Northeastern University-CPS

Boston Crime Analysis

Course: ALY6040 20935
Winter Quarter A, 2018

Presented by:

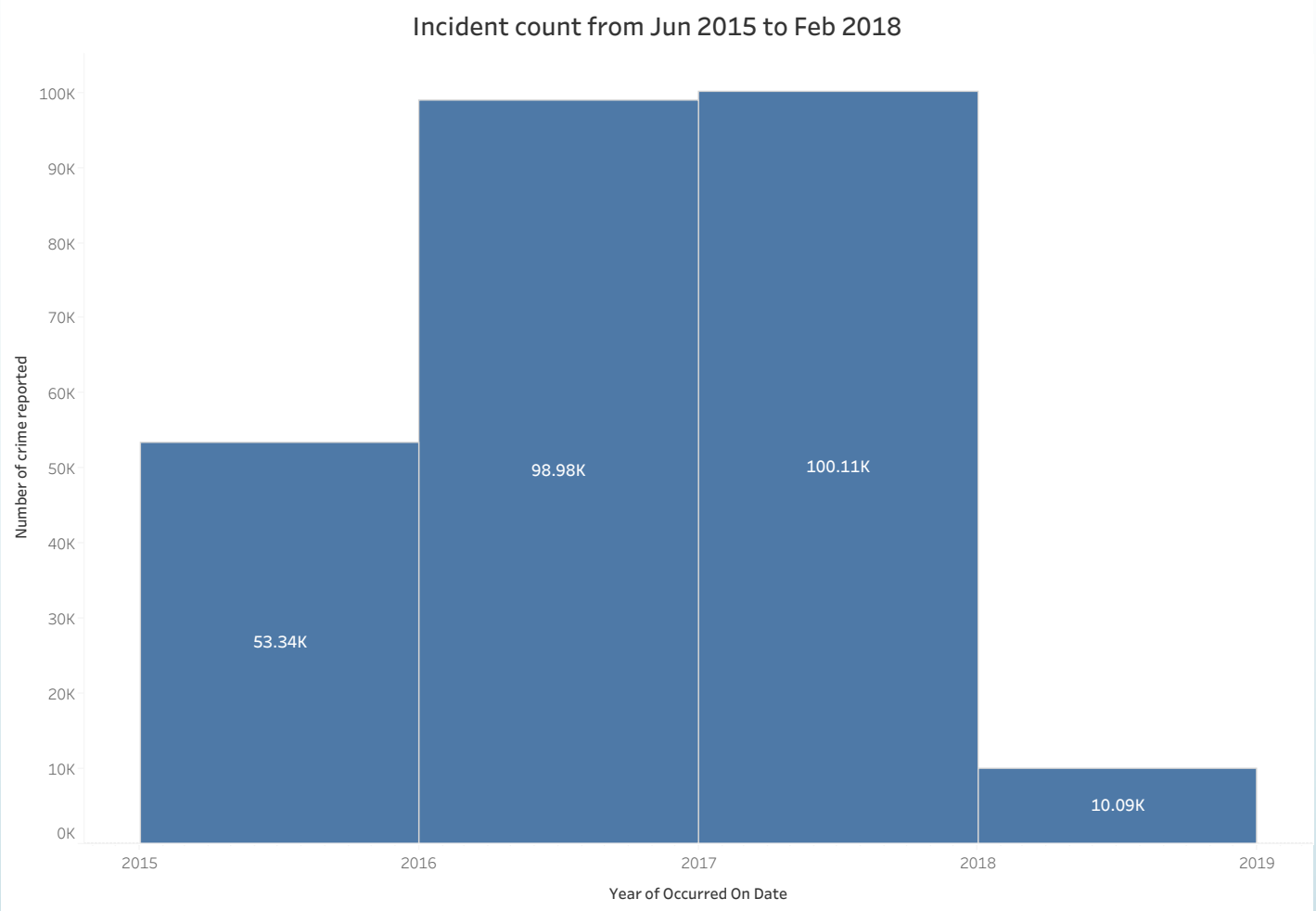
Abhishek Jaiswal, PaoTing Kung

Agenda

- Introduction
- About Data
- Crime Data Analysis
  - Crime Frequency by Category
  - Crime by Time
  - Crime by Week
  - Crime by Month
- Data Modeling
  - Code
  - Model data
  - Evaluation
- Recommendations

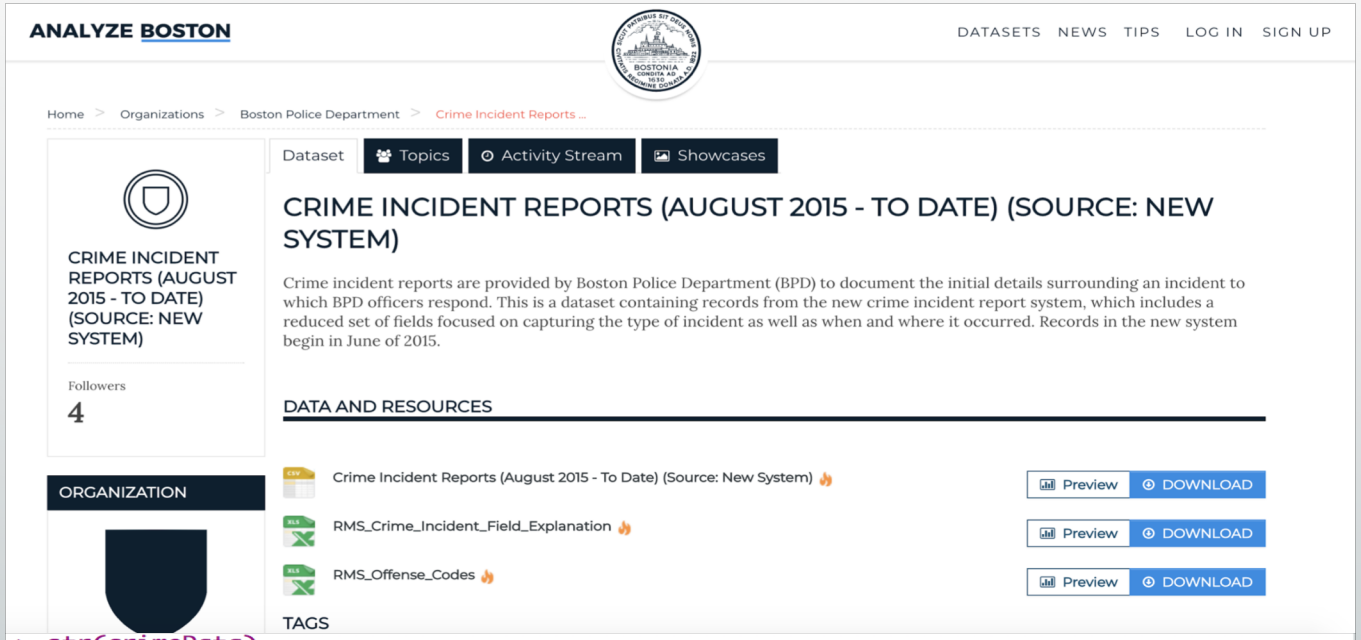
Introduction

- Increasing trend in Crime Rate in Boston
- Need an improvement in Law enforcement
- Involvement of computer-aided technology
- Utilization of vast crime data available
- Use of data Science
- Trend analysis, correlation, & data modeling
- Predictive policing



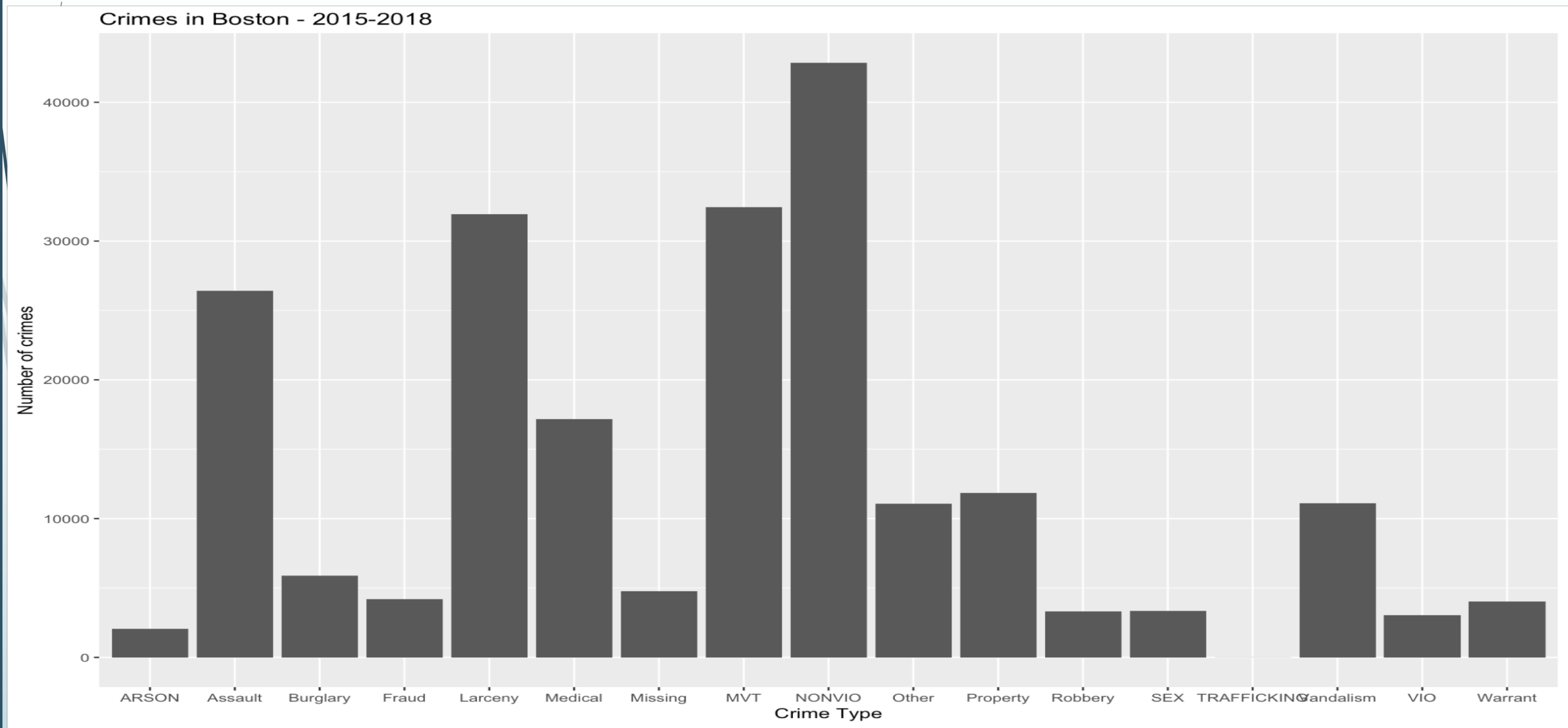
About Data

- Data extracted from “data.boston.gov”
- Total of 17 variables
- Data cleaning
  - Removing duplicate using subset () function
  - Removing records with null values
- Bucketing crime time into four time intervals
- Converting 64 crime categories into 16 types
- Formatting shooting data as “Y” to “1”, and null to “0”

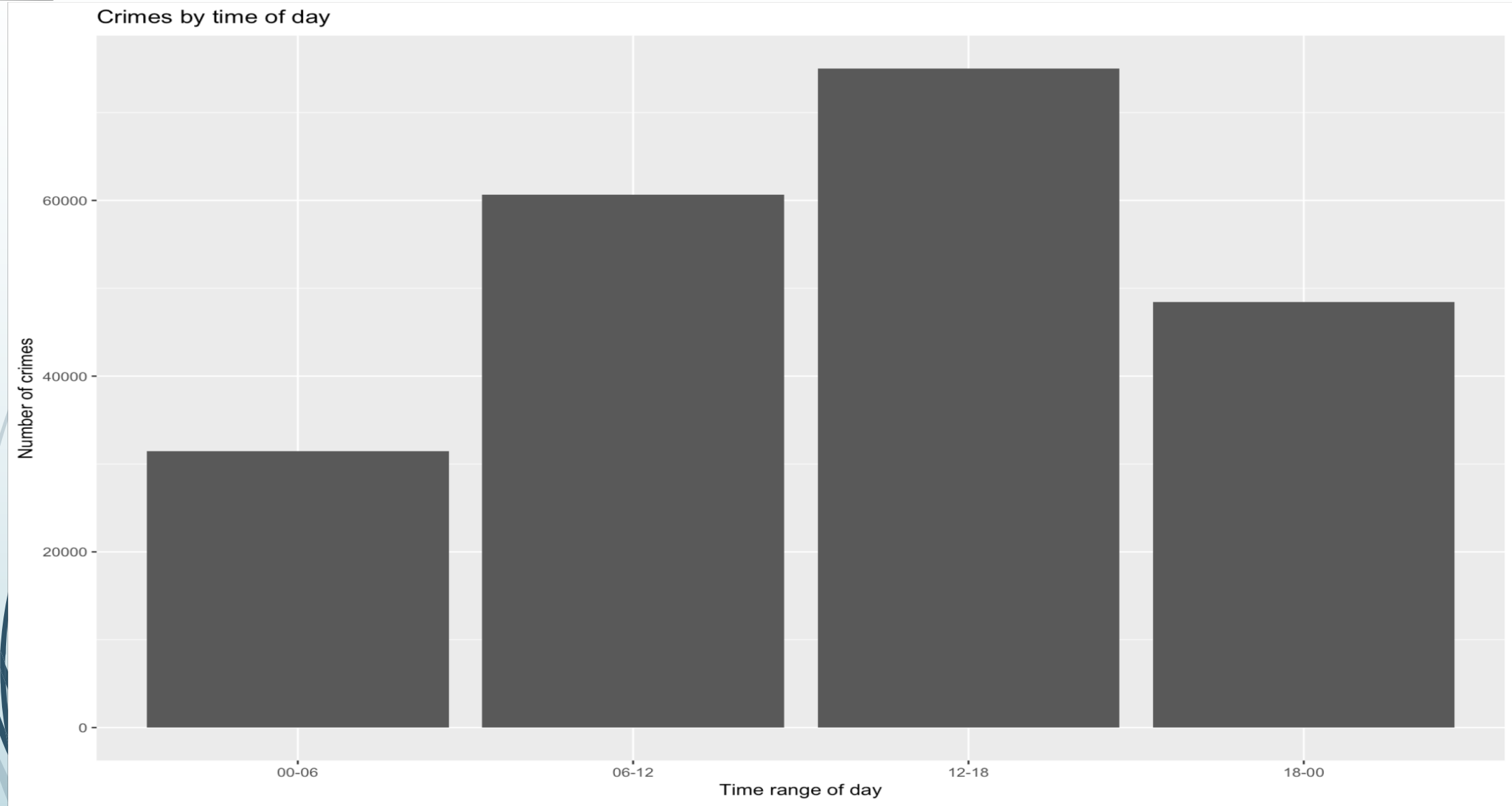



```
> str(crimeData)
'data.frame':   258776 obs. of  17 variables:
 $ INCIDENT_NUMBER    : Factor w/ 229659 levels "142052550","I010370257-00",..: 229659 2
29658 229657 229656 229655 229654 229653 229652 229651 229650 ...
 $ OFFENSE_CODE       : int  617 3109 802 3831 2647 3160 614 3801 3006 3006 ...
 $ OFFENSE_CODE_GROUP : Factor w/ 67 levels "Aggravated Assault",..: 35 50 62 44 47 20 3
6 44 41 41 ...
 $ OFFENSE_DESCRIPTION: Factor w/ 243 levels "A&B HANDS, FEET, ETC.  - MED. ATTENTION RE
Q.",..: 147 213 21 156 221 95 149 160 215 215 ...
 $ DISTRICT           : Factor w/ 13 levels "","A1","A15",..: 9 4 5 1 6 6 7 4 11 5 ...
 $ REPORTING_AREA     : int  796 12 613 NA 428 944 254 24 510 588 ...
 $ SHOOTING           : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1 1 1 1 ...
 $ OCCURRED_ON_DATE   : Factor w/ 189699 levels "2015-06-15 00:00:00",..: 189696 189699
189698 189697 189684 189695 189690 189694 189693 189689 ...
 $ YEAR               : int  2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
 $ MONTH              : int  1 1 1 1 1 1 1 1 1 1 ...
 $ DAY_OF_WEEK        : Factor w/ 7 levels "Friday","Monday",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ HOUR               : int  20 21 21 21 18 20 19 20 20 19 ...
 $ UCR_PART           : Factor w/ 5 levels "","Other","Part One",..: 3 4 5 4 5 4 3 4 4 4
...
 $ STREET             : Factor w/ 4468 levels "","  ALBANY ST ",..: 1944 3617 3189 1 2808
4295 270 2416 4214 1857 ...
 $ Lat                : num  42.3 42.4 42.3 42.3 42.3 ...
 $ Long               : num  -71.1 -71 -71.1 -71.1 -71.1 ...
 $ Location           : Factor w/ 17554 levels "(-1.00000000, -1.00000000)",..: 13582 17
442 11624 10387 3224 4489 7062 16754 5852 10604 ...
```

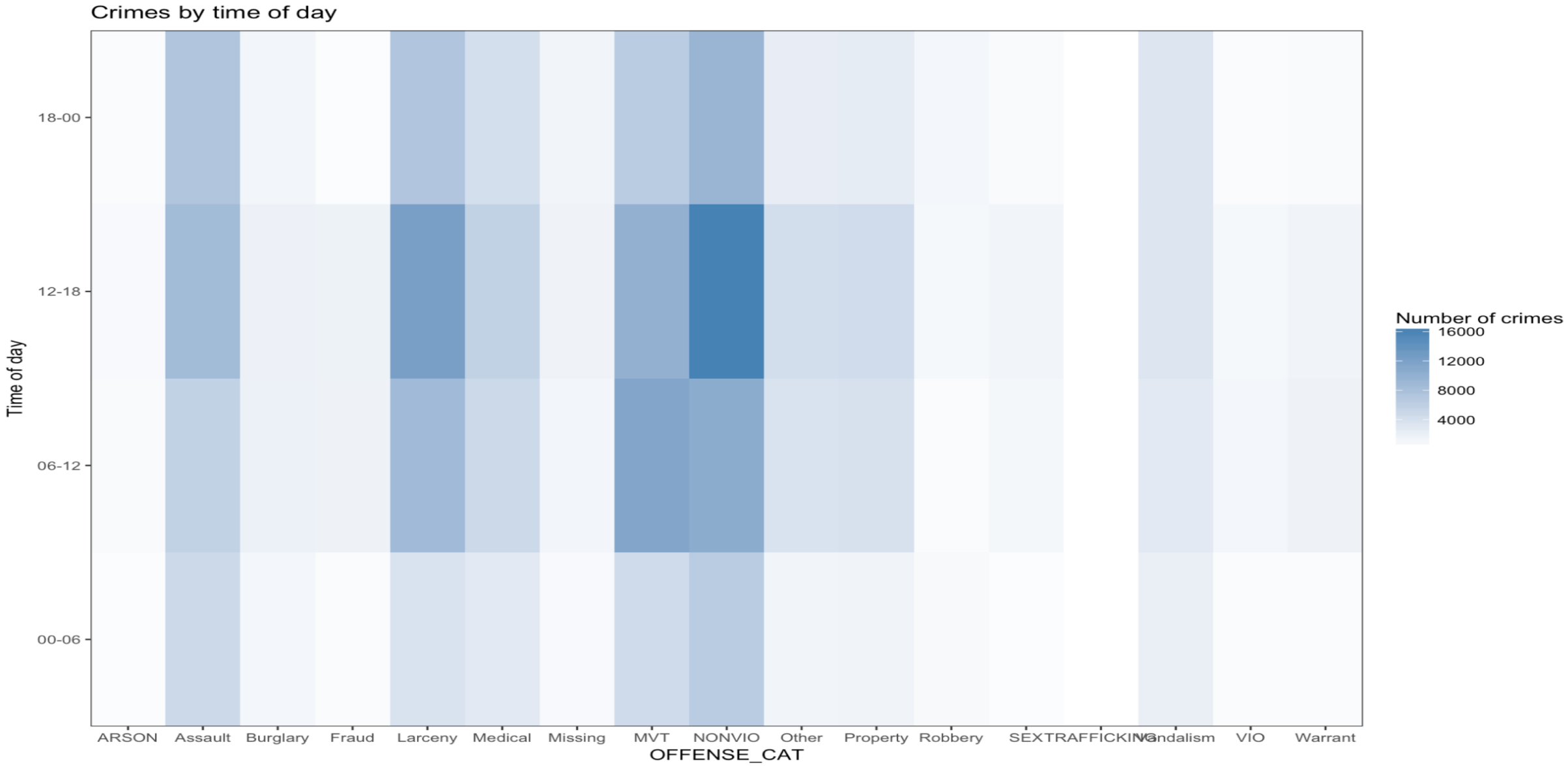
Crime Data Analysis – Crime Frequency



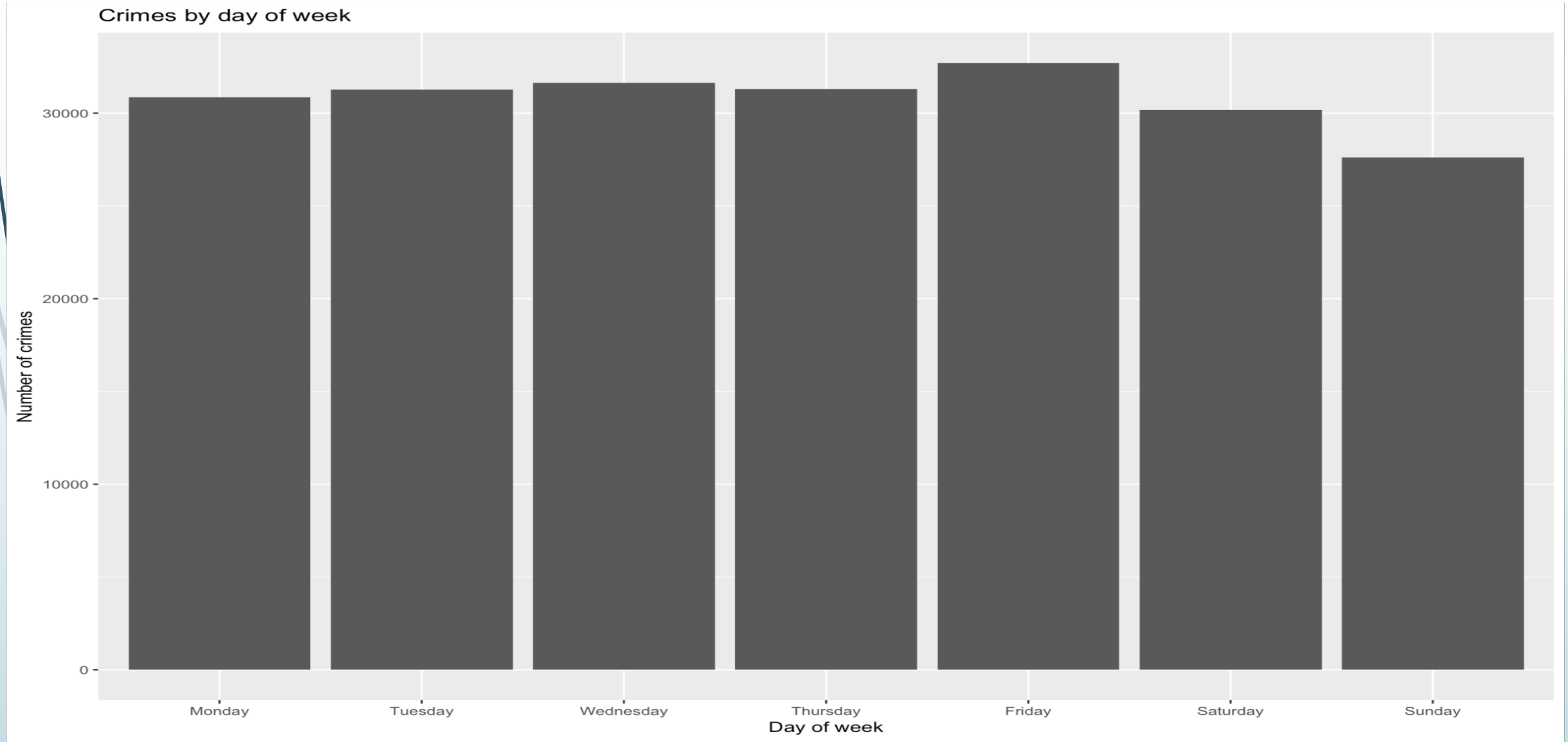
Crime Data Analysis – Crime by Time



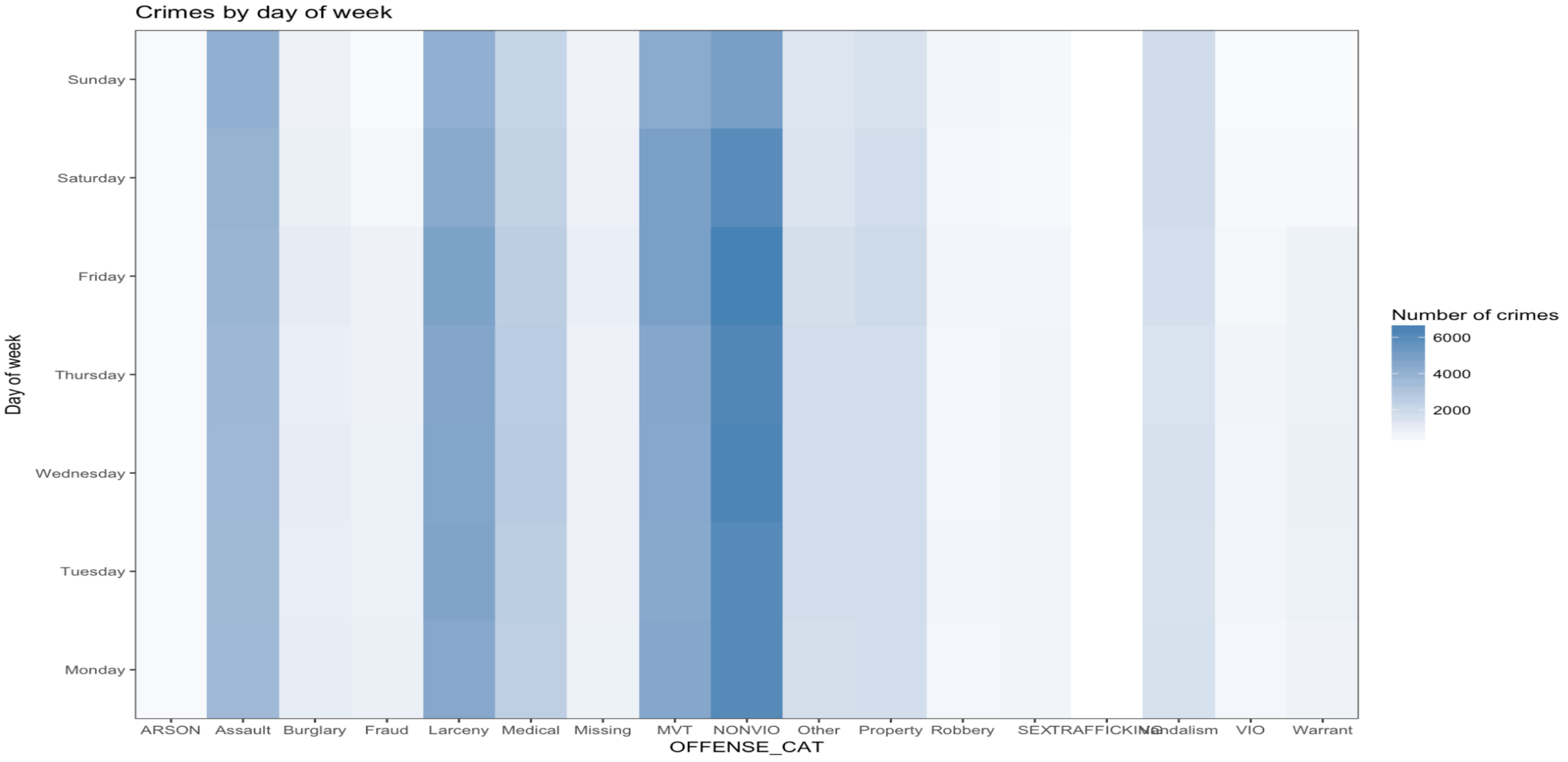
Crime Data Analysis – Crime by Time/Type



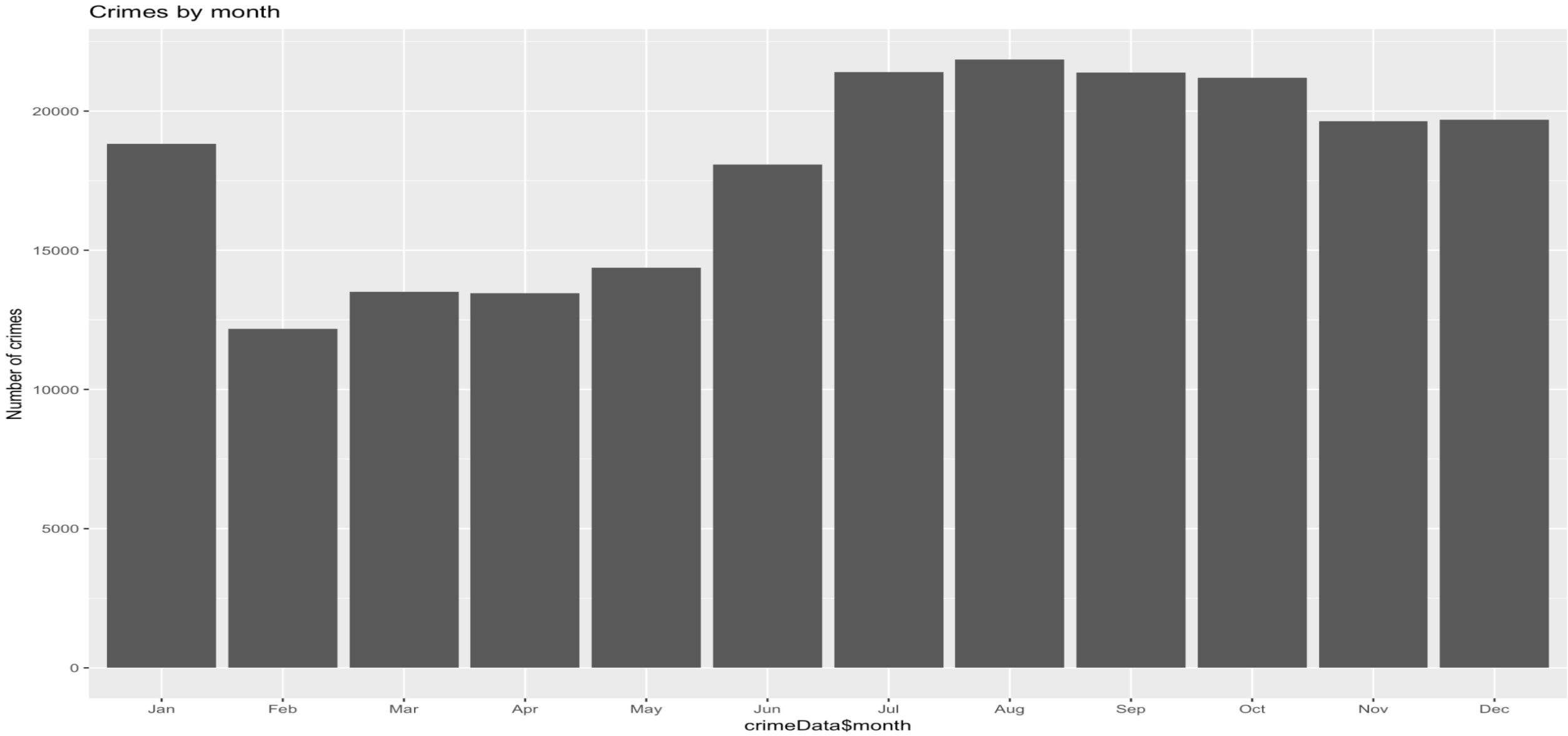
Crime Data Analysis – Crime by Week



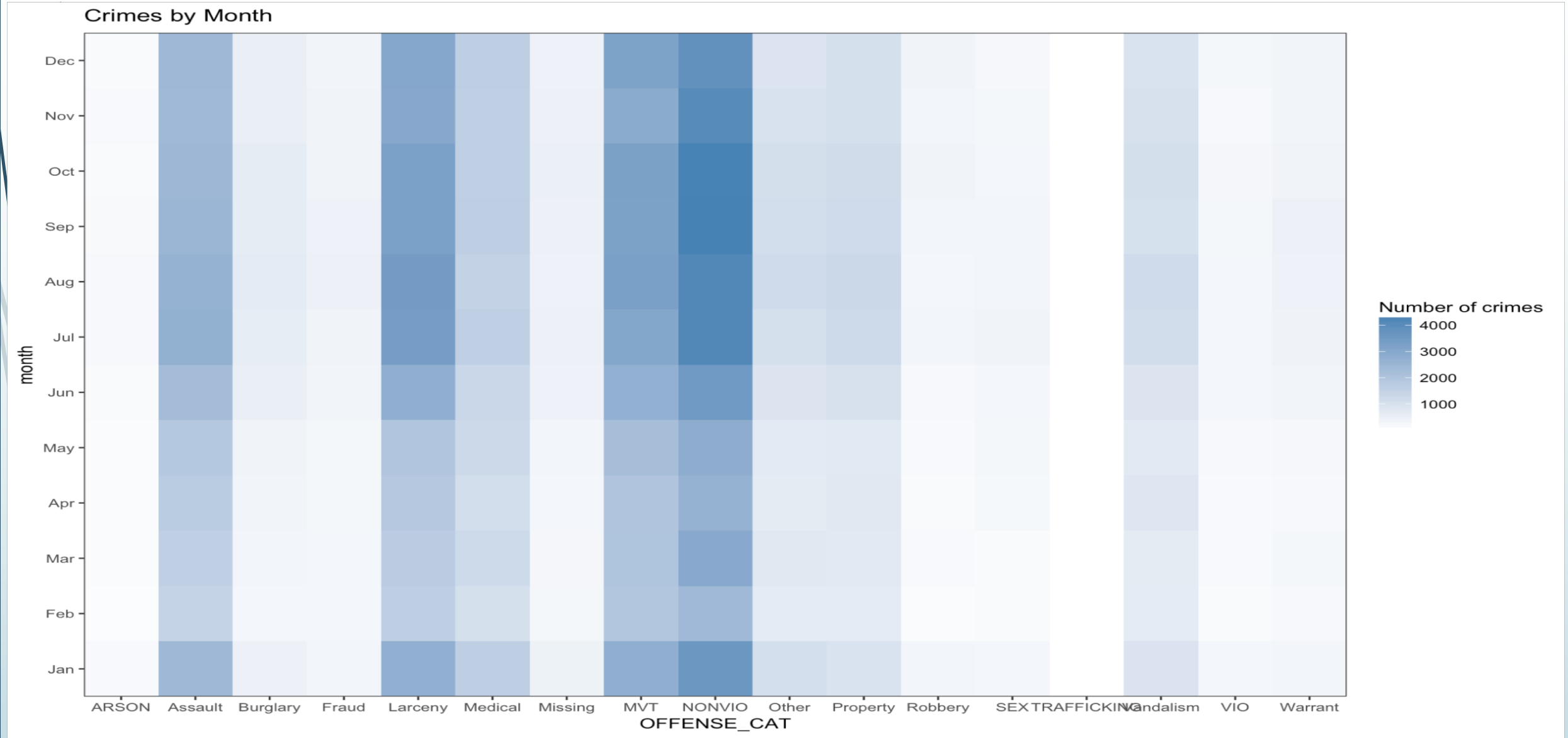
Crime Data Analysis – Crime by Week/Type



Crime Data Analysis – Crime by Month

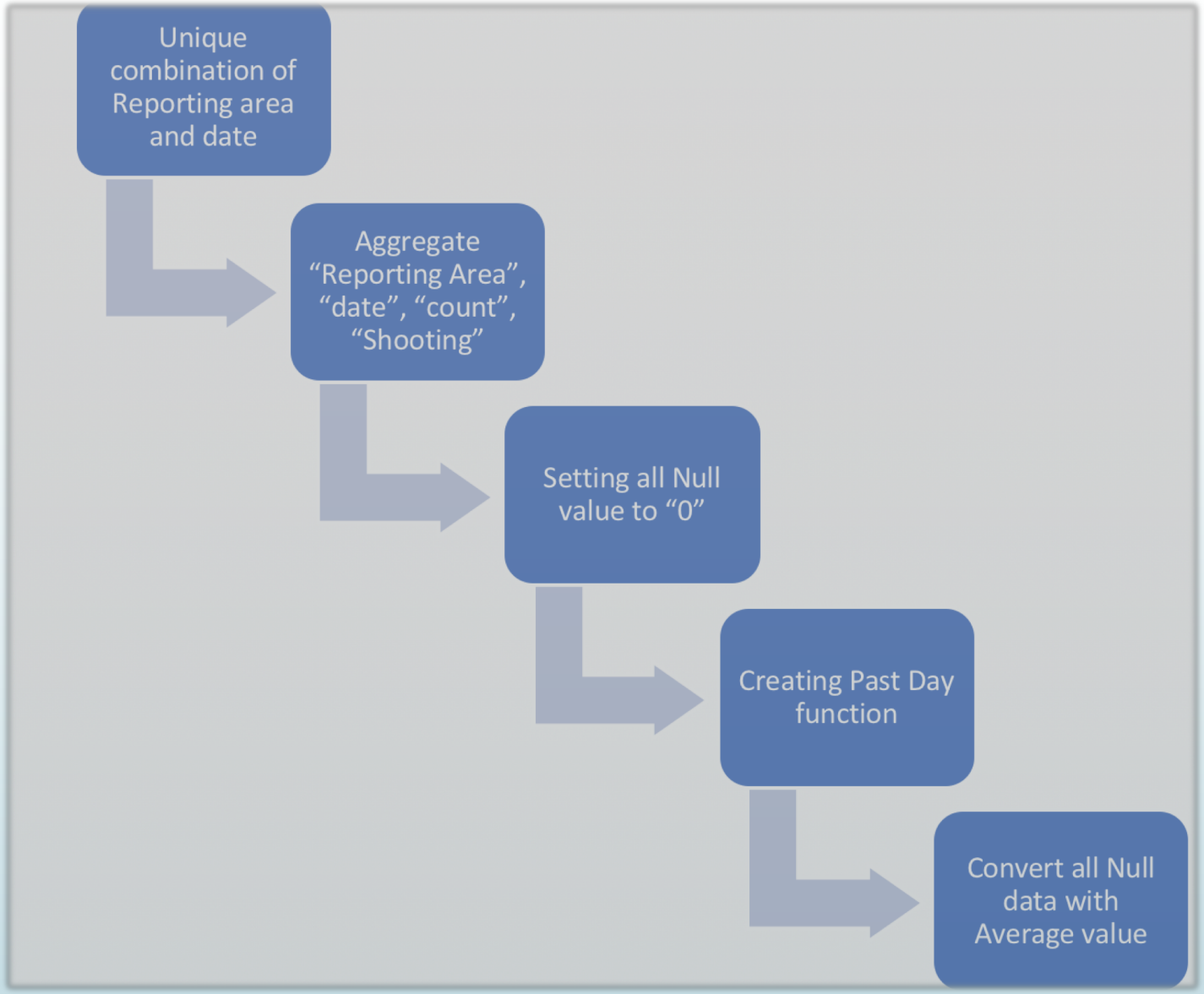


Crime Data Analysis – Crime by Month/Type



Data Modeling

- Determine the data to create the model
- Pre defined reporting area is considered instead of location
- Can not predict when and where crime will happen
- Predictive shooting
- Multivariate regression model with negative binomial distribution of errors



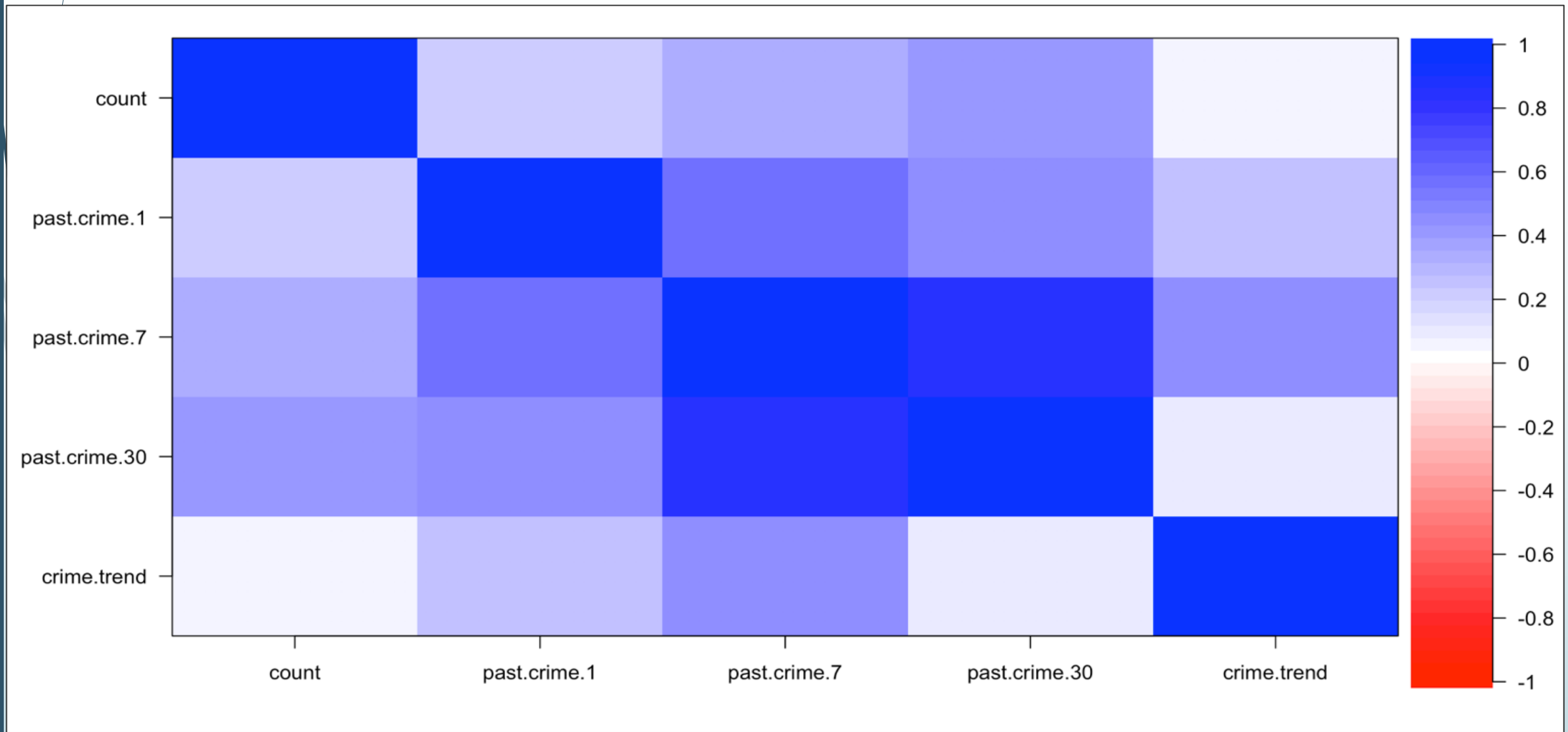
Data Modeling - Code

```
model.data <- aggregate(crimeData.agg[, c('count', 'SHOOTING')], by= list(crimeData.agg$REPORTING_AREA, as.character(crimeData.agg$date)),
                        FUN= sum)
names(model.data) <- c("reportingArea", "dates", "count", "SHOOTING")
model.data <- merge(temp4, model.data, by= c('reportingArea', 'dates'), all.x= TRUE)
model.data$count[is.na(model.data$count)] <- 0
model.data$SHOOTING[is.na(model.data$SHOOTING)] <- 0
model.data$day <- weekdays(as.Date(model.data$dates), abbreviate= TRUE)
model.data$month <- months(as.Date(model.data$dates), abbreviate= TRUE)
pastDays <- function(x) { c(0, rep(1, x))}
model.data$past.crime.1 <- ave(model.data$count, model.data$reportingArea, FUN= function(x) filter(x, pastDays(1), sides= 1))
model.data$past.crime.7 <- ave(model.data$count, model.data$reportingArea, FUN= function(x) filter(x, pastDays(7), sides= 1))
model.data$past.crime.30 <- ave(model.data$count, model.data$reportingArea, FUN= function(x) filter(x, pastDays(30), sides= 1))
meanNA <- function(x){
  mean(x, na.rm= TRUE)
}
model.data$past.crime.1 <- ifelse(is.na(model.data$past.crime.1),
                                  meanNA(model.data$past.crime.1), model.data$past.crime.1)
model.data$past.crime.7 <- ifelse(is.na(model.data$past.crime.7), meanNA(model.data$past.crime.7), model.data$past.crime.7)
model.data$past.crime.30 <- ifelse(is.na(model.data$past.crime.30), meanNA(model.data$past.crime.30), model.data$past.crime.30)

model.data$past.shooting.30 <- ave(model.data$SHOOTING, model.data$reportingArea, FUN= function(x) filter(x, pastDays(30), sides= 1))
model.data$past.shooting.30 <- ifelse(is.na(model.data$past.shooting.30), meanNA(model.data$past.shooting.30), model.data$past.shooting.30)

cor(model.data$past.crime.30, model.data$past.shooting.30)
cor(model.data$past.crime.30, model.data$past.crime.7)

model.data$crime.trend <- ifelse(model.data$past.crime.30 == 0, 0, model.data$past.crime.7/model.data$past.crime.30)
```

Data Modeling – Model data

reportingArea	dates	count	SHOOTING	day	month	past.crime.1	past.crime.7	past.crime.30	past.shooting.30	crime.trend	season
13	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
14	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
15	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
16	2015-06-15	1	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
17	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
18	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
19	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
20	2015-06-15	2	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
21	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
22	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
23	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
24	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
25	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
26	2015-06-15	1	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
27	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
28	2015-06-15	3	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
29	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
30	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer
31	2015-06-15	0	0	Mon	Jun	0.2560966	1.793597	7.697577	0.01852606	0.233008	summer

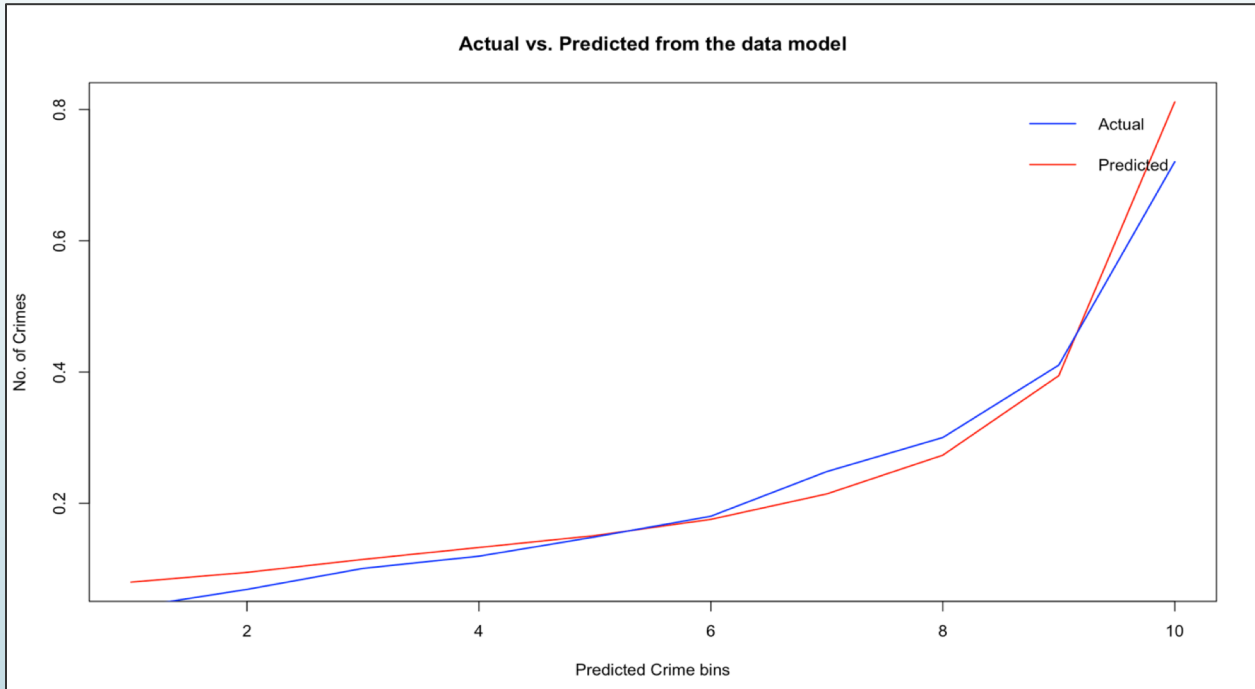
Data Modeling – Correlation Matrix Plot



Data Modeling – Evaluation

- 90% of data for Build and 10% for test
- Mean and variance not same, meaning data is widespread.
- Used glm.nb () function to create the Model.
- Used predict function to compare values to actual values using RMSE method
- Actual and Predictive values merge in between
- Actual and Predictive values diverge at extremes.

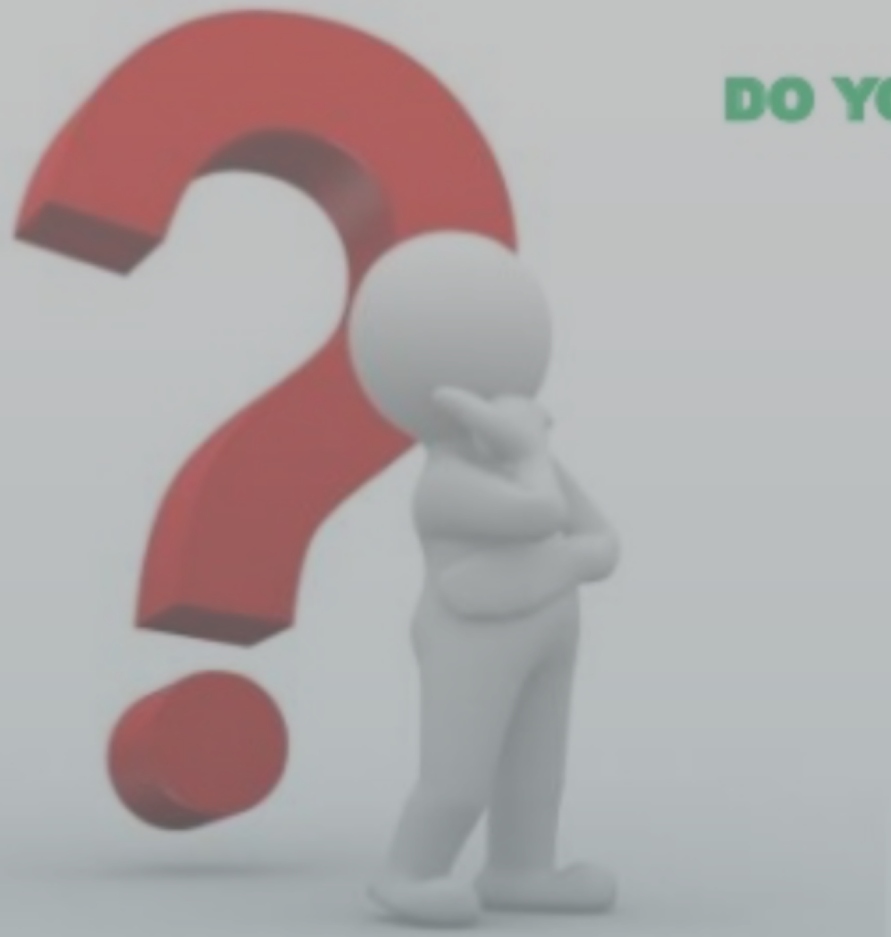
```
validate <- data.frame(test.data$count, crime.model.pred)
names(validate) <- c("actual", "predicted")
validate$bucket <- with(validate, cut(predicted, breaks= quantile(predicted, probs= seq(0, 1, 0.1)),
                                      include.lowest= TRUE, labels= c(1:10)))
validate <- aggregate(validate[, c('actual', 'predicted')], by=
                        list(validate$bucket), FUN = mean)

plot(validate$predicted, col= "red", type= "l", lwd= 1.5, ylab= "No. of Crimes", xlab= "Predicted Crime bins",
     main= "Actual vs. Predicted from the data model")
      lines(validate$actual, col= "blue", lwd= 1.5)
        legend("topright", c("Actual", "Predicted"), col= c("blue", "red"), lwd= c(1.5, 1.5), bty= "n")
```



Recommendations

- Better is to deal with different crime separately to create a better Model
- Spatial dimension included in the Model
- Prediction with smaller time intervals would be far precise than current
- Use of geo maps to locate high alert zones

THANK YOU.....



DO YOU HAVE ANY QUESTİONS ?