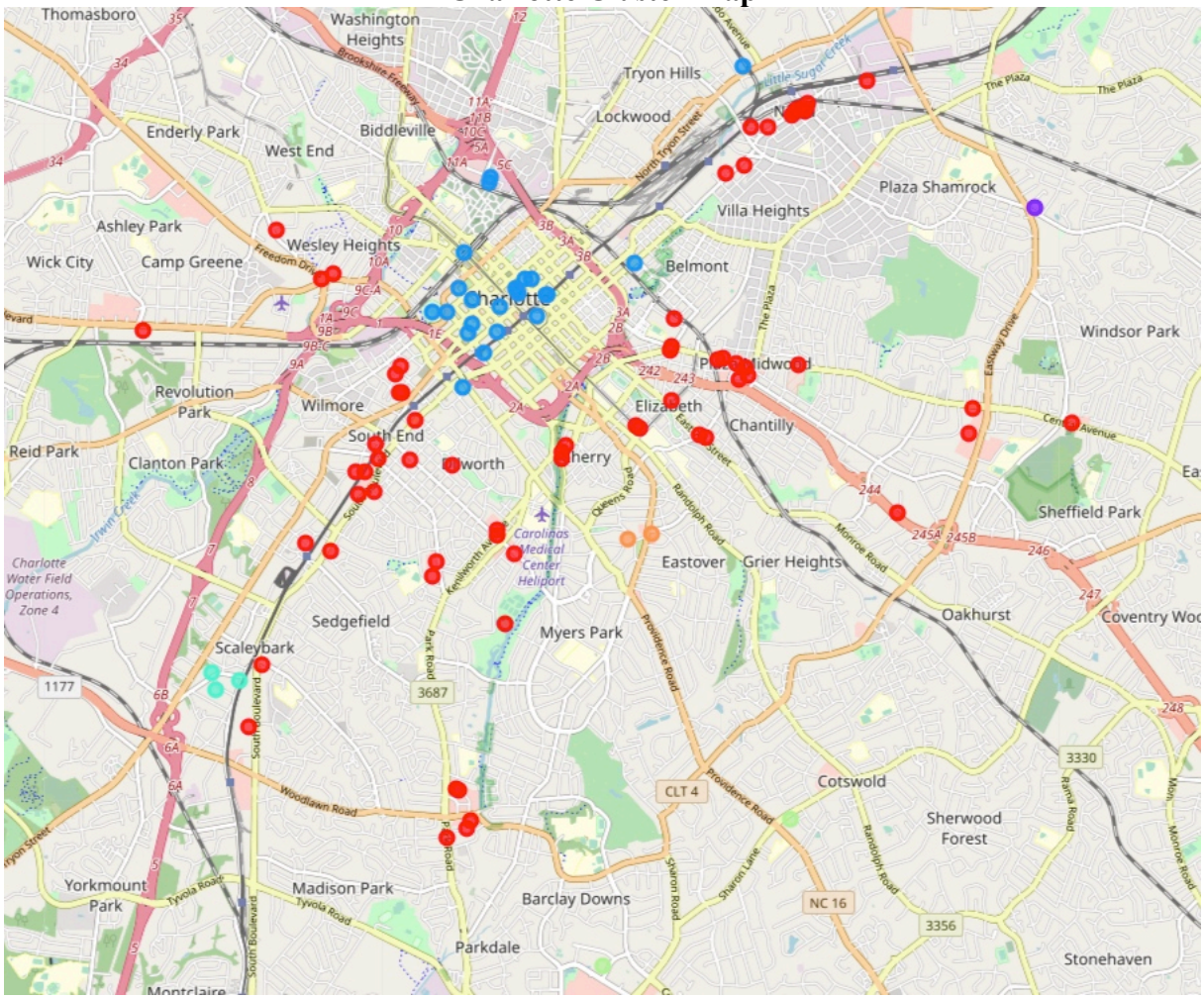
Postal codes 28204 and 28211 list their most common venue as being a Latin American Restaurant, with it being the second most common for 28205, behind Brewery / Bar. Latin American Restaurants rank as the 7th most common venue for Postal Code 28202, with Arts and Entertainment being the most common venue, making it a very attractive opportunity with less saturation and high potential for family traffic.

Finding the optimal value for k is important because if it is too low, the model will be too complex and if it is too high, it will be an overly generalized model, neither of which would be ideal for predictive purposes.



*Setting the number of clusters: based on the elbow method, the graph appears to indicate that 3 could be the optimal k, however several iterations of this clustering analysis produced an outcome that made more sense when using 6 clusters.

Charlotte Cluster Map

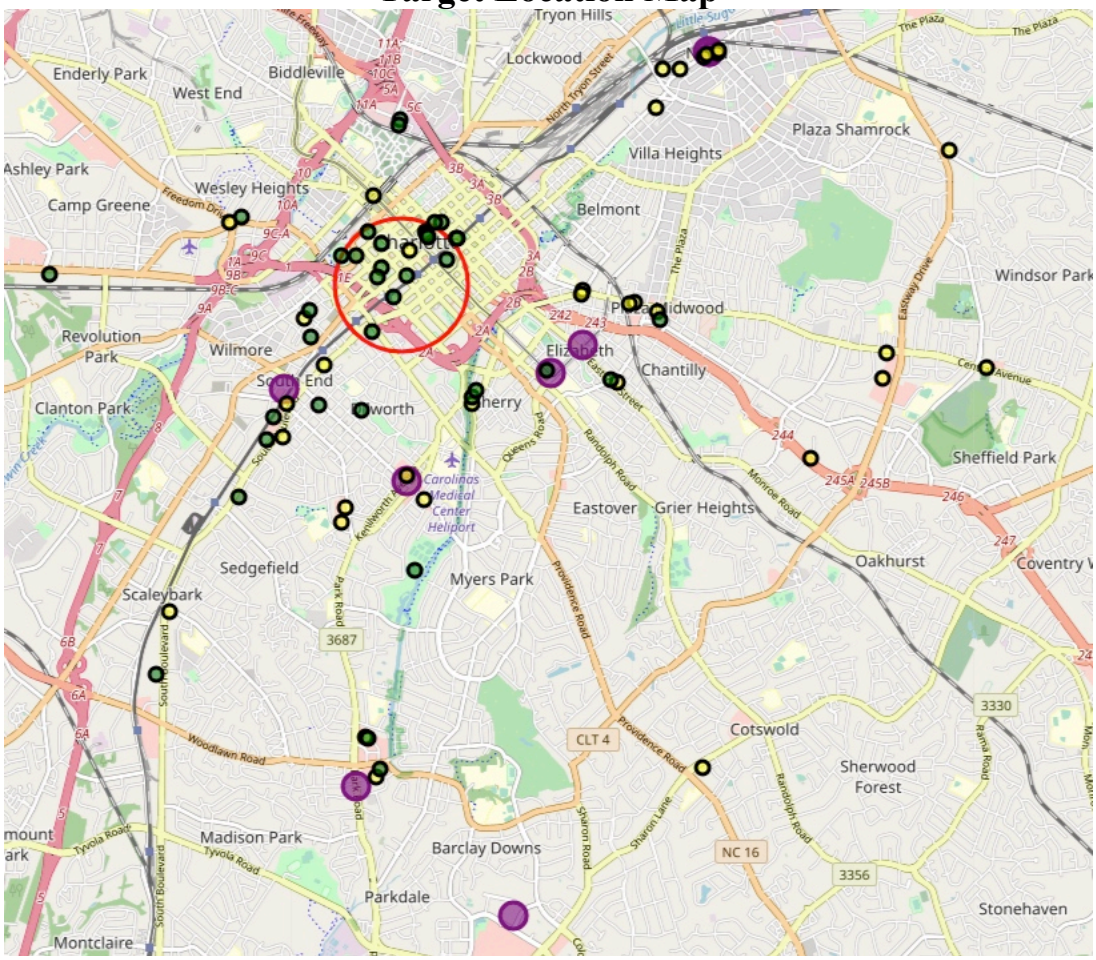


The clusters appear to spiral out from the center of Charlotte, with blue in the center (cluster 2), then a heavy concentration of red (cluster 0), with light blue (cluster 3), orange (cluster 5), purple (cluster 1) and light green (cluster 4) sprinkled here and there.

Results

Based on the family entertainment options, the proximity to downtown businesses and lack of other Latin American-themed restaurants in the cluster, cluster 2 is the ideal cluster.

Target Location Map



This map highlights the lack of penetration that Latin American-themed restaurants have in an area heavily dominated by a thriving arts and entertainment section of Charlotte. Uptown also hosts the corporate headquarters of several fortune 500 companies, most notably Bank of America and Wells Fargo. Therefore, not only could any restaurant expect a steady stream of customers on nights and weekends, but also during the lunch hour, making this a perfect location for the proposed empanada quick service restaurant.

Discussion

As mentioned above, there were some notable gaps that require further investigation. For example, how many restaurants are already in the target area and how can this information be sourced, if not via Foursquare? Why are gyms and hotels not listed in Foursquare? Would the clustering be different if accurate neighborhood labels were available and could be relied upon rather than using postal codes?

It is recommended that further cursory analysis is conducted in the spirit of due diligence to ensure that answering these questions would not fundamentally change the outcome of this analysis.

Conclusion

Assuming that no additional information surfaces during the due diligence phase, it is recommended that Uptown Charlotte be the next expansion target for the empanada franchise. This analysis used location information from Foursquare to find the most common venues per neighborhood, the heaviest concentration of each type of venue and isolated the area with the least saturation of Latin American restaurants.

Having identified the target area, discussions concerning the ideal physical location based on the cost of rent, foot traffic, proximity to parking, amenities, etc., can begin.