Comparison of Crime Categories vs Venue Categories in NYC

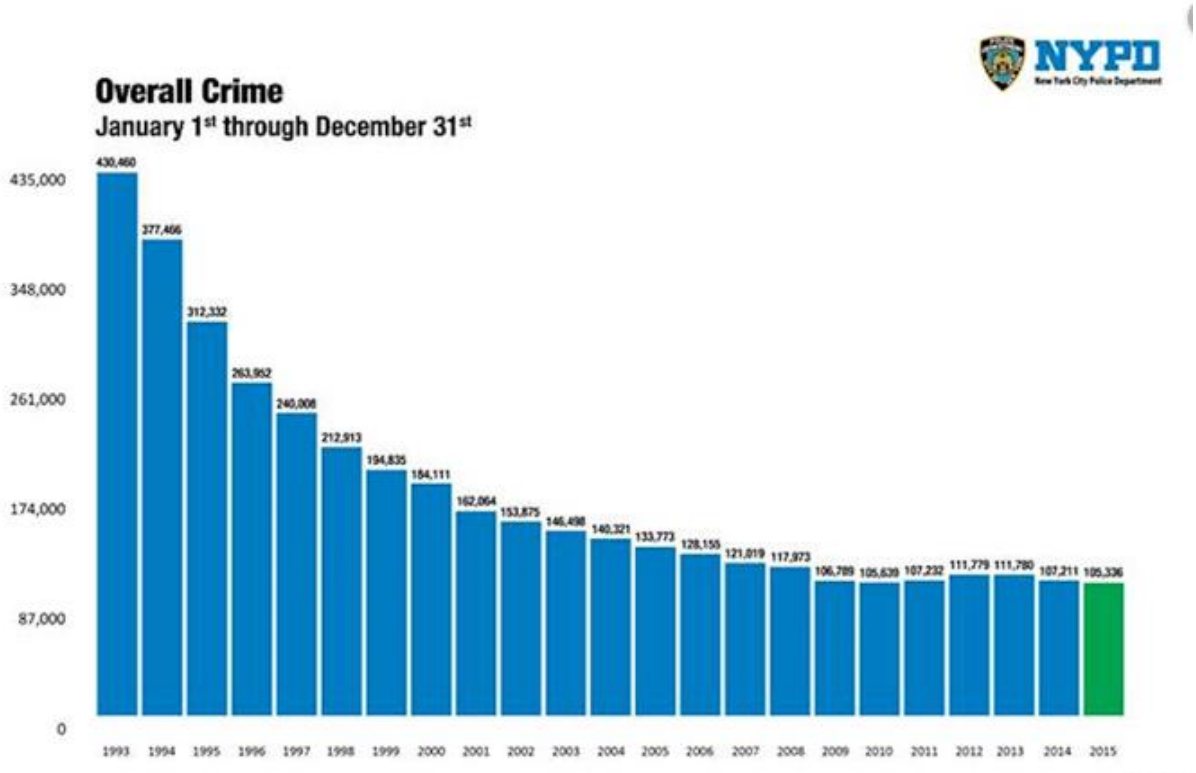
Marc D

August 2019

1. Introduction

1.1. Background

The crime rate in NYC has consistently decreased in the last several decades as shown in the figure below provided by NYPD.



But there are still, as shown above over 100,000 crimes committed every year and the decrease seems to have leveled off in the last few years. Reducing the crime rate stays a priority for the city.

1.2. Problem Description

It would be interesting to understand what are the areas with higher crime rate and study whether the type of venues that are established in those areas are in any way correlated to the type or number of arrests performed in those areas. Having that knowledge could help improve the allocation of security resources and focus safety messages to patrons.

It is understood the study could yield inconclusive results or no result at all. It is still interesting to try to answer the question and demonstrate the tools and technics learned throughout the data science course.

1.3. Report Audience and Stakeholders

If valuable insight is found out of the data through the study, the completed report could be presented to venue owners whose business is located in risk areas.

2. Data

2.1. Arrest Data

The site https://opendata.cityofnewyork.us/ provides useful public data about NYC. In particular it contains data listing all the arrests in NYC since 2013: NYPD Arrests Data (Historic) – Note: you need to create an account (it's free) to access the data.

This is a considerable amount of data that goes back several years, it contains millions of records. The venue data that we have access to through Foursquare (described in the next section) is current data, not historic data, so we will limit our study to the arrests recorded in the first 6 months of 2019. The site provides these data as well in a separate data set: NYPD Arrests Data (Year to Date). Reducing the data set could impact the study but it will help us manage the limited computing resource that are available to us for free (IBM Watson)

The Arrests data set contains a log of all the arrests performed in NYC from Jan-1-2019 to June-30-2019. There are a dozen field included for each record. The field of interests for this exercise are:

- OFNS_DESC which is the type of offense committed
- Latitude and Longitude: The coordinate of the location where the arrest took place. we will use these 2 fields to query Foursquare to get a list of venues around that location.

2.2. Venue Data

We will use the Foursquare API (https://developer.foursquare.com/) to explore the venues around the arrest locations. The Venue Category is the field we are interested in to try to corollate to the number of arrests or the type of offense that triggered the arrest near that venue.

We will set a radius of 200 meter around the arrest locations in the call to the API. We limit the number of locations we will explore (less than 1000) to manage the restrictions we are under related the number of calls we can performed to the Foursquare API.

3. Methodology

After loading the arrest data set, we cleansed the data by performing the following:

- Removed arrest types that were not relevant to the problem;
    - Arrests with generic or cryptic description
    - Arrests related to parking violation

- Dropped records in the data set that have undefined values

- We ended up with close to 100,000 arrests locations after cleansing the data. These locations were clustered geographically into 100 groups.

- The Foursquare API was called on the cluster center for each cluster to get a list of venues and their category around the arrests clusters

- Correlations between arrests types and venue category was established. Correlations greater than 0.7 were retained and a regression plot was created for each of them.

- A Ridge model was built to try to see if the number of arrests around a location could be predicted depending on the types of venues that exists around that location.

4. Results

Pretty weak correlations were found between arrest type and venue type forr the following features:

```
correlation between ALCOHOLIC BEVERAGE CONTROL LAW and Theme Park is 0.7 -
their total counts are respectively: 240 and 12

correlation between ALCOHOLIC BEVERAGE CONTROL LAW and Theme Park Ride / A
ttraction is 0.7 - their total counts are respectively: 240 and 12

correlation between BURGLAR'S TOOLS and Lingerie Store is 0.65 - their tot
al counts are respectively: 360 and 15

correlation between GAMBLING and Tea Room is 0.67 - their total counts are
respectively: 325 and 19

correlation between GRAND LARCENY and Cycle Studio is 0.62 - their total c
ounts are respectively: 5033 and 15

correlation between PETIT LARCENY and Lingerie Store is 0.62 - their total
counts are respectively: 11187 and 15
there are 6 correlations
```

The linear and polynomial model that were built to establish relation between the number of arrests and the type of venues in NYC yielded very poor results as shown below

	Training	Testing
Linear	0.39	0.13
Polynomial	0.81	0.21

5. Discussion

Although correlations were found, they were not very strong. The figure below shows a regression plot for the variables that were found to have a Pearson correlation greater than 0.6. The points are not scattered in a regular way and many of them lie on the x axis.

The linear model performed very poorly. It was trained with different values for alpha (0.001,0.1,1,10,100) but the best R^2 result obtained was 0.13 on the test set as shown below:

```
In [177]: # Ridge Model
          reg = Ridge(alpha=Grid1.best_estimator_.alpha,normalize=Grid1.best_estimator_.normalize)
          reg.fit(X_train, y_train)

Out[177]: Ridge(alpha=10, copy_X=True, fit_intercept=True, max_iter=None,
              normalize=True, random_state=None, solver='auto', tol=0.001)

In [178]: reg.score(X_train,y_train)

Out[178]: 0.393702455076674

In [179]: reg.score(X_validate,y_validate)

Out[179]: 0.2582428019592309

In [180]: reg.score(X_test,y_test)

Out[180]: 0.013365815819114557
```

A polynomial regression was also tried, it performed a little better

```
In [168]: reg.score(X_train,y_train)
Out[168]: 0.8106851444335533

In [169]: reg.score(X_validate,y_validate)
Out[169]: 0.31002248064803606

In [170]: reg.score(X_test,y_test)
Out[170]: 0.21390965612162693
```

.

We used 100 cluster for Kmeans for the geographical classification of the locations. Using a larger number of clusters may have produced better results. 100 was chosen because of the large amount of time it took to run the algorithm.

6 Conclusion

We can not conclude at this point there is a clear correlation between arrests and venue types around the areas that were studied. A deeper study could be conducted with multiyear arrests records. As mentioned in the introduction, we were not able to use a bigger data set because of system power limitations (the kernel kept crashing with a bigger data set). However the results obtained so far show that it is very unlikely there is a clear relation between crime and venue types.