# Predicting Crime Activity in Washington DC Neighborhoods

*Chris Robinson*

*December 2020*

## Introduction

### Background

Washington DC is a city that offers 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. Understanding and predicting patterns of crime activity is key enabler for improving public safety. This study will make use of location data for Washington DC neighborhoods to develop features that capture key aspects of human activity correlated with access to a variety of local venues. These metrics will then be used to analyze crime patterns and predict crime activity in Washington DC neighborhoods.

### Problem Statement

Predict patterns of local crime activity in Washington DC neighborhoods from data on local neighborhood venues. These predictions will be useful in improving public safety by informing the general public and public safety officials on how crime activity is associated with number and diversity of local neighborhood venues.

## 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.

Data on local venues in the various Washington DC neighborhoods will be obtained via API calls to FourSquare.com

### Data cleaning and preparation

Extraneous data in the raw .CSV files downloaded from DC government websites was removed using Excel and saving the reduced data set as .CSV file into 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.

### How data will be used to address the problem statement

The DC 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 category count
* Neighborhood venue category data will be used for a cluster analysis. The cluster results will be combined with crime statistics and analyzed for patterns.
* Venue entropy score – measures the diversity of an area as captured by the categories of the venues within that area.

The DC crime data will be used to generate the following labels:

* Total number of criminal incidents reported in each location for 2019
* Total number of violent criminal incidents reported in each location for 2019
* Total number of property related criminal incidents reported in each location for 2019