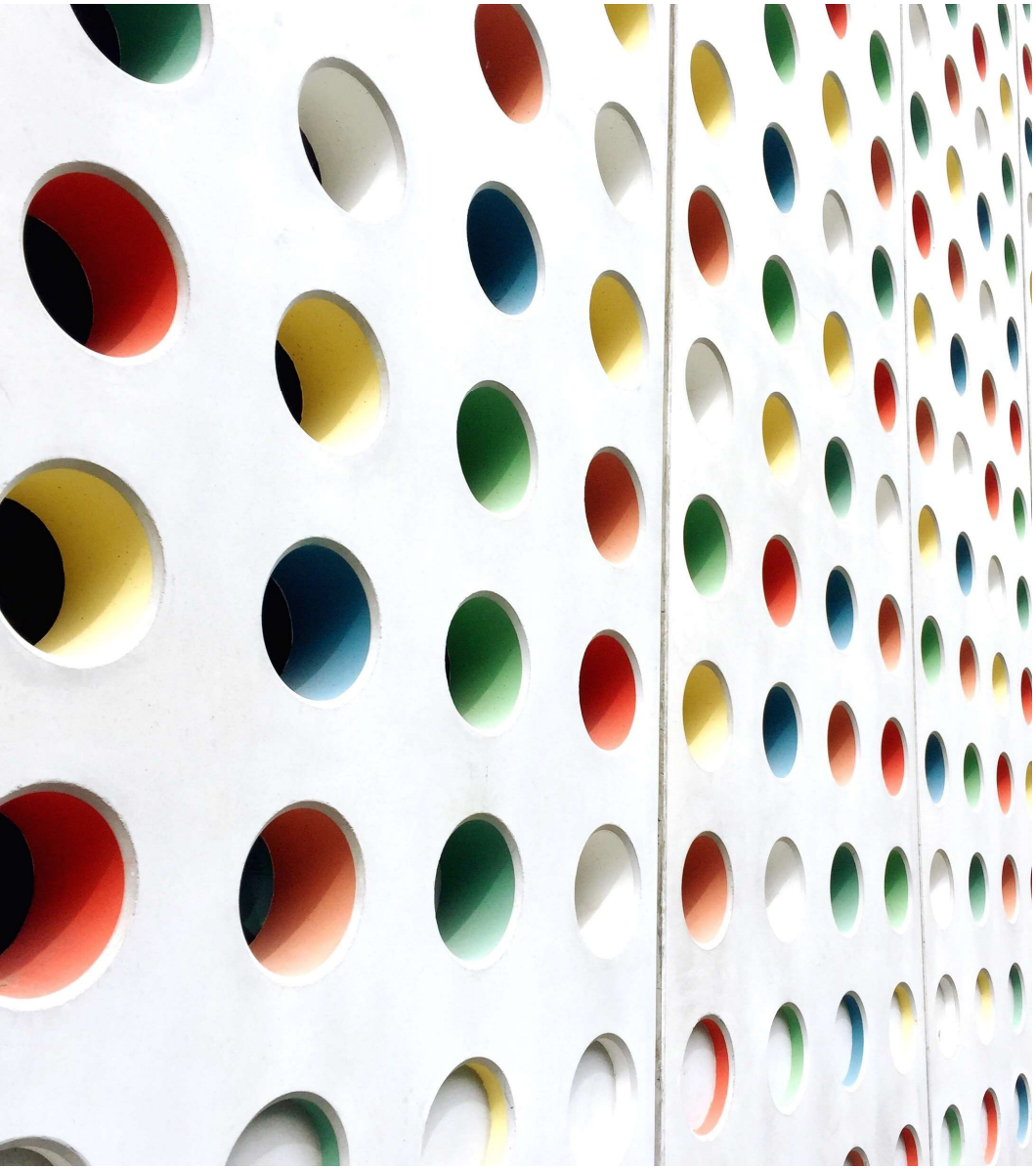


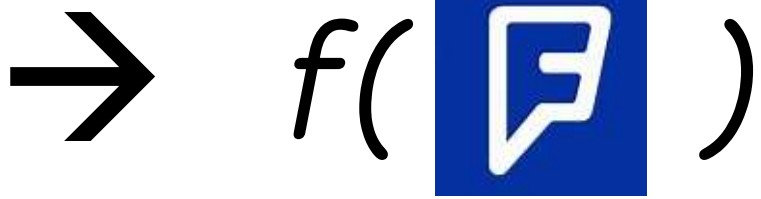
Predicting Crime from Neighborhood Venue Data

DATA SCIENCE CAPSTONE PROJECT

Predicting Crime Patterns Can Improve Public Safety

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

- Washington DC neighborhood labels and geographic position data are available for download in .CSV format from *Open Data DC* (https://opendata.dc.gov)
- 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
- 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
- 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

- Extraneous data in the raw .CSV files downloaded from DC and NYC government websites was removed using Microsoft Excel.
- 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.

Using Data to Address the Problem

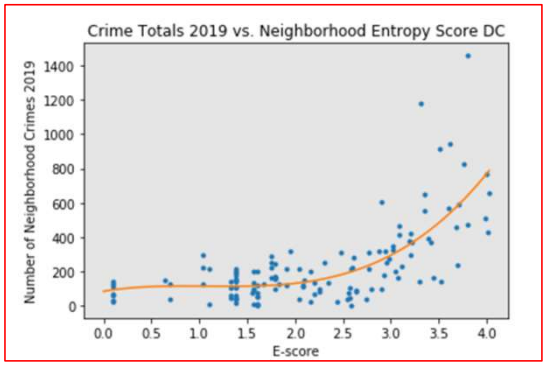
- The DC and NYC neighborhood location data will be used to retrieve venue data from FourSquare.com.
- Feature selection and engineering:
  - Four Square - Neighborhood venue count, venue category counts, and venue category frequencies
  - Derived neighborhood venue entropy score – measures the diversity 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)

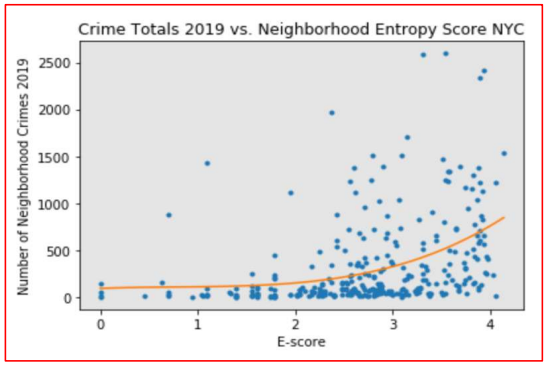
$$Escore = -\sum \frac{N_c}{N} \times \log \frac{N_c}{N}$$

  - $N_c = Number\ of\ venues\ of\ category\ c\ \in C, where\ C\ is\ the\ set\ of\ venue\ cateagories$
  - $N = Total\ number\ of\ venues\ in\ a\ neighborhood$

- Neighborhood venue data was also combined with neighborhood crime data for both cluster and regression analysis.

Exploring the Data

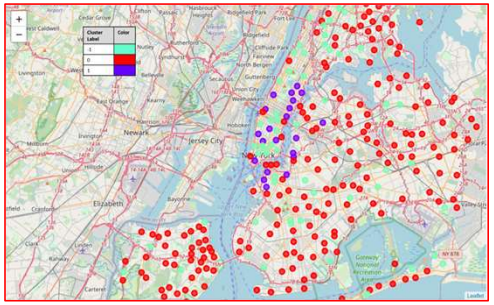


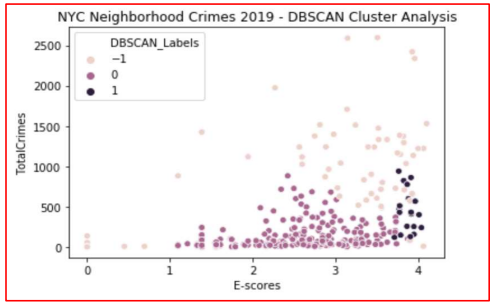


A rough non-linear correlation between crime totals and neighborhood venue entropy score

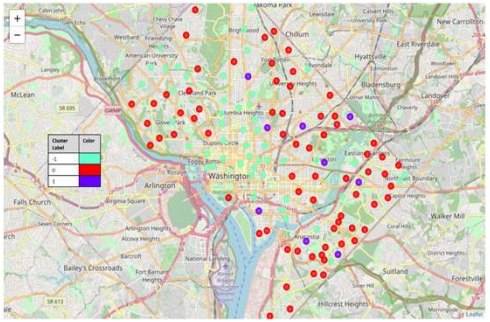
DBSCAN Cluster Analysis

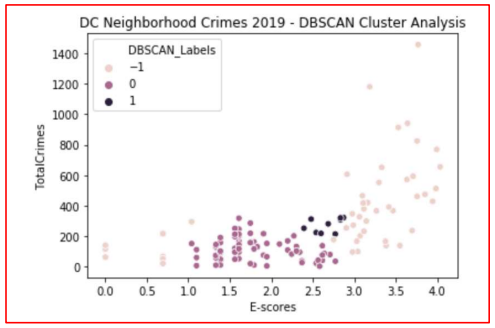
New York City





Washington DC





Outlier neighborhoods (-1 label) less predictable (both high and low crime neighborhoods) based on Total Venues and E-score.

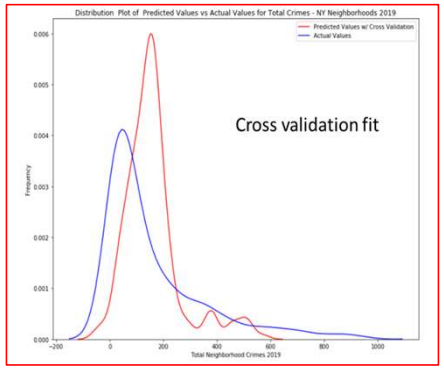
DC has similar pattern to NYC. Outlier neighborhoods are more prevalent in vibrant, inner city locations such area extending from the National Mall to Georgetown and up to Dupont Circle

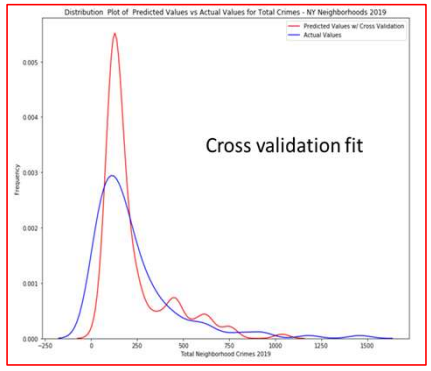
Most **(0 labeled)** neighborhoods have very low average annual crime totals in which crime totals are somewhat independent of E-score.

Few (**1 labeled**) neighborhoods have moderately high crime totals on average with very high E-scores on average with low variability.

Multi-Variate Polynomial Regression

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





R^2 results using model cross-validation for different degrees of the multi-variate polynomial fit did not perform well with the set of features (*Total Venues* and *E-score*).

The distribution plots also indicate that the polynomial regression models do not perform well on the data sets.

- Models do not capture neighborhoods with high crime totals (right tail of the distribution).

Conclusion

More work to do!

Modeling crime activity from neighborhood venue data does offer useful insights, but likely needs to be combined with other neighborhood features to improve predictive capability.

- 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 R2 scores from cross-validation of the regression models at various polynomial degrees.

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.