Applied Data Science Capstone Project

Exploring and Analyzing Neighborhoods in Charlotte, North Carolina

IBM Data Science Professional Certificate

The IBM Data Science Professional program consists of 9 online courses that provide exposure to data tools and skills including open source tools and libraries, Python, databases, SQL, data visualization, data analysis, statistical analysis, predictive modeling, and machine learning algorithms.
This Professional Certificate has a strong emphasis on applied learning.

Introduction

Charlotte is the most populous city in the state of North Carolina.

A real estate development client has commissioned phase one data analysis of nineteen Charlotte neighborhoods with the goal of determining the most popular venues in each area. This client is exploring the feasibility of starting a business venture in Charlotte. However, they want to know which venues are currently successful and where they are located. This information will allow them to perform a phase two deep dive analysis.

Data

A dataset with the required information could not be located, therefore a small dataset was manually created using Google Maps to identify neighborhood names and GPS coordinates.

This data was then imported from CSV format into a Pandas dataframe.
After importing the data, Geopy was used to create a base map for Charlotte, North Carolina. Folium was used to place the neighborhoods into the base map.

Foursquare location data was used to identify the most popular venues in each area.
K-Means clustering was then applied to group the neighborhoods according to their similarity to each other.

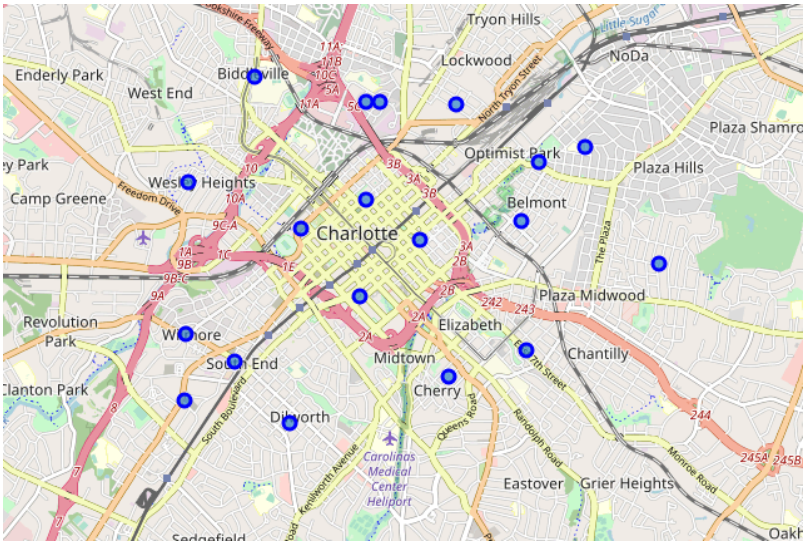
Methodology

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

Load the manually created dataset:

	Neighborhood	Latitude	Longitude
0	Belmont	35.228498	-80.820533
1	Biddleville	35.244668	-80.857104
2	Cherry	35.211191	-80.830472
3	Dilworth	35.205974	-80.852310
4	Elizabeth	35.214095	-80.819952
5	First Ward	35.226395	-80.834404
6	Fourth Ward	35.230943	-80.841794
7	Greenwood	35.241831	-80.840049
8	Lockwood	35.241594	-80.829456
9	Phifer Heights	35.235135	-80.818139
10	Second Ward	35.220138	-80.842736
11	Southend	35.212767	-80.859850
12	Third Ward	35.227657	-80.850846
13	Wesley Heights	35.232926	-80.866287
14	Wilmore	35.215888	-80.866588
15	Villa Heights	35.236783	-80.811898
16	Greenville	35.241829	-80.841868
17	Brookhill	35.208392	-80.866825
18	Plaza Midwood	35.223746	-80.801713

A manually created dataset was imported into a pandas dataframe. There are nineteen distinct neighborhoods.



A base map for Charlotte, North Carolina USA was created and neighborhoods added using GeoPy. GeoPy is a Python client for several popular geocoding web services. geopy makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources.

Foursquare location data was then used to find nearby venues for our group of neighborhoods.

There are 135 unique venue categories in the charlotte_venues dataframe.

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Belmont	9	9	9	9	9	9
Biddleville	4	4	4	4	4	4
Brookhill	26	26	26	26	26	26
Cherry	42	42	42	42	42	42
Dilworth	3	3	3	3	3	3
Elizabeth	28	28	28	28	28	28
First Ward	20	20	20	20	20	20
Fourth Ward	57	57	57	57	57	57
Greenville	17	17	17	17	17	17
Greenwood	6	6	6	6	6	6
Lockwood	4	4	4	4	4	4
Phifer Heights	5	5	5	5	5	5
Plaza Midwood	5	5	5	5	5	5
Second Ward	43	43	43	43	43	43
Southend	42	42	42	42	42	42
Third Ward	30	30	30	30	30	30
Villa Heights	14	14	14	14	14	14
Wesley Heights	17	17	17	17	17	17
Wilmore	3	3	3	3	3	3

One hot encoding was used to convert the categorical variables:

One hot encoding *is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.*

	Neighborhood	Advertising Agency	American Restaurant	Antique Shop	Arcade	Art Museum	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	...	Thrift / Vintage Store	Vegetarian / Vegan Restaurant	Vietnames Restauran
0	Belmont	0	0	0	0	0	0	0	0	0	...	0	0	0
1	Belmont	0	0	0	0	0	0	0	0	0	...	0	0	0
2	Belmont	0	0	0	0	0	0	0	0	0	...	0	0	0
3	Belmont	0	0	0	0	0	0	0	0	0	...	0	0	0
4	Belmont	0	0	0	0	0	0	0	0	0	...	0	0	0

5 rows × 136 columns

The neighborhoods were then grouped by rows and the mean of the frequency of occurrence for each category was calculated.

	Neighborhood	Advertising Agency	American Restaurant	Antique Shop	Arcade	Art Museum	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	...
0	Belmont	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.111111	0.000000	0.000000	...
1	Biddleville	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	...
2	Brookhill	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	...
3	Cherry	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	...
4	Dilworth	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	...
5	Elizabeth	0.000000	0.035714	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	...
6	First Ward	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.050000	0.000000	...
7	Fourth Ward	0.000000	0.070175	0.000000	0.0	0.000000	0.000000	0.000000	0.017544	0.000000	...
8	Greenville	0.000000	0.058824	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	...
9	Greenwood	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	...
10	Lockwood	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	...

```
----Belmont----
                          venue  freq
0                   Deli / Bodega  0.11
1               Athletics & Sports  0.11
2  Southern / Soul Food Restaurant  0.11
3                      Skate Park  0.11
4                      Food Truck  0.11


----Biddleville----
                 venue  freq
0                 Park  0.25
1   Fried Chicken Joint  0.25
2  Gym / Fitness Center  0.25
3                 Café  0.25
4            Nightclub  0.00
```

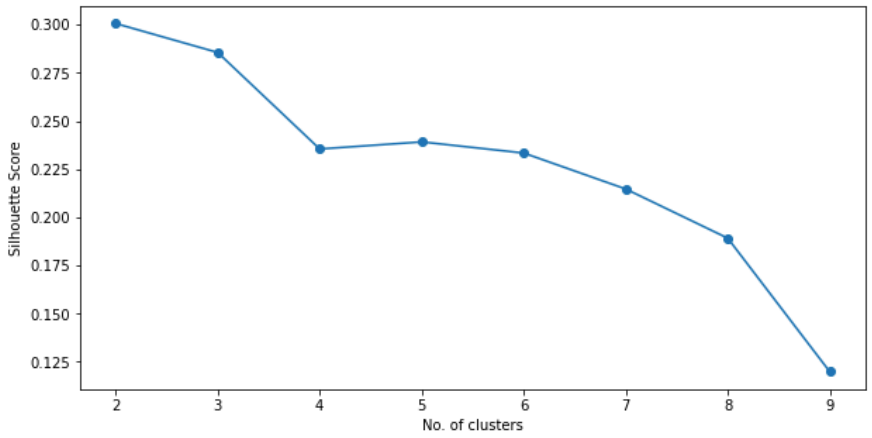
The top five venue categories for each neighborhood was then identified.

A new dataframe was created and the top five most common venues in each neighborhood was identified.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Belmont	Food Truck	Southern / Soul Food Restaurant	Dance Studio	Deli / Bodega	Restaurant
1	Biddleville	Gym / Fitness Center	Park	Café	Fried Chicken Joint	Yoga Studio
2	Brookhill	Coffee Shop	Pet Store	Brewery	Furniture / Home Store	Pool Hall
3	Cherry	Pizza Place	Sandwich Place	Boutique	Fast Food Restaurant	Mediterranean Restaurant
4	Dilworth	Park	Bank	Massage Studio	Yoga Studio	Furniture / Home Store
5	Elizabeth	Sandwich Place	Spa	Pet Store	Cajun / Creole Restaurant	Shipping Store
6	First Ward	Hotel	Cocktail Bar	Creperie	Farmers Market	Food
7	Fourth Ward	American Restaurant	Restaurant	Irish Pub	Theater	Lounge

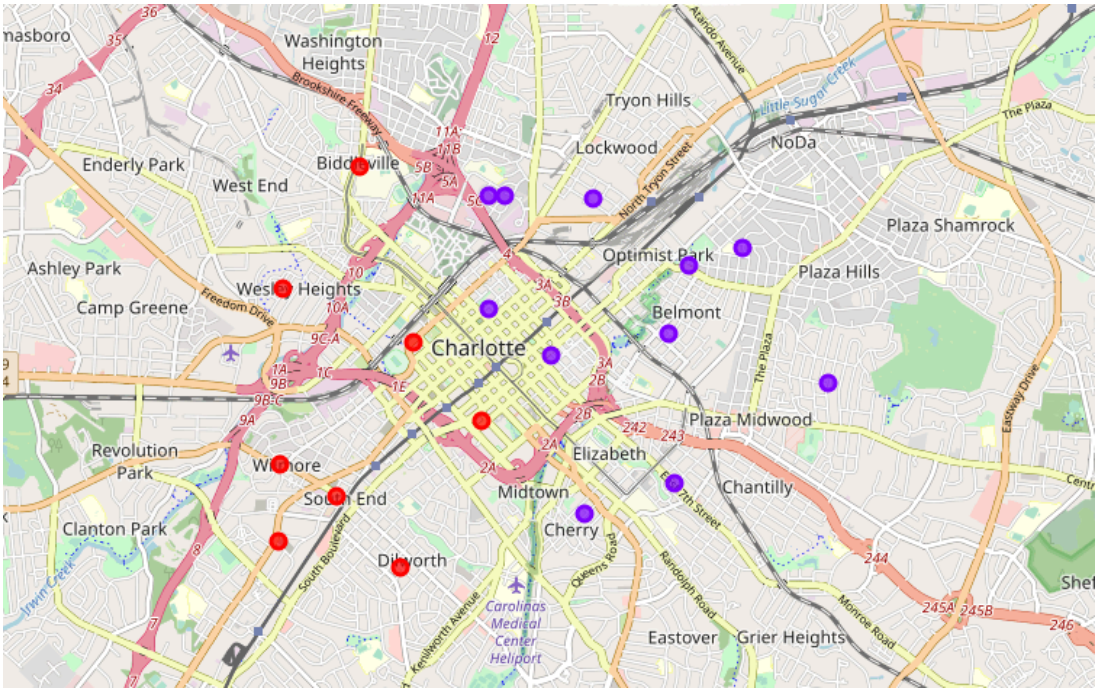
The neighborhoods were then clustered using k-mean. Silhouette analysis was used to identify the optimal number of clusters. In this case, two.

Silhouette Analysis will be used to determine the optimal number of K clusters per dataset, Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].



Plotted Silhouette Score indicates that the optimal number of clusters is two.

K-Means Clustering is an unsupervised machine learning algorithm. The objective of K-means is to group similar data points together and discover underlying patterns.



Results

The results of this initial analysis shows that restaurants are the most popular venues across all of the surrounding neighborhoods. The next step for this client is to perform a phase two deep dive to scout locations and to determine which cuisine has the highest probability of success.

Discussion

Based on the results of this initial analysis, my observations are that to allow for future scalability, a larger area should be analyzed to gather more data points. A two cluster grouping may be two small to gain sufficient insight.

Conclusion

In conclusion, the client was satisfied that they have sufficient information to plan an enterprise in the immediate area with the selected neighborhoods. However, if they desire to expand into the surrounding areas, more analysis is necessary.