Quantitative Methods

Is there a temporal pattern for crime?
A temporal analysis of assaults in New York City

1. Introduction

Much of the existing research in criminology can be characterized by one of two approaches. The larger share of research focuses on the characteristics of the offender, such as psychological or social factors. The other part of criminology looks at a criminal offense as an event. The most fundamental literature using the latter approach was written by M. Felson and L. E. Cohen, who also coined the term routine activity theory.[1] Routine activity theory focuses on the circumstances, more specifically on the relation of a criminal offense to space and time.

This coursework aims to investigate patterns over time in order to find out if time could serve as a predictor for assault in the third degree in New York City. Assault in the third degree is the weakest criminal offense in the family of assaults. By definition, it involves causing physical injury to another person or the intent or recklessness to do so.[2] 11% of all reported offenses were assaults in the third degree, making it the third most common criminal offense in New York City in 2016 (Fig. 1).

Part one provides an overview over the used dataset, part two focuses on the methods to explore daily and monthly temporal patterns and part three presents the results. Part four explores approaches for further improvement and discusses possible explanations.

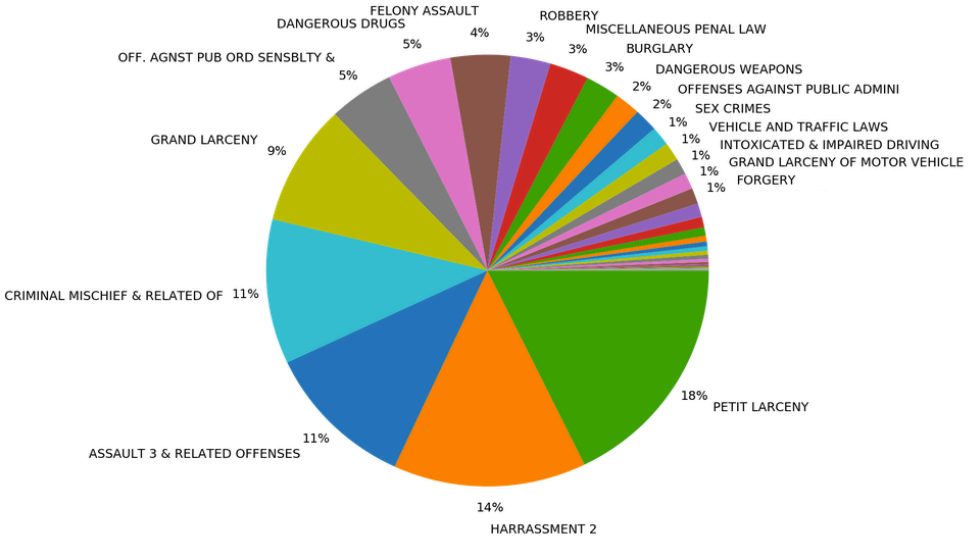


Fig 1: Crime types by occurence (darkblue: assault)

2. Data

The entire data set is taken from the NYC open data platform and is accessed via the Soda API. The data set is maintained by the Mayor's Office of Data Analytics (MODA) and the Department of Information Technology and Telecommunication

[1] Cohen L. E, Felson M., 1979. Social Change and Crime Rate Trends: A Routine Activity Approach. American Sociological Review,Vol. 44, No. 4, pp. 588-608

[2] New York Consolidated Laws, Penal Law - PEN § 120.00. http://codes.findlaw.com/ny/penal-law/pen-sect-120-00.html (accessed 1.9.18).

(DoITT) in New York City.[3] The full data set contains 13 columns and 574'040 rows of information on all reported crimes in New York City in 2016. For the purpose of this paper, only the reported assaults in the third degree are selected, leaving us with 52'182 rows. The selected variable is the reported start time of an assault recorded as day, hour and minute intervals.

3. Methodology

When looking at patterns, time series literature differentiates between trend and seasonality patterns. A trend can be loosely described as a long-term change in the mean or tendency of the data. Seasonality occurs when a series is influenced by seasonal factors. Seasonality can also apply to shorter intervals than yearly seasons.[4] In this regard, the following two hypotheses about the temporal dependency of assault in the third degree are made:

Hypothesis 1: There are daily seasonal patterns for assault rates over a month.
Hypothesis 2: There are hourly seasonal patterns for assault rates over a day.

When looking at the data, it is evident that the reported assaults also follow a non-linear trend reaching its inflection point in summer time with smaller fluctuations around the mean throughout the year (Fig. 2). However, this coursework will only investigate patterns on the level of months and days as it difficult to make assumptions of the yearly behavior with data limited to a year.

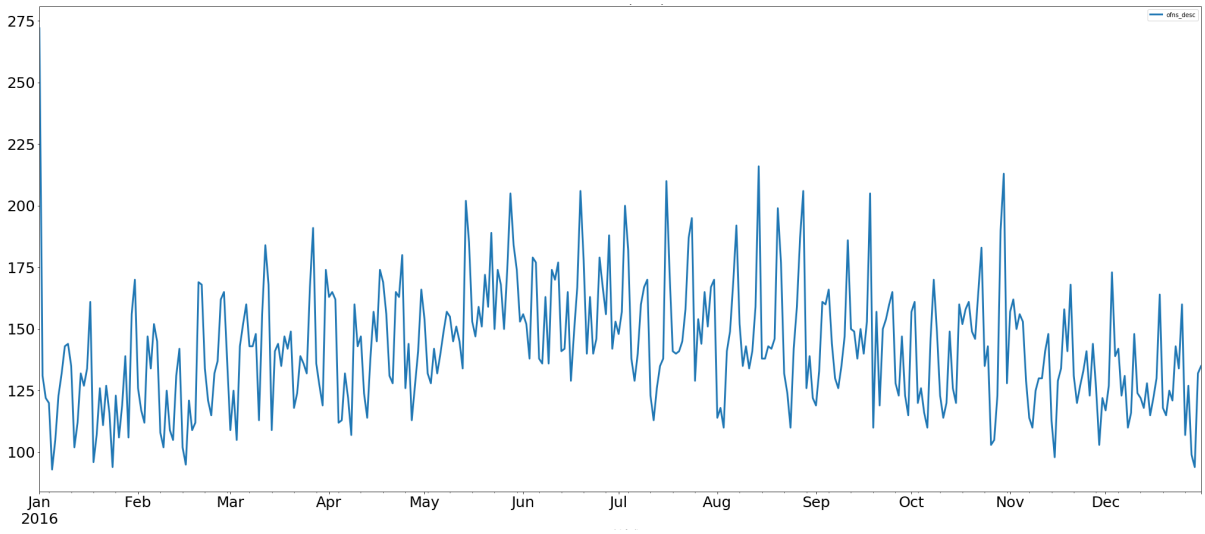


Fig 2: Total assault rates over the year 2016

Autocorrelation

Due to the temporal aspect errors are not considered independent in time series.[5] Thus, the many statistical methods that require independent errors are not suitable for time series.[6] One way to examine a temporal aspect is by looking at the autocorrelation between the number of occurrences at a certain time, Y_t, and the

[3] https://opendata.cityofnewyork.us/
[4] Chatfield, C., 2016. The Analysis of Time Series: An Introduction, Sixth Edition. CRC Press.
[5] Pennsylvania State University, Stat 501: Lesson 14: Time Series and Autocorrelation. https://onlinecourses.science.psu.edu/statprogram/node/138
[6] Fox, J., Weisberg, S., 2011. An R Companion to Applied Regression. SAGE.

number of occurrences at the lagged instances of Y_{t-j}. The j-th autocorrelation of a series Y_t and the series lagged values Y_{t-j} is calculated by

$$r = \frac{\sum_{t=1}^{T}(Y_t - \bar{Y})(Y_{t-j} - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

where $\bar{Y}$ is the sample average. The assumption in autocorrelation is that the observations are equally spaced in time.[7]

Curve fitting with non-linear least squares regression

The initial visualizations (Fig. 3,4) suggest that the behavior of assault rates over days and months follow an oscillating curve. This non-linear behavior could therefore be modeled after a sinus curve. The sinusoidal function is given by:

$$\bar{y} = a * \sin(bt + c) + d$$

Where $\bar{y}$ is the estimated value, a the amplitude, b the period, t the time, c the phase, and d the intercept, i.e. the population mean. [8]

A first, a rough estimate for all parameters is made by simply looking at the data and the visualized autocorrelation. In an iterative process, non-linear least squares regression optimizes the first estimate by minimizing the calculated residuals, i.e. the differences between the observed data points and the first estimate. [9]

Thus, prior to performing a least squares regression, the residuals of the estimated curve and the observed values have to be calculated by subtracting the observed values y: [10]

$$r = (a * \sin(bx + c) + d) - y$$

In a last step, the fitted curve is tested on some observed values.

Hypothesis 1) daily seasonal patterns

In order to test hypothesis 1, autocorrelation is used to identify a correlation between days and its magnitude and frequency. For that purpose, the time interval is defined as a day and the assaults committed on each day are counted and grouped by day. The number of chosen lags is 28 days, i.e. a month. Based on the observations from the autocorrelation, the curve will be modeled by using the assault rates of an averaged day for the timespan of a month as the dependent variable.

Hypothesis 2) hourly seasonal patterns

The same method is applied to investigate hourly patterns. Thus, the assaults are counted and grouped by full hours. The number of chosen lags is 24 hours, i.e. a

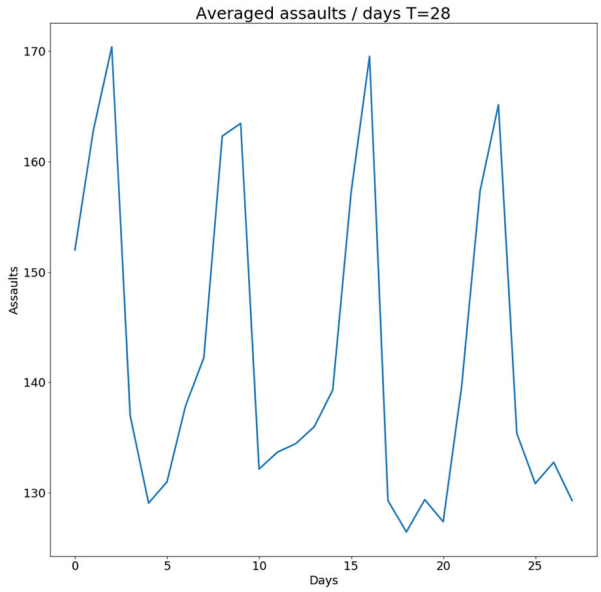
[7] 1.3.5.12. Autocorrelation [http://www.itl.nist.gov/div898/handbook/eda/section3/eda35c.htm (accessed 1.5.18).

[8] Hudock, M. Section 6.6 - Phase Shift; Sinusoidal Curve Fitting. http://www.matthewhudock.com/Math2412Links.htm (accessed 1.5.18).

[9] Brown, A. M., 2000. A step-by-step guide to non-linear regression analysis of experimental data using a Microsoft Excel spreadsheet. Computer Methods and Programs in Biomedicine 65,191–200

[10] Spector, P., 2010. Calculating the Jacobean and Residuals of a Nonlinear Regression Model. Statistics 243, UC Berkeley.

day. Again, the autocorrelation will serve as a base to model the time series over an averaged day from hour 0 to hour 24.



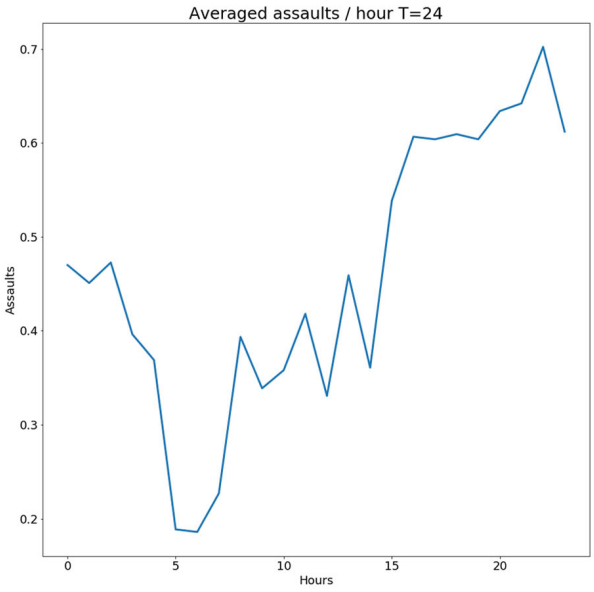


Fig 3: Averaged assault rates over a month (0 = Friday). Fig 4: Averaged assault rates over a day

5. Results

Hypothesis 1) daily seasonal patterns

Figure 3 displays a tendency for assault rates to steadily peak towards the weekend and reaching a low point around Tuesday. The autocorrelation plot (Fig. 5) indicates that the time interval after lag 1 at which correlation above 0.5 occurs is approximately 7 days. Therefore, the first estimate for the frequency of the sinusoidal function is:

$$b = \frac{2\pi}{7}$$

An estimate for the amplitude a, the phase c and the population mean d can be derived from the original data (Fig. 3). All estimated parameters are summarized in the table below (Parameters 1). By performing a least squares regression, the optimized model is obtained (Parameters 2). Both the estimate and the optimized curve underfit around the peaks in the data (Fig. 7).

Parameters 1: Estimate		***Parameters 2: Fitted***	
a_1	(max(y)-min(y))/2	a_2	17.430
b_1	2π / 7	b_2	0.889
c_1	0	c_2	0.461
d_1	mean(y)	d_2	142.609
RMS	49.617	RMS	38.214

The optimized curve is now overlaid with a two selected summer and winter months. Although the fitted curve captures the oscillation to some extent, it does not perform well as a predictor (Fig. 9).

Hypothesis 2) hourly seasonal patterns

The number of assaults per hour reaches a low point around 06:00 and then steadily increases towards nighttime with a peak around 22:00 (Fig. 4). Equally to the previous model, first estimates about all parameters are made. The autocorrelation plot (Fig. 6) indicates that the time interval of one oscillation is roughly 24 hours, therefore the frequency is:

$$b = \frac{2\pi}{24}$$

Again, the starting values for the amplitude a, the phase c and the population mean d of the sample y is estimated (Fig. 1). The resulting parameters are summed up in the table below (Parameters 1). Plotting the first result suggests that the estimated curve underfits the data (Fig. 8, in red). By performing a least squares regression a new optimized model is obtained (Parameters 2). The root mean square error (RMS) is significantly lower for the optimized curve compared to the estimate (Fig. 8, in blue).

Parameters 1: Estimate		***Parameters 2: Fitted***	
a_1	(max(y) - min(y))/2	a_2	-0.185
b_1	2π / 24	b_2	0.233
c_1	0	c_2	-0.100
d_1	mean(y)	d_2	0.469
RMS	0.606	RMS	0.282

Based on the results of the model for the daily seasonal patterns, it can be assumed that the hourly model is equally unable to capture the actual behavior of assault rates in 2016.

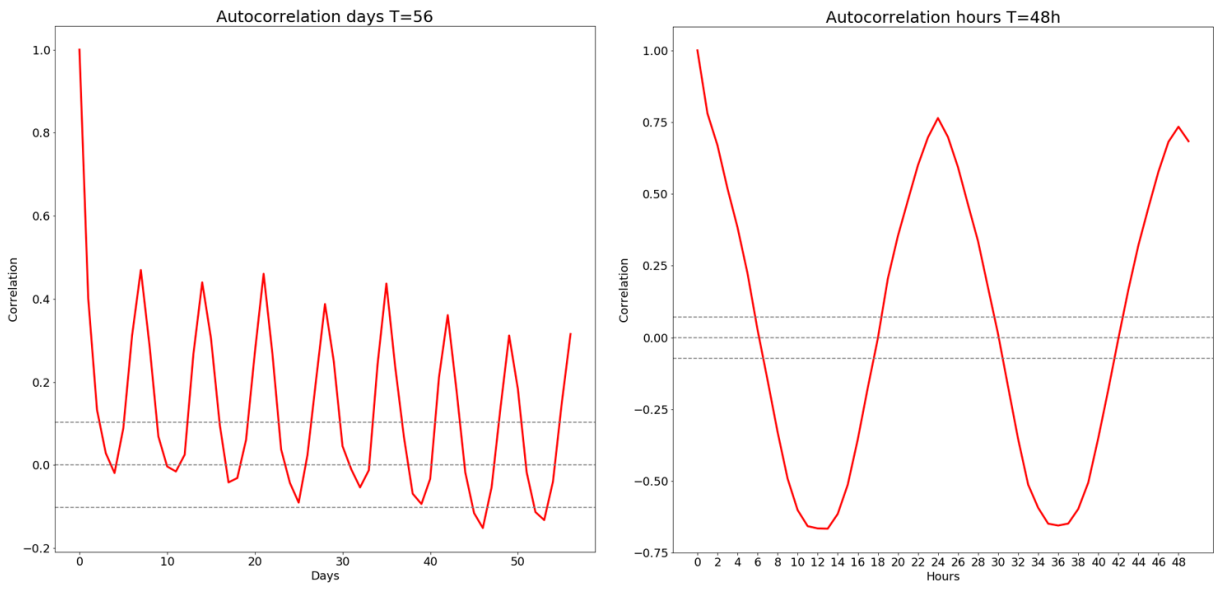
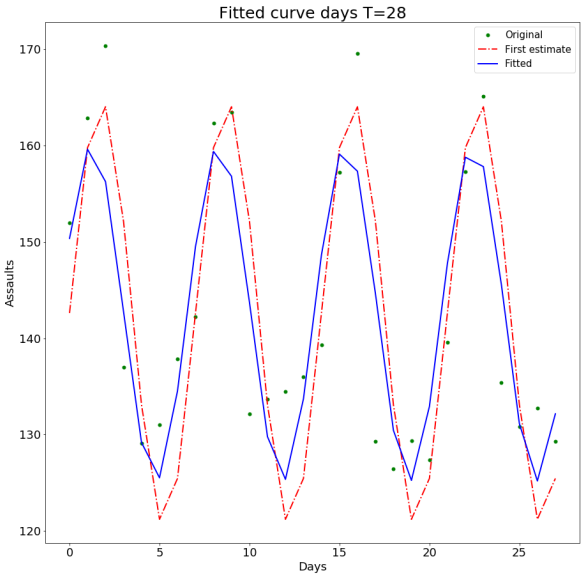


Fig 5: Autocorrelation days (over 56 days). Fig 6: Autocorrelation hours (over 48 hours)



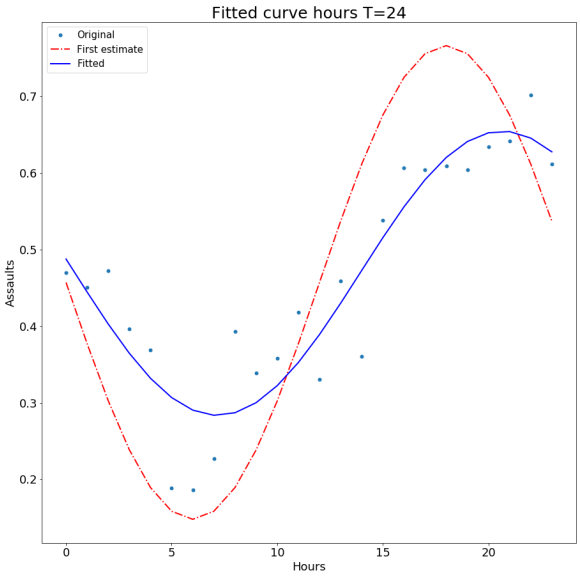


Fig 7: Estimated curve (red), fitted curve (blue) for a month. Fig 8: Estimated curve (red), fitted curve (blue) for a day.

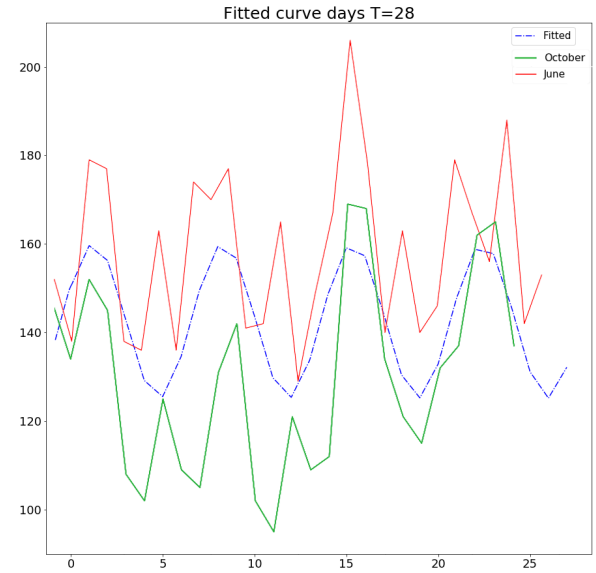


Figure 9: Fitted curve for a month (blue), June assaults (red), October assaults (green)

Limitations to the model

Attempting to model a time series on averages over a short period results in a very a rough approximation. There are many underlying influences shaping a time series that have to be taken into account. According to a study by the Pew Center of Research, violent crime in the United States has been steadily falling since 1990.[11] A time series with the length of only a year is unable to captivate the declining macro trend.

Furthermore, the assaults reported do not represent an accurate image of the assault rates in New York. The same study has found that as many as 47% of the tracked crimes are unreported.

Although a time might give an indication about the behavior of assault rates, it cannot explain a causal relationship. According to Box and Tiao (1975) time series modelling can tell us only if there „is evidence that change in the series of the kind expected

[11] Gramlich, J., 2017. 5 facts about crime in the U.S. Pew Research Center.

actually occurred, and, if so, what can be said of the nature and magnitude of the change".[12]

Conclusion

The analysis done within the framework of this coursework confirms that the rates of assaults to the third degree in New York City follow a temporal pattern, both on for a daily period with peaks in the evening hours and a monthly period, with peaks on weekends. One explanation for the increased assault rates in the evening and the weekend might lie within the spatial characteristics of the offenses. According to the data, every second assault is committed in a residence and every third assault on the street.[13] It can be said that, despite a temporal relationship, using time as a predictor proves difficult due to a multitude of influences.

References

Cohen L. E, Felson M., 1979. Social Change and Crime Rate Trends: A Routine Activity Approach. American Sociological Review,Vol. 44, No. 4, pp. 588-608

New York Consolidated Laws, Penal Law - PEN § 120.00. http://codes.findlaw.com/ny/penal-law/pen-sect-120-00.html (accessed 1.9.18).

Chatfield, C., 2016. The Analysis of Time Series: An Introduction, Sixth Edition. CRC Press.

Pennsylvania State University, Stat 501: Lesson 14: Time Series and Autocorrelation. https://onlinecourses.science.psu.edu/statprogram/node/138

Fox, J., Weisberg, S., 2011. An R Companion to Applied Regression. SAGE.

1.3.5.12. Autocorrelation [http://www.itl.nist.gov/div898/handbook/eda/section3/eda35c.htm (accessed 1.5.18).

Hudock, M. Section 6.6 - Phase Shift; Sinusoidal Curve Fitting. http://www.matthewhudock.com/Math2412Links.htm (accessed 1.5.18).

Brown, A. M., 2000. A step-by-step guide to non-linear regression analysis of experimental data using a Microsoft Excel spreadsheet. Computer Methods and Programs in Biomedicine 65,191–200

Spector, P., 2010. Calculating the Jacobean and Residuals of a Nonlinear Regression Model. Statistics 243, UC Berkeley.

Gramlich, J., 2017. 5 facts about crime in the U.S. Pew Research Center.

Vujić, S., Commandeur, J.J.F., Koopman, S.J., 2016. Intervention time series analysis of crime rates: The case of sentence reform in Virginia. Economic Modelling 57, 311–323. https://doi.org/10.1016/j.econmod.2016.02.017

Andresen, M.A., Malleson, N., 2015. Intra-week spatial-temporal patterns of crime. Crime Sci 4, 12. https://doi.org/10.1186/s40163-015-0024-7

The dataset used in this coursework can be accessed under: https://opendata.cityofnewyork.us/

[12] Vujić, S., Commandeur, J.J.F., Koopman, S.J., 2016. Intervention time series analysis of crime rates: The case of sentence reform in Virginia. Economic Modelling 57, 311–323. https://doi.org/10.1016/j.econmod.2016.02.017

[13] This statistic is taken directly from the data set (see Appendix, Fig. 1)

Appendix

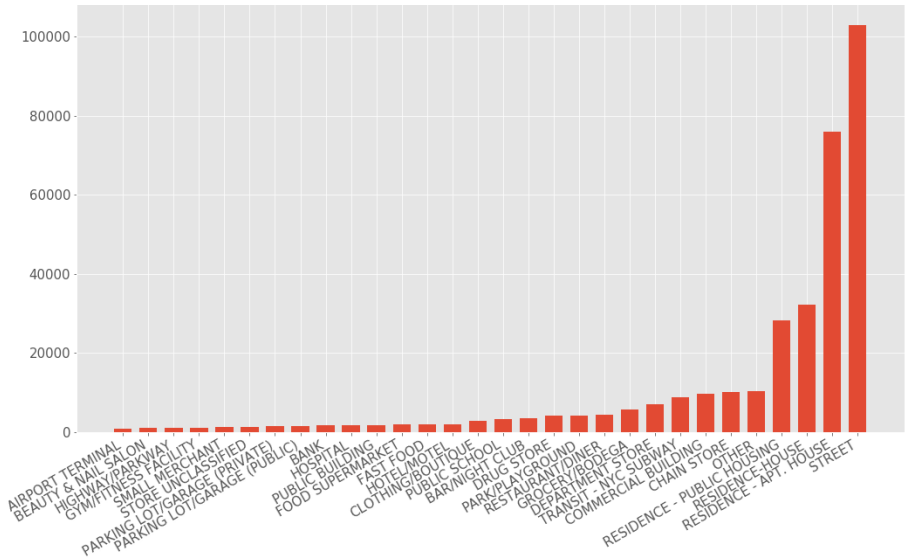


Figure 1: Assault locations