Predicting Neighborhood Crime Activity Using FourSquare Venue Data

Chris Robinson

January 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 a feature that provides a measure of the heterogeneity of the neighborhood. 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 MS 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 will be done to convert date-time strings in the DC crime data to date-time objects. This will be done within the project Jupyter notebook 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 will be used to address the problem statement

The analysis will use crime data for 2019. Though 2020 data is available and more recent, it will not be used since I assume 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)
    - $Escore = -\sum \frac{N_c}{N} \times \log \frac{N_c}{N}$
    - $N_c = Number\ venues\ of\ category\ c \in C, where\ C\ is\ the\ set\ of\ venue\ cateagories$
    - $N = Total\ number\ of\ venues\ in\ a\ neighborhood$
- Neighborhood venue data will be used for k-means cluster analysis based on the distribution of most frequent venue categories in each neighborhood. The k-means cluster results will be combined with crime statistics and analyzed for patterns.

- Neighborhood venue data will also be combined with neighborhood crime data for both cluster and regression analysis. For cluster analysis of the combined data, I will use DBSCAN clustering 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.

I will use multi-variate polynomial regression and non-linear regression to construct models to predict annual neighborhood crime incidents based on neighborhood venue count and neighborhood venue entropy scores. DC and NYC data will be used separately and combined to train and test various models. The various models will be evaluated using R^2 scoring and comparing predicted distributions of crime incidents against actual distributions.