CHICAGO CRIME

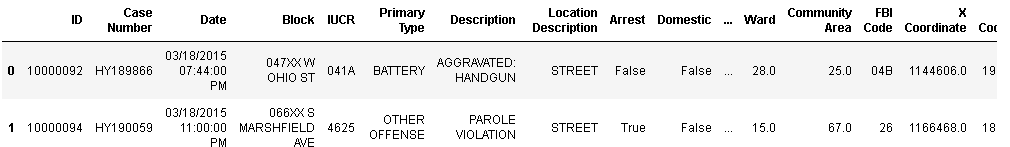
# Introduction

Our main motive is to predict the crime in 2018 and explore the data in such a way that it can be used by CPD. The data for this analysis is fetched from the website <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>. It contains historical data also from year 2001 till 2018. This dataset reflects reported incidents of crime. There are total 22 fields in the data.

Fields in the data –

| **Column Name** | | **Description** | | **Type** |
| --- | --- | --- | --- | --- |
| **ID** | Unique identifier for the record. | | Number | |  |
| **Case Number** | The Chicago Police Department RD Number (Records Division Number), which is unique to the incident. | | Plain Text | |  |
| **Date** | Date when the incident occurred. this is sometimes a best estimate. | | Date & Time | |  |
| **Block** | The partially redacted address where the incident occurred, placing it on the same block as the actual address. | | Plain Text | |  |
| **IUCR** | The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at <https://data.cityofchicago.org/d/c7ck-438e>. | | Plain Text | |  |
| **Primary Type** | The primary description of the IUCR code. | | Plain Text | |  |
| **Description** | The secondary description of the IUCR code, a subcategory of the primary description. | | Plain Text | |  |
| **Location Description** | Description of the location where the incident occurred. | | Plain Text | |  |
| **Arrest** | Indicates whether an arrest was made. | | Checkbox | |  |
| **Domestic** | Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act. | | Checkbox | |  |
| **Beat** | Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at <https://data.cityofchicago.org/d/aerh-rz74>. | | Plain Text | |  |
| **District** | Indicates the police district where the incident occurred. See the districts at <https://data.cityofchicago.org/d/fthy-xz3r>. | | Plain Text | |  |
| **Ward** | The ward (City Council district) where the incident occurred. See the wards at <https://data.cityofchicago.org/d/sp34-6z76>. | | Number | |  |
| **Community Area** | Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at <https://data.cityofchicago.org/d/cauq-8yn6>. | | Plain Text | |  |
| **FBI Code** | Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at <http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html>. | | Plain Text | |  |
| **X Coordinate** | The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block. | | Number | |  |
| **Y Coordinate** | The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block. | | Number | |  |
| **Year** | Year the incident occurred. | | Number | |  |
| **Updated On** | Date and time the record was last updated. | | Date & Time | |  |
| **Latitude** | The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block. | | Number | |  |
| **Longitude** | The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block. | | Number | |  |
| **Location** | The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block. | | Location | |  |

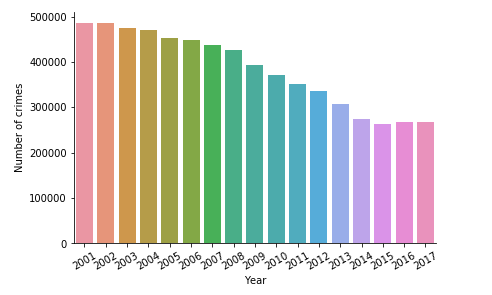
Snapshot of Data



# Data Exploration

Let’s explore the data from a different angle –

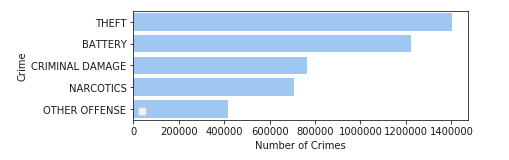
Let’s see the number of crimes over the years –



From the above graph it is almost clear that the crime rate is decreasing over the years.

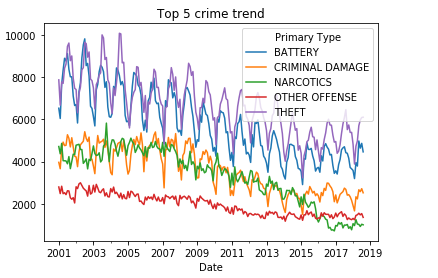
Slicing and Dicing the data from different angle.

Top 5 types of primary crimes are - THEFT, BATTERY, CRIMINAL DAMAGE, NARCOTICS, OTJER OFFENSE



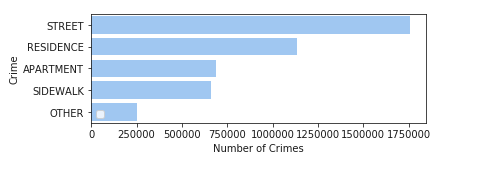
Let’s see their trend over the years –

From the below graph we can say that rate of the Primary type crime is decreasing over the years especially Battery and theft crime. This information can be used by CPD, with this analysis, they can focus on which type of crime they need to focus more in order to see the decreasing trend in future.



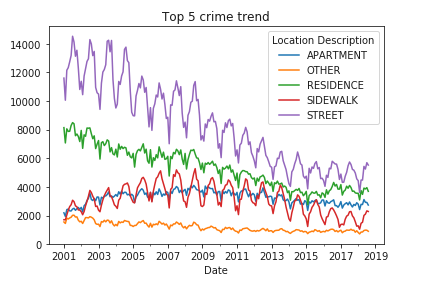
Similarly,

Let’s now see the same information from location angle. From the below graph we can say that top five crime locations are – STREET, RESIDENCE, APARTMENT, SIDEWALK and OTHER.



Let’s see the trend of crime of top 5 locations over the years.

From the past information we already know that crime rate was decreasing over the years, the below graphs reflects the similar kind of information. We can see that there is a significant decline in street crime. Residence crime is also showing declining trend. Although there seems to be little decline in other locations. Now our CPD has the information from this analysis, to decide where to focus more and how they have been performing in their past.



# Data Modelling

Clearly this is a time series problem, there are multiple ways to solve this problem. Either we can use linear regression, Prophet, mutlti output regressor and Gradient Boosting Regressor( we can also use KNN).

**The basic steps for Linear Regression are**

Steps for that –

1) Smoothening and normalizing the data.

2) Implementing linear regression using sklearn.

We can also use Prophet library –

**The basic steps for Prophet are -**

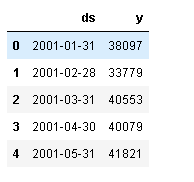
Prophet uses Fourier series and is very strong in time series prediction having seasonality and trend patterns. Clearly from the above graph it is clear we have trend and seasonality in our crime pattern

Steps 1) Preprocessing of data – since it accepts the data in the form of date column as “ds” and target variable as “y”.

Aggregating the data monthly – with new column names as –

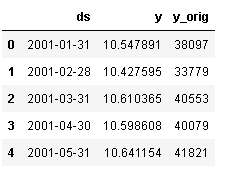
“ds” – date parameter (Input variable)

“y” – Count of crimes (Output variable)



Before we can do any analysis with this data, we need to log transform the ‘y’ variable to a try to convert non-stationary data to stationary. This also converts trends to more linear trends.

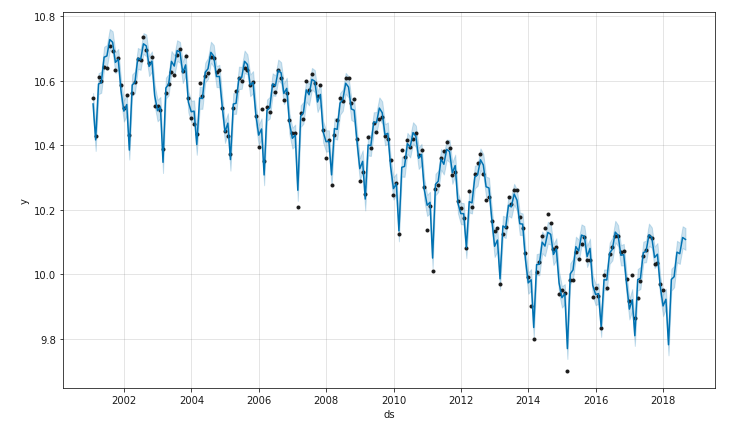
Here y is after taking log (using NumPy – np.log () function). Later on, we will convert this by using exp () – taking exponential



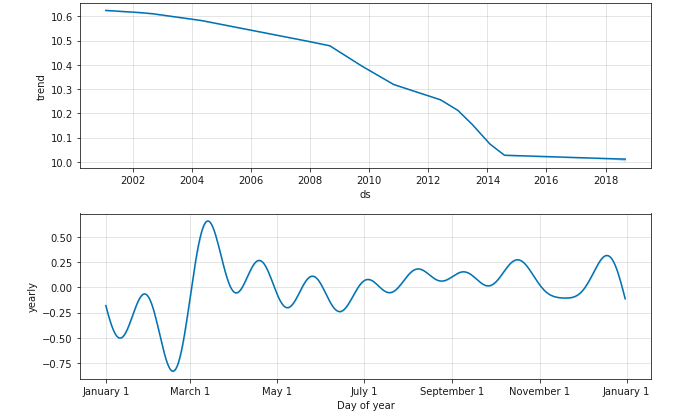
**Finally making predictions –**

Making predictions for the year 2018. Predictions are then made on a data frame with a column ds containing the dates for which a prediction is to be made. The predict method will assign each row in future a predicted value which it names yhat.

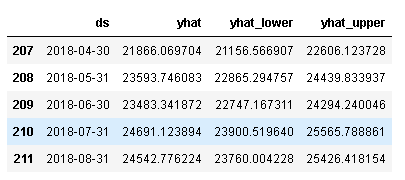
Let’s see how the forecast plot looks like –



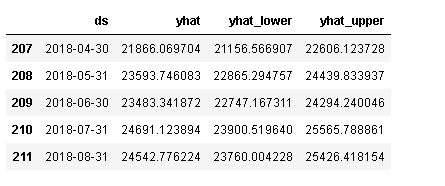
Forecast broken down to trend and yearly seasonality



Snapshot of how forecast looks like -



But this is not we what, remember we took the log initially, now it’s time to take the exponential in order to get the right form of output. Here is the snapshot after taking exponential –



**Calculating root mean square error –**

To see whether our model is performing good or not or to compare it with other models, it is advisable to calculate some sort of metric to figure out the accuracy of model.

***RMSE: 649.529007***

***Although it is not good. We can perform more tuning or implement other time series methodology to find a better model. This is one of the methods.***

# Future work –

Let's try some other method which groups the predictions based on Crime primary type - multivariate regression problem

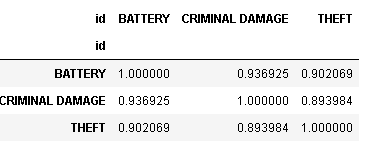
For this I am considering top three categories of crime type - "THEFT" "BATTERY" and "CRIMINAL DAMAGE"

First step –

1. Find the correlation between the dependent variables, if they are highly correlated than we cannot run independent regression on them (Technically we can, but it is not advisable – accuracy would be poor)
   1. For this I created correlation matrix

**Correlation matrix -**

Clearly it shows high correlation between battery crime, criminal damage and theft. Therefore, we cannot run a model independently on each of them Technically we can run a model but the accuracy will not be good enough. So, it is advisable not to run models on each of them separately.



1. Since they are highly correlated, I am trying Multi output regressor and Gradient boosting regressor. I have to figure out a way to implement these methods on time series data.

# Takeaways –

In this report I have tried to figure out the top 5 types of crime and their locations, which can alert CPD and they can work on these areas to make the city more secure.

Also, I have tried predicting the crimes in 2018, which can help us in identifying the number of crimes and based on that data, police force can be hired and prepared for emergency situations.

**Thank you**