Analysis of Crime Data in Chicago

DSO 510 - Business Analytics (16323)
Professor Mohammed Salem Alyakoob

Team 3
Fall Semester 2021

Chinmayi Bengaluru Prakash
Daniel Strangio
Hemanth Mallagatta Ravishankar
Naveen Kumar Manjunatha
Sravanthi Kuchibhotla
Vicky Choi

Phase 1: Problem Definition and Scope

Problem Statement

"Our objective is to use data driven approach to make Chicago a safer place by identifying factors that contribute to crime to assist in deploying first responder resources more effectively."

To Fulfill Our Problem Statement, Our Approach Will Be The Following:

- *First, we investigate using inferential visualizations to determine different patterns and trends from the data to provide insights on how when and where crimes occur.*
- *Second, we determine whether occurrence of an arrest is influenced by predictors namely, time of day and type of crime.*
- *Third, we perform regression analysis to predict Monthly Average Total Crime Incidents and determine whether a crime incident will lead to an arrest or not.*

Leveraging the Data Generating Process To Understand Chicago Crime

Sources For Our Data and Research

- **Chicago Police Department:**
  Crime Statistics | Chicago Police Department
- **Chicago Crime Data:**
  https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data

Data Generating Process: Leveraging the Data Collected by the Chicago Police Dept.

Chicago in 2020

- Chicago is home to a population of 2,747,356.

Compared to 2019:

➔ Shootings are up 39.5% YoY.

➔ Shooting victims are up 41% YoY.

➔ Murders are up 36% YoY.

Our Dataset

- Our dataset is a 3-year compilation of crime information (Nov '18 - Oct '21).
  - 685,150 Rows of crime data, each row representing a unique crime incident.
  - 22 Columns of Information
- Categorical variables include Crime Description, Location Description, Community, Area, etc.

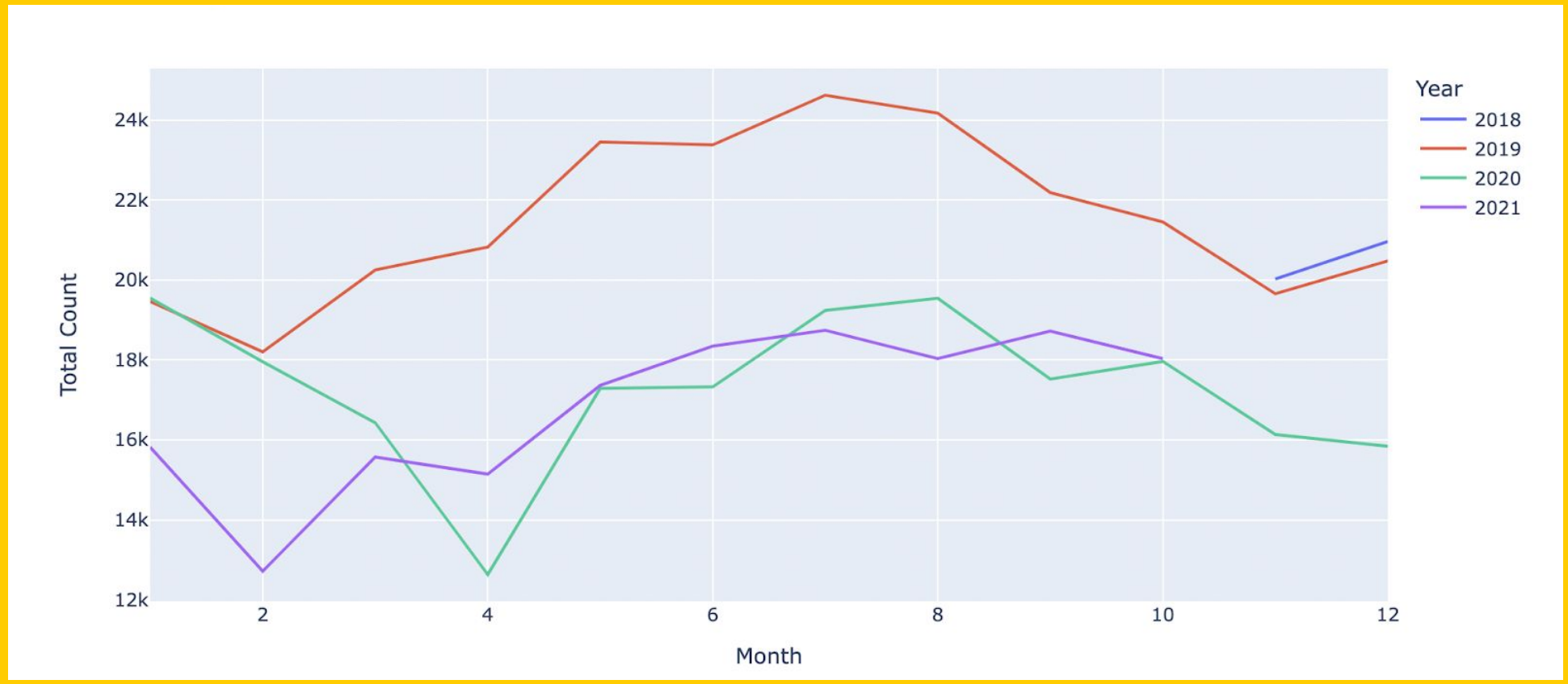
Representation of Our Dataset

Case Number	Date	Block	IUCR	Primary Type	Description	Location Description
JC216048	11/01/2018 12:00:00 AM	062XX S MARSHFIELD AVE	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	APARTMENT
JC216504	11/01/2018 12:00:00 AM	076XX S PAXTON AVE	1130	DECEPTIVE PRACTICE	FRAUD OR CONFIDENCE GAME	RESIDENCE
JC175542	11/01/2018 12:00:00 AM	052XX S MAY ST	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	OTHER
JC297842	11/01/2018 12:00:00 AM	081XX S INGLESIDE AVE	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	APARTMENT

Domain Expertise: Chicago's Finest, Chicago PD

Visualizing Crime Counts on a Line Chart

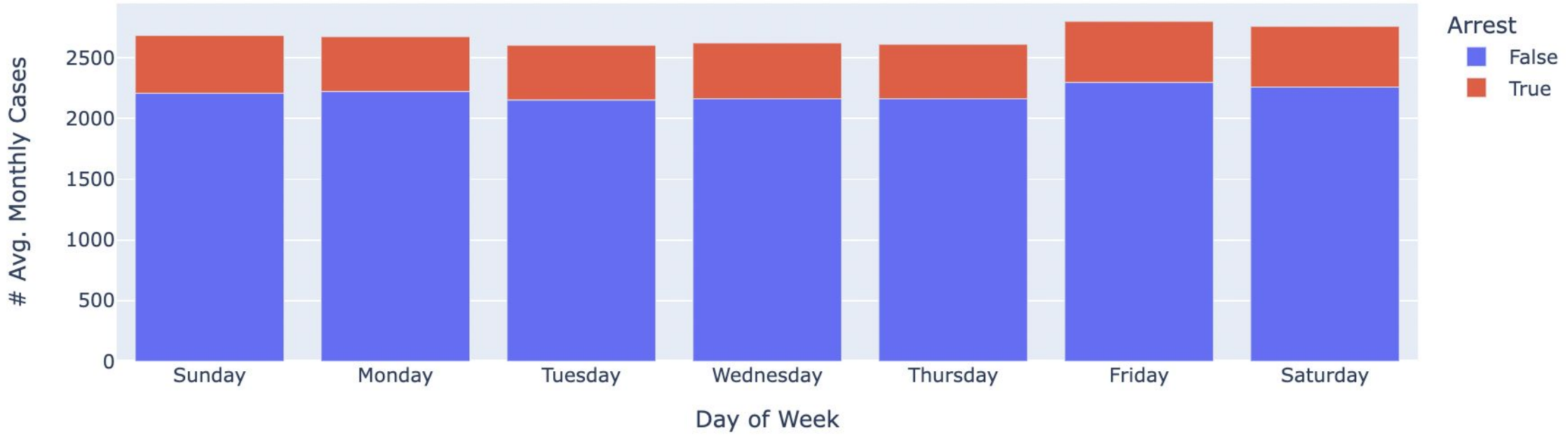
Crime Patterns by Month of the Year



Phase 2: EDA and Hypothesis Testing

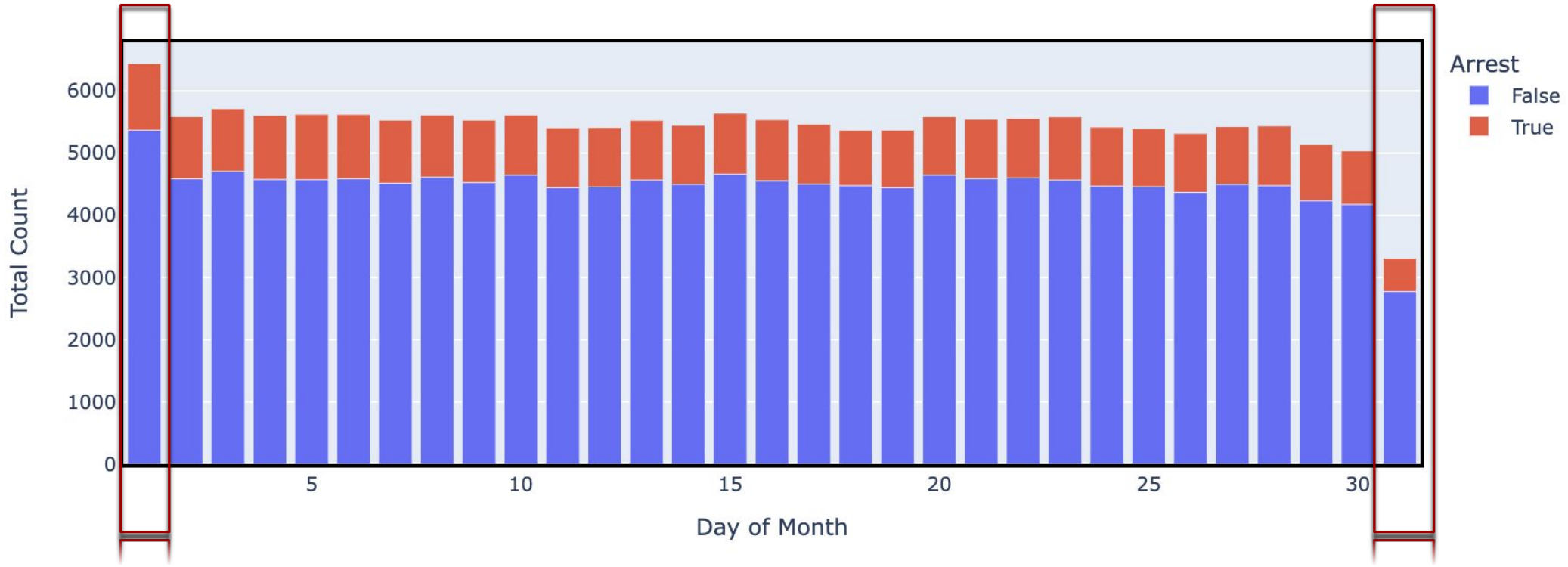
Visualizing # Cases by Day of Week Using a Bar Plot

On average, Fridays and Saturdays have a higher number of crimes committed that lead to arrests, which could be due to an increase in people attending late Night Parties / Events.



Visualizing # of Cases by Day of Month Using a Bar Plot

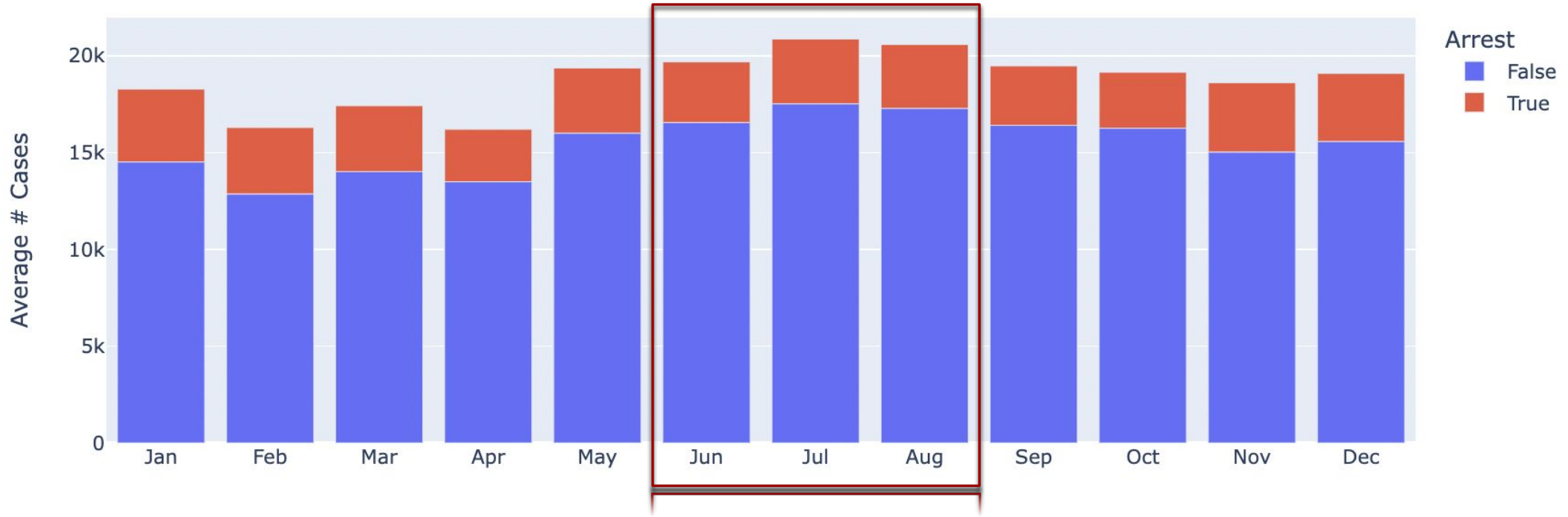
Based on the bar graph, the daily number of crime incidents is uniformly distributed. Day 1 has a higher case count than the rest, indicating a possible issue with the data generating process.



** The last day is the lowest as not all months have 31 days*

Using Bar Graph to Visualize # of Cases and Arrests

Based on the bar graph, the highest average # of cases occur during the summer months. Crime appears to be a seasonal offense and a social event. Warmer weather in the downtown area might enable people into situations where they are interacting more often, usually with friends.



Depicting Crime Frequency Across Chicago Using Heatmaps

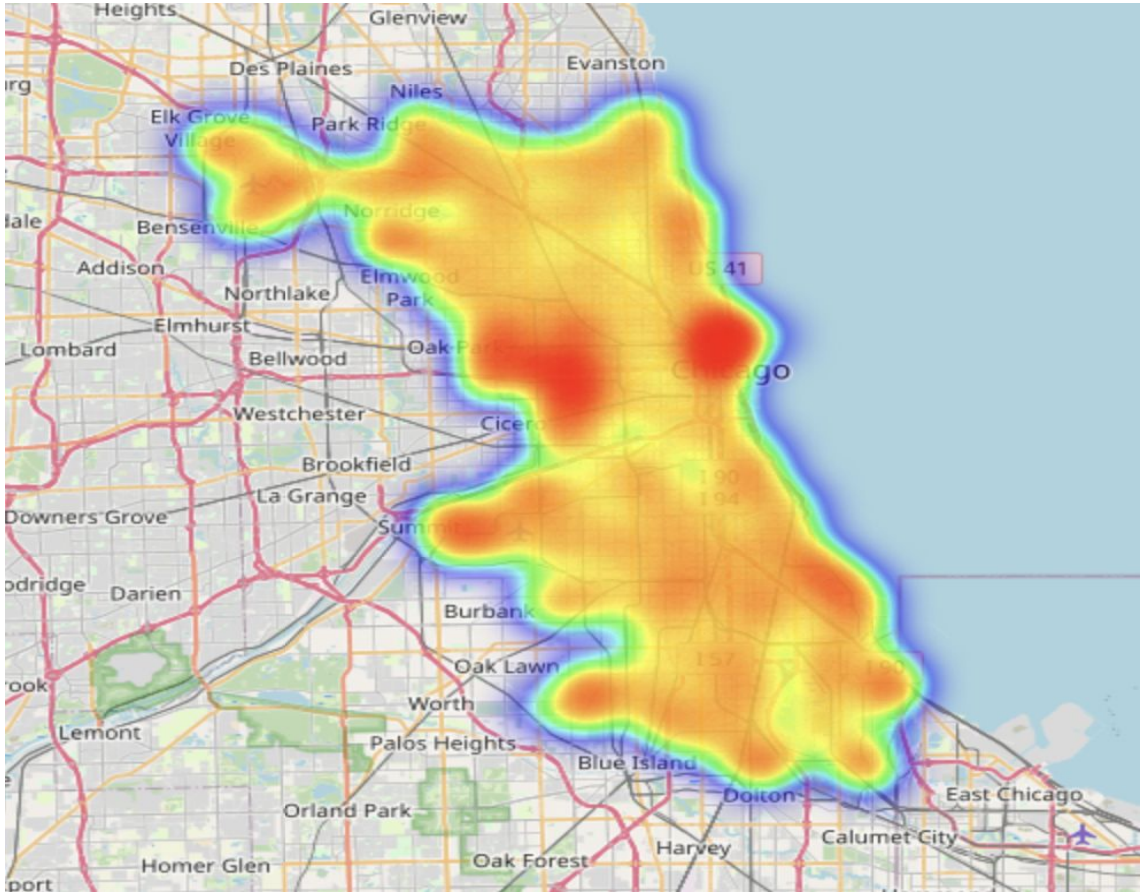
Concentrating Resources By Ward Based on Severe Crime Density

Heatmap *depicts the frequency of crimes across Wards in Chicago between Nov '18 and Oct '21*

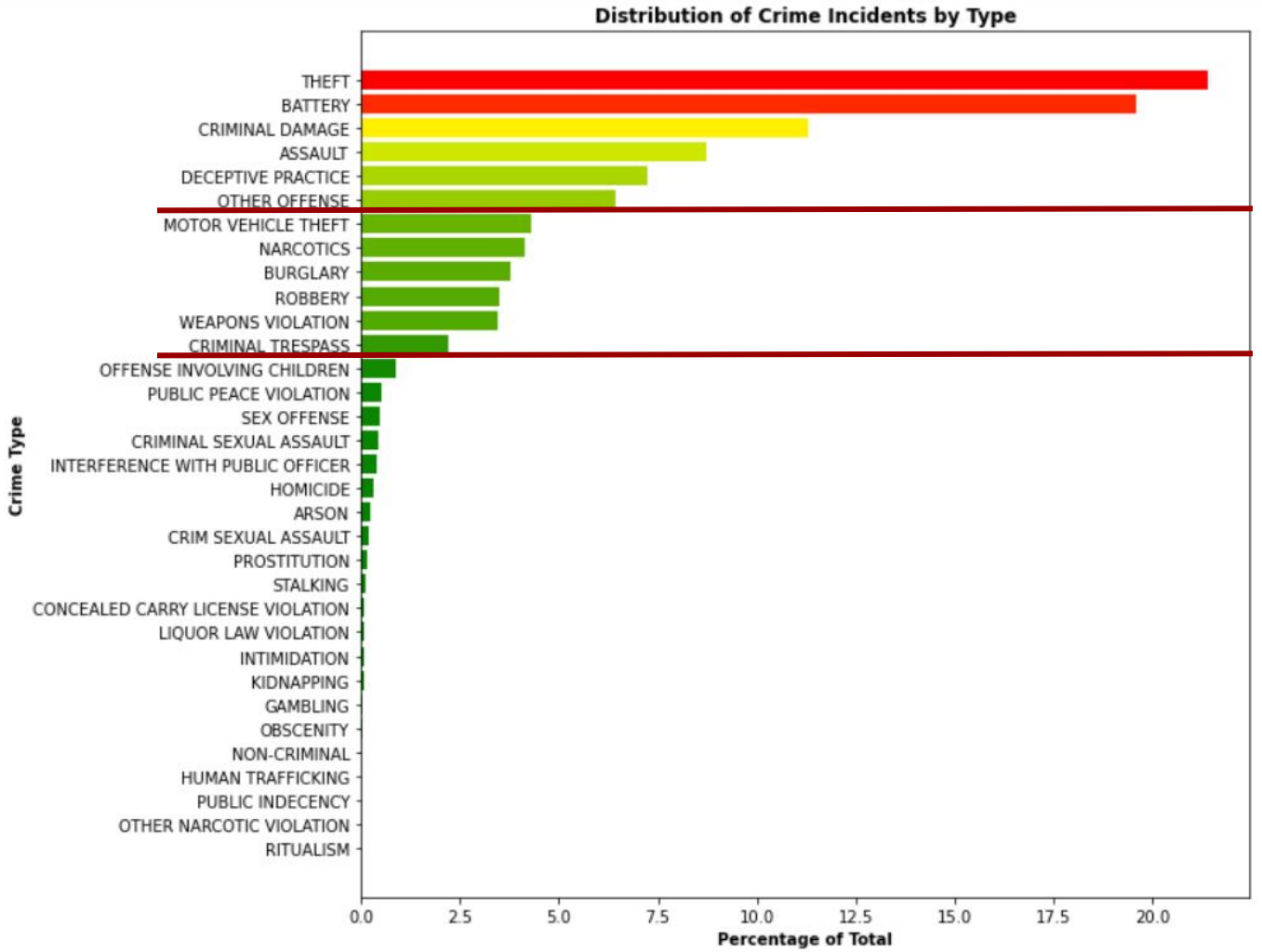
- Crime rates are higher in and around the downtown region.

Future Intent

- *Dynamic mapping of crime incidents, by severity, enable First Responder Leadership to plan resource deployment requirements.*
- *Proactive policing can also deter or mitigate criminal escalation if deployment of resources (critical mass) is more efficient.*
- *Data collection and trend analysis, can turn dynamic mapping into future AI driven crime forecaster.*



Why Use A Data Driven Approach?



High Frequency

Medium Frequency

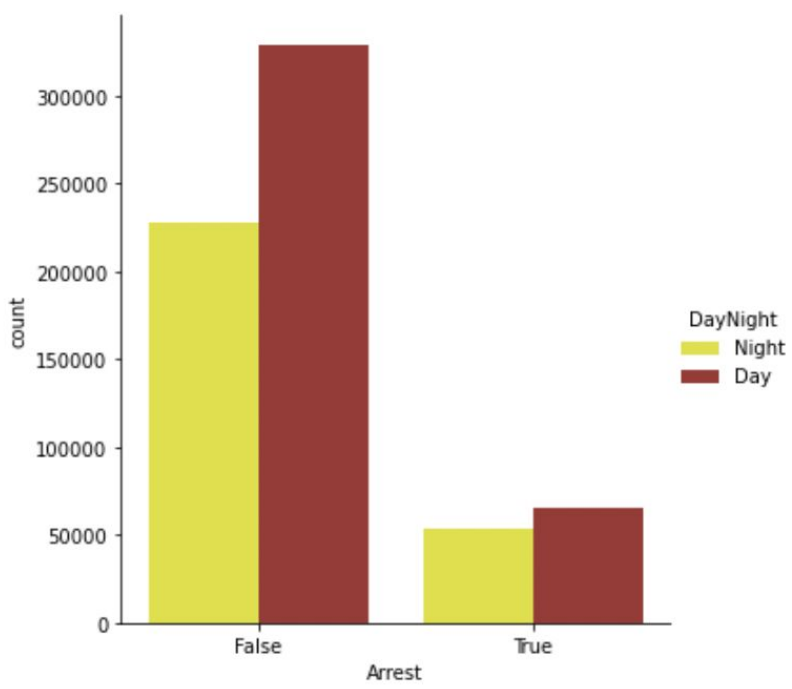
Low Frequency

- We separated Crime Incidents into Categories of Low, Medium, and High based on Frequency.
- We use a data driven approach as a useful tool to highlight where the preponderance of crime occurs.
- This enables First Responder Leaders to plan how best to deploy police forces across the Wards of the City. The alternative is an approach based on guesswork and heuristics.

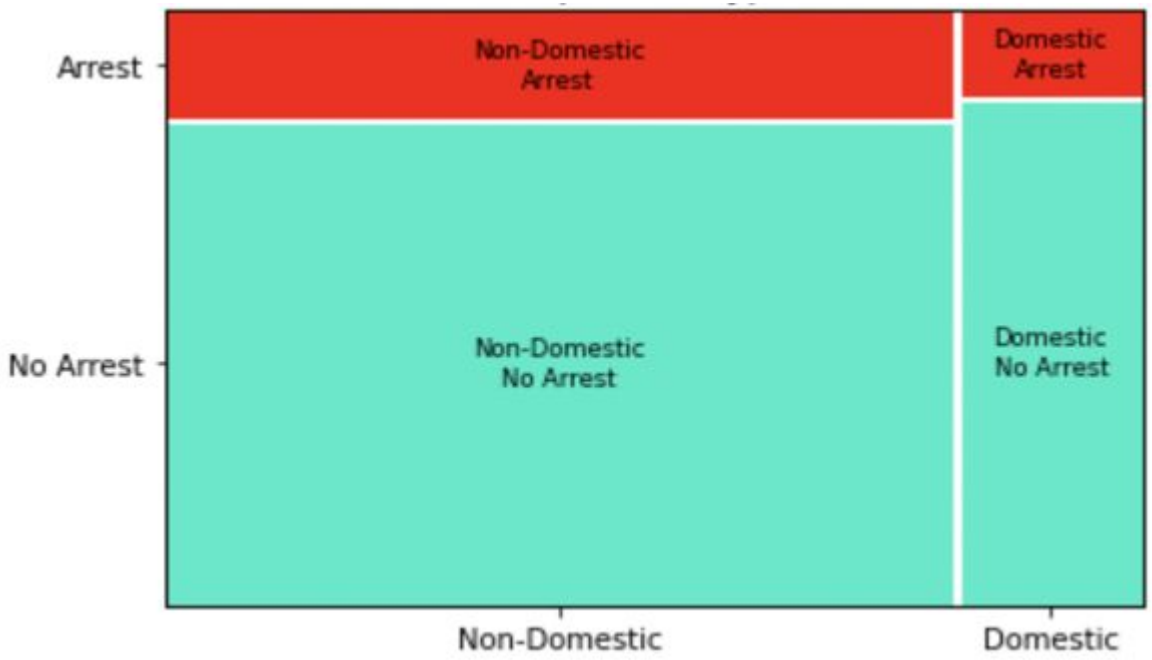
When to Deploy First Responder Resources

Statistically Relevant Predictor Variables for Improving Model Fidelity

Arrest Vs. Time of Day



Arrest By Type of Incident



Two Sample Hypothesis Testing

Testing the Influence of Factors on Occurrences of Arrest

i) Night Vs. Day

Ho: $p\ Arrest_{Night} <= p\ Arrest_{Day}$

Ha: $p\ Arrest_{Night} > p\ Arrest_{Day}$

Test Results:

Z observed = 24.48

P < 0.05

Conclusion:

At 95% CI, since p-value is less than 0.05 we can reject the null hypothesis and conclude the that the proportion of incidents that lead to an arrest is greater at night than during the day.

ii) Domestic Vs. Non-Domestic

Ho: $p\ Arrest_{Non\text{-}Domestic} <= p\ Arrest_{Domestic}$

Ha: $p\ Arrest_{Non\text{-}Domestic} > p\ Arrest_{Domestic}$

Test Results:

Z observed = 33.01

P < 0.05

Conclusion:

At 95% CI, since p-value is less than 0.05 we can reject the null hypothesis and conclude the that the proportion of arrests is greater in non-domestic incidents than domestic incidents.

Phase 3: Predictive Modeling

Logistic Regression Model:
Objective: Predicting if an Incident will Lead to an Arrest or Not

Defining Our Variables

Dependent Variable

- **Arrest (Y/N)**

Independent Variables

- **Day (Day of the month)**
- **DayNight_Night**
- **Domestic_Domestic**
- **Dayofweek_Monday**
- **Dayofweek_Tuesday**
- **Dayofweek_Wednesday**
- **Dayofweek_Thursday**
- **Dayofweek_Friday**
- **Dayofweek_Saturday**
- **Frequency_Low**
- **Frequency_Mid**
- **Severity_Low**
- **Severity_Mid**

Explanation of Variables

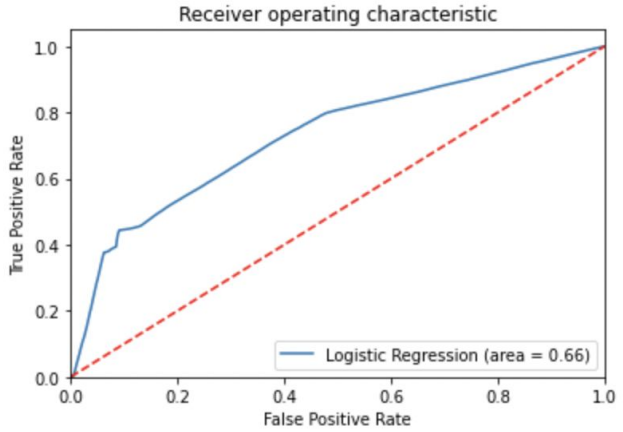
- DayNight_Night indicates the crime incidents that took place between 6pm to 6am, and the rest were tagged to be DayNight_Day
- Domestic indicates the crime incidents that took place indoors (Apartments, Residence, Parking etc.)
- Dayofweek indicates the day of occurrence of the crime incident
- Frequency is based on the Distribution of Crime incidents by type (Refer slide 12)
- Severity is based on the crime levels set by the state (source: CRIME SEVERITY LEVELS)

Logistic Regression Model Results

Employing a **Logistic Regression Model** to Predict Whether a Crime Incident Leads to an Arrest or Not

```
Optimization terminated successfully.
         Current function value: 0.602024
         Iterations 6
                         Results: Logit
==================================================================
Model:               Logit             Pseudo R-squared:  0.131
Dependent Variable:  Arrest            AIC:               937881.2444
Date:                2021-12-05 20:50  BIC:               938031.5980
No. Observations:    778918            Log-Likelihood:    -4.6893e+05
Df Model:            12                LL-Null:           -5.3990e+05
Df Residuals:        778905            LLR p-value:       0.0000
Converged:           1.0000            Scale:             1.0000
No. Iterations:      6.0000
------------------------------------------------------------------
                      Coef.  Std.Err.      z      P>|z|   [0.025  0.975]
------------------------------------------------------------------
Day                  -0.0051   0.0002  -20.5596  0.0000  -0.0056 -0.0046
DayNight_Night        0.0147   0.0049    3.0014  0.0027   0.0051  0.0242
Domestic_Domestic     0.0285   0.0066    4.3384  0.0000   0.0156  0.0414
dayofweek_Friday     -0.1521   0.0079  -19.1924  0.0000  -0.1676 -0.1366
dayofweek_Monday     -0.1999   0.0081  -24.7073  0.0000  -0.2157 -0.1840
dayofweek_Saturday   -0.1332   0.0079  -16.7720  0.0000  -0.1488 -0.1176
dayofweek_Thursday   -0.1588   0.0081  -19.6013  0.0000  -0.1747 -0.1429
dayofweek_Tuesday    -0.1508   0.0081  -18.6270  0.0000  -0.1667 -0.1349
dayofweek_Wednesday  -0.1471   0.0081  -18.2241  0.0000  -0.1630 -0.1313
Frequency_Low         1.1913   0.0113  104.9884  0.0000   1.1691  1.2135
Frequency_Mid         0.4638   0.0082   56.8515  0.0000   0.4478  0.4798
Severity_Low         -0.5478   0.0053 -103.3946  0.0000  -0.5582 -0.5375
Severity_Mid          1.5156   0.0108  139.9895  0.0000   1.4944  1.5368
==================================================================
```

	Precision	Recall	F1-Score
0	0.62	0.87	0.72
1	0.78	0.46	0.57
Accuracy			0.66
Macro Avg	0.7	0.66	0.65
Weighted Avg	0.7	0.66	0.65



Multiple Linear Regression: Predicting Monthly Total Arrests

Defining Our Variables

Aggregated Dataset:

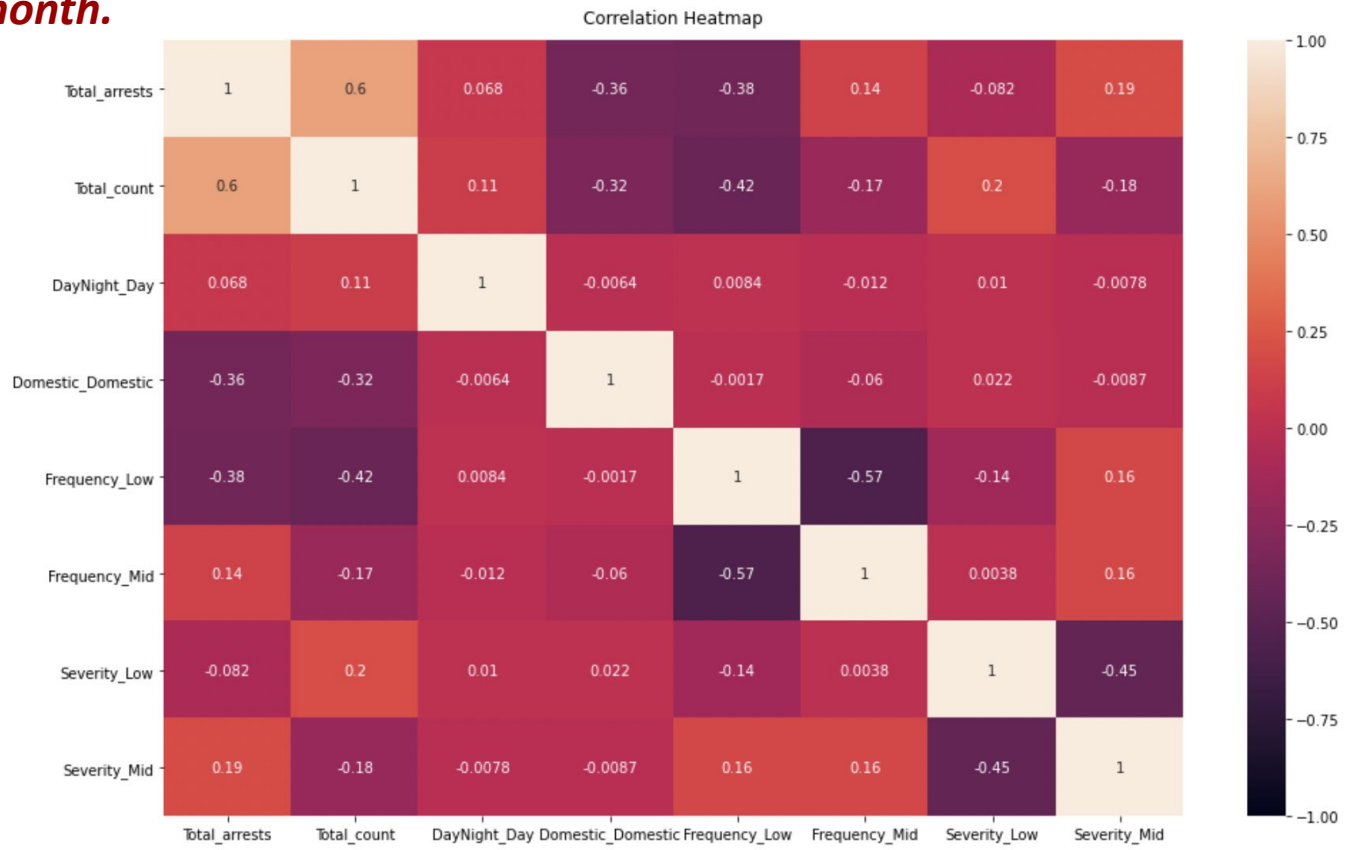
The parent dataset is now aggregated at a monthly level to predict average number of arrests per month.

Dependent Variables

- **Total Arrests**

Independent Variables

- **Total_count (incidents per month)**
- **DayNight_Day**
- **Domestic_Domestic**
- **Frequency_Low**
- **Frequency_Mid**
- **Severity_Low**
- **Severity_Mid**



Looking at Total Arrests as a Dependent Variable, Frequency_Low and Domestic_Domestic are highly correlated.

Multiple Linear Regression Model Results

Predicting # Crime Incidents (Viz. Results Before & After Eliminating Non-Significant Variables)

OLS Regression Results

Dep. Variable:	Total_arrests	**R-squared:**	0.502
Model:	OLS	**Adj. R-squared:**	0.497
Method:	Least Squares	**F-statistic:**	104.0
Date:	Sun, 05 Dec 2021	**Prob (F-statistic):**	7.74e-105
Time:	20:51:01	**Log-Likelihood:**	-4565.8
No. Observations:	729	**AIC:**	9148.
Df Residuals:	721	**BIC:**	9184.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	69.5293	18.030	3.856	0.000	34.131	104.928
Total_count	0.0958	0.007	13.387	0.000	0.082	0.110
DayNight_Day	6.4319	9.569	0.672	0.502	-12.355	25.219
Domestic_Domestic	-59.3948	10.975	-5.412	0.000	-80.941	-37.849
Frequency_Low	-50.4224	17.404	-2.897	0.004	-84.590	-16.255
Frequency_Mid	41.7816	16.458	2.539	0.011	9.470	74.093
Severity_Low	-43.4933	11.086	-3.923	0.000	-65.258	-21.729
Severity_Mid	103.8221	13.349	7.777	0.000	77.614	130.031

High p-value indicates that DayNight_Day is non-significant

OLS Regression Results

Dep. Variable:	Total_arrests	**R-squared:**	0.502
Model:	OLS	**Adj. R-squared:**	0.498
Method:	Least Squares	**F-statistic:**	121.3
Date:	Sun, 05 Dec 2021	**Prob (F-statistic):**	8.43e-106
Time:	20:51:01	**Log-Likelihood:**	-4566.0
No. Observations:	729	**AIC:**	9146.
Df Residuals:	722	**BIC:**	9178.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	71.4314	17.800	4.013	0.000	36.485	106.378
Total_count	0.0965	0.007	13.609	0.000	0.083	0.110
Domestic_Domestic	-59.0398	10.958	-5.388	0.000	-80.553	-37.527
Frequency_Low	-49.1374	17.292	-2.842	0.005	-83.085	-15.189
Frequency_Mid	42.6980	16.395	2.604	0.009	10.510	74.886
Severity_Low	-43.7491	11.075	-3.950	0.000	-65.493	-22.006
Severity_Mid	103.3355	13.325	7.755	0.000	77.176	129.495

Omnibus:	297.783	**Durbin-Watson:**	2.042
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1803.982
Skew:	1.731	**Prob(JB):**	0.00
Kurtosis:	9.885	**Cond. No.**	7.38e+03

VIF Values

Variance inflation factor (VIF) is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis. All of our explanatory variables have a VIF score lower than 5, which means there is low correlation among the independent variables.

Features	VIF
Severity_Low	2.08
Severity_Mid	1.87
Frequency_Low	1.75
Frequency_Mid	1.72
Domestic_Domestic	1.5

Recommendations

Based on our findings .. .

Summer months of June, July and August, along with evening hours, concentrated on non-domestic related incidents appear to materialize in an elevated number of incidents leading to arrests made, in the Downtown Ward area.

In an effort to optimize the safety and welfare of citizens and property, we recommend that the Police leverage predictive modeling to estimate the rates of criminal activities.

Academy training cycles and staffing hours of personnel should be provisioned proportionally to handle the critical mass of incidents.