CHICAGO CRIME DATA ANALYSIS

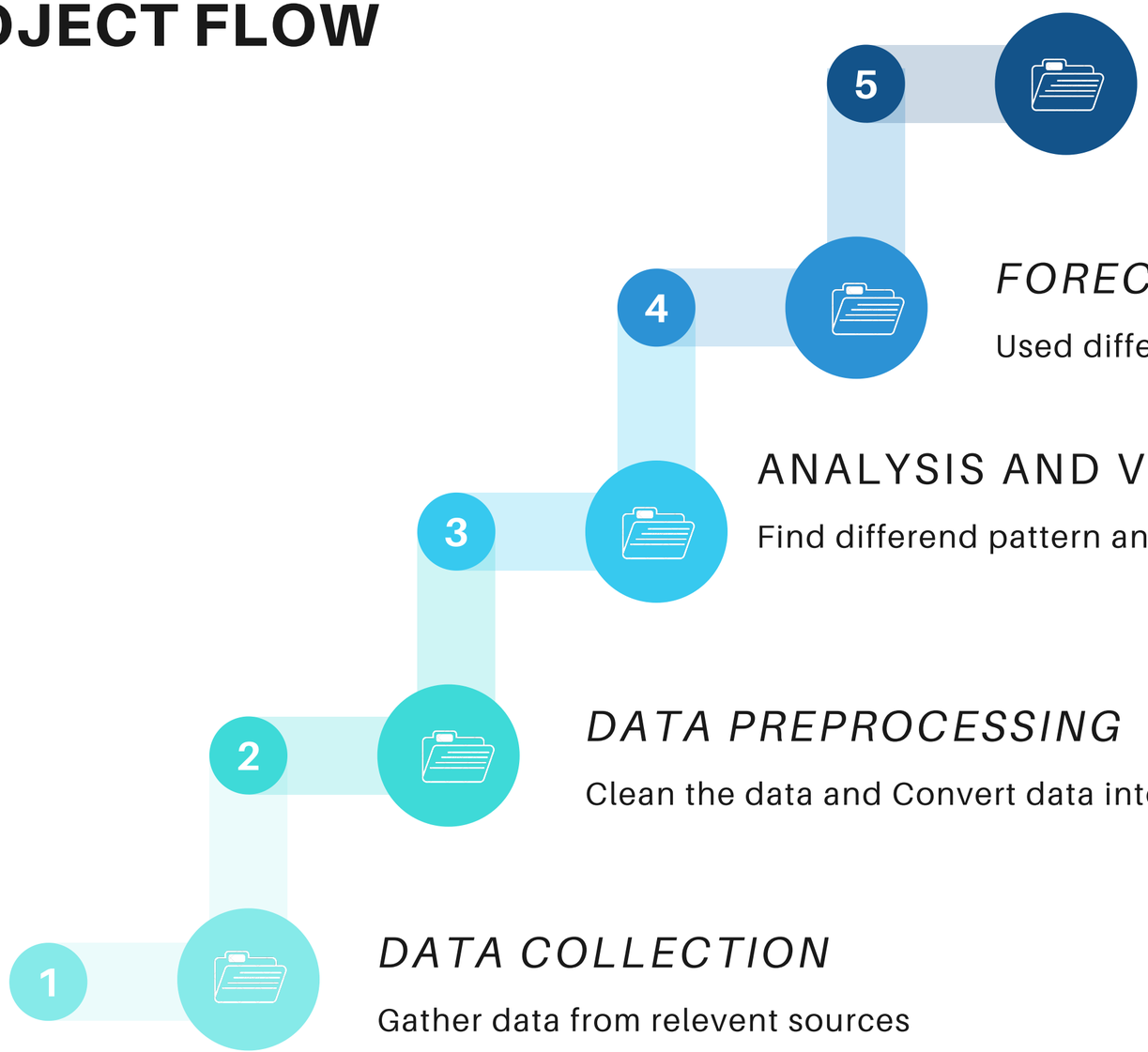
OVERVIEW

Main aim of this project is to analyze and forecast crime data of Chicago city in USA

This project analyzes crime data with the help of "Ivarious visualizations for easy understanding

Used 2010 to 2020 years’ of crime data from US government website to forecast future crime rate.

PROJECT FLOW



ERROR CALCULATION

Check error rate of each forecasting model eith the help of RMSE matrics

FORECASTING

Used different forecasting model for future prediction

ANALYSIS AND VISUALIZATIONS

Find differend pattern and flow of data

DATA PREPROCESSING

Clean the data and Convert data into appropriate format

DATA COLLECTION

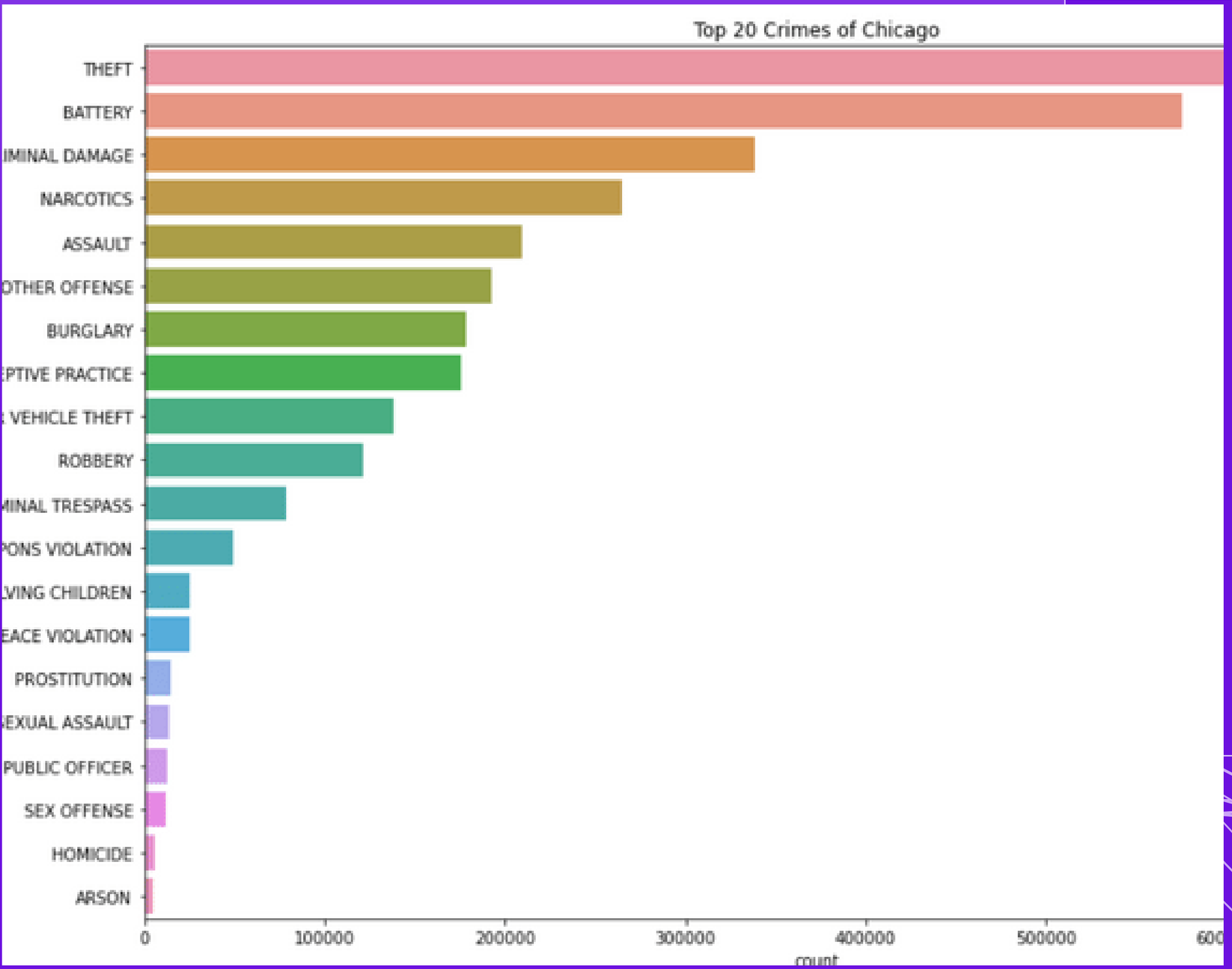
Gather data from relevent sources

ABOUT DATA SET

The dat set contains 3167036 rows and 11 columns

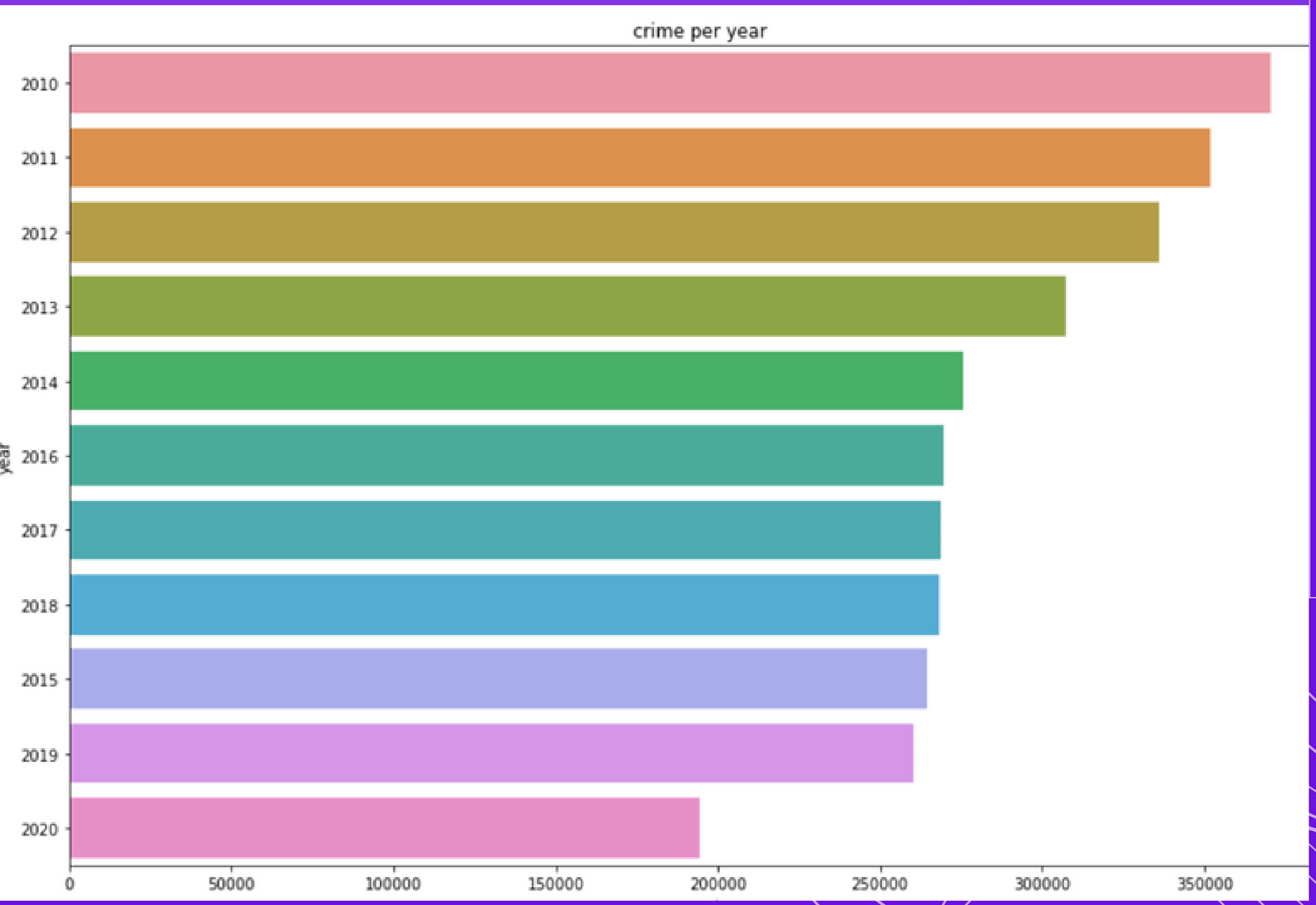
#	Column	Dtype
0	ID	int64
1	date	datetime64[ns, UTC]
2	primary_type	object
3	description	object
4	location_description	object
5	arrest	bool
6	domestic	bool
7	beat	int64
8	community_area	float64
9	year	object
10	location	object
11	Month	object
12	Day	object
13	hours	int64

DATA ANALYSIS AND VISUALIZATION



VISUALIZATION OF TOP 20 CRIMES

Theft and battery was most occuering crimes in Chicago city.
High rate of battery indicate the physically violent community

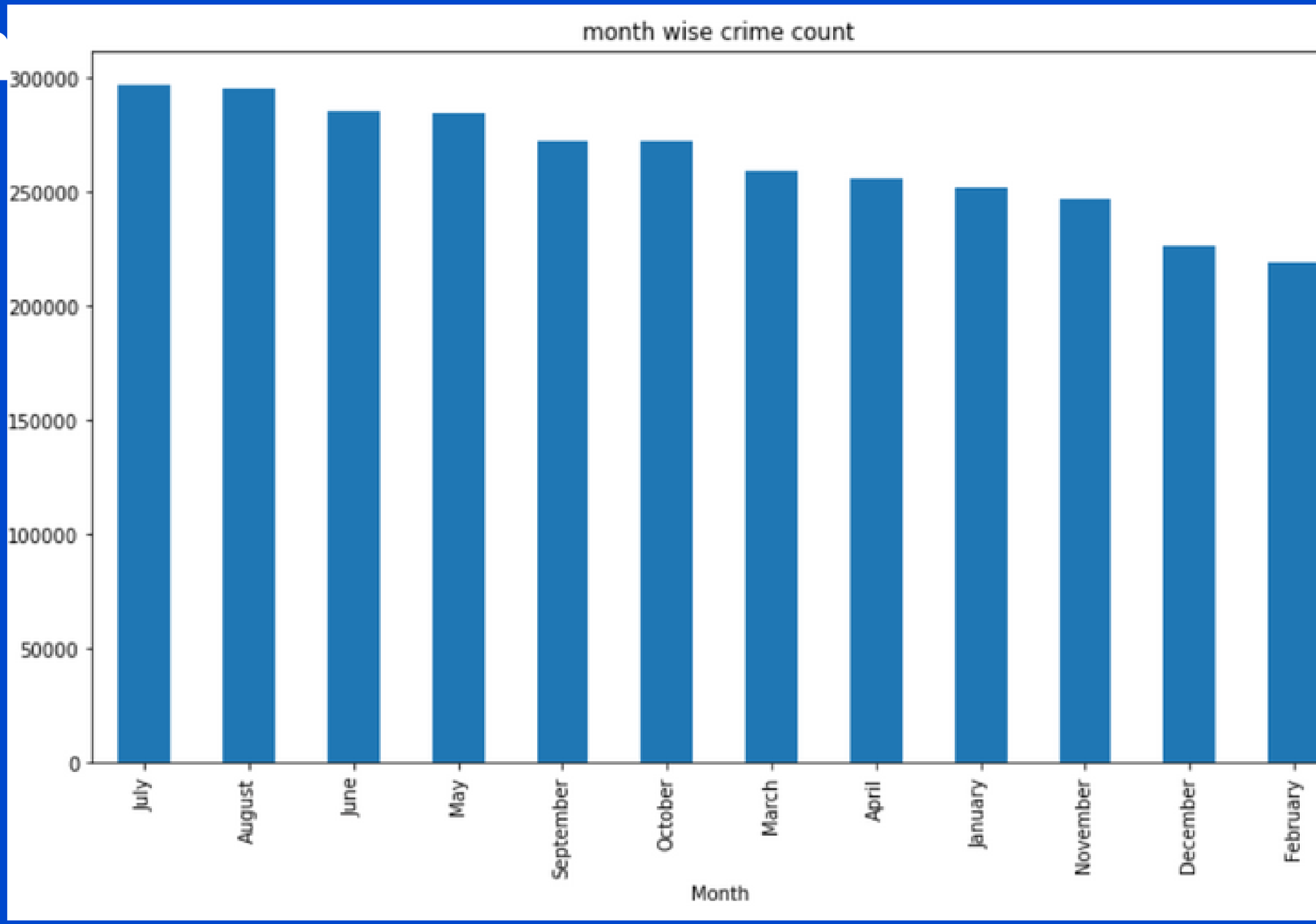


VISUALIZATION OF CRIME COUNT IN YEAR WISE

Most of the crime occured during the years 2010 2013. After 2013 we can see that crime rate was decreasing

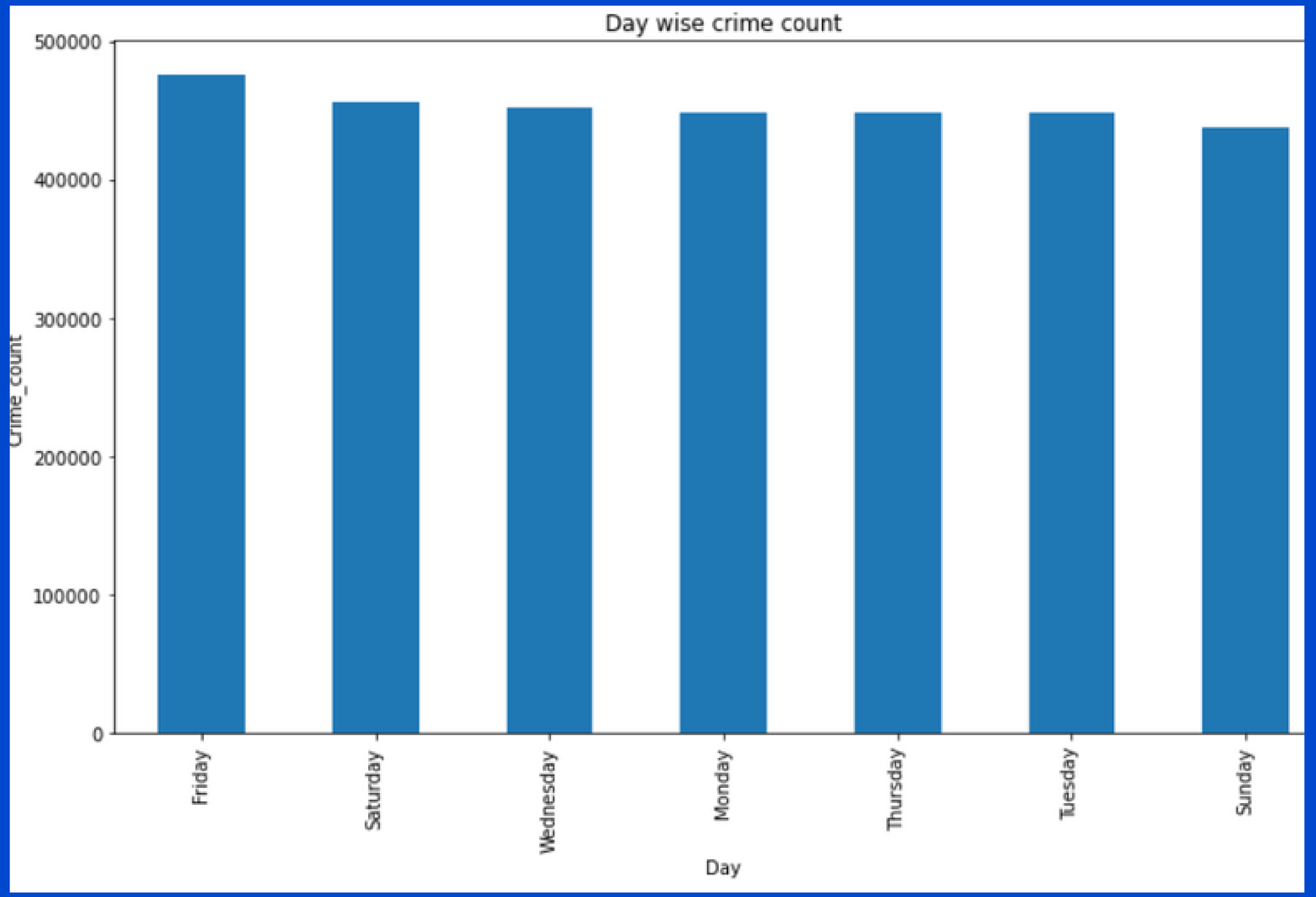
visualization of crime count in Month wise

Most of the crime occured during months june july august. Crime rate during winter time was very less



visualization of crime count in Day wise

Here we can see that crime count on friday is slightly high compared to other days.Remaining days has equal distribution of crime.

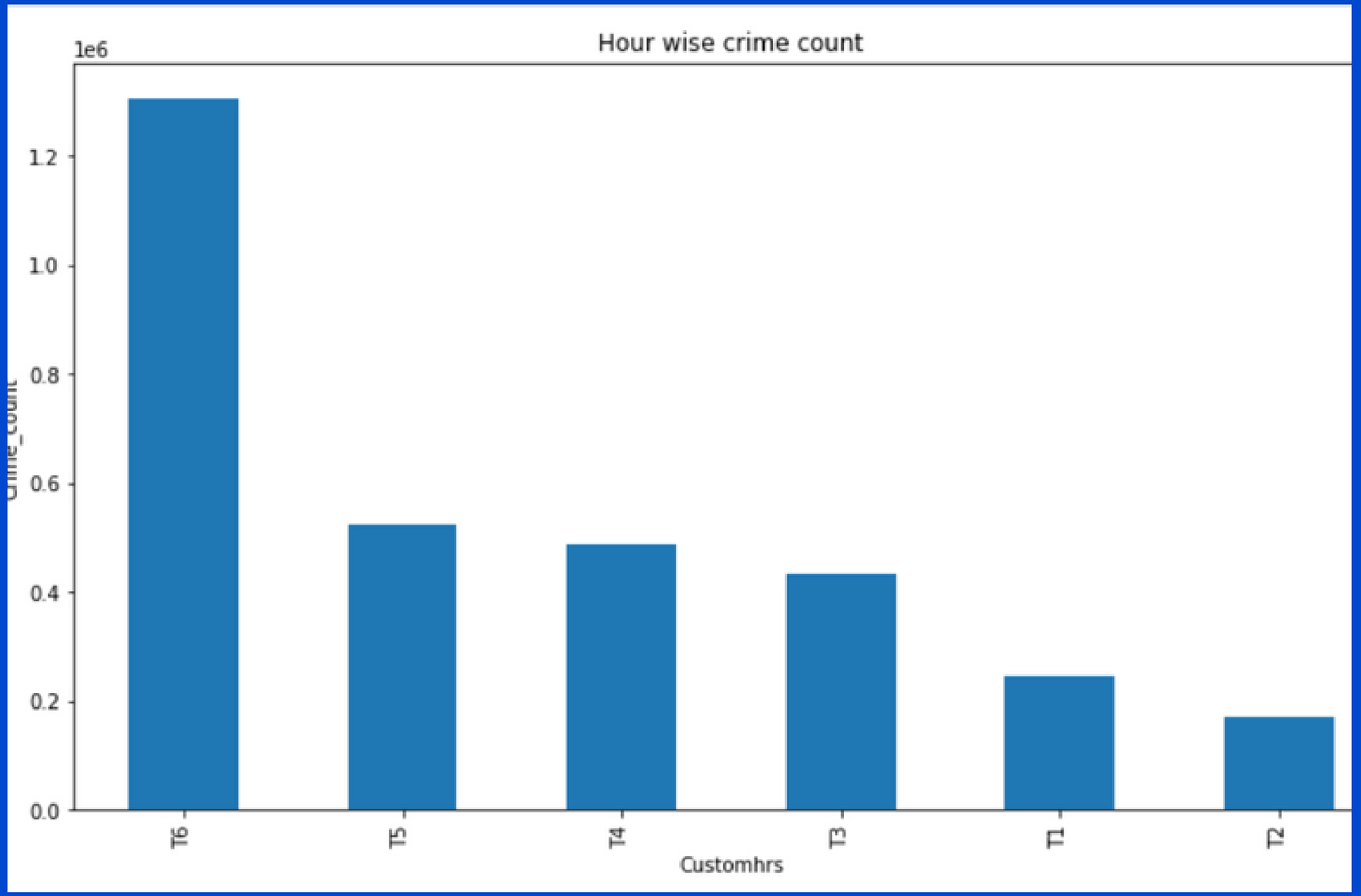


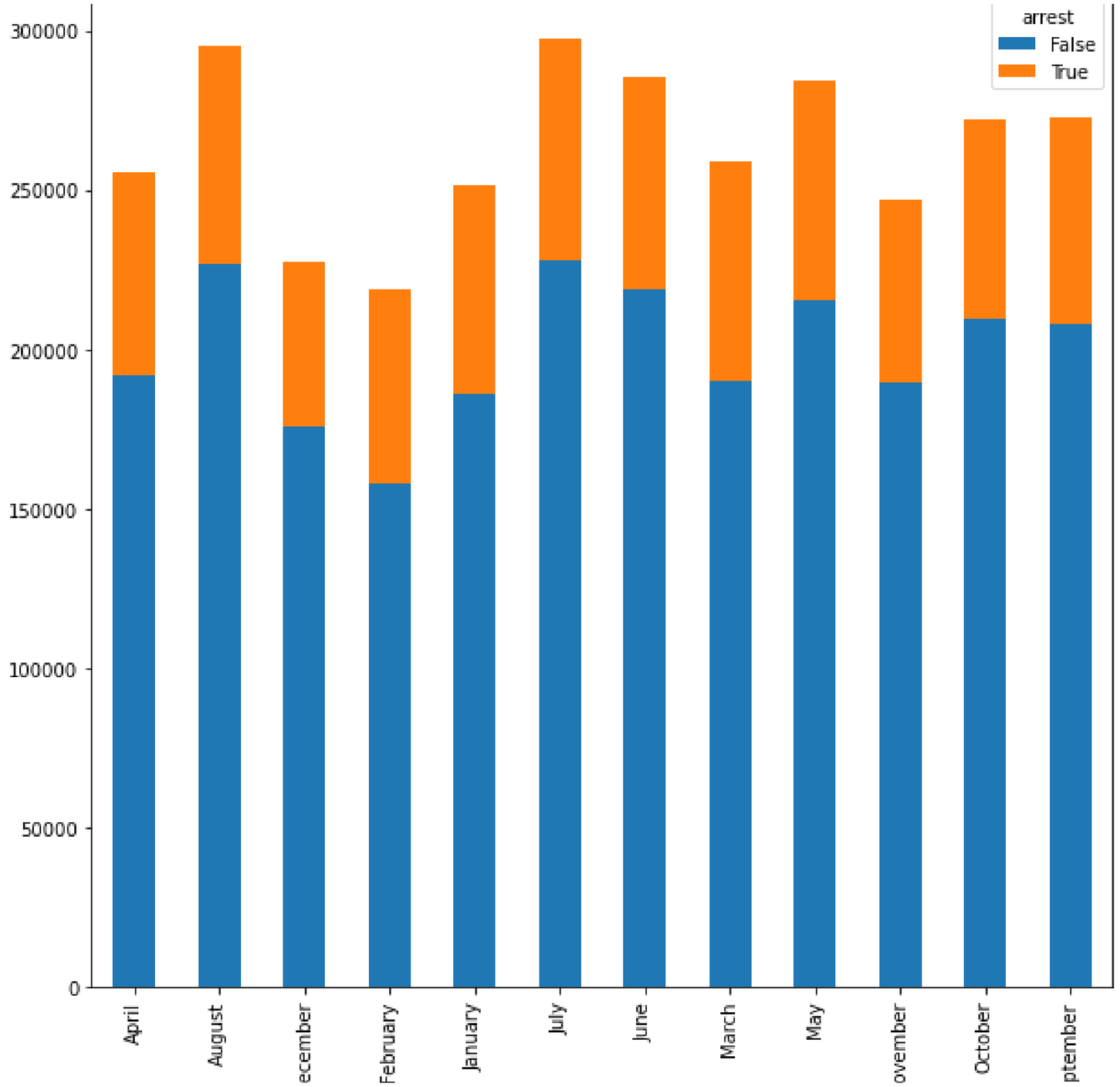
→

Visualizations of hourly occurrence of crime.

T1: 12 AM TO 4 PM
T2: 4 AM TO 8 AM
T3: 8 AM TO 12 PM
T4: 12 PM TO 4 PM
T5: 4 PM TO 8 PM
T6: 8 PM TO 12 AM

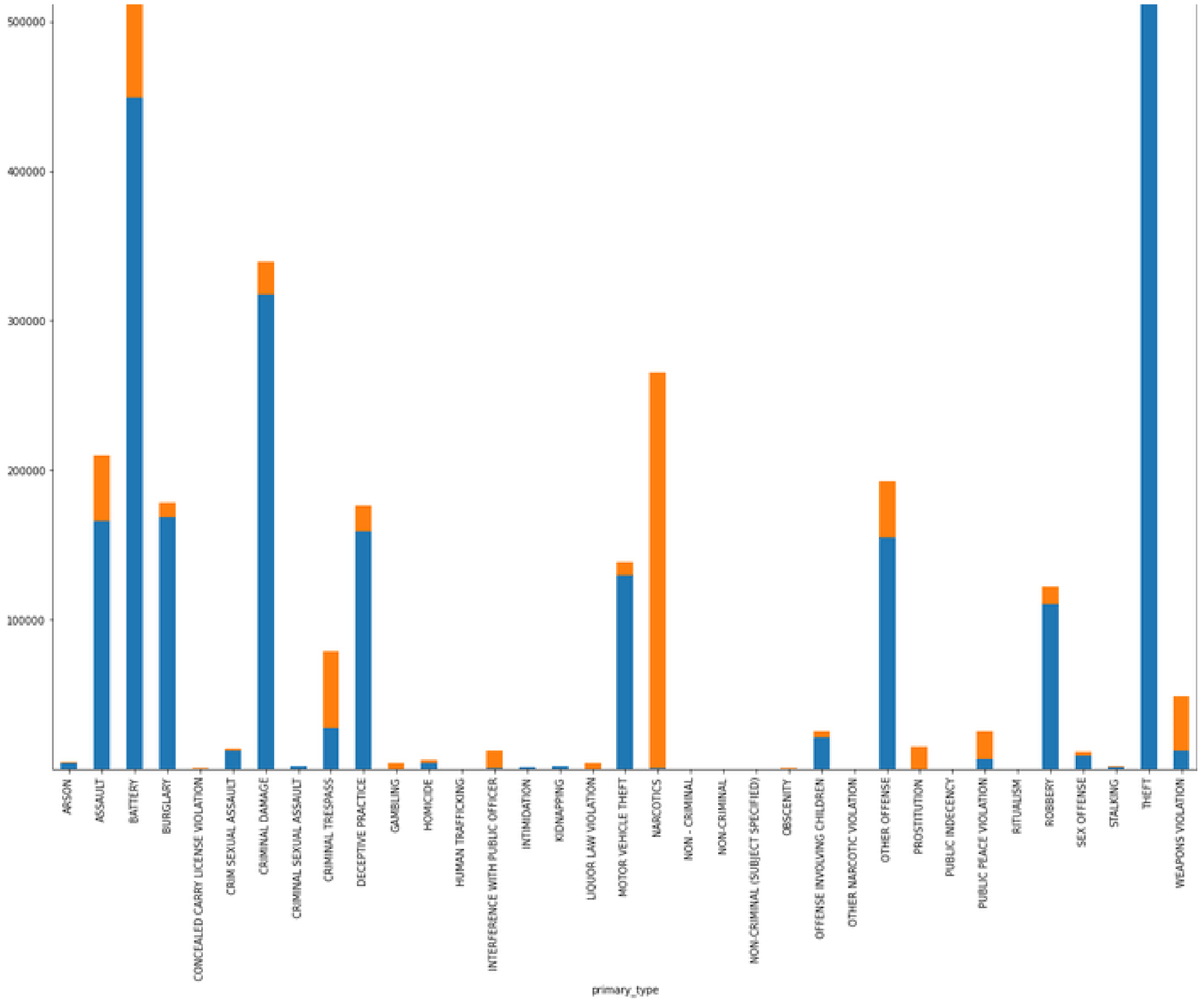
More number of crimes happening at night from 8 pm to 12 pm The above visualization helps to understand that residents of Chicago need to be safe during nights.





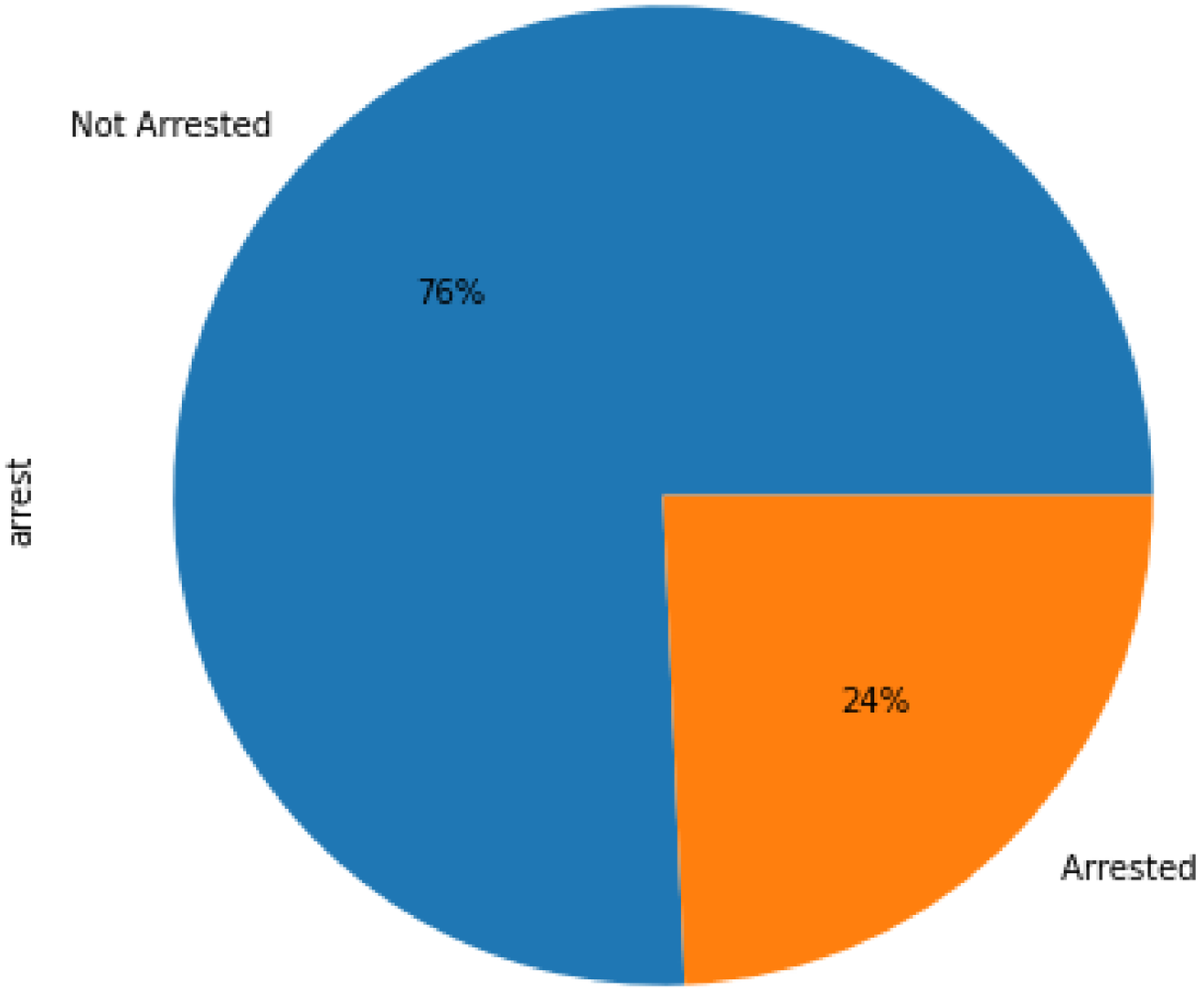
ANALYSIS OF ARREST IN MONTH WISE

Here we have analysed arrest of each month . True indicate arrest was made and false indicate not made.in this visualization we can see that the cout of False is very high



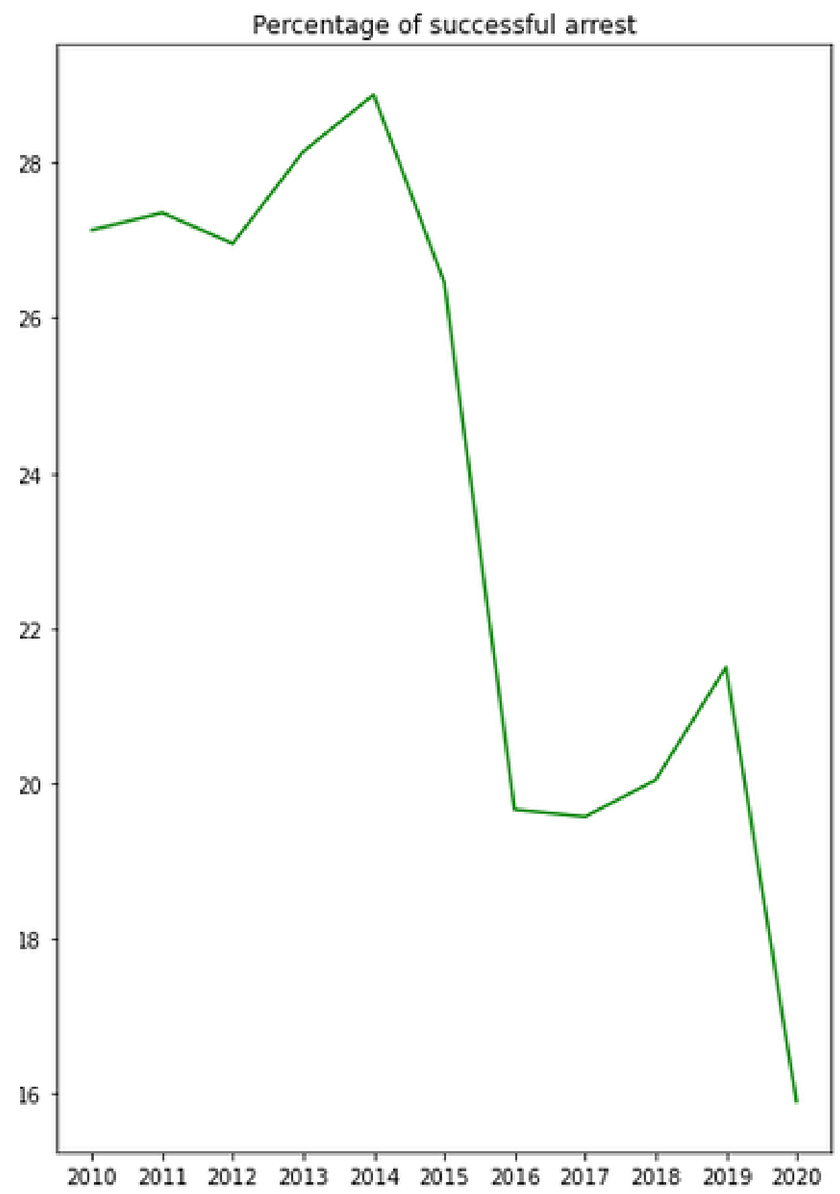
#ANALYSIS OF ARREST IN PRIMARY_TYPE WISE

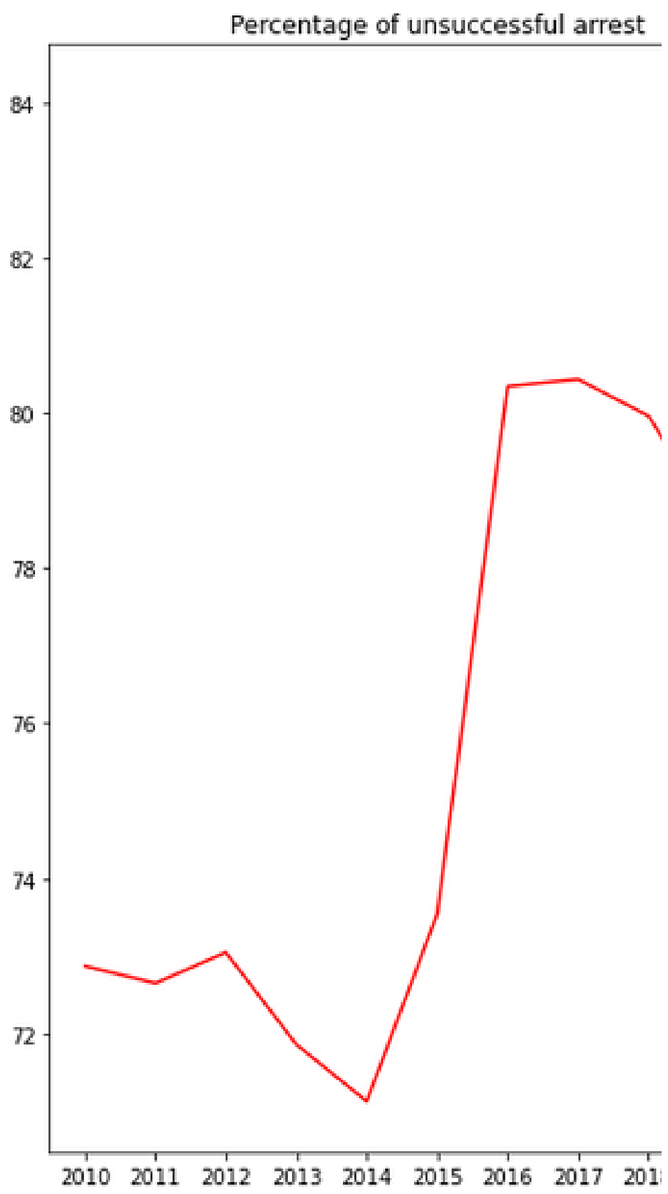
narcotics has more number of arrest



PERCENTAGE OF ARREST USING PIE CHART

in pie chart representation we can see that 76% of crimes has no arrest



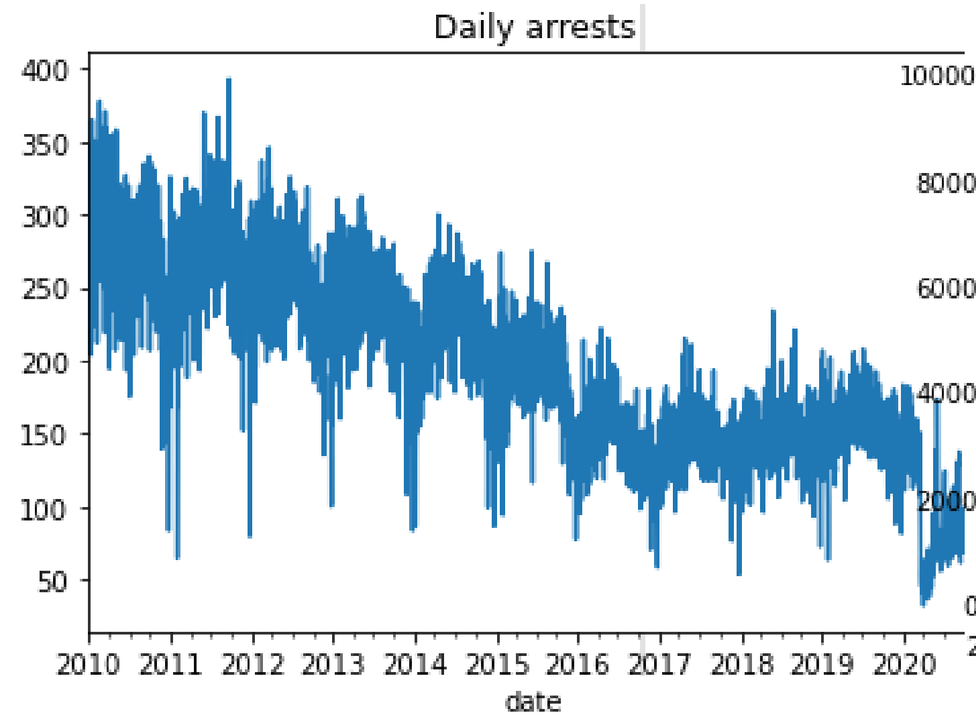


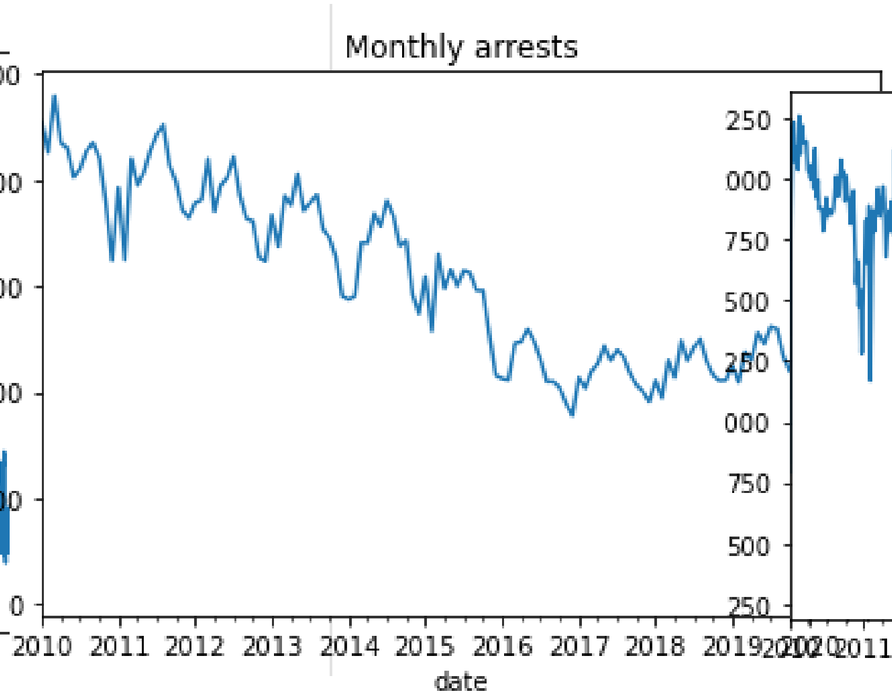
ARREST PERCENTAGES PER YEAR

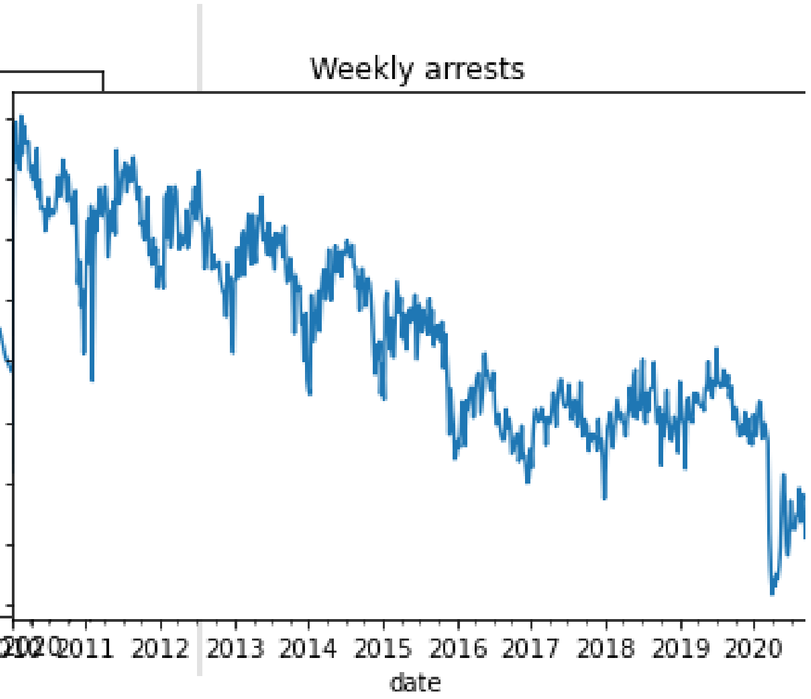
Year 2016 to 2020 unsuccesful arrest percentage is high t

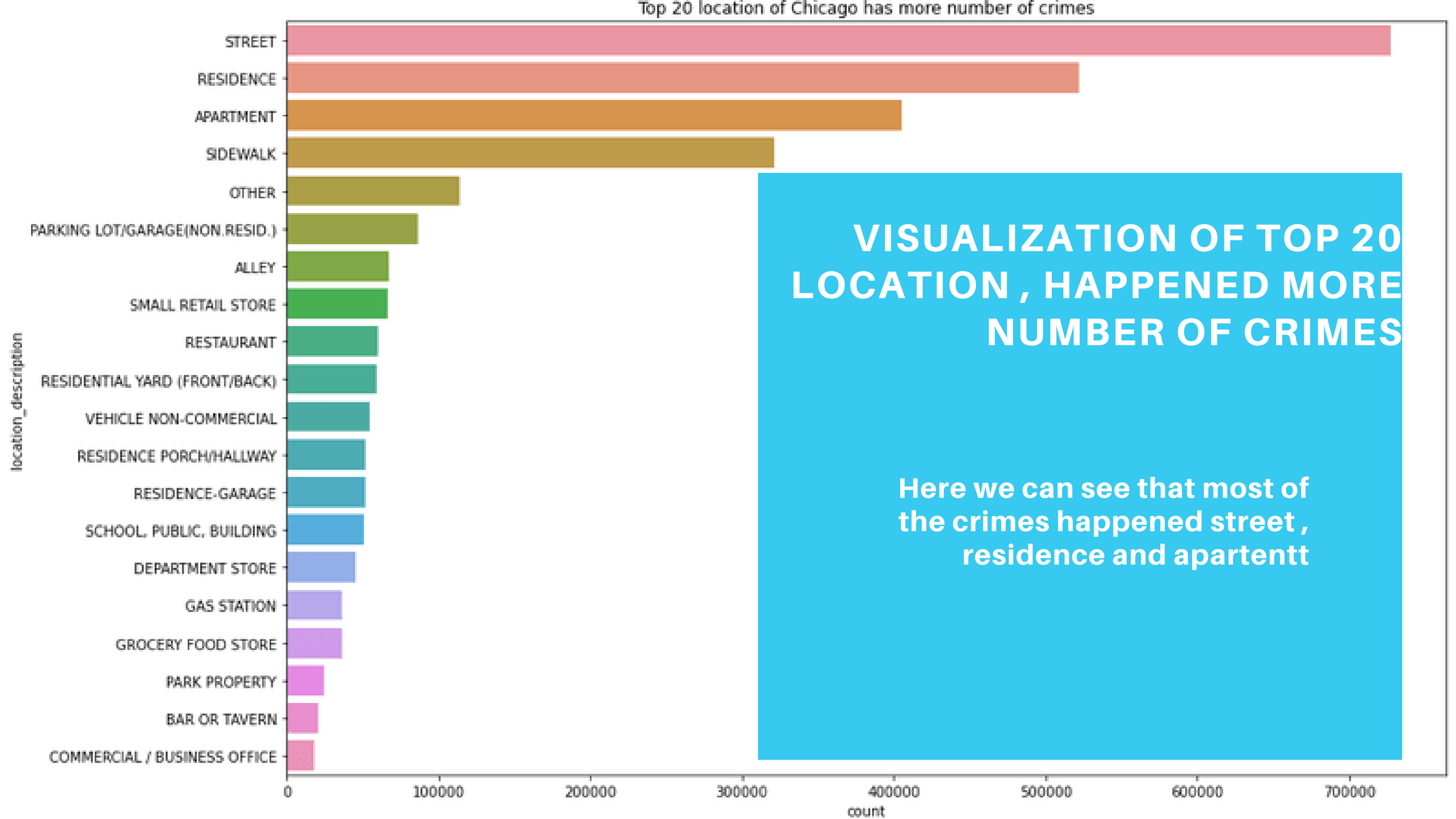
PLOTTED ARREST ON THE BASIS OF MONTH,WEEK AND DAY

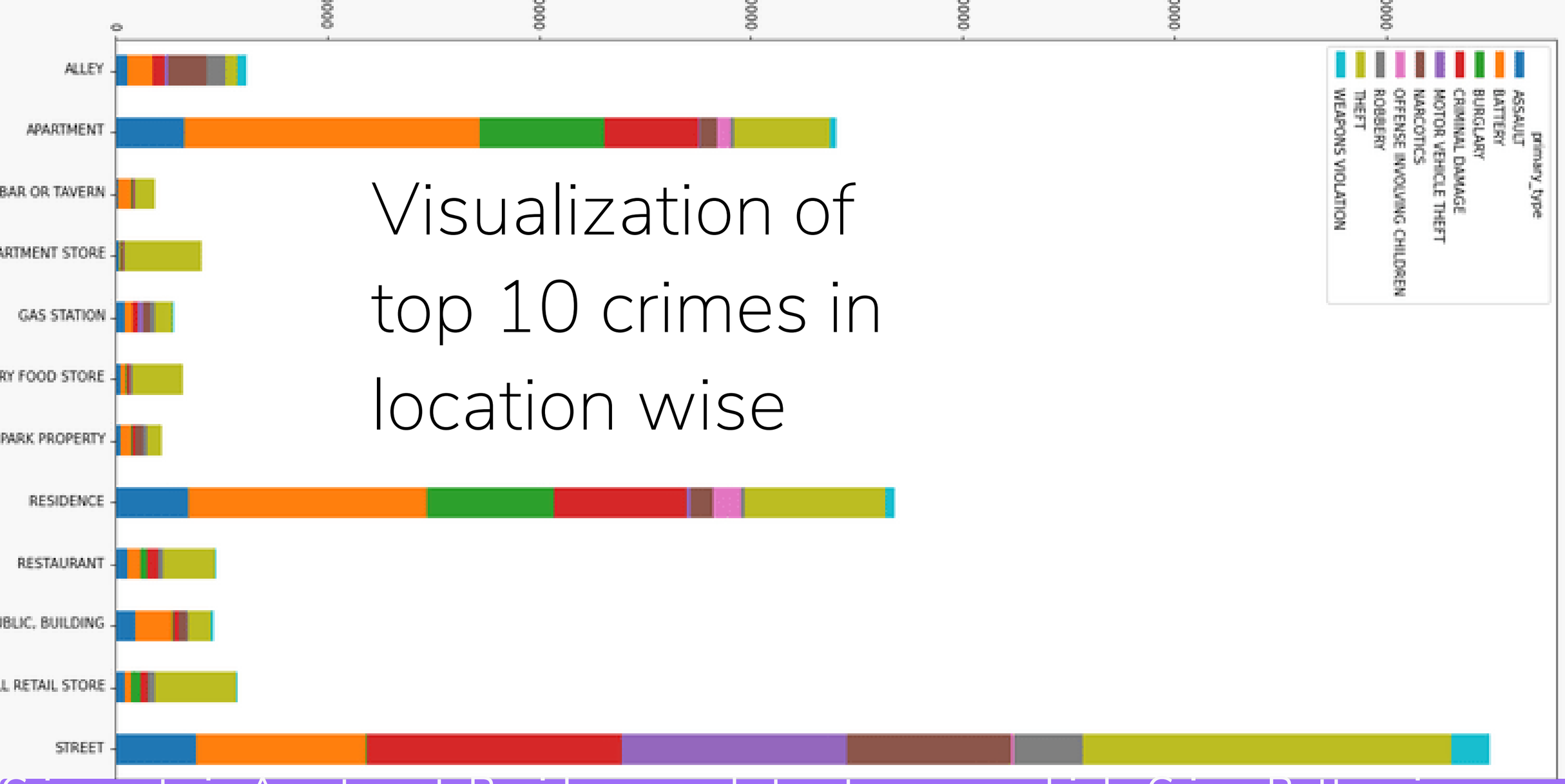
Here we can see that a downward trend on arrest. in all cases successful arrest rate is decreasingr



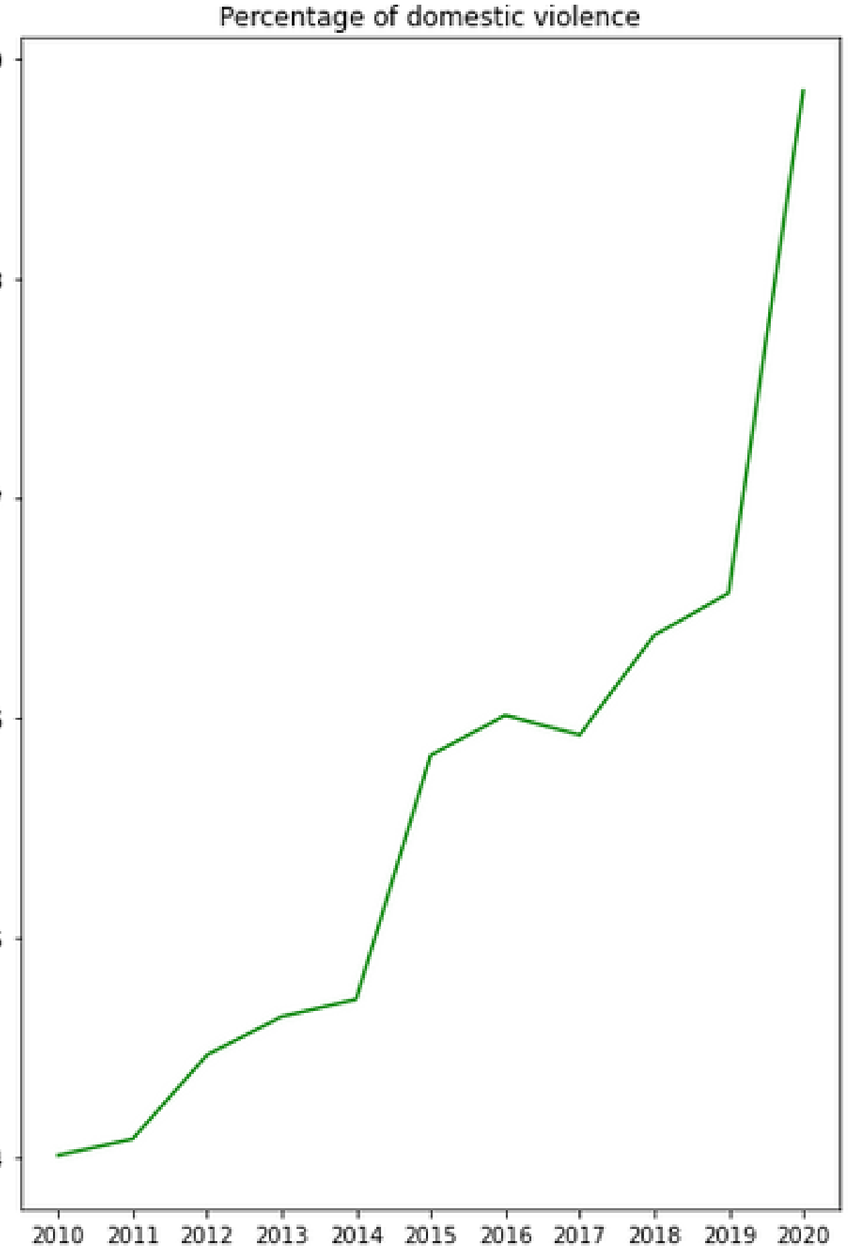


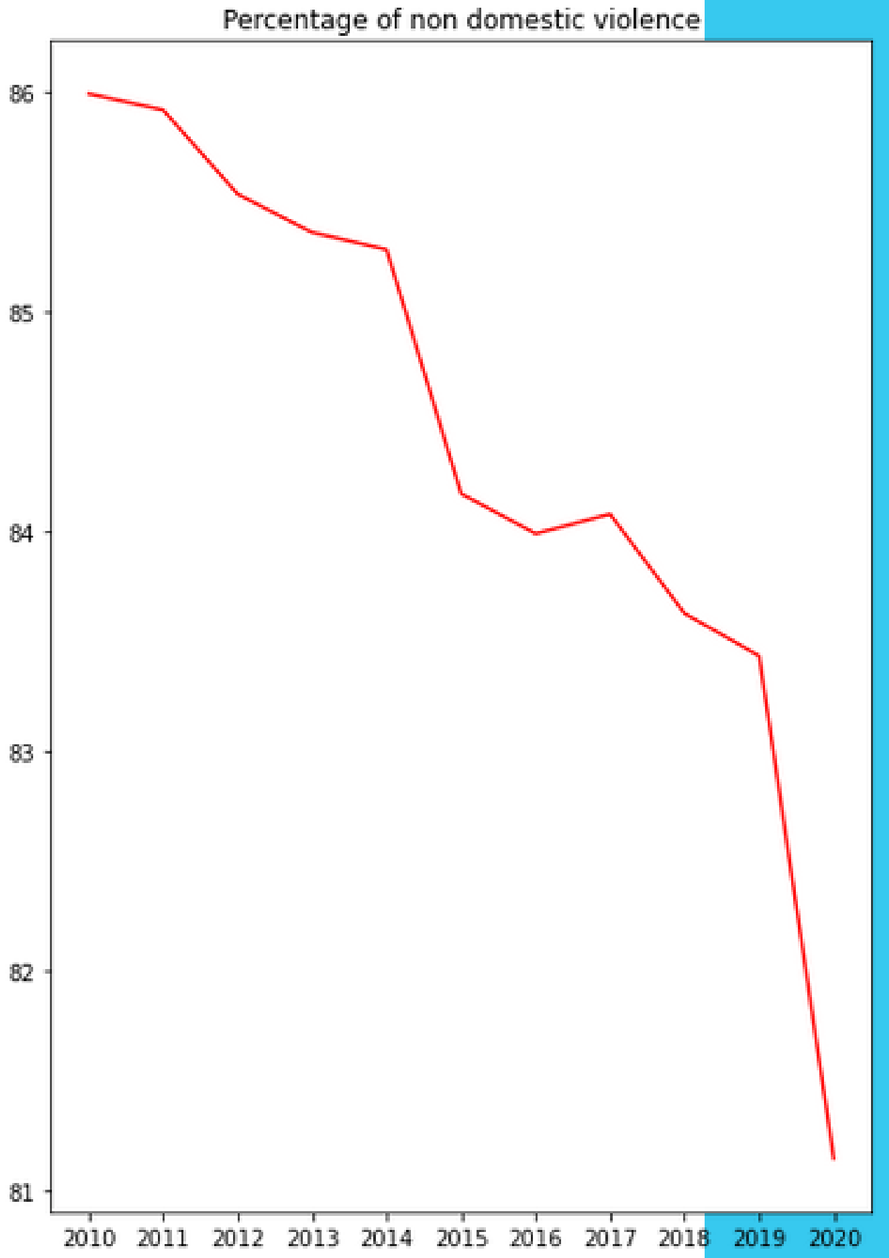



Top 20 location of Chicago has more number of crimes
location_description
STREET
RESIDENCE
APARTMENT
SIDEWALK
OTHER
PARKING LOT/GARAGE(NON.RESID.)
ALLEY
SMALL RETAIL STORE
RESTAURANT
RESIDENTIAL YARD (FRONT/BACK)
VEHICLE NON-COMMERCIAL
RESIDENCE PORCH/HALLWAY
RESIDENCE-GARAGE
SCHOOL, PUBLIC, BUILDING
DEPARTMENT STORE
GAS STATION
GROCERY FOOD STORE
PARK PROPERTY
BAR OR TAVERN
COMMERCIAL / BUSINESS OFFICE
0
100000
200000
300000
400000
500000
600000
700000
count
VISUALIZATION OF TOP 20 LOCATION , HAPPENED MORE NUMBER OF CRIMES
Here we can see that most of the crimes happened street , residence and apartentt



Crime rate in Apartment, Residence and street was very high. Crime Battery in Residence and Apartment is very high. so people who living in those should be more carefull



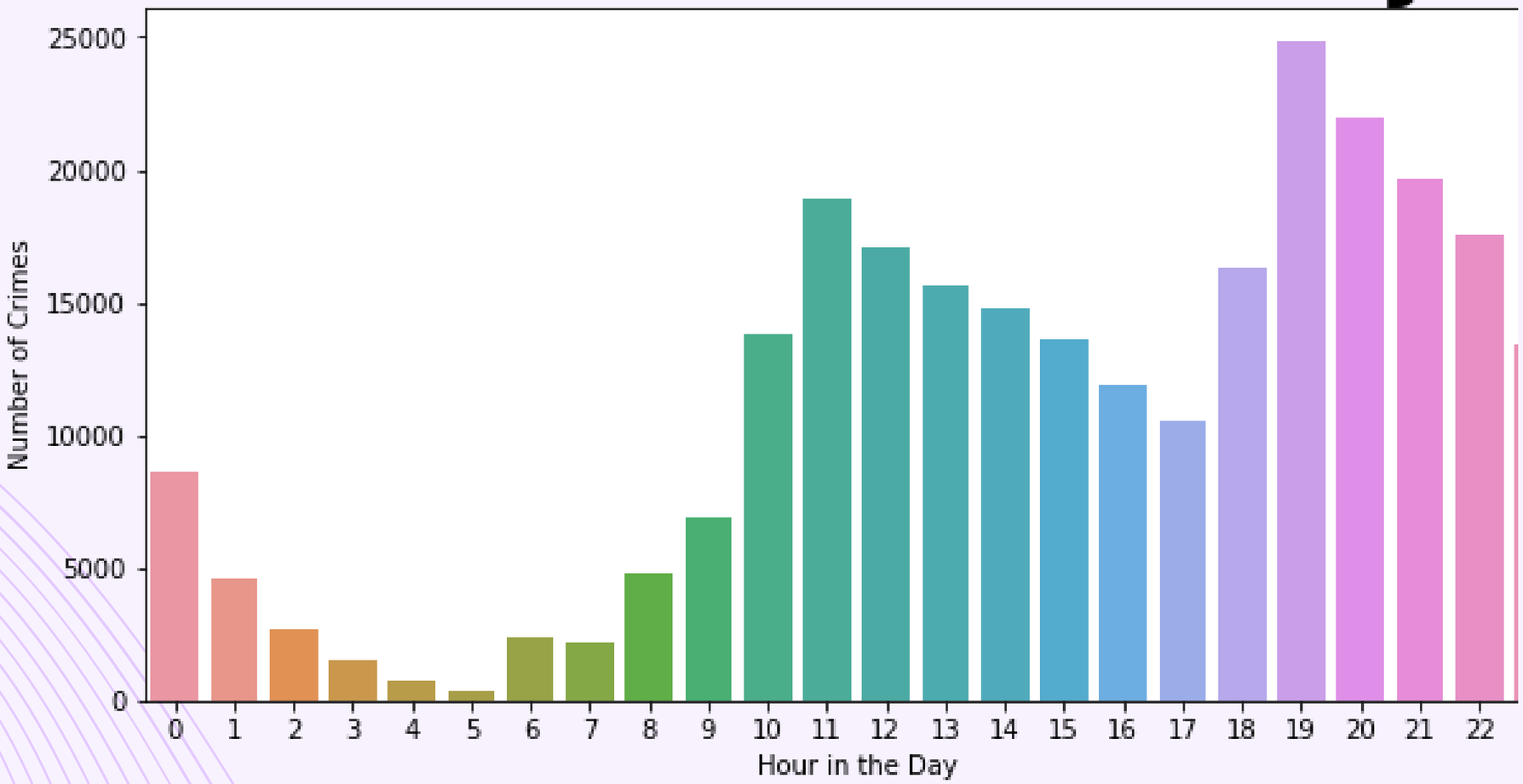


ANALYSIS OF DOMESTIC RELATED CRIME PERCENTAGE

here we can see that every year domestic violence increase. in 2020 domestic violence was in peak position

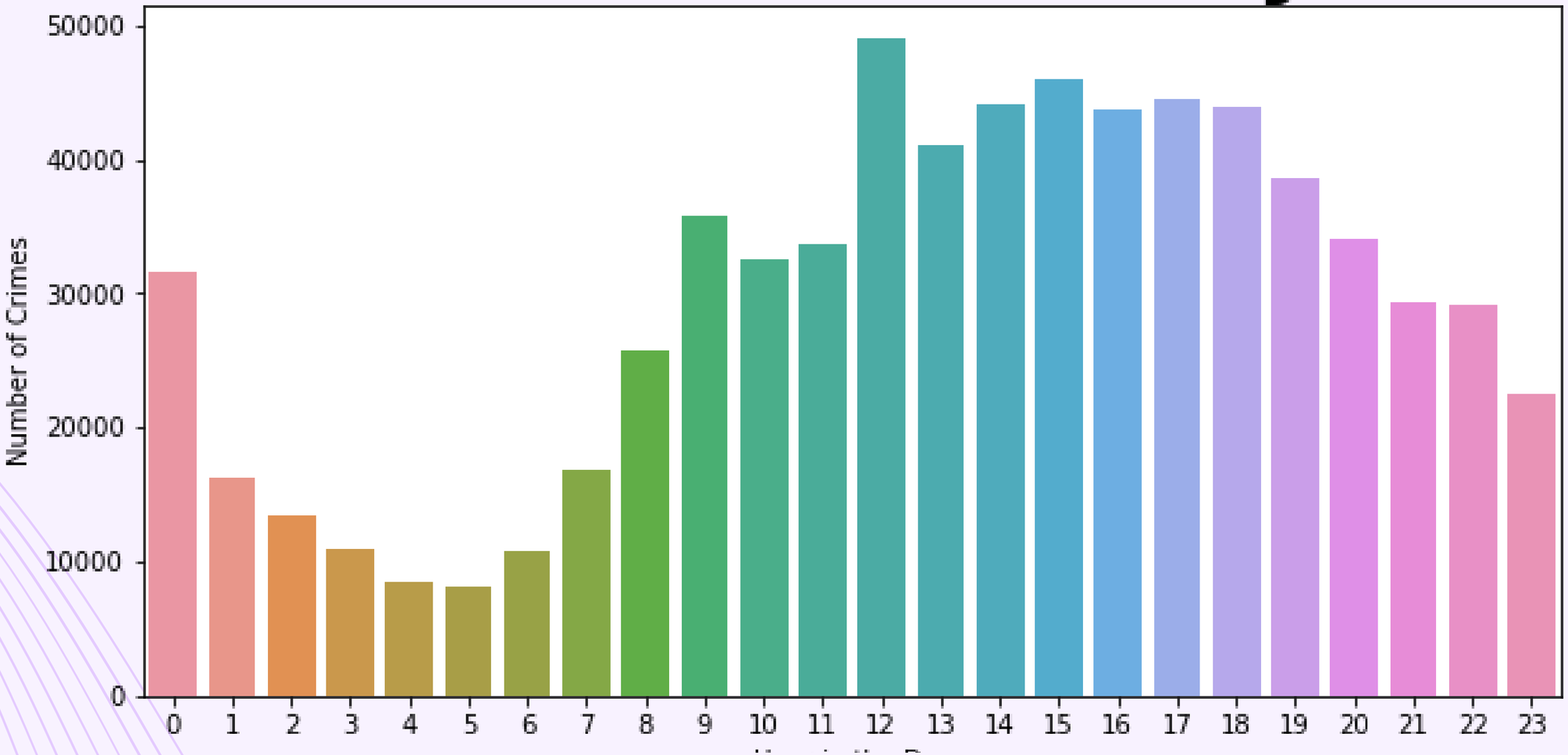
PATTERN OF DIFFERENT CRIME OVER A DAY

Narcotics over a day



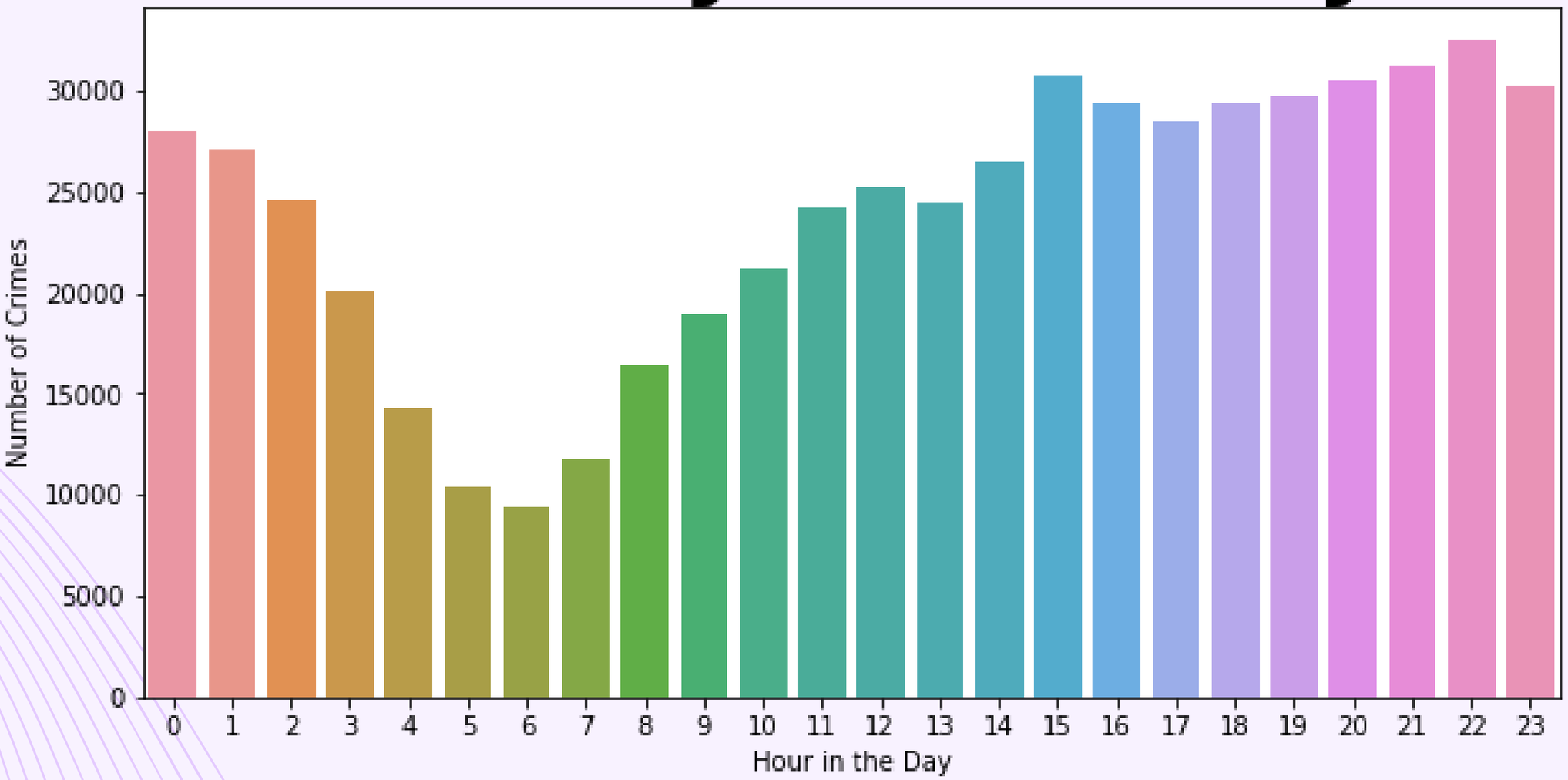
Most of the Narcotics rate increases at mid night

Theft over a day



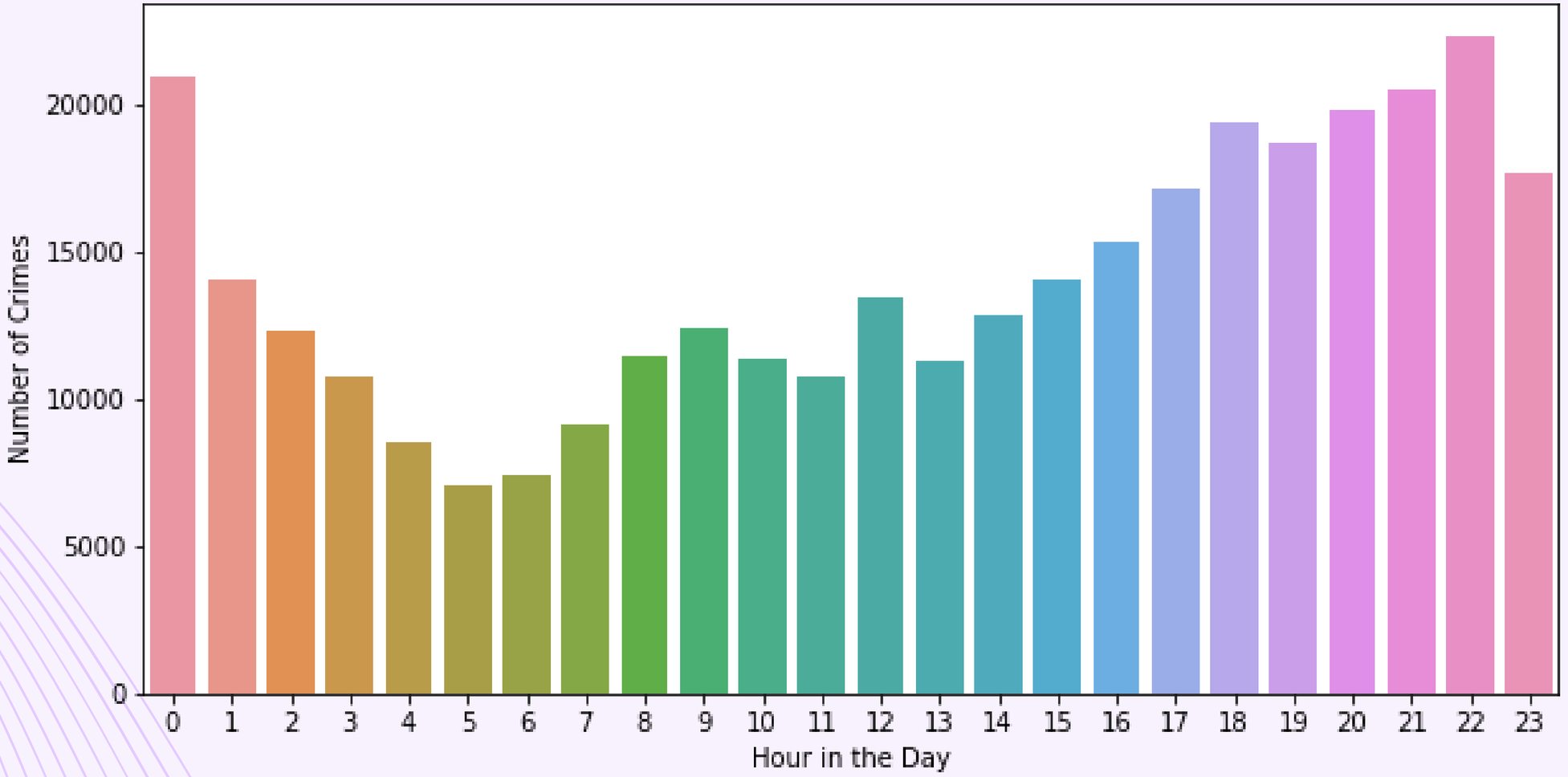
Most of the Theft rate increases at day time

Battery over a day



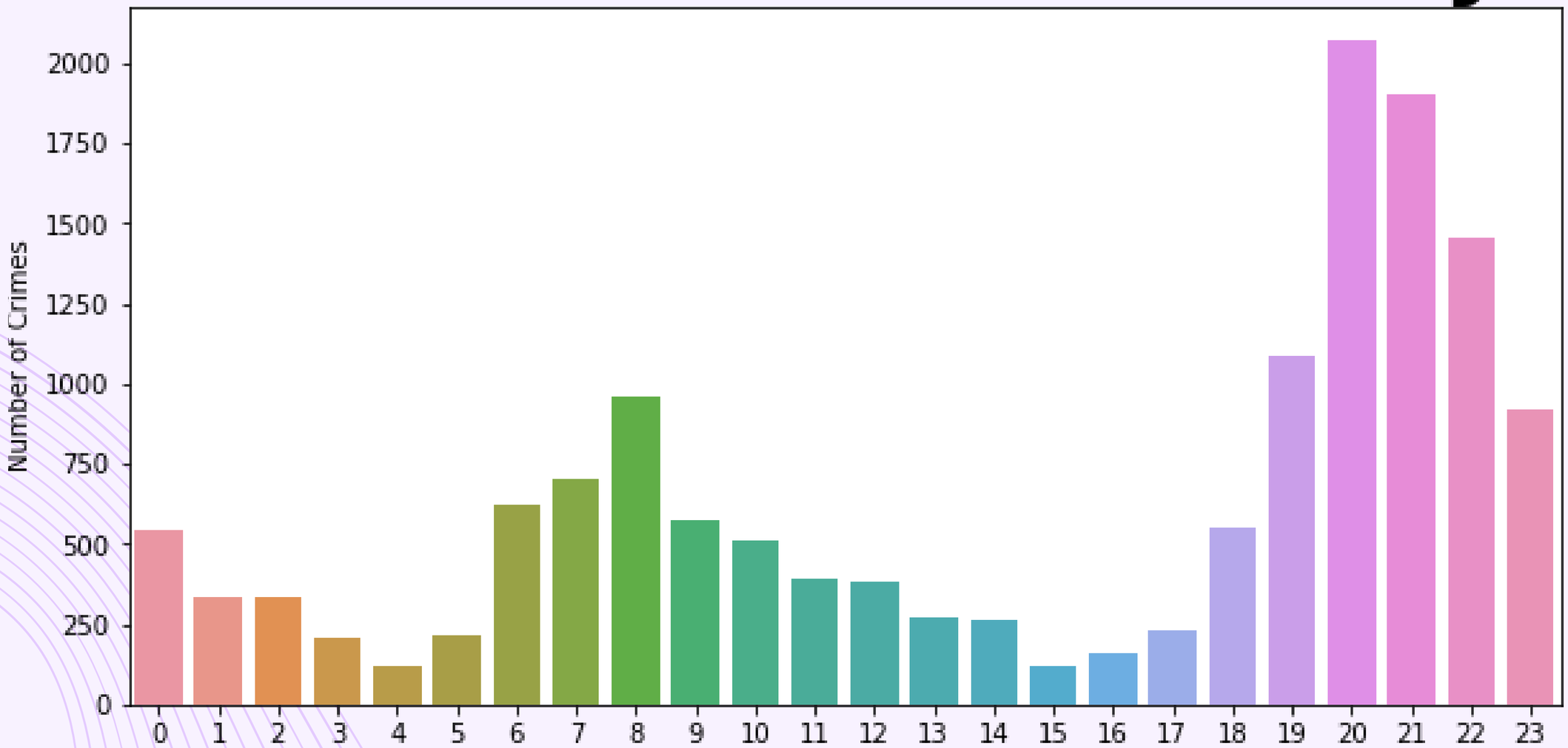
Most of the Battery rate increases after sunset

CRIMINAL DAMAGE over a day



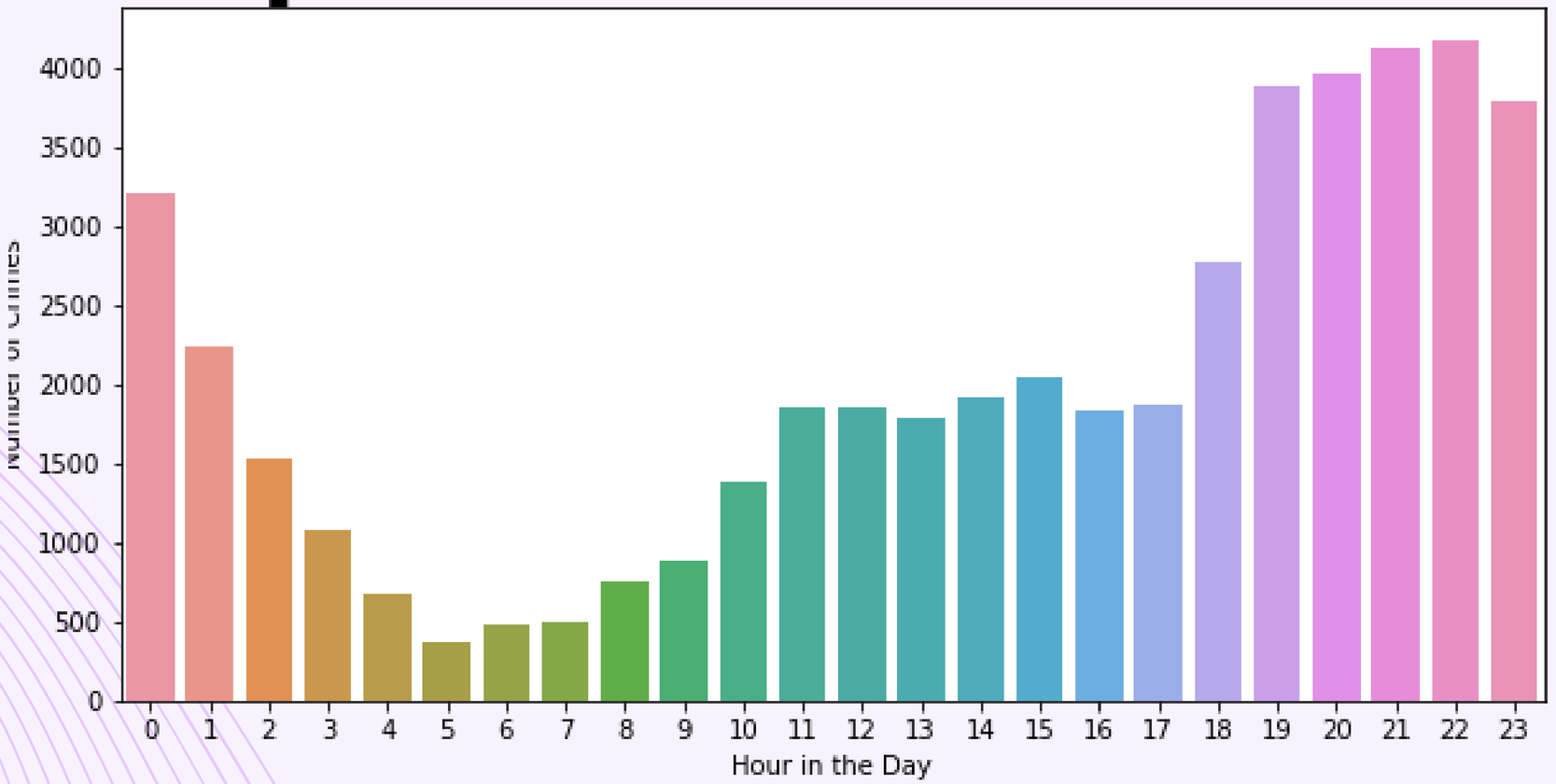
Most of the Criminal damages rises after sunset

Prostitution over a day



Most of the Criminal damages rises midnight

Weapon violation over a day



Most of the Weapon violation increases during midnight

Time Series Forecasting of crime Data

Persistence / Base model

Also called naive forecasr. This is where obsevation from the previous time step is used as the prediction for the observation at the next step

ARIMA

ARIMA is a forecasting technique that estimates the future values of a time series.Removes Residual auto correlation. Data shoul be stationary

SARIMA(Seasonal ARIMA)

.An extension to ARIMA that supports the direct modeling of the seasonal component of the series is called SARIMA.

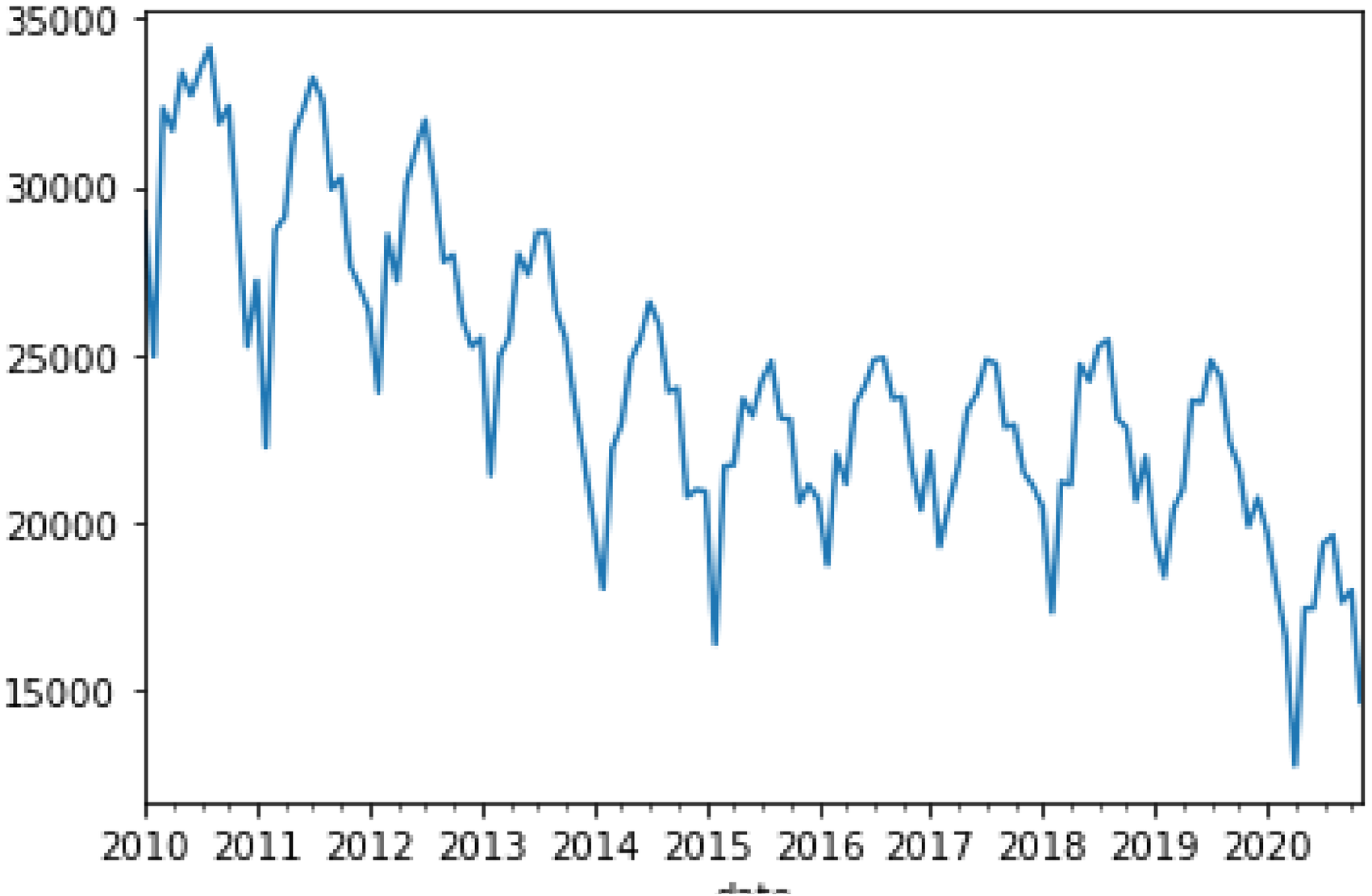
Holt-Winters Method

Forecast method based on smoothing. It is more suitable with data that contains high seasonality and trend

Holt-Linear model

allow the forecasting of data with a trend.

Persistence / Base model



Persistence / Base model

```
>Predicted=22101.000, Expected=19288.000
>Predicted=19288.000, Expected=20549.000
>Predicted=20549.000, Expected=21679.000
>Predicted=21679.000, Expected=23365.000
>Predicted=23365.000, Expected=23843.000
>Predicted=23843.000, Expected=24848.000
>Predicted=24848.000, Expected=24724.000
>Predicted=24724.000, Expected=22833.000
>Predicted=22833.000, Expected=22905.000
>Predicted=22905.000, Expected=21474.000
>Predicted=21474.000, Expected=21084.000
>Predicted=21084.000, Expected=20460.000
>Predicted=20460.000, Expected=17328.000
>Predicted=17328.000, Expected=21210.000
>Predicted=21210.000, Expected=21127.000
>Predicted=21127.000, Expected=24707.000
>Predicted=24707.000, Expected=24211.000
>Predicted=24211.000, Expected=25244.000
```

The baseline prediction for time series forecasting is called the naive forecast or persistence

The observation from the previous time step is used as the prediction for the observation at the next step.

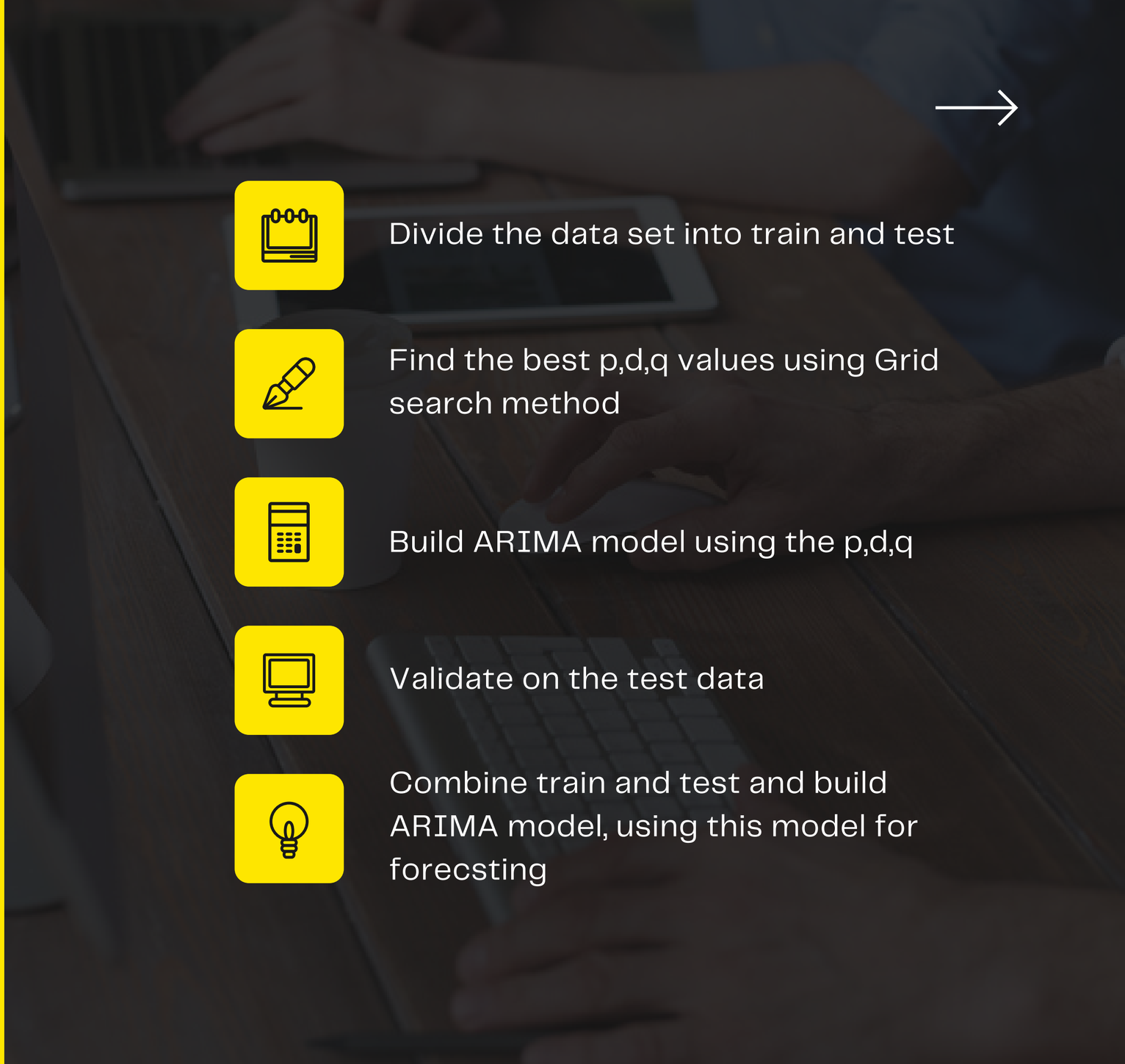
ARIMA MODEL

ARIMA means AutoRegressive Integrated Moving Average’

ARIMA is a Univariate Time Series Forecasting, means forecasting use only the previous values of the time series to predict its future values

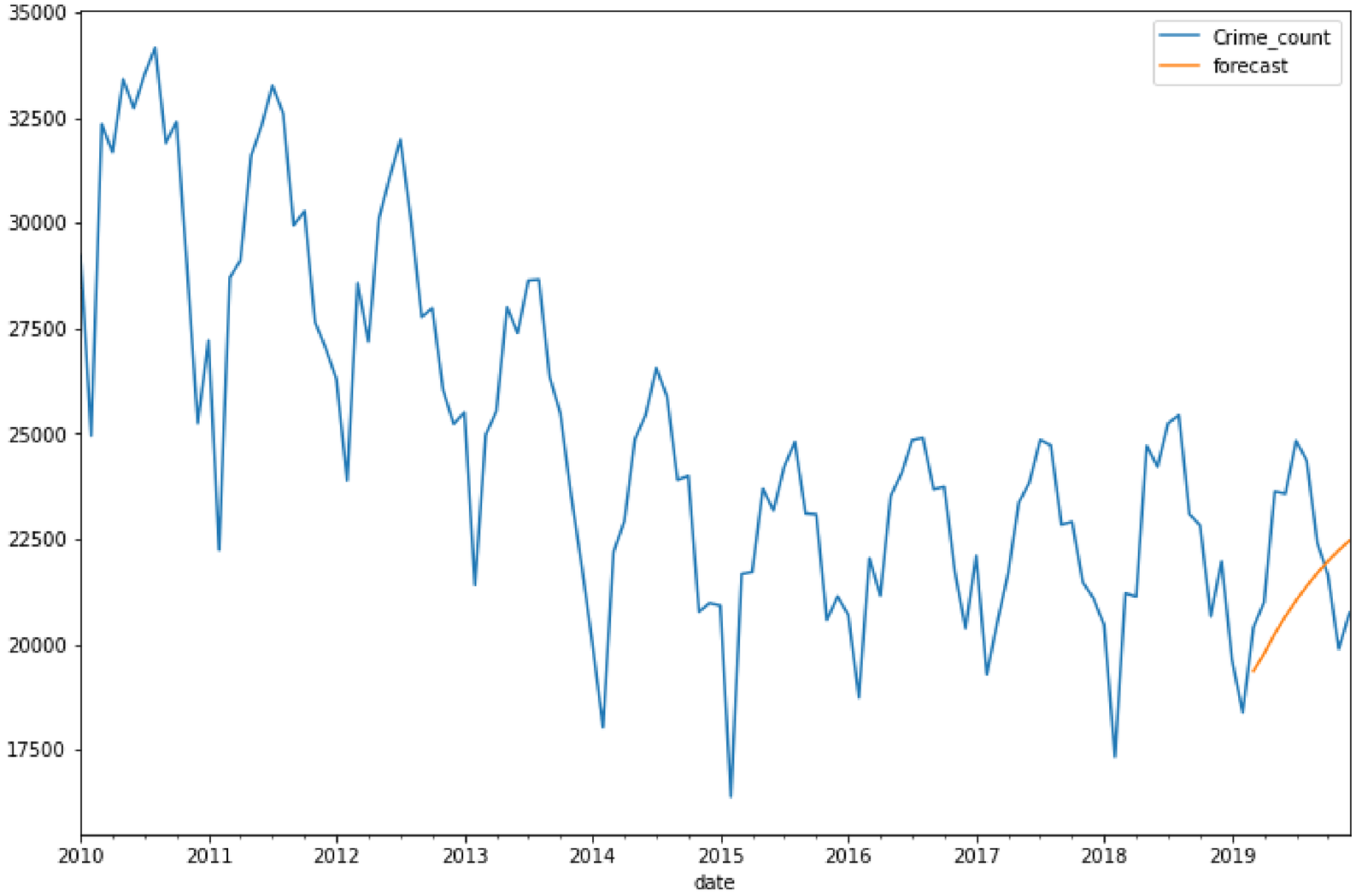
ARIMA contain three parts:
p:order of the AR term
d:number of differencing required to make the time series stationary
q: order of the MA term
:

Model building flow of ARIMA model



- Divide the data set into train and test
- Find the best p,d,q values using Grid search method
- Build ARIMA model using the p,d,q
- Validate on the test data
- Combine train and test and build ARIMA model, using this model for forecsting

Forecasting 12 months data



Limitations of ARIMA

- ARIMA does not support time series with a seasonal component.
- Time seires data should be seasonal
- ARIMA expects data that is either not seasonal or has the seasonal component removed

SARIMA MODEL

Seasonal Autoregressive Integrated Moving Average or Seasonal ARIMA

SARIMA contains two parameters: Trends and seasonal elements

TREND:

Three trend elements:
p: Trend autoregression order.
d: Trend difference order.
q: Trend moving average order.

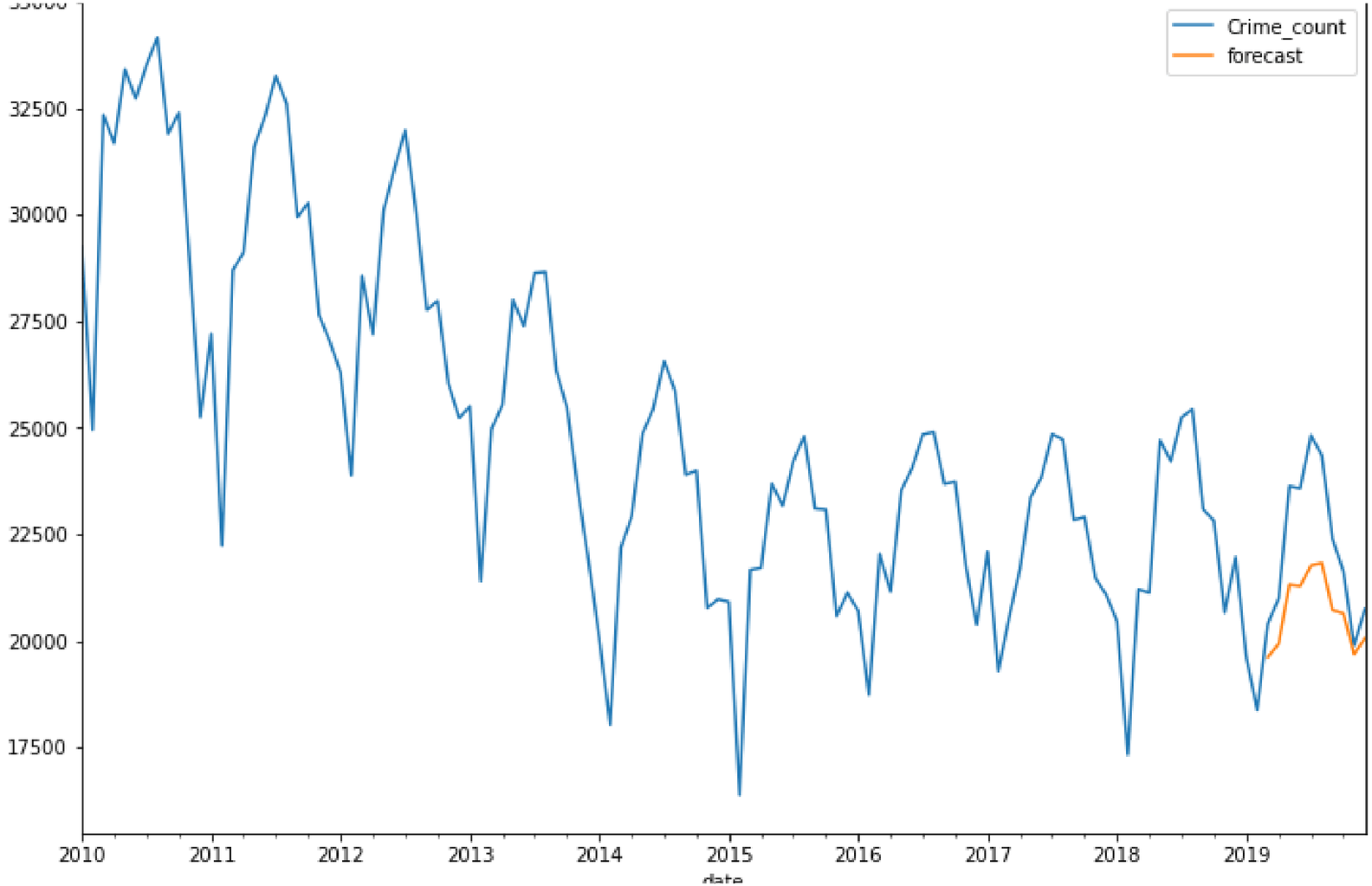
Seasonal Elements:

Four seasonal elements:
P: Seasonal autoregressive order
.D: Seasonal difference order.
Q: Seasonal moving average order.
m: The number of time steps for a single seasonal period.

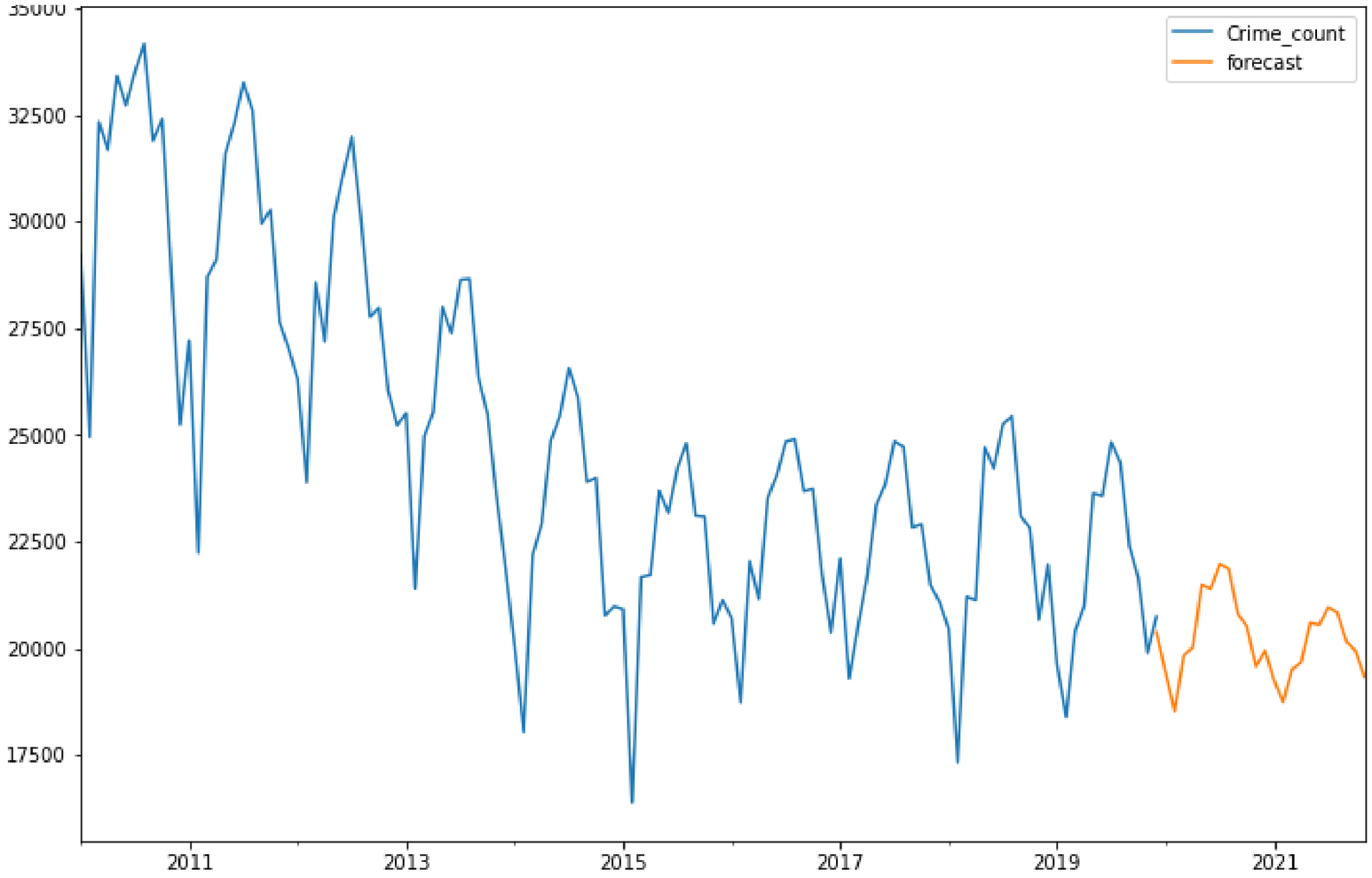
SARIMA syntax

SARIMA(p,d,q)(P,D,Q)m

Forecasting 12 months data using SARIMA



Forecasting 24 months data using SARIMA



Forecast method based on smoothing

Two forecasting methods based on forecasting:

Moving averages
Exponential smoothing.

Exponential smoothing

Holt-Linear methods
Holt-Winter methods

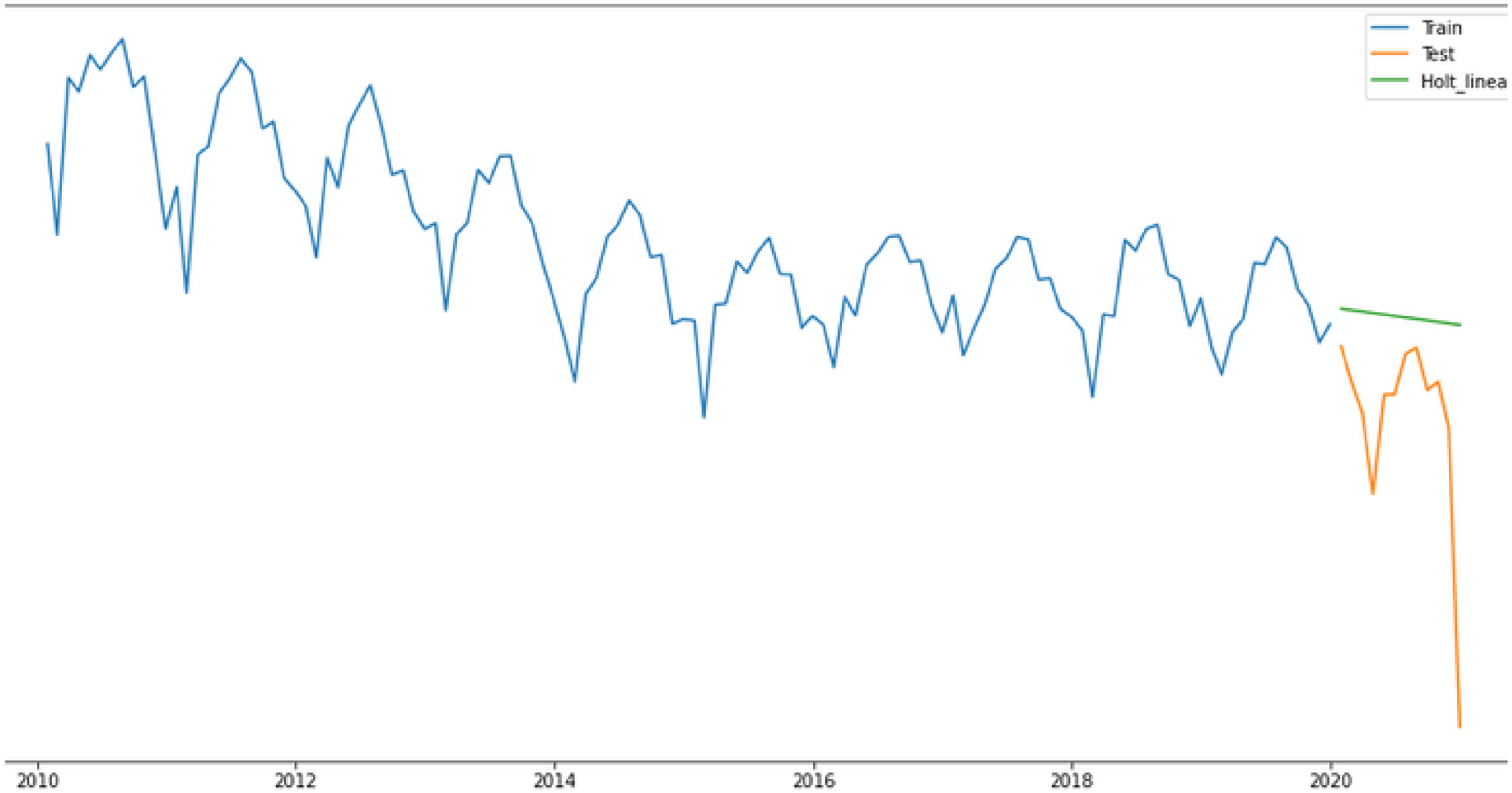
Holt-Linear method

Holt-Linear methods also called Holt extended simple exponential smoothing to allow the forecasting of data with a trend

Holt-Winter method

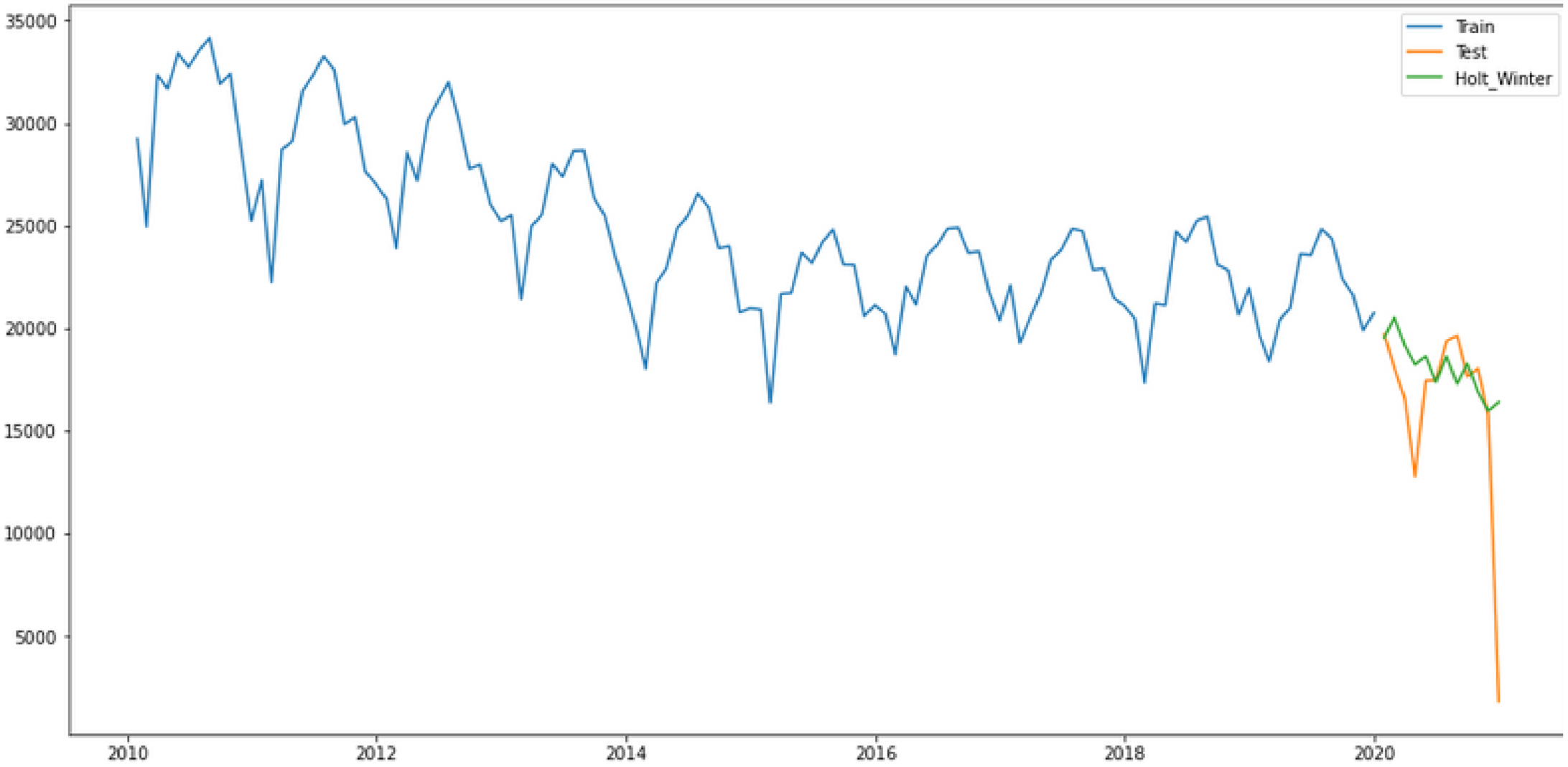
HHere using Holt-Winter's Additive Seasonal Method

Holt–Linear method




```
fit1 = Holt(Train['Crime_count']).fit(smoothing_level=0.3,smoothing_slope = 0.1)
y_hat_avg['Holt_linear'] = fit1.forecast(len(Test["Crime_count"]))
```

Holt–Winter method




```
fit2 = ExponentialSmoothing((Train['Crime_count']),seasonal_periods=7 ,trend='add', seasonal='add').fit()
y_hat_avg['Holt_Winter'] = fit2.forecast(len(Test["Crime_count"]))
```

Error Measurement

forecasts the crime rate with different methods such asARIMA,SARIMA Holt-Linear and Holt-Winter

One of the better ways to measure error for forecasting techniques is Mean Absolute Percentage Error (MAPE)

Here SARIMA has lowest MAPE value. So consider SARIMA as best forecasting model

	MODEL	MAPE_Values
0	ARIMA_model	0.089741
1	SARIMA_model	0.067450
2	HOLT_LINEAR_model,	0.751642
3	HOLT_WINTER_model	0.751642

Final Observations

1. *Analysed 2010 to 2020 crime data of Chiccago city in USA.*
2. *In 2010 to 2020 , there is a downword trend , that means crime rate is decreasing*
3. *Theft and battery was most occuring crimes in Chiccago city*
4. *Most of the Crime occured 2010 to 2013, after 2013 we can see that crime rate was decreasing*
5. *Crime rate is very less at the time of Winter*

Final Observations

1. *Most of the visualization states that summers are dangerous in Chiccago, because crime rate rise during summer*
2. *Friday night from 8 pm to 12 pm have high crime rate*
3. *Most of the crime ,arrest rate was very less*
4. *76% of cases, marked no arrest*
5. *Most of the crime occured in street, residence and apartment*

Final Observations

1. *Battery crime rate is very high in apartment and residence*
2. *Domestic violence increases after 2017*
3. *According to forecasting results for the year 2021 Crimes are decreasing, with around 20,000 crimes per month*
4. *In 201 monthly crime rate was around 35,000*
5.