IBM Applied Data Science Capstone

Restaurant Analysis in Charlotte, NC, USA

By: Shyam Sridharan

1. Introduction

1.1 Background

Charlotte, North Carolina is a growing city with approximately 873,000+ residents residing in its many neighborhoods. In recent years Charlotte has seen a dramatic increase in population dude to becoming a new banking capitol attracting many businesses to the city. With this surge in business, Charlotte has become a diverse city home to many restaurants of various cuisines with newer restaurants opening frequently. Thanks to apps such as UberEats, DoorDash, Postmates, etc. it has never been easier to enjoy these restaurants from the comfort of your home especially during Covid-19 pandemic where the safest measures are social distancing and enjoying these meals at one's home.

1.2 Business Problem

Given a city and its neighborhoods can we gather information on the restaurants in the area in ordered to be better informed on hotspots for multinational cuisine and locations in need of more variety using data driven analysis?

1.3 Target Audience

We should market this project towards:

- Consumers looking to enjoy a variety of options for cuisines moving to Charlotte
- Entrepreneurs looking to open restaurants in the city
- Drivers for Food Delivery apps trying to maximize amount of deliveries

2. Data

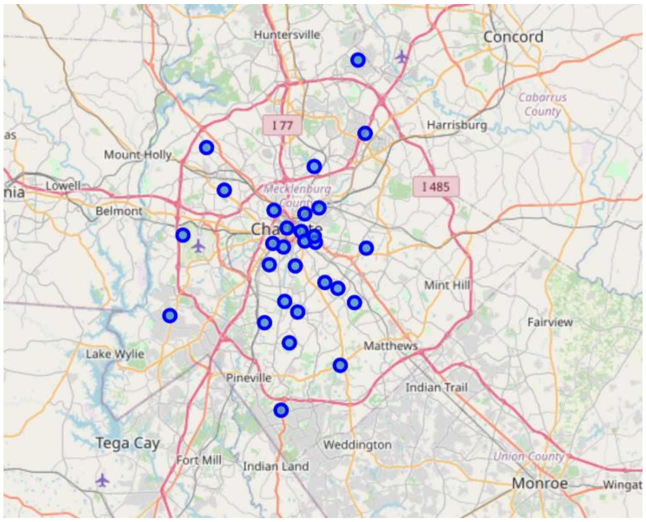
2.1 Data Preparation

The Data Necessary to complete this project is as follows:

- Neighborhood Data for Charlotte NC

- This was procured from Wikipedia from the following site: https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Charlotte,_North_Carolina using the BeautifulSoup package
- Then take some necessary steps to clean the data so that only the names of each neighborhood remain

- Geographical Coordinate Data for each of the Neighborhoods
  - Using the geocoder package we are able to procure each of the Longitude and Latitudes for the perspective Neighborhoods
  - Then attach this location data to the existing Neighborhoods dataframe so now we have three columns; Neighborhood, Latitude, Longitude

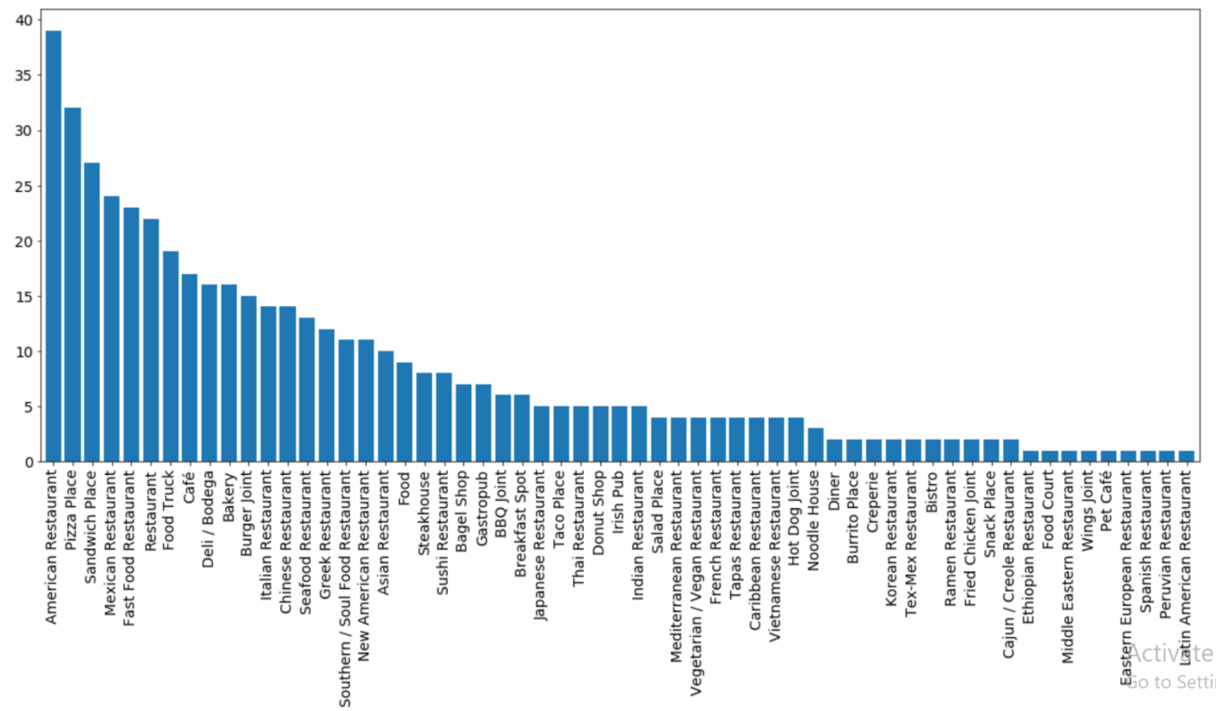


Folium map showing Charlotte neighborhoods with Geographical Coordinates imposed on a map of Charlotte, North Carolina

- Restaurant Venue Data based on Neighborhoods
  - This where the API use comes into play, FourSquare is a service with a large database consisting of venues of all type for countries across the world. What we must do is create a free, or paid if API is used extensively, developer account so we can connect to the API
  - We use our developer credential to make a GET request for restaurant venues by specifying the categoryId for food: “4d4b7105d754a06374d81259” as well as setting a custom radius and limit of results
  - After this API call, we now have a dataframe with 481 venues denoted with their perspective venue category and neighborhood location.

2.2 Exploratory Data Analysis

Now that we have dataset with all of out venues, we are using for the analysis portion of the project we inspect the dataset to see the distribution of our venue categories. These categories will denote the types of restaurants in our dataset i.e. American Restaurant, Pizza Place, etc. To do this we get the value counts for all the venues and display them in a bar chart.



From this we see a large portion of restaurants are American Restaurants, Pizza Places, Sandwich Places, Mexican Restaurants, and Fast Food. This is a similar distribution of restaurants to most of the cities in the country but we will keep that in mind moving forward with our analysis.

3. Modelling

3.1 Model Selection

The next step moving forward is by understanding our data and our objective what modelling approach will help us to solve our problem moving forward and generate the insights needed. Seeing as the goal of our project is to create distinction between the neighborhoods so that the target audience may know moving forward what area lean towards certain restaurant styles, a clustering approach would behoove our analysis. By using clustering, we can group similar neighborhoods together so that we can understand which areas are similar.

3.2 Data Preparation

Now that we know that clustering is the approach to take for this problem, we need to prepare our dataset for modelling. As of know the dataset is very large with 481 rows with each being a single venue. We can focus our analysis rather than on individual venues on venue categories so we can determine what types of restaurants are popular in given neighborhoods. From the exploratory data analysis, we know that we have 60 distinct venue categories so what we did was first one hot encode each row of the dataset or in this case each venue. Turning each observation into a row with the neighborhood and 60 columns where 59 of them are zero and the correct venue category for each venue will be the only column with a value of 1. This transformed dataset has the shape (481, 61). From here we group together all the venues by each distinct neighborhood so now each observation has a proportion value for each of the venue categories in terms of all the venues in the neighborhood. Since many of the venue categories will have small proportional values or a value of zero, we transform the dataset so that for each observation we have the neighborhood and its top 10 venue categories.

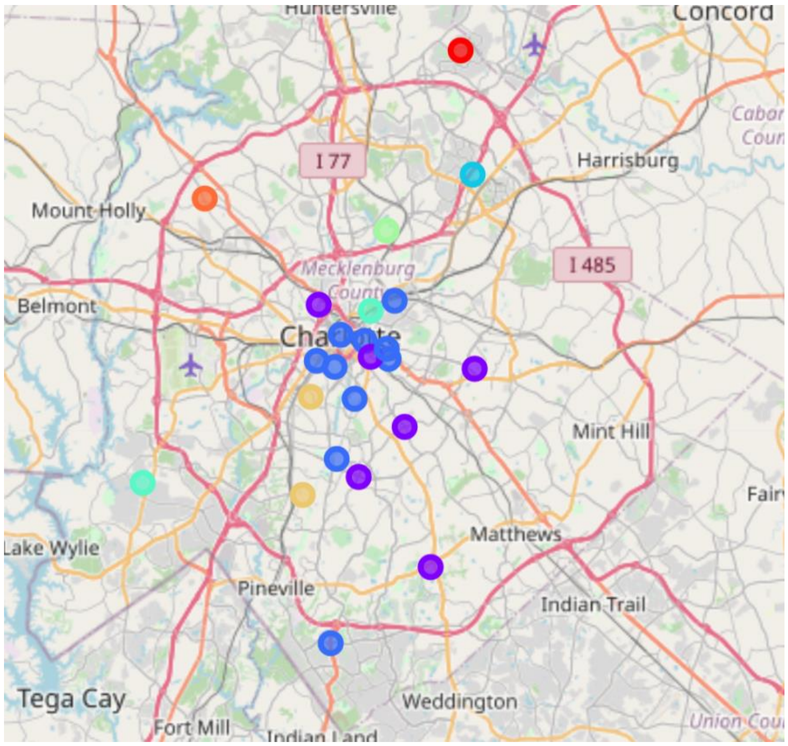
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Ballantyne	American Restaurant	Pizza Place	Asian Restaurant	Burger Joint	Italian Restaurant	Indian Restaurant	Food Truck	Mexican Restaurant	Bakery	Breakfast Spot
1	Biddleville	Café	Fried Chicken Joint	Mediterranean Restaurant	Food	Wings Joint	Diner	Greek Restaurant	Gastropub	French Restaurant	Food Truck
2	Brooklyn	Deli / Bodega	Pizza Place	Food	Southern / Soul Food Restaurant	Restaurant	Mexican Restaurant	Seafood Restaurant	Burger Joint	Taco Place	Food Truck
3	Chantilly	Bakery	Restaurant	Pizza Place	Deli / Bodega	Chinese Restaurant	Southern / Soul Food Restaurant	Sandwich Place	Greek Restaurant	Food Truck	Vietnamese Restaurant
4	Cotswold	Pizza Place	Fast Food Restaurant	Chinese Restaurant	Breakfast Spot	Mexican Restaurant	Greek Restaurant	Sandwich Place	Café	Donut Shop	Burger Joint
5	Coulwood	Gastropub	Mexican Restaurant	Southern / Soul Food Restaurant	Wings Joint	Deli / Bodega	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food
6	Derita	Southern / Soul Food Restaurant	Wings Joint	Deli / Bodega	Greek Restaurant	Gastropub	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food
7	Dilworth	American Restaurant	BBQ Joint	Food Truck	Mediterranean Restaurant	Food	Fast Food Restaurant	Seafood Restaurant	Deli / Bodega	Gastropub	Fried Chicken Joint
8	Eastland	Seafood Restaurant	Fast Food Restaurant	Chinese Restaurant	Mexican Restaurant	Caribbean Restaurant	Sandwich Place	Indian Restaurant	Breakfast Spot	Burger Joint	Gastropub

3.3 Model Deployment

For clustering we will use the sklearn package for kmeans clustering, sklearn is a widely used package for model deployment and machine learning in general with several models built in already just to be called by keyword and specifying arguments. For kmeans clustering all we must specify is the dataset we are committing the clustering analysis on and specifying the number of clusters. For the model remove the neighborhood column and then the dataset will be ready. From there run the cluster experimenting with different number of clusters and interpret the results.

4. Results

After using kmeans clustering several times with various different number of cluster we settled on 8 clusters to properly show the results. Of the 8 clusters; clusters 0, 3, 4, 5, 6, 7 were made up of 2 or less neighborhoods while clusters 1 and 2 were made up of many. We visualized the clusters in another folium map to see the pattern geographically. Clusters 1 and 2 are purple and blue respectively.



5. Conclusions

After completing the cluster analysis, I took a look at the neighborhoods in each cluster to try an observe a pattern as to why the clusters were formed each time, I chose a specific cluster number. It was finally at 8 clusters I saw that the clusters with fewer neighborhoods existed because most of these neighborhoods had food options that existed in all of the neighborhoods such as American restaurants, Fast Food, Mexican restaurants, Burger joints, Hot Dog Joints etc. What was different about clusters 1 and 2 is that these were the neighborhoods with much more variety and multicultural food options such as French, Indian, and different European and Asian cuisines. In fact, with fewer clusters all of the neighborhoods in cluster 1 and 2 would be merged into one cluster. The largest difference between cluster 1 and 2 was in cluster 1 there were more popular options for multicultural restaurants. From this I conclude that If one were looking to open newer multicultural restaurants, they would be popular in clusters 1 and 2 but if you were looking to stand out, they should open in the clusters where its mostly normal American dining. If one were looking for areas to live with vastly different cuisines then they should look for living in the neighborhoods in cluster 1. As for those looking to become deliverers for food service apps look to cluster 2 as many are in close proximity with a variety of options.

6. Future Improvements

For the future these are some of things that I wish to look into and employ into this project to further improve upon it:

- Find other APIs to connect and find more data such as a ratings API
- Do more proximity analysis into neighborhoods and connect to Housing listings to in theory find houses near multiple venues
- Create a web application allowing for the functionality to be interactive
- Procure more venues to look at to get more specific results
- Looking into menus for a more in-depth look concerning dishes similar or dissimilar
- Scrape venues from social media