**Capstone Project**

**Finding a New York City Neighborhood similar to DC Shaw**

**R. A**

**May 2020**

# Introduction

Sophie and her family have been living in the Shaw neighborhood in Washington DC for Many years. They have come to appreciate the precious mix of new restaurants, art gallery and entertainment that has come to define this neighborhood. They enjoy the quiet and peaceful atmosphere and the convenient proximity of small shops and fitness facilities.

Sophie was just offered her dream job in New York City and is due to start next month. She's excited to start her new job but feels very sad to leave her Shaw neighborhood she loves so much.

As she's planning her move and looking for a new place to call home in NYC, we are going to use the new skills we learned in this class to help her choose a neighborhood that is the most similar to her beloved Shaw.

# Data Acquisition and Cleaning

## 2.1 Data Sources

For this project we will mainly use Foursquare location data to help describe the landscape of our DC and New York neighborhoods in terms and venues and activities. We also include to demographic and crime data to help break the tie if more one NY neighborhood matches Shaw.

* We use the Foursquare API to explore DC’s Shaw and New York city neighborhood and identify the venue categories in each neighborhood.
* The Shaw neighborhood demographic and crime data is extracted from several tables in <https://opendata.dc.gov/>
* The New York city neighborhood demographic and crime data is extracted from several tables in <https://opendata.cityofnewyork.us/>

## 2.2 Data Cleaning

Data was collected for the various Open Data web sites for New York and Washington DC and combine it with venues data from Foursquare data.

In order to get consistent data across all our neighborhoods, we decided to use demographic data from the 2010 Census for availability reasons.

We had access to crime data by neighborhood for DC. For New York City, the crime data was available at the precinct level. Unfortunately, we didn’t have an obvious way of uniquely assigning precincts to neighborhoods. We decided to use borough level crime for New York city.

Foursquare returns a very granular categorization of the venues. To allow a more meaningful analysis we grouped the venues into 25 disjointed categories.



# Methodology

In order to identify neighborhoods in New York City that better match the Shaw neighborhood, we first need to draw a profile of the Shaw. We use the Foursquare tool **explore** function to get the most common venue categories in the Shaw Neighborhood. We then explore the New York City venues using the categories highlighted in the Shaw neighborhood.

For each neighborhood, we calculate the frequency of each category and the per capita frequency for better comparison.

We finally run a K-means clustering algorithm on a population comprised of Shaw and New York City neighborhood using the per capita number of selected categories as features. The Neighborhood in the cluster containing Shaw are candidates for our recommendation. The 2010 crime rate will be used as tie breaker to select a unique neighborhood.

# Exploratory Data Analysis

## 4.1 Exploring the DC Shaw Neighborhood

Shaw is a central neighborhood in the Northwest quadrant of Washington, D.C., United States. Shaw and the U Street Corridor historically have been the city's black social, cultural, and economic hub, witness to Martin Luther King, Jr., Malcolm X, and numerous riots, marches, and protests that fought to achieve racial equality in Shaw and the entirety of America.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Neighborhood** | **Latitude** | **Longitude** | **pop\_2010** | **Crime\_2010\_per1000** |
| Shaw | 38.921257 | -77.023415 | 12,174 | 17.00 |

A picture containing text, map

Description automatically generated

Fig 1: DC Neighborhood Map – Green Dot represents the Shaw neighborhood

Our Foursquare exploration shows the following distribution by venue category:



## 4.2 Exploring New York City Neighborhoods

For this project, we’re considering New York City neighborhood tabulation areas (NTA). We have a total of 170 NTAs across 5 boroughs.



Our Foursquare exploration returns the following counts by borough for the categories selected in the Shaw neighborhood.



A close up of a map

Description automatically generated

Fig 2. Map Showing New York City Neighborhoods

# Cluster Analysis

Our cluster analysis is based on per capita count for the NYC and Shaw neighborhood for the following categories:

* RESTAURANT
* BAR
* THEATER / SHOW
* GYM / SPORT
* SPECIALTY STORE / SERVICE

## 5.1 Determining the Number of Clusters

One method to determine the optimal number of clusters is the Elbow method. In this method, the sum of distances of observations from their cluster centroids, called Within-Cluster-Sum-of-Squares (WCSS). This is computed as

A picture containing clock, table

Description automatically generated

Yi is centroid for observation Xi.

We then plot a line chart of the SSE for each value of cluster k. If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best.A close up of a map

Description automatically generated

Fig 3. Elbow graph showing the optimal number of clusters

According to the Elbow graph we determine the clusters number as 6.

## 5.1 Analysis the Clusters

The cluster analysis shows neighborhood groupings based on different value ranges of our selected venue categories. The cluster analysis shows that the Shaw neighborhood is assigned to cluster number 4.

Based on those values, we can characterize the different clusters.

The tables below show various statistics for each cluster and the corresponding description.



A picture containing game

Description automatically generated

Fig 4. This graph visualize the mean values for each cluster



A picture containing text, map

Description automatically generated

Fig 5. Map showing NYC neighborhoods colored by their cluster

## 5.2 Looking at Cluster 4

The Shaw neighborhood is assigned to cluster number 4. This cluster is comprised of the neighborhood the high to moderate concentration of all selected venues categories.

The New York City neighborhoods in this cluster are located in Brooklyn and Manhattan and listed in the table below:



From this cluster, we select the neighborhood the highest venue concentration for most categories and the lower crime rate.

A close up of a map

Description automatically generated

Fig 6. Map of the Cluster 4 with the selected neighborhood in green (Fort Greene)

Our recommendation is for Sophie to relocate to the neighborhood of **Fort Greene – Brooklyn**.