12/4/2016

By: Trever Alexander & Jiahao Deng

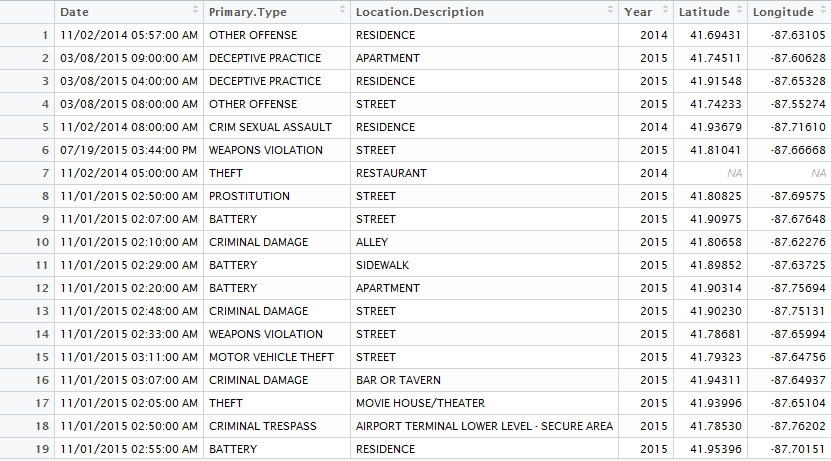
Analyzing Data for business intelligence

Chicago Crime

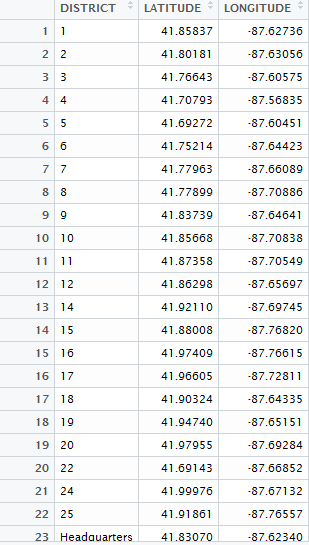
**Introduction**

Our analysis is based on two datasets the Chicago crime datasets and the Chicago police station dataset. The Chicago crime dataset contains detailed information including the types of crimes and coordinates of each crime. The dataset is for the time period from 2001 to 2016. The Chicago police station contains information about all the 23 police stations in the city of Chicago including the coordinates of each police station. The crime dataset contains more then 6 million rows of data. Therefore, for the convenience of our study, we used the subset that contains crimes for the most recent five years from 2012 to 2016 for most of our analysis. The following are parts of our datasets with the column that we have selected for the purpose of our study:

1. Crime Dataset



1. Police station dataset



**Data retrieving and cleaning**

1. **Data Retrieving**

At first, we tried to retrieved the dataset by using the API provided by the website with the token we acquired when we signed up on the website. The key code for this is the following:

**install.packages("RSocrata")**

**library(RSocrata)**

**df <- read.socrata("https://data.cityofchicago.org/resource/6zsd-86xi $$app\_token=aW2fHanT9bC9OyB5m0tqqg0Ns")**

Given the huge volume of our dataset, this method is very unreliable and many times it crashed our computer systems when we ran it. In order to get more reliable results, we manually downloaded the dataset and build our analysis based on it.

1. **Data Cleaning**

We followed two steps to clean our datasets:

First, we remove the rows that have more than one missing values.

Second, we remove the outliers in our crime dataset. We defined outliers as the ones that fall outside the range of Chicago. We calculated the distance from each row of our crime dataset to the central point of Chicago. We utilized the longitude and latitude columns in the crime dataset and used a function **"distm"** to calculate the distance (all distances are in meters). This is a function inside the package "**geosphere**" and it accurately calculates the distance between two points on earth. The following the code:

**distance <- vector()**

**for (i in 1:nrow(crime)){**

**distance <- append(distance,distm(c(crime$Longitude[i],crime$Latitude[i]), c(**-**87.623177,41.881832), fun = distHaversine))**

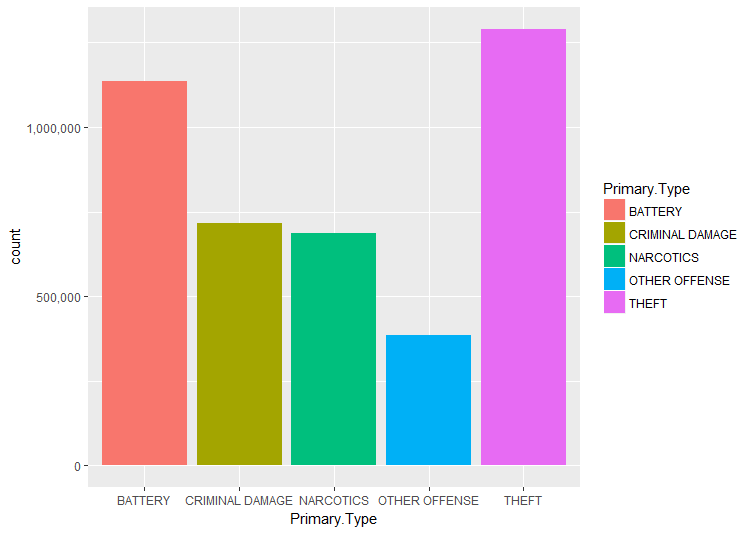
**}**

After we find the distances between each row to the center of Chicago, we removed the ones that are clearly out of range. In this case, the one with a distance more than 20,000 meters.

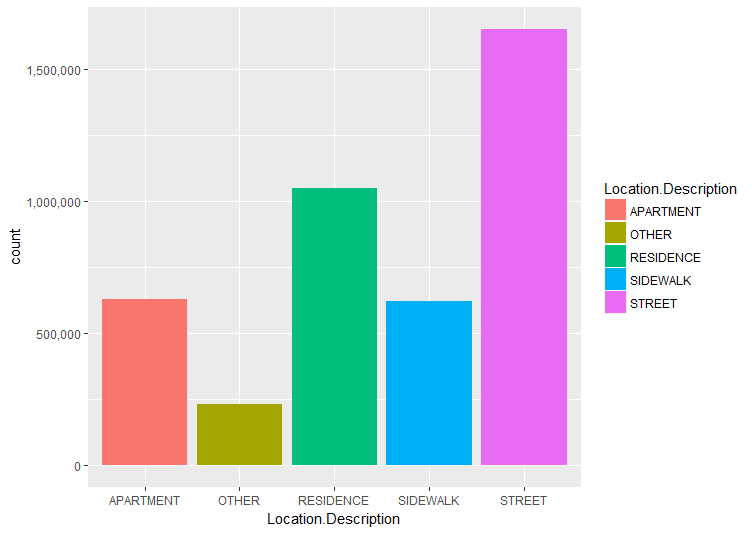
**Analysis**

**1.Basic Summary of crimes**

**Top 5 criminal offenses**

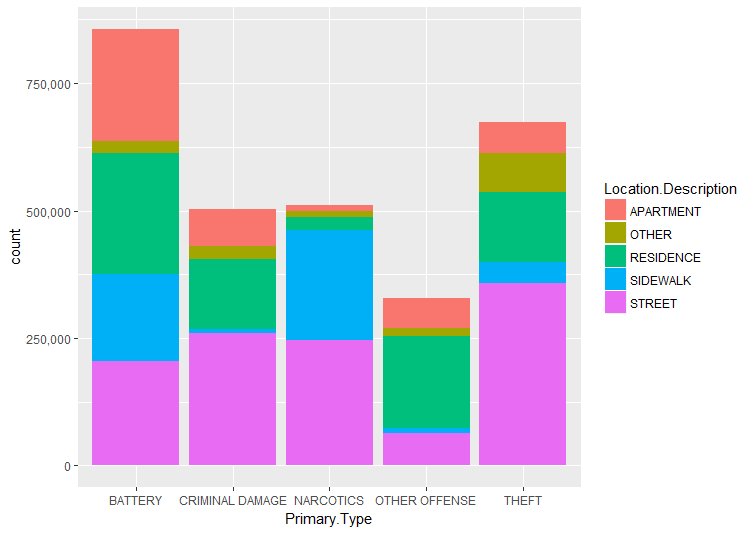


In this visualization the primary goal was to outline the top 5 criminal offenses in our data set. As you can see criminal theft and battery were the most common types of crimes to occur in Chicago followed by criminal damage, narcotics, and then other offenses.



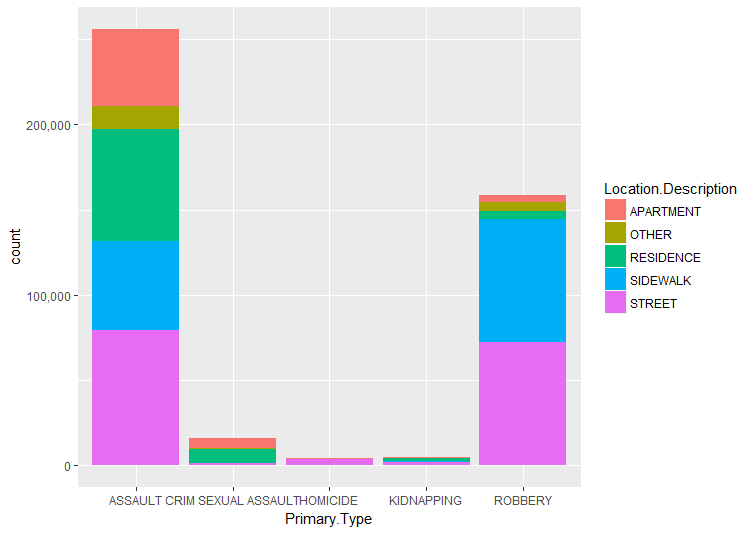
The goal for this visualization was to identify the top 5 locations where crimes are occurring. As you can see the most common place for a crime to occur in Chicago is on the Street. The next common places a crime will occurs is in someone's residence, on the sidewalk, in an apartment, and other miscellaneous locations.

**Top 5 crimes by top 5 locations of crimes for the last five years**



Our objective for this map was to visualize the top 5 crimes in respect to the top 5 locations. This map is a great visualization because it takes the first two maps and combines them together so you can get an overall idea of where each crimes is more likely to happen.

**Top 5 Violent crimes by top 5 crime locations for the last 5 years**



In this plot we took the top 5 violent crimes over a 5-year period and analyzed where they were most likely to occur at. Based on our data you can see that assault and robbery have the highest chance of occurring. This makes a lot of sense because it would be very rare to see homicide, kidnapping, and sexual assault crimes happen as much as crimes like assault and robbery. The top location for assaults to happen at is inside someone's residence or on the street while the top location for robbery to happen at is on the street and on the sidewalk.

**2. Time Series Analysis**

In order to analyze how number of crimes change over time, we built time series analysis graphs by using the "Date" column in our crime dataset. In order to make the time series work, we need to convert the date to a standard format and calculate the number of crimes that happened on each day from 2001 using "Group by" function. **The code we used is the following:**

**time <- read.csv("crime.csv)**

**time$Date <-substr(as.character(time$Date),1,10)**

**time$Date <-as.Date(time$Date,format ="%m/%d/%Y" )**

**group1<- group\_by(time,Date)**

**sum1 <- summarise(group1,frequency=n())**

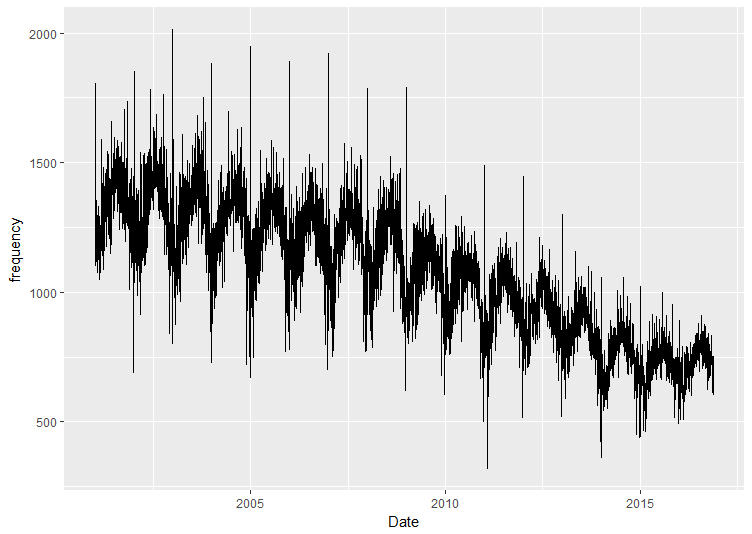
**Time series for all crime from 2001**

**Code:**

**library(reshape2)**

**df <- melt(sum1, id="Date")**

**ggplot(sum4,aes(x=Date,y=frequency) + geom\_line()**



This chart represents a time series of all the crimes that happened between 2001-2016. As you can see the frequency of all crimes are decreasing over the years.

**Time series for violent crimes since 2001**

**Code:**

**violent <- c("ASSAULT","HOMICIDE","CRIM SEXUAL ASSAULT","KIDNAPPING","ROBBERY")**

**time1<- subset(time,Primary.Type %in% violent)**

**time1$Date <-substr(as.character(time1$Date),1,10)**

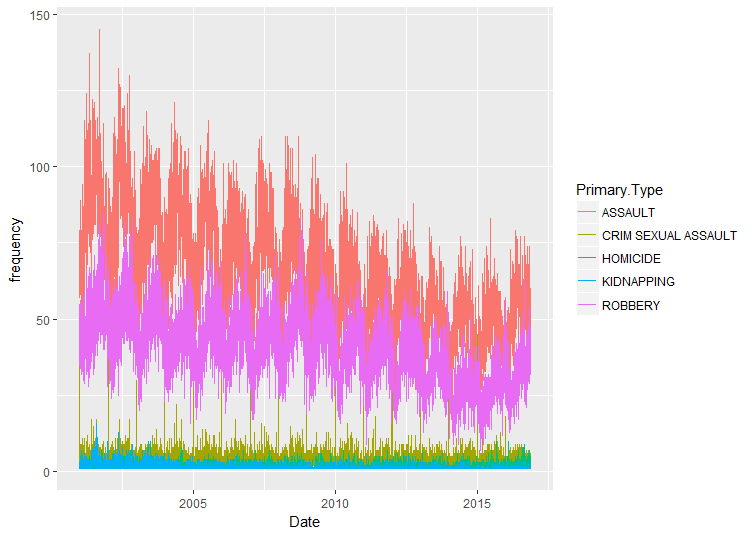
**time1$Date <-as.Date(time1$Date,format ="%m/%d/%Y" )**

**group4 <- group\_by(time1,Date,Primary.Type)**

**sum4 <- summarise(group4,frequency=n())**

**df <- melt(sum4, id="Date")**

**ggplot(sum4,aes(x=Date,y=frequency,colour=Primary.Type,group=Primary.Type)) + geom\_line()**



In this time series plot we focused our attention only on the top 5 violent crimes and analyzed how they increased/decreased over time. As you can see homicide, kidnapping, and sexual assault stayed constant while assault and robbery decreased for many years before it began to rise again in 2015.

**3. Geographical analysis**

**Heatmap for all crimes for the last 5 years**

**Code:**

**chicago <- get\_map(location='chicago',zoom = 11)**

**mapdata <-crime**

**mapdata$Longitude <- round(as.numeric(crime$Longitude), 2)**

**mapdata$Latitude <- round(as.numeric(crime$Latitude), 2)**

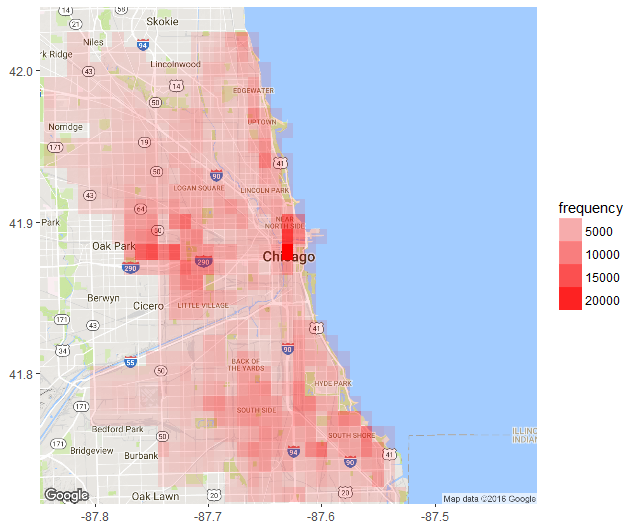
**group1 <- group\_by(mapdata, Longitude, Latitude)**

**sum1 <- summarise(group1,frequency=n())**

**chicagomap <- ggmap(chicago)**

**chicagomap+ geom\_tile(data =sum1, aes(x = Longitude, y = Latitude, alpha = frequency),**

**fill = 'red') + theme(axis.title.y = element\_blank(), axis.title.x = element\_blank())**



This heat map shows where all the crime is happening in Chicago. If you look closely you can see that the highest crime area is in the Chicago Northside close to the loop area and also in Oak Park.

**Heatmap for all violent crimes over the past 5 years**

**Code:**

**sub1 <- subset(crime,Primary.Type %in% violent,Year>=2012)**

**sub1$Latitude <-round(sub1$Latitude,8)**

**sub1$Longitude <- round(sub1$Longitude,8)**

**group3 <- group\_by(sub1,Latitude,Longitude,Primary.Type)**

**sum3 <- summarise(group3,frequency=n())**

**chicagomap +**

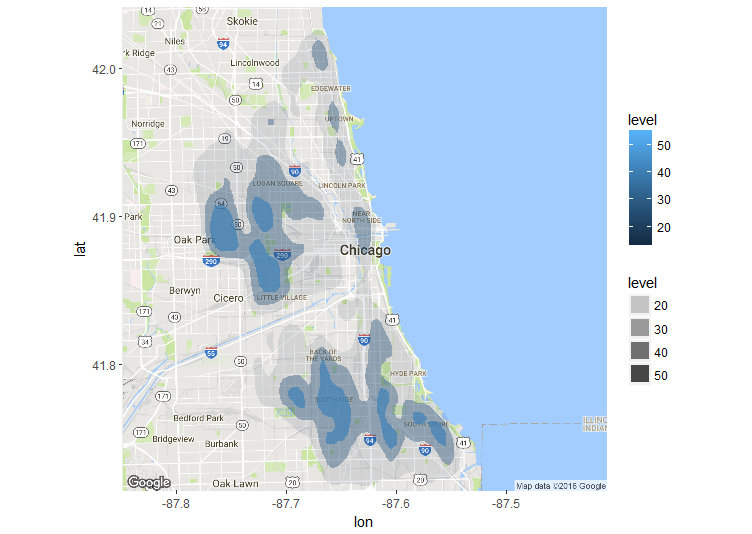
**stat\_density2d(**

**aes(x = Longitude, y = Latitude, fill = ..level..,**

**alpha = ..level..),**

**size = 2, bins = 4, data = sum3,**

**geom = "polygon")**

 This heat map shows where the top 5 most violent crimes are occurring at in the city of Chicago. As you can see the West side and South side of Chicago are the most violent areas in the city.

**Heatmap for change of violent crimes over the past 3 years**

**Code:**

**sub7<- subset(crime, Year>=2014 & Primary.Type %in% violent)**

**group7 <- group\_by(sub7,Latitude,Longitude,Year)**

**sum7 <- summarise(group7,frequency=n())**

**sum7 <- sum7[complete.cases(sum7),]**

**q1<-chicagomap +**

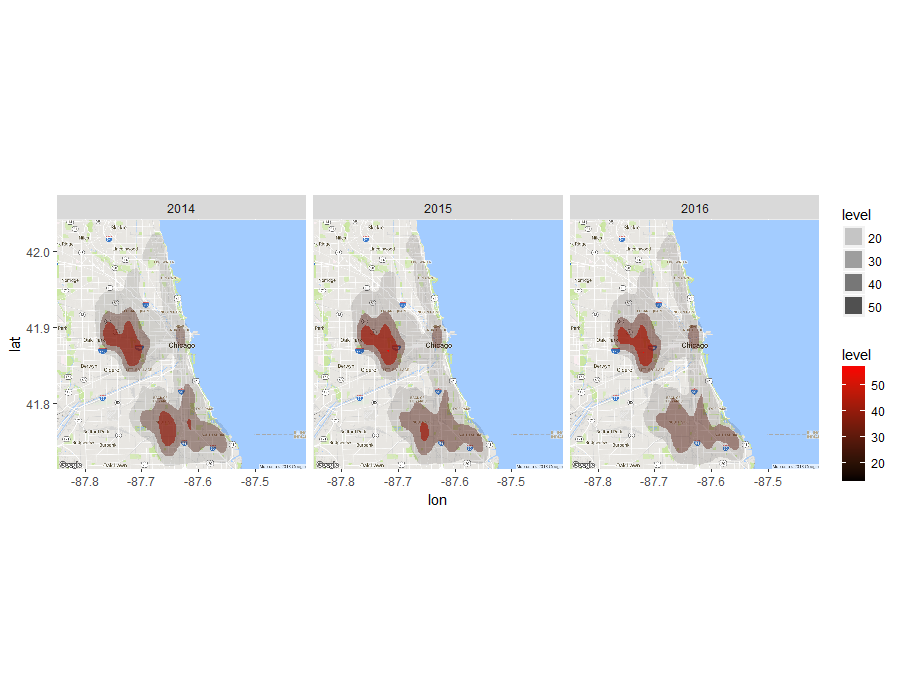
**stat\_density2d(**

**aes(x = Longitude, y = Latitude, fill = ..level..,**

**alpha = ..level..),**

**size = 2, bins = 4, data = sum7,**

**geom = "polygon")+scale\_fill\_gradient(low = "black",high= "red")+facet\_grid(~Year)**



This heatmap shows the most recent 3 years and analyzes the location where the most violent crimes are occurring. As you can see in the Southside of Chicago crimes are significantly decreasing over the years. In the Westside of Chicago, the rate of violent crimes does not seem to change at all over the years.

**4. Analysis of crimes to the nearest police station**

Next, we analyze how crimes are affected when they are In a closer proximity of a police station over the last five years. In order to get this, first we need to calculate the distance between every single row in the crime dataset to the nearest police station. In addition, for further analysis, we need to assign the police station number with closest distance to each crime instance. **The following is the code to achieve both purposes:**

**save1 <- vector()**

**mindis <- vector()**

**stationnum <- vector()**

**for(i in crimelocation){**

**for (a in policelocation){**

**save1=append(save1,distm(i, a, fun = distHaversine))**

**if(length(save1)==length(policelocation)){**

**mindis <-append(mindis,min(save1))**

**stationnum <-which(save1==min(save1))**

**save1 <- vector()**

**}**

**}**

**}**

**crime$distance <- mindis**

**crime$station\_num <- stationnum**

**Density of the violent crimes to the distance between each crime to the nearest police station over the last five years**

**Code:**

**violent <- c("ASSAULT","HOMICIDE","CRIM SEXUAL ASSAULT","KIDNAPPING","ROBBERY")**

**crime1$distance**

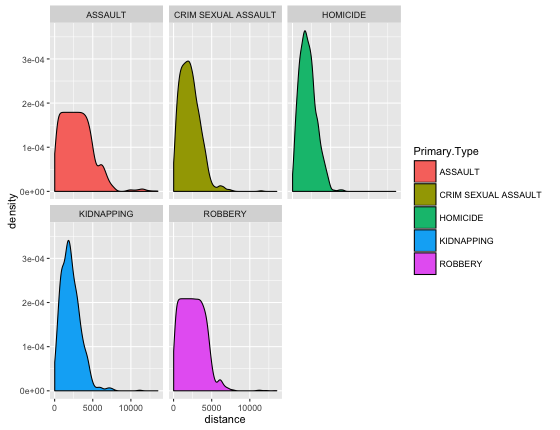
**crime1 <- subset(crime,Primary.Type %in% violent)**

**crime1$distance <- round(crime1$distance)**

**group8<-group\_by(crime1,Primary.Type,distance)**

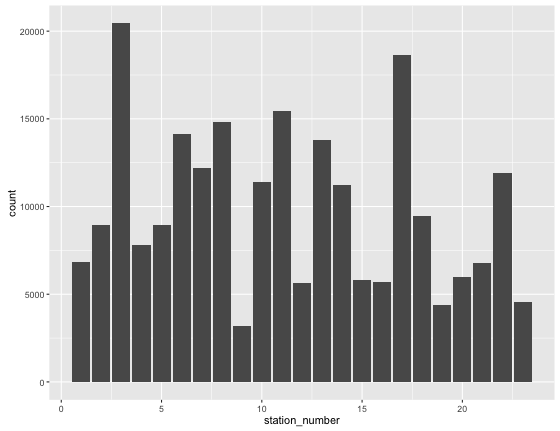
**sum8 <- summarise(group8,frequency=n())**

**qplot(distance,data = sum8,fill=Primary.Type,geom="density",facets= ~Primary.Type)**



This density plot identifies the top 5 violent crimes in respect to the proximity of the police station. As you can see most of the violent crimes occur within 2300 meters of a police station.

**Number of crimes within each station in 2016**



This bar chart analyzes how crimes are affected when they are in a closer proximity to a police station. As you can see station 3 and station 17 seem to have the highest amounts of crimes occurring within the proximity of a police station. This is valuable information because these stations can implement placing more patrols around these areas so they can decrease crime or they can increase their budget to hire more officers for this specific locations.

**Map of different violent crimes around the police station with the most crimes (police station number 3) in 2016**

**Code:**

**zoom <- ggmap(get\_map(policelocation[[3]],zoom = 15))**

**df3 <-subset(crime,Year==2016&Primary.Type %in% violent)**

**zoom +**

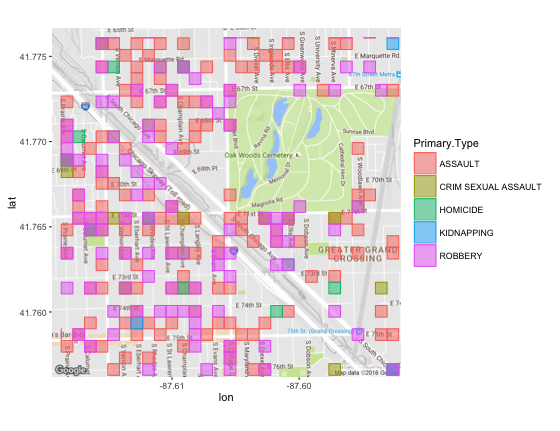
**stat\_bin2d(**

**aes(x = Longitude , y = Latitude, colour = Primary.Type,**

**fill = Primary.Type),**

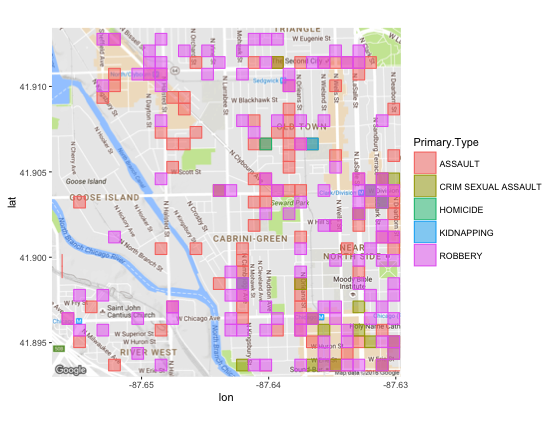
**size = .5, bins = 30, alpha = 1/2,**

**data = df3)**



This map zooms in on location 3 which is one of the police station locations that has a lot of crimes that occurs within its proximity. As you can see in this area robbery and assault are the most common types of crimes. Next you will notice that homicide, criminal sexual assault, and kidnapping are the rarer crimes to occur in this proximity.

**Map of different violent crimes around the police station with the second most crimes (police station number 17) in 2016**

Lastly, this map zooms in on police station number 17 which was also identified as high risk area. As you can see assault and robbery are also the top crimes within the proximity of this location but you will also notice that this area has more sexual assault crimes happening when being compared to location 3.

**Final thoughts**

After analyzing the Chicago crime dataset and Police station dataset we were amazed by the accuracy of the results. As you may know, in Chicago crimes are surging. By analyzing these datasets, it gave us an overview on what exactly is happening in the city of Chicago. Despite the numerous reports of homicides occurring in Chicago we realized that it is not really increasing as many predicated. Based off our time series plots we were able to pinpoint that homicide has remained at a constant rate over the years. Also, we noticed that the overall crime has been decreasing and the only crimes that are increasing in the city of Chicago is assault and robbery. If we were to provide Chicago Police Department with advice, we would advise them to place more patrols around station 3 and 17 because these areas are seeing significantly higher crime rates in respect to the other stations.

**Packages We used**

library(ggplot2)

library(devtools)

library(reshape2)

library(scales)

library(dplyr)

library(ggmap)

library(devtools)

library(reshape2)

library(scales)

library(geosphere)