Crime Analytics

Team 4

Rijuta Wagh

Siddhi Kulkarni

Akshay Gavandi

Project Idea and Motivation

1

Residential safety places nearby the university

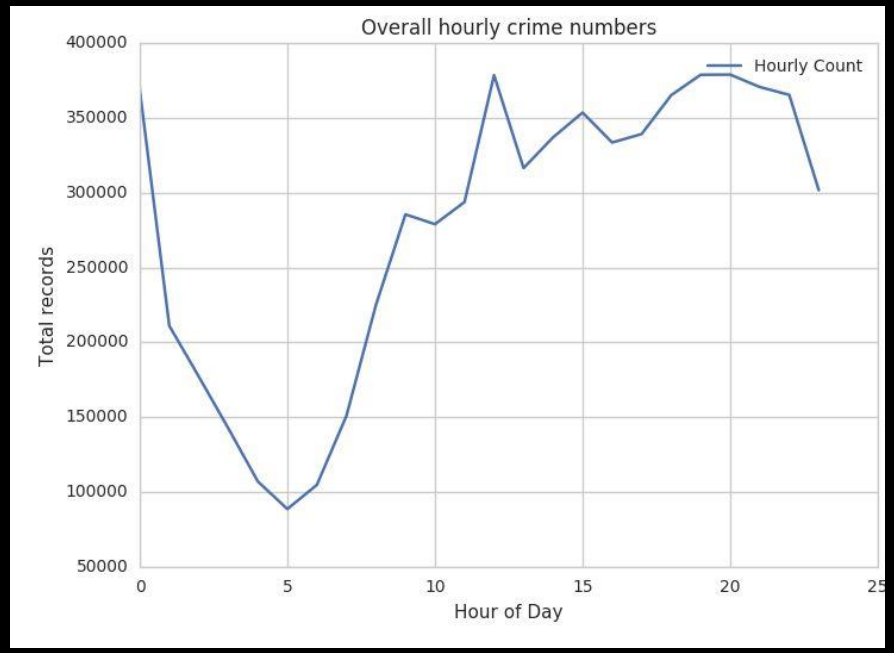
2

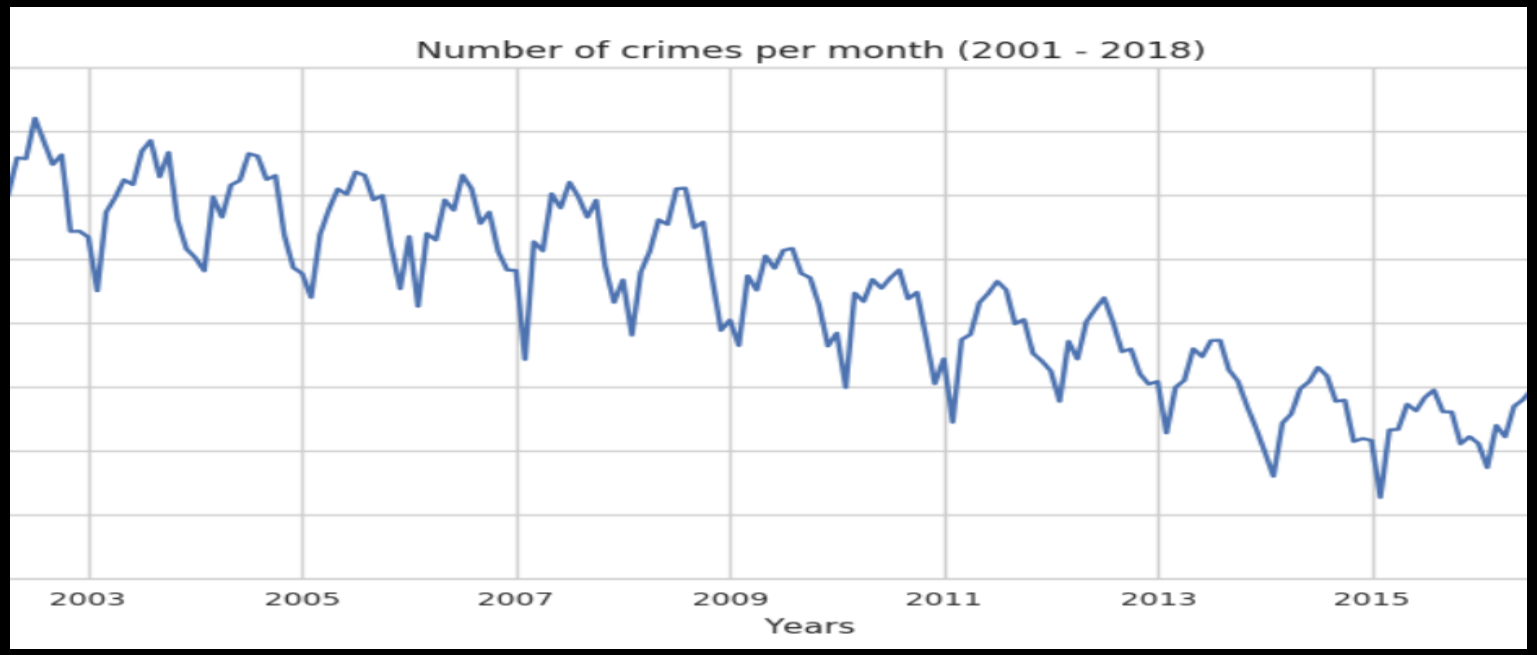
Understand and analyze the criminal activities in city area

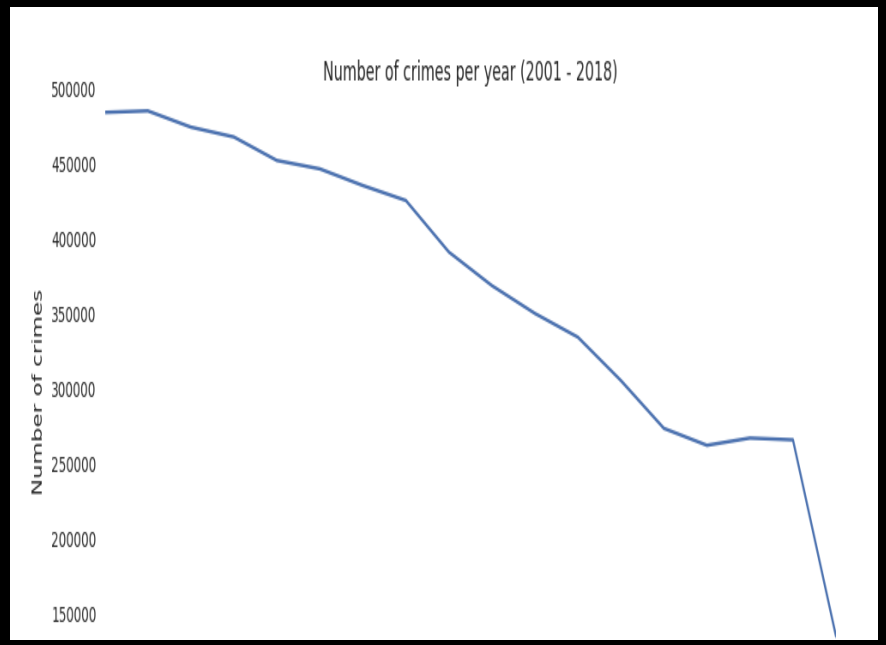
Data Source and Analytics Platform

- Crime incidences in city of Chicago from 2001 to present
- https://catalog.data.gov/dataset/crimes-2001-to-present-398a4
  - Dataset size: 1.6GB
  - 6.65 million reported crimes
- Databricks
  - Faster processing
  - Handles big data

Exploratory Data Analysis







Yearly - Decreasing crime rate

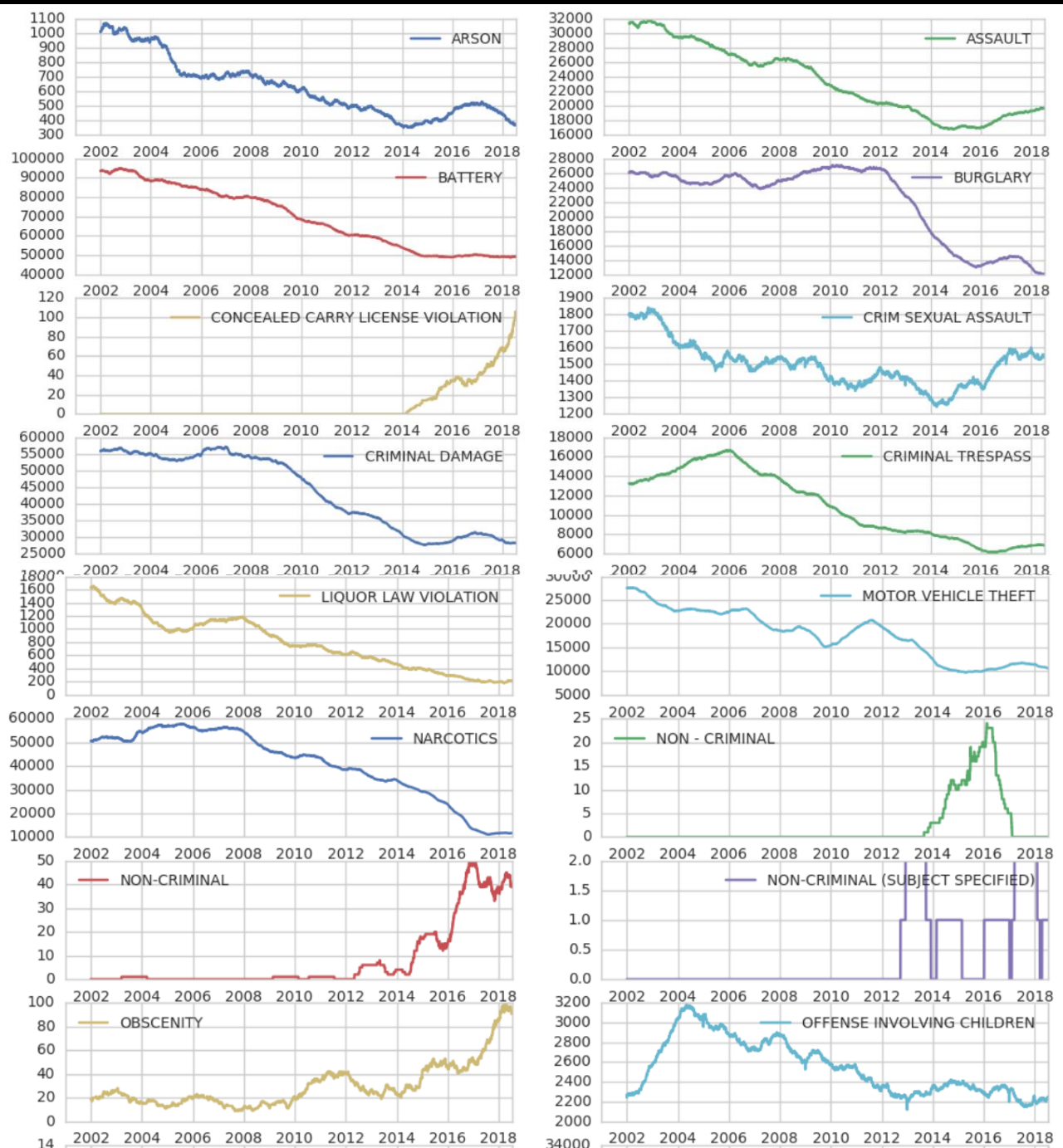
Monthly - Periodic and decreasing rate

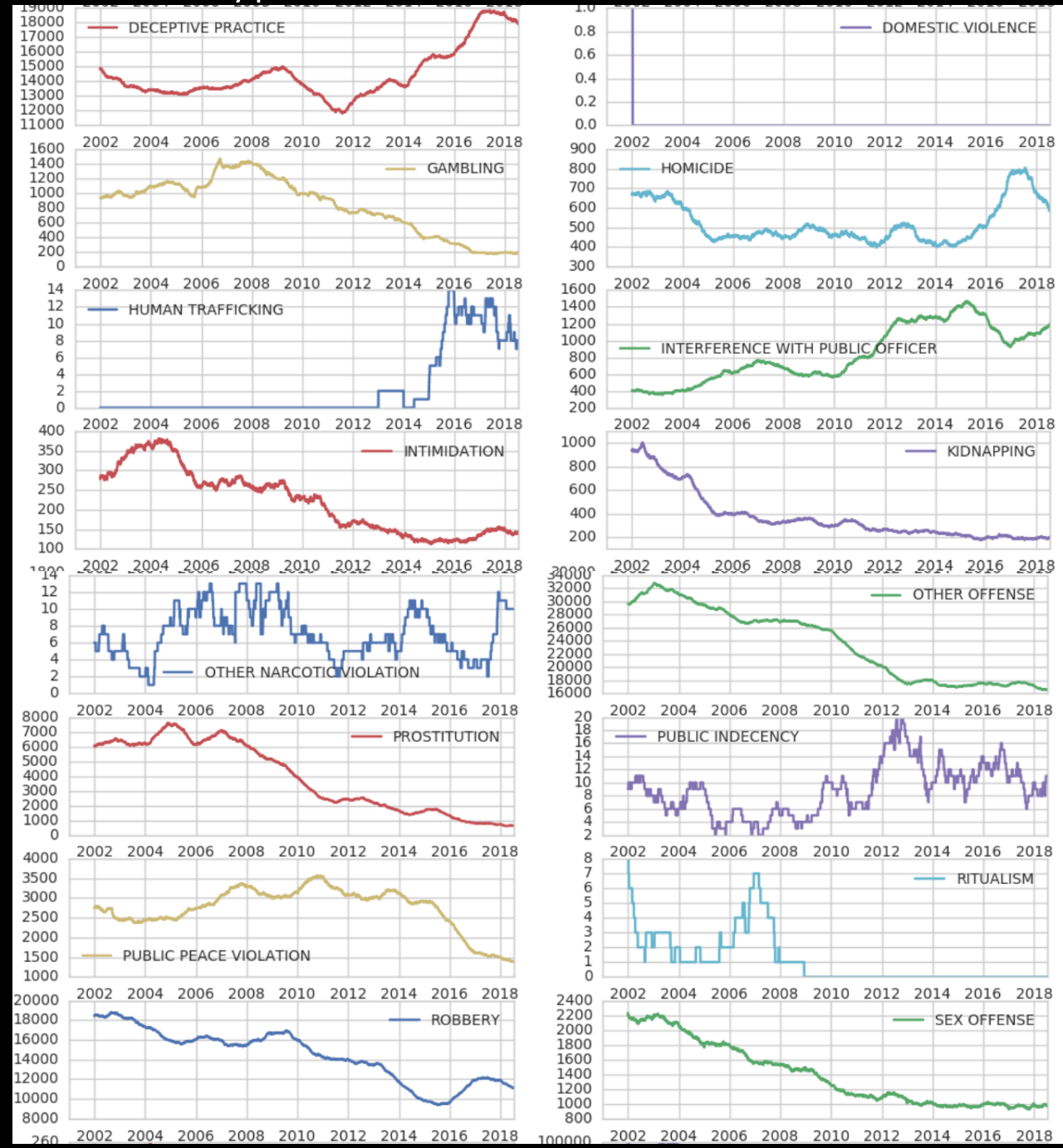
Hourly – Maximum crime rates at 12 noon and midnight.

Minimum crime rate at 5am

Crime Occurrence

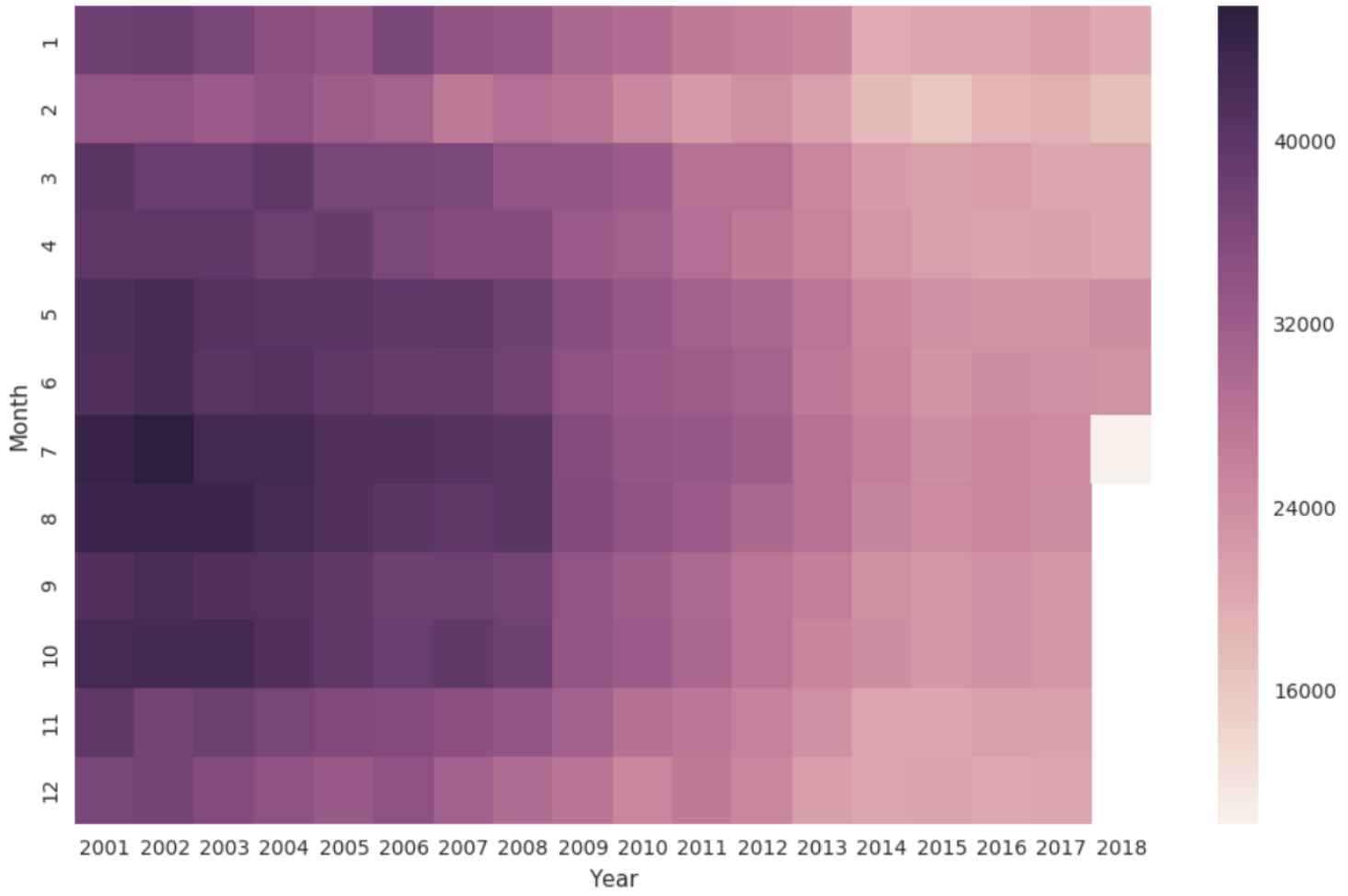
Trends for Various Crime Types

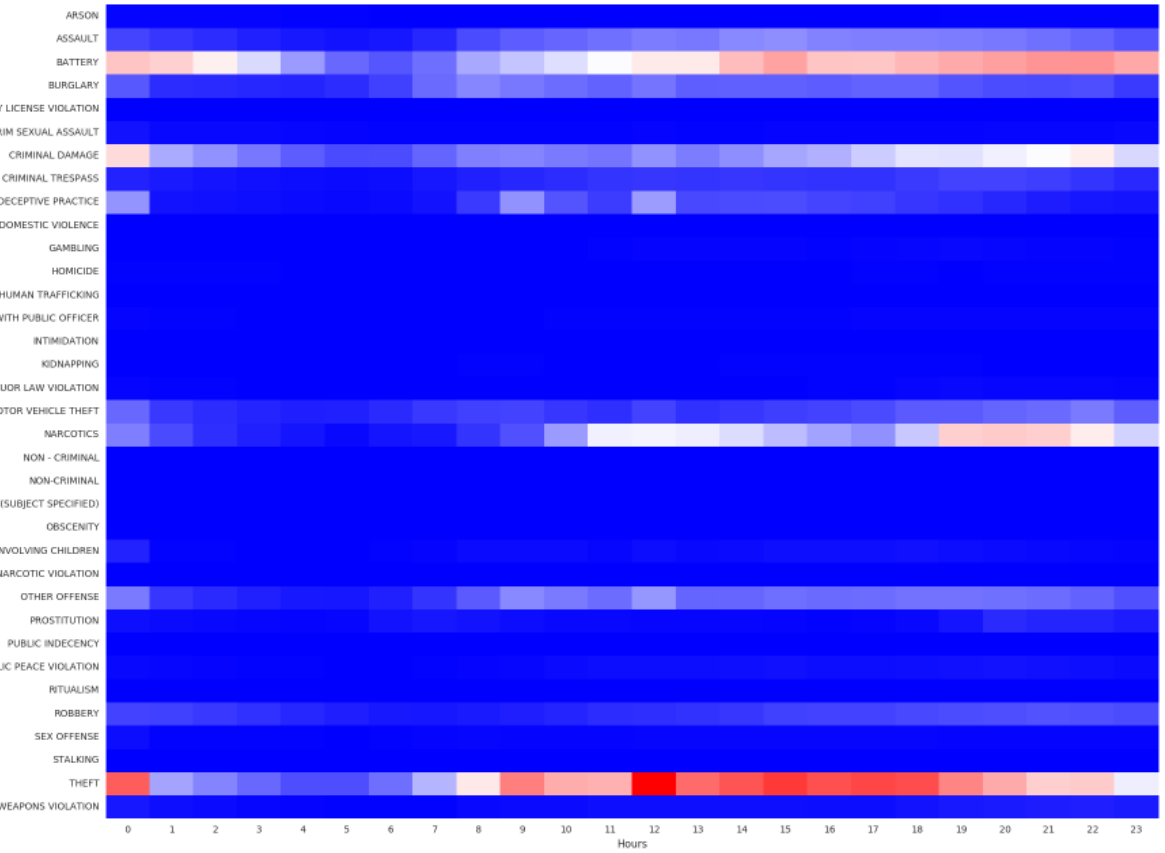


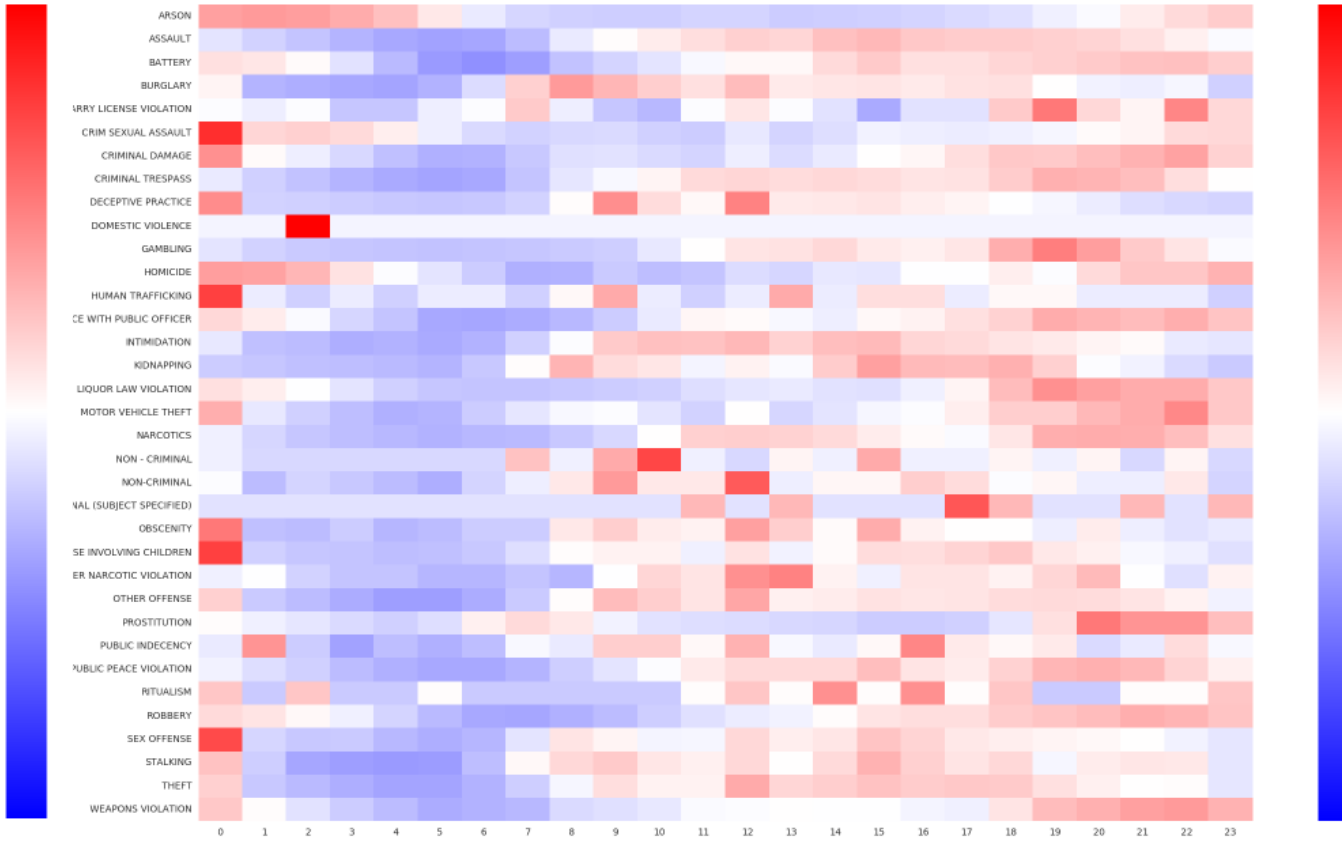


Heatmap Monthly Crime count

- Gradual decrease in crime rates
- Most crimes occur between May and October
- February has the lowest crime count



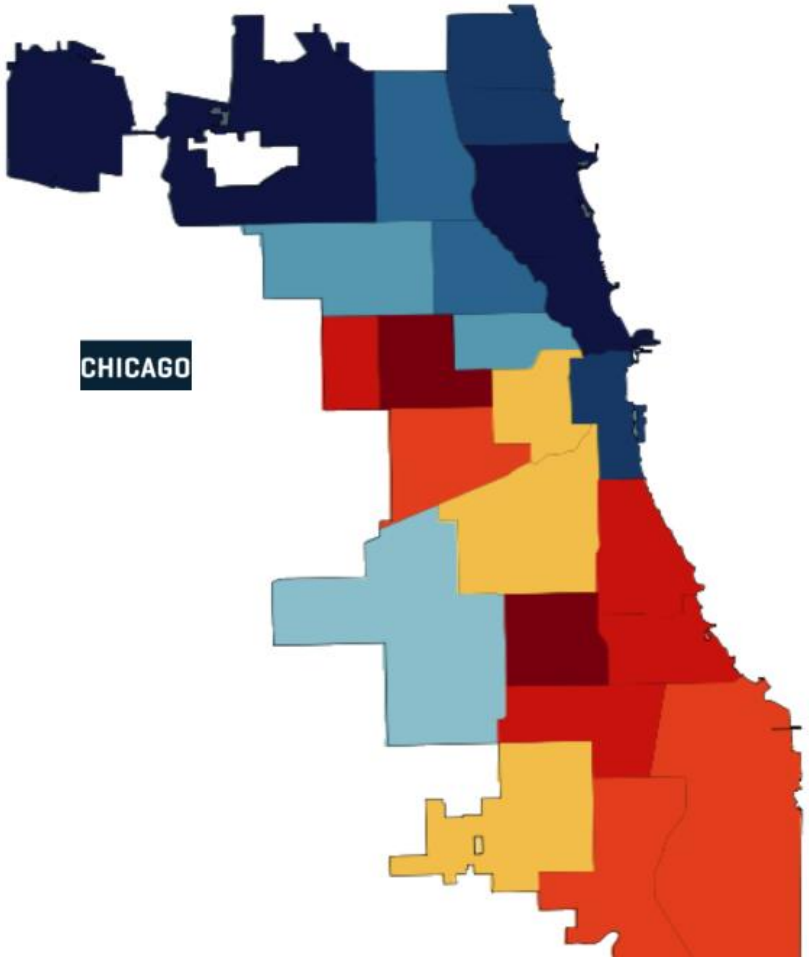


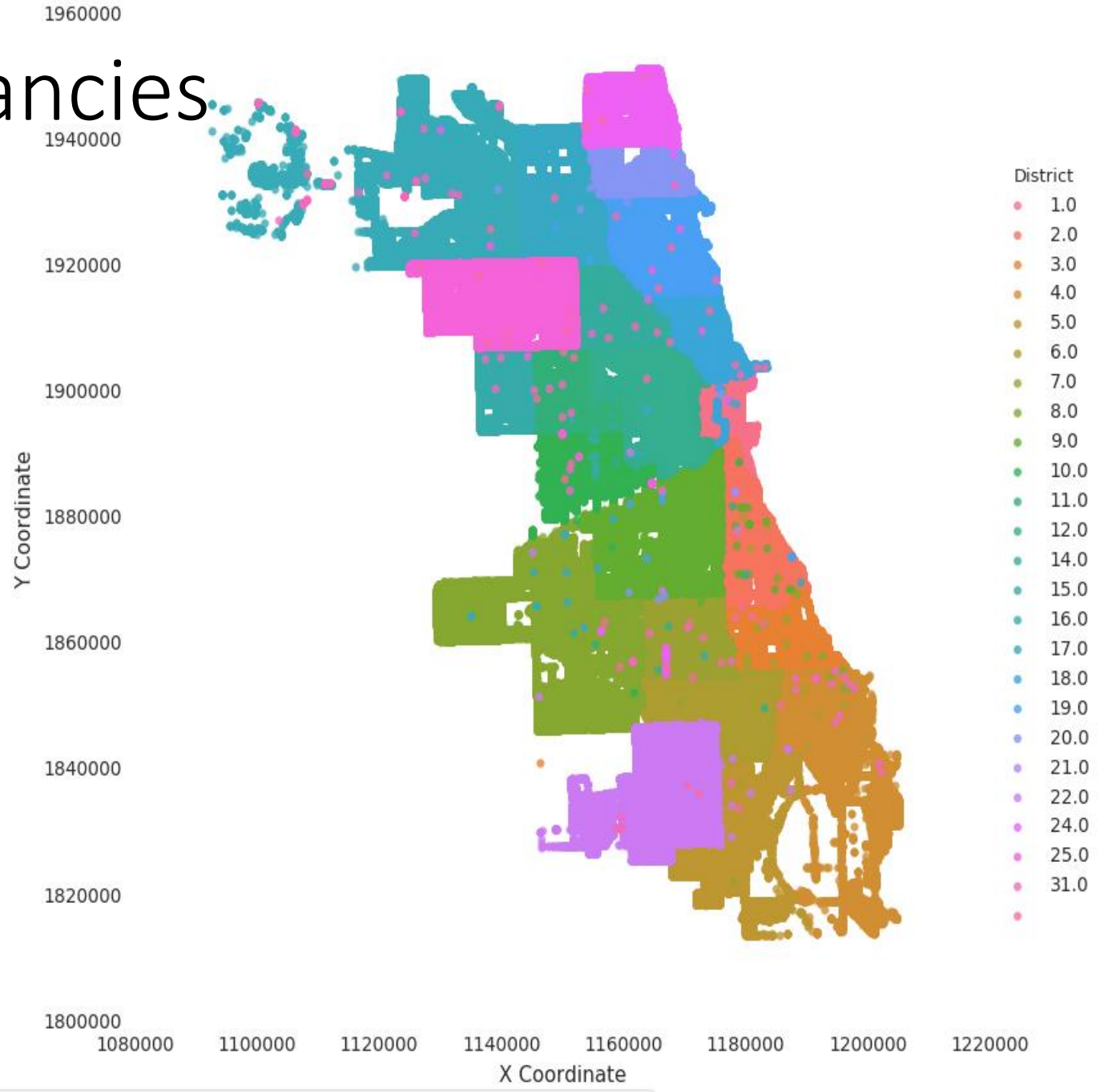


Hourly Crime Type

Before and After Renormalization

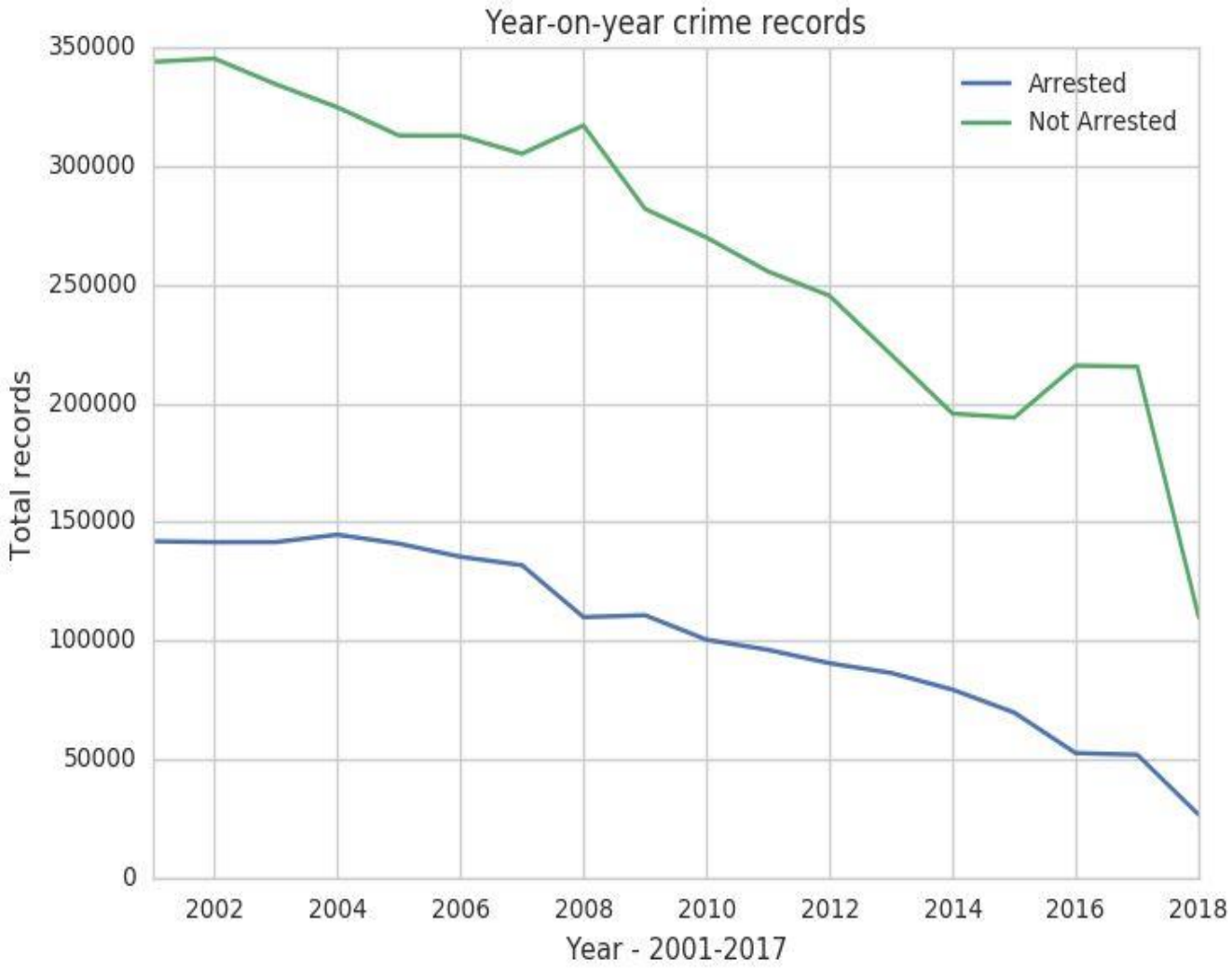
Finding Data Discrepancies





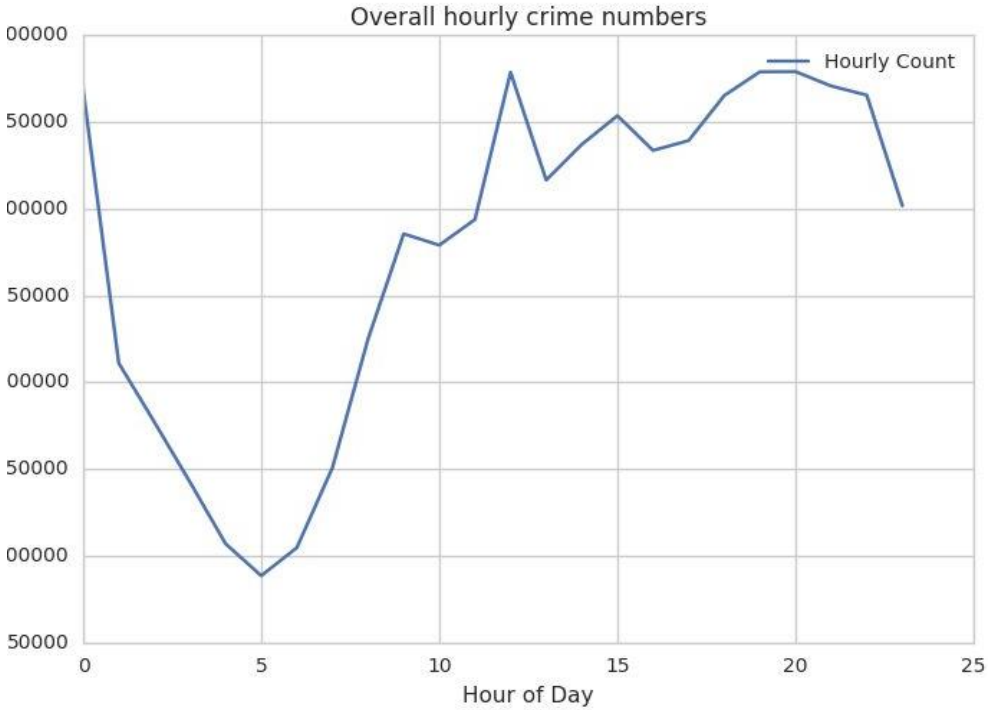
Evolution of Crime

- Decrease in crime rate
- Increase in percent arrest
- Gap between crime rate and arrest rate is reducing meaning more criminals are arrested for the crime committed



Peak Time for Criminal Activities

- 12AM, 12PM and 8PM
- Least criminal activities at 5AM



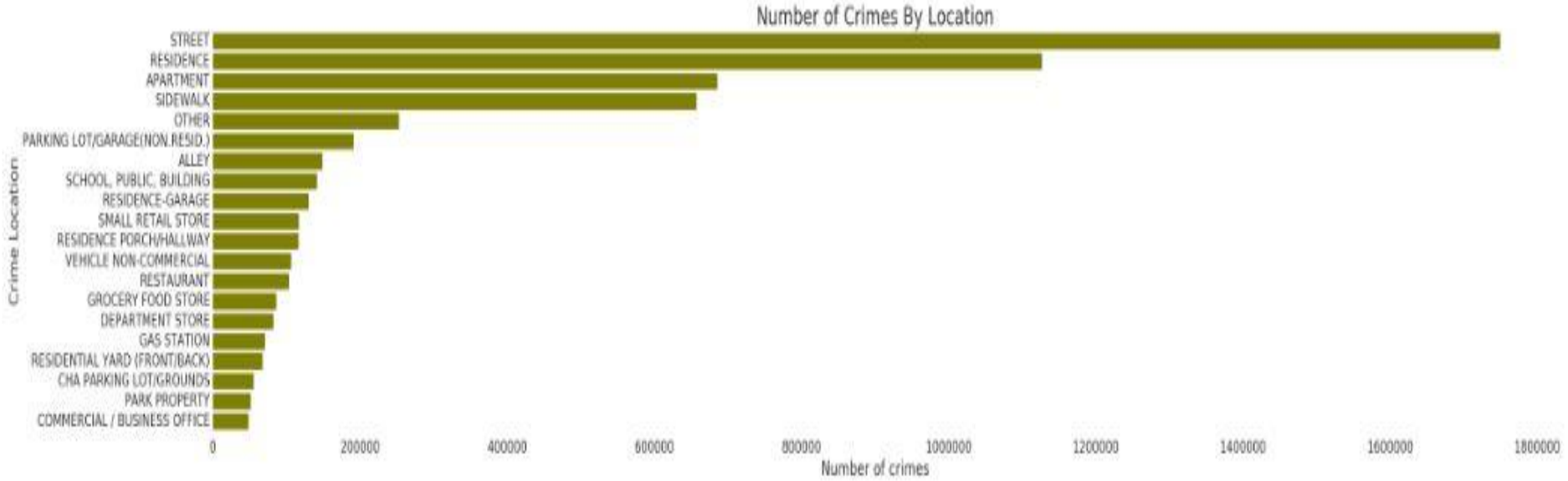
Top 10 unsafe Locations

```
1 crimes.groupBy(['location_description']).count().orderBy('count', ascending=False).show(10)
```

▸ (1) Spark Jobs

location_description	count
STREET	1749370
RESIDENCE	1126950
APARTMENT	685713
SIDEWALK	657105
OTHER	252611
PARKING LOT/GARAG...	191109
ALLEY	148926
SCHOOL, PUBLIC, B...	141360
RESIDENCE-GARAGE	130152
SMALL RETAIL STORE	116787

only showing top 10 rows



Worst Days of the Month

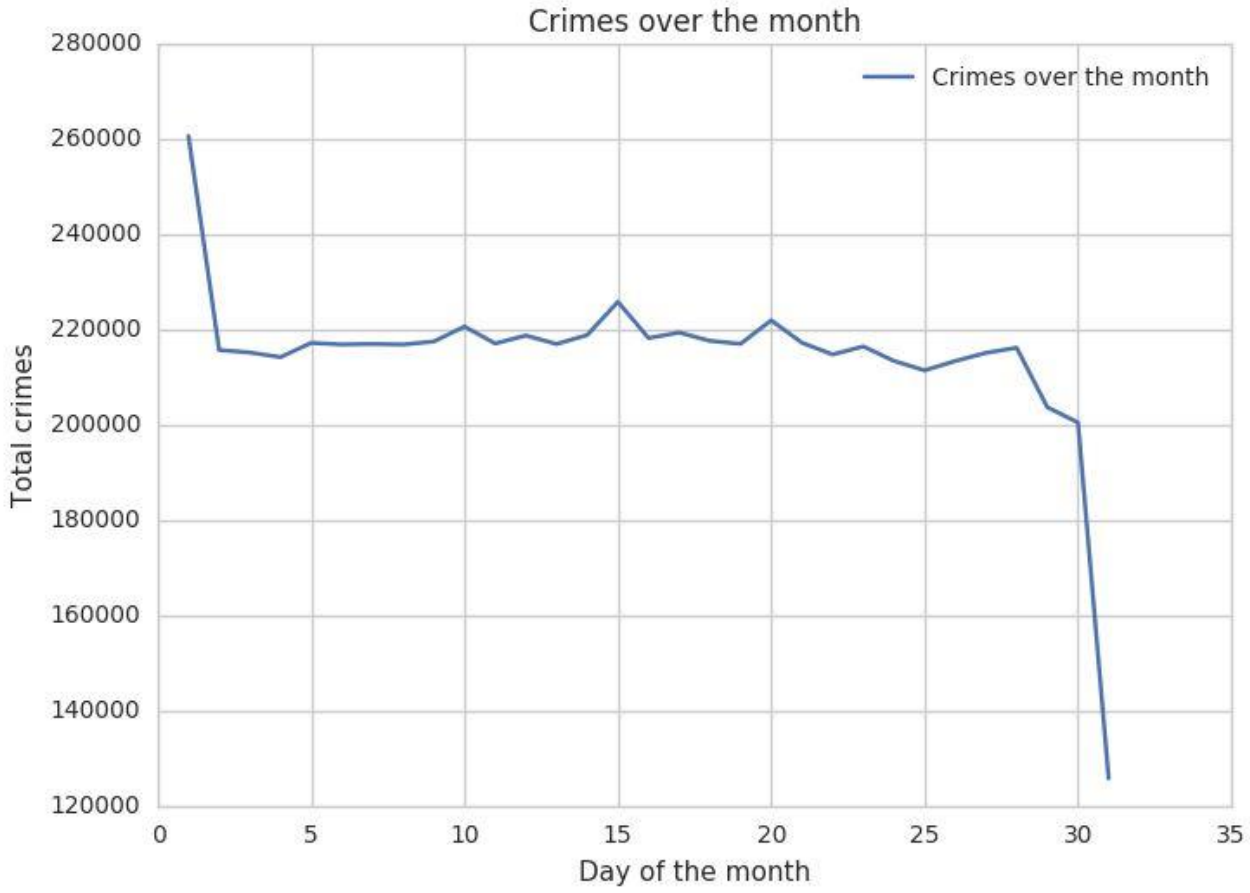
- Highest crime rates are observed at start of the month (Pay Day)
- Theft being the most contributing crime

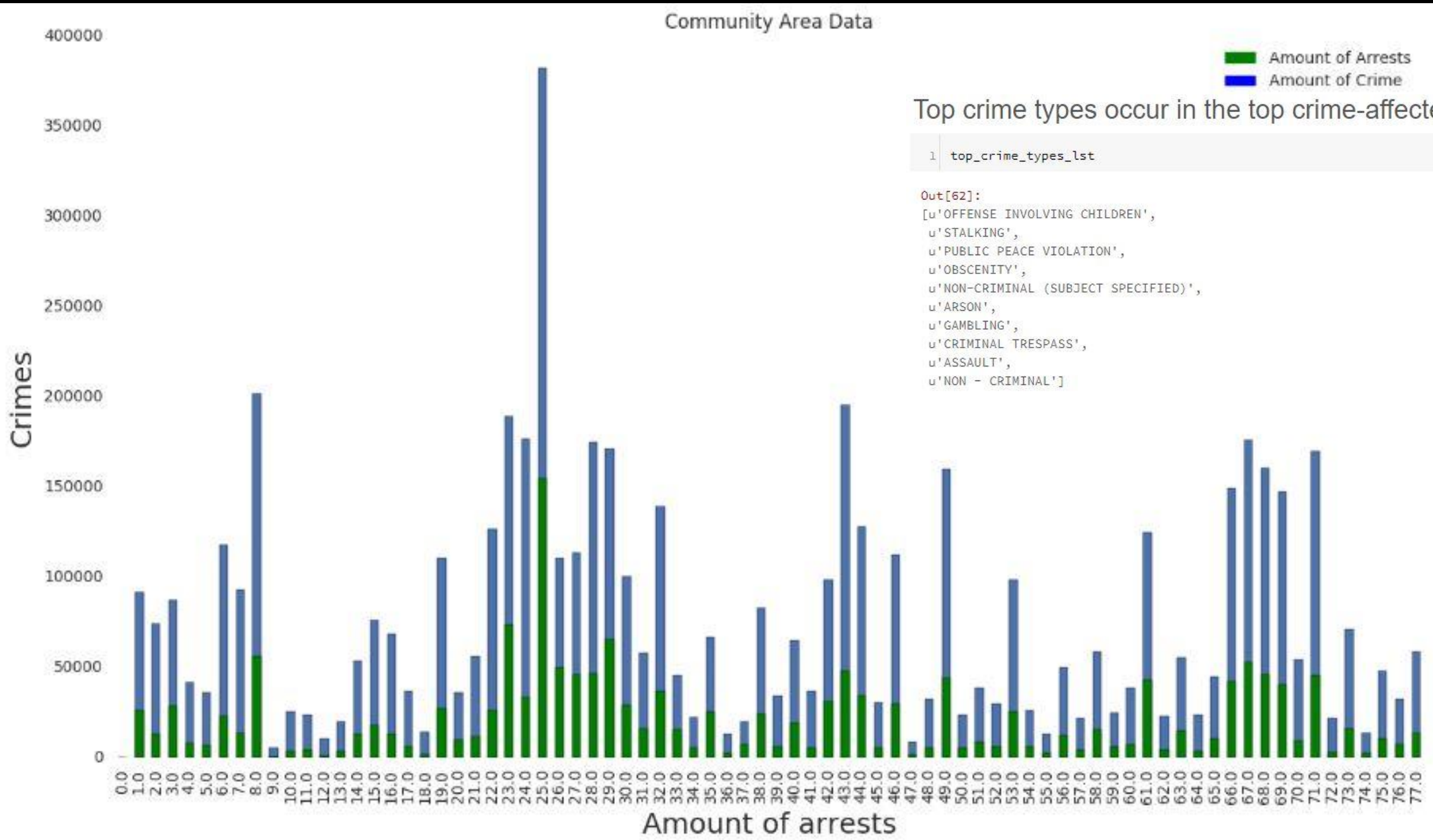
Top 10 worst days of the month

```
1  month_day_crime_counts_pddf.sort_values(by='count', ascending=False).head(10)
2
```

Out[170]:

	count	month_day
0	260658	1
14	225861	15
19	221969	20
9	220673	10
16	219414	17
13	218890	14
11	218791	12
15	218247	16
17	217672	18
8	217533	9





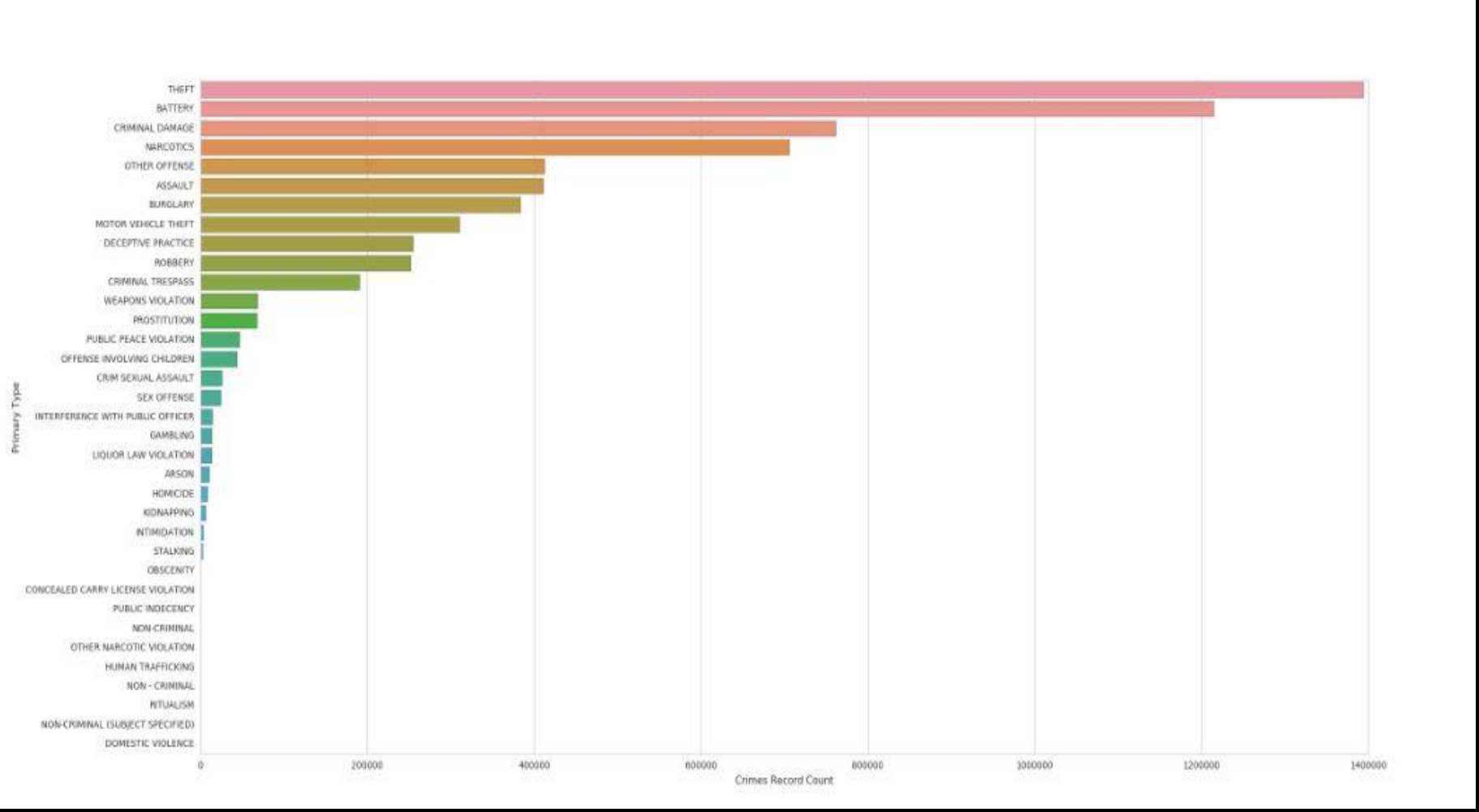
Top crime types occur in the top crime-affected areas

```
1 top_crime_types_lst
```

```
Out[62]:
[u'OFFENSE INVOLVING CHILDREN',
 u'STALKING',
 u'PUBLIC PEACE VIOLATION',
 u'OBSCENITY',
 u'NON-CRIMINAL (SUBJECT SPECIFIED)',
 u'ARSON',
 u'GAMBLING',
 u'CRIMINAL TRESPASS',
 u'ASSAULT',
 u'NON - CRIMINAL']
```

Feature Selection For Logistic Regression

- Location Description
- Arrest
- Domestic
- Beat
- District
- Ward
- Community Area
- Hour
- Week
- Day
- Year
- Month
- Date



Model Accuracy = 69.37%

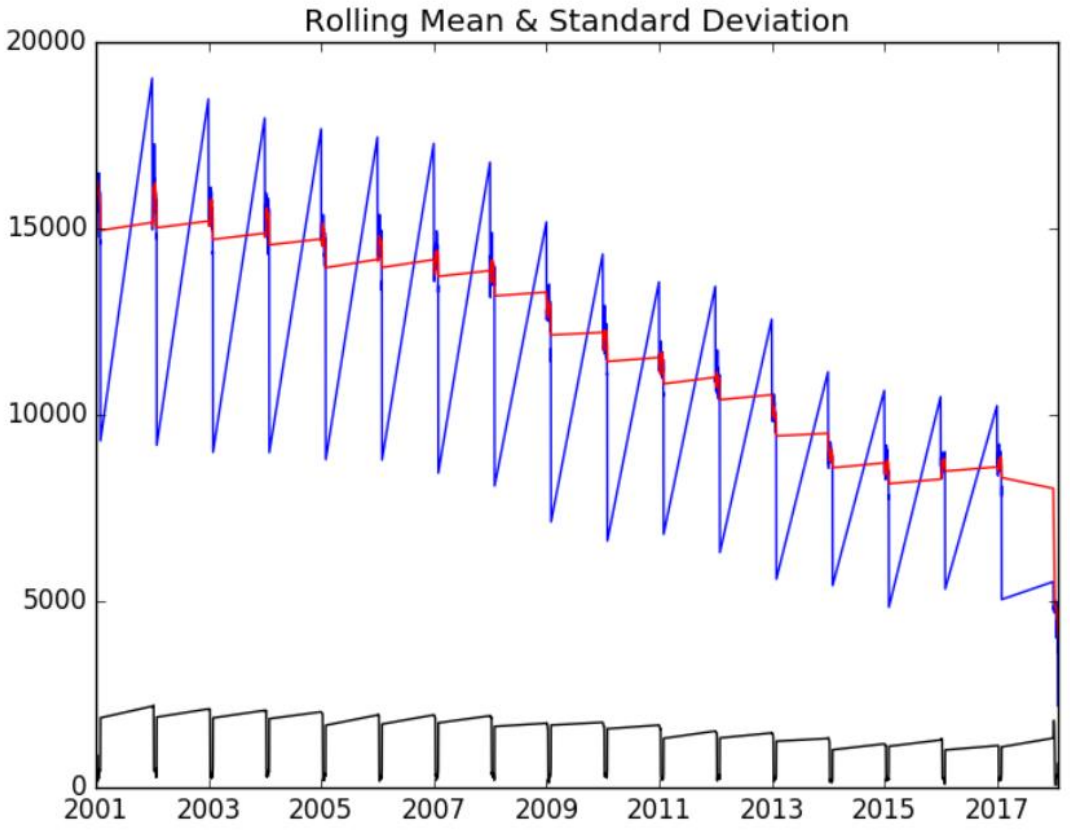
Primary Type	recall_rate	true_positive_rate	f_measure	false_positive_rate	precision_rate
Theft	0.933	0.933	0.778	0.12	0.6674
Narcotis	0.837	0.837	0.68	0.073	0.573

Time Series Analysis ARIMA Modelling

- Average and Standard Deviation of crime count by date
- Stationarity Check

```
1 check_stationarity_data(x)
```

```
Results of Dickey-Fuller Test:
Test Statistic                    0.489843
p-value                           0.984561
#Lags Used                        6.000000
Number of Observations Used     551.000000
Critical Value (5%)              -2.866800
Critical Value (1%)              -3.442274
Critical Value (10%)             -2.569571
dtype: float64
```

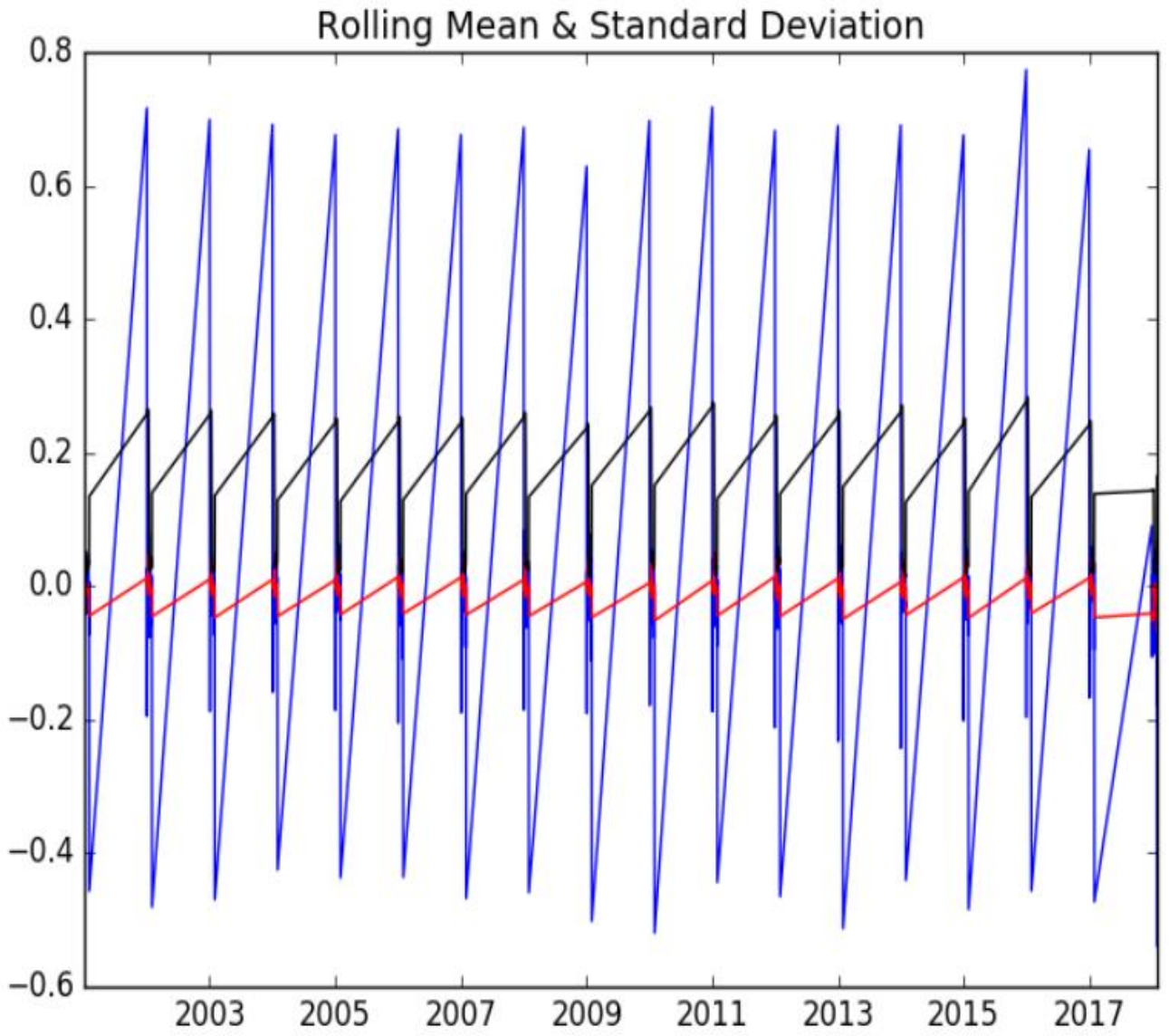


- Calculated difference of moving average and log value of crime count
- Calculated P values and statistics

```
1 check_stationarity_data(ts_log_ewma_diff)
```

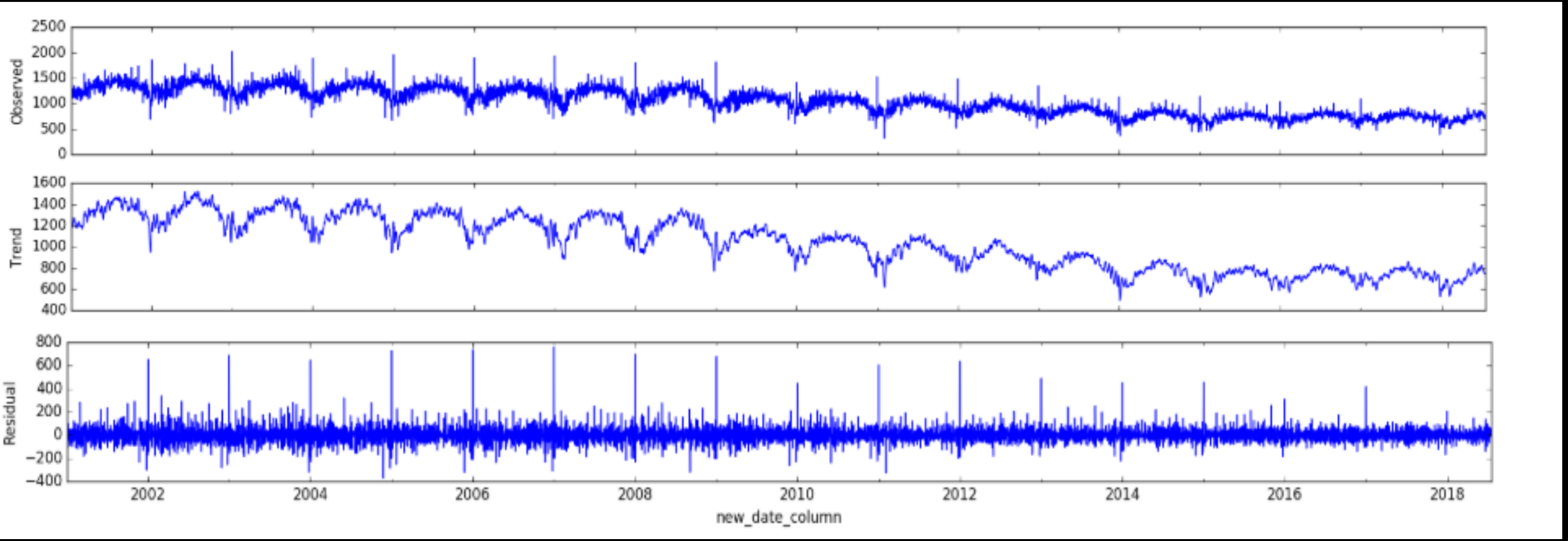
```
Results of Dickey-Fuller Test:
Test Statistic                  -2.797224
p-value                          0.058671
#Lags Used                       5.000000
Number of Observations Used    552.000000
Critical Value (5%)             -2.866790
Critical Value (1%)             -3.442252
Critical Value (10%)            -2.569566
dtype: float64
```

```
1 ts_log_diff.dropna(inplace=True)
2 check_stationarity_graph(ts_log_diff)
```

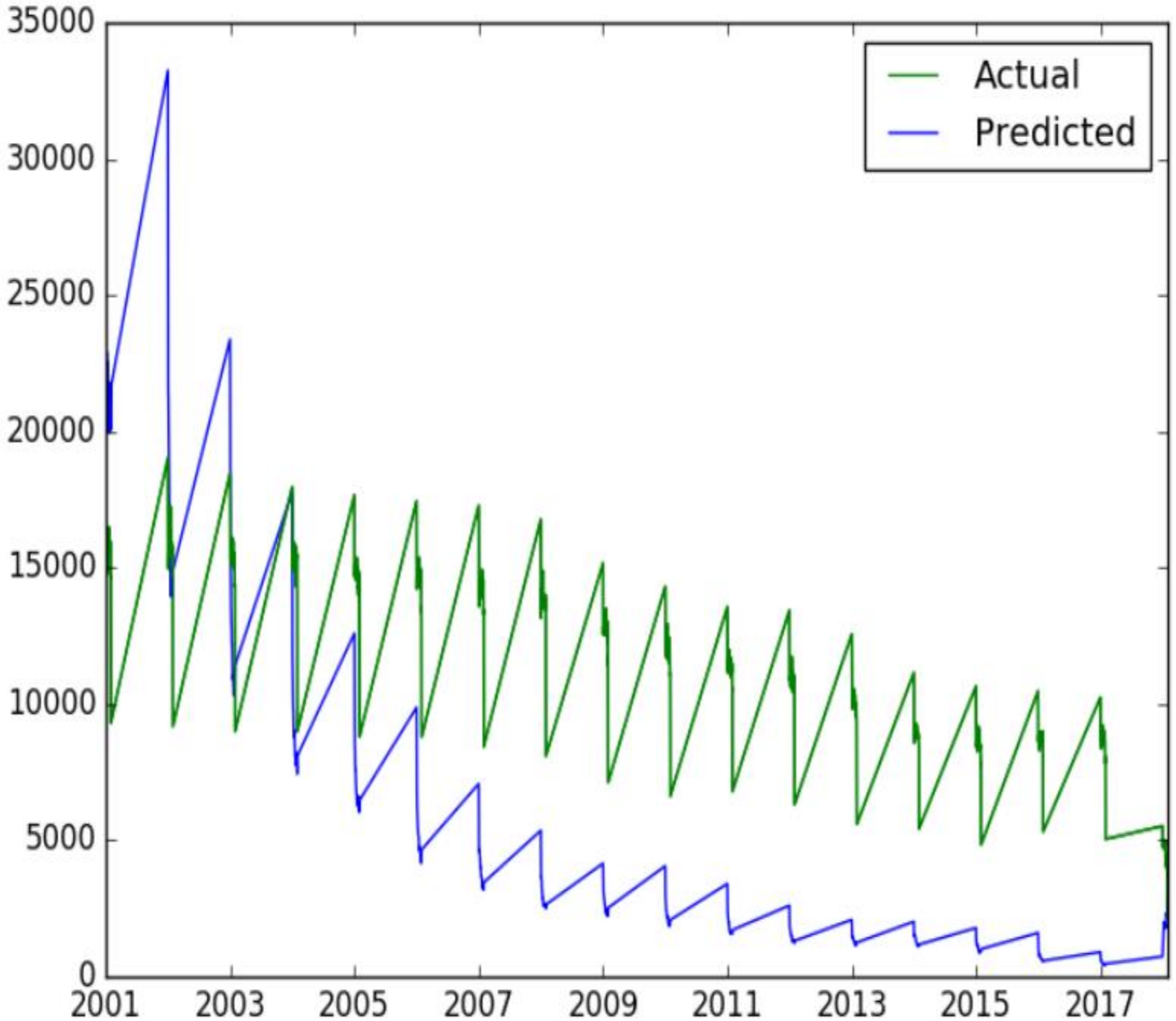


Time Series Decomposition

Trend – Shows increasing and decreasing crime values in the time series
Residual – Accounts for random variations in series

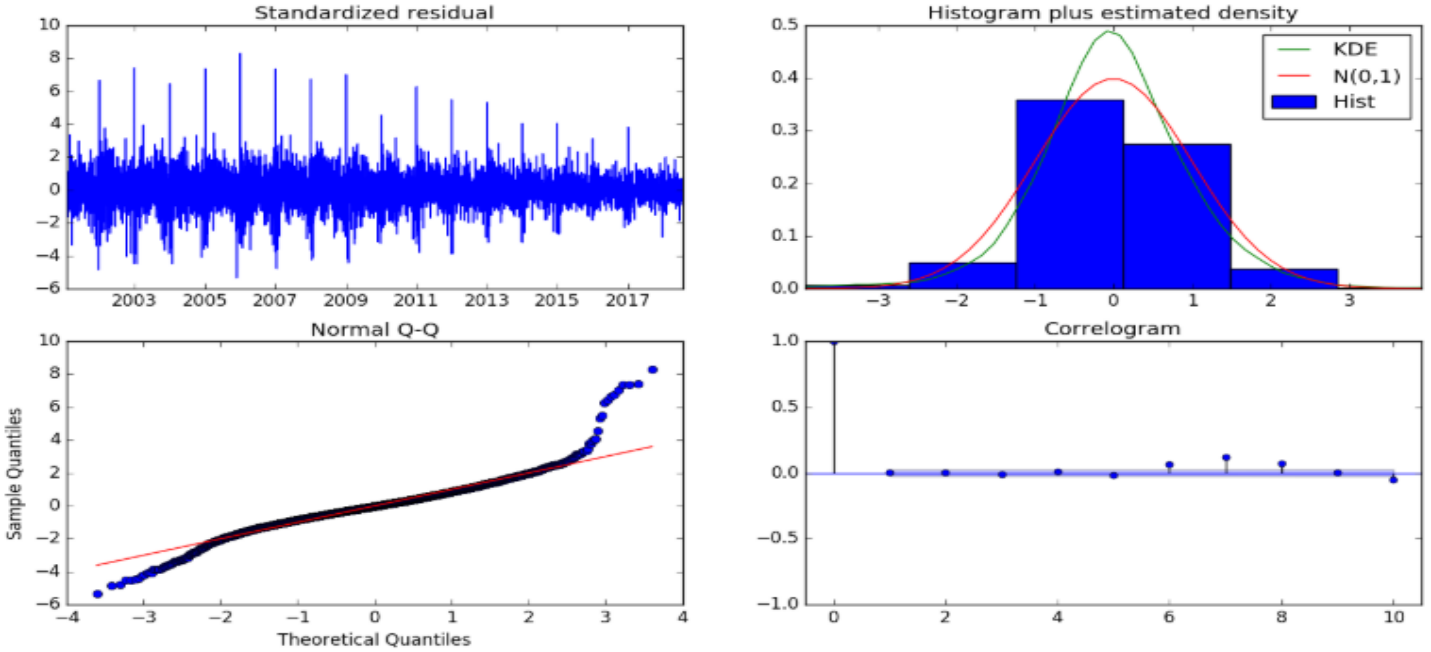


Poor ARIMA Model performance



Seasonal ARIMA Model (SARIMAX)

- Improving the crime prediction using SARIMAX by choosing the best parameter combination
- SARIMAX(1,1,1)x(1,1,1,12) yields the lowest AIC value 75783.3

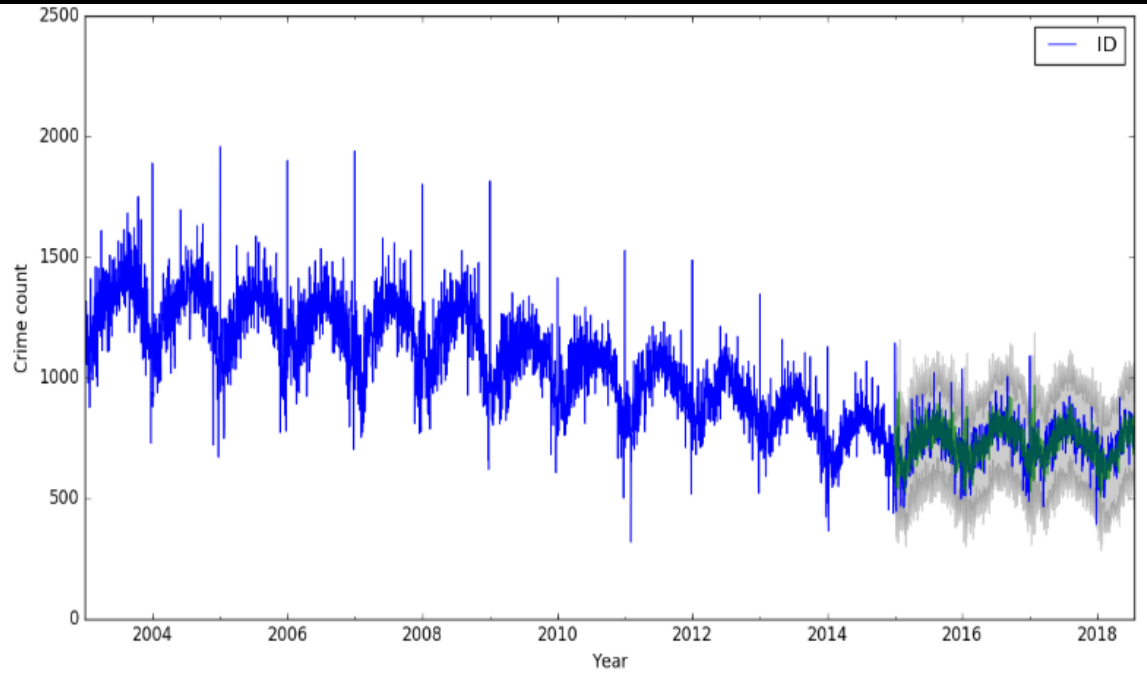


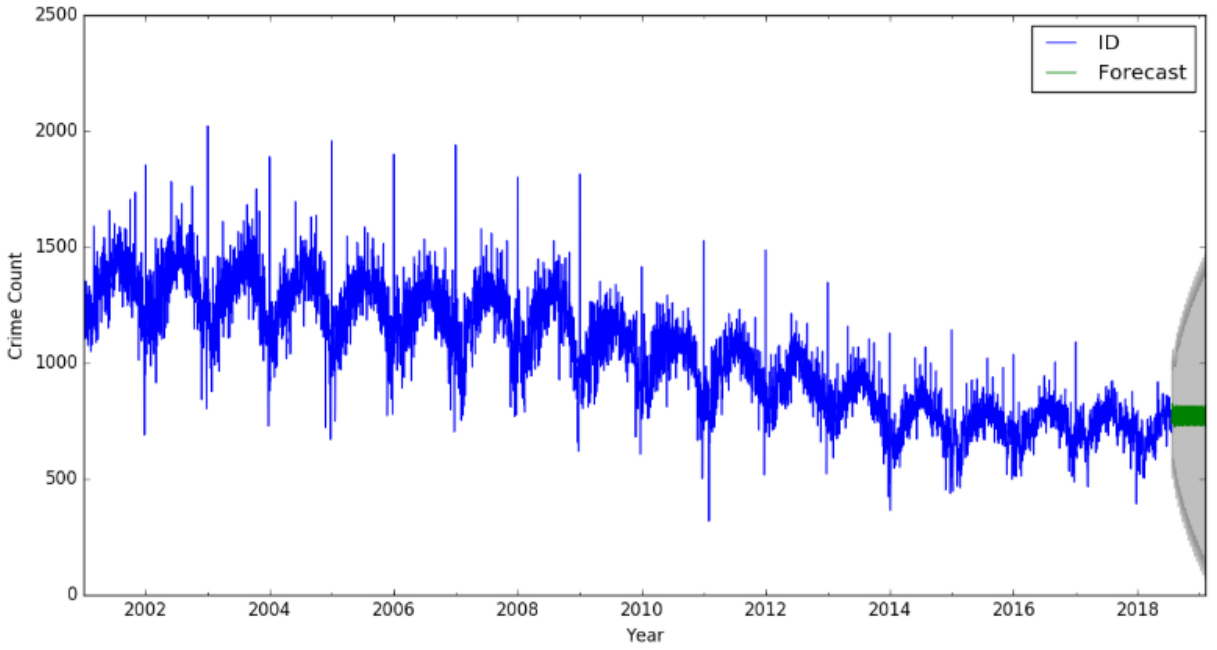

```
mod = sm.tsa.statespace.SARIMAX(crime_per_date,
                                order=(1, 1, 1),
                                seasonal_order=(1, 1, 1, 12),
                                enforce_stationarity=False,
                                enforce_invertibility=False)
results = mod.fit()
print(results.summary().tables[1])
```

```
/databricks/python/local/lib/python2.7/site-packages/statsmodels/tsa/base/tsa_model.py:171:
  % freq, ValueWarning)
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2375      0.009     25.115      0.000       0.219       0.256
ma.L1         -0.9046      0.005   -165.217      0.000      -0.915      -0.894
ar.S.L12      -0.0698      0.013     -5.561      0.000      -0.094      -0.045
ma.S.L12      -1.0000      1.367     -0.732      0.464      -3.679       1.679
sigma2      8302.0806   1.14e+04      0.731      0.465   -1.39e+04    3.06e+04
==============================================================================
```

Producing and Visualizing Crime Forecasts

- Compared the observed crime count to the forecasted crime count
- Forecast start at 2015-01-01 to present
- Captured seasonality towards the end of the year





References

- https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
- https://www.kaggle.com/currie32/crimes-in-chicago
- https://datascienceplus.com/spark-dataframes-exploring-chicago-crimes/
- https://datascienceplus.com/spark-dataframes-exploring-chicago-crimes/
- https://github.com/ajitkoduri/Chicago-Crime-Analysis/blob/master/Chicago%20Sex%20Crimes%20Analysis.ipynb
- https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b

Conclusion

- Exploratory Analysis on Criminal Activities in Chicago City
- Found data discrepancies and corrected them
- Most occurring crimes by hour, month, year crime type and location
- Calculated Crime and Arrest Rates using Logistic Regression with 69% accuracy score
- Time Series Forecasting using ARIMA and improving model accuracy Seasonal ARIMA model to predict future crime rate
- Expand the analysis for any demographic region

```
if questions:
    try:
        answer()
    except RuntimeError:
        pass
else:
    print('Thank You.')
```