Final Project

Thanh Nguyen-Duong

2/11/2020

Overall, write a coherent narrative that tells a story with the data as you complete this section.

Chicago is a city in the State of Illinois, filled with many landmarks and it is the third most populous city in the United States. However, due to the high number of populations in the city, crime and traffic collision rates also increased. These unfortunate events can negatively affect the image and how tourists viewed the city. In the datasets, there are many types of crimes recorded by the City of Chicago as well as traffic crashes and collisions from 2015-2019. To further understand the relationship and to get the whole map outline the area where crimes most exist, the Chicago police stations dataset is included.

This research projects will help and tell a story of how the growing percentage of school drop outs can influence the crime rates. Furthermore, an increased in the population of Chicago over time can also have an impact on traffic collisions over the years. Even though, the percentage of school drop outs focused mainly on 2011 and 2012, it still provides a glimpse to how drop outs rates can influence the increased crime rates in Chicago.

After finished analyzing the datasets, I found that there was not really a significant relationship between the percentage of drop outs in 2012 and the influence on its crime rates for that particular year. There are many other factors that can influence the crime rates in Chicago. However, there was a significant relationship between the numbers of traffic collision occurrences and months of the year. As the month increases, the traffic collision occurrence also increase. The highest occurrences in traffic crashes tend to be around winter time (October - December), where the weather is not too friendly and roads are covered with snow and low visibility road conditions do not help in term of reducing the crash occurrences.

Summarize the problem statement you addressed.

1. School drop out rates can have an influence on Chicago crime rates, and as the drop out rate increase the crime rate will also increase.
2. Chicago traffic collisions increase over time, and as the weather start to transition into the winter, traffic collisions will also increase.

Summarize how you addressed this problem statement (the data used and the methodology employed).

1. Gathering all the datasets from the listed sources, I needed to clean, remove/impute any missing values, create, append, subset, and other data manipulations to be able to fully evaluate and identify relationships between each variable.
2. Using cleaned data frames, I was able to perform different visualization models and linear regression to identify relationships for different variables.

a. Chicago Police Stations Dataset Chicago Police Stations Dataset

b. Chicago Traffic Crashes Dataset Chicago Traffic Crashes Dataset

c. Chicago Public School Dataset Chicago Public School Dataset

d. Chicago Crime Dataset Chicago Crime Dataset

Summarize the interesting insights that your analysis provided.

For 2012, there was no significant relationship between crime rates and drop outs rates in Chicago. On the other hand, the relationship between traffic collision occurrences and month of the year was positively proportional. As the weather gets colder, the road and visibility conditions can decrease, thus, creating more crashes than summer where the sun is out. To me, this was a surprise because I thought there would be more accidents in the summer where tourists are most likely to visit.

Summarize the implications to the consumer (target audience) of your analysis.

This research was done with no intention of biases. The results of my analysis were intended to show which crime type occurred the most and the areas of Chicago where crimes are most active. Over the years, Chicago crime rate has gone down but traffic collisions will continue

to increase as Chicago population increases.

Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

The limitations of this analysis is due to the limited amount of present data. These datasets only recorded up to 2017/2018 which we may not get a complete understanding of what Chicago's crime rates and traffic collision rates are like in the present years. Furthermore, the dropout rates for this analysis was 2011 and 2012 which were the only two that were in the dataset, by having more drop out rates in other years, we can fully understand whether or not if drop out rates have any influence on crime rates in Chicago.

In addition, submit your completed Project using R Markdown or provide a link to where it can also be downloaded from and/or viewed.

Below is my completed project that I used for this analysis

```
# Loading Packages
library(ggplot2)
library(tidyverse)
library(ggmap)
library(zoo)
library(dplyr)
library(lubridate)
library(data.table)
library(ggrepel)
library(sp)
library(rgeos)
library(reshape2)
```

```
# Getting datasets
Police_Stations <- read.csv("C:/Users/Tommy/Desktop/Bellevue Edu/DSC 520, Statistics for Data Science/DSC 52
0 Final Project Dataset/police-stations.csv", header = TRUE)

HighSchool_Report <- read.csv("C:/Users/Tommy/Desktop/Bellevue Edu/DSC 520, Statistics for Data Science/DSC
520 Final Project Dataset/chicago-public-schools-high-school-progress-report-card-2012-2013.csv", header = T
RUE)

Chicago_Crime <- read.csv("C:/Users/Tommy/Desktop/Bellevue Edu/DSC 520, Statistics for Data Science/DSC 520
Final Project Dataset/Crimes_2001_to_present.csv")

Traffic_Crash <- read.csv("C:/Users/Tommy/Desktop/Bellevue Edu/DSC 520, Statistics for Data Science/DSC 520
Final Project Dataset/Traffic_Crashes_-_Crashes.csv", header = TRUE)
```

```
# Crime dataset clean
crime_data <- data.frame(Chicago_Crime$Primary.Type, Chicago_Crime$District, Chicago_Crime$Year, Chicago_Cri
me$ID)

names(crime_data) <- c("CrimeType", "District", "Year", "ID")
```

```
# Police Stations Dataset clean
police_stations_data <- data.frame(Police_Stations$DISTRICT, Police_Stations$LATITUDE, Police_Stations$LONGI
TUDE, Police_Stations$DISTRICT.NAME)

names(police_stations_data) <- c("District", "Latitude", "Longitude", "District Name")
```

```
# School Dataset Clean
school_data <- data.frame(HighSchool_Report$School.ID, HighSchool_Report$One.Year.Dropout.Rate.2011...Percen
t, HighSchool_Report$One.Year.Dropout.Rate.2012...Percent)

names(school_data) <- c("SchoolID", "DropoutRate2011", "DropoutRate2012")

# Cleaning up School ID so all ID are in sequential order
school_data[1:92, 1] = c(1:92)
```

```
# Traffic Collisions dataset clean
crash_data <- data.frame(Traffic_Crash$CRASH_DATE, Traffic_Crash$CRASH_MONTH, Traffic_Crash$CRASH_HOUR, Traf
fic_Crash$FIRST_CRASH_TYPE)
names(crash_data) <- c("CrashDate", "CrashMonth", "CrashHour", "CrashType")

# converting class (factor) to date class
crash_data$CrashDate <-as.Date(crash_data$CrashDate, format = "%m/%d/%Y")

# Extracting only years and paste back into the crash_data data frame
a<-year(crash_data$CrashDate) # Only years are extracted
crash_data$Year<-paste(a,sep= ",")

# 2014 data were cut off from this data frame (lack of sample size)
crash_data <- crash_data[-c(261261:261265),]
```

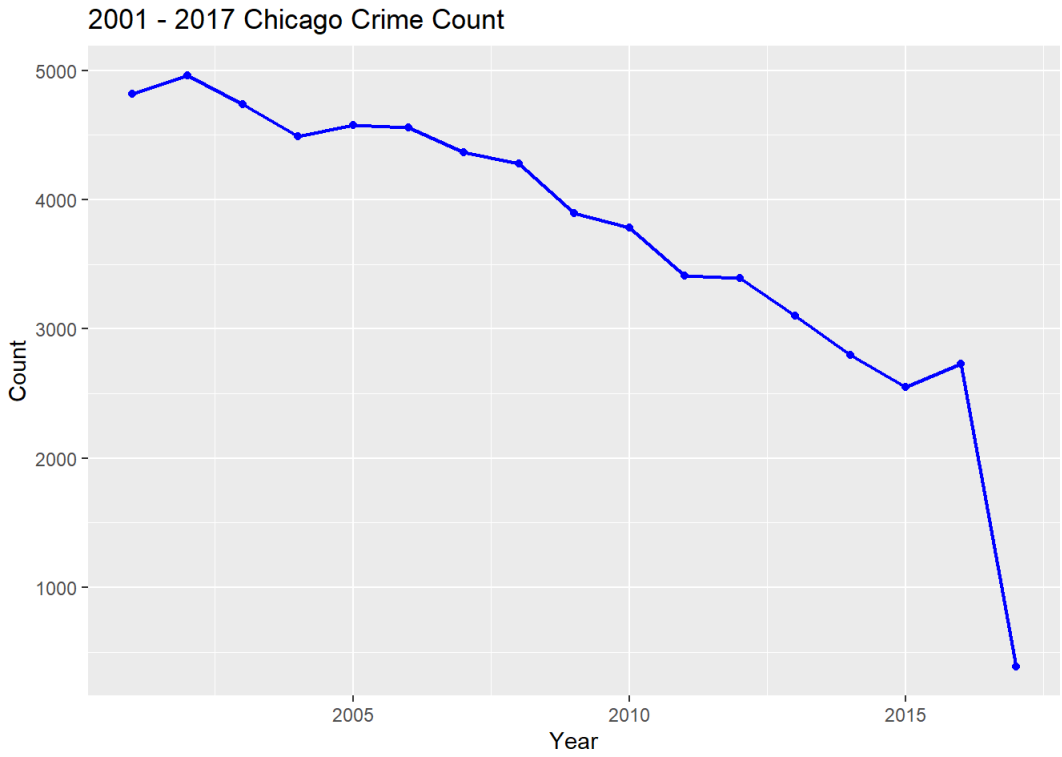
```
# Omit missing values
na.omit(crime_data)
na.omit(crash_data)
na.omit(school_data)
na.omit(police_stations_data)
```

Data Exploration

How has crime evolved over time in the city of Chicago?

Based on the graph, it looks like the crime rate in Chicago is decreasing over the year, especially after 2015, the crime count has drastically decreased.

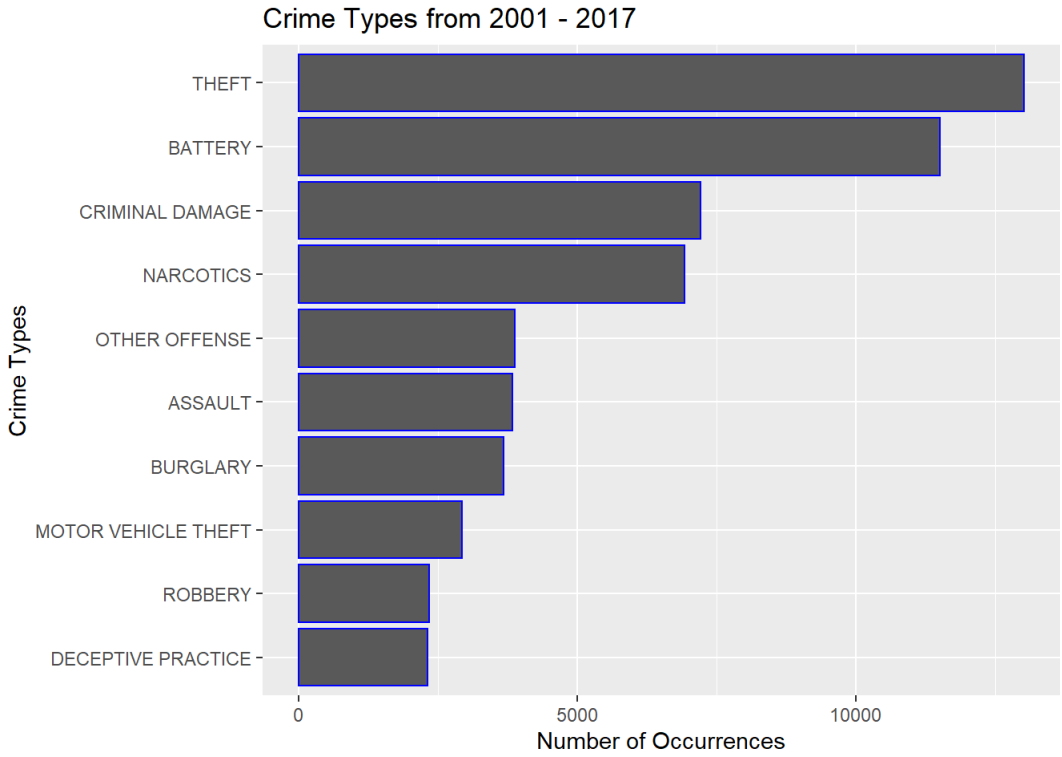
```
Crime_Count <- crime_data %>% count(Year)
ggplot(data= Crime_Count, aes(x=Year, y = n)) + geom_line(color = "blue", size = .8) + geom_point(color = "
blue") + labs(title = "2001 - 2017 Chicago Crime Count", y = "Count", x = "Year")
```



What are the distributions of the Different Types of Crimes?

Based on the bar plot, theft and battery were the top two crimes that have occurred in Chicago throughout the years (from 2001 - 2017).

```
## CType_Count <- count(crime_data, c("CrimeType"))
CType_Count <- crime_data %>% count(CrimeType)%>%top_n(10)


ggplot(data= CType_Count, aes(x= reorder(CrimeType, n), y = n)) + geom_bar(stat = "identity", color = "blue"
)+ coord_flip() + labs(title = "Crime Types from 2001 - 2017", y = "Number of Occurrences", x = "Crime Type
s")
```

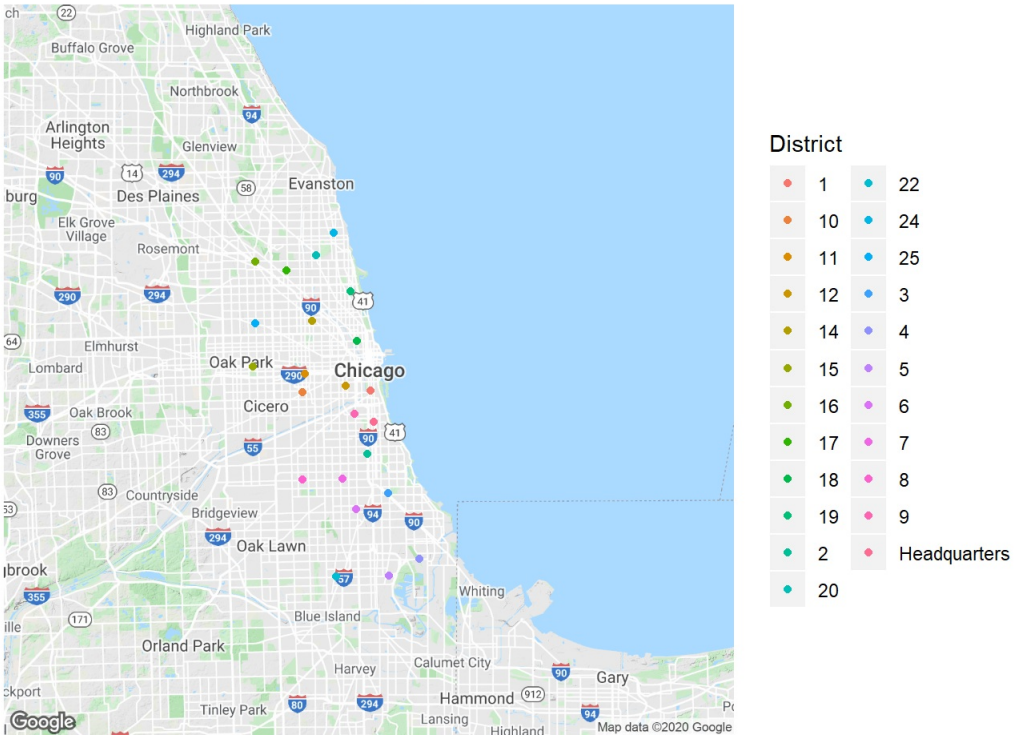


Map of all Police District Locations in Chicago

We can visually see how police districts are dispersed throughout the City of Chicago with its headquarter being in the epicenter of Chicago.

```
ggmap::register_google(key = "AIzaSyBCTcgVKReBhR_2cIxYo6WBDodBvUP8tu0")

Chicago_map <- qmap("chicago", zoom = 10, color = "color", size = .1) + ggtitle("Police District Locations
In Chicago")
Chicago_map +
geom_point(aes(x = Longitude, y = Latitude, colour = District),  data = police_stations_data)
```

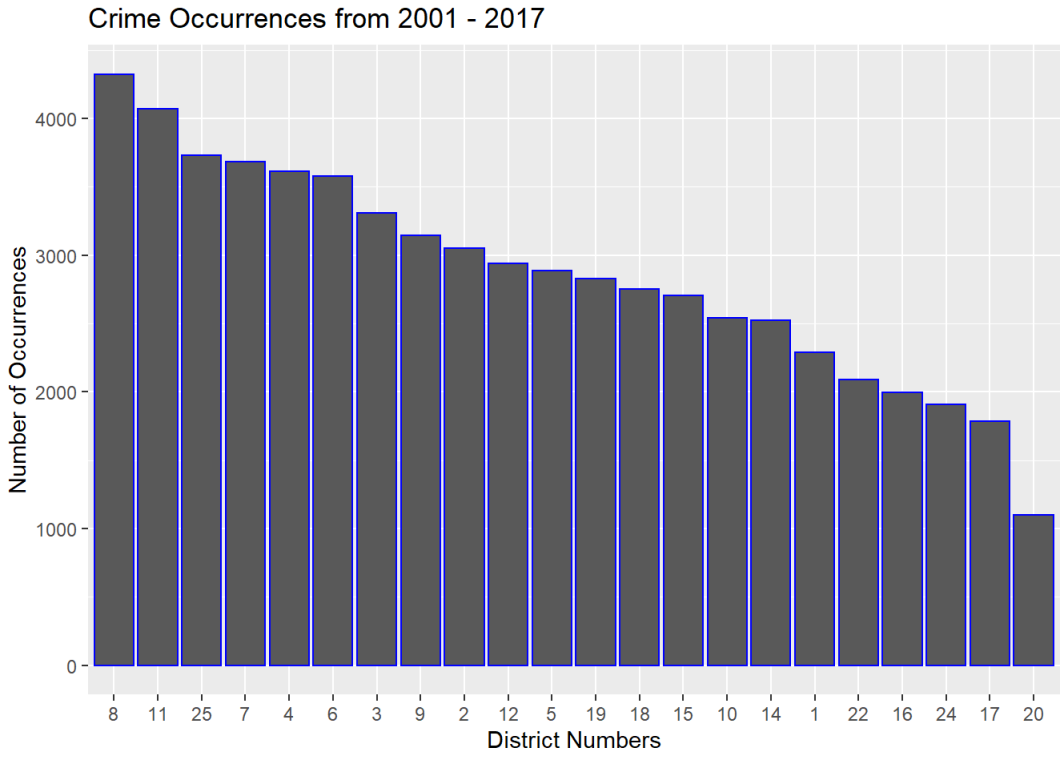
Police District Locations In Chicago



Crime Occurrence Distributions per District In Chicago from 2001 - 2017

From the bar graph, District 8 had the highest number of crime occurrences from 2001 - 2017, making it the most dangerous district with more than 4000 incidences. Oppositely, district 20 is the safest district in Chicago with just slightly more than 1000 incidences.

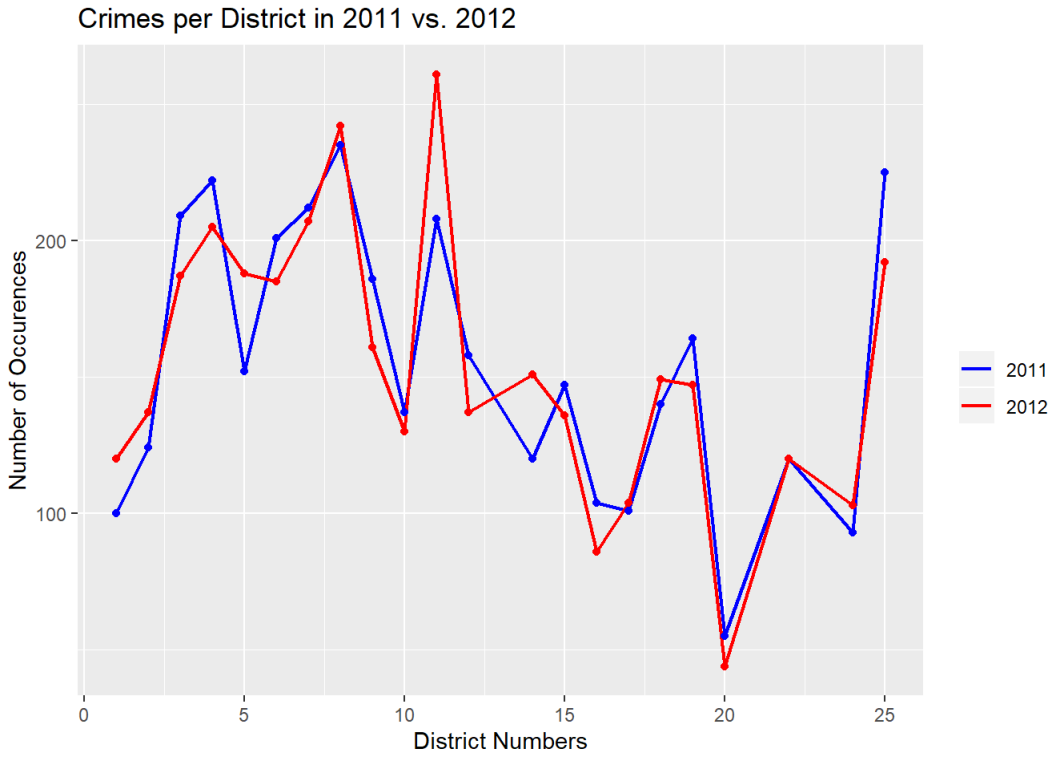
```
# Getting count for crime occurrences per district
crime_districts <- crime_data %>% count(District)

ggplot(data = crime_districts, aes(x= reorder(District, -n), y = n)) + geom_bar(stat = "identity", color = "blue")+ labs(title = "Crime Occurrences from 2001 - 2017", y = "Number of Occurrences", x = "District Numbers")
```



Crime Occurrences per District in 2011 vs. 2012

Based on the line plot comparing between crimes per district in 2011 and 2012. In 2012, district 11 had the highest crime rate whereas district 20 had the lowest crime rate. In 2011, the district that had the lowest crime rate was also district 20 and highest crime rate was district 8. Overall trend for number of crime occurrences per police district is roughly the same for the year 2011 and 2012.

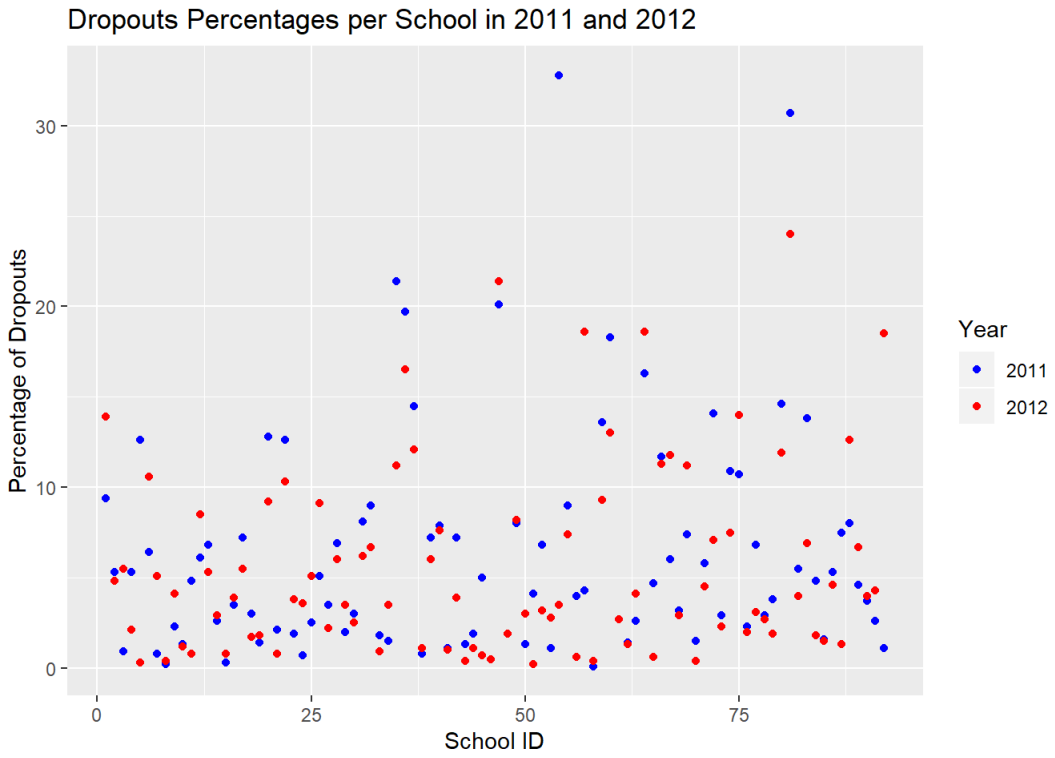
```
library(sqldf)
library(gsubfn)
library(proto)
library(RSQLite)

# crime dataset count for 2011
crime2011 <- sqldf("select * from crime_data where year > '2010' and year < '2012'")
crime2011_count <- crime2011 %>% count(District)

# crime dataset count for 2012
crime2012 <- sqldf("select * from crime_data where year > '2011' and year < '2013'")
crime2012_count <- crime2012 %>% count(District)

ggplot() +
  geom_line(data = crime2011_count, aes(x = District, y = n, color = "2011"), size = 0.8) +
  geom_line(data = crime2012_count, aes(x = District, y = n, color = "2012"), size = 0.8) +
  labs(title = "Crimes per District in 2011 vs. 2012", x = "District Numbers", y = "Number of Occurences") +
geom_point(data = crime2011_count, aes(x = District, y = n), color = "blue") + geom_point(data = crime2012_
count, aes(x = District, y = n), color = "red") + scale_colour_manual("",
                      breaks = c("2011", "2012"),
                      values = c("blue", "red"))
```



Dropouts Percentages per School in 2011 and 2012

The overall trend is hard to recognize whether 2011 has more percentage of dropout students compared to 2012 or not. However, majority of schools in 2012 saw a decrease in the percentage of dropout students (based on the majority of red data points are near the bottom of the graph).

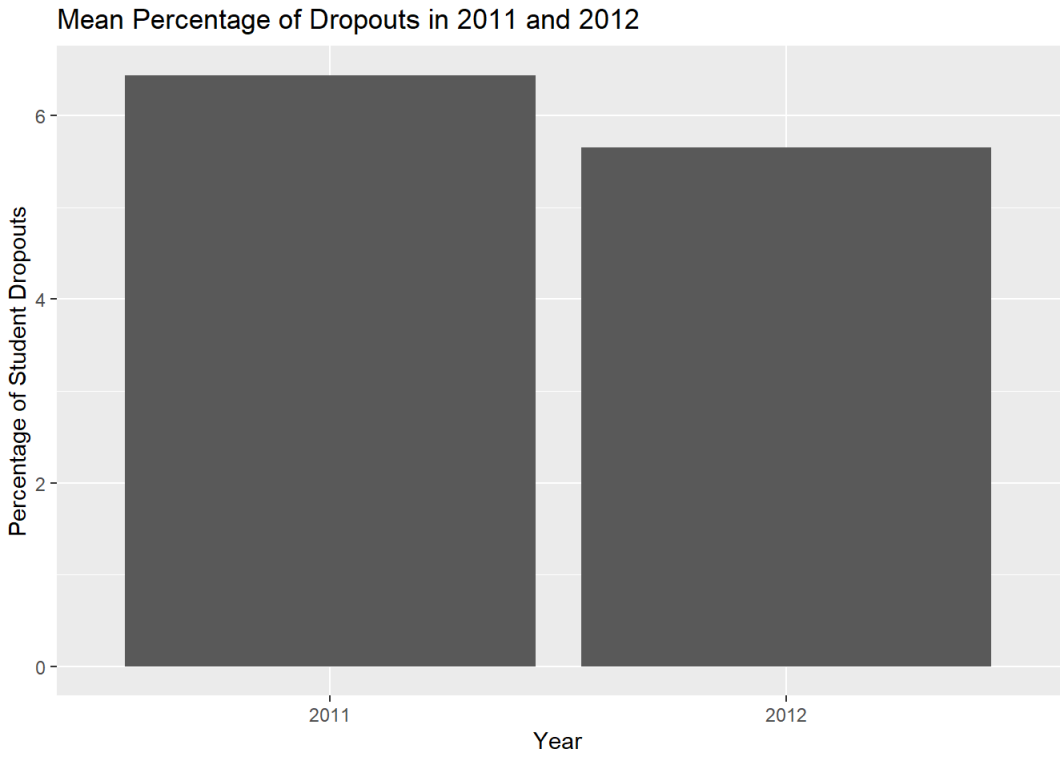
```
ggplot(data = school_data) +
  geom_point(aes(x = SchoolID, y = DropoutRate2011, color = "2011")) +
  geom_point(aes(x = SchoolID, y = DropoutRate2012, color = "2012")) + labs(title = "Dropouts Percentages p
er School in 2011 and 2012", x = "School ID", y = "Percentage of Dropouts")  + scale_colour_manual("Year",
                      breaks = c("2011", "2012"),
                      values = c("blue", "red"))
```



Mean Percentage of Dropouts in 2011 and 2012

Looking closer at this dataset comparison between percentage of dropouts in 2011 and 2012, the total mean percentage of dropouts in 2011 (6.44) is higher than the total mean percentage of dropouts in 2012 (5.66).

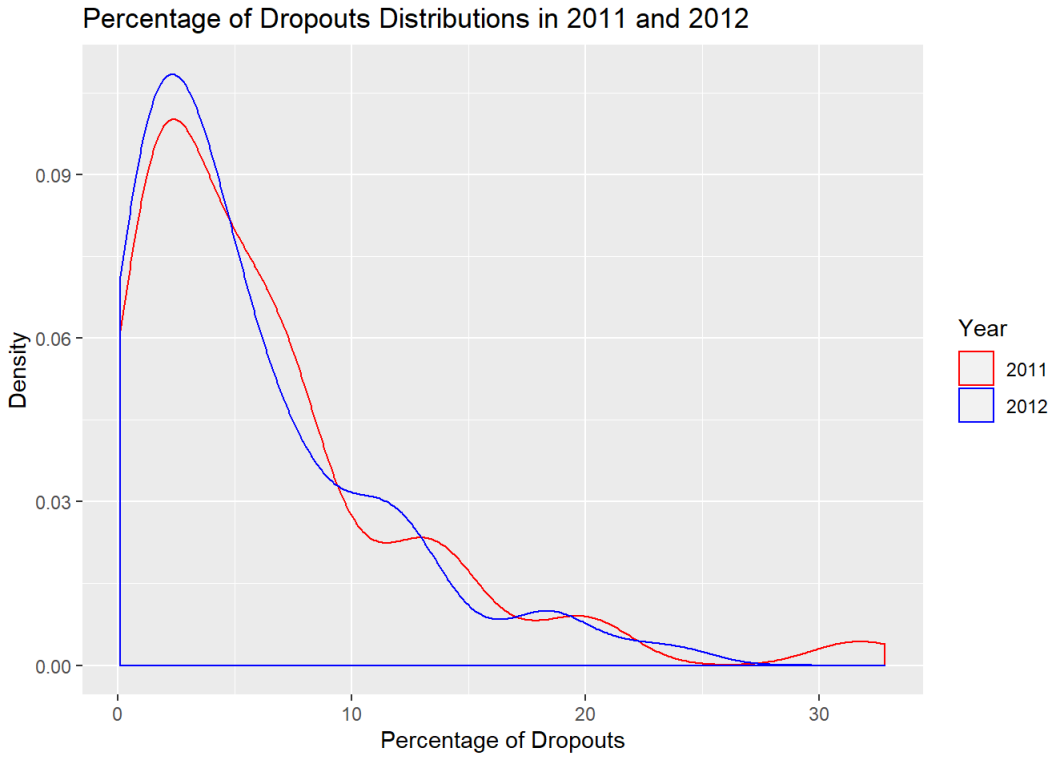
```
# mean drop out for 2011
meandropout2011 <- mean(school_data[["DropoutRate2011"]], na.rm = TRUE)

# mean drop out for 2012
meandropout2012 <- mean(school_data[["DropoutRate2012"]], na.rm = TRUE)

# create a data frame for these two columns
meandropout_total <-data.frame(
  Year = c("2011", "2012"),
  Value = c(meandropout2011, meandropout2012))

ggplot(data = meandropout_total, aes(x = Year, y = Value)) + geom_bar(stat = "identity") + labs(title = "Mea
n Percentage of Dropouts in 2011 and 2012", x = "Year", y = "Percentage of Student Dropouts")
```



Percentage of Dropouts Distributions in 2011 and 2012

The distribution of dropouts for both 2011 and 2012 can be seen as right-skewed.

```
# standard deviations for 2011 and 2012

sddropout2011 <- sd(school_data[["DropoutRate2011"]], na.rm = TRUE)
sddropout2012 <- sd(school_data[["DropoutRate2011"]], na.rm = TRUE)

ggplot(data = school_data) + geom_density(aes(x = DropoutRate2011, kernel = "gaussian", color = "2011", line
= 0.8)) + geom_density(aes(x = DropoutRate2012, kernel = "gaussian", color = "2012")) + scale_colour_manual
("Year",
                       breaks = c("2011", "2012"),
                       values = c("red", "blue")) + labs(title = "Percentage of Dropouts Distributions in 20
11 and 2012", x = "Percentage of Dropouts", y = "Density")
```



Regression Analysis to figure out relationship between Student Dropout Rates for 2012 and crime count for each Types of Crime

From the regression results, the percentage of student dropouts in 2012 did not have any influence on the total crime count for that year as well as the different types of crimes

```
crime2012_2 <- sqldf("select * from crime_data where year > '2011' and year < '2013' LIMIT 92")

df1 <- data.frame(crime2012_2)
df1$dropoutrate2012 <- school_data$DropoutRate2012

library(lm.beta)

crimeRegression <- lm( `dropoutrate2012` ~ ID + `CrimeType`,  data = df1)
summary(crimeRegression)
```

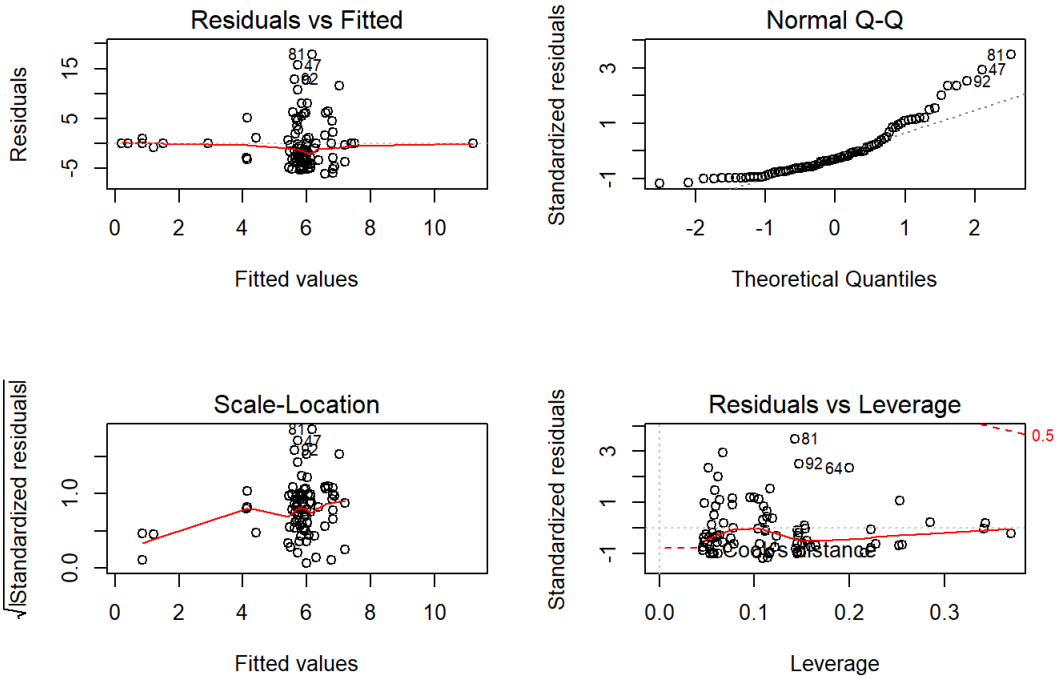
```
##
## Call:
## lm(formula = dropoutrate2012 ~ ID + CrimeType, data = df1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.164 -3.320 -1.181  1.214 17.824
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        1.389e+01  3.552e+01   0.391    0.697
## ID                                -9.011e-07  4.141e-06  -0.218    0.828
## CrimeTypeBATTERY                  -1.994e-01  2.430e+00  -0.082    0.935
## CrimeTypeBURGLARY                 -2.082e-01  2.960e+00  -0.070    0.944
## CrimeTypeCRIM SEXUAL ASSAULT      -2.934e+00  6.107e+00  -0.480    0.632
## CrimeTypeCRIMINAL DAMAGE          -4.712e-01  3.041e+00  -0.155    0.877
## CrimeTypeCRIMINAL TRESPASS         1.332e+00  5.907e+00   0.226    0.822
## CrimeTypeDECEPTIVE PRACTICE       -5.586e+00  5.968e+00  -0.936    0.352
## CrimeTypeGAMBLING                 -5.876e+00  5.923e+00  -0.992    0.324
## CrimeTypeLIQUOR LAW VIOLATION     -4.725e+00  5.912e+00  -0.799    0.427
## CrimeTypeMOTOR VEHICLE THEFT      -1.787e+00  3.565e+00  -0.501    0.618
## CrimeTypeNARCOTICS                 5.436e-01  2.728e+00   0.199    0.843
## CrimeTypeOFFENSE INVOLVING CHILDREN  4.969e+00  5.913e+00   0.840    0.403
## CrimeTypeOTHER OFFENSE             9.286e-01  3.256e+00   0.285    0.776
## CrimeTypeROBBERY                  -5.065e+00  3.866e+00  -1.310    0.194
## CrimeTypeTHEFT                    -3.052e-01  2.480e+00  -0.123    0.902
## CrimeTypeWEAPONS VIOLATION         1.373e+00  5.944e+00   0.231    0.818
##
## Residual standard error: 5.525 on 75 degrees of freedom
## Multiple R-squared:  0.08719,    Adjusted R-squared:  -0.1075
## F-statistic: 0.4477 on 16 and 75 DF,  p-value: 0.9633
```

```
coef_lmbeta1 <- lm.beta(crimeRegression)
coef_lmbeta1
```

```
##
## Call:
## lm(formula = dropoutrate2012 ~ ID + CrimeType, data = df1)
##
## Standardized Coefficients::
##                           (Intercept)                                    ID
##                            0.00000000                           -0.02744390
##                   CrimeTypeBATTERY                   CrimeTypeBURGLARY
##                           -0.01628885                           -0.01184694
##        CrimeTypeCRIM SEXUAL ASSAULT           CrimeTypeCRIMINAL DAMAGE
##                           -0.05826170                           -0.02392784
##          CrimeTypeCRIMINAL TRESPASS        CrimeTypeDECEPTIVE PRACTICE
##                            0.02645231                           -0.11093615
##                   CrimeTypeGAMBLING      CrimeTypeLIQUOR LAW VIOLATION
##                           -0.11668745                           -0.09383772
##        CrimeTypeMOTOR VEHICLE THEFT                 CrimeTypeNARCOTICS
##                           -0.06979377                            0.03240610
## CrimeTypeOFFENSE INVOLVING CHILDREN             CrimeTypeOTHER OFFENSE
##                            0.09868472                            0.04031890
##                   CrimeTypeROBBERY                    CrimeTypeTHEFT
##                           -0.17229492                           -0.02318863
##          CrimeTypeWEAPONS VIOLATION
##                            0.02726880
```

Plot the Crime Regression Results

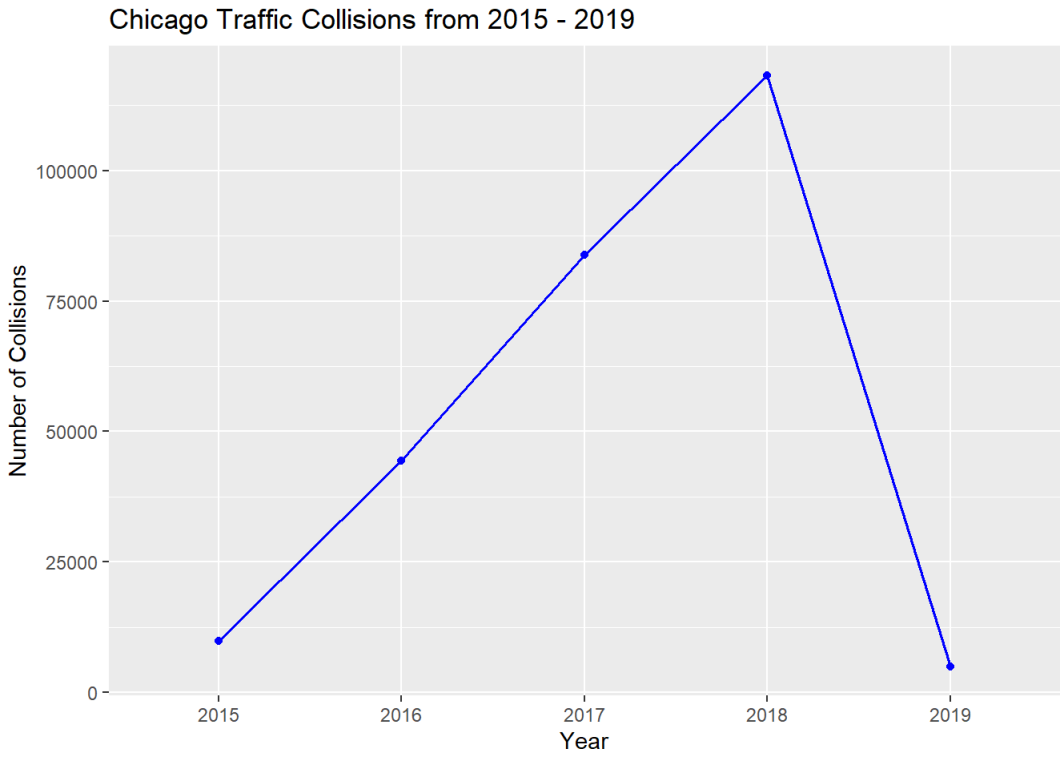
```
par(mfrow=c(2,2))
plot(crimeRegression)
```



2015 - 2019 Traffic Collisions in Chicago

2018 had the highest numbers of traffic collisions in Chicago, meanwhile 2019 had the lowest number of collisions.

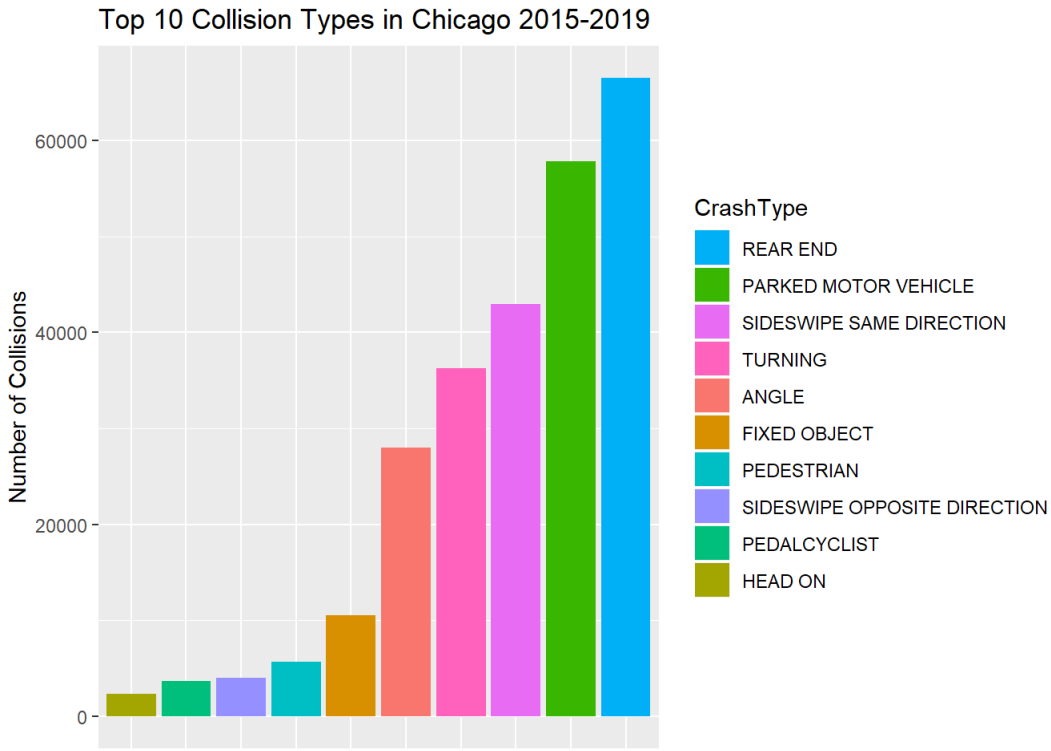
```
crash_count <- crash_data %>% count(Year)

ggplot(data = crash_count, aes(Year, n))+ geom_point(color = "blue")+ geom_line(group = 1, size = 0.7, color
= "blue")+ labs(title = "Chicago Traffic Collisions from 2015 - 2019", y = "Number of Collisions", x = "Yea
r")
```



Top 10 Traffic Collision Types in Chicago 2015 - 2019

From the figure, rear end collision occurred the most with more than 60,000 incidences from 2015 - 2019. Oppositely, head on collisions ranked the lowest among the top 10 collisions types.

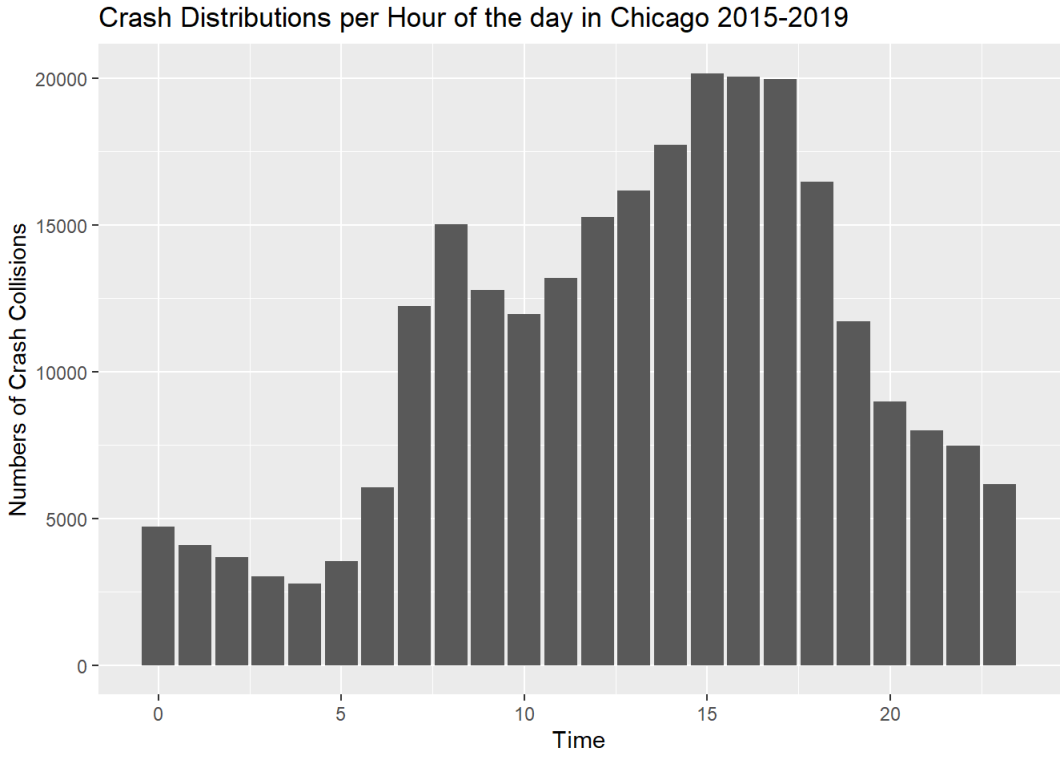
```
crashType_count <- crash_data %>% count(CrashType)%>%top_n(10)
legend_ord <- levels(with(crashType_count, reorder(CrashType, -n)))


ggplot(data = crashType_count, aes(x = reorder(CrashType, n), y = n, fill = CrashType)) + geom_bar(stat = "identity") + scale_fill_discrete(breaks=legend_ord) +  labs(title = "Top 10 Collision Types in Chicago 2015-2019", y = "Number of Collisions", x = "Collision Types") +  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



What Time of Day Do Most Collisions Occur in Chicago?

Based on the figure, the times in which traffic collisions occur the most from 2015-2019 were from 3-5pm, which translated to rush hours.

```
crashHour_count <- crash_data %>% count(CrashHour)
ggplot(data = crashHour_count, aes(x = CrashHour, y = n))+ geom_bar(stat = "identity") + labs(title = "Cras
h Distributions per Hour of the day in Chicago 2015-2019", x = "Time", y = "Numbers of Crash Collisions")
```



Relationships between collision occurrences and Months from 2015 - 2019

We can see, based on all the models and graphs, there is a positive relationship between collision occurrences and months of the year. Crashes tend to happen more in the month of October, November and December where Chicago experience winter season with snowy weather.

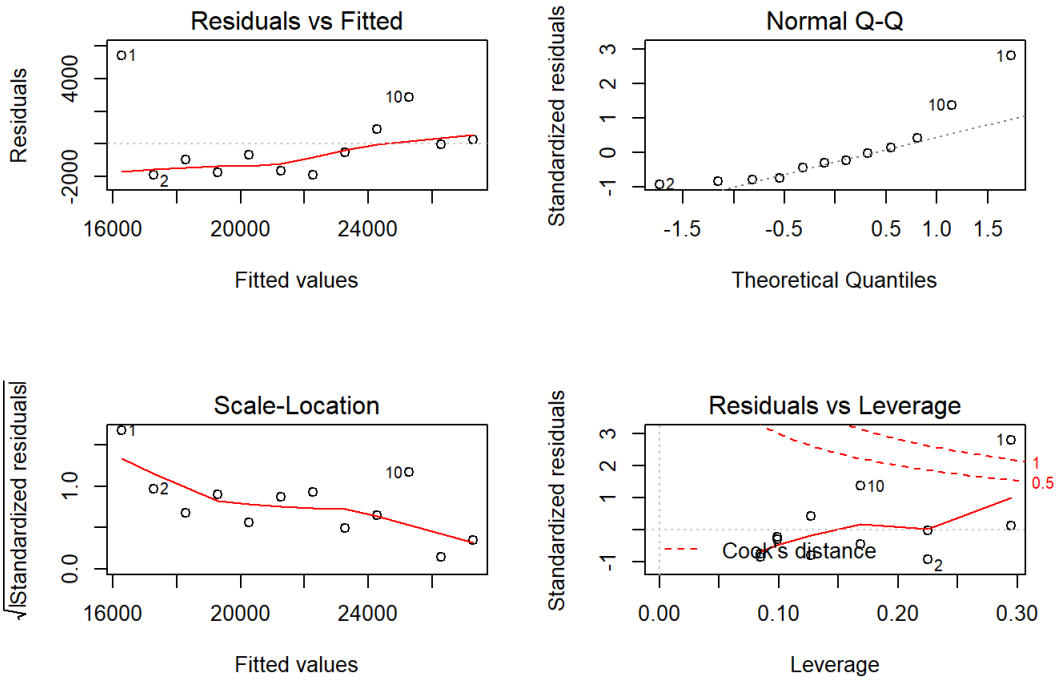
```
crashmonth_count <- crash_data %>% count(CrashMonth)


trafficregression <- lm( n~ CrashMonth,  data = crashmonth_count)
summary(trafficregression)
```

```
##
## Call:
## lm(formula = n ~ CrashMonth, data = crashmonth_count)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1907.4 -1691.7  -600.7   402.8  5417.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15264.0     1418.6  10.760 8.09e-07 ***
## CrashMonth    1001.2      192.7   5.194 0.000405 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2305 on 10 degrees of freedom
## Multiple R-squared:  0.7296, Adjusted R-squared:  0.7025
## F-statistic: 26.98 on 1 and 10 DF,  p-value: 0.0004046
```

```
coef_lmbeta1 <- lm.beta(trafficregression)
coef_lmbeta1
```

```
##
## Call:
## lm(formula = n ~ CrashMonth, data = crashmonth_count)
##
## Standardized Coefficients::
## (Intercept)  CrashMonth
##   0.0000000   0.8541606
```

```
par(mfrow=c(2,2))
plot(trafficregression)
```




```
predicted_df <- data.frame(n_predict = predict(trafficregression, crashmonth_count))

df_crashmont_count <- data.frame(crashmonth_count)
df_crashmont_count$n_predict <- predicted_df$n_predict

ggplot(data = df_crashmont_count) + geom_line(aes(x = CrashMonth, y = n_predict, color = "blue")) + geom_po
int(data = df_crashmont_count, aes(x = CrashMonth, y = n_predict) , color = "blue") +


  geom_line(aes(x = CrashMonth, y = n, color = "red")) + labs(color = "Legend\n", title = "Actual vs. Predi
cted Linear Regression Model\n
Crash Occurrences per month in Chicago from 2015-2019", x = "Months", y = "Collision Occurrences" ) + scale_
color_manual(labels = c("Predicted Model", "Actual Model"), values = c("blue", "red")) + geom_point(data =
df_crashmont_count, aes(x = CrashMonth, y = n) , color = "red")
```

Actual vs. Predicted Linear Regression Model

Crash Occurrences per month in Chicago from 2015-2019