IBM Applied Data Science Capstone Project

OPENING AN INDIAN RESTAURANT IN WASHINGTON D.C.

Sidhant Mishra |Coursera Capstone Project | 07/22/2021

Introduction

- ## Background

  There are tons of factors to be considered when you are planning to open a new restaurant out of which location is one of the major factors. Whether you're opening your first full-service restaurant, your second, or your 50th, it's important to understand what to look out for when choosing a new restaurant location. It is of utmost importance to determine the most strategic location in order to attract customers and maximize profit.

- ## Business Problem

  Let us choose a hypothetical situation:

  A client having little to no experience in the restaurant and hospitality business seeks to open a restaurant specializing in Indian cuisine in Washington D.C. area. Being from the DC area himself, he kind of understands the demand which Indian food has and certainly believes that given the right strategic location, he could make it a successful one.

  Taking into account the price level at which the restaurant will operate, the intent is to find an optimal location in the DC neighborhood, where there is significant demand for other cuisines and which already has a good footfall that can be capitalized.

  The objective of this capstone project is to locate the optimal neighborhood for operation using unsupervised machine learning techniques. We would be using the data available on the internet to extract all neighborhoods of Washington D.C and then use Foursquare API to get an idea of different places of interest in each neighborhood.

- ## Target Audience

  Entrepreneurs seeking to establish a new restaurant of a certain niche would be able to get an idea of how choosing an optimal location would give them a competitive advantage and help stay ahead in the game.

Data Acquisition, Cleaning and Methodology

- ## Data Sources

  To perform this analysis, the following set of data would be required

  1. List of Washington D.C. neighborhoods.
  2. Geo coordinates of all the neighborhoods.
  3. Top venues in each of the neighborhoods.

  The list of neighborhoods can be scraped from here. Geographical coordinates for each neighborhood can be obtained using the geocoder tool in the notebook. Data pertaining to top venues would be retrieved using Foursquare API. One has to register for a Foursquare developer account here to access their API credentials.

- ## Data Cleaning and Methodology

  Data downloaded from the web page would be stored in a data frame. We would only be utilizing data that do not contain any junk or unnecessary values that would create problems going forward. In order to do so, we would need to make the data ready for analysis.

  1. Only the cells that have an assigned neighborhood would be processed and the rest would be ignored.
  2. Sometimes the geocoder package fails to retrieve geo coordinates of a neighborhood. In that case, the neighborhoods that would not have a corresponding geo coordinate would be ignored from any further analysis going forward.

  Once the data is available for use, we would be exploring the neighborhoods using Foursquare API and then use machine learning technique (K- means clustering) and map visualization (Folium) to create and visualize different neighborhood clusters.

Business Logic and Methodology

- ## Business Logic

  Since we are looking for the most optimal neighborhoods as our probable locations for the restaurant, it is important that we place some parameters on the basis of which we are going to assess neighborhoods/clusters.

  1. Using the Foursquare API's explore function we would be able to return the neighborhoods that have frequently occurring Indian restaurants. The higher the frequency, the more the competition. The assumption of our analysis is that the barrier of entry to establish a new restaurant in a competitive market is high as existing Indian restaurants may have the competitive advantage of brand loyalty. Therefore, we would not be exploring such neighborhoods for this particular analysis.
  2. We would be prioritizing neighborhoods that already have presence of cafes, restaurants specializing in other cuisines, bars, coffee shops etc. Such neighborhoods would already have a guaranteed footfall and we can easily capitalize on it.

- ## Methodology

  ### 1. Use Web Scrapping to get the list of neighborhoods

  We would be using the list of Washington D.C neighborhoods available in this webpage and then use web scrapping technique by implementing beautiful soup packages to get the neighborhoods list and store it in a python dataframe. The data would look something like this:

	neighborhood
1	Neighborhoods in Northeast
2	Adams Morgan
3	Anacostia
4	Barney Circle
5	Barry Farm
6	Benning Ridge
7	Berkley
8	Blagden Alley-Naylor Court Historic District
9	Bloomingdale
10	Brightwood Park

2. Use GeoPy to get the coordinates of respective neighborhoods

By using GeoPy, we can retrieve the coordinates for each of the neighborhoods. The coordinates can then be split to get the latitude and longitude values.

Note : Due to the unpredictability of GeoPy, the coordinates could not be retrieved for some of the neighborhoods, hence we had to exclude them from any further analysis.

	Neighborhood	Latitude	Longitude
0	Adams Morgan	38.921500	-77.042199
1	Anacostia	38.862581	-76.984441
2	Barney Circle	38.880121	-76.985114
3	Barry Farm	38.859804	-76.996971
4	Benning Ridge	38.881351	-76.938630
5	Berkley	38.952072	-77.098358
6	Blagden Alley-Naylor Court Historic District	38.907115	-77.024980
7	Bloomingdale	38.916778	-77.011365
8	Brightwood Park	38.956734	-77.027992
9	Brightwood	38.965633	-77.027115
10	Brookland	38.932832	-76.984226

3. Explore neighborhoods

In the next step, we would be exploring the neighborhoods using FourSquare API i.e. top venues would be retrieved for each of the neighborhoods. The data from Foursquare is received in JSON format and is then converted to Python dataframe. Venues are collected within a radius of 500m from the point of neighborhood coordinates. The top venues for each district are then grouped as below:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Adams Morgan	55	55	55	55	55	55
Anacostia	6	6	6	6	6	6
Barney Circle	36	36	36	36	36	36
Barry Farm	7	7	7	7	7	7
Benning Ridge	4	4	4	4	4	4
Berkley	3	3	3	3	3	3
Blagden Alley-Naylor Court Historic District	57	57	57	57	57	57
Bloomingdale	32	32	32	32	32	32
Brightwood	21	21	21	21	21	21
Brightwood Park	34	34	34	34	34	34

4. Analyze each neighborhood

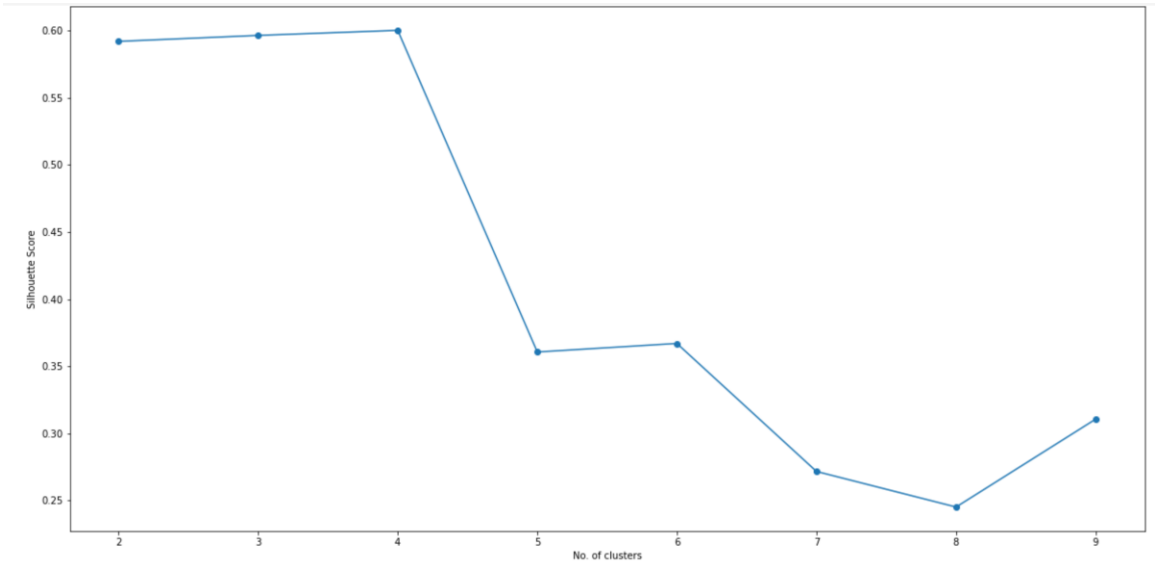
We would now be analyzing each of the neighborhoods and check the venue categories of the top venues that we had retrieved previously. After that, by using one hot encoding, we create dummy variables for each of the categories.

Once the one hot encoded dataset is available to us, we perform normalization on the dataset and get the top ten most common venues for each neighborhood.

	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	Adams Morgan	Bar	Coffee Shop	Ice Cream Shop	Mediterranean Restaurant	BBQ Joint	Asian Restaurant	Cocktail Bar	Diner	Burger Joint	Japanese Restaurant
1	Anacostia	Convenience Store	American Restaurant	Grocery Store	Gym	Sandwich Place	History Museum	Drugstore	Dumpling Restaurant	Eastern European Restaurant	Electronics Store
2	Barney Circle	Sandwich Place	Liquor Store	Gym / Fitness Center	Intersection	Coffee Shop	Harbor / Marina	Mobile Phone Shop	Food Court	Fast Food Restaurant	Snack Place
3	Barry Farm	Bus Stop	Convenience Store	Rental Car Location	Metro Station	Intersection	Basketball Court	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Falafel Restaurant
4	Benning Ridge	Convenience Store	Burger Joint	Park	Insurance Office	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Fish & Chips Shop	Exhibit
5	Berkley	Yoga Studio	Park	Business Service	Ethiopian Restaurant	Eastern European Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Escape Room	Event Space

5. Cluster Neighborhoods

For clustering, K-means method will be applied. To be able to select the optimal number of clusters, the silhouette score will be used.



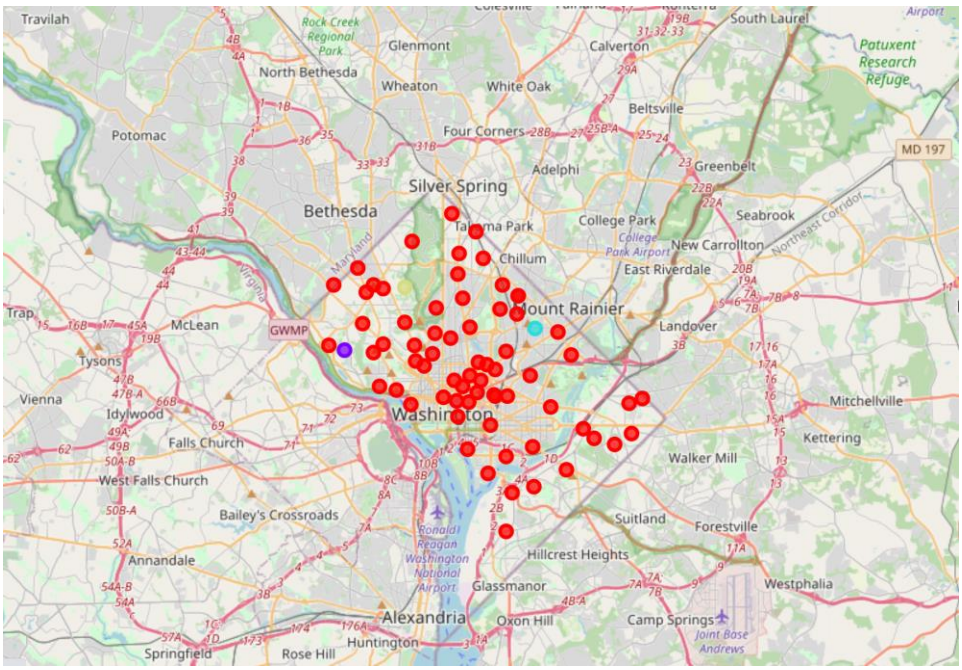
As we can see, the score is highest for k=4. So, there would be a total of 4 clusters.

In the next step, we would be running the K-means algorithm for k=4. When done, we add the cluster labels to the dataset which we had previously created comprising of top venues. This would then be merged with the dataframe that was created in the beginning comprising of neighborhood names and coordinates.

As defined in the business logic, we would not want to consider neighborhoods that already have Indian restaurants. Hence we would be removing those neighborhoods from our dataframe. Below image shows the list of neighborhoods that have Indian restaurants in the top 10 most frequently occurring venues.

	Neighborhood	Latitude	Longitude	Cluster Labels	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
23	Dupont Circle	38.912423	-77.041251	0.0	Pizza Place	Italian Restaurant	Spa	Steakhouse	Thai Restaurant	Cocktail Bar	Cosmetics Shop	Bistro	Park	Indian Restaurant
40	Judiciary Square	38.896086	-77.016681	0.0	Theater	Bar	Ramen Restaurant	American Restaurant	Indian Restaurant	Cocktail Bar	Italian Restaurant	Japanese Restaurant	Food Truck	Dive Bar
50	Mount Vernon Square	38.902531	-77.022948	0.0	Hotel	American Restaurant	Italian Restaurant	Bar	Coffee Shop	Cocktail Bar	New American Restaurant	Mediterranean Restaurant	Indian Restaurant	Gym / Fitness Center
59	Penn Quarter	38.895896	-77.022268	0.0	American Restaurant	Theater	Indian Restaurant	Monument / Landmark	Pizza Place	Cocktail Bar	Salad Place	Plaza	Italian Restaurant	Basketball Stadium
67	Sheridan-Kalorama	38.912155	-77.050659	0.0	Spa	Park	History Museum	Coffee Shop	Salon / Barbershop	Bagel Shop	Nail Salon	Ice Cream Shop	Tea Room	Indian Restaurant
71	Swampoodle	38.903528	-77.002316	0.0	Coffee Shop	Sandwich Place	Gym / Fitness Center	Indian Restaurant	Pharmacy	Liquor Store	Food Truck	Grocery Store	Hotel	Yoga Studio
76	West End	38.907056	-77.049699	0.0	Hotel	Gym	Indian Restaurant	Coffee Shop	Café	Turkish Restaurant	Hotel Bar	Bar	New American Restaurant	Pizza Place

After removing the above neighborhoods, we get the final list which can be visualized using Folium map



Results

- ## Understanding the clusters

By looking at the data, we could figure out that Cluster 1(Cluster label 0) is the one that we are most interested in.

1. Cluster 1 (Cluster Label 0)

Our first cluster contains cafes, coffee shops, restaurants of different cuisines, pubs etc. This is the most dominant cluster and contains most of the neighborhoods of Washington D.C. By looking at the venue category type, its clear that the cluster contains all the commercial neighborhoods that are busy areas where people come for work, tourism etc. Hence considering the heavy footfall and presence of other businesses, the neighborhoods in this cluster appear to be the best ones for the restaurant to be opened. The below image shows first five neighborhoods belonging to this cluster.

	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	Adams Morgan	Bar	Coffee Shop	Ice Cream Shop	Mediterranean Restaurant	BBQ Joint	Asian Restaurant	Cocktail Bar	Diner	Burger Joint	Japanese Restaurant
1	Anacostia	Convenience Store	American Restaurant	Grocery Store	Gym	Sandwich Place	History Museum	Drugstore	Dumpling Restaurant	Eastern European Restaurant	Electronics Store
2	Barney Circle	Sandwich Place	Liquor Store	Gym / Fitness Center	Intersection	Coffee Shop	Harbor / Marina	Mobile Phone Shop	Food Court	Fast Food Restaurant	Snack Place
3	Barry Farm	Bus Stop	Convenience Store	Rental Car Location	Metro Station	Intersection	Basketball Court	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Falafel Restaurant
4	Benning Ridge	Convenience Store	Burger Joint	Park	Insurance Office	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Fish & Chips Shop	Exhibit

2. Cluster 2 (Cluster Label 1)

This cluster has just one neighborhood that has mostly museums and exhibition centers and hence wouldn't be ideal for a restaurant business.

	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
32	Foxhall	Museum	Zoo Exhibit	Exhibit	Electronics Store	Empanada Restaurant	Entertainment Service	Escape Room	Ethiopian Restaurant	Event Space	Falafel Restaurant

3. **Cluster 3 (Cluster Label 2)**

   This cluster has just one neighborhood that has houses and exhibition centers and appears to be a residential area, hence wouldn't be ideal for a restaurant business.

	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
10	Brookland	Boarding House	Zoo Exhibit	Exhibit	Electronics Store	Empanada Restaurant	Entertainment Service	Escape Room	Ethiopian Restaurant	Event Space	Falafel Restaurant

4. **Cluster 4 (Cluster Label 3)**

   This cluster has just one neighborhood that has art exhibition centers and event spaces, hence wouldn't be ideal for a restaurant business.

	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
29	Forest Hills	Public Art	Event Space	Dumpling Restaurant	Eastern European Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Escape Room	Ethiopian Restaurant	Zoo Exhibit

Discussion

By analyzing each of the clusters, it is evident that cluster 1 appears to be the promising one for the new restaurant to be opened. By looking at the venue category type, it's clear that the cluster contains all the commercial neighborhoods that are busy areas where people come for work, tourism etc. Hence considering the heavy footfall and presence of other businesses, the neighborhoods in this cluster appear to be the best ones for the restaurant to be opened.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, analyzing and clustering the data into 4 different clusters using various machine learning techniques based on the venue category types and finally visualizing them using Folium map. We were able to provide recommendations to our client on the ideal neighborhoods for opening an Indian restaurant. The findings of this project would help potential restaurateurs to capitalize on opportunities present in high potential locations while avoiding competition at the same time.