Predicting Neighborhood Crime Activity Using FourSquare Venue Data

Chris Robinson

February 2021

Introduction

Background

Washington DC (DC) and New York City (NYC) are cities that offer a wide variety of local venues to satisfy the social, cultural, economic, spiritual, and day-to-day needs of its residents and visitors. As with most large cities, public safety is an ongoing challenge. Predicting patterns of criminal activity is key enabler for improving public safety. This study will make use of local venue data for DC and NYC neighborhoods to derive features that provide a measure of the activity and heterogeneity of the neighborhood venues. I will test this feature as a predictor of criminal activity using both cluster analysis and regression analysis.

Problem Statement

Predict patterns of local crime activity in large city neighborhoods based on statistics that measure the number and variety of local neighborhood venues. These predictions will be useful in improving public safety by informing the general public and public safety officials on how criminal activity is associated with the distribution of venues in a local neighborhood.

Data acquisition and cleaning

Data sources

Washington DC neighborhood labels and geographic position data are available for download in .CSV format from *Open Data DC* (https://opendata.dc.gov). The Washington DC government shares hundreds of datasets via this website.

Data on DC crime incidents was download from the DC Metropolitan Police Department website at https://dcatlas.dcgis.dc.gov/crimecards/all:crimes/all:weapons/2:years/citywide:heat. This data was also obtained in .CSV format.

NYC neighborhood location data was obtained from the "Segmenting and Clustering Neighborhoods in New York City" lab that is part of this IBM Data Science capstone course. The neighborhood dataframe was downloaded from the lab notebook as a .CSV file into my capstone project Git repository.

Data on NYC crime incidents for 2019 was downloaded from NYC Open Data website at https://opendata.cityofnewyork.us/. The data was accessed through the following API call *https://data.cityofnewyork.us/resource/8h9b-rp9u.csv?$where=arrest_date between '2019-01-*

01T00:00:00' and '2020-01-
01T00:00:00'&$$app_token=3RN3HuHQhtXYNzT6OwDSVDKAK&$limit=1000000

Data on local venues in the various DC and NYC neighborhoods will be obtained via API calls to FourSquare.com directly from the project Jupyter notebook.

Data cleaning and preparation

Extraneous data in the raw .CSV files downloaded from DC and NYC government websites was removed using Microsoft Excel. Some preliminary filtering of data was also done in Excel. The .CSV files were then saved with the partially cleaned data to the Git repository established for this project.

Additional cleaning was done to convert date-time strings in the DC crime data to date-time objects. This will be done in Jupyter notebooks using Python. This was necessary in order to filter on crime incidents that occurred in 2019. For the NYC crime incident data, the initial API call limited the response to only incidents that occurred in 2019. Python Pandas was used to extract only the columns needed for subsequent modeling, analysis and interaction with the FourSquare API.

FourSquare neighborhood venue data was downloaded and cleaned through reuse of the Python routines provided in the cluster analysis lab for the IBM™ *Applied Data Science Capstone* course. (See https://labs.cognitiveclass.ai/tools/jupyterlab/lab/tree/labs/DS0701EN/DS0701EN-3-3-2-Neighborhoods-New-York-py-v1.0.ipynb?lti=true)

How data was used to address the problem statement

The analysis used crime data for 2019. Though 2020 data is available and more recent, it was not be used since it is assumed there has likely been significant changes in crime patterns due to COVID-19 (testing this assumption would be an interesting follow on project).

The DC and NYC neighborhood location data will be used to retrieve venue data from FourSquare.com. The venue data will be used to generate the following features:

- Neighborhood venue count, venue category counts, and venue category frequencies
- Neighborhood venue entropy score – measures the diversity of an area as captured by the variety of venue categories within that area (see https://www.cs.uic.edu/~urbcomp2013/urbcomp2016/papers/Exploring.pdf)
    - $Escore = -\sum \frac{N_c}{N} \times \log \frac{N_c}{N}$
    - $N_c = Number\ of\ venues\ of\ category\ c \in C, where\ C\ is\ the\ set\ of\ venue\ cateagories$
    - $N = Total\ number\ of\ venues\ in\ a\ neighborhood$
- Neighborhood venue data was also combined with neighborhood crime data for both cluster and regression analysis. For cluster analysis of the combined data, DBSCAN was used for cluster analysis for its ability to identify outliers.

The DC and NYC crime data will be used to generate the following feature/label:

- Total number of criminal incidents reported in each location/neighborhood for 2019

To count the number of crime incidents for each neighborhood, I will use a Python function to loop through each neighborhood and count incidents within 500 meters of the neighborhood location using the respective crime incident data sets. Figures 1 and 2 below displays the header rows of the baseline data frame for Washington DC and New York City.

	Neighborhood	Total Venues	E-scores	Latitude	Longitude	TotalCrimes	ViolentCrimes	PropertyCrimes
0	16th Street Heights	14	2.639057	38.950315	-77.033559	78	4	74
1	Adams Morgan	57	3.790662	38.920472	-77.042391	473	39	434
2	American University Park	2	0.693147	38.947612	-77.090250	35	4	31
3	Arboretum	15	2.615631	38.914860	-76.972490	86	13	73
4	Barnaby Woods	4	1.386294	38.975433	-77.060174	36	0	36

Figure 1 Washington DC 2019

	Neighborhood	Total Venues	E-scores	Borough	Latitude	Longitude	TotalCrimes
0	Allerton	33	3.026753	Bronx	40.865788	-73.859319	211
1	Annadale	8	1.667462	Staten Island	40.538114	-74.178549	12
2	Arden Heights	4	1.386294	Staten Island	40.549286	-74.185887	19
3	Arlington	5	1.609438	Staten Island	40.635325	-74.165104	120
4	Arrochar	22	2.689178	Staten Island	40.596313	-74.067124	77

Figure 2 New York City 2019

I used multi-variate polynomial regression to construct models to predict annual neighborhood crime incidents using *neighborhood venue count* and *neighborhood venue entropy scores* as features. DC and NYC data were used to train and test the regression models. The various models will be evaluated using R^2 scoring and comparing predicted distributions of crime incidents against actual distributions.

Exploring the Data

Relationship Between Neighborhood Venue Entropy and Total Annual Crimes

To better understand the relationship between crime and the variety and quantity of venues in a neighborhood location, I created the scatter plots below. These plots only look at total annual crime for 2019 versus neighborhood venue entropy score (*E-score*). Figure 3 shows a scatter plot for Washington DC and figure 4 for New York City. The plots indicate a rough non-linear relationship between crime totals for 2019 and the neighborhood *E-score*. Additionally, the scatter plot indicates that cluster patterns may exist.

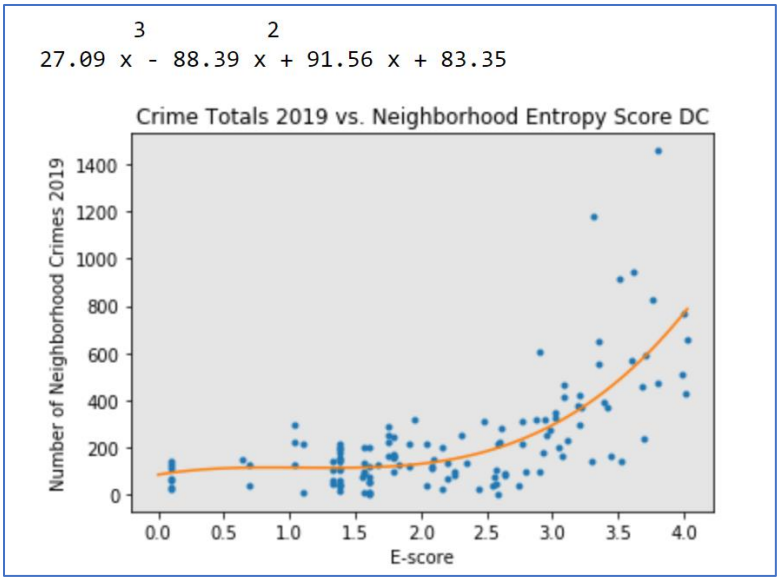

Figure 3 Exploring Relationship between Neighborhood Crime Totals and Venue Entropy Score - Washington DC

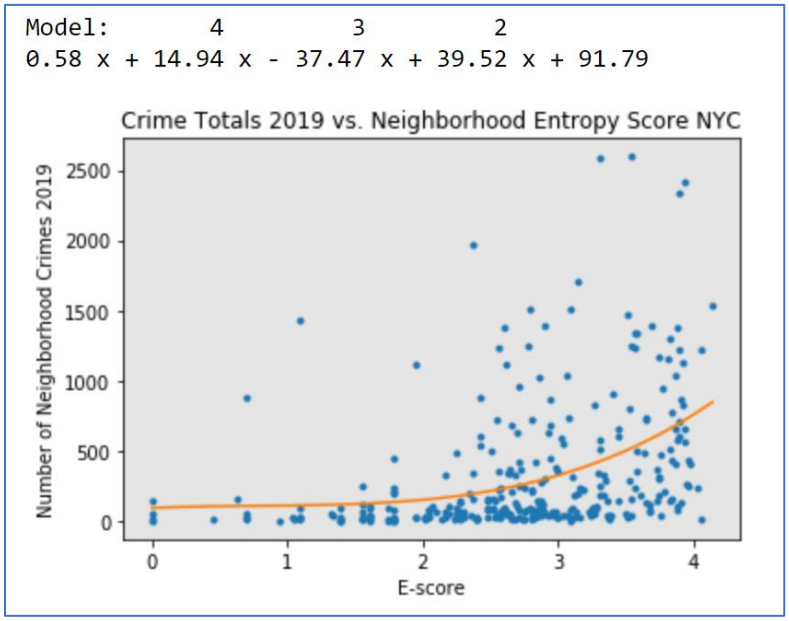

Figure 4 Exploring Relationship between Neighborhood Crime Totals and Venue Entropy Score – New York City

Exploring the data with k-means cluster analysis

To get a sense for patterns in the data, I conducted a k-means cluster analysis on both the DC and NYC data. I used both *squared erro*r and *silhouette score* to try to identify an optimal k value. This proved inconclusive as is shown in figure 5 below. As a result, I decided to conduct a Density Based Scan (DBSCAN) cluster analysis. This DBSCAN analysis achieved better results,

providing useful insight into the crime patterns. The DBSCAN analysis is summarized in the following section.

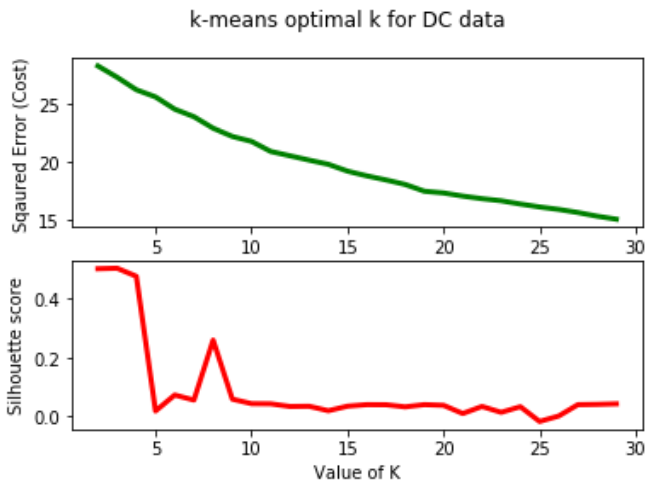

Figure 5 Finding the optimal number of clusters for k-means analysis for DC data

Visualizing neighborhood crime patterns through density based (DBSCAN) cluster analysis

I used *Total Crimes*, *Total Neighborhood Venues*, and *Venue Entropy Score (E-score)* as features for a DBSCAN cluster analysis and then visualized clusters on a map to see how neighborhood crime patterns related to geographic location. The DBSCAN cluster analysis was evaluated using a silhouette score. Silhouette score is measure of the separation distance between the resulting clusters (see https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html). Through trial and error, I evaluated combinations of the DBSCAN parameters *eps* (search distance) and *min_samples* (minimum number of samples within search distance) to find the best clustering result around the feature set described above.

Figure 6 shows the results for New York City neighborhoods. The -1 label indicates outliers. One interpretation, considering Figure 6, is that in the outlier neighborhoods crime activity is less predictable based on *Total Venues* and *E-score.* And from the location pattern of outliers, these less predictable neighborhoods are more prevalent in vibrant, inner city locations such as Mid-town and lower Manhattan and the area of Washington DC extending from the National Mall to Georgetown and up to Dupont Circle (see figure 8).

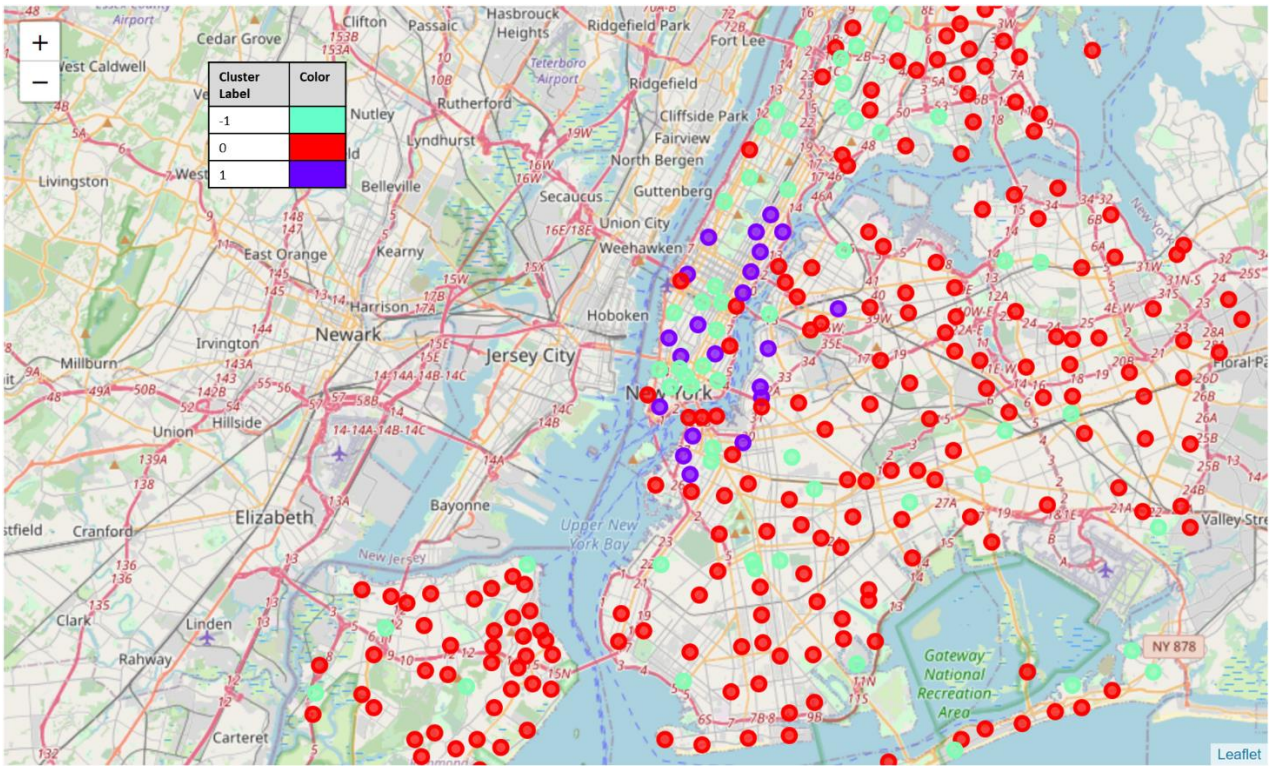

Figure 6 NYC Neighborhood Clusters Based on Venue Features and Annual Crime Totals

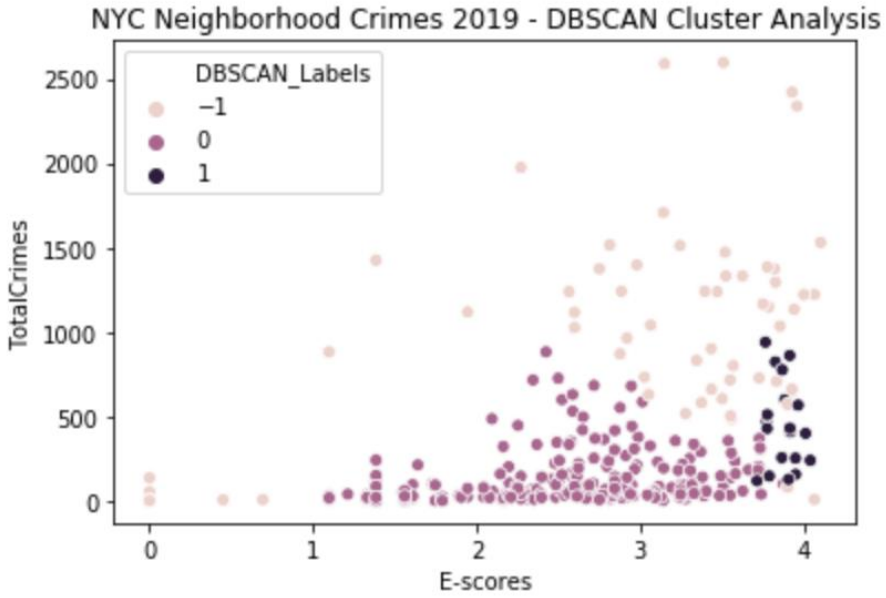

Figure 7 NYC Neighborhood DBSCAN Clusters - Total Crimes vs. E-scores Scatter Plot

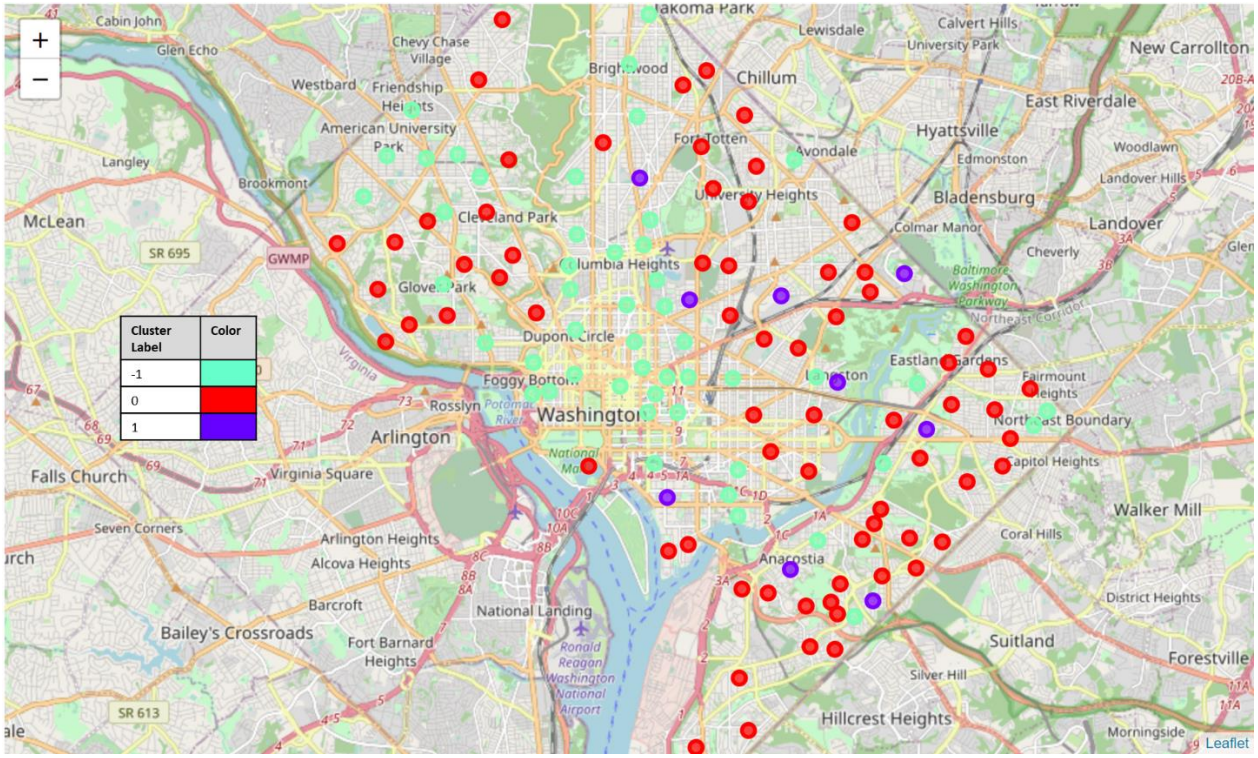

Figure 8 5 DC Neighborhood Clusters Based on Venue Features and Annual Crime Totals

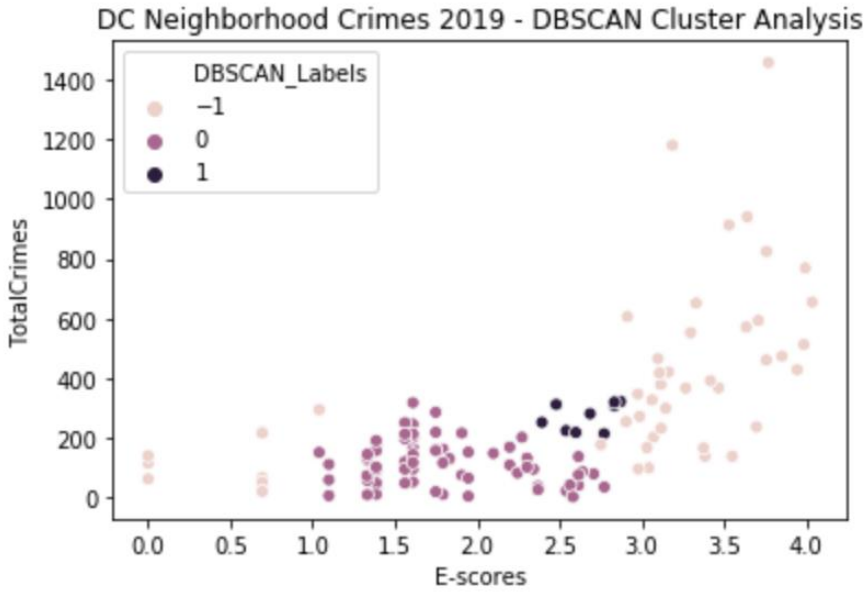

Figure 9 DC Neighborhood DBSCAN Clusters - Total Crimes vs. E-scores Scatter Plot

Both New York City and Washington DC show similar density-based clustering around the selected feature set (*Total Annual Neighborhood Crimes*, *Total Neighborhood Venue Count*, and *E-score*). This is show in the bar charts (figure 10) below as follows:

- **-1** labeled outlier neighborhoods have high crime totals on average, with moderately high E-scores on average, but very high variability in E-scores.
- **0** labeled neighborhoods have very low average annual crime totals and moderate average E-scores with moderate variability. A large majority of neighborhoods are labeled this way in both NYC and DC.
- **1** labeled neighborhoods have moderately high crime totals on average with very high E-scores on average with low variability.

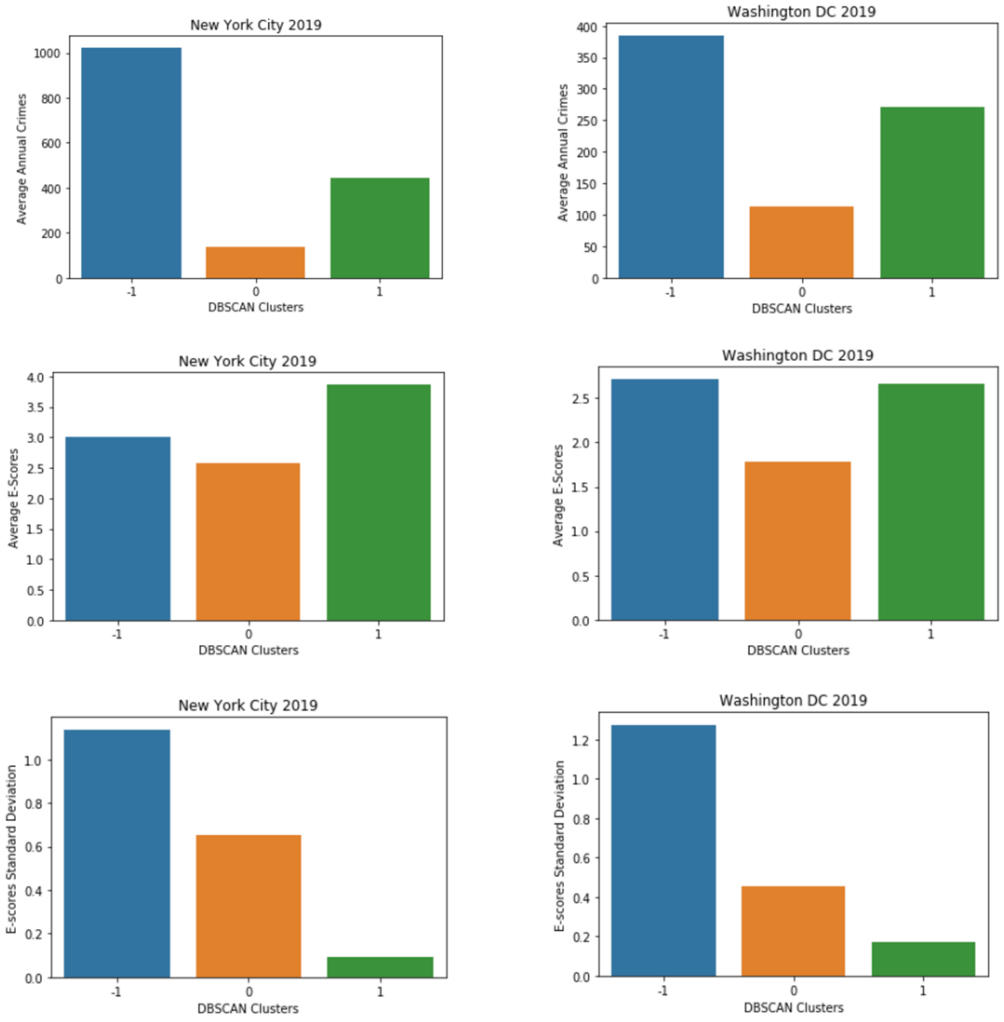

Figure 10 Analysis of Feature Statistics and Cluster Labels - NYC & DC Crime vs. Venue Characteristics

Predictive Modeling

Multi-variate Polynomial Regression with Cross-Validation

Next, I constructed a regression model to predict annual neighborhood crime totals using neighborhood *Total Venues* and *Venue Entropy Score (E-score)* as features. The regression model was evaluation using R^2 score, cross-validation, and distribution plots of actual versus predicted annual neighborhood crime totals.

Figure 11 below shows the mean R^2 results using model cross-validation for different degrees of the multi-variate polynomial fit. It is clear that the polynomial regression does not perform well with the set of features (*Total Venues* and *E-score*). The distribution plots (figure 12) also indicate that the polynomial regression models do not perform well on the data sets, indicating the models do not capture neighborhoods with high crime totals (right tail of the distribution).

Degree for Multi-Variate Polynomial Regression	Mean R^2 Score from Cross-Validation	
	NYC Data	DC Data
1	0.140	0.312
2	0.167	0.165
3	0.191	-0.561
4	0.164	-1.795
5	0.084	-5.200

Figure 11 Cross-Validation R2 Scores for Polynomial Regression Fit to Predict Neighborhood Crime Totals

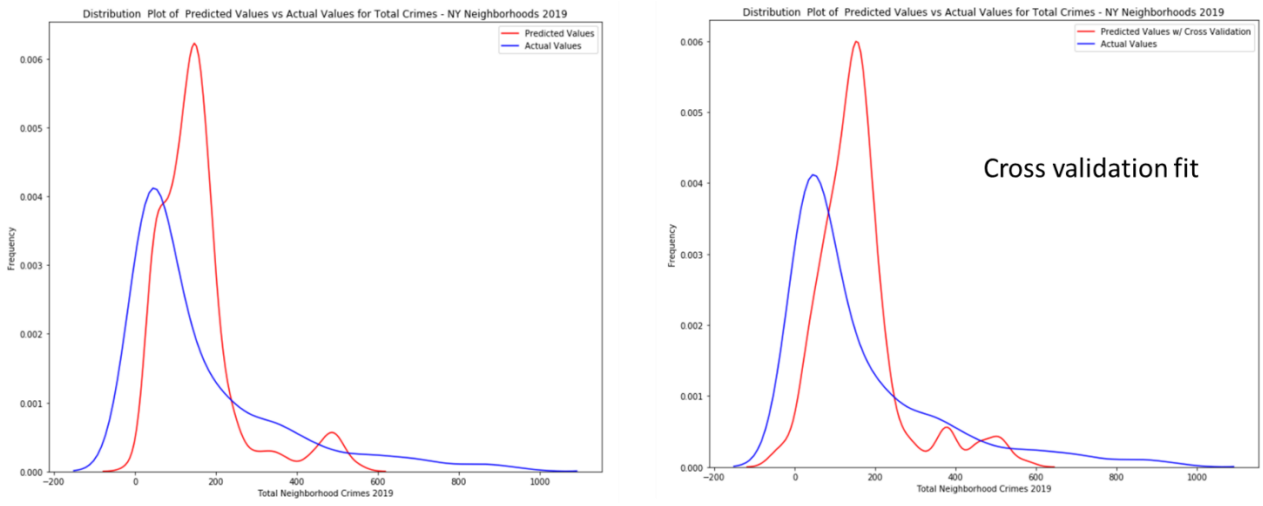

Figure 12 Distribution Plots of Actual vs. Predicted Neighborhood Crime Totals NYC

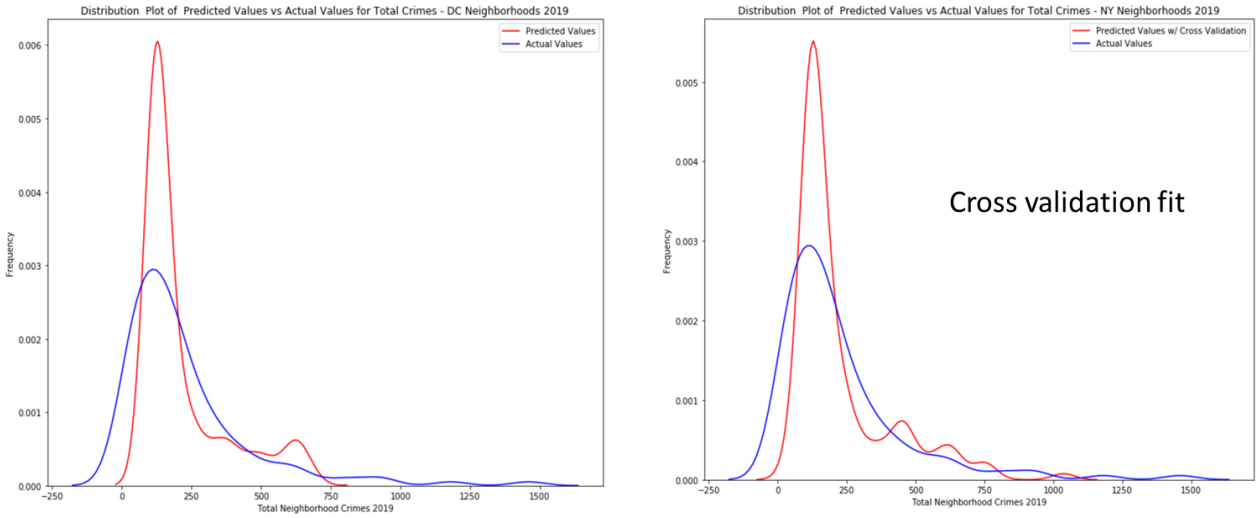

Figure 13 Distribution Plots of Actual vs. Predicted Neighborhood Crime Totals DC

Conclusions

An analysis was conducted to explore the use of FourSquare venue data as a predictor of criminal activity in neighborhoods in New York City and Washington DC.

DBSCAN cluster analysis using *Total Neighborhood Crimes for 2019*, *Total Neighborhood Venues*, and *Neighborhood Venue Entropy Scores* as features was moderately useful in providing insights into crime patterns, especially when cluster labels were visualized geographically.

Multi-variate polynomial regression to predict *Total Neighborhood Crimes* from the feature set of *Total Neighborhood Venues* and *Neighborhood Venue Entropy* did not perform well based on average R^2 scores from cross-validation of the regression models at various polynomial degrees. The NYC regression model performed best at degree 2 and the DC regression at degree 1. Future efforts to improve the model should investigate adding features. These features could be derived from neighborhood socio-economic data and proximity of public safety resources, for example.