NYC Arrests - Report

Reporter: Cat Chenal

Email: catchenal@gmail.com

[GitHub (https://github.com/CatChenal/NYCData)](https://github.com/CatChenal/NYCData)

Summary:

Even though the number of arrests has steadily dropped in the 2015-2018 period (~ 30% per year), there is a remarkable constance in key factors:

a. Type of arrests: only 8 types are common to all years' Top 5 ranking, while the dataset reports over 350 different violations of the NYS penal code.
b. Criminality has its days! The lowest levels occur on Sundays, the highest on Wednesday.

Dataset quality: ***NYPD Arrests Data (Historic)***

Missing data were imputed if possible. The clean dataset is reduced by 220 uncategorized arrests to 4,798,119, but There are still 8,650 records with an empty 'pd_desc' field. The top 3 Precincts with uncategorized arrests are: P14 (53), P103 (10) and P73 (6), amounting to 69 out of 220, or over 31%.

Has the arrest rate been decreasing from 2015-2018?

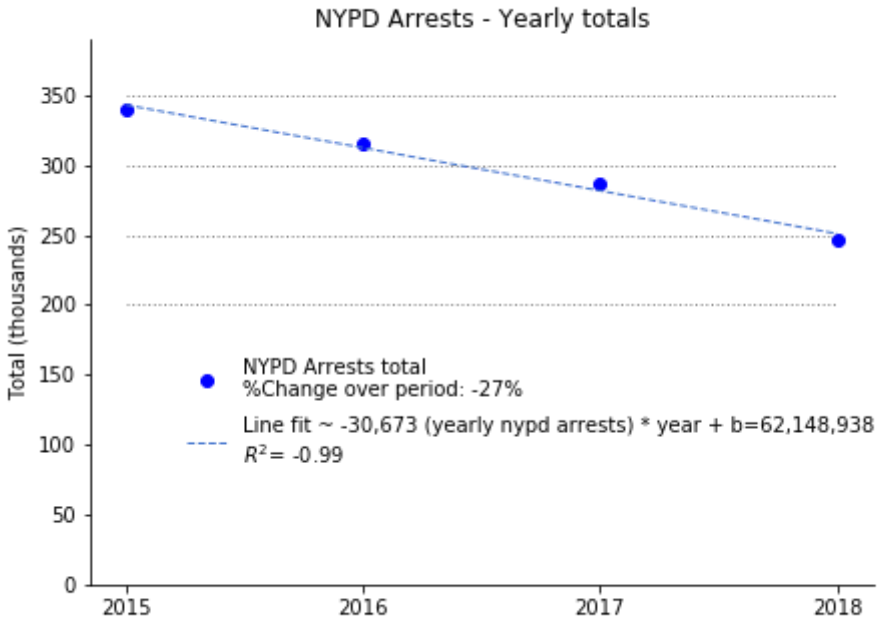

Figure 1 - ***NYPD Arrests Data (Historic)*** *(8h9b-rp9u (https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u/))*: Yearly total number of arrests (2015 - 2018).

Yes: The percentage decrease in the number of NYPD arrests over the last four years (2015 - 2018) is 27%.Additionally, the yearly totals have decreased at a steady pace over that period. As depicted in Figure 1, the linear change in the yearly totals amounts to a decrease of over 30,000 arrests per year.
(*Statistical note: the high coefficient of determination (R^2) ascertains the data is properly modelled by a line.*)

What are the top 5 most frequent arrests as described in the column 'pd_desc' in 2018?

These are the Top 5 arrests in 2018:

2018	count
pd_desc	
ASSAULT 3	26611
LARCENY,PETIT FROM OPEN AREAS,UNCLASSIFIED	23405
TRAFFIC,UNCLASSIFIED MISDEMEAN	14856
ASSAULT 2,1,UNCLASSIFIED	11763
CONTROLLED SUBSTANCE, POSSESSION 7	9982

When traced over the 2015-2018 period, most display a remarkable stability (measured as the percentage change ('%change') between the first and last year):

	2015	2016	2017	2018	%change
ASSAULT 3	27631.0	26961.0	26281.0	26611.0	-0.04
LARCENY,PETIT FROM OPEN AREAS,UNCLASSIFIED	25772.0	23871.0	23020.0	23405.0	-0.09
TRAFFIC,UNCLASSIFIED MISDEMEAN	14435.0	15231.0	16298.0	14856.0	0.03
ASSAULT 2,1,UNCLASSIFIED	12048.0	12127.0	11911.0	11763.0	-0.02
CONTROLLED SUBSTANCE, POSSESSION 7	17420.0	14177.0	12341.0	9982.0	-0.43

In fact, there are only 2 types of arrest that changed by over 5%, with 'CONTROLLED SUBSTANCE, POSSESSION 7' having the largest drop (-43%). This persistence is clearly depicted in Figure 2:

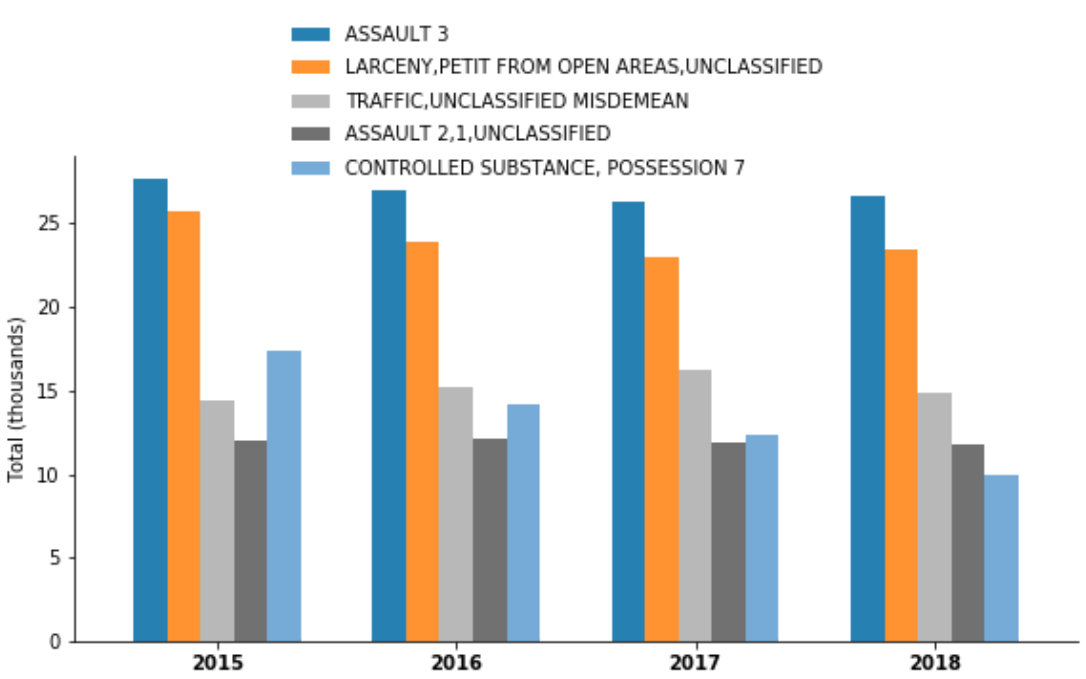

Figure 2 - ***NYPD Arrests Data (Historic)*** *(8h9b-rp9u (https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u/))*: Evolution of the 2018 Top 5 arrests over the period (2015 - 2018).

The persistence of a few number of arrest types with high count is also occurring when we consider the Top 5 arrests
in each year: there are 350 (properly categorized) arrests in the 2015-2018 period, but only 8 that appear among the Top 5 yearly rankings:

Arrests description in any yearly Top 5 ranking	2015	2016	2017	2018
ASSAULT 3	2nd: 27,631	2nd: 26,961	**1st: 26,281**	**1st: 26,611**
LARCENY,PETIT FROM OPEN AREAS,UNCLASSIFIED	3rd: 25,772	3rd: 23,871	2nd: 23,020	2nd: 23,405
TRAFFIC,UNCLASSIFIED MISDEMEAN		5th: 15,231	5th: 16,298	3rd: 14,856
ASSAULT 2,1,UNCLASSIFIED				4th: 11,763
CONTROLLED SUBSTANCE, POSSESSION 7	4th: 17,420			5th: 9,982
THEFT OF SERVICES, UNCLASSIFIED	**1st: 29,833**	**1st: 27,234**	3rd: 19,825	
MARIJUANA, POSSESSION 4 & 5		4th: 18,117	4th: 17,964	
NY STATE LAWS,UNCLASSIFIED VIOLATION	5th: 16,899			

Total number of arrests per day and month in 2018

Criminality has its days! The lowest level occurs on Sundays, the highest on Wednesday (true for all years):

(Monthly min: green, max: orange.)

Month	1	2	3	4	5	6	7	8	9	10	11	12
Day												
Sun	2203	2043	2263	2554	2061	1826	2619	2206	2293	1858	1859	2103
Mon	3273	2687	2593	2875	2300	2275	2976	2319	2115	2753	2131	2280
Tue	4381	3530	3467	3201	3746	2896	3642	2909	2608	3555	2634	2356
Wed	4764	3668	2795	3484	4293	3169	3026	4066	3017	3751	2872	2687
Thu	3261	3913	4268	3479	4031	3079	3144	3834	2884	2924	3260	2652
Fri	3211	3137	3727	3051	3193	3382	2899	3619	2732	2642	3144	2483
Sat	2789	2746	3349	2683	2412	2985	2569	2427	2859	2024	2096	2834

If we think of arrests as a sample of total crime, is there more crime in precinct 19 (Upper East Side) than precinct 73 (Brownsville)?

Precinct 19 accounts for less than 1% of all crimes and of those in the lastfour years (0.85% and 0.92%, respectively) while the percentages for Precinct 79 are 2.79% and 2.40%, i.e. over twice as much. As the population of Community District 8, where Precinct 19 resides, is slightly over 2.5 larger than that of CD 16, where Precinct 73 is located, this is not a population effect: there is more crime in Precinct 73 than in Precinct 19.

Given the available data, what model would you build to predict crime to better allocate NYPD resources?

Predicitve crime modelling is an active area of research due to its data, design and ethical challenges. These spatio-temporal models have stringent requirements for their effectiveness: For the purpose of allocating patrol officers to needed areas in a timely fashion, the model would need to predict the locations (within a reasonable distance variability) *together with* a long enough "time-ahead" window. For instance a model capable of predicting a crime hot-spot at the block level would be useless if it could not do so *at least* 48 hours prior the predicted activity. As the _NYPD Arrests Data_ dataset does not have time data, other sources, such as 911 call logs, are needed.

Part of the data challenge is to decide which input data to use: while weather and "crowd signals" from social media are probably useful, the use of socio-economic data may introduce bias. The more varied the sources, the wider the skills needed to process and anlyze the data as they take many forms: text, sound, video, etc. Part of the modeling challenge is to identify which features are necessary to achieve a given goal. In this context, the minimal type of information is location, date and time and the age-group distibutions (at the census tract level if possible) since the 24-44 age group is the most represented in the dataset.

Whether the model is parametric (i.e. regression) or neural (machine learning), the first step is to establish a baseline against which other models (issued from different feature selection, for example) will be rated.