# Chicago Crime Exploratory Data Analysis

## Course Project Phase 1

### Miguel Fernandez

The following analysis represents Phase 1 of the BAN-502 course project at the University of North Carolina - Wilmington. This analysis will examine crime data from Chicago, Illinois. The data set contains 15,000 observations from January 1, 2018 through December 30, 2018. Phase 1 will consist of data cleaning and exploratory analysis to identify patterns and variables that may help inform a model that will attempt to predict the Arrest variable. Model building will be completed in Phase 2.

# Import libraries  
library(tidyverse)  
library(lubridate)  
library(ggmap)  
library(sf)  
library(stringr)  
library(VIM)  
library(wrapr)  
library(gridExtra)  
library(janitor)  
  
# Set global theme  
theme\_set(theme\_bw())  
  
# Set behavior for wrapr  
apply\_left.gg <- function(pipe\_left\_arg,  
 pipe\_right\_arg,  
 pipe\_environment,  
 left\_arg\_name,  
 pipe\_string,  
 right\_arg\_name) {  
 pipe\_right\_arg <- eval(pipe\_right\_arg,  
 envir = pipe\_environment,  
 enclos = pipe\_environment)  
 pipe\_left\_arg + pipe\_right\_arg   
}

# Read in data  
df <- read\_csv("chicago2.csv")  
  
# Convert Dates  
# Drop columns of no interest  
df <- df %>%  
 mutate(Date = mdy\_hms(Date)) %>%  
 select(-c(ID, `Case Number`, `Updated On`,  
 `X Coordinate`, `Y Coordinate`, Location))

We will preview the structure of the data and generate quick summary statistics.

str(df)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 15000 obs. of 17 variables:  
## $ X1 : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Date : POSIXct, format: "2018-09-13 14:00:00" "2018-07-06 23:00:00" ...  
## $ Block : chr "050XX S LAWNDALE AVE" "028XX W CERMAK RD" "064XX S FRANCISCO AVE" "035XX W 12TH PL" ...  
## $ IUCR : chr "1320" "0460" "1310" "0820" ...  
## $ Primary Type : chr "CRIMINAL DAMAGE" "BATTERY" "CRIMINAL DAMAGE" "THEFT" ...  
## $ Description : chr "TO VEHICLE" "SIMPLE" "TO PROPERTY" "$500 AND UNDER" ...  
## $ Location Description: chr "STREET" "STREET" "APARTMENT" "STREET" ...  
## $ Arrest : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Domestic : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Beat : chr "0821" "1023" "0823" "1021" ...  
## $ District : chr "008" "010" "008" "010" ...  
## $ Ward : num 14 12 15 24 3 9 14 21 25 9 ...  
## $ Community Area : num 62 30 66 29 38 50 57 71 31 49 ...  
## $ FBI Code : chr "14" "08B" "14" "06" ...  
## $ Year : num 2018 2018 2018 2018 2018 ...  
## $ Latitude : num 41.8 41.9 41.8 41.9 41.8 ...  
## $ Longitude : num -87.7 -87.7 -87.7 -87.7 -87.6 ...

summary(df)

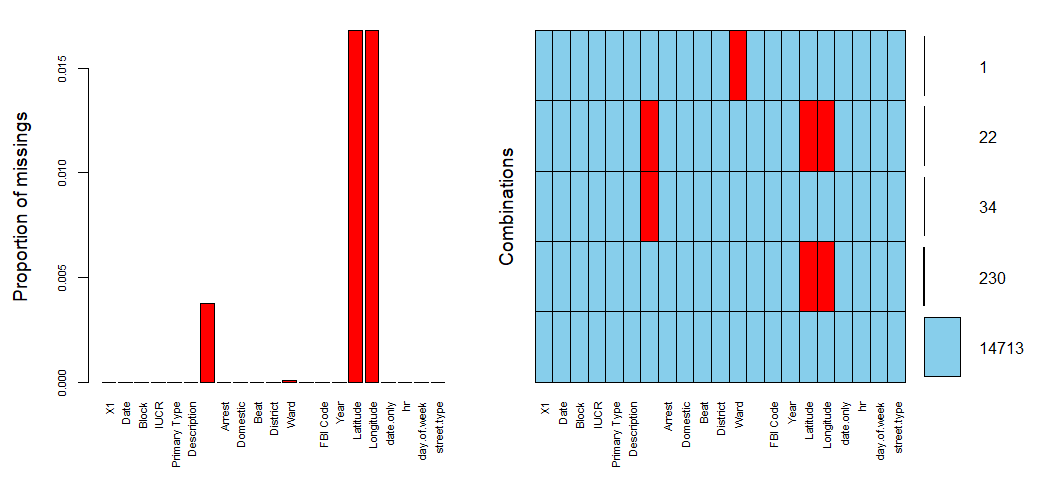
## X1 Date Block   
## Min. : 1 Min. :2018-01-01 00:00:00 Length:15000   
## 1st Qu.: 3751 1st Qu.:2018-04-11 22:23:00 Class :character   
## Median : 7500 Median :2018-07-05 14:30:00 Mode :character   
## Mean : 7500 Mean :2018-07-04 05:10:47   
## 3rd Qu.:11250 3rd Qu.:2018-09-26 09:48:30   
## Max. :15000 Max. :2018-12-30 23:40:00   
##   
## IUCR Primary Type Description Location Description  
## Length:15000 Length:15000 Length:15000 Length:15000   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Arrest Domestic Beat District   
## Mode :logical Mode :logical Length:15000 Length:15000   
## FALSE:11980 FALSE:12550 Class :character Class :character   
## TRUE :3020 TRUE :2450 Mode :character Mode :character   
##   
##   
##   
##   
## Ward Community Area FBI Code Year   
## Min. : 1.00 Min. : 0.00 Length:15000 Min. :2018   
## 1st Qu.:10.00 1st Qu.:23.00 Class :character 1st Qu.:2018   
## Median :24.00 Median :32.00 Mode :character Median :2018   
## Mean :23.33 Mean :36.54 Mean :2018   
## 3rd Qu.:35.00 3rd Qu.:53.00 3rd Qu.:2018   
## Max. :50.00 Max. :77.00 Max. :2018   
## NA's :1   
## Latitude Longitude   
## Min. :41.64 Min. :-87.93   
## 1st Qu.:41.77 1st Qu.:-87.71   
## Median :41.87 Median :-87.66   
## Mean :41.84 Mean :-87.67   
## 3rd Qu.:41.91 3rd Qu.:-87.63   
## Max. :42.02 Max. :-87.53   
## NA's :252 NA's :252

We can see there is information about the type of crime, whether or not it was domestic or if the offender was arrested. There are also several geographic variables including block, beat, district and ward. We will create additional variables to help inform the analysis.

# Create variables  
df <- df %>%  
 mutate(  
 date.only = as.Date(Date),  
 hr = hour(Date),  
 day.of.week = weekdays(Date),  
 street.type = str\_extract(Block, '\\w+$')  
 )

Let’s check for missing values in the data set.

vim\_plot = aggr(df, numbers = TRUE,  
 prop = c(TRUE, FALSE), cex.axis = 0.7)



There appears to be a column with missing values but the name of the variable does not appear on the plot. Let’s get a list of column names to determine which ones are missing from the plot.

colnames(df)

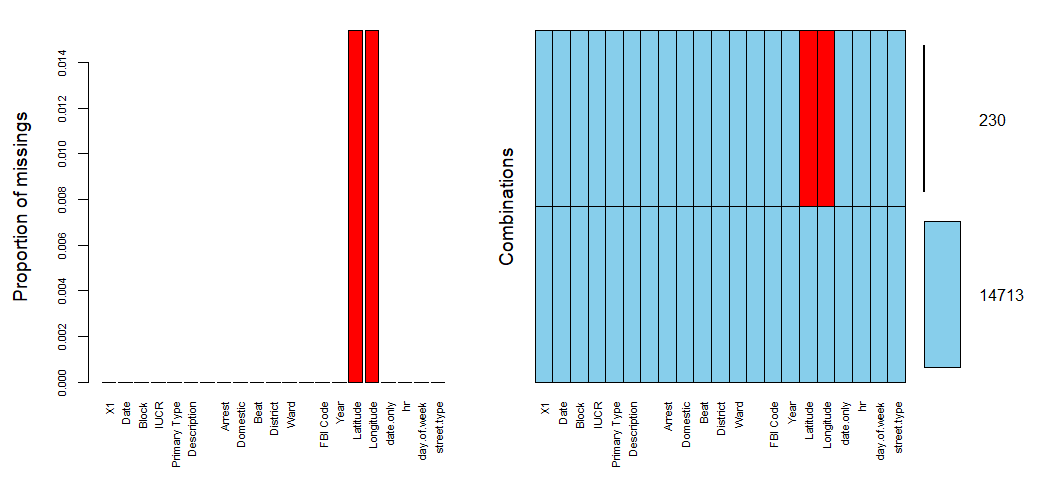
## [1] "X1" "Date" "Block"   
## [4] "IUCR" "Primary Type" "Description"   
## [7] "Location Description" "Arrest" "Domestic"   
## [10] "Beat" "District" "Ward"   
## [13] "Community Area" "FBI Code" "Year"   
## [16] "Latitude" "Longitude" "date.only"   
## [19] "hr" "day.of.week" "street.type"

Location Description and Community Area are the two variables that do not appear in the plot and the former is the one with missing values. There is also one missing value from the Ward variable. Because there are so few missing values in the data set which is large, we are going to remove the observations with missing values from Location Description and Ward columns. We are not performing any geographic analysis so the missing coordinate pairs are currently of no concern.

df <- df %>%  
 drop\_na("Location Description", "Ward")

Let’s make sure the correct rows were removed.

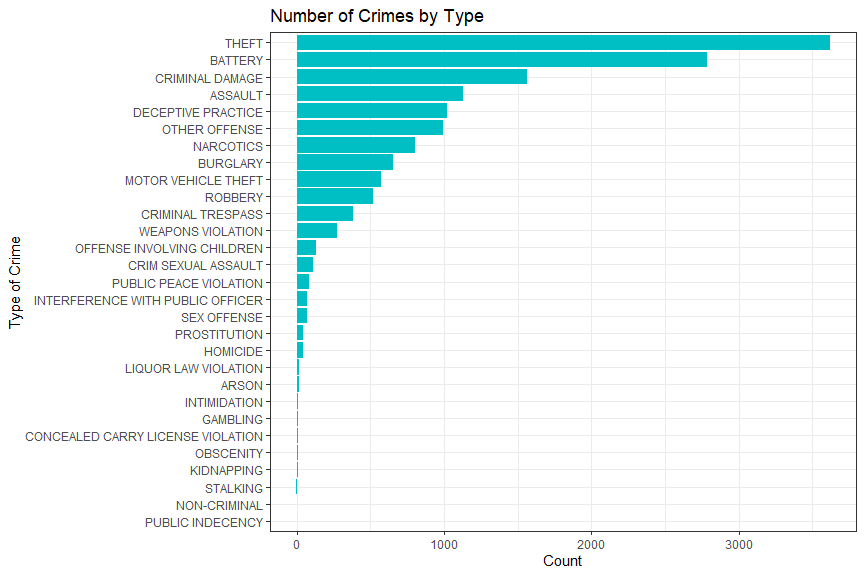
vim\_plot = aggr(df, numbers = TRUE,  
 prop = c(TRUE, FALSE), cex.axis = 0.7)



Several of the character variables need to be converted to factors before further analysis can be done.

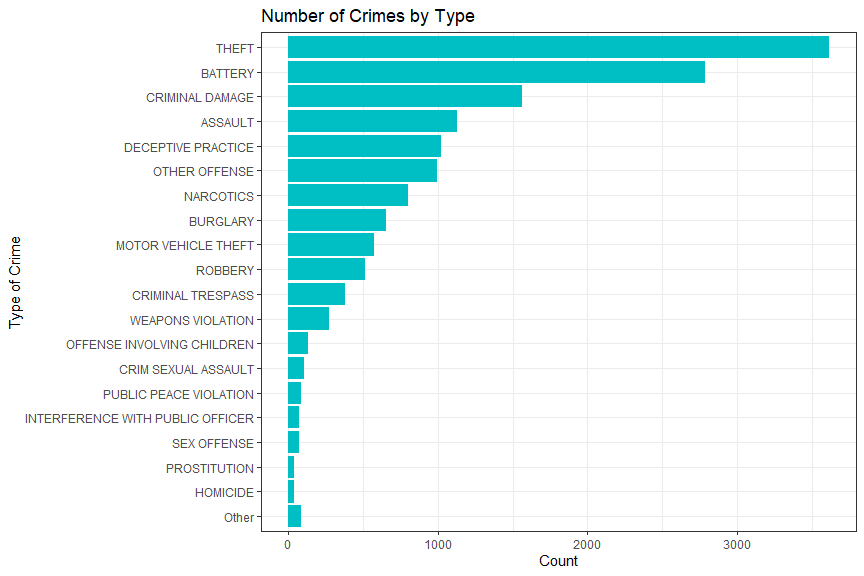
df <- df %>%  
 mutate(  
 IUCR = as\_factor(IUCR),  
 Description = as\_factor(Description),  
 Arrest = as\_factor(as.character(Arrest)),  
 Arrest = fct\_recode(Arrest, "Yes" = "TRUE", "No" = "FALSE"),  
 Domestic = as\_factor(as.character(Domestic)),  
 Domestic = fct\_recode(Domestic, "Yes" = "TRUE", "No" = "FALSE"),  
 Beat = as\_factor(Beat),  
 District = as\_factor(District),  
 Ward = as\_factor(Ward),  
 hr = as\_factor(hr),  
 day.of.week = as\_factor(day.of.week),  
 street.type = as\_factor(street.type),  
 Primary.Type = as\_factor(`Primary Type`),  
 Community.Area = as\_factor(`Community Area`),  
 FBI.Code = as\_factor(`FBI Code`),  
 Location.Description = as\_factor(`Location Description`)  
 ) %>%  
 select(-c(`Primary Type`, `Community Area`,  
 `FBI Code`, `Location Description`))

# Sort the Type by count  
df$Primary.Type <- reorder(df$Primary.Type, df$Primary.Type, FUN = length)  
  
ggplot(df, aes(x = Primary.Type)) +  
 geom\_bar(fill = "#00BFC4") +  
 ggtitle("Number of Crimes by Type") +  
 xlab("Type of Crime") +  
 ylab("Count") +  
 coord\_flip()



There seem to be several types of crimes with only a few observations. We will group those levels with fewer than 25 observations into a new level called Other.

# Sort the Type by count  
df$Primary.Type <- reorder(df$Primary.Type, df$Primary.Type, FUN = length)  
# Replace levels with fewer than 25 observations with Other  
levels(df$Primary.Type)[table(df$Primary.Type) < 25] <- 'Other'  
  
# Recreate plot from above to confirm grouping  
ggplot(df, aes(x = Primary.Type)) +  
 geom\_bar(fill = "#00BFC4") +  
 ggtitle("Number of Crimes by Type") +  
 xlab("Type of Crime") +  
 ylab("Count") +  
 coord\_flip()



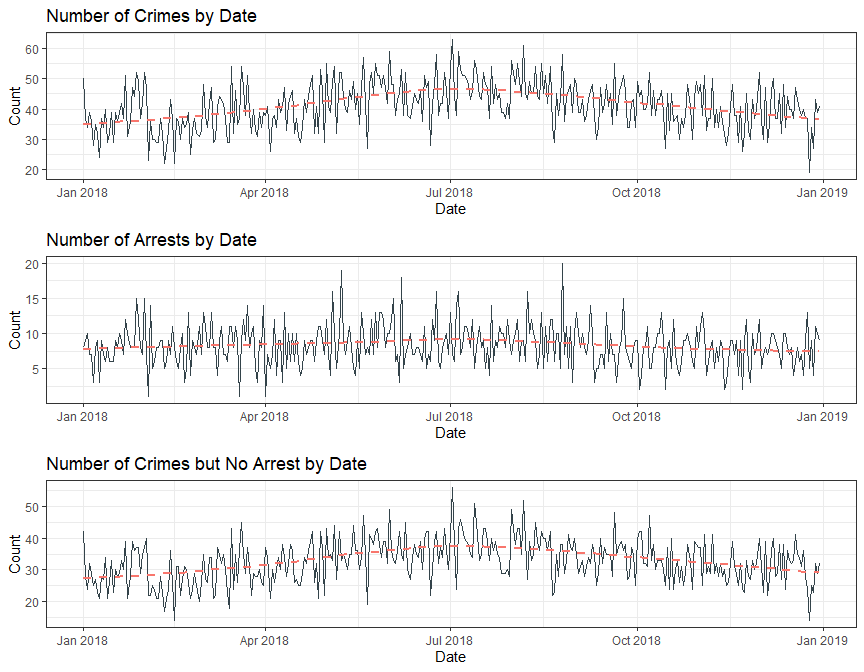
The plot above shows Theft to be the most common type of crime followed by Battery, Criminal Damage and Assult. Using the tabyl function from the janitor package we can generate some quick summary statistics on the Primary.Type variable.

tabyl(df$Primary.Type, sort = TRUE)

## df$Primary.Type n percent  
## Other 84 0.005621361  
## HOMICIDE 41 0.002743760  
## PROSTITUTION 42 0.002810681  
## SEX OFFENSE 71 0.004751389  
## INTERFERENCE WITH PUBLIC OFFICER 73 0.004885231  
## PUBLIC PEACE VIOLATION 86 0.005755203  
## CRIM SEXUAL ASSAULT 108 0.007227464  
## OFFENSE INVOLVING CHILDREN 132 0.008833568  
## WEAPONS VIOLATION 272 0.018202503  
## CRIMINAL TRESPASS 381 0.025496888  
## ROBBERY 516 0.034531219  
## MOTOR VEHICLE THEFT 571 0.038211872  
## BURGLARY 657 0.043967075  
## NARCOTICS 802 0.053670615  
## OTHER OFFENSE 995 0.066586362  
## DECEPTIVE PRACTICE 1021 0.068326307  
## ASSAULT 1131 0.075687613  
## CRIMINAL DAMAGE 1560 0.104396707  
## BATTERY 2784 0.186307970  
## THEFT 3616 0.241986214

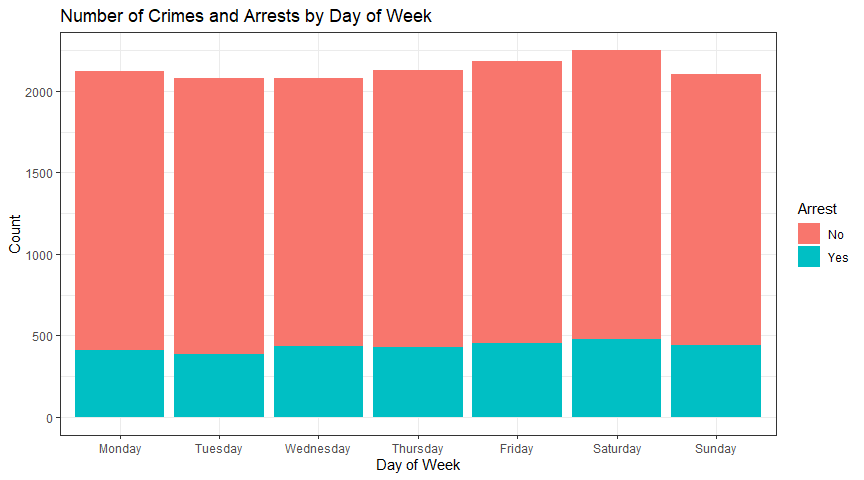
We can see that thefts account for nearly a quarter of the data set with more than 3,600 cases in 2018. Less than one percent of the crimes were homicides, with 41 observations in the data set.

# Crimes by Date  
p1 <- df %.>%  
 group\_by(., date.only) %.>%  
 tally(.) %.>%  
 rename(., "Num.of.Crimes" = n) %.>%  
 ggplot(., aes(x = date.only, y = Num.of.Crimes)) %.>%  
 geom\_line(color = '#33434A') %.>%  
 geom\_smooth(color = "#F8766D",  
 se = FALSE,  
 linetype = "dashed") %.>%  
 ggtitle("Number of Crimes by Date") %.>%  
 xlab("Date") %.>%  
 ylab("Count")  
  
# Arrests by Date  
p2 <- df %.>%  
 filter(., Arrest == "Yes") %.>%  
 group\_by(., date.only) %.>%  
 tally(.) %.>%  
 rename(.,"Num.of.Crimes" = n) %.>%  
 ggplot(., aes(x = date.only, y = Num.of.Crimes)) %.>%  
 geom\_line(color = "#33434A") %.>%  
 geom\_smooth(color = "#F8766D",  
 se = FALSE,  
 linetype = "dashed") %.>%  
 ggtitle("Number of Arrests by Date") %.>%  
 xlab("Date") %.>%  
 ylab("Count")  
  
# Non Arrests by Date  
p3 <- df %.>%  
 filter(., Arrest == "No") %.>%  
 group\_by(., date.only) %.>%  
 tally(.) %.>%  
 rename(.,"Num.of.Crimes" = n) %.>%  
 ggplot(., aes(x = date.only, y = Num.of.Crimes)) %.>%  
 geom\_line(color = "#33434A") %.>%  
 geom\_smooth(color = "#F8766D",  
 se = FALSE,  
 linetype = "dashed") %.>%  
 ggtitle("Number of Crimes but No Arrest by Date") %.>%  
 xlab("Date") %.>%  
 ylab("Count")  
  
grid.arrange(p1, p2, p3, ncol = 1)



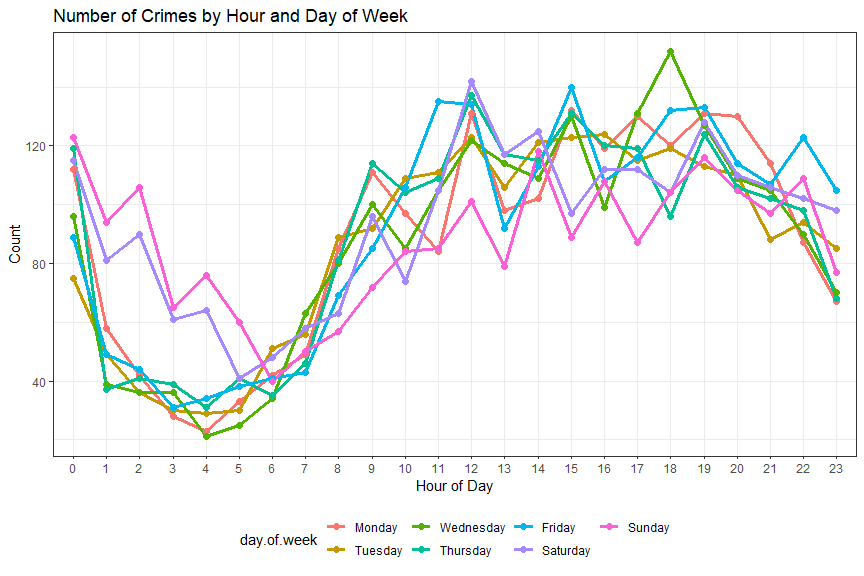
The three plots above show the number of crimes by date, the number of arrests by date and the number of crimes but no arrests by date. There seems to be more crimes during the summer months, peaking around July. However, the second plot reveals that the number of arrests remains constant around 7 or 8 arrests a day, even in the summer months when crimes increase. There also appears to be a chop in the data with severe swings from low to high. This might be caused by more crimes occurring on the weekends. Another interesting pattern visible in the plots is the spike at the beginning on the year. This spike is only present in the total crimes and no arrest plot. We can conclude that there are several more crimes on New Year’s Eve but they are relatively minor as the number of arrests remain low.

# Set order for day of week variable  
df$day.of.week <- factor(df$day.of.week,  
 levels = c("Monday", "Tuesday",  
 "Wednesday", "Thursday",  
 "Friday", "Saturday", "Sunday"))  
  
df %.>%  
 group\_by(., day.of.week) %.>%  
 ggplot(., aes(x = day.of.week, fill = Arrest)) %.>%  
 geom\_bar() %.>%  
 ggtitle("Number of Crimes and Arrests by Day of Week") %.>%  
 xlab("Day of Week") %.>%  
 ylab("Count")



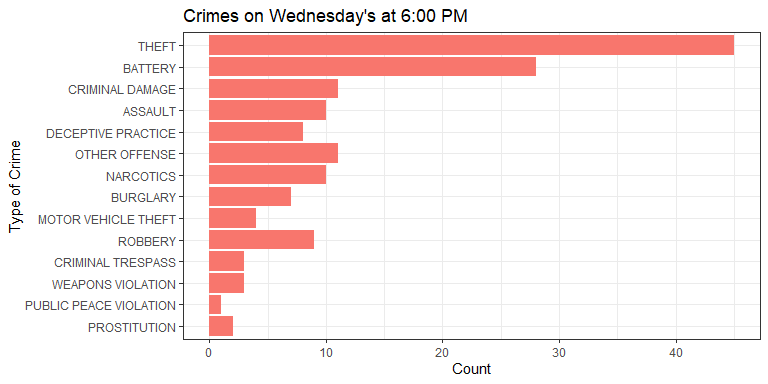
There does appear to be a slight increase in the number crimes and arrests on the weekends. Let’s see what the data looks like by hour of the day and day of week..

# Set order for day of week variable  
df$day.of.week <- factor(df$day.of.week,  
 levels = c("Monday", "Tuesday",  
 "Wednesday", "Thursday",  
 "Friday", "Saturday", "Sunday"))  
  
df %.>%  
 group\_by(., hr, day.of.week) %.>%  
 tally(.) %.>%  
 rename(., "Num.of.Crimes" = n) %.>%  
 ggplot(., aes(x = hr, y = Num.of.Crimes,  
 group = day.of.week,  
 color = day.of.week)) %.>%  
 geom\_line(size = 1.25) %.>%  
 geom\_point(size = 2.1) %.>%  
 theme(legend.position = "bottom") %.>%  
 ggtitle("Number of Crimes by Hour and Day of Week") %.>%  
 xlab("Hour of Day") %.>%  
 ylab("Count")



The plot reveals fairly consistent patterns across the weekdays. There are some exceptions, however. For instance, there tends to be more crimes in the early morning hours of Saturday and Sunday between 1:00 AM and 5:00 AM. Intuitively, late night on Friday’s and Saturday’s are higher than other days as they carry into early mornings of Saturday and Sunday. A sharp rise in the number of crimes observed occurs between 7:00 AM and 9:00 AM. One possible explanation for this is as people begin to wake up, they realize a crime was committed overnight and they report it. A second possilbe explanation is that everyone is now going about their day and as more people are out, there is a higher likelihood that a crime will be committed. The plot does support this hypothesis as the number of crimes is high throughout the day and into the evening. The number of reported crimes drops around 1:00 AM for most days when the population returns home and is in for the night.  
The more crimes tend to occur at 6:00 PM on Wednesdays. This is surprising. One would think the most crimes would occur on Friday’s or Saturdays in the evening. Let’s examine the crimes on Wednesday at 6:00 PM further.

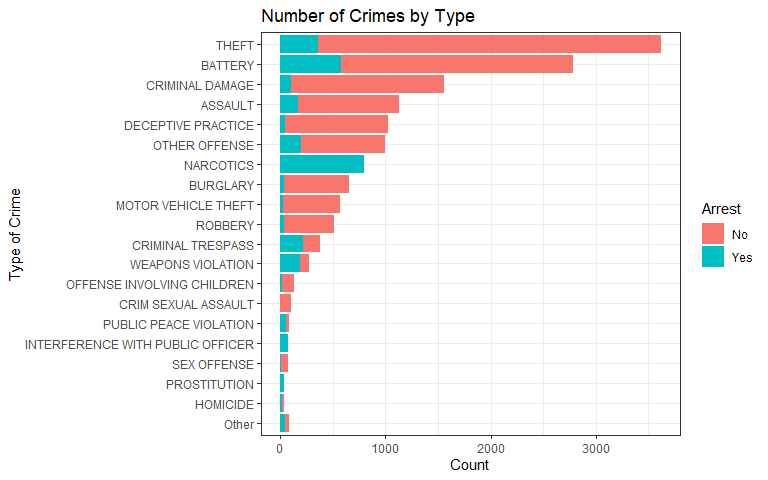
df %.>%  
 filter(., hr == 18, day.of.week == "Wednesday") %.>%  
 ggplot(., aes(x = Primary.Type)) %.>%  
 geom\_bar(fill = "#F8766D") %.>%  
 coord\_flip() %.>%  
 ggtitle("Crimes on Wednesday's at 6:00 PM") %.>%  
 xlab("Type of Crime") %.>%  
 ylab("Count")



The pattern of crimes on Wednesday’s at 6:00 PM is consistent with the overall pattern. Theft, battery and criminal damage are the top three types of crime.

Let’s take a plot from above showing crimes by type, but now let’s see the type of crime by whether or not an arrest was made.

ggplot(df, aes(x = Primary.Type, fill = Arrest)) +  
 geom\_bar() +  
 ggtitle("Number of Crimes by Type") +  
 xlab("Type of Crime") +  
 ylab("Count") +  
 coord\_flip()



The plot shows some very interesting results. It appears as if nearly, if not all crimes dealing with narcotics end in arrest. While there are fewer crimes involving interference with a public officer, those crimes all end with an arrest too. We’ll create a proportion table to better understand the results.

t2 = table(df$Primary.Type, df$Arrest)  
round(prop.table(t2, margin = 1),2)

##   
## No Yes  
## Other 0.44 0.56  
## HOMICIDE 0.66 0.34  
## PROSTITUTION 0.00 1.00  
## SEX OFFENSE 0.92 0.08  
## INTERFERENCE WITH PUBLIC OFFICER 0.00 1.00  
## PUBLIC PEACE VIOLATION 0.35 0.65  
## CRIM SEXUAL ASSAULT 0.97 0.03  
## OFFENSE INVOLVING CHILDREN 0.88 0.12  
## WEAPONS VIOLATION 0.32 0.68  
## CRIMINAL TRESPASS 0.43 0.57  
## ROBBERY 0.93 0.07  
## MOTOR VEHICLE THEFT 0.96 0.04  
## BURGLARY 0.95 0.05  
## NARCOTICS 0.00 1.00  
## OTHER OFFENSE 0.80 0.20  
## DECEPTIVE PRACTICE 0.95 0.05  
## ASSAULT 0.85 0.15  
## CRIMINAL DAMAGE 0.94 0.06  
## BATTERY 0.79 0.21  
## THEFT 0.90 0.10

The table confirms our observation from above. Narcotics and interferring with a public officer have a 100 percent arrest rate. A majority of other crimes tend to end with no arrest. We created a variable that extracted the ending of the Block variable that attempts to capture the type of street the crime occurred on. Let’s filter for some of the most common street types.

# Subset street type  
s.type <- df %>%  
 filter(street.type == "AVE" |  
 street.type == "RD" |  
 street.type == "PL" |  
 street.type == "ST" |  
 street.type == "DR" |  
 street.type == "BLVD")  
  
# Refactor street type variable  
s.type$street.type <- factor(s.type$street.type)  
  
# Generate table of street type  
tabyl(s.type$street.type, sort = TRUE)

## s.type$street.type n percent  
## AVE 7588 0.51965484  
## RD 514 0.03520066  
## PL 357 0.02444871  
## ST 5277 0.36138885  
## DR 336 0.02301055  
## BLVD 530 0.03629640

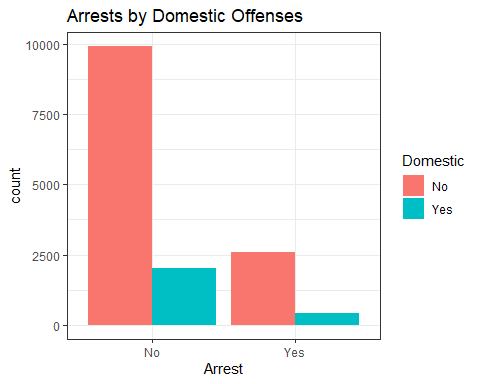
This table contains the most common values for street type. We can see that avenues account for just over 50 percent of crimes. The second most common street type location for crimes is street. Let’s break those down further and see how street type influences arrest.

t3 = table(s.type$street.type, s.type$Arrest)  
round(prop.table(t3, margin = 1),2)

##   
## No Yes  
## AVE 0.81 0.19  
## RD 0.77 0.23  
## PL 0.87 0.13  
## ST 0.78 0.22  
## DR 0.83 0.17  
## BLVD 0.82 0.18

We can see that the street suffix does not have much of an effect on the arrest variable. Arrest/no arrest are approximately split 20/80 across all street types. Let’s now look at the Domestic crime variable against whether or not someone was arrested.

# Examine the Arrest Variable against Domestic  
ggplot(df, aes(x = Arrest, fill = Domestic)) +  
 geom\_bar(position = "dodge") +  
 ggtitle("Arrests by Domestic Offenses")



The plot above reveals that there are far more none arrest observations than there are incidents that ended with an arrest. We can also conclude that domestic crimes represent a small portion of the data set. Let’s get a quantitative measure of how many arrests were made in 2018 compared to crimes but no arrests.

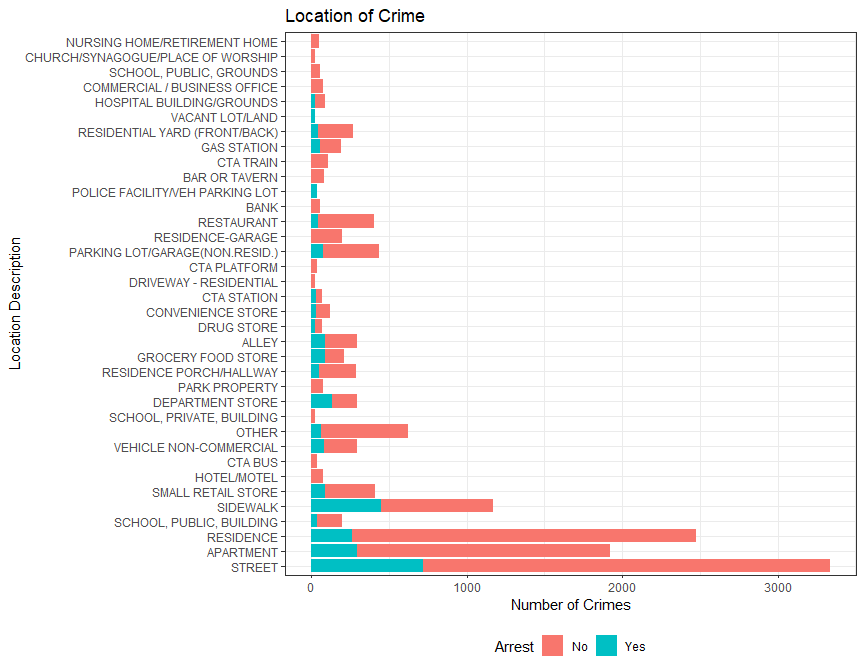
tabyl(df$Arrest, sort = TRUE)

## df$Arrest n percent  
## No 11924 0.7979656  
## Yes 3019 0.2020344

The table reveals interesting results. We can see that same split between arrests and no arrests made that is present in the street type variable. This is further evidence that street type would not be an effective predictor for the arrest variable.

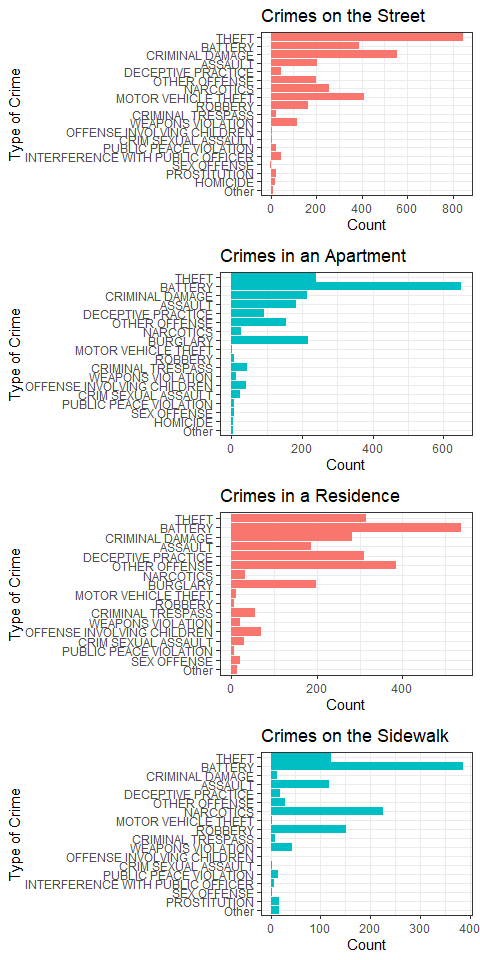
The variable Location.Description contains information about where the crime occurred. Perhaps this might help inform the model that will be constructed later.

df %>%  
 group\_by(., Location.Description, Arrest) %.>%  
 tally(.) %.>%  
 rename(., "Num.of.Crimes" = n) %.>%  
 filter(., Num.of.Crimes >= 25) %.>%  
 ggplot(., aes(x = Location.Description,  
 y = Num.of.Crimes,  
 fill = Arrest)) %.>%  
 geom\_col() %.>%  
 ggtitle("Location of Crime") %.>%  
 xlab("Location Description") %.>%  
 ylab("Number of Crimes") %.>%  
 theme(legend.position = "bottom") %.>%  
 coord\_flip()



Looking at the location description for where a crime occurred, we can see much more variation than the street type variable we just looked at. There are 104 different location descriptions in the data set but we filtered out those with less than or equal to 25 crimes. Because of this filtering, some locations appear to have no arrests when in fact there may be less than 25 arrests made. Looking at the plot, we can see that the majority of crimes occurred on the street with a small percentage ending in arrest. The same pattern is true for several of the top locations including residences and apartments. There are a few exceptions. For example, crimes that occurred on sidewalks tend to have a higher percentage of arrests made than other locations. Let’s select the top four locations for crimes and see if a certain crime dominates those locations.

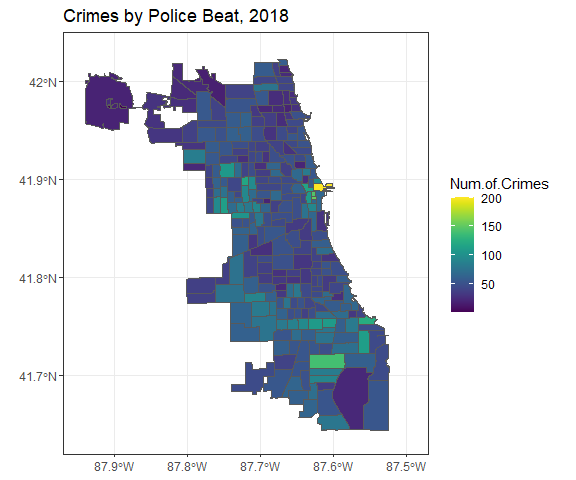
street <- df %.>%  
 filter(., Location.Description == "STREET") %.>%  
 ggplot(., aes(x = Primary.Type)) %.>%  
 geom\_bar(fill = "#F8766D") %.>%  
 ggtitle("Crimes on the Street") %.>%  
 xlab("Type of Crime") %.>%  
 ylab("Count") %.>%  
 coord\_flip()  
  
apt <- df %.>%  
 filter(., Location.Description == "APARTMENT") %.>%  
 ggplot(., aes(x = Primary.Type)) %.>%  
 geom\_bar(fill = "#00BFC4") %.>%  
 ggtitle("Crimes in an Apartment") %.>%  
 xlab("Type of Crime") %.>%  
 ylab("Count") %.>%  
 coord\_flip()  
  
res <- df %.>%  
 filter(., Location.Description == "RESIDENCE") %.>%  
 ggplot(., aes(x = Primary.Type)) %.>%  
 geom\_bar(fill = "#F8766D") %.>%  
 ggtitle("Crimes in a Residence") %.>%  
 xlab("Type of Crime") %.>%  
 ylab("Count") %.>%  
 coord\_flip()  
  
swlk <- df %.>%  
 filter(., Location.Description == "SIDEWALK") %.>%  
 ggplot(., aes(x = Primary.Type)) %.>%  
 geom\_bar(fill = "#00BFC4") %.>%  
 ggtitle("Crimes on the Sidewalk") %.>%  
 xlab("Type of Crime") %.>%  
 ylab("Count") %.>%  
 coord\_flip()  
  
grid.arrange(street, apt, res, swlk, ncol = 1)



We can see that thefts occur in large numbers on the street. Criminal damage and motor vehicle thefts also occur mostly in the street. We see similar patterns between apartments and residences. This is intuitive as both are living spaces. There are many cases of battery that occur in living spaces. Another interesting pattern is that crimes on the sidewalk resemble the pattern of living spaces more than it does crimes that occur on the street.

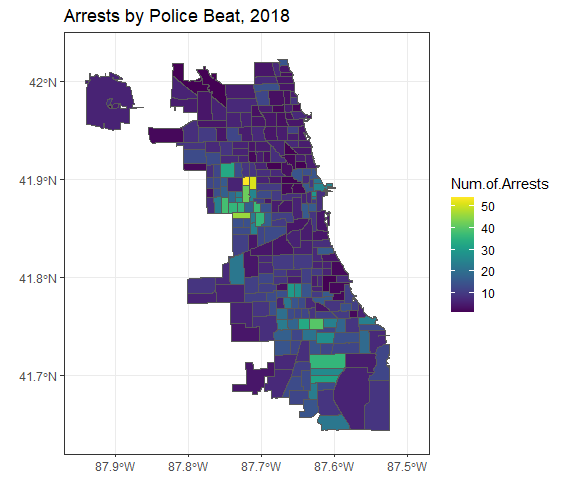
A large part of crime analysis is analzying the spatial distribution of incidents. We’ll use the sf package for R to accomplish some preliminary mapping.

# Read in shapefile  
sdf <- st\_read("police\_beat\_bndry/police\_beat\_bndry.shp",  
 stringsAsFactors = FALSE, quiet = TRUE)  
  
# Group by Beat ID  
by\_beat <- df %>%  
 group\_by(Beat) %>%  
 tally() %>%  
 rename("Num.of.Crimes" = n)  
  
# Merge by\_beat and sdf by Beat ID  
df1 <- merge(x = by\_beat, y = sdf[ , c("beat\_num", "geometry")],  
 by.x = "Beat", by.y = "beat\_num", all.x = TRUE)  
# Convert df1 to sf object  
df1 <- st\_as\_sf(df1)  
  
# Map crimes by beat  
ggplot() +   
 geom\_sf(data = df1) + aes(fill = Num.of.Crimes) +  
 scale\_fill\_gradientn(colors = viridis::viridis(20)) +  
 coord\_sf(xlim = c(-87.97, -87.47),  
 ylim = c(41.62, 42.05),  
 expand = FALSE) +  
 ggtitle("Crimes by Police Beat, 2018")



The map above visualizes the number of crimes by beat. We can see there are a few outliers present. The bright yellow polygon on the east side of the plot represents 200 crimes in the span 2018. We can also see there are a few beats with around 150 crimes reported but the majority of beats appear to have around 100 crimes of less. Let’s look at a similar map but now we’ll focus on the number of arrests made.

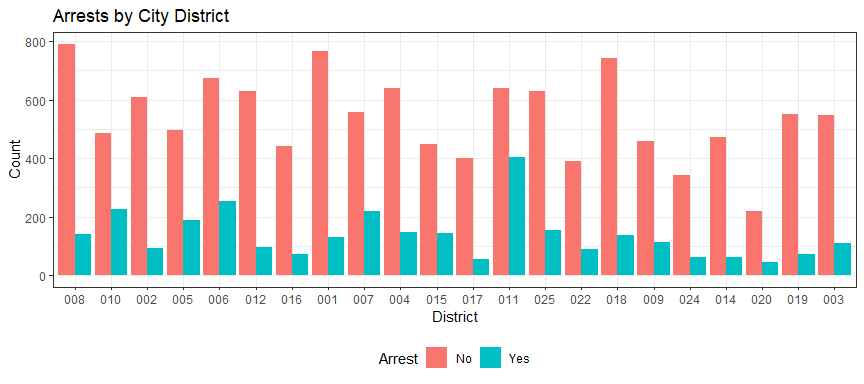
# Filter for arrests and group by beat  
beat\_arrest <- df %>%  
 filter(Arrest == "Yes") %>%  
 group\_by(Beat) %>%  
 tally() %>%  
 rename("Num.of.Arrests" = n)   
  
# Merge by\_beat and sdf by Beat ID  
df2 <- merge(x = beat\_arrest, y = sdf[ , c("beat\_num", "geometry")],  
 by.x = "Beat", by.y = "beat\_num", all.x = TRUE)  
# Convert df1 to sf object  
df2 <- st\_as\_sf(df2)  
  
# Map arrests by beat  
ggplot() +   
 geom\_sf(data = df2) + aes(fill = Num.of.Arrests) +  
 scale\_fill\_gradientn(colors = viridis::viridis(20)) +  
 coord\_sf(xlim = c(-87.97, -87.47),  
 ylim = c(41.62, 42.05),  
 expand = FALSE) +  
 ggtitle("Arrests by Police Beat, 2018")



Looking at the beats by arrest, we see a very different story. There appears to be a hotspot of arrests to the west that is concentrated around a few neighborhoods. There are also fewer arrests made in the northern portion of the city and approximately 30 to 40 arrests made in several beats to the south. We can also see that there is an irregular shaped beat in the north-west corner in the first plot but is missing in the second plot indicating that no arrests were made in that beat. That appears to be the only beat with zero arrests.

We do not have a geomtry file for city districts but let’s create a quick bar plot to see the distribution of arrests across distric variable.

# Examine the Arrest Variable against Domestic  
ggplot(df, aes(x = District, fill = Arrest)) +  
 geom\_bar(position = "dodge") +  
 ggtitle("Arrests by City District") +  
 ylab("Count") +  
 theme(legend.position = "bottom")



Looking back on the data, we know that around 80 percent of crimes have had zero arrests made. We also know that crimes vary widely by type. There is a slight up-tick in crimes over the summer months but the number of arrests remain constant. The number of crimes and arrests does not vary much by day of week but we can see trends given the hour of the day. Utilizing day of week and hour might prove to be beneficial to a predictive model. There is much variation in the location of the crime as well as whether or not an arrest was made. Using the location description field might prove informative to a model. Analyzing the data spatially, we were able to identify police beats that were outliers. This level of analysis may be too fine grain. We see more variation looking at the distribution of crimes and arrests by city districts which might be a suitable model variable. However, this is all speculation and we will not know how well the variable will perform in the model until we complete the next phase of the project.