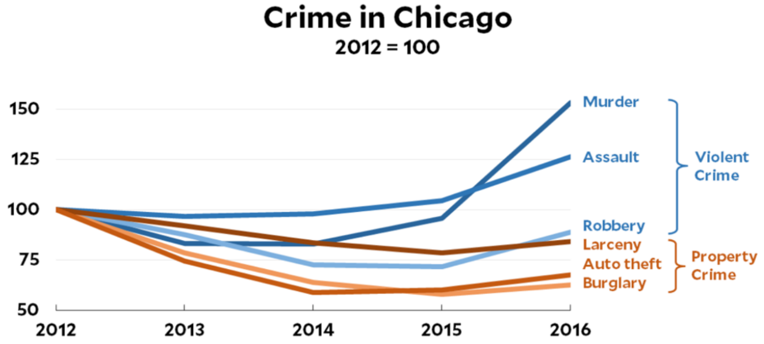
Abstract

Chicago and other major cities continue to get a lot of attention as they struggle with high crime rates. Chicago is currently ranked as one of the most crime-affected cities in the U.S and has witnessed a 50% increase in its murder rate in a year period. Law enforcement agencies usually face a complex problem when trying to efficiently assign resources to minimize crime overtime and across different city locations. The following project aims to review different data mining methods to effectively visualize crime trends and locations. Additionally, taking into account that crime analysis and prediction is a delicate subject, this report also intends to promote a discussion about community engagement and other crime reduction measures.

Introduction

Crime analysis includes looking at data from two important dimensions: space and time. Space involves observing the characteristics of a particular region and surrounding areas. On the other hand, time involves observing and analyzing an event overtime. Using real crime data for the city of Chicago and several data mining models, we will explore different scenarios to provide options to address a real-world problem.

In this project, we will analyze Chicago crime data to understand different crime trends and identify relevant hot spots. By analyzing several characteristics of particular regions and observing criminal incidents overtime, we will attempt to formulate a solution and we will provide several recommendations. For this purpose, several models were trained and then were tested to guide us through our effort of finding effective predictive policing strategies. There has been a lot of debate regarding this topic due to ethical concerns and the risk of targeting vulnerable areas. Further discussion will be provided in the conclusions.



Source: FBI Uniform Crime Reporting

Problem Statement and Data Sources

In order to identify crime trends overtime and popular crime locations, we would need to formulate and explore several important questions. As previously stated, it would be important to analyze the data from spatial and temporal dimensions. Understanding the nature of the problem would be critical to design a workable solution within a given set of constrains. Did crime increase or decrease overtime? Which month shows the most criminal activity? Which location reports the most crimes? These are just a number of relevant questions that would guide us through our mission of analyzing Chicago crime data to solve a complex problem: trying to efficiently assign resources to minimize crime overtime and across different city locations.

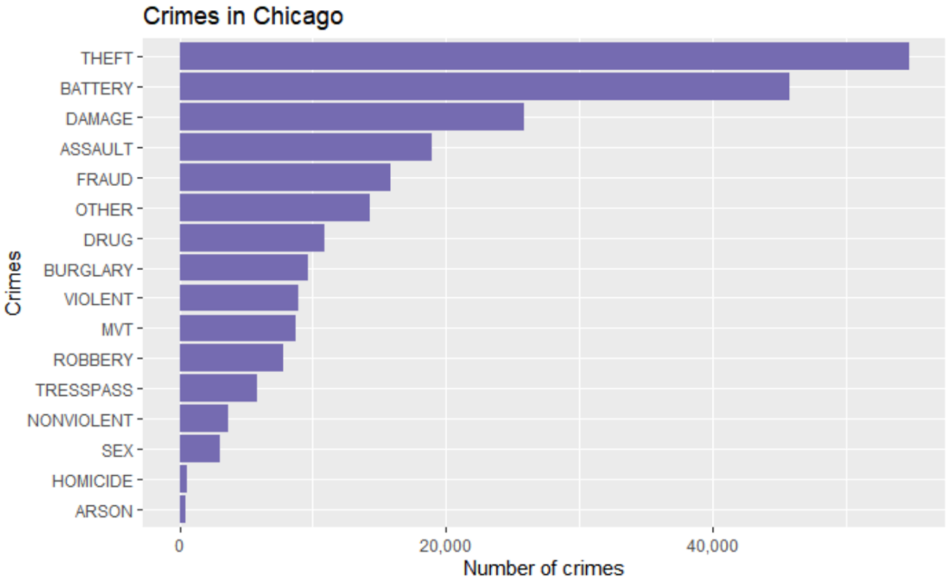
For this project, we will use crime data for the city of Chicago, which is publicly available from their open data portal.
https://data.cityofchicago.org/.
The city of Chicago documents reported incidents of crime and the information is available from the year 2001 onwards.
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data
For the data related to crime location, hot spots and the Chicago police departments beats.
https://data.cityofchicago.org/Public-Safety/BoundariesPolice-Beats/kd6k-pxkv.

The Chicago crime dataset is organized with 22 variables. In our process to analyze this dataset, some fields will be more important than others. Some relevant attributes include:
id = Unique identifier for the record
case_number = The Chicago Police Department Records Division Number, unique to the incident.
date = Date when the incident occurred.
block = Partially redacted address where the incident occurred.
iucr = Illinois Unifrom Crime Reporting code (directly linked to primary_type and description)
primary_type = The primary description of the IUCR code.
location_description = Description of the location where the incident occurred.
arrest = Indicates whether an arrest was made.
domestic = Indicates whether the incident was domestic-related.
beat = Indicates the police beat where the incident occurred.
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.

Data Pre-Processing and Exploratory Data Analysis

It's important to understand how the data is organized and what fields are present in the table. We started by investigating the internal structure and doing some initial analysis. Our dataset originally included 22 variables; however, to find an accurate prediction model, we needed to remove unnecessary features with missing information. The variables with most NA values, with near zero variance and redundant IDs were also removed. For this dataset, each incident has a unique identifier and it's stored in the "*case_number*" variable. The "*case_number*" variable should have all unique values; however, after doing some exploration, we can see that some instances are duplicated. These duplicated rows need to be removed. Other preprocessing work included formatting the "*date_of_occurence*" variable, so it recognizes the values as date & time and not just characters.

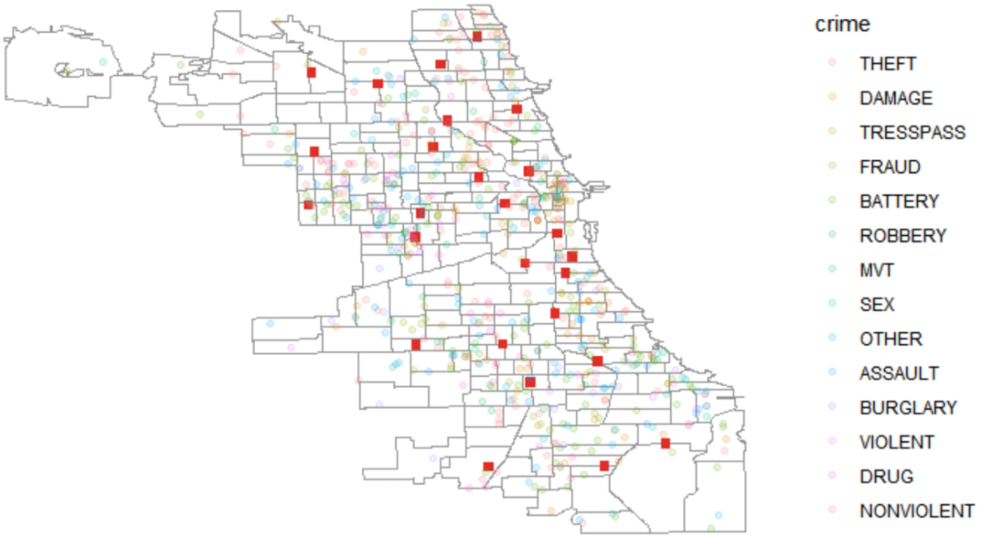
The data contains 32 crime types. I order to simplify the analysis process, it was necessary to merge similar categories. After some initial review, we can see that there is a prevalence of different crimes with theft and battery being more frequent.



The frequency of crimes is not really consistent throughout the day. There are usually peaks of crime activity during different time frames. In order to better understand the time window of criminal activity, it was necessary to group the time stamps related to a crime. The distribution of crime incidents across the day suggests that crimes are more frequent during the latter half of the day. Refer to *Figure 1 on Appendix (Only relevant plots will be included in the body of the report).*

We can use the date of incidence to determine which day of the week as shown in *Figure 2 on Appendix*, and which month of the year the crime occurred. Refer to *Figure 3 on Appendix*. It is possible that there is a pattern in the way crimes occur depending on the day of the week and month. With respect to the time dimension, it seems that later during the day, end of the week, and summer months experienced more criminal activity.

As previously discussed, crime analysis includes looking at data from two important dimensions: space and time. We just did an initial review observing and analyzing crime events overtime. However, we also need to study the space element which involves observing the characteristics of a particular region and surrounding areas. Crimes vary considerably with respect to geographies, usually there will be zones that will observe higher criminal activity compared to the others. These areas are often the focus for effective predictive policing. With the location of each incident, we can observe criminal patterns in the city of Chicago.



The multiple dots across the city of Chicago represent different crimes. The red squares on the plot represent the police stations. A map with just the police stations can be reviewed in *Figure 4 on Appendix.* A quick analysis indicates that several police stations are located around active crime areas.

Additionally, as part of our preliminary analysis, we reviewed some basic summary statistics of our predictors as shown on *Figure 5 on Appendix*.

Proposed Methodology

In this project, we are going to compare the performance of different data mining methods for our crime prediction problem. First, we need to determine the level of data in which we are going to build the model and the independent variables that would better assist with our process. For each recorded crime incident in the city of Chicago, we have the location, date, type, beat and ward. The location information is very specific, but it's important to understand that not all data can be transformed to be part of the solution. Understanding the nature of the problem is critical for designing a solution considering our set of constraints.

We are aiming to assist with predictive policing, and we need to construct our models with a focus on the geographical area and time. Our area could be a block or a zip code, and the time frame could be a day or a month. In order to achieve this goal, we will work with the following models:

Multivariate Regression: This model is an extension of multiple regression with one dependent variable and multiple independent variables. Based on the number of independent variables, we try to predict the output. This model was selected because of its simplicity and general understanding. In terms of accuracy, a better solution could be to deal with different crime types separately and generate predictions for smaller time periods.

Poisson Regression: This model is best used for modeling events where the outcomes are counts. More specifically, to count data like non-negative integer values that count something, the number of times an event occurs during a given timeframe or the number of people that were victims of a crime.

Multivariate Adaptive Regression Splines: A non-parametric regression method that builds multiple linear regression models across the range of predictor values. The MARS algorithm is an extension of linear models that makes no assumptions about the relationship between the response variable and the predictor variables.

Recursive Partitioning and Regression Tree: A non-parametric model, widely used in regression and classification problems. It's used to identify groups of observations with similar values of parameters of the model of interest. The iterative process splits the data into partitions.

Analysis and Results

In this section, we will execute the mentioned data mining methods and will explore the key findings of our results. Our modeling dataset has all the crime incidents that were recorded in the past twelve months, but during this time period, there were locations with no criminal activity. For these observations, we will indicate that there were no crimes and no arrests. As previously discussed, there are fluctuations in crime incidents depending on days of the week or months; therefore, using those two variables as predictor could play an important role in our models. The "*beat*" variable has some missing values, but just removing the complete rows would lead to loss of important information. As a solution, it was necessary to impute the missing observations with appropriate estimates. For modeling purposes, several predictor variables were created, including interactions. It was necessary to get a set of variables that explains the variation in the dependent variable. The *Figure 6 on Appendix* describes the correlation between the dependent variable and the independent variables.

As part of our analysis, we will check the performance of our models on our testing dataset. First, we would use the root mean squared error (RMSE) to decide if we should add an additional variable. We will compare their scores to the actual values by using the (RMSE). This metric will help us review how far we are on average from the actual values. The first score was RMSE = 2.1721; however, after including a relevant variable, the score increased to RMSE = 2.5387. For our RMSE, a better performance is indicated by a smaller value, so this result advised us to avoid adding more variables. Additional information can be reviewed in *Figure 7 on Appendix*.

The following tables report the evaluation metrics for each model. The misclassification error indicates the percentage of times that our model fails a prediction. The true positive rate (sensitivity) tells us the percentage that our model accurately makes the prediction.

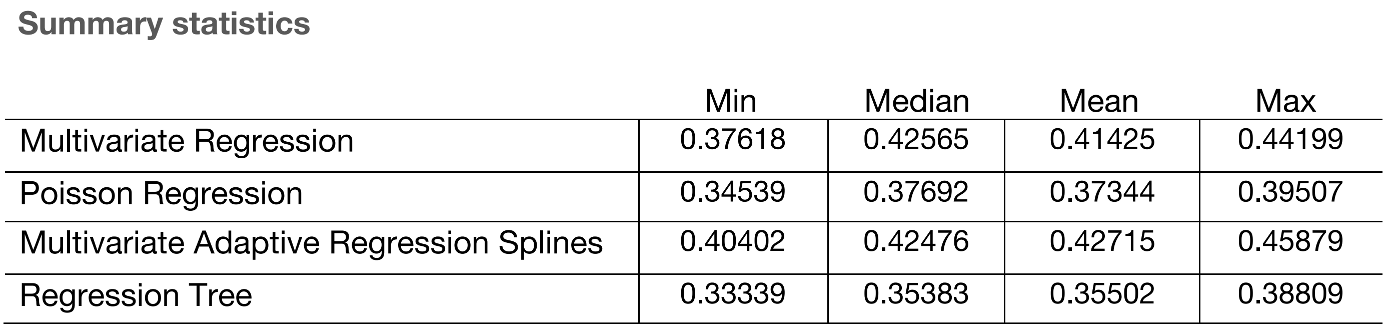
Evaluation metrics

	Testing error	Sensitivity
Multivariate Regression	0.26165	0.71311
Poisson Regression	0.35231	0.51602
Multivariate Adaptive Regression Splines	0.19323	0.77174
Regression Tree	0.20766	0.82335

By evaluating our results, we can observe that Multivariate Adaptive Regression Splines and the Regression Tree models have the best performance. For the misclassification error, the best score will be indicated by a smaller value. It seems that the Multivariate Adaptive Regression Splines outperforms the rest with a testing error of around 19.32%. The next model with a strong performance is the Regression Tree with a testing error of approximately 20.7%, followed by the Multivariate Regression with a testing error of 26.1% and the last one would be the Poisson Regression.

In terms of the true positive rate, we can notice that the Poisson Regression model again have a poor performance. For the sensitivity, we evaluate a good performance based on a larger value. In this case, the Regression Tree has the best performance with a sensitivity rate of 82.3%. It is followed by the Multivariate Adaptive Regression Splines model with a rate of around 77.1% and then by the Multivariate Regression model with a sensitivity rate of approximately 71.3%.

As part of our evaluation, the following table reports the summary statistics based on cross-validation R squared.



Summary statistics

	Min	Median	Mean	Max
Multivariate Regression	0.37618	0.42565	0.41425	0.44199
Poisson Regression	0.34539	0.37692	0.37344	0.39507
Multivariate Adaptive Regression Splines	0.40402	0.42476	0.42715	0.45879
Regression Tree	0.33339	0.35383	0.35502	0.38809

After reviewing the statistics table, it seems that the Multivariate Adaptive Regression Splines model performs the best with a mean of 0.42. By comparing different metrics, the Multivariate Adaptive Regression Splines model and the Regression Tree model achieve the best results. However, it’s important to note that if we care about the true positive rate, which I believe we should because we are dealing with crime prediction, the Regression Tree model has the strongest performance.

Conclusions

As previously acknowledged, predictive policing is a controversial and debatable topic. I took the challenge of carefully reviewing the available information prior to starting the "*Chicago Crime Prediction and Analysis Project*". I understand that the context behind crime and policing goes beyond the realm of data mining and statistical learning. The purpose behind these results is to just provide a source of information, so law enforcement agencies can efficiently manage their resources. I believe that machine learning technology should be just another tool available, and not a substitute for community engagement and other crime reduction measures.

It has been a lengthy process. We handled, cleaned, and explored crime data. We built and tested predictive models; however, there still are some issues and improvements that we need to address. The discussed models attempted to predict expected number of crimes in different locations in the city of Chicago. These strategies provide a general guidance, but there are several limitations. Crimes tend to follow a pattern during a particular day, week, or month; therefore, the results could be static and, in the end, not really helpful. In order to have more accurate results, an option could be to explore smaller time intervals. Another problem is that these models would not have the ability to differentiate the gravity among crimes. All crimes are not equal and violent crimes need special attention, so those should be separated from non-violent activity. Additionally, being too specific regarding locations, could create the risk of ignoring other areas that should also be getting attention.

As previously discussed, I hope this project would encourage others to start a discussion about ethical concerns in data analysis. I think that these conversations could motivate analysts to explore different ideas to make machine learning technologies more inclusive, and then use these tools to improve the safety of our communities.

Lessons Learned

After a thoughtful reflection, I believe that selecting such a sensible topic for my project, pushed me to really take into account the ethical aspects of data analytics. My project started with a focus in data mining models for predictive policing; however, after exploring the data and selecting variables, I noticed how important it was to remain as neutral as possible. This project made me realize that it's our duty to responsibly use the technology available to make a positive impact in society.

References

- For this project, we used crime data for the city of Chicago, which is publicly available from their open data portal.
  https://data.cityofchicago.org/.
- The city of Chicago documents reported incidents of crime and the information is available from the year 2001 onwards.
  https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data
- For the data related to crime location, hot spots and the Chicago police departments beats.
  https://data.cityofchicago.org/Public-Safety/BoundariesPolice-Beats/kd6k-pxkv.

- Information regarding concepts and definitions:
  https:/developer.ibm.com/articles/cc-models-machine-learning/ and Wikipedia.

- Information regarding machine learning parameters and statistical learning in r:
  https://www.kaggle.com/camnugent/introduction-to-machine-learning-in-r-tutorial and
  https://cran.r-project.org/web/views/MachineLearning.html

- Information and articles regarding ethics in predictive policing:
  https://www.eff.org/deeplinks/2020/09/technology-cant-predict-crime-it-can-only-weaponize-proximity-policing

- Statistical analysis, models, and visualizations:
  Computer – MacBook Pro macOS Catalina with processor 2.4GHz Intel Core i7
  Software - RStudio Version 1.4.1103. 2009-2021, PBC

Appendix

Graphs and visualizations:

Figure 1.

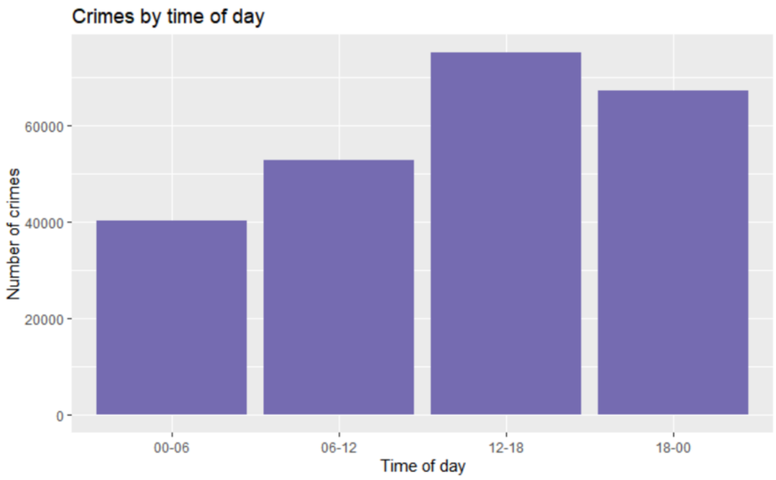


Figure 2.

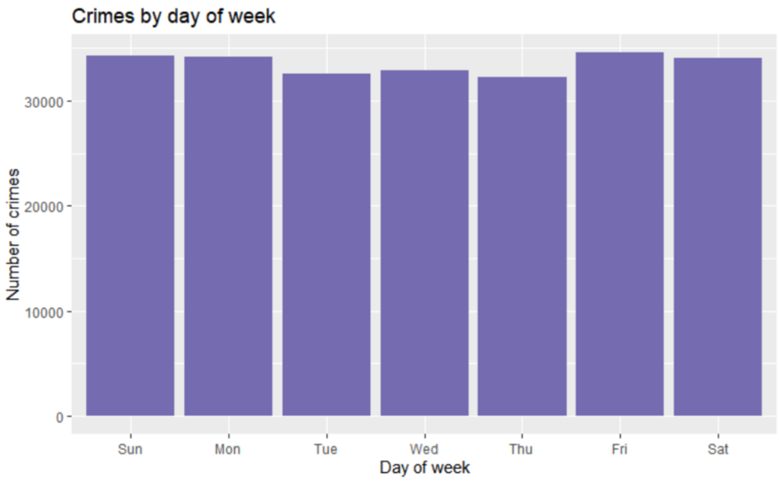


Figure 3.

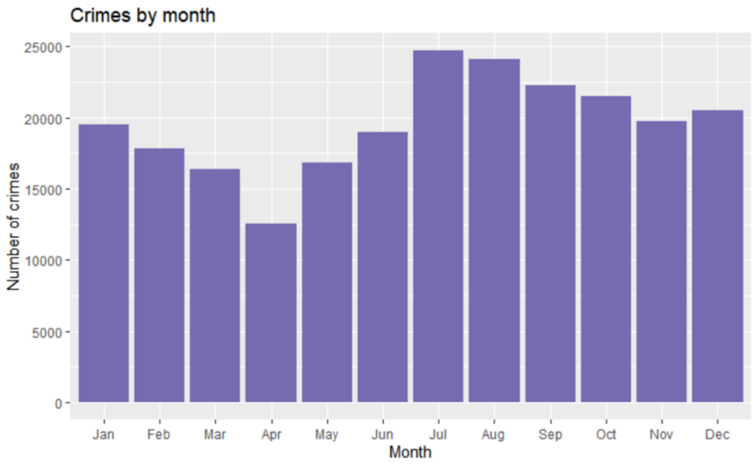


Figure 4.

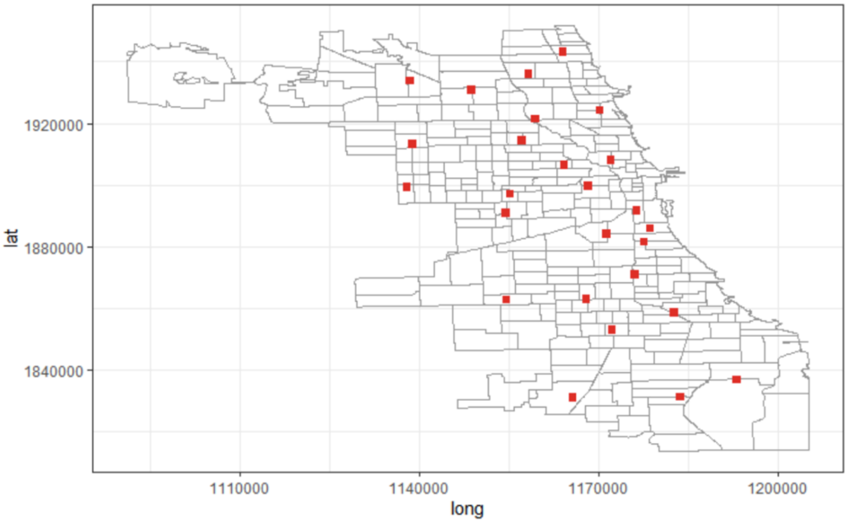
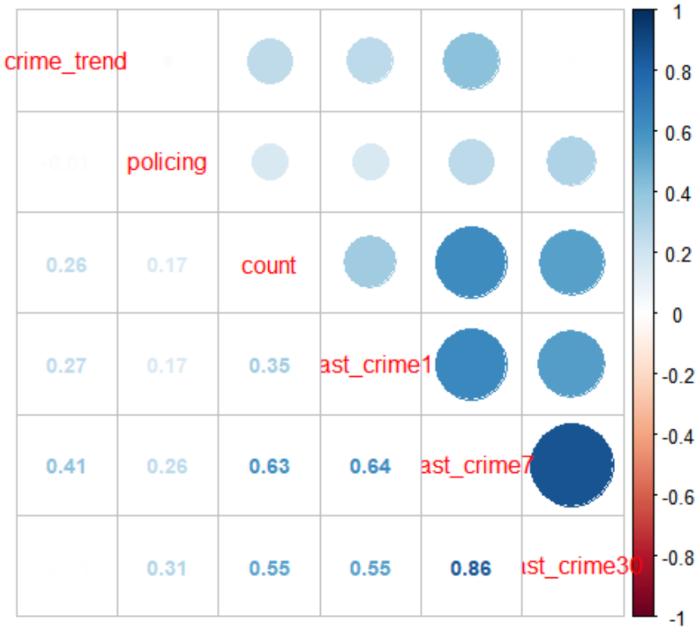
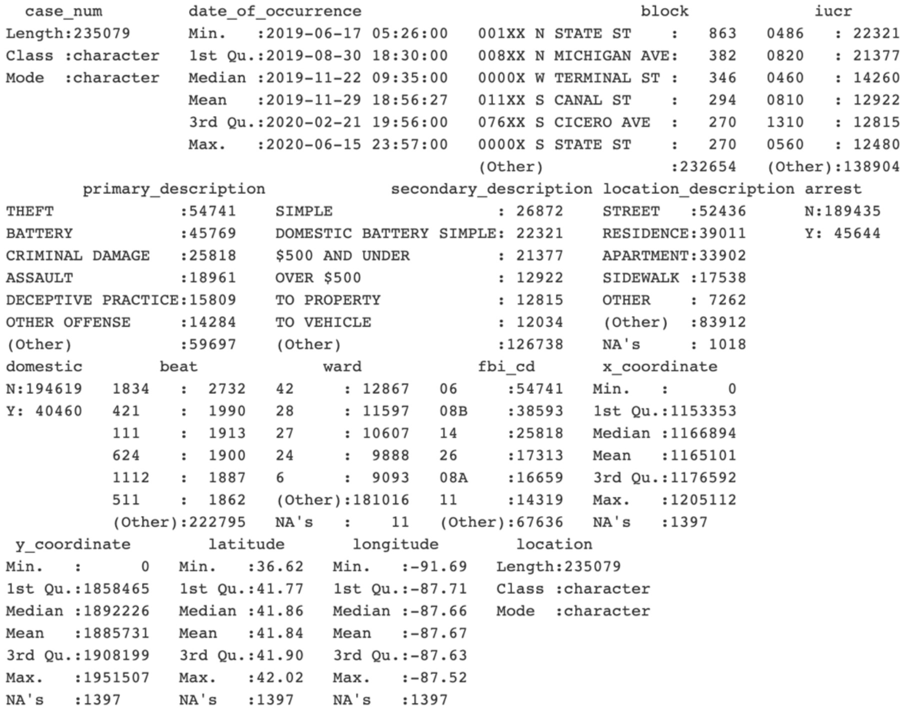


Figure 6.



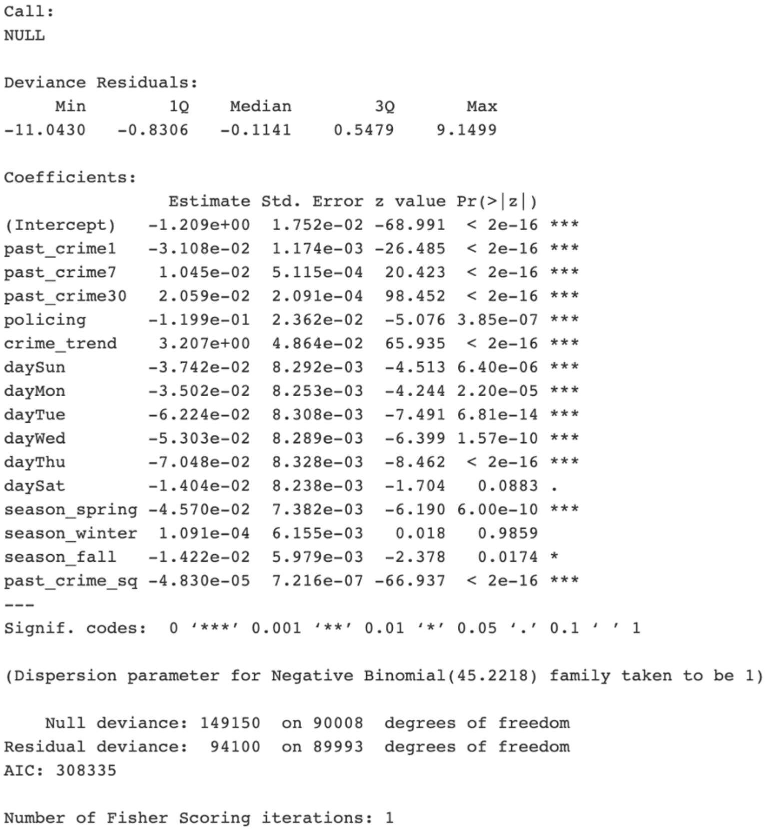
Summary statistics:

Figure 5.



```
  case_num         date_of_occurrence                            block             iucr
Length:235079      Min.   :2019-06-17 05:26:00   001XX N STATE ST    :   863   0486   : 22321
Class :character   1st Qu.:2019-08-30 18:30:00   008XX N MICHIGAN AVE:   382   0820   : 21377
Mode  :character   Median :2019-11-22 09:35:00   0000X W TERMINAL ST :   346   0460   : 14260
                   Mean   :2019-11-29 18:56:27   011XX S CANAL ST    :   294   0810   : 12922
                   3rd Qu.:2020-02-21 19:56:00   076XX S CICERO AVE  :   270   1310   : 12815
                   Max.   :2020-06-15 23:57:00   0000X S STATE ST    :   270   0560   : 12480
                                                 (Other)             :232654   (Other):138904
        primary_description            secondary_description location_description arrest
THEFT             :54741    SIMPLE                 : 26872    STREET   :52436      N:189435
BATTERY           :45769    DOMESTIC BATTERY SIMPLE: 22321    RESIDENCE:39011      Y: 45644
CRIMINAL DAMAGE   :25818    $500 AND UNDER         : 21377    APARTMENT:33902
ASSAULT           :18961    OVER $500              : 12922    SIDEWALK :17538
DECEPTIVE PRACTICE:15809    TO PROPERTY            : 12815    OTHER    : 7262
OTHER OFFENSE     :14284    TO VEHICLE             : 12034    (Other)  :83912
(Other)           :59697    (Other)                :126738    NA's     : 1018
domestic        beat             ward           fbi_cd       x_coordinate
N:194619   1834   :  2732   42     : 12867   06     :54741   Min.   :      0
Y: 40460   421    :  1990   28     : 11597   08B    :38593   1st Qu.:1153353
           111    :  1913   27     : 10607   14     :25818   Median :1166894
           624    :  1900   24     :  9888   26     :17313   Mean   :1165101
           1112   :  1887   6      :  9093   08A    :16659   3rd Qu.:1176592
           511    :  1862   (Other):181016   11     :14319   Max.   :1205112
           (Other):222795   NA's   :    11   (Other):67636   NA's   :1397
 y_coordinate        latitude       longitude        location
Min.   :      0   Min.   :36.62   Min.   :-91.69   Length:235079
1st Qu.:1858465   1st Qu.:41.77   1st Qu.:-87.71   Class :character
Median :1892226   Median :41.86   Median :-87.66   Mode  :character
Mean   :1885731   Mean   :41.84   Mean   :-87.67
3rd Qu.:1908199   3rd Qu.:41.90   3rd Qu.:-87.63
Max.   :1951507   Max.   :42.02   Max.   :-87.52
NA's   :1397      NA's   :1397    NA's   :1397
```

Figure 7.



```
Call:
NULL

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-11.0430   -0.8306   -0.1141    0.5479    9.1499

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.209e+00  1.752e-02 -68.991  < 2e-16 ***
past_crime1   -3.108e-02  1.174e-03 -26.485  < 2e-16 ***
past_crime7    1.045e-02  5.115e-04  20.423  < 2e-16 ***
past_crime30   2.059e-02  2.091e-04  98.452  < 2e-16 ***
policing      -1.199e-01  2.362e-02  -5.076 3.85e-07 ***
crime_trend    3.207e+00  4.864e-02  65.935  < 2e-16 ***
daySun        -3.742e-02  8.292e-03  -4.513 6.40e-06 ***
dayMon        -3.502e-02  8.253e-03  -4.244 2.20e-05 ***
dayTue        -6.224e-02  8.308e-03  -7.491 6.81e-14 ***
dayWed        -5.303e-02  8.289e-03  -6.399 1.57e-10 ***
dayThu        -7.048e-02  8.328e-03  -8.462  < 2e-16 ***
daySat        -1.404e-02  8.238e-03  -1.704   0.0883 .
season_spring -4.570e-02  7.382e-03  -6.190 6.00e-10 ***
season_winter  1.091e-04  6.155e-03   0.018   0.9859
season_fall   -1.422e-02  5.979e-03  -2.378   0.0174 *
past_crime_sq -4.830e-05  7.216e-07 -66.937  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(45.2218) family taken to be 1)

    Null deviance: 149150  on 90008  degrees of freedom
Residual deviance:  94100  on 89993  degrees of freedom
AIC: 308335

Number of Fisher Scoring iterations: 1
```