Exploring the Preferences of Robbery Crimes

Xinyu Ma

1. Introduction

1.1 Background

Robbery is one of the most common violent crimes in New York City. Statistics show, from 2000 to 2019, there are an average of about 20,000 robberies reported in NYC each year, ranging from 12,913 in 2018 to 32,562 in 2000. Robbery makes up for at least 13% of all violent crimes reported in NYC. However, the frequency robbery occurs in different areas of the city varies. It would be valuable for both business owners as well as the local police department to know in which areas there is a higher risk of robbery.

1.2 Problem

The risk of robbery in a specific area of a city is dependent on various factors. One of them is the distribution of different categories of business and stores in that area. Based on this assumption, this project aims to find a correlation between the robbery risk and distribution of business categories and to apply it to predict the robbery risks in the future, even for areas where the data of crime history is not available.

2. Data Acquisition and Cleaning

2.1 Data Sources

All reported crimes in 2018 are available through the Open Data NYC program by NYC government agencies. The Raw data can be downloaded from the following url:

https://data.cityofnewyork.us/api/views/8h9b-rp9u/rows.csv?accessType=DOWNLOAD

The numbers of different categories of venues in a specific area is obtained by using the explore endpoint of Foursquare API.

2.2 Data Cleaning and Usage

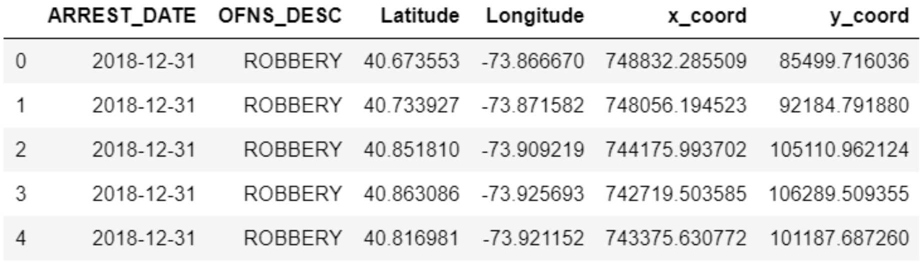
For the crime data, the downloaded raw data contains all sorts of crimes that have been reported and recorded. Given that this project is focused on robbery, the first step is to exclude all the records that are not a robbery crime from the dataset. Secondly, the dataset contains crime records starting from January 1, 2006 to December 31, 2018. Since there could have been a huge change to a neighborhood during the 12 years of timespan, only the records in 2018 were used for the project for a better accuracy. Finally, since the crime data is utilized to locate the areas with high density of robbery crimes, only the latitude and longitude of each record are kept while other features are dropped from the dataset.

Venue information for different neighborhoods can be obtained through Foursquare API. The explore endpoint is used for the purpose of this project. Results from each API call were summarized together into a table showing the number of different categories of venues in different neighborhoods.

3. Methodology and Results

3.1 DBSCAN Clustering of Robbery Records

Before DBSCAN was started, the longitudes and latitudes of each robbery record is projected into x-y plane for future convenience and consistency.



	ARREST_DATE	OFNS_DESC	Latitude	Longitude	x_coord	y_coord
0	2018-12-31	ROBBERY	40.673553	-73.866670	748832.285509	85499.716036
1	2018-12-31	ROBBERY	40.733927	-73.871582	748056.194523	92184.791880
2	2018-12-31	ROBBERY	40.851810	-73.909219	744175.993702	105110.962124
3	2018-12-31	ROBBERY	40.863086	-73.925693	742719.503585	106289.509355
4	2018-12-31	ROBBERY	40.816981	-73.921152	743375.630772	101187.687260

Figure 1. Dataframe ready for DBSCAN clustering

The optimal parameters for DBSCAN were set to be eps = 0.02 and min_samples = 20. After DBSCAN clustering, 6753 records of robbery were clustered, according to their spatial distances, into 82 clusters, while 2271 records were not clustered into any one of them. According to the visualization of those clustered and un-clustered records, the clustered areas showed a much higher density of robbery occurrence while the un-clustered showed a more scattered and much more even distribution over the map. This implies the occurrence of robbery in those high-density areas are more likely to have connections to the surrounding environment of those areas, while other robberies have a higher chance to be individual cases and are less related to the surroundings.

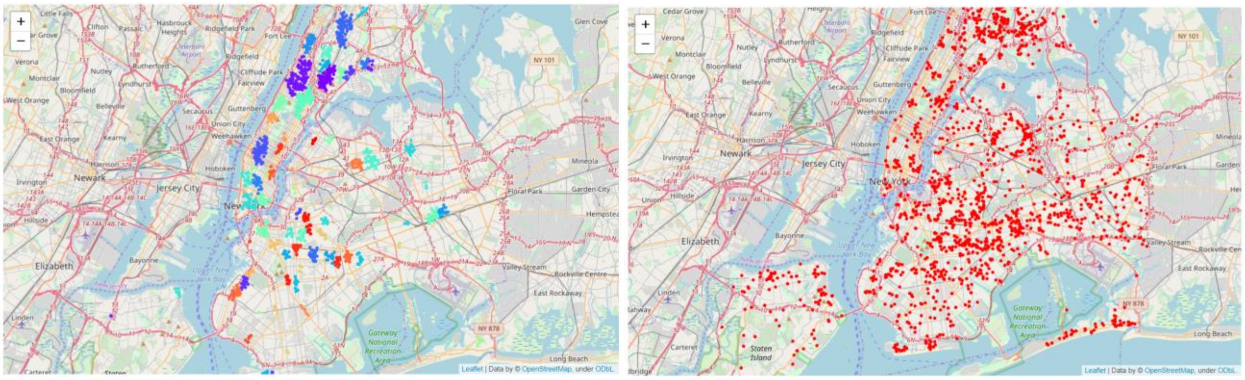


Figure 2. Visualization of clustered (left) and scattered (right) robberies on the map of New York City

3.2 Density and Centroid of Robbery Occurrences

With the robbery records clustered, we need to quantify the density of robberies in each cluster. To begin with, the area of each cluster needs to be calculated, which can be represented by the area of the convex hull of each cluster. However, there are 10 clusters, in which all robberies occurred at only one or two locations. These clusters do not have a convex hull. Hence their areas and density cannot be calculated. These clusters will also be dropped out of the final dataset and finally there are 72 clusters remaining for further investigation.

Besides the density, the centroid of each clusters also needs to be calculated. This could be easily done by taking the mean of the latitudes and longitudes of the robbery records that fall into each cluster. All these data about each cluster, including the latitudes and longitudes of their centroids and their density of robbery occurrences, are summarized and tabulated for future uses.

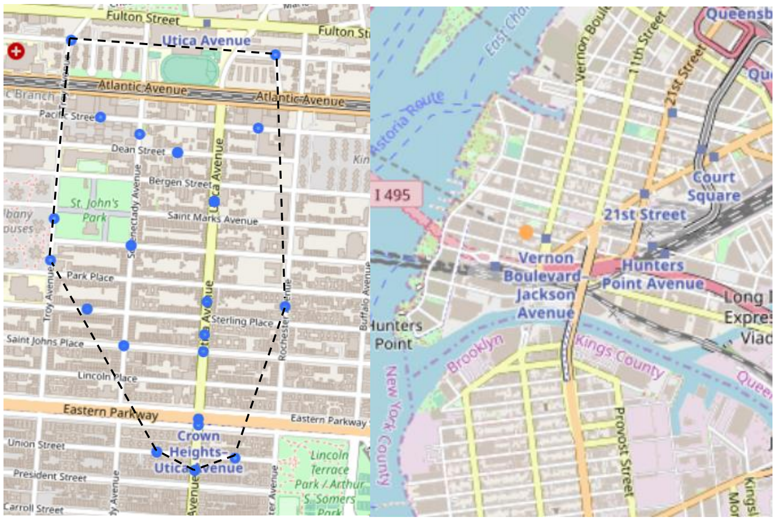


Figure 4. A cluster and its convex hull shown by black dashed line (left) and a cluster which does not have a convex hull due to all robberies occurred at a single location

3.3 Preparing Dataset for Regression and Exploratory Analysis

Using Foursquare API calls, venues information in each cluster were gathered and grouped by the category of venues. The frequency of each category of venues were calculated for each cluster. All these data were tabulated and merged with the robbery density data to give the final dataset for further regression analysis.

However, there are 342 different categories of venues, i.e. 342 features, in the final dataset. It is not much meaningful to directly perform regression analysis on all those 342 features. Thus, an exploratory analysis of the dataset was first performed to find which features have the highest impact on the density of robbery in each cluster by calculating the correlation coefficient of each feature to the density of robberies. The results showed “Drugstore”, “Music Store”, “Steakhouse”, and “Convenience Store” have the most positive correlation coefficients and “Gym / Fitness Center”, “Cocktail Bar”, “Indian Restaurant” have the most negative correlation coefficients.

The regression plots of robbery density against these 8 features were shown respectively in figure 6. It is obvious that none of those features showed a very strong correlation with the robbery density in that area. A major reason of this issue is the lack of datapoints. A potential solution to this issue is to include more data from more cities and more periods of time. However, due to the scope and time limit of this project, it is not included in this report but would be interesting and meaningful for further investigated.

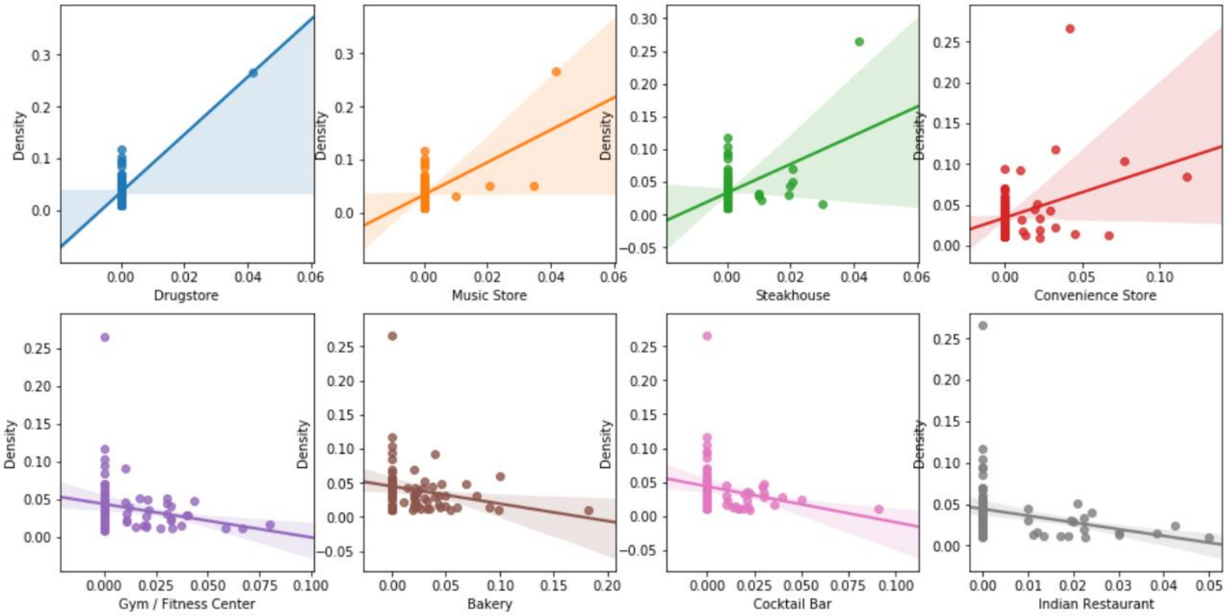


Figure 6. Regression plot of robbery density against the 4 most positively correlated features and the 4 most negatively correlated features

3.4 Regression Analysis of Prediction

With the above-mentioned explorative analysis, the 4 most positively correlated features and the 4 most negatively correlated features were selected to perform the final regression analysis. The dataset was split into a training set which contains 54 samples and a test set which contains 18 samples. Both linear regression and k-Nearest Neighbor regression algorithms were used and compared.

3.4.1 Linear Regression

Linear regression model was fitted on the training set and applied to the prediction of test set. The RMSE (root of mean squared error) was 0.0197 and the R^2 score is 0.336, which showed there are some extent of correlation between these features and robbery density, but the model is not very accurate in doing predictions. Nevertheless, the model still showed some consistency between the predicted values and true values.

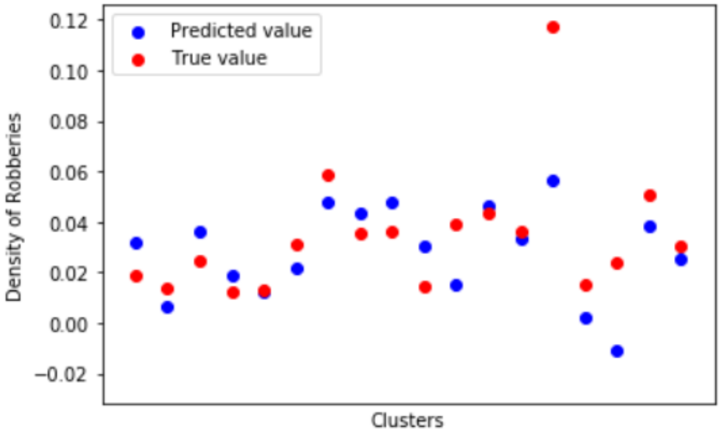


Figure 7. Predicted density by linear regression model and true density

3.4.2 k-Nearest Neighbors Regression

k-Nearest Neighbors regression was also used on the same dataset as a comparison. First, the optimal number of neighbors k was determined by observing the RMSE and R^2 score changing from k = 1 to k = 10. At k = 5, the RMSE reached the lowest at 0.0189 and the R^2 reached the highest at 0.388. Compared to linear regression model, k-Nearest Neighbors regression model gave slightly better results. Although the overall accuracy is still not quite satisfactory, the model still showed some extent of consistency between the predicted values and the true values.

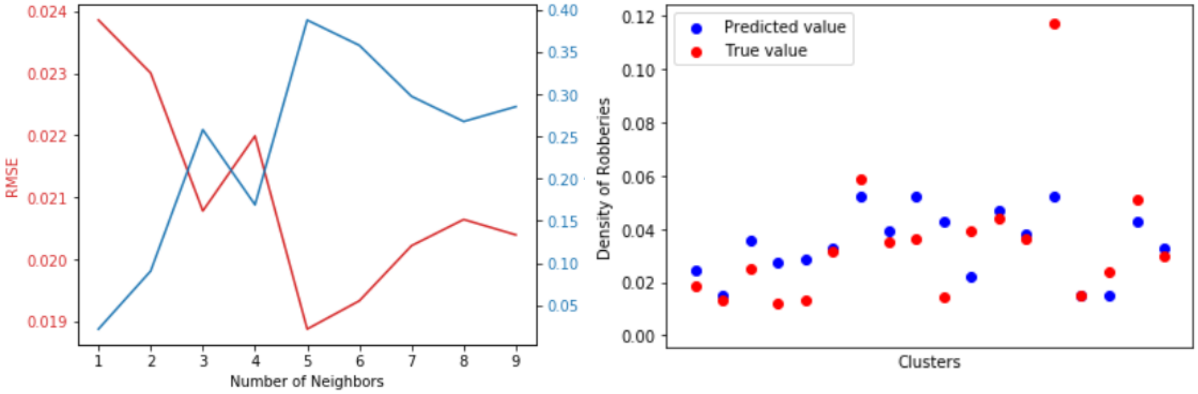


Figure 8. Looking for the optimal number of neighbors for k-Nearest Neighbors model (left) and the predicted density by k-Nearest Neighbors model (k = 5) and true density (right)

4. Discussion

From the exploratory analysis and the linear regression model, it is shown that the frequency of drugstores, music stores, steak houses and convenience stores are the 4 most positively correlated venues to the density of robberies in that area, while gyms / fitness centers, bakery, cocktail bars and Indian restaurants are the 4 most negatively correlated venues to the density of robberies. It is recommended that neighborhoods with higher frequency of the former 4 kinds of venues or lower frequency of the latter 4 kinds of venues should be more careful about potential robberies.

Corresponding predictive models were also established to predict the potential robbery risk of an area given the distribution of different kinds of venues in that area. The predicted values could be useful reference for local police department as well as local shop owners.

5. Conclusion

Through DBSCAN clustering of the spatial distribution of robberies in New York City, 6753 robbery records in 2018 were clustered into 82 clusters, with 72 clusters able to give convex hull. The distribution of different kinds of venues in these 72 clusters were further investigated. Two predictive regression models were established to predict the potential robbery risk of a given neighborhood.

However, both models greatly suffer from the lack of datasets, resulting in the unsatisfactory accuracy of predictions. Still, both models showed quite some consistency between the predicted values and the true values of the density of robberies, which demonstrated the value of this methodology. An improvement to the accuracy could be anticipated given more data fed to the models.