Capstone Project 1 Statistical Data Analysis

By: Charlie Lee
Date: 2/24/2020

To represent the occurrence of each crime per year, the counts were accumulated monthly. This provided a consistent sample size of 12 for all years.

Shapiro-Wilk Test

To determine if each dataset was normally distributed, the Shapiro-Wilk test was conducted first. Since Shapiro-Wilk is based on a one-tailed test, significance level of α = 0.05 was chosen. Null hypothesis was rejected when the p-value was less than α. Null hypothesis and alternative hypothesis were:

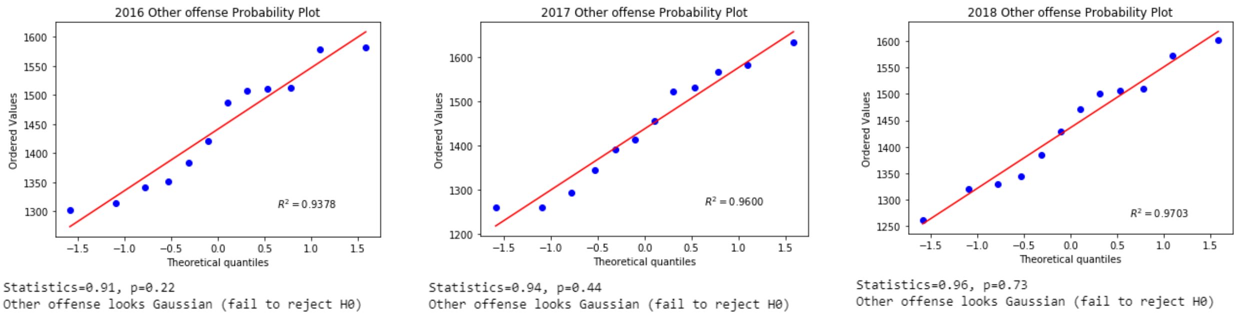
Shapiro-Wilk Test Hypothesis

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 1 shows one of the results from 29 unique crime categories 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.



```
Statistics=0.91, p=0.22
Other offense looks Gaussian (fail to reject H0)
```

```
Statistics=0.94, p=0.44
Other offense looks Gaussian (fail to reject H0)
```

```
Statistics=0.96, p=0.73
Other offense looks Gaussian (fail to reject H0)
```

Fig 1. Example of a crime (other offense) where null hypothesis was failed to be rejected for 3 consecutive years

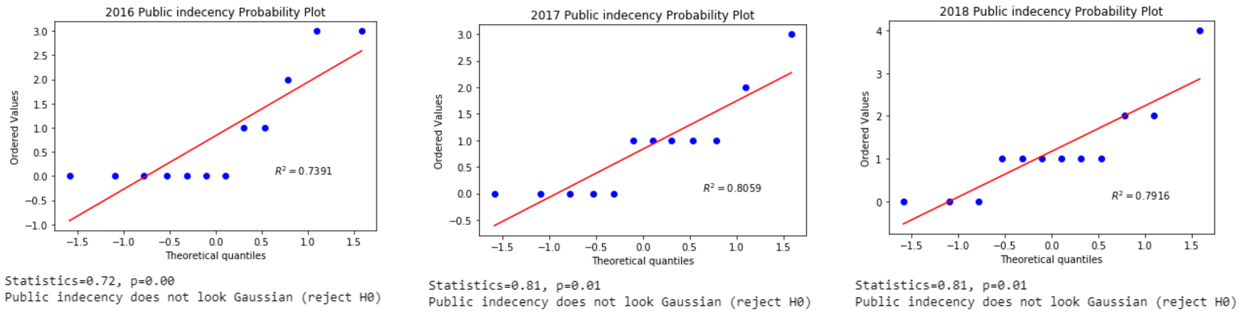
Table 1 shows the results from the Shapiro-Wilk test in which the null hypothesis was failed to be rejected. The outcome was 21 out of 29 crimes with a Gaussian distribution curve.

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 2 and 3 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.

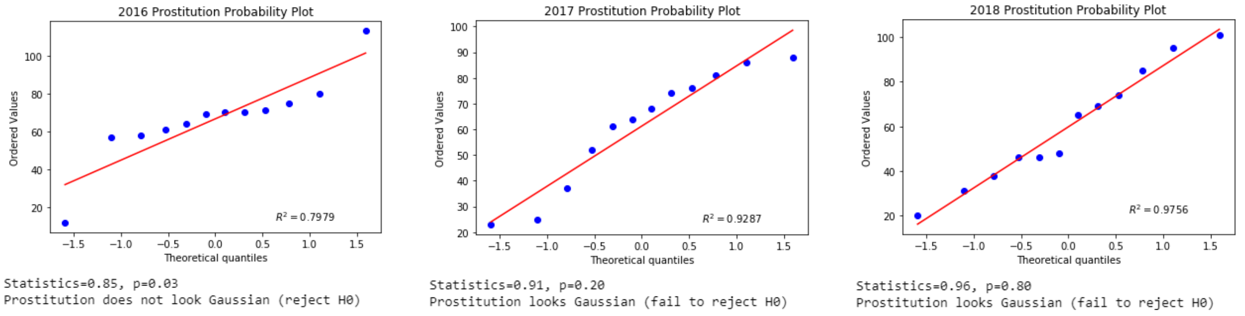


```
Statistics=0.72, p=0.00
Public indecency does not look Gaussian (reject H0)
```

```
Statistics=0.81, p=0.01
Public indecency does not look Gaussian (reject H0)
```

```
Statistics=0.81, p=0.01
Public indecency does not look Gaussian (reject H0)
```

Fig 2. Example of a crime (public indecency) where null hypothesis was rejected for 3 consecutive years



```
Statistics=0.85, p=0.03
Prostitution does not look Gaussian (reject H0)
```

```
Statistics=0.91, p=0.20
Prostitution looks Gaussian (fail to reject H0)
```

```
Statistics=0.96, p=0.80
Prostitution looks Gaussian (fail to reject H0)
```

Fig 3. 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.

One-way ANOVA Test

To determine if a certain crime had mean difference between the 3 years, one-way ANOVA test was conducted on the selected crimes (Table 1). Significance level of α = 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 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 premature 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**: No trend from 2016 to 2017, but increased from 2017 to 2018.
- **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.

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. S

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

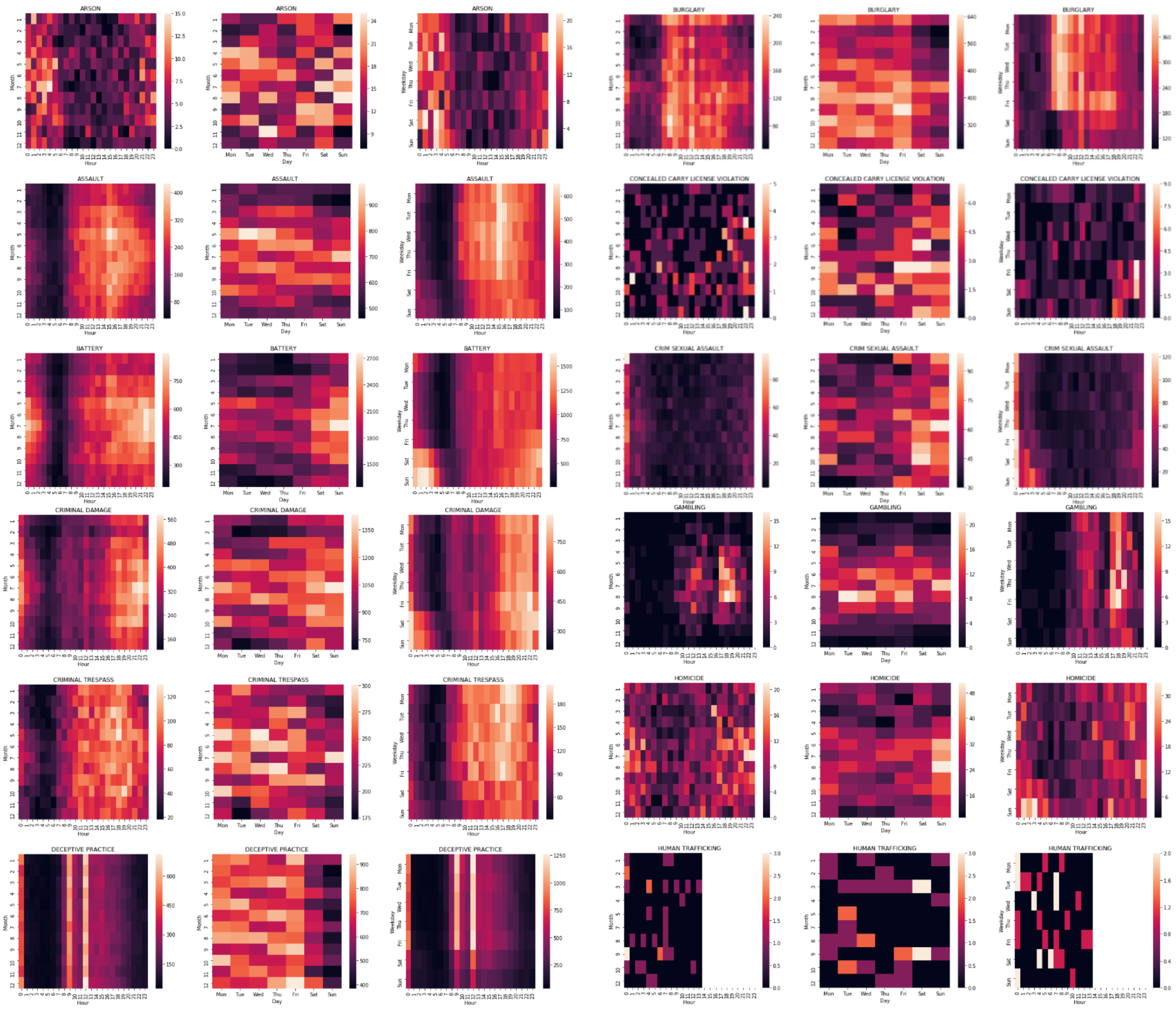
* Crime which H_0 was rejected.

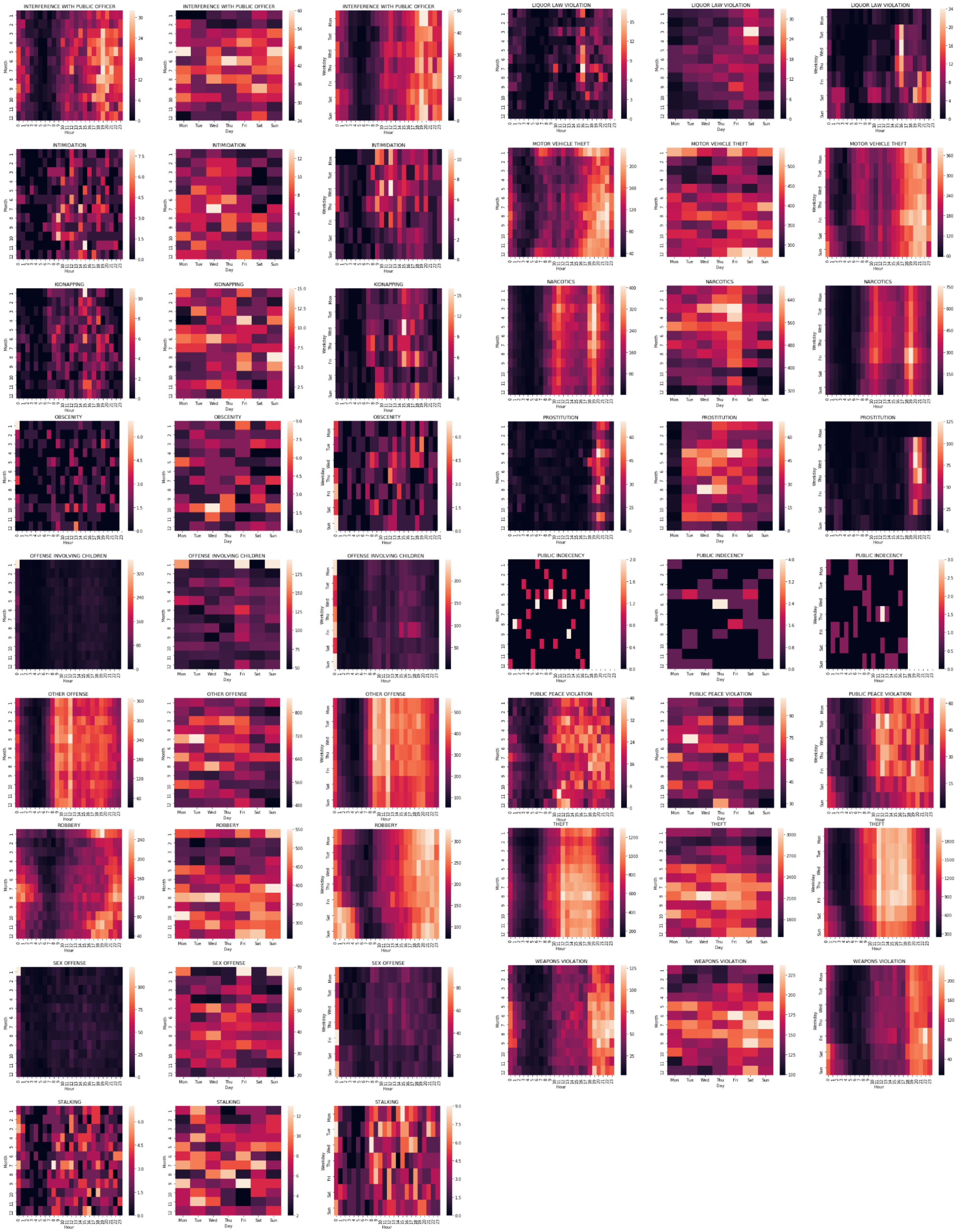
Note: Non-rounded critical χ^2 was used for the test

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. 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.




INTERFERENCE WITH PUBLIC OFFICER
INTERFERENCE WITH PUBLIC OFFICER
INTERFERENCE WITH PUBLIC OFFICER
LIQUOR LAW VIOLATION
LIQUOR LAW VIOLATION
LIQUOR LAW VIOLATION
INTIMIDATION
INTIMIDATION
INTIMIDATION
MOTOR VEHICLE THEFT
MOTOR VEHICLE THEFT
MOTOR VEHICLE THEFT
KIDNAPPING
KIDNAPPING
KIDNAPPING
NARCOTICS
NARCOTICS
NARCOTICS
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
Month
Hour
Day
Weekday
Mon
Tue
Wed
Thu
Fri
Sat
Sun