Do certain crimes occur more during a certain month in Chicago?

By: Charlie Lee

Date: 3/23/2020

TABLE OF CONTENTS

ABSTRACT

This report is not affiliated with or endorsed by any organization. The findings and conclusions are for informational purposes only. The main purpose of this report is to analyze and determine if certain crimes occur more during a certain month or not. This report scrutinizes crime dataset from the city of Chicago. Although the dataset contains enormous amounts of data from 2001 to current, only the data from 2016 to 2018 were used. Crime types reported in this document uses the same wording as defined by the city of Chicago. The process of wrangling and cleaning the data are described in this report. Exploratory data analyses are presented by the visualizations. Applying appropriate statistical inferences to the cleaned data are introduced. Machine learning such as PCA (principal component analysis), K-nearest neighbors (KNN), and random forest were fit with the best tuned parameters to predict the problem.

1. Introduction

1.1 Problem Statement

Crime data are used by law enforcement in ways where it provides predictions for resource allocation, budget formulation, planning, and other various purposes. The crime data benefits politicians, researchers, criminal justice professionals to comprehend crime and society. Also, chambers of commerce and tourism agencies review crime data to see how it impacts the particular geographic jurisdiction they serve at. The crime data are notorious by justice professionals to learn about nature, cause, and movement of crime over time.

With all that said, crime data are used in security and police work in an attempt to reduce criminal activities. Law enforcement can provide safer communities if they can foresee what type of specific crime are likely to occur at a certain month. Such crime data can bring us one step closer to prevent crime rather than reacting to them. This report will present analysis of crimes to see if certain crimes occur more during a certain month or not.

1.2 Client - Northern District of Illinois | Department of Justice

This report was prepared for the Northern District of Illinois - Department of Justice, who is mainly concerned about the crimes in the city of Chicago.

From this data report, the client will be able to:

1. Budget more accurately.
    a) Accurately budget sufficient money to the proper locations and programs to save money and provide safer communities.
    b) Help determine which programs in a certain community can receive criminal justice grant.
2. Efficient allocation of resources Improve resource allocation during certain time at a given location.
    a) Improve resource allocation during certain time at a given location.
    b) Determine when and where more or less police officers will be required.
3. Predictive policing and initiative assessment.

a) Help law enforcement to anticipate increased risk of a certain crime during specific months, therefore by being able to prevent from occurring.
b) Insight of whether a certain crime increases or decreases during specific months and implement changes if needed.

1.3 Dataset

Dataset was acquired from the city of Chicago data portal (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data). Data was extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. This dataset reflects reported incidents of crime (with the exception of murders) that occurred in the city of Chicago from 2001 to present, minus the most recent seven day. The dataset gets updated on a daily basis.

This report analyzed the dataset from years 2016, 2017, and 2018. 2019 was not included because per dataset, these crimes may be based upon preliminary information and may be changed at a later date based upon additional investigation. Since the dataset was acquired in early January 2020, dataset of 2019 was considered to be premature. Instead, 2019 dataset will be solely processed as a test set to see if the data shows trends as predicted (if applicable). 2019 test set will be acquired at a later date to reduce any false data provided by the preliminary information. Crime types will be categorized by the same way the city of Chicago reports in their dataset (called "Primary Type" in the dataset column). Each crime occurrence is recorded with the following information:

- *ID*: Unique identifier for the record.
- *Case Number*: The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
- *Date*: Date when the incident occurred. This is sometimes a best estimate.
- *Block*: The partially redacted address where the incident occurred, placing it on the same block as the actual address.
- *IUCR*: The Illinois Uniform Crime Reporting Code. This is directly linked to the Primary Type and Description.
- *Primary Type*: The primary description of the IUCR code.
- *Description*: The secondary description of the IUCR code, a subcategory of the primary description.

- *Location Description*: Description of the location where the incident happened.
- *Arrest*: Indicates whether an arrest was made.
- *Domestic*: Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
- *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
- *District*: Indicates the police district where the incident occurred.
- *Ward*: The ward (City Council district) where the incident occurred.
- *Community Area*: Indicates the community area where the incident occurred. Chicago has 77 community areas.
- *FBI Code*: Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
- *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.
- *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.
- *Year:* Year the incident occurred.
- *Updated On*: Date and time the record was last updated.
- *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.
- *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.
- *Location*: Combination of latitude and longitude.

2. Data wrangling and cleaning

2.1 Missing values

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...
0	10457545	HZ190268	2016-03-17 06:00:00	006XX W OHARE ST	1812	NARCOTICS	POSS: CANNABIS MORE THAN 30GMS	GOVERNMENT BUILDING/PROPERTY	True	False	...
1	10425678	HZ156460	2016-02-18 19:41:45	043XX W GLADYS AVE	2027	NARCOTICS	POSS: CRACK	RESIDENCE	True	False	...
2	10538622	HZ283084	2016-05-27 15:00:00	023XX N CLARK ST	0620	BURGLARY	UNLAWFUL ENTRY	ATHLETIC CLUB	False	False	...
3	11914609	JC538182	2016-01-20 00:01:00	039XX W 66TH ST	1562	SEX OFFENSE	AGG CRIMINAL SEXUAL ABUSE	RESIDENCE	True	True	...
4	10883231	JA193135	2016-05-23 00:00:00	016XX S MORGAN ST	1751	OFFENSE INVOLVING CHILDREN	CRIM SEX ABUSE BY FAM MEMBER	RESIDENCE	False	True	...

Fig 1. First 5 rows of the 2016, 2017, and 2018 raw dataset.

Unnecessary columns such as: *Arrest*, *Domestic*, *X Coordinate*, *Y Coordinate*, *Updated On*, and *Location* were dropped. Upon observing the data types for the columns (Figure 2), it was observed that *District*, *Ward*, and *Community Area* were in floats rather than integer which were not correct.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 805540 entries, 0 to 267790
Data columns (total 16 columns):
ID                    805540 non-null int64
Case Number           805540 non-null object
Date                  805540 non-null datetime64[ns]
Block                 805540 non-null object
IUCR                  805540 non-null object
Primary Type          805540 non-null object
Description           805540 non-null object
Location Description  802267 non-null object
Beat                  805540 non-null int64
District              805539 non-null float64
Ward                  805535 non-null float64
Community Area        805538 non-null float64
FBI Code              805540 non-null object
Year                  805540 non-null int64
Latitude              795572 non-null float64
Longitude             795572 non-null float64
dtypes: datetime64[ns](1), float64(5), int64(3), object(7)
```

Fig 2. Data types of the columns before data wrangling and cleaning

Missing values observed in the *Location Description*, *Latitude* and *Longitude* were dropped since the rest of the analyses used *District* as the location variable. Prior to converting the data types, missing values had to be taken care of. There were: *District* (1), *Ward* (5), and *Community Area* (2) missing values. For each of these missing values, its record was searched within the dataset to look for other records with the same conditions (same *District*, *Ward*, and/or

Community Area). After searching under those conditions, the missing values were then replaced by the most frequent or the only value. The data type of the columns was changed to integer afterwards. Also, it was observed that three of the *Community Area* had value of 0 which was not correct (there is no community area 0). They were replaced using the same method as the missing values.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 805540 entries, 0 to 267790
Data columns (total 16 columns):
ID                    805540 non-null int64
Case Number           805540 non-null object
Date                  805540 non-null datetime64[ns]
Block                 805540 non-null object
IUCR                  805540 non-null object
Primary Type          805540 non-null object
Description           805540 non-null object
Location Description  802267 non-null object
Beat                  805540 non-null int64
District              805540 non-null int64
Ward                  805535 non-null float64
Community Area        805540 non-null int64
FBI Code              805540 non-null object
Year                  805540 non-null int64
Latitude              795572 non-null float64
Longitude             795572 non-null float64
dtypes: datetime64[ns](1), float64(3), int64(5), object(7)
```

Fig 3. Data types of columns after data wrangling and cleaning

2.2 Crime types

Upon observing all of the different crime types (Figure 4), three were dropped and one was combined to the other type. 'Non-criminal', 'non-criminal (subject specified)', and 'non – criminal' types were dropped due to crimes being specified as non-criminal. 'Other narcotic violation' was changed to 'Narcotics' to generalize the crime and to remove the low count sample. Afterwards, there were 29 unique crime types remaining.

THEFT	191228	SEX OFFENSE	3159
BATTERY	149324	PROSTITUTION	2253
CRIMINAL DAMAGE	87883	HOMICIDE	2054
ASSAULT	58444	ARSON	1333
DECEPTIVE PRACTICE	57707	LIQUOR LAW VIOLATION	686
OTHER OFFENSE	51776	GAMBLING	581
BURGLARY	39030	STALKING	574
NARCOTICS	38419	KIDNAPPING	563
ROBBERY	33520	INTIMIDATION	455
MOTOR VEHICLE THEFT	32662	CONCEALED CARRY LICENSE VIOLATION	254
CRIMINAL TRESPASS	20028	OBSCENITY	223
WEAPONS VIOLATION	13593	NON-CRIMINAL	122
OFFENSE INVOLVING CHILDREN	6908	PUBLIC INDECENCY	34
CRIM SEXUAL ASSAULT	4862	HUMAN TRAFFICKING	33
PUBLIC PEACE VIOLATION	4476	OTHER NARCOTIC VIOLATION	16
INTERFERENCE WITH PUBLIC OFFICER	3329	NON-CRIMINAL (SUBJECT SPECIFIED)	6
		NON - CRIMINAL	5

Fig 4. Total occurrences of all crimes from 2016, 2017, and 2018

A new dataset was generated to rearrange and count all 29 crimes by the year and the month when it occurred. Another dataset was generated to organize all 29 crimes by the year, month, hour, weekday, district, and community area occurred.

3. Initial observations

3.1 Crime occurrences by each year

Upon observation of the bar graphs (see Appendix A) for the crime occurrences of all types, separated by each year, initial observations were made:

- *Arson:* Decreased over the years.
- *Assault*: Increased over the years.
- *Battery*: No significant trend over the years.
- *Burglary*: Decreased every year.
- *Concealed carry license violation*: Increased over the years.
- *Crim sexual assault*: No significant trend over the years.
- *Criminal damage*: Decreased over the years.
- *Criminal trespass*: Increased over the years
- *Deceptive practice*: No significant trend over the years.
- *Gambling*: No significant trend over the years.
- *Homicide*: Decreased over the years.
- *Human trafficking*: No significant trend over the years.
- *Interference with public officer*: Increased over the years
- *Intimidation*: Increased over the years.
- *Kidnapping*: Decreased over the years.
- *Liquor law violation*: Decreased from 2016 to 2017, but increased in 2018.
- *Motor vehicle theft*: Increased from 2016 to 2017, but decreased in 2018.
- *Narcotics*: Decreased from 2016 to 2017, but increased back in 2018.
- *Obscenity*: Increased from 2016 to 2017, no different from 2017 to 2018.
- *Offense involving children*: No significant trend over the years.
- *Other offense*: No significant trend over the years.
- *Prostitution*: Decreased over the years.
- *Public indecency*: No significant trend over the years.
- *Public peace violation*: Decreased over the years.
- *Robbery*: Decreased significantly in the year of 2018 compared to the past two years.
- *Sex offense*: Increased over the years.
- *Stalking*: Increased over the years.
- *Theft*: Increased over the years.
- *Weapons violation*: Significantly increased over the years.

3.2 Crime occurrences by each month and year

From visualization of the time series plot (see Appendix B), following observations were made regarding the trend between the months for all of the crime types:

- *Arson*: No significant trend between the months.
- *Assault*: Lowest occurred in January, highest occurred in May.
- *Battery*: Lowest occurred in February, highest occurred between May and July.
- *Burglary*: No significant trend between the months.
- *Concealed carry license violation*: No significant trend between the months.
- *Crim sexual assault*: Highest occurred in July.
- *Criminal damage*: Lowest occurred in February, highest occurred in July.
- *Criminal trespass*: Highest occurred between May and July
- *Deceptive practice*: No significant trend between the months.

- *Gambling*: Increased from April to July (highest), then started to decrease
- *Homicide*: No significant trend between the months.
- *Human trafficking*: No significant trend between the months. Not enough data.
- *Interference with public officer*: No significant trend between the months.
- *Intimidation*: No significant trend between the months.
- *Kidnapping*: No significant trend between the months.
- *Liquor law violation*: No significant trend between the months.
- *Motor vehicle theft*: Lowest between February and April, then increased.
- *Narcotics*: No significant trend between the months.
- *Obscenity*: No significant trend between the months.
- *Offense involving children*: Highest occurred in January.
- *Other offense*: Highest occurred in May.
- *Prostitution*: Highest occurred in April.
- *Public indecency*: No significant trend over the years. Not enough data.
- *Public peace violation*: No significant trend between the months.
- *Robbery*: Lowest occurred in February, then started to continuously increase until August (highest).
- *Sex offense*: No significant trend between the months.
- *Stalking*: No significant trend between the months.
- *Theft*: Lowest occurred in February, then started to continuously increase until August (highest). After August, it started to decrease. The trend between robbery and theft are very similar.
- *Weapons violation*: No significant trend between the months.

3.3 Crime occurrences by the hour

Following observations were made for each crime by observing the histogram (see Appendix C).

- *Arson*: Mostly occurred between 12AM and 6AM.
- *Assault*: Increased from 6AM to 4PM (highest at 4PM) and decreased.
- *Battery*: Lowest from 5AM to 7AM. Increased from 7AM to 11PM and decreased from 11PM to 5AM.
- *Burglary*: Significantly increased from 6AM to 8AM.
- *Concealed carry license violation*: Highest between 10PM to 11PM.
- *Crim sexual assault*: Majority of the crime occurred between 12AM and 1AM.
- *Criminal damage*: No significant trend was observed.
- *Criminal trespass*: No significant trend was observed.
- *Deceptive practice*: Majority of the crime occurred between 12AM to 1AM, 9AM to 10AM, and 12PM to 1PM.
- *Gambling*: Highest between 6PM to 9PM.
- *Homicide*: No significant trend was observed.
- *Human trafficking*: Mostly occurred between 12AM to 1AM.
- *Interference with public officer*: Mostly occurred between 6PM to 9PM.
- *Intimidation*: No significant trend was observed.
- *Kidnapping*: Mostly occurred between 3PM to 7PM.
- *Liquor law violation*: Mostly occurred between 5PM to 10PM.
- *Motor vehicle theft*: Increased continuously from 5AM to 11PM.
- *Narcotics*: Mostly occurred from 10AM to 1PM and 6PM to 9PM.
- *Obscenity*: Highest between 12AM to 1AM.
- *Offense involving children*: Majority of the crime occurred from 12AM to 1AM.

- *Other offense*: No significant trend was observed.
- *Prostitution*: Mostly occurred from 6PM to 12AM.
- *Public indecency*: No significant trend was observed. Not enough data.
- *Public peace violation*: No significant trend was observed
- *Robbery*: Lowest from 4AM to 6AM. Increased from 6AM to 7PM and decreased from 7PM to 4AM.
- *Sex offense:* Highest from 12AM to 1AM.
- *Stalking*: No significant trend was observed
- *Theft*: Lowest from 7AM to 9AM. Increased from 9AM to 11PM and decreased from 11PM to 7AM.
- *Weapons violation*: Majority of the crime occurred from 6PM to 1AM.

4. Statistical data analysis

4.1 Shapiro-Wilk test

Sample means over the years had to be compared statistically to provide the confidence that the total occurrences of each crime for 2019 will have no significant difference from 2016, 2017, and 2018. To do this, Shapiro-Wilk test was conducted first to determine if each dataset was normally distributed. This statistical test reports total occurrences of each crime per month over the 3 years. Significance level of $\alpha = 0.05$ was chosen. Null hypothesis was rejected when the p-value was less than α. Null hypothesis and alternative hypothesis were:

Shapiro-Wilk Test Hypothesis

H_0 : The sample is a Gaussian distribution

H_a : The sample is not a Gaussian distribution

This hypothesis test was conducted for each crime for each year. Along with this hypothesis, the P-P plot (probability plot) was plotted to visualize and compare the empirical cumulative distribution function with a specified theoretical cumulative distribution function.

Figure 5 shows one of the results from 29 unique crimes where the null hypothesis was failed to be rejected for 3 consecutive years. Along with the p-value, P-P plot was analyzed to confirm the normality.

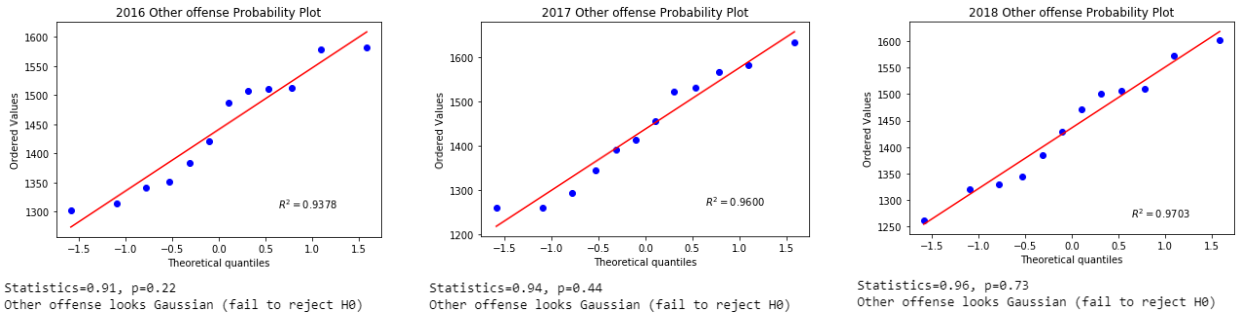


Fig 5. Example of a crime (other offense) where null hypothesis was failed to be rejected for 3 consecutive years

Table 1 shows the results from the Shapiro-Wilk test in which the null hypothesis was failed to be rejected. The outcome was 21 out of 29 crimes with a Gaussian distribution curve.

Table 1. Shapiro-Wilk test (H_0 failed to be rejected)

	2016		2017		2018	
Crime	Statistic	P	Statistic	P	Statistic	P
Arson	0.98	0.98	0.93	0.42	0.91	0.24
Assault	0.87	0.07	0.91	0.19	0.98	0.96
Battery	0.93	0.43	0.93	0.40	0.96	0.79
Burglary	0.89	0.11	0.96	0.84	0.97	0.94
Concealed carry license violation	0.87	0.07	0.93	0.39	0.98	0.98
Criminal damage	0.94	0.52	0.98	0.97	0.93	0.42
Deceptive practice	0.91	0.20	0.91	0.19	0.96	0.81
Gambling	0.90	0.18	0.94	0.50	0.88	0.09
Homicide	0.97	0.88	0.92	0.28	0.94	0.48
Interference with public officer	0.94	0.54	0.94	0.53	0.95	0.64
Intimidation	0.96	0.83	0.90	0.17	0.96	0.77
Kidnapping	0.93	0.42	0.98	0.96	0.91	0.21
Liquor law violation	0.88	0.08	0.89	0.13	0.90	0.17
Motor vehicle theft	0.92	0.25	0.96	0.83	0.98	0.99
Narcotics	0.88	0.09	0.97	0.89	0.97	0.87
Other offense	0.91	0.22	0.94	0.44	0.96	0.73
Public peace violation	0.93	0.43	0.93	0.41	0.95	0.69
Robbery	0.94	0.49	0.93	0.40	0.98	0.98
Sex offense	0.95	0.58	0.89	0.12	0.88	0.10
Theft	0.96	0.76	0.97	0.94	0.96	0.73
Weapons violation	0.92	0.30	0.95	0.59	0.95	0.63

Figure 6 and 7 displays two examples where the null hypothesis was rejected for at least one of the years. P-value was smaller than α and was also confirmed by the P-P plot.

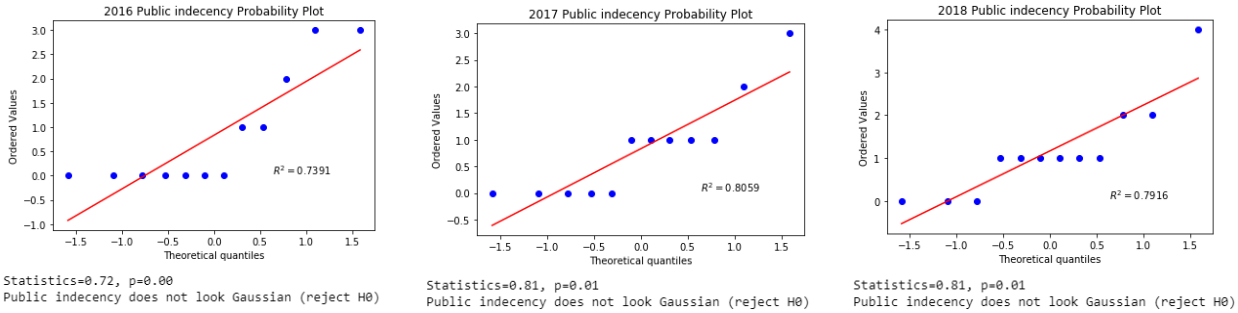


Fig 6. Example of a crime (public indecency) where null hypothesis was rejected for 3 consecutive years

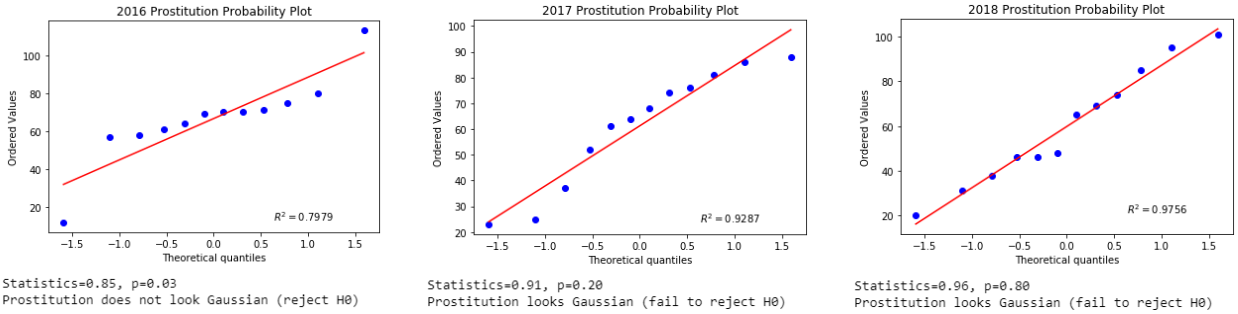


Fig 7. Example of a crime (prostitution) where null hypothesis was rejected for one or two years

Table 2 shows the results from the Shapiro-Wilk test which the null hypothesis was rejected. The outcome was 8 out of 29 crimes with a non-Gaussian distribution.

Table 2. Shapiro-Wilk test (H_0 rejected)

	2016		2017		2018	
Crime	Statistic	P	Statistic	P	Statistic	P
Crim sexual assault	0.96	0.83	0.95	0.65	0.85	0.04*
Criminal trespass	0.89	0.11	0.94	0.45	0.85	0.03*
Human trafficking	0.82	0.02*	0.73	0.00*	0.85	0.03*
Obscenity	0.79	0.01*	0.73	0.00*	0.94	0.53
Offense involving children	0.96	0.78	0.82	0.02*	0.90	0.17
Prostitution	0.85	0.03*	0.91	0.20	0.96	0.80
Public indecency	0.72	0.00*	0.81	0.01*	0.81	0.01*
Stalking	0.95	0.61	0.99	1.00	0.84	0.03*

* p-value < 0.05.

The crimes which the H_0 failed to be rejected (Gaussian distribution) were selected to be tested for one-way ANOVA test. The crimes which the H_0 was rejected (non-Gaussian distribution) were selected to be tested for Kruskal-Wallis H test.

4.2 One-way ANOVA test

To determine if a certain crime had mean difference in occurrences between the 3 years, one-way ANOVA test was conducted on the selected crimes (Table 1). Significance level of $\alpha = 0.05$ was chosen for this test. Post hoc comparisons between the groups were not analyzed since determination of which group or groups were different was not of interest.

One-way ANOVA Test Hypothesis

H_0 : The sample has no statistically significant differences between the group means.
H_a : The sample has significant differences between the group means.

Table 3. One-way ANOVA test (H_0 failed to be rejected)

Crime	F-Statistic	Sig.
Assault	1.31	0.28
Battery	0.09	0.91
Criminal damage	2.68	0.08
Deceptive practice	0.29	0.75
Gambling	0.02	0.98
Intimidation	1.82	0.18
Kidnapping	1.20	0.32
Liquor law violation	2.25	0.12
Narcotics	2.43	0.10
Other offense	0.01	0.99
Public peace violation	2.64	0.09
Sex offense	1.36	0.27
Theft	0.69	0.51

The crimes listed in the Table 3 (H_0 failed to be rejected) provides pretty good implication that the years in close proximity to 2016, 2017, and 2018 will have no significant differences in sample means. This is safe to assume under the assumption that no law has passed which affected a specific crime count to either increase or decrease significantly.

Table 4. One-way ANOVA test (H_0 rejected)

Crime	F-Statistic	Sig.
Arson	8.31	0.00
Burglary	6.34	0.00
Concealed carry license violation	27.54	0.00
Homicide	4.30	0.02
Interference with public officer	11.45	0.00
Motor vehicle theft	4.39	0.02
Robbery	7.55	0.00
Weapons violation	21.06	0.00

H_0 rejected crimes (Table 4) indicates that the group size was either too low (low crime occurrences for that crime overall) or that total count of a crime have been increasing or decreasing noticeably over the 3 years. These crimes were compared to the bar graph analysis. Below are the observations which were made initially:

- *Arson*: Decreased over the years.
- *Burglary*: Decreased every year.
- *Concealed carry license violation*: Increased over the years.
- *Homicide*: Decreased over the years.

- *Interference with public officer*: Increased over the years.
- *Motor vehicle theft*: Increased from 2016 to 2017, but decreased in 2018.
- *Robbery*: Decreased significantly in the year of 2018 compared to the past two years.
- *Weapons violation*: Significantly increased over the years.

These initial observations provide some insights as to why the null hypothesis could have been rejected.

4.3 Kruskal-Wallis H test

Kruskal-Wallis H test was performed on the nonparametric crimes (Table 2) since they could not be tested via one-way ANOVA test. Post hoc comparisons between the groups were not analyzed since determination of which group or groups were different was not of interest.

Kruskal-Wallis H Test Hypothesis

H_0 : The population median of all of the groups are equal.
H_a : The population median of all of the groups are not equal.

Instead of having α level of 0.05 to be the determining factor for testing the null hypothesis, critical chi square value was chosen to be compared to H statistics. For 2 degrees of freedom and α level of 0.05, critical chi square value was 5.9915. If the critical chi-square value was less than the H statistic, null hypothesis was rejected. If the chi-square value was more than the H statistic, null hypothesis was failed to be rejected.

Table 5. Kruskal-Wallis H test

Crime	H-Statistic	Sig.	Critical χ^2
Crim sexual assault	0.43	0.81	5.99
Criminal trespass	6.41*	0.04	5.99
Human trafficking	0.47	0.79	5.99
Obscenity	5.90	0.05	5.99
Offense involving children	0.74	0.69	5.99
Prostitution	0.42	0.81	5.99
Public indecency	1.34	0.51	5.99
Stalking	1.27	0.53	5.99

* Crime which H_0 was rejected.
Note: Non-rounded critical χ^2 was used for the test

4.4 Heat map – Month, hour, and weekday correlations

Heat maps were generated to present correlations between month vs. hour, month vs. day and weekday vs. hour (see Appendix D). From the above heatmap, it is observed that and heatmap are very similar. This indicates that the variable 'hour' is very robust in measuring the crime rates. Just because 'hour' seems to be the most significant variable, it does not mean the other two variables (Month and Weekday) can be ignored.

For example, looking at Weekday vs. Hour heatmap for the offense involving children, the crime mostly occurred on Friday, Sunday, and Monday. You could think to yourself there's no pattern between the weekdays. But looking at our calendar, 1st of January for 2016, 2017, and 2018 were Friday, Sunday and Monday respectively. By observing the other two heatmaps (Month vs. Hour and Month vs. Day), we can see that this crime mostly occurred on January 1st, between 12 AM and 1 AM.

5. Machine learning

5.1 Feature selection

Following features (independent variables) used from the dataset: *Hour, Weekday, IUCR, Description, Beat, District, Ward, Community Area, and FBI Code* (refer to Section 1.3). *Month* was the dependent variable. *Location description* was excluded from the features because *IUCR* is directly linked to *Primary Type* and *Description*. *Longitude* and *Latitude* were excluded as well since information from *Beat, District, Ward,* and *Community Area* provided sufficient information.

5.2 Pre-processing

Even though *IUCR* and *FBI Code* seemed to contain only integers (see Fig 9), there were some values with an alphabet attached. These along with the *Description* were preprocessed using the Label Encoder to transform non-numerical labels to numerical labels.

	Month	Hour	Weekday	IUCR	Description	Beat	District	Ward	Community Area	FBI Code
44	2	9.000000	2	0810	OVER $500	834	8	18	70	06
56	2	8.400000	5	0820	$500 AND UNDER	1532	15	28	25	06
89	3	22.000000	6	0890	FROM BUILDING	1831	18	42	8	06
147	1	22.000000	2	0810	OVER $500	723	7	20	68	06
150	7	0.016667	4	0810	OVER $500	1822	18	27	8	06

Fig 9. Dataframe before the Label Encoder (5 selected rows from *Theft*)

	Month	Hour	Weekday	IUCR	Description	Beat	District	Ward	Community Area	FBI Code
44	2	9.000000	2	0	5	834	8	18	70	0
56	2	8.400000	5	1	0	1532	15	28	25	0
89	3	22.000000	6	8	3	1831	18	42	8	0
147	1	22.000000	2	0	5	723	7	20	68	0
150	7	0.016667	4	0	5	1822	18	27	8	0

Fig 10. Dataframe after the Label Encoder (5 selected rows from *Theft*)

5.3 Principal component analysis (PCA)

As a preliminary model, *Theft* was selected to see if PCA could show clear visualization of different months within the crime. The features were preprocessed to scale into a distribution with a mean value of 0 and standard deviation of 1.

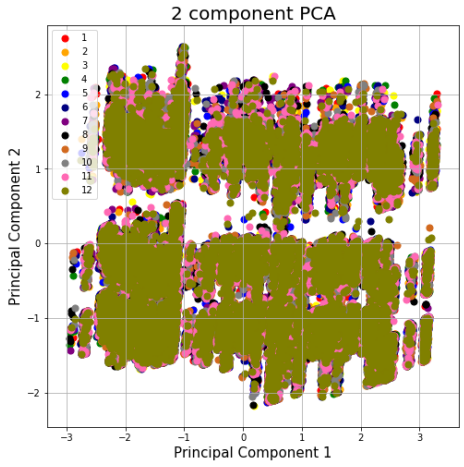


Fig 11. 2 components PCA for *Theft*

Upon the analysis, it was determined that PCA could not distinguish different months within the model.

5.4 K-nearest neighbors (KNN) classifier

Since the one of the main goal for the client was to budget more accurately, this analysis was performed over each district for all crimes. Prior to performing KNN, there were some crimes which either has never occurred in a certain district, or had too few occurrences to properly perform the test. The minimal threshold for this analysis was set to 50 occurrences for each crime per district. Crimes below the threshold were not analyzed and reported as "Crime has neither occurred in the district nor had too few occurrences." Each crime was tested per district with the training size of 70% and test size of 30%. For this analysis, the number of neighbors was set to 12 (number of months). Accuracy classification scores were reported for each crime in the district (see Appendix E). The scores were very poor. The highest accuracy score reported was 0.53 (*Prostitution*, District 25) while the lowest being 0.00.

5.5 Random Forest

Unlike the KNN classifier analysis, random forest analysis was performed by combining all districts together. It was not cost effective to analyze each district while tuning the parameters and performing cross validation. The parameters for the algorithm were tuned by using grid search with cross validation. Total of 9 parameter combinations were tested to fit for the best model performance. Accuracy scores and confusion matrix were reported (see Table 6 and Appendix F). The results were very poor.

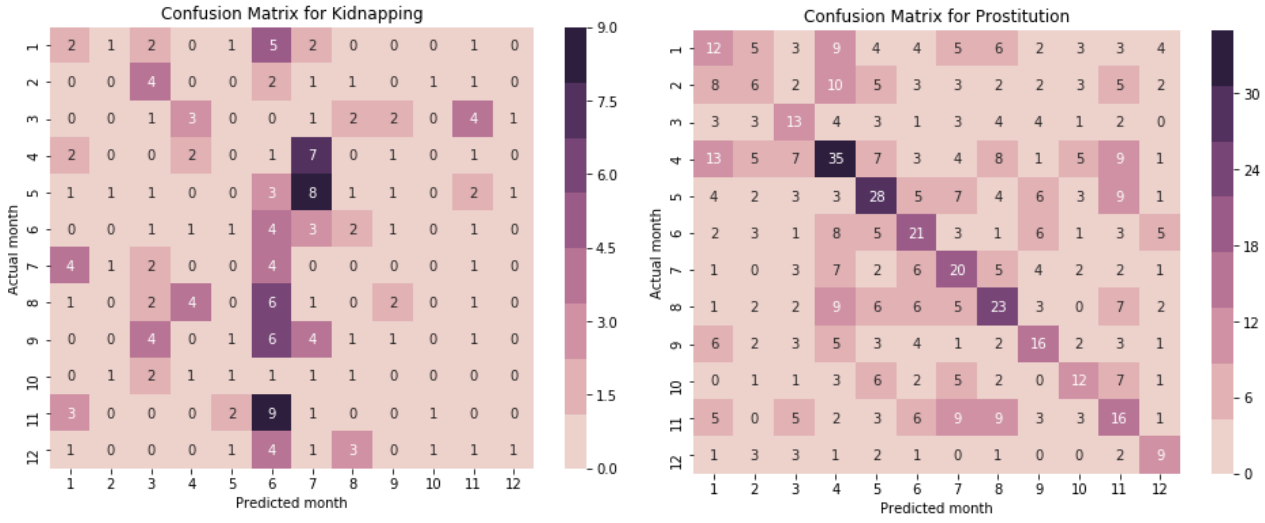


Fig 12. Confusion matrix of the worst (left) and the best (right) results from the random forest

Table 6. Random forest model accuracy scores

Crime	Accuracy score	Crime	Accuracy score
Arson	0.10	Liquor law violation	0.18
Assault	0.11	Motor vehicle theft	0.11
Battery	0.11	Narcotics	0.12
Burglary	0.11	Obscenity	0.09
Concealed carry license violation	0.17	Offense involving children	0.15
Crim sexual assault	0.11	Other offense	0.11
Criminal damage	0.11	Prostitution	0.31
Criminal trespass	0.11	Public indecency	*
Deceptive practice	0.10	Public peace violation	0.11
Gambling	0.19	Robbery	0.12
Homicide	0.14	Sex offense	0.12
Human trafficking	*	Stalking	0.11
Interference w/ public officer	0.08	Theft	0.11
Intimidation	0.09	Weapons violation	0.10
Kidnapping	0.07		

* Not enough observations for the analysis

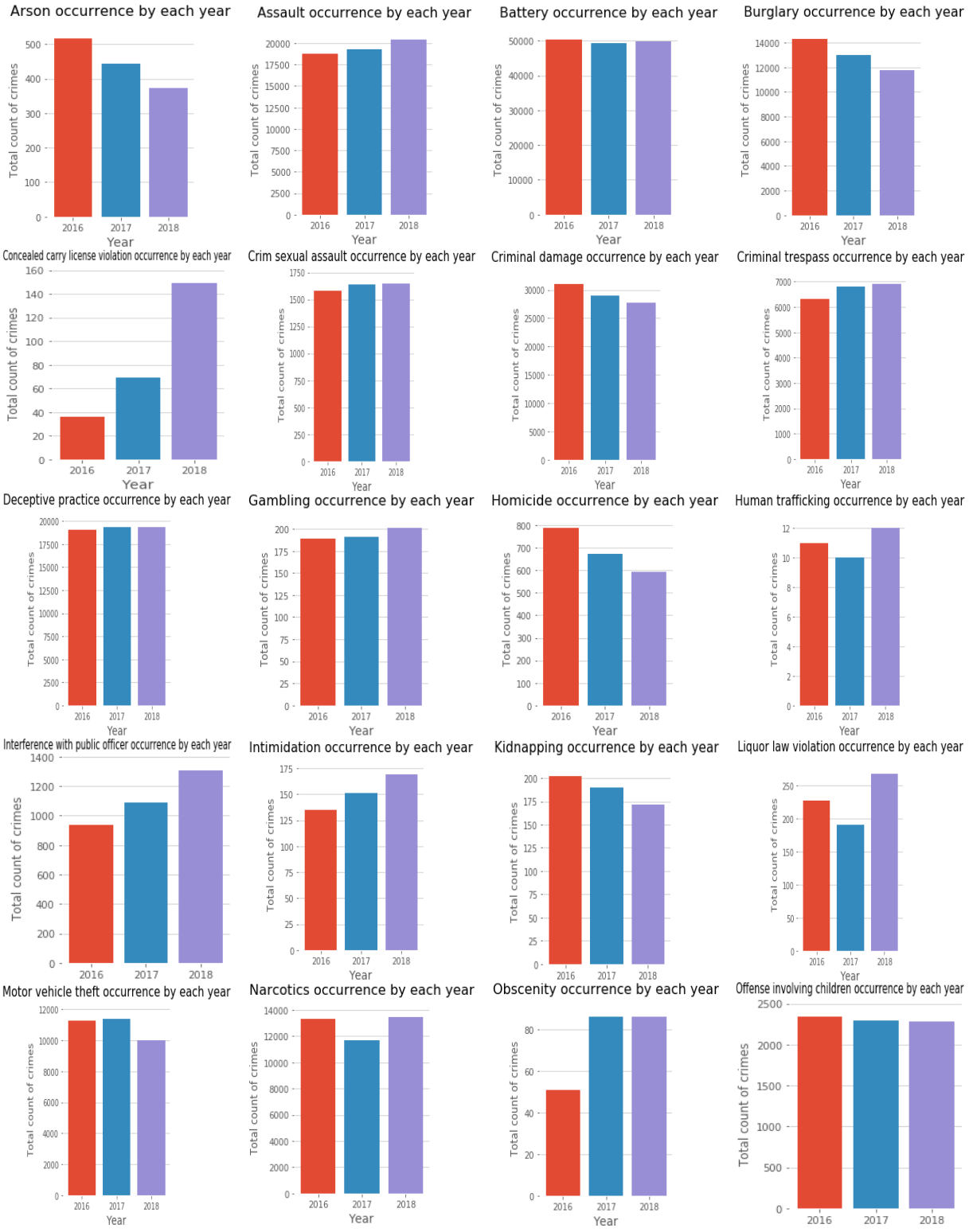
6. Conclusion

Machine learning could not properly predict if certain crime occurs more during a certain month. The results were very poor and no correlations were observed between the months and its features. Since all three years were combined for the analysis, the validity of the machine learning results for the crimes listed in Table 1 (Shapiro-Wilk test) are stronger than those not listed.

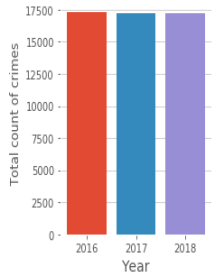
Month-to-month trend can be observed more visibly by referring to the graphs plotted in the Appendix B and the observations noted in Section 3.2. Although some crimes don't present obvious trend for all years, some crimes do. For example, *Theft* continuously increases from February (lowest) to August (highest) for all three consecutive years. From the heatmap, as portrayed in Appendix D, it appears to be that hour is the most significant variable when it comes to frequency of crimes.

APPENDIX A

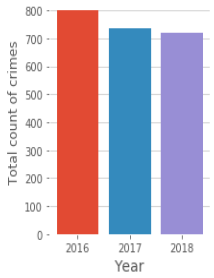
Bar graphs of crime occurrences by each year



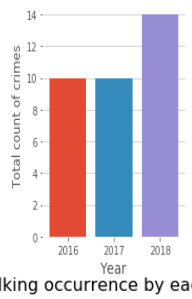
Other offense occurrence by each year



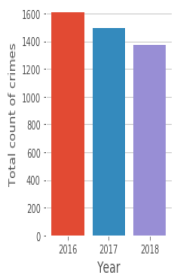
Prostitution occurrence by each year



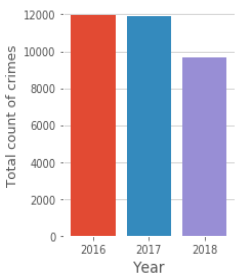
Public indecency occurrence by each year



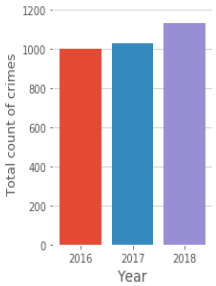
Public peace violation occurrence by each year



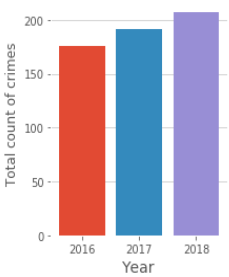
Robbery occurrence by each year



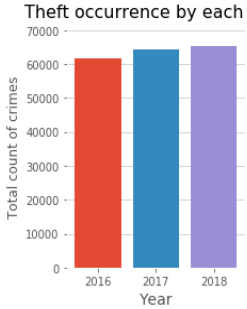
Sex offense occurrence by each year



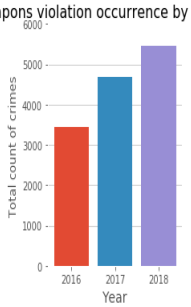
Stalking occurrence by each year



Theft occurrence by each year

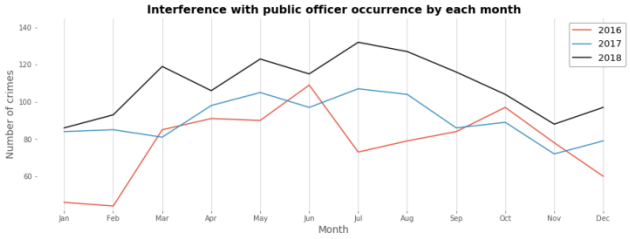


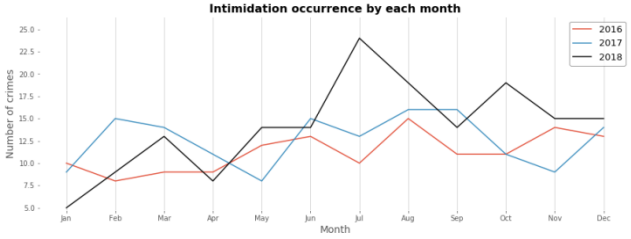
Weapons violation occurrence by each year

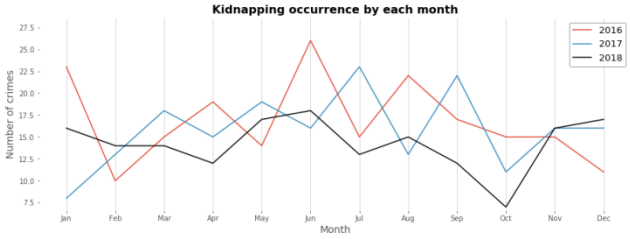


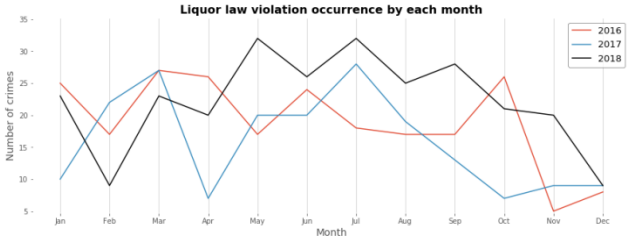
APPENDIX B

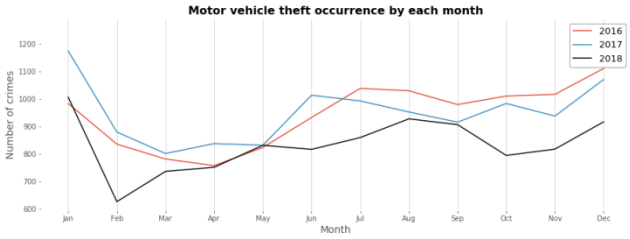
Time-series plot of crime occurrences by each month and year

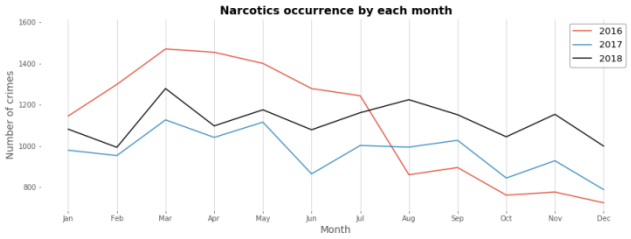

Interference with public officer occurrence by each month
Number of crimes
Month
2016
2017
2018

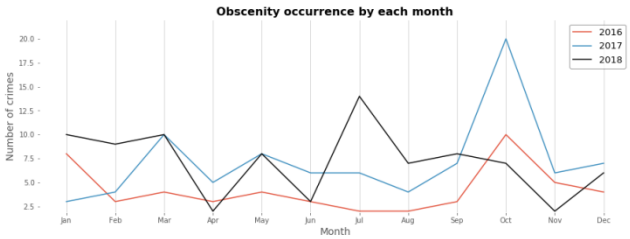

Intimidation occurrence by each month
Number of crimes
Month
2016
2017
2018

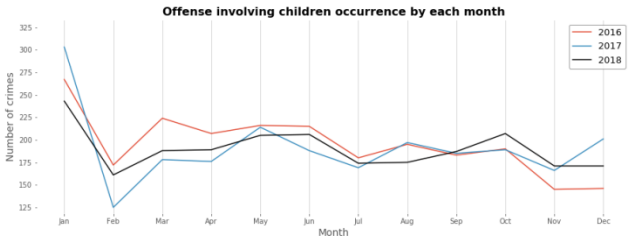

Kidnapping occurrence by each month
Number of crimes
Month
2016
2017
2018

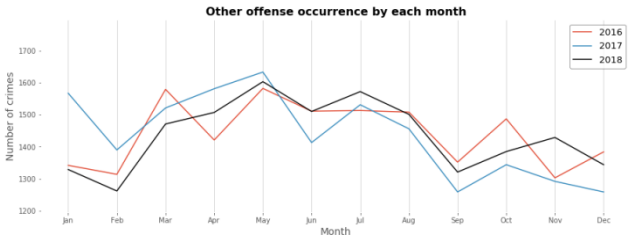

Liquor law violation occurrence by each month
Number of crimes
Month
2016
2017
2018

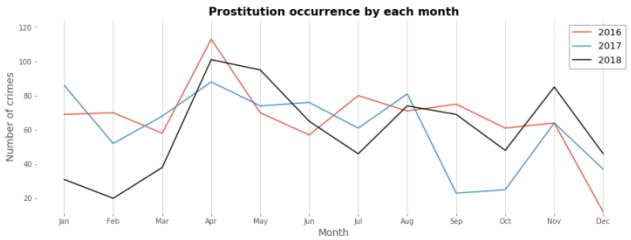

Motor vehicle theft occurrence by each month
Number of crimes
Month
2016
2017
2018

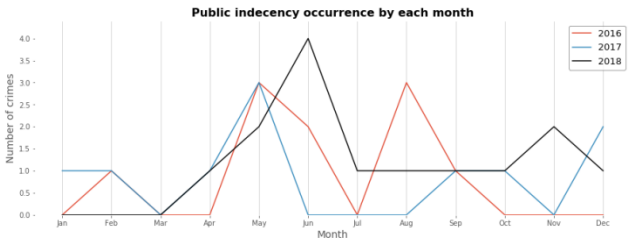

Narcotics occurrence by each month
Number of crimes
Month
2016
2017
2018

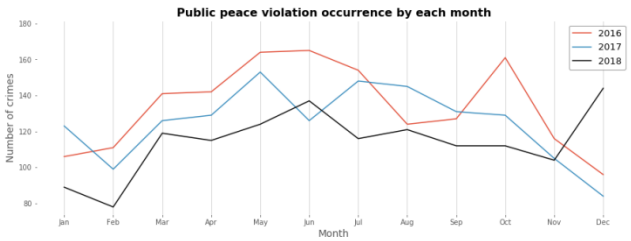

Obscenity occurrence by each month
Number of crimes
Month
2016
2017
2018

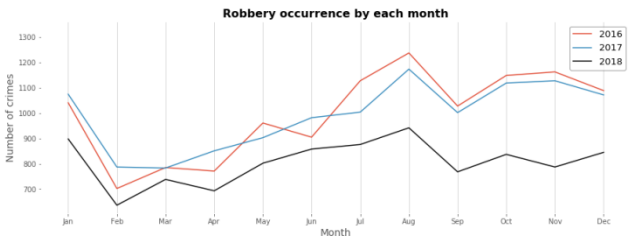

Offense involving children occurrence by each month
Number of crimes
Month
2016
2017
2018

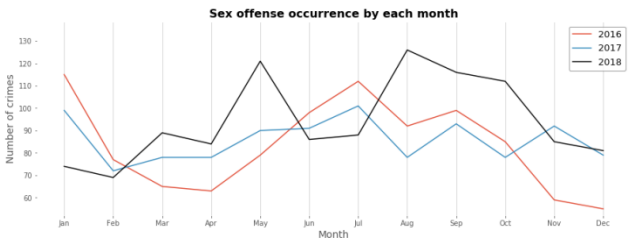

Other offense occurrence by each month
Number of crimes
Month
2016
2017
2018

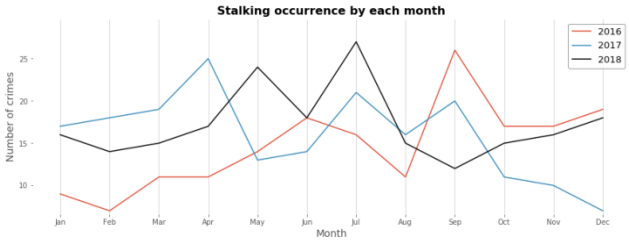

Prostitution occurrence by each month
Number of crimes
Month
2016
2017
2018

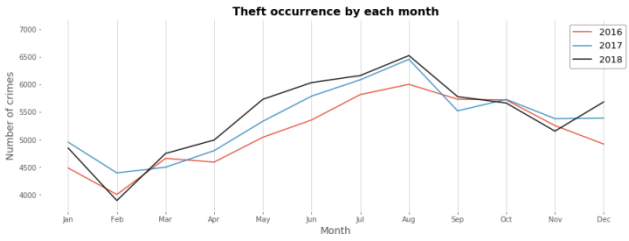

Public indecency occurrence by each month
Number of crimes
Month
2016
2017
2018

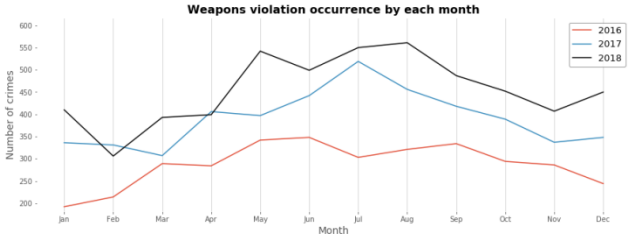

Public peace violation occurrence by each month
Number of crimes
Month
2016
2017
2018


Robbery occurrence by each month
Number of crimes
Month
2016
2017
2018


Sex offense occurrence by each month
Number of crimes
Month
2016
2017
2018

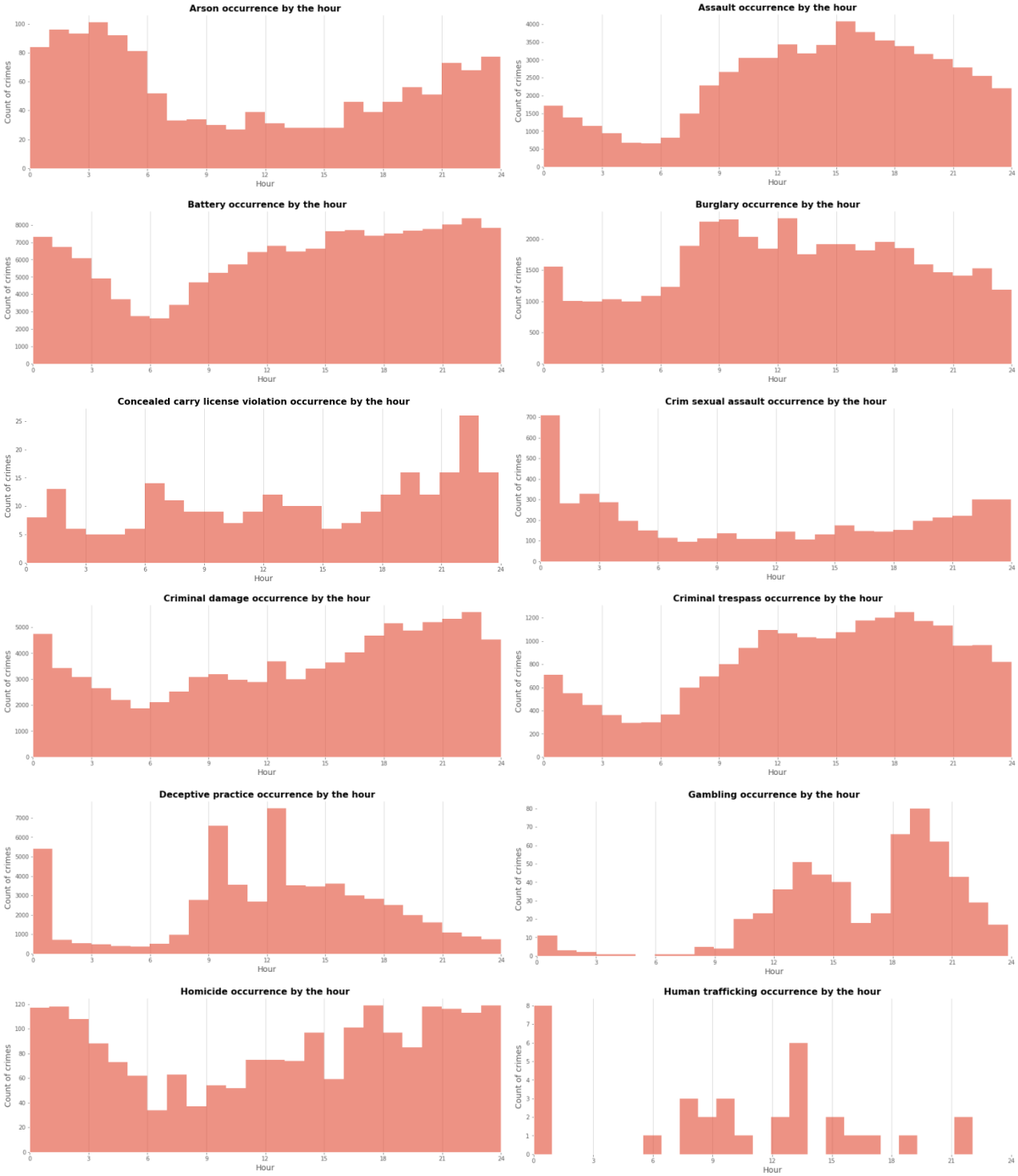

Stalking occurrence by each month
2016
2017
2018
Number of crimes
Month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

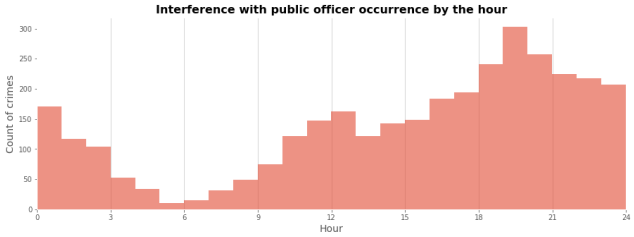

Theft occurrence by each month
2016
2017
2018
Number of crimes
Month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

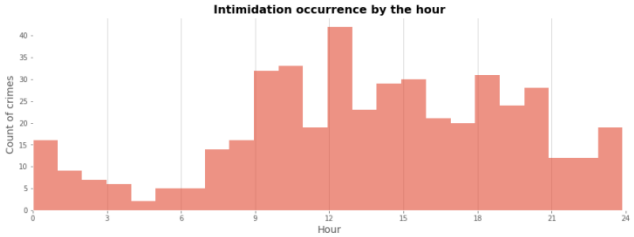

Weapons violation occurrence by each month
2016
2017
2018
Number of crimes
Month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

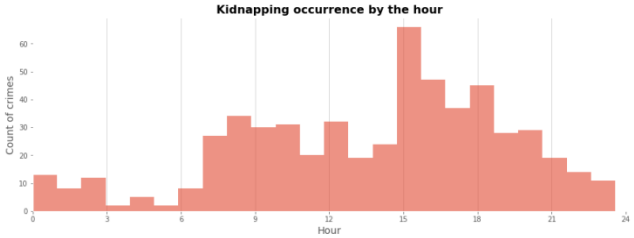
APPENDIX C

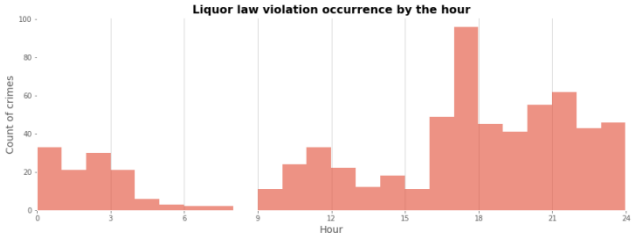
Histogram of crime occurrences by hour

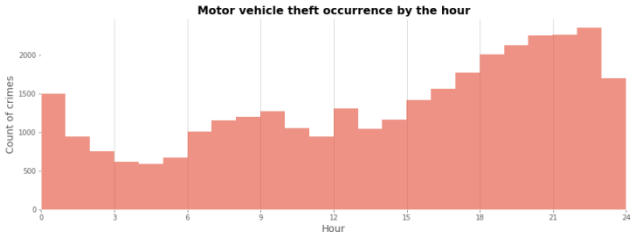


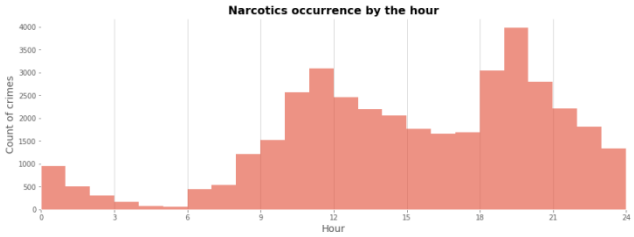

Interference with public officer occurrence by the hour
Count of crimes
Hour

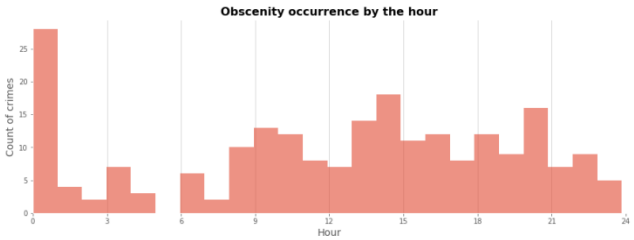

Intimidation occurrence by the hour
Count of crimes
Hour

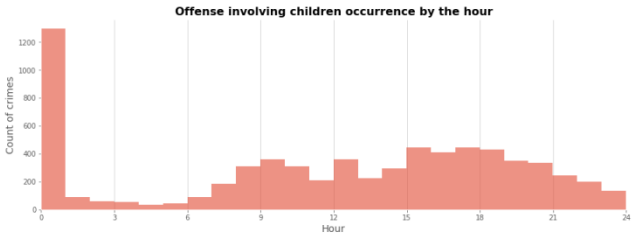

Kidnapping occurrence by the hour
Count of crimes
Hour

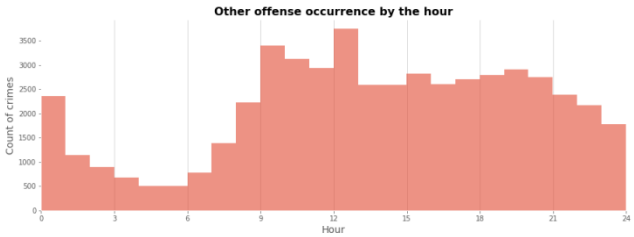

Liquor law violation occurrence by the hour
Count of crimes
Hour

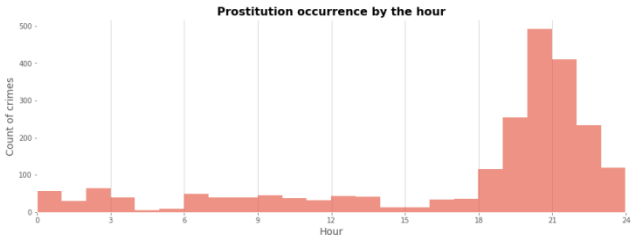

Motor vehicle theft occurrence by the hour
Count of crimes
Hour

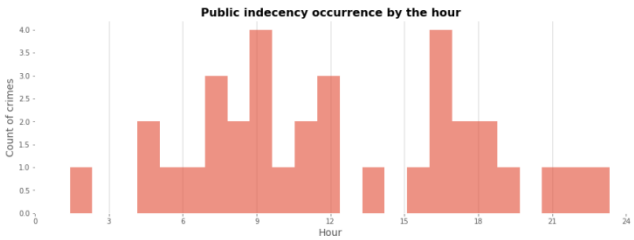

Narcotics occurrence by the hour
Count of crimes
Hour

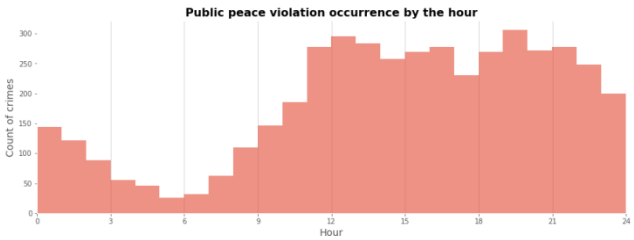

Obscenity occurrence by the hour
Count of crimes
Hour

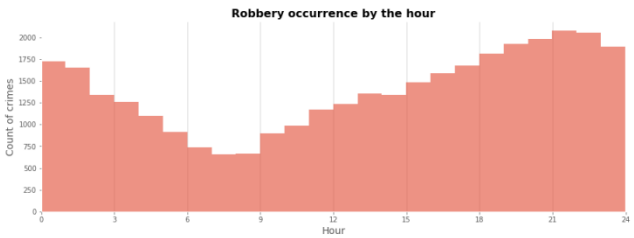

Offense involving children occurrence by the hour
Count of crimes
Hour

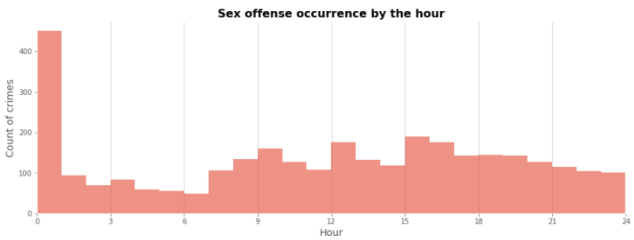

Other offense occurrence by the hour
Count of crimes
Hour

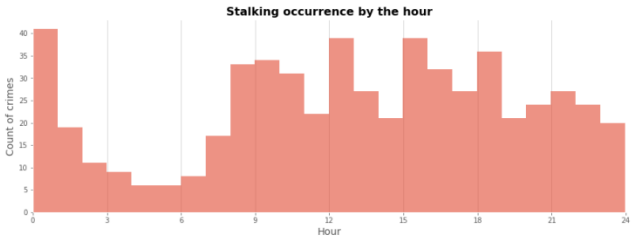

Prostitution occurrence by the hour
Count of crimes
Hour

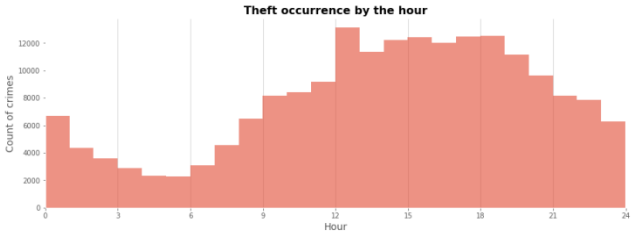

Public indecency occurrence by the hour
Count of crimes
Hour

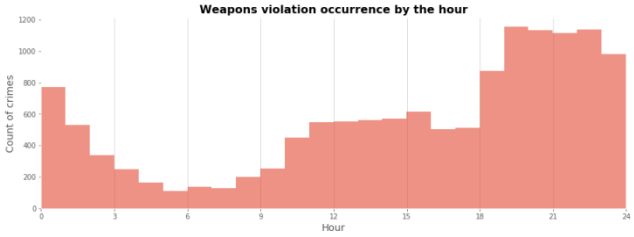

Public peace violation occurrence by the hour
Count of crimes
Hour


Robbery occurrence by the hour
Count of crimes
Hour


Sex offense occurrence by the hour
Count of crimes
Hour

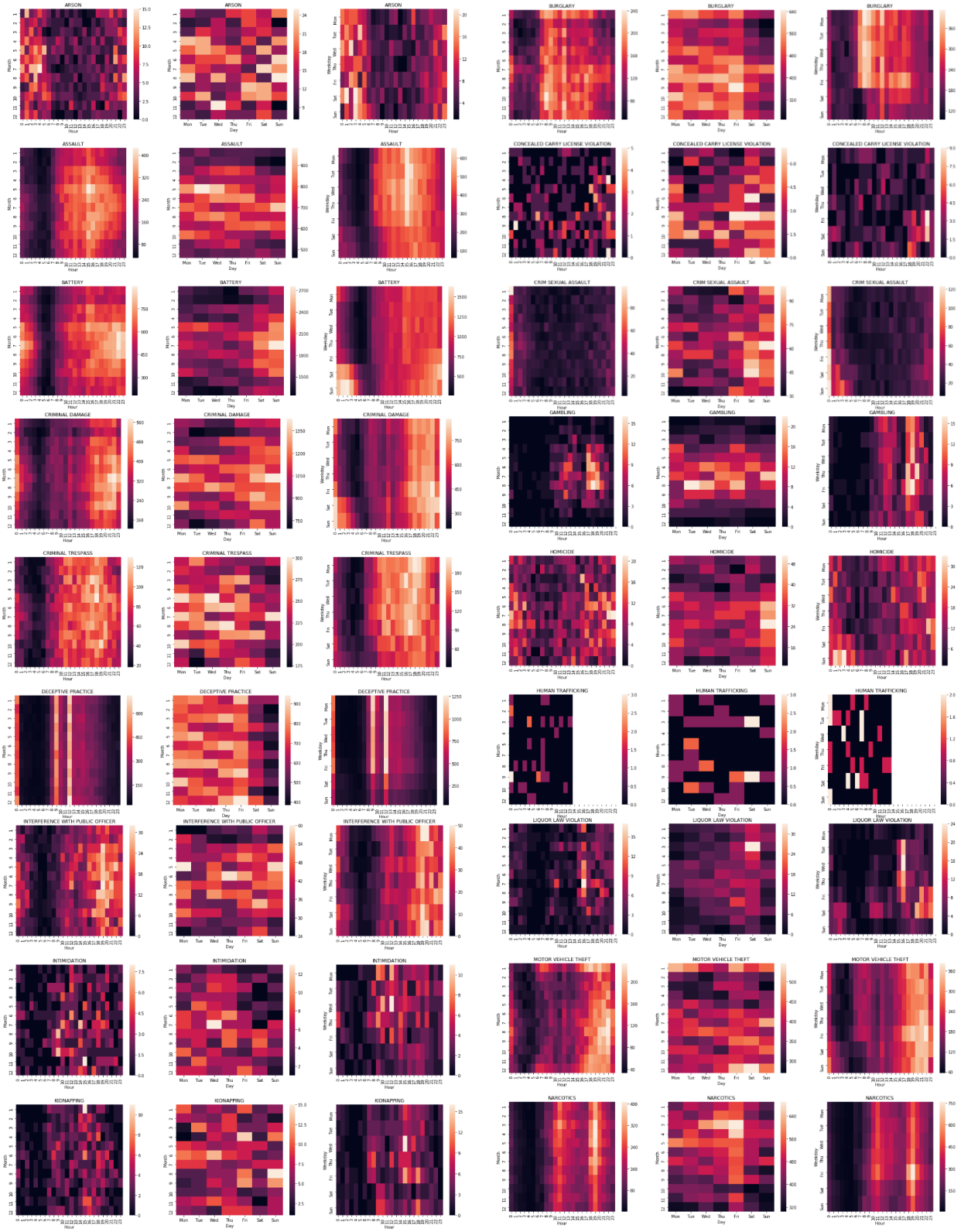

Stalking occurrence by the hour
Count of crimes
Hour

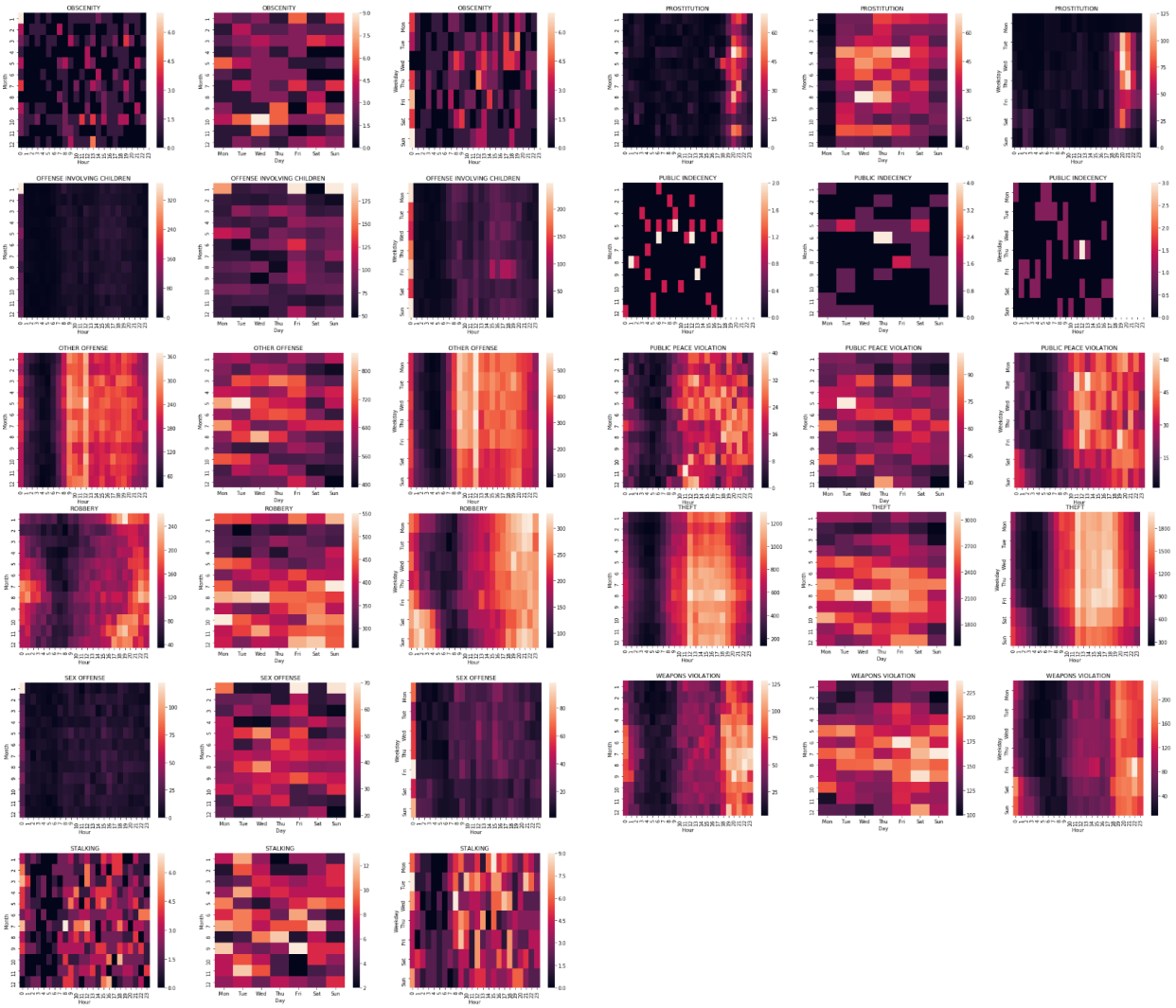

Theft occurrence by the hour
Count of crimes
Hour


Weapons violation occurrence by the hour
Count of crimes
Hour

APPENDIX D

Heat map correlation of month, hour, and weekday




OBSCENITY
OBSCENITY
OBSCENITY
PROSTITUTION
PROSTITUTION
PROSTITUTION
OFFENSE INVOLVING CHILDREN
OFFENSE INVOLVING CHILDREN
OFFENSE INVOLVING CHILDREN
PUBLIC INDECENCY
PUBLIC INDECENCY
PUBLIC INDECENCY
OTHER OFFENSE
OTHER OFFENSE
OTHER OFFENSE
PUBLIC PEACE VIOLATION
PUBLIC PEACE VIOLATION
PUBLIC PEACE VIOLATION
ROBBERY
ROBBERY
ROBBERY
THEFT
THEFT
THEFT
SEX OFFENSE
SEX OFFENSE
SEX OFFENSE
WEAPONS VIOLATION
WEAPONS VIOLATION
WEAPONS VIOLATION
STALKING
STALKING
STALKING

APPENDIX E

Accuracy scores from KNN classifier analysis

	District										
Crime	1	2	3	4	5	6	7	8	9	10	11
Arson	*	*	*	*0.15*	*0.07*	*0.05*	*0.09*	*0.03*	*0.11*	*0.11*	*0.23*
	12	14	15	16	17	18	19	20	22	24	25
	*	*	*0.22*	*	*	*	*	*	*	*	*0.06*
Assault	1	2	3	4	5	6	7	8	9	10	11
	0.09	*0.08*	*0.08*	*0.08*	*0.09*	*0.09*	*0.08*	*0.89*	*0.10*	*0.10*	*0.09*
	12	14	15	16	17	18	19	20	22	24	25
	0.08	*0.09*	*0.10*	*0.09*	*0.06*	*0.09*	*0.08*	*0.09*	*0.08*	*0.06*	*0.08*
Battery	1	2	3	4	5	6	7	8	9	10	11
	0.08	*0.09*	*0.10*	*0.09*	*0.09*	*0.09*	*0.08*	*0.09*	*0.08*	*0.09*	*0.10*
	12	14	15	16	17	18	19	20	22	24	25
	0.08	*0.09*	*0.09*	*0.09*	*0.09*	*0.09*	*0.09*	*0.08*	*0.08*	*0.10*	*0.08*
Burglary	1	2	3	4	5	6	7	8	9	10	11
	0.07	*0.09*	*0.09*	*0.08*	*0.09*	*0.09*	*0.10*	*0.09*	*0.12*	*0.10*	*0.09*
	12	14	15	16	17	18	19	20	22	24	25
	0.11	*0.10*	*0.10*	*0.12*	*0.10*	*0.14*	*0.10*	*0.09*	*0.12*	*0.09*	*0.08*
Concealed carry license violation	1	2	3	4	5	6	7	8	9	10	11
	*	*	*	*	*	*	*	*	*	*	*
	12	14	15	16	17	18	19	20	22	24	25
	*	*	*	*0.18*	*	*	*	*	*	*	*
Crim sexual assault	1	2	3	4	5	6	7	8	9	10	11
	0.15	*0.09*	*0.06*	*0.10*	*0.05*	*0.08*	*0.10*	*0.05*	*0.09*	*0.10*	*0.12*
	12	14	15	16	17	18	19	20	22	24	25
	0.12	*0.11*	*0.08*	*0.07*	*0.17*	*0.08*	*0.10*	*0.17*	*0.06*	*0.06*	*0.09*
Criminal damage	1	2	3	4	5	6	7	8	9	10	11
	0.10	*0.12*	*0.09*	*0.09*	*0.10*	*0.10*	*0.08*	*0.10*	*0.08*	*0.10*	*0.10*
	12	14	15	16	17	18	19	20	22	24	25
	0.10	*0.10*	*0.09*	*0.07*	*0.10*	*0.08*	*0.09*	*0.14*	*0.09*	*0.10*	*0.10*
Criminal trespass	1	2	3	4	5	6	7	8	9	10	11
	0.11	*0.12*	*0.12*	*0.10*	*0.10*	*0.11*	*0.08*	*0.14*	*0.08*	*0.05*	*0.11*
	12	14	15	16	17	18	19	20	22	24	25
	0.11	*0.05*	*0.10*	*0.11*	*0.08*	*0.09*	*0.07*	*0.10*	*0.08*	*0.10*	*0.11*
Deceptive practice	1	2	3	4	5	6	7	8	9	10	11
	0.09	*0.09*	*0.10*	*0.09*	*0.08*	*0.08*	*0.05*	*0.07*	*0.09*	*0.12*	*0.10*
	12	14	15	16	17	18	19	20	22	24	25
	0.09	*0.10*	*0.10*	*0.09*	*0.09*	*0.09*	*0.10*	*0.08*	*0.09*	*0.07*	*0.09*
Gambling	1	2	3	4	5	6	7	8	9	10	11
	*	*	*	*	*	*	*0.14*	*	*	*0.24*	*0.27*
	12	14	15	16	17	18	19	20	22	24	25
	*	*	*0.18*	*	*	*	*	*	*	*	*
Homicide	1	2	3	4	5	6	7	8	9	10	11
	*	*0.11*	*0.16*	*0.10*	*0.00*	*0.04*	*0.14*	*0.10*	*0.18*	*0.25*	*0.15*
	12	14	15	16	17	18	19	20	22	24	25
	0.05	*	*0.16*	*	*	*	*	*	*0.14*	*	*0.00*
Interference with public officer	1	2	3	4	5	6	7	8	9	10	11
	0.00	*0.06*	*0.12*	*0.06*	*0.09*	*0.12*	*0.09*	*0.07*	*0.06*	*0.05*	*0.08*
	12	14	15	16	17	18	19	20	22	24	25
	0.11	*	*0.13*	*	*	*0.26*	*0.12*	*	*0.10*	*0.06*	*0.07*
Liquor law violation	1	2	3	4	5	6	7	8	9	10	11
	*	*	*	*	*	*	*	*	*	*0.28*	*
	12	14	15	16	17	18	19	20	22	24	25
	0.00	*	*	*	*	*0.10*	*0.26*	*	*	*	*
Motor vehicle theft	1	2	3	4	5	6	7	8	9	10	11

	0.08	*0.11*	*0.07*	*0.06*	*0.08*	*0.07*	*0.11*	*0.09*	*0.10*	*0.09*	*0.09*
	12	14	15	16	17	18	19	20	22	24	25
	0.07	*0.12*	*0.09*	*0.13*	*0.08*	*0.13*	*0.08*	*0.06*	*0.12*	*0.10*	*0.09*
Narcotics	1	2	3	4	5	6	7	8	9	10	11
	0.12	*0.13*	*0.11*	*0.10*	*0.11*	*0.09*	*0.11*	*0.10*	*0.11*	*0.11*	*0.12*
	12	14	15	16	17	18	19	20	22	24	25
	0.10	*0.08*	*0.10*	*0.13*	*0.10*	*0.07*	*0.11*	*0.10*	*0.09*	*0.14*	*0.10*
Offense involving children	1	2	3	4	5	6	7	8	9	10	11
	0.12	*0.12*	*0.10*	*0.09*	*0.10*	*0.11*	*0.08*	*0.11*	*0.13*	*0.11*	*0.07*
	12	14	15	16	17	18	19	20	22	24	25
	0.14	*0.17*	*0.13*	*0.12*	*0.14*	*0.08*	*0.16*	*0.16*	*0.13*	*0.10*	*0.17*
Other offense	1	2	3	4	5	6	7	8	9	10	11
	0.11	*0.09*	*0.08*	*0.11*	*0.10*	*0.09*	*0.10*	*0.08*	*0.11*	*0.08*	*0.12*
	12	14	15	16	17	18	19	20	22	24	25
	0.10	*0.10*	*0.11*	*0.09*	*0.10*	*0.09*	*0.09*	*0.12*	*0.07*	*0.10*	*0.11*
Prostitution	1	2	3	4	5	6	7	8	9	10	11
	*	*	*	*	*0.15*	*	*0.29*	*0.16*	*0.42*	*	*0.18*
	12	14	15	16	17	18	19	20	22	24	25
	*	*	*	*	*	*	*	*	*	*	*0.53*
Public peace violation	1	2	3	4	5	6	7	8	9	10	11
	0.08	*0.11*	*0.08*	*0.16*	*0.11*	*0.11*	*0.13*	*0.07*	*0.16*	*0.05*	*0.15*
	12	14	15	16	17	18	19	20	22	24	25
	0.08	*0.16*	*0.08*	*0.06*	*0.10*	*0.12*	*0.12*	*0.06*	*0.18*	*0.12*	*0.19*
Robbery	1	2	3	4	5	6	7	8	9	10	11
	0.10	*0.11*	*0.10*	*0.09*	*0.12*	*0.09*	*0.07*	*0.09*	*0.11*	*0.10*	*0.10*
	12	14	15	16	17	18	19	20	22	24	25
	0.11	*0.13*	*0.11*	*0.11*	*0.10*	*0.11*	*0.10*	*0.09*	*0.09*	*0.10*	*0.09*
Sex offense	1	2	3	4	5	6	7	8	9	10	11
	0.12	*0.03*	*0.11*	*0.18*	*0.12*	*0.18*	*0.19*	*0.21*	*0.15*	*0.10*	*0.09*
	12	14	15	16	17	18	19	20	22	24	25
	0.07	*0.06*	*0.09*	*0.05*	*0.09*	*0.07*	*0.12*	*0.04*	*0.09*	*0.14*	*0.08*
Theft	1	2	3	4	5	6	7	8	9	10	11
	0.11	*0.09*	*0.10*	*0.09*	*0.09*	*0.09*	*0.09*	*0.08*	*0.09*	*0.09*	*0.09*
	12	14	15	16	17	18	19	20	22	24	25
	0.10	*0.09*	*0.07*	*0.11*	*0.08*	*0.10*	*0.10*	*0.09*	*0.09*	*0.10*	*0.09*
Weapons violation	1	2	3	4	5	6	7	8	9	10	11
	0.03	*0.11*	*0.12*	*0.07*	*0.10*	*0.10*	*0.07*	*0.12*	*0.08*	*0.10*	*0.10*
	12	14	15	16	17	18	19	20	22	24	25
	0.12	*0.09*	*0.05*	*0.07*	*0.11*	*0.16*	*0.08*	*0.17*	*0.07*	*0.06*	*0.07*

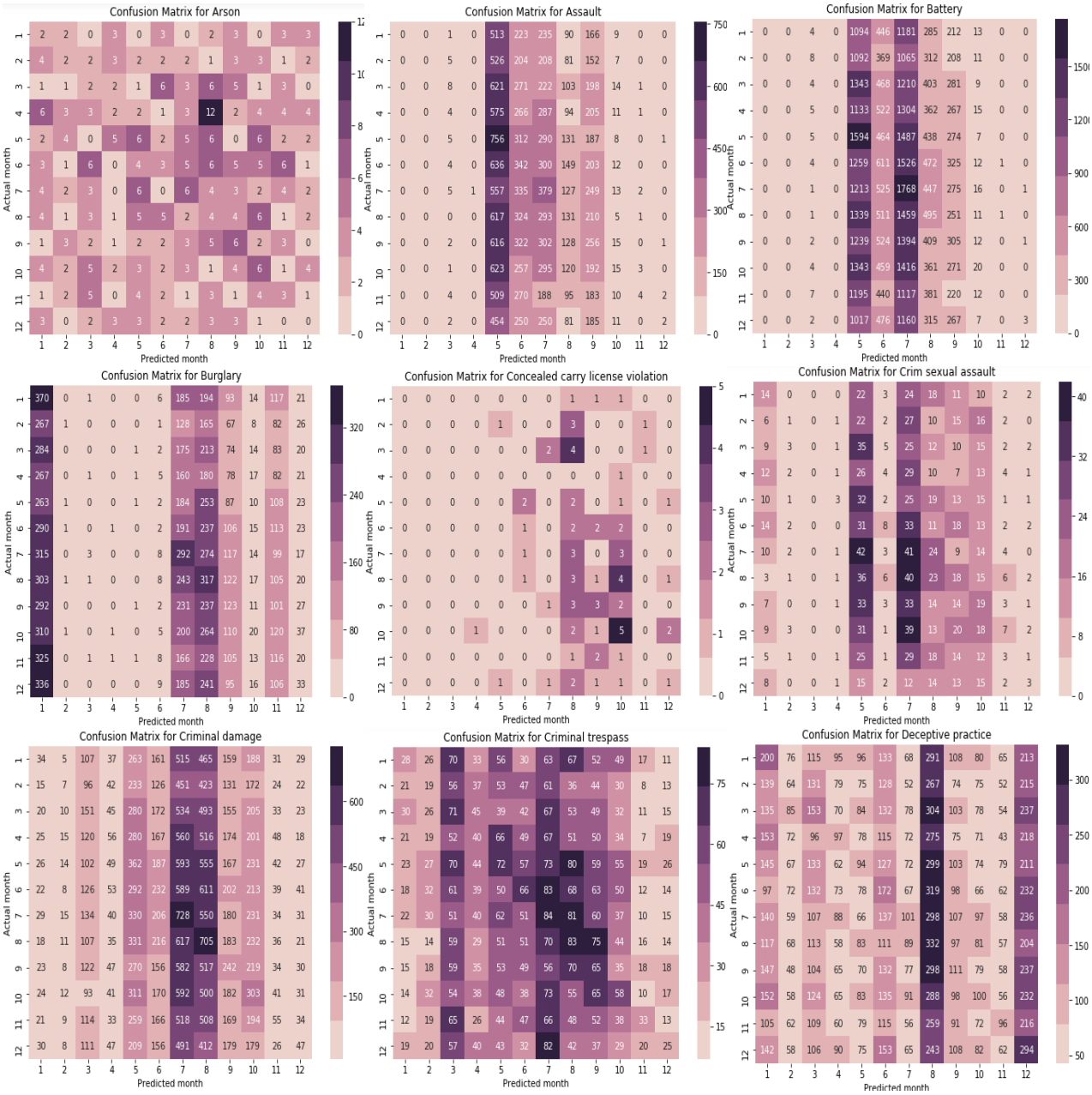
* Indicates crime has neither occurred in the district nor had too few occurrences.

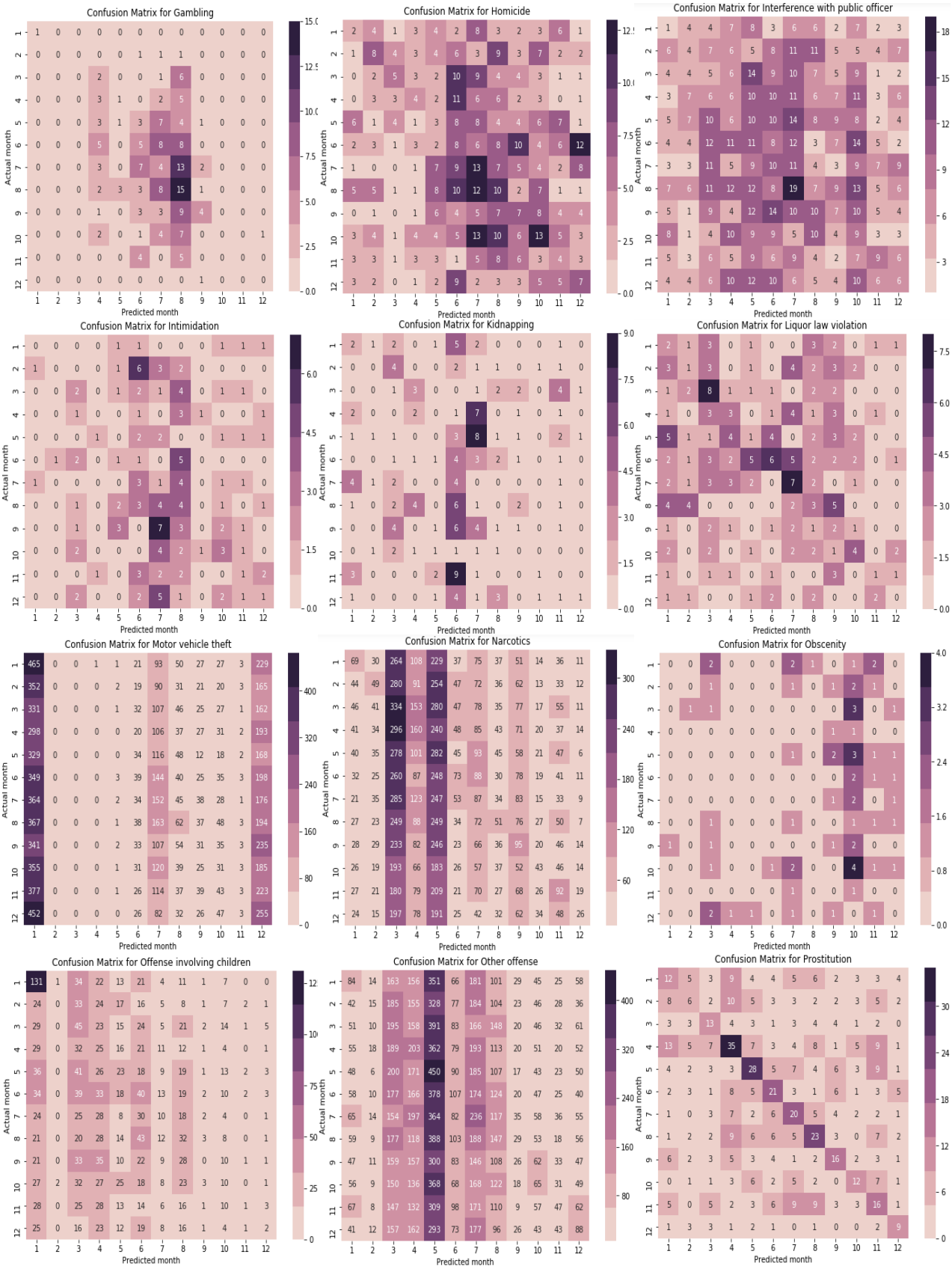
** *Human trafficking*, *Intimidation*, *Kidnapping*, *Obscenity*, *Public indecency*, and *Stalking* had too few occurrences in all districts and therefore and omitted from the table.

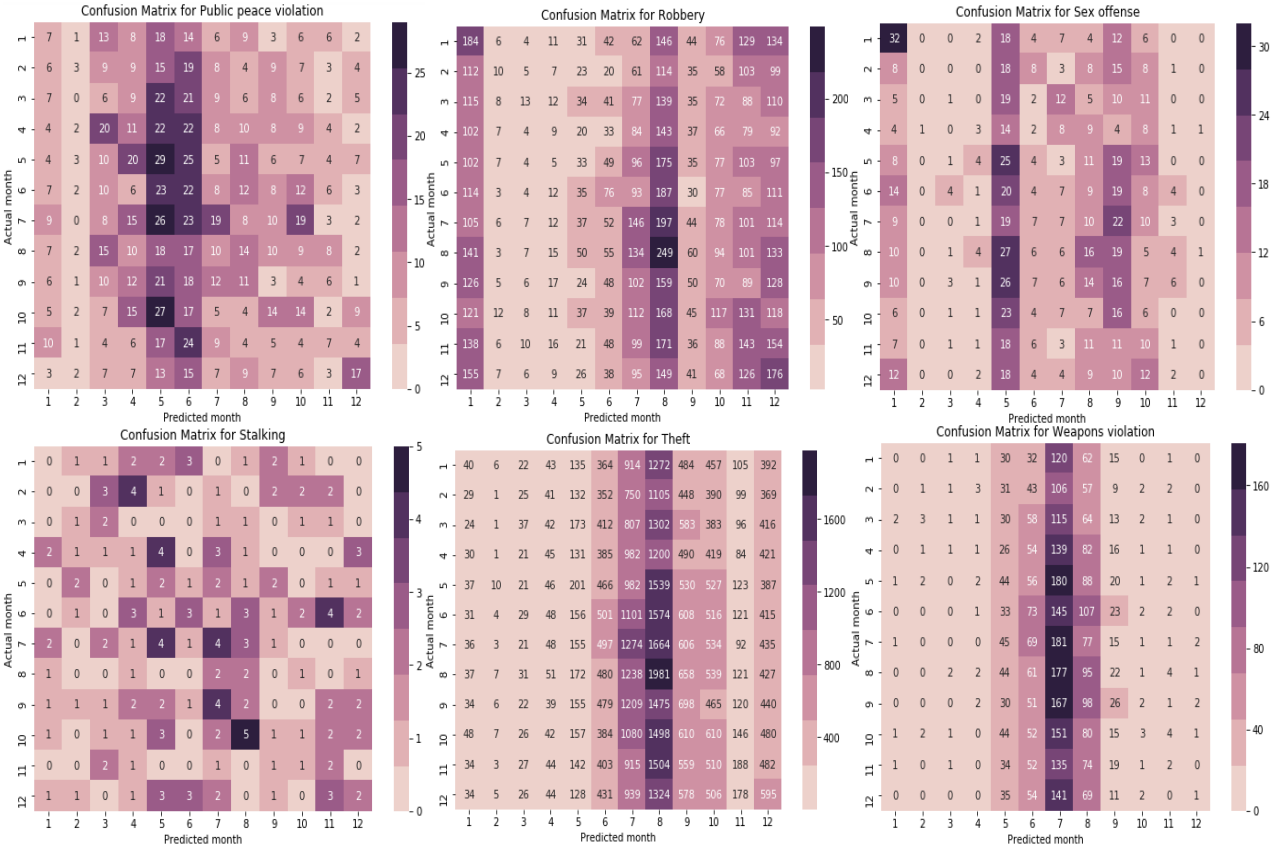
*** District 31 was omitted from the table due to too few occurrences for all crimes.

APPENDIX F

Confusion matrix for all crimes from random forest model




Confusion Matrix for Gambling
Confusion Matrix for Homicide
Confusion Matrix for Interference with public officer
Confusion Matrix for Intimidation
Confusion Matrix for Kidnapping
Confusion Matrix for Liquor law violation
Confusion Matrix for Motor vehicle theft
Confusion Matrix for Narcotics
Confusion Matrix for Obscenity
Confusion Matrix for Offense involving children
Confusion Matrix for Other offense
Confusion Matrix for Prostitution
Actual month
Predicted month


Confusion Matrix for Public peace violation
Confusion Matrix for Robbery
Confusion Matrix for Sex offense
Confusion Matrix for Stalking
Confusion Matrix for Theft
Confusion Matrix for Weapons violation
Actual month
Predicted month