DSC630 Final Project - Crime Analysis - Part1

Amie Davis

27 June, 2020

Data Sources:

Uniform Crime Reporting Program Data: National Incident-Based Reporting System, [United States], 2016; United States Federal Bureau of Investigation; Inter-university Consortium for Political and Social Research (ICPSR), University of Michigan; https://www.icpsr.umich.edu/icpsrweb/NACJD/NIBRS/
Geodetic Data for US Cities: https://simplemaps.com/data/us-cities

Load Libraries

```
library(readr)
library(ggplot2)
library(lubridate)
library(dplyr)
require(scales)
```

1. Prepare Data

a) Import the Data

```
# Load victim data
vic_data <- read_tsv("Data/UCR_2016/Victim_Segment/37065-0005-Data.tsv",
  col_types = cols(
  .default = col_character(),
  V4024 = col_double(),
  V4010 = col_character()
) )

# Load precinct data
bat_data <- read_tsv("Data/UCR_2016/Batch_Header_Segment/37065-0001-
Data.tsv")
```

b) Review Features

```
str(vic_data)

str(bat_data)
```

c) Derived Features

```
# Convert date fields to date stamps
vic_date_data <- vic_data %>% mutate(V4005 = ymd(V4005))

# Split date fields into separate columns using lubridate package
vic_new_data <- vic_date_data %>% mutate (VIC_INC_YEAR = year(V4005),
                                  VIC_INC_MONTH = month(V4005),
                                  VIC_INC_DAY = day(V4005),
                                  VIC_INC_DOW = weekdays(V4005))

# Convert categorical variables to factors
vic_new_data$VIC_INC_MONTH <- factor(vic_new_data$VIC_INC_MONTH)
vic_new_data$VIC_INC_DOW <- factor(vic_new_data$VIC_INC_DOW)

vic_new_data$V4007 <- factor(vic_new_data$V4007)
vic_new_data$V4008 <- factor(vic_new_data$V4008)
vic_new_data$V4009 <- factor(vic_new_data$V4009)
vic_new_data$V4010 <- factor(vic_new_data$V4010)
vic_new_data$V4011 <- factor(vic_new_data$V4011)
vic_new_data$V4012 <- factor(vic_new_data$V4012)
vic_new_data$V4013 <- factor(vic_new_data$V4013)
vic_new_data$V4014 <- factor(vic_new_data$V4014)
vic_new_data$V4015 <- factor(vic_new_data$V4015)
vic_new_data$V4016 <- factor(vic_new_data$V4016)
vic_new_data$V4017 <- factor(vic_new_data$V4017)
vic_new_data$V4017A <- factor(vic_new_data$V4017A)
vic_new_data$V4017B <- factor(vic_new_data$V4017B)
vic_new_data$V4019 <- factor(vic_new_data$V4019)
vic_new_data$V4020 <- factor(vic_new_data$V4020)
vic_new_data$V4021 <- factor(vic_new_data$V4021)
vic_new_data$V4022 <- factor(vic_new_data$V4022)
vic_new_data$V4023 <- factor(vic_new_data$V4023)
vic_new_data$V4024 <- factor(vic_new_data$V4024)
vic_new_data$V4025 <- factor(vic_new_data$V4025)
vic_new_data$V4026 <- factor(vic_new_data$V4026)
vic_new_data$V4027 <- factor(vic_new_data$V4027)
vic_new_data$V4028 <- factor(vic_new_data$V4028)
vic_new_data$V4029 <- factor(vic_new_data$V4029)
vic_new_data$V4030 <- factor(vic_new_data$V4030)
vic_new_data$V4032 <- factor(vic_new_data$V4032)
vic_new_data$V4034 <- factor(vic_new_data$V4034)
vic_new_data$V4036 <- factor(vic_new_data$V4036)
vic_new_data$V4038 <- factor(vic_new_data$V4038)
vic_new_data$V4040 <- factor(vic_new_data$V4040)
vic_new_data$V4042 <- factor(vic_new_data$V4042)
vic_new_data$V4044 <- factor(vic_new_data$V4044)
vic_new_data$V4046 <- factor(vic_new_data$V4046)
vic_new_data$V4048 <- factor(vic_new_data$V4048)
vic_new_data$V4050 <- factor(vic_new_data$V4050)
```

```
#head(vic_new_data)

# Convert categorical variables to factors

# Location Groupings
bat_data$BH002 <- factor(bat_data$BH002)
bat_data$BH008 <- factor(bat_data$BH008)
bat_data$BH009 <- factor(bat_data$BH009)
bat_data$BH010 <- factor(bat_data$BH010)
bat_data$BH011 <- factor(bat_data$BH011)
bat_data$BH012 <- factor(bat_data$BH012)
bat_data$BH013 <- factor(bat_data$BH013)

# Fed District
bat_data$BH015 <- factor(bat_data$BH015)
bat_data$BH016 <- factor(bat_data$BH016)

# MSA and Country Codes
bat_data$BH020 <- factor(bat_data$BH020)
bat_data$BH021 <- factor(bat_data$BH021)
bat_data$BH023 <- factor(bat_data$BH023)
bat_data$BH025 <- factor(bat_data$BH025)
bat_data$BH028 <- factor(bat_data$BH028)
bat_data$BH029 <- factor(bat_data$BH029)
bat_data$BH032 <- factor(bat_data$BH032)
bat_data$BH033 <- factor(bat_data$BH033)
bat_data$BH036 <- factor(bat_data$BH036)
bat_data$BH037 <- factor(bat_data$BH037)

# FIPS County Codes
bat_data$BH054 <- factor(bat_data$BH054)
bat_data$BH055 <- factor(bat_data$BH055)
bat_data$BH056 <- factor(bat_data$BH056)
bat_data$BH057 <- factor(bat_data$BH057)
bat_data$BH058 <- factor(bat_data$BH058)

#head(bat_data)
```

d) Re-Label data fields

```
names(vic_new_data)[names(vic_new_data) == "V4001"] <- "ARR_SEG_LEVEL"   #NOT
NEEDED - ALWAYS 4 FOR VICTIM FILE
names(vic_new_data)[names(vic_new_data) == "V4002"] <- "VIC_STATE_CODE"
names(vic_new_data)[names(vic_new_data) == "V4003"] <- "ORI"
names(vic_new_data)[names(vic_new_data) == "V4004"] <- "INC_NUM"
names(vic_new_data)[names(vic_new_data) == "V4005"] <- "VIC_INC_DATE" #DATE
names(vic_new_data)[names(vic_new_data) == "V4006"] <- "VIC_SEQ_NUM"

names(vic_new_data)[names(vic_new_data) == "V4007"] <- "OFF_CODE01"
```

```r
names(vic_new_data)[names(vic_new_data) == "V4008"] <- "OFF_CODE02"
names(vic_new_data)[names(vic_new_data) == "V4009"] <- "OFF_CODE03"
names(vic_new_data)[names(vic_new_data) == "V4010"] <- "OFF_CODE04"
names(vic_new_data)[names(vic_new_data) == "V4011"] <- "OFF_CODE05"
names(vic_new_data)[names(vic_new_data) == "V4012"] <- "OFF_CODE06"
names(vic_new_data)[names(vic_new_data) == "V4013"] <- "OFF_CODE07"
names(vic_new_data)[names(vic_new_data) == "V4014"] <- "OFF_CODE08"
names(vic_new_data)[names(vic_new_data) == "V4015"] <- "OFF_CODE09"
names(vic_new_data)[names(vic_new_data) == "V4016"] <- "OFF_CODE10"
names(vic_new_data)[names(vic_new_data) == "V4017"] <- "VICTIM_TYPE"
#CAT - 9
names(vic_new_data)[names(vic_new_data) == "V4017A"] <- "ACT_TYPE_OFFC"
#CAT - 11 (numeric)
names(vic_new_data)[names(vic_new_data) == "V4017B"] <- "ASSG_TYPE_OFFC"
#CAT - 7
names(vic_new_data)[names(vic_new_data) == "V4017C"] <- "ORI_OTHER"

names(vic_new_data)[names(vic_new_data) == "V4018"] <- "AGE_OF_VICTIM"
#NUM (00-UNK, 99-99+)
names(vic_new_data)[names(vic_new_data) == "V4019"] <- "SEX_OF_VICTIM"
#CAT (M/F)
names(vic_new_data)[names(vic_new_data) == "V4020"] <- "RACE_OF_VICTIM"
#CAT - 5 VALUES, U=UNK
names(vic_new_data)[names(vic_new_data) == "V4021"] <- "ETHNIC_OF_VIC"
#CAT (H/N/U)
names(vic_new_data)[names(vic_new_data) == "V4022"] <- "VIC_RESIDENT"
#CAT (R/N/U)

names(vic_new_data)[names(vic_new_data) == "V4023"] <- "ASSAULT_CIRC1"
#CAT - 17
names(vic_new_data)[names(vic_new_data) == "V4024"] <- "ASSAULT_CIRC2"
#CAT - 17
names(vic_new_data)[names(vic_new_data) == "V4025"] <- "JUST_HOM_CIRC"
#CAT - 7

names(vic_new_data)[names(vic_new_data) == "V4026"] <- "INJURY_TYPE1"
#CAT - 8
names(vic_new_data)[names(vic_new_data) == "V4027"] <- "INJURY_TYPE2"
#CAT - 8
names(vic_new_data)[names(vic_new_data) == "V4028"] <- "INJURY_TYPE3"
#CAT - 8
names(vic_new_data)[names(vic_new_data) == "V4029"] <- "INJURY_TYPE4"
#CAT - 8
names(vic_new_data)[names(vic_new_data) == "V4030"] <- "INJURY_TYPE5"
#CAT - 8

names(vic_new_data)[names(vic_new_data) == "V4031"] <- "OFF_NUM_KEY1"
names(vic_new_data)[names(vic_new_data) == "V4032"] <- "REL_TO_OFF1"
#CAT - 26
```

```r
names(vic_new_data)[names(vic_new_data) == "V4033"] <- "OFF_NUM_KEY2"
names(vic_new_data)[names(vic_new_data) == "V4034"] <- "REL_TO_OFF2"
#CAT - 26
names(vic_new_data)[names(vic_new_data) == "V4035"] <- "OFF_NUM_KEY3"
names(vic_new_data)[names(vic_new_data) == "V4036"] <- "REL_TO_OFF3"
#CAT - 26
names(vic_new_data)[names(vic_new_data) == "V4037"] <- "OFF_NUM_KEY4"
names(vic_new_data)[names(vic_new_data) == "V4038"] <- "REL_TO_OFF4"
#CAT - 26
names(vic_new_data)[names(vic_new_data) == "V4039"] <- "OFF_NUM_KEY5"
names(vic_new_data)[names(vic_new_data) == "V4040"] <- "REL_TO_OFF5"
#CAT - 26
names(vic_new_data)[names(vic_new_data) == "V4041"] <- "OFF_NUM_KEY6"
names(vic_new_data)[names(vic_new_data) == "V4042"] <- "REL_TO_OFF6"
#CAT - 26
names(vic_new_data)[names(vic_new_data) == "V4043"] <- "OFF_NUM_KEY7"
names(vic_new_data)[names(vic_new_data) == "V4044"] <- "REL_TO_OFF7"
#CAT - 26
names(vic_new_data)[names(vic_new_data) == "V4045"] <- "OFF_NUM_KEY8"
names(vic_new_data)[names(vic_new_data) == "V4046"] <- "REL_TO_OFF8"
#CAT - 26
names(vic_new_data)[names(vic_new_data) == "V4047"] <- "OFF_NUM_KEY9"
names(vic_new_data)[names(vic_new_data) == "V4048"] <- "REL_TO_OFF9"
#CAT - 26
names(vic_new_data)[names(vic_new_data) == "V4049"] <- "OFF_NUM_KEY10"
names(vic_new_data)[names(vic_new_data) == "V4050"] <- "REL_TO_OFF10"
#CAT - 26

names(vic_new_data)[names(vic_new_data) == "V4051"] <- "NUM_RECS_PER_VICTIM"

#str(vic_new_data)

names(bat_data)[names(bat_data) == "BH002"] <- "NUM_STATE_CODE"    # CAT
(numeric)
names(bat_data)[names(bat_data) == "BH006"] <- "DATE_WENT_NBIRS"   # DATE
names(bat_data)[names(bat_data) == "BH003"] <- "ORI"
names(bat_data)[names(bat_data) == "BH007"] <- "CITY_NAME"
names(bat_data)[names(bat_data) == "BH008"] <- "STATE_ABBR"        #CAT
names(bat_data)[names(bat_data) == "BH009"] <- "POP_GROUP"         #CAT - 23
names(bat_data)[names(bat_data) == "BH010"] <- "CTRY_DIVISION"     #CAT - 10
names(bat_data)[names(bat_data) == "BH011"] <- "CTRY_REGION"       #CAT - 5
names(bat_data)[names(bat_data) == "BH012"] <- "AGENCY_IND"        #CAT - 9
(numeric)
names(bat_data)[names(bat_data) == "BH013"] <- "CORE_CITY"         #CAT (Y/N)
names(bat_data)[names(bat_data) == "BH015"] <- "FBI_OFFICE"
names(bat_data)[names(bat_data) == "BH016"] <- "JUDICIAL_DIST"
names(bat_data)[names(bat_data) == "BH019"] <- "CURRENT_POP1"
names(bat_data)[names(bat_data) == "BH020"] <- "UCR_COUNTY_CD1"
names(bat_data)[names(bat_data) == "BH021"] <- "MSA_CD1"
names(bat_data)[names(bat_data) == "BH022"] <- "LAST_POP1"
```

```
names(bat_data)[names(bat_data) == "BH023"] <- "CURRENT_POP2"
names(bat_data)[names(bat_data) == "BH023"] <- "UCR_COUNTY_CD2"
names(bat_data)[names(bat_data) == "BH025"] <- "MSA_CD2"
names(bat_data)[names(bat_data) == "BH026"] <- "LAST_POP2"
names(bat_data)[names(bat_data) == "BH027"] <- "CURRENT_POP3"
names(bat_data)[names(bat_data) == "BH028"] <- "UCR_COUNTY_CD3"
names(bat_data)[names(bat_data) == "BH029"] <- "MSA_CD3"
names(bat_data)[names(bat_data) == "BH030"] <- "LAST_POP3"
names(bat_data)[names(bat_data) == "BH031"] <- "CURRENT_POP4"
names(bat_data)[names(bat_data) == "BH032"] <- "UCR_COUNTY_CD4"
names(bat_data)[names(bat_data) == "BH033"] <- "MSA_CD4"
names(bat_data)[names(bat_data) == "BH034"] <- "LAST_POP4"
names(bat_data)[names(bat_data) == "BH035"] <- "CURRENT_POP5"
names(bat_data)[names(bat_data) == "BH036"] <- "UCR_COUNTY_CD5"
names(bat_data)[names(bat_data) == "BH037"] <- "MSA_CD5"
names(bat_data)[names(bat_data) == "BH038"] <- "LAST_POP5"
names(bat_data)[names(bat_data) == "BH054"] <- "FIPS_COUNTY1"
names(bat_data)[names(bat_data) == "BH055"] <- "FIPS_COUNTY2"
names(bat_data)[names(bat_data) == "BH056"] <- "FIPS_COUNTY3"
names(bat_data)[names(bat_data) == "BH057"] <- "FIPS_COUNTY4"
names(bat_data)[names(bat_data) == "BH058"] <- "FIPS_COUNTY5"

#str(bat_data)
```

e) Join Datasets

```
# Join victim data with ORI reference data
comb_df <- merge(vic_new_data, bat_data, by="ORI")

#head(comb_df)
#summary(comb_df)
```

f) Drop unneeded columns

```
# After analysis of the codebook, the following fields are not needed and
will be removed.
# Dropping file id number, since it is a constant.
# Using state code from batch file vice victim file
# Only using location information from batch file
# Relationship victim has to offender is out of scope of this project, so the
related fields will be removed.
# Each reporting district can associate up to 5 metropolitan areas.  For the
scope of this project, I will limit to the 1st area.

  comb_df[ ,c(
    "ARR_SEG_LEVEL",
    "VIC_SEQ_NUM",
    "VIC_STATE_CODE",
    "ORI_OTHER",
    "OFF_NUM_KEY1",
    "OFF_NUM_KEY2",
```

```
    "OFF_NUM_KEY3",
    "OFF_NUM_KEY4",
    "OFF_NUM_KEY5",
    "OFF_NUM_KEY6",
    "OFF_NUM_KEY7",
    "OFF_NUM_KEY8",
    "OFF_NUM_KEY9",
    "OFF_NUM_KEY10",
    "REL_TO_OFF1",
    "REL_TO_OFF2",
    "REL_TO_OFF3",
    "REL_TO_OFF4",
    "REL_TO_OFF5",
    "REL_TO_OFF6",
    "REL_TO_OFF7",
    "REL_TO_OFF8",
    "REL_TO_OFF9",
    "REL_TO_OFF10",
    "BH001",
    "BH004",
    "BH005",
    "BH014",
    "BH017",
    "BH018",
    "BH024",
    "BH039",
    "BH040",
    "BH041",
    "BH042",
    "BH043",
    "BH044",
    "BH045",
    "BH046",
    "BH047",
    "BH048",
    "BH049",
    "BH050",
    "BH051",
    "BH052",
    "BH053",
    "BH059",
    "BH060",
    "DATE_WENT_NBIRS",
    "CURRENT_POP2",
    "UCR_COUNTY_CD2",
    "MSA_CD2",
    "LAST_POP2",
    "CURRENT_POP3",
    "UCR_COUNTY_CD3",
    "MSA_CD3",
```

```
    "LAST_POP3",
    "CURRENT_POP4",
    "UCR_COUNTY_CD4",
    "MSA_CD4",
    "LAST_POP4",
    "CURRENT_POP5",
    "UCR_COUNTY_CD5",
    "MSA_CD5",
    "LAST_POP5",
    "FIPS_COUNTY2",
    "FIPS_COUNTY3",
    "FIPS_COUNTY4",
    "FIPS_COUNTY5"
)] <- list(NULL)
```

g) Summary Statistics

```
summary(comb_df)
```

Observations:
6034725 Victim Records
Some fields have nothing but N/A values
Outliers in Victim Date Field

```
# After analysis of the summary statistics, the following fields are not
needed and will be removed.
# Removing fields without data - only N/A values exist

  comb_df[ ,c(
    "OFF_CODE08",
    "OFF_CODE09",
    "OFF_CODE10",
    "CURRENT_POP4",
    "UCR_COUNTY_CD4",
    "MSA_CD4",
    "LAST_POP4",
    "CURRENT_POP5",
    "UCR_COUNTY_CD5",
    "MSA_CD5",
    "LAST_POP5",
    "FIPS_COUNTY4",
    "FIPS_COUNTY5"
)] <- list(NULL)

#summary(comb_df)
```

h) Remove Outliers

```
# Remove outliers with incident dates from 2015.  Look at 2016 incidents
only.
clean_data <- subset(comb_df, VIC_INC_YEAR == 2016)

summary(clean_data$VIC_INC_DATE)

##         Min.      1st Qu.       Median         Mean      3rd Qu.
Max.
## "2016-01-01" "2016-04-06" "2016-07-04" "2016-07-02" "2016-09-29" "2016-12-
31"

#str(clean_data)
#5951120 records
```

i) Handle NA Values in Numeric Fields

```
# 1) Victim's Age
clean_data$AGE_OF_VICTIM <- as.numeric(clean_data$AGE_OF_VICTIM)
summary(clean_data$AGE_OF_VICTIM)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    25.0    36.0    38.7    51.0    99.0 1759383

# Distribution is slightly skewed towards mean, so will impute with median
# Impute NAs with median age
clean_data$AGE_OF_VICTIM[is.na(clean_data$AGE_OF_VICTIM)] <-
median(clean_data$AGE_OF_VICTIM, na.rm = TRUE)

print('Imputed Summary')

## [1] "Imputed Summary"

summary(clean_data$AGE_OF_VICTIM)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    30.0    36.0    37.9    45.0    99.0

# 2) Population Fields
summary(comb_df$CURRENT_POP1)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   22346   68434  173074  209914 1105798

summary(comb_df$LAST_POP1)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   21966   67665  171378  206884 1106066

# Replace 0 with mean population
clean_data$CURRENT_POP1[clean_data$CURRENT_POP1 == 0] <- NA
clean_data$LAST_POP1[clean_data$LAST_POP1 == 0] <- NA
```

```r
# Impute with mean population
clean_data$CURRENT_POP1[is.na(clean_data$CURRENT_POP1)] <-
mean(clean_data$CURRENT_POP1, na.rm = TRUE)
clean_data$LAST_POP1[is.na(clean_data$LAST_POP1)] <-
mean(clean_data$LAST_POP1, na.rm = TRUE)

print('Imputed Summary')
```

```
## [1] "Imputed Summary"
```

```r
summary(clean_data$CURRENT_POP1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       8   27113   80215  180044  209914 1105798
```

```r
summary(clean_data$LAST_POP1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       8   27100   79461  178357  206884 1106066
```

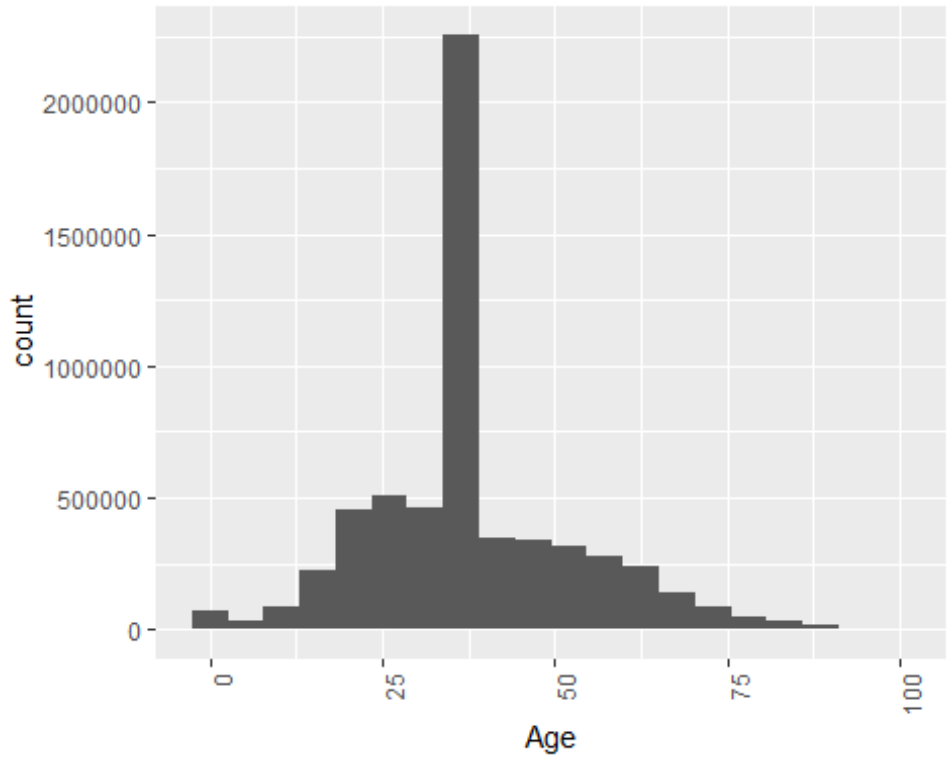
j) Apply filters

```r
# Create subset to include only records marked as justifiable homicide
just_hom_df <- filter(clean_data, OFF_CODE01 == "09C" | OFF_CODE02 == "09C" |
OFF_CODE03 == "09C" |
                      OFF_CODE04 == "09C" | OFF_CODE05 == "09C" | OFF_CODE06 ==
"09C" | OFF_CODE07 == "09C")

#summary(just_hom_df)
# There are 217 incidents of justifiable homicide reported to UCR in 2016

# Create subset to include only records marked as aggravated assault
agg_asst_df <- filter(clean_data, OFF_CODE01 == "13A" | OFF_CODE02 == "13A" |
OFF_CODE03 == "13A" |
                      OFF_CODE04 == "13A" | OFF_CODE05 == "13A" | OFF_CODE06 ==
"13A" | OFF_CODE07 == "13A")

#summary(agg_asst_df)
# There are 235,811 incidents of aggravated assaults reported to UCR in 2016
```

2. EDA - Review Distributions

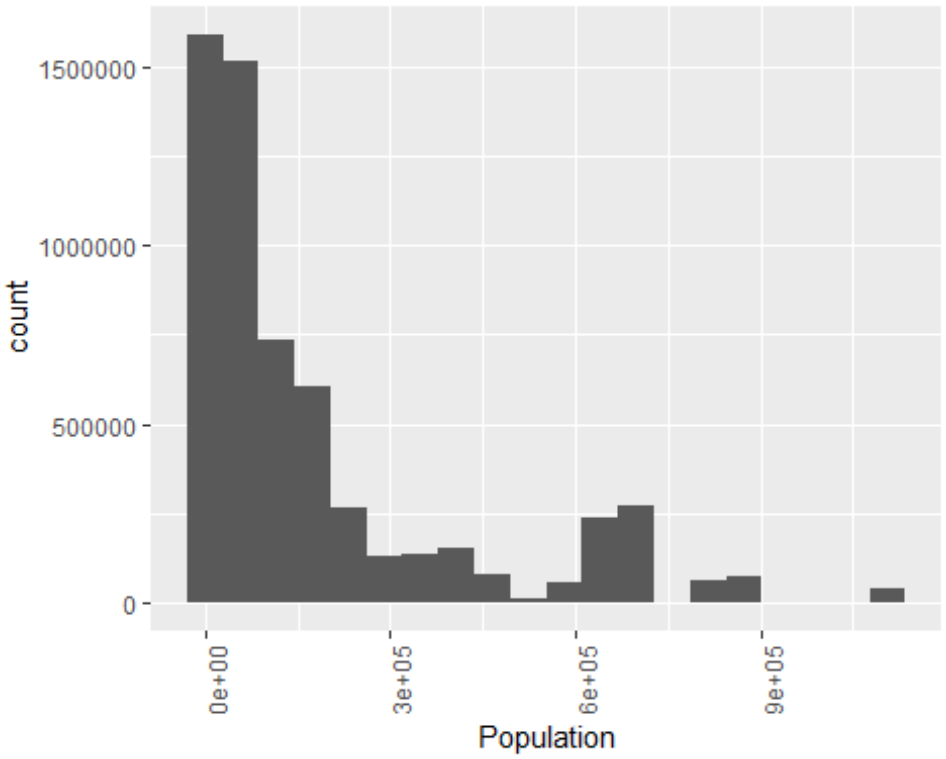
a) Plot Histograms for Numeric Vars

```r
ggplot(clean_data, aes(x=AGE_OF_VICTIM)) +
    geom_histogram(bins=20) +
    labs(x="Age") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



The peak is inflated to to imputation of na values. Distribution is fairly normal, but is slightly left-skewed.

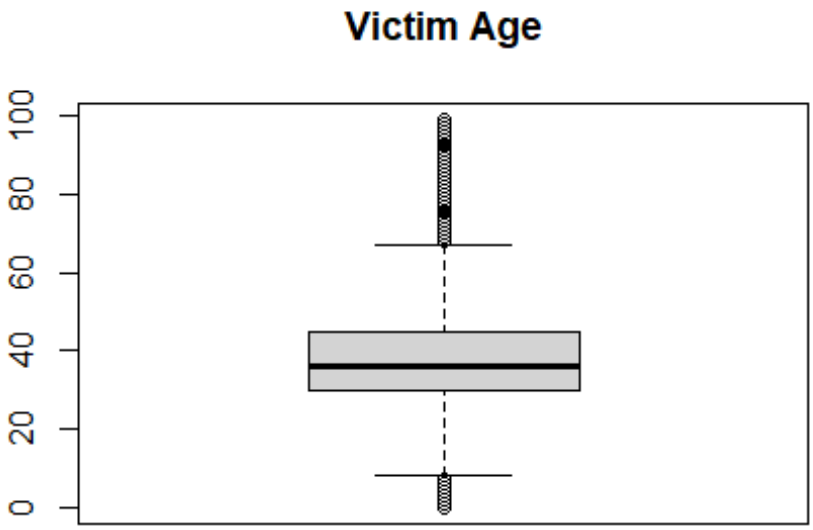
```
ggplot(clean_data, aes(x=CURRENT_POP1)) +
    geom_histogram(bins=20) +
    labs(x="Population") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



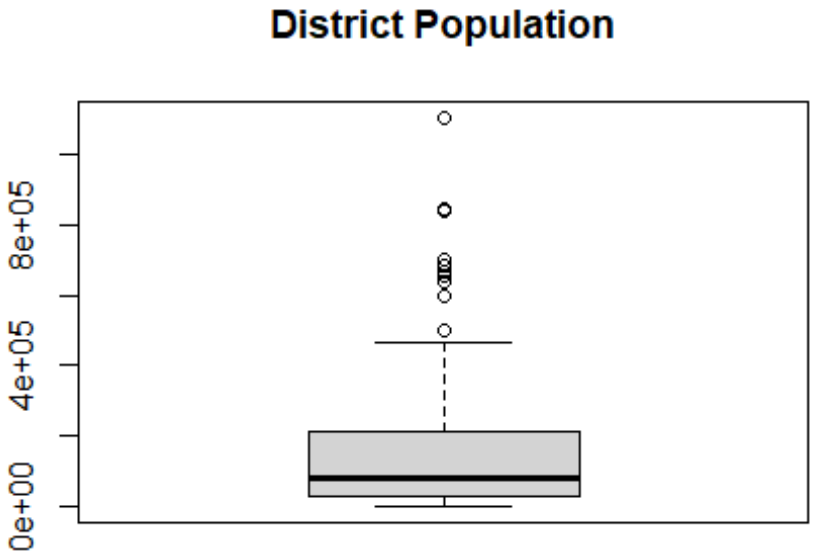
Distribution is left-skewed.

b) Box Plots for Numeric Features

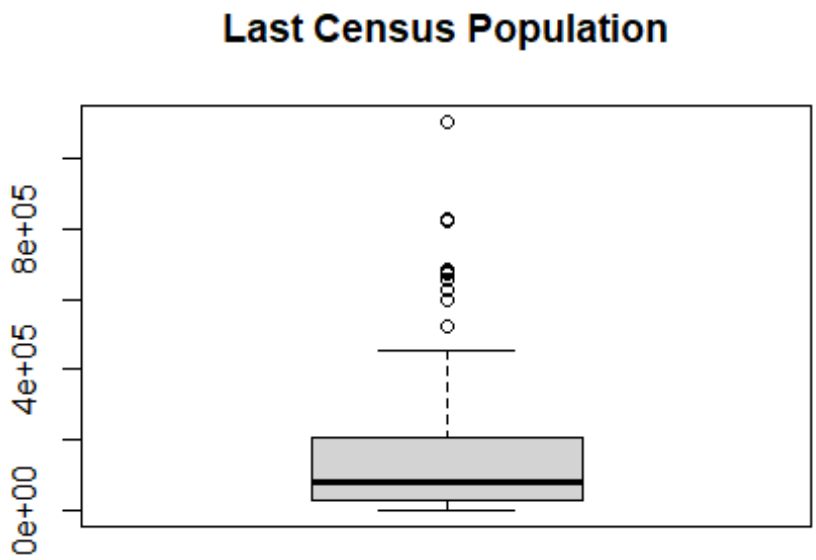
```
boxplot(clean_data$AGE_OF_VICTIM,
  main="Victim Age")
```




```r
boxplot(clean_data$CURRENT_POP1,
  main="District Population")
```

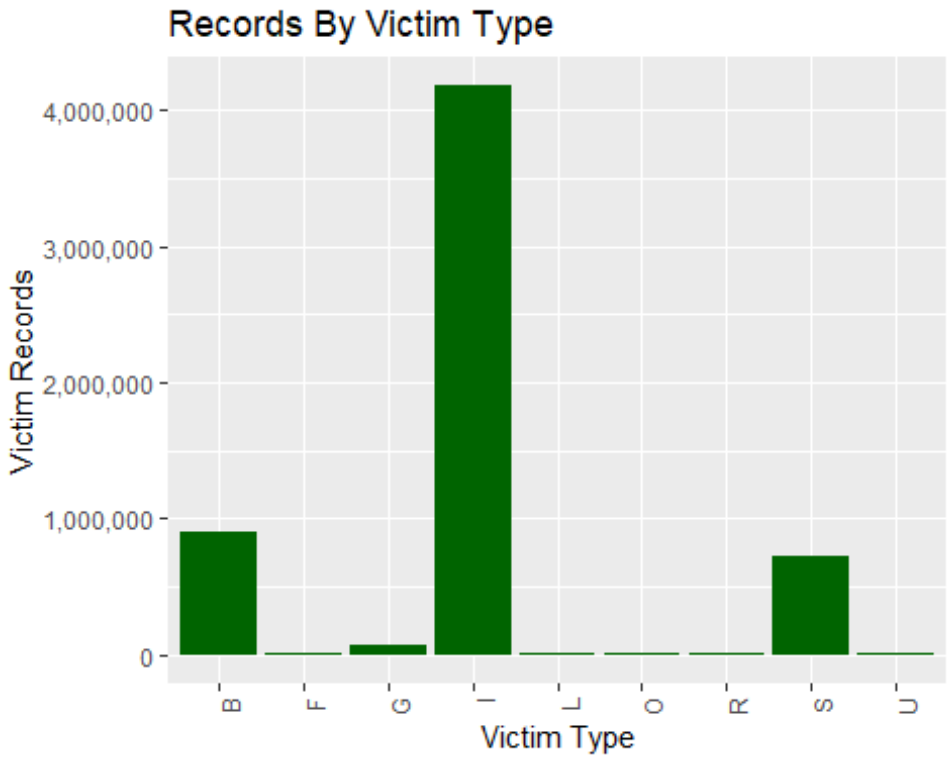


```r
boxplot(clean_data$LAST_POP1,
  main="Last Census Population")
```



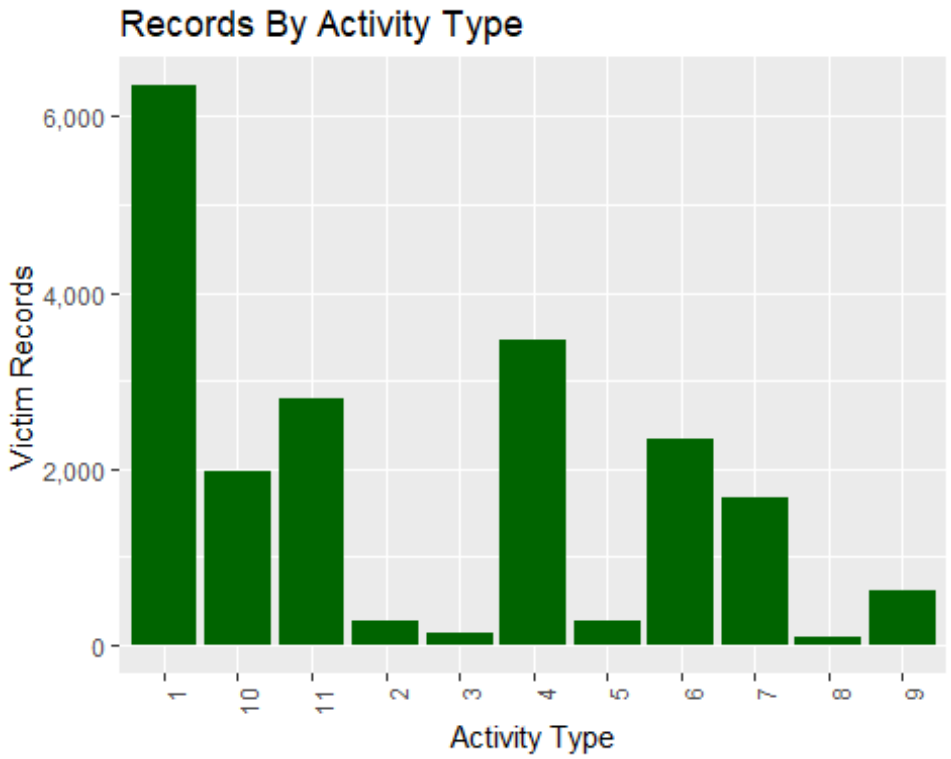
c) Histograms for Categorical Features

```r
# Victim Type
p <- ggplot(clean_data, aes(x=VICTIM_TYPE)) +
     geom_bar(fill="dark green") +
     labs(x="Victim Type", y="Victim Records", title="Records By Victim
Type") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



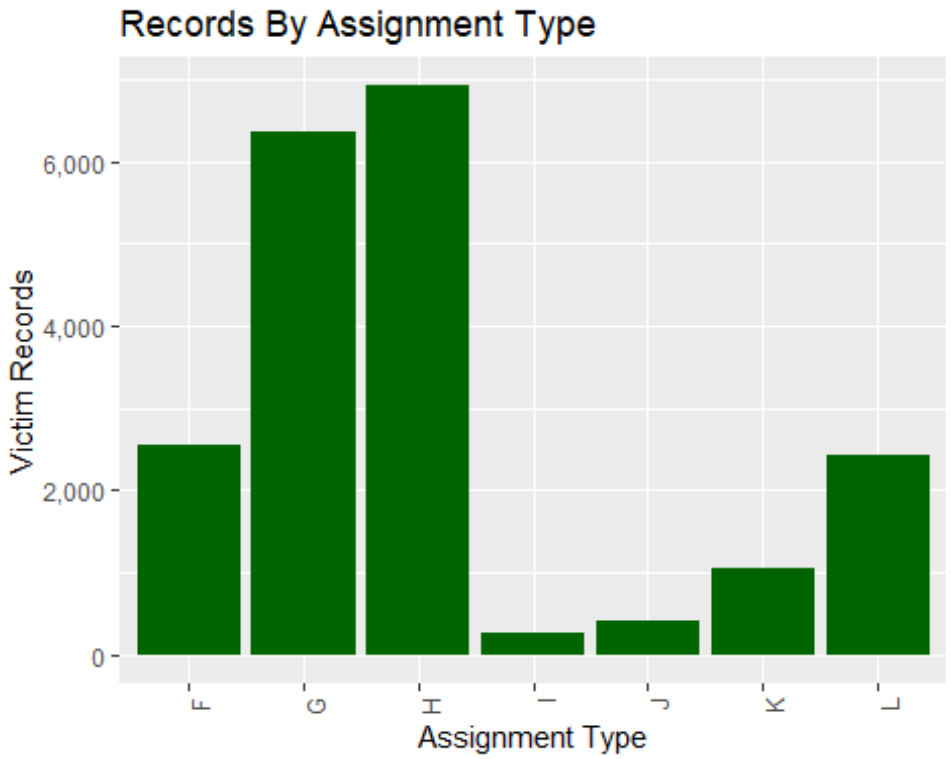
I = Individual
B = Business
S = Society/Public

```
# Activity Type
# Most values are NA.  Removed NA to review remaining fields
p <- ggplot(data=subset(clean_data, !is.na(ACT_TYPE_OFFC)),
aes(x=ACT_TYPE_OFFC)) +
     geom_bar(fill="dark green") +
     labs(x="Activity Type", y="Victim Records", title="Records By Activity
Type") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



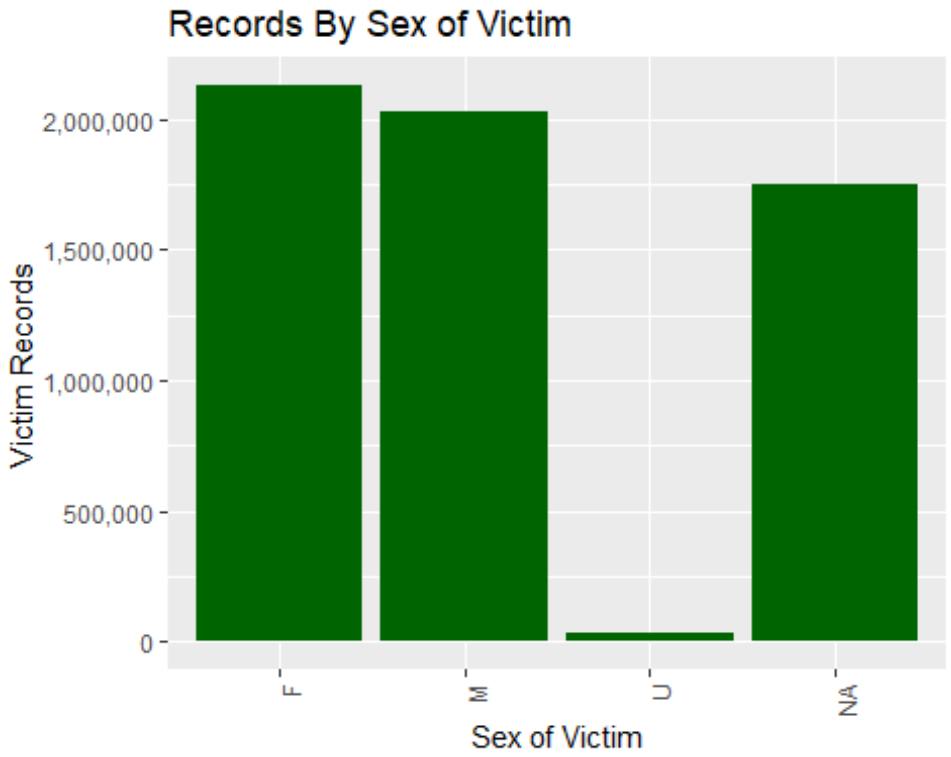
This field is only used when an officer is assaulted or killed in the line of duty
1 = Respond disturbance call (Family quarrels, person with firearm, etc.)
2 = Burglaries in progress or pursuing burglary suspects
3 = Robberies in progress or pursuing robbery suspects
4 = Attempting other arrests
5 = Civil disorder (Riot, mass disobedience)
6 = Handling, transporting, custody of prisoners
7 = Investigating suspicious persons or circumstances
8 = Ambush-no warning
9 = Mentally deranged assailant
10 = Traffic pursuits and stops
11 = All other

```
# Assignment Type

# Most values are NA.  Removed NA to review remaining fields
p <- ggplot(data=subset(clean_data, !is.na(ASSG_TYPE_OFFC)),
aes(x=ASSG_TYPE_OFFC)) +
     geom_bar(fill="dark green") +
     labs(x="Assignment Type", y="Victim Records", title="Records By
Assignment Type") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```

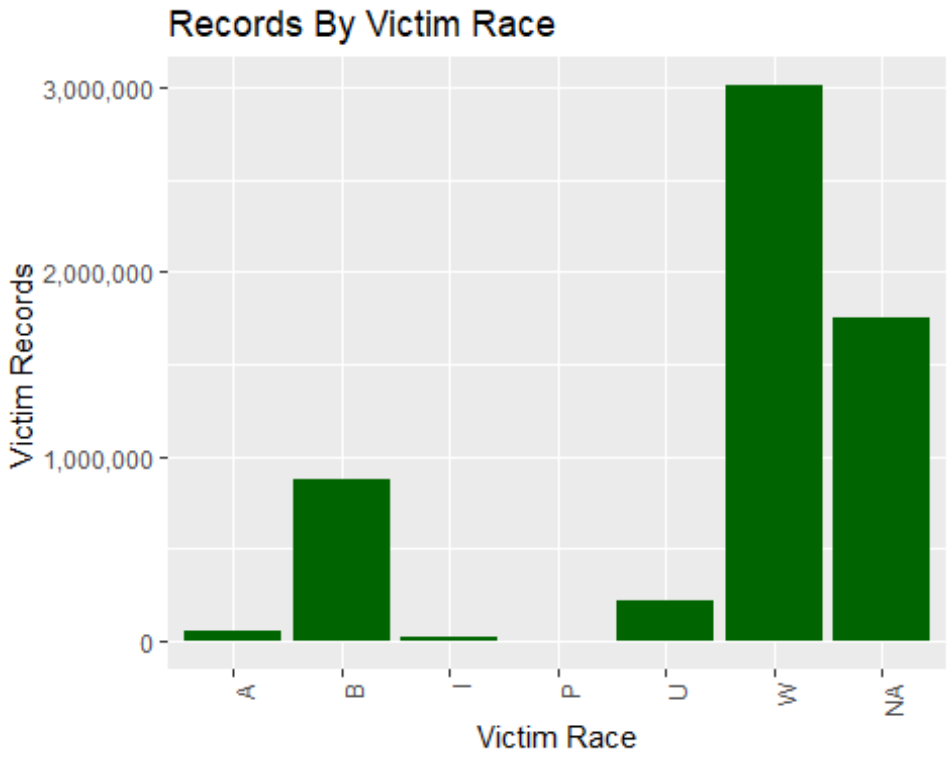


F = Two-officer vehicle
G = One-officer vehicle (alone)
H = One-officer vehicle (assisted)
I = Detective or special assignment (alone)
J = Detective or special assignment (assisted)
K = Other (alone)
L = Other (assisted)

```
# Sex of Victim
p <- ggplot(clean_data, aes(x=SEX_OF_VICTIM)) +
     geom_bar(fill="dark green") +
     labs(x="Sex of Victim", y="Victim Records", title="Records By Sex of
Victim") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```

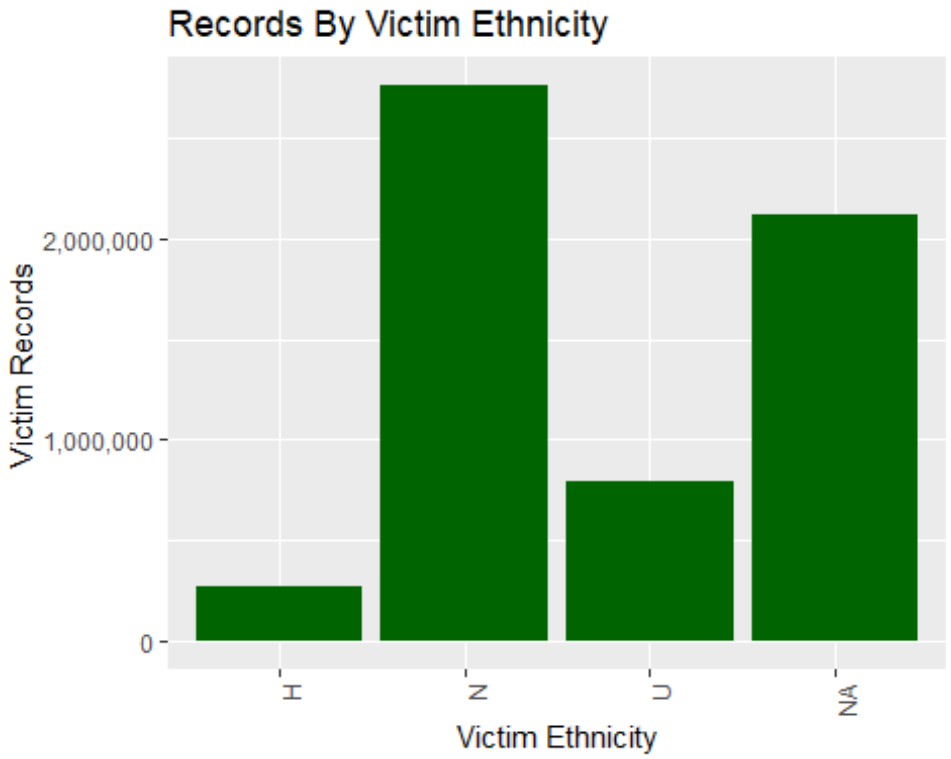



```
# Race of Victim
p <- ggplot(clean_data, aes(x=RACE_OF_VICTIM)) +
     geom_bar(fill="dark green") +
     labs(x="Victim Race", y="Victim Records", title="Records By Victim
Race") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



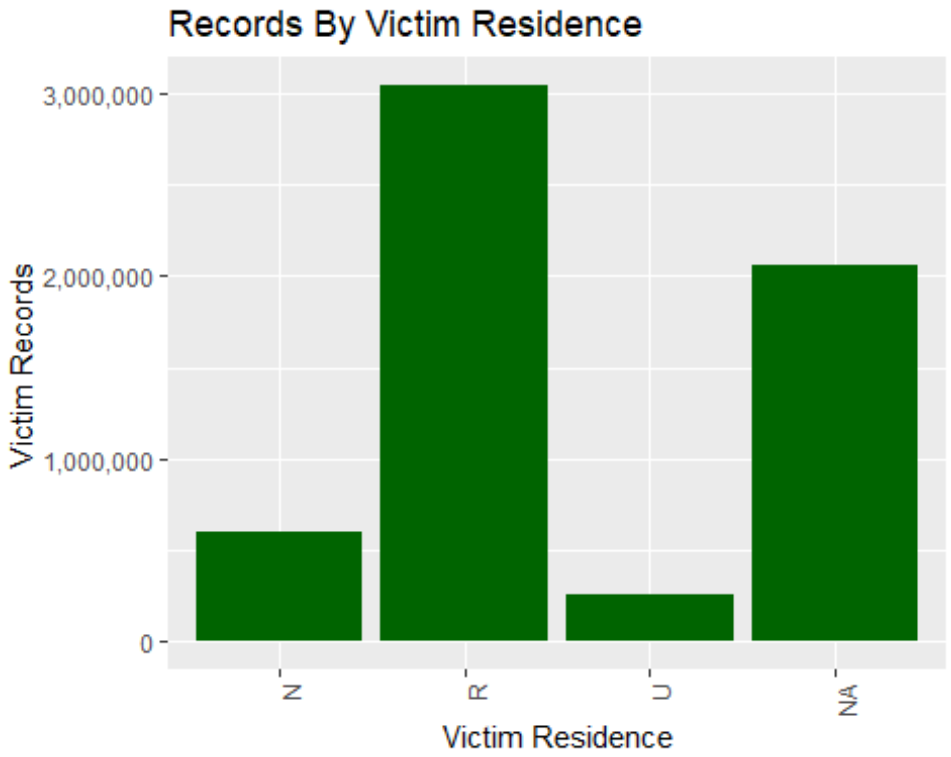
W = White
B = Black or African American
I = American Indian or Alaska Native
A = Asian
P = Native Hawaiian or Other Pacific Islander
U = Unknown

```
p <- ggplot(clean_data, aes(x=ETHNIC_OF_VIC)) +
     geom_bar(fill="dark green") +
     labs(x="Victim Ethnicity", y="Victim Records", title="Records By Victim
Ethnicity") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



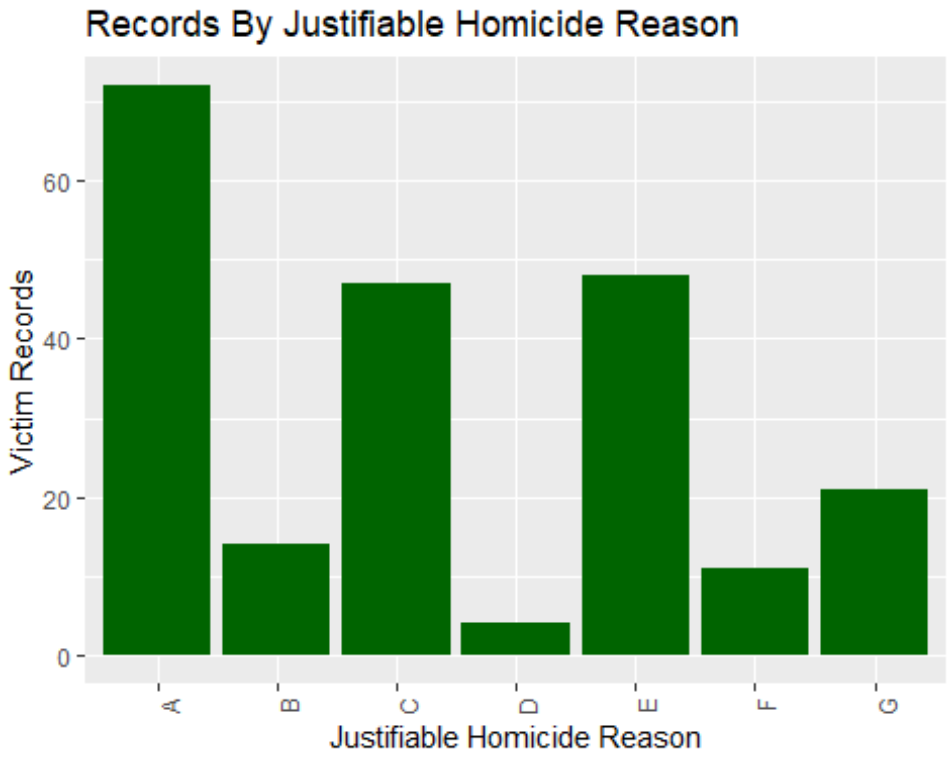
H = Hispanic or Latino Origin
N = Not of Hispanic or Latino Origin
U = Unknown

```
p <- ggplot(clean_data, aes(x=VIC_RESIDENT)) +
     geom_bar(fill="dark green") +
     labs(x="Victim Residence", y="Victim Records", title="Records By Victim
Residence") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



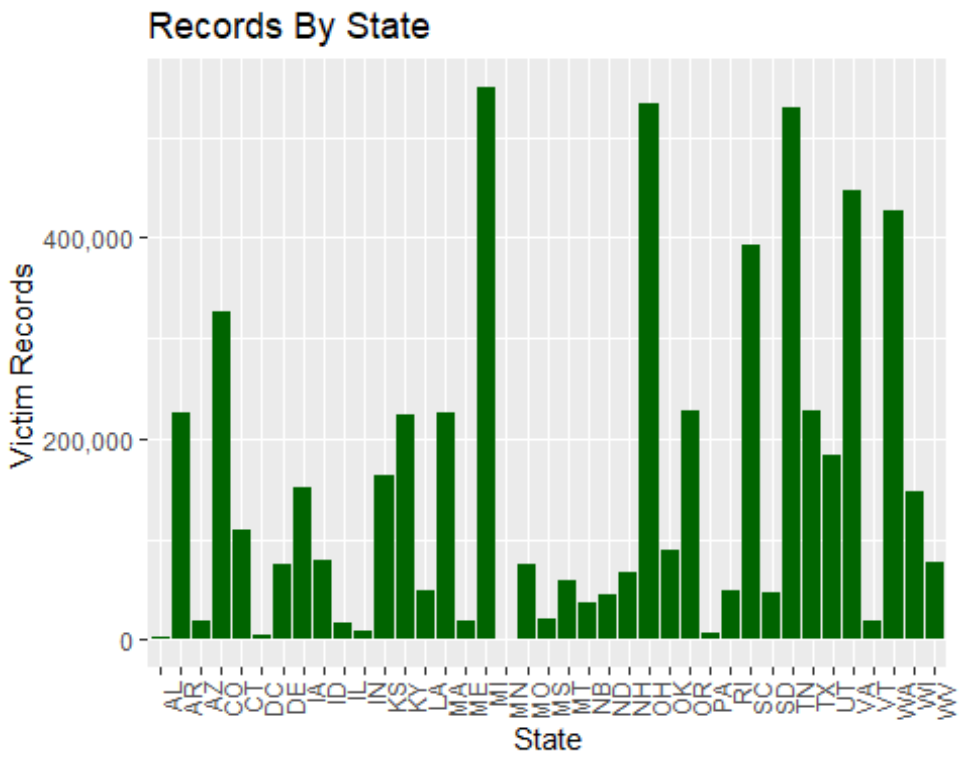
R = Resident. Victim is a resident of the reporting precinct.
N = Nonresident
U = Unknown

```
# Justifiable Homicide

# Most values are NA.  Removed NA to review remaining fields
p <- ggplot(data=subset(clean_data, !is.na(JUST_HOM_CIRC)),
aes(x=JUST_HOM_CIRC)) +
     geom_bar(fill="dark green") +
     labs(x="Justifiable Homicide Reason", y="Victim Records", title="Records
By Justifiable Homicide Reason") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



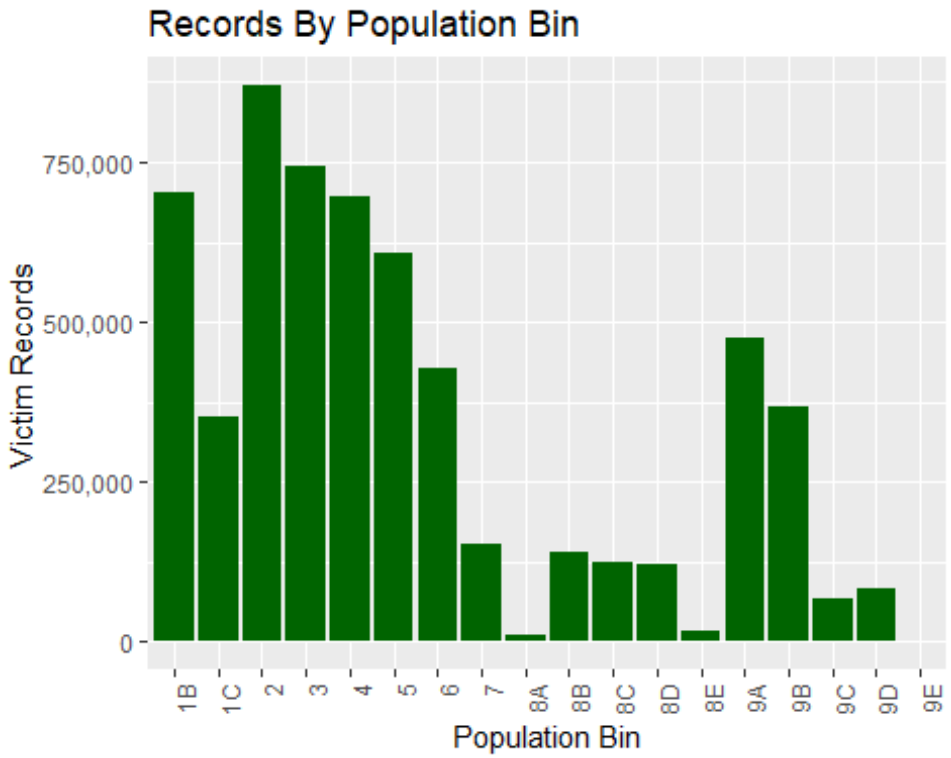
This field is only used if assault circumstance is justifiable homicide)
A=Criminal Attacked Police Officer and That Officer Killed Criminal
B=Criminal Attacked Fellow Police Officer and Criminal Killed by Another Police Officer
C=Criminal Attacked a Civilian
D=Criminal Attempted Flight From a Crime
E=Criminal Killed In Commission of a Crime
F=Criminal Resisted Arrest
G=Unable to Determine/Not Enough Information

```
# State

p <- ggplot(clean_data, aes(x=STATE_ABBR)) +
     geom_bar(fill="dark green") +
     labs(x="State", y="Victim Records", title="Records By State") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



Note that some states may have switched over to the NBIRS system. I may want to focus on states with the highest reported incidents.

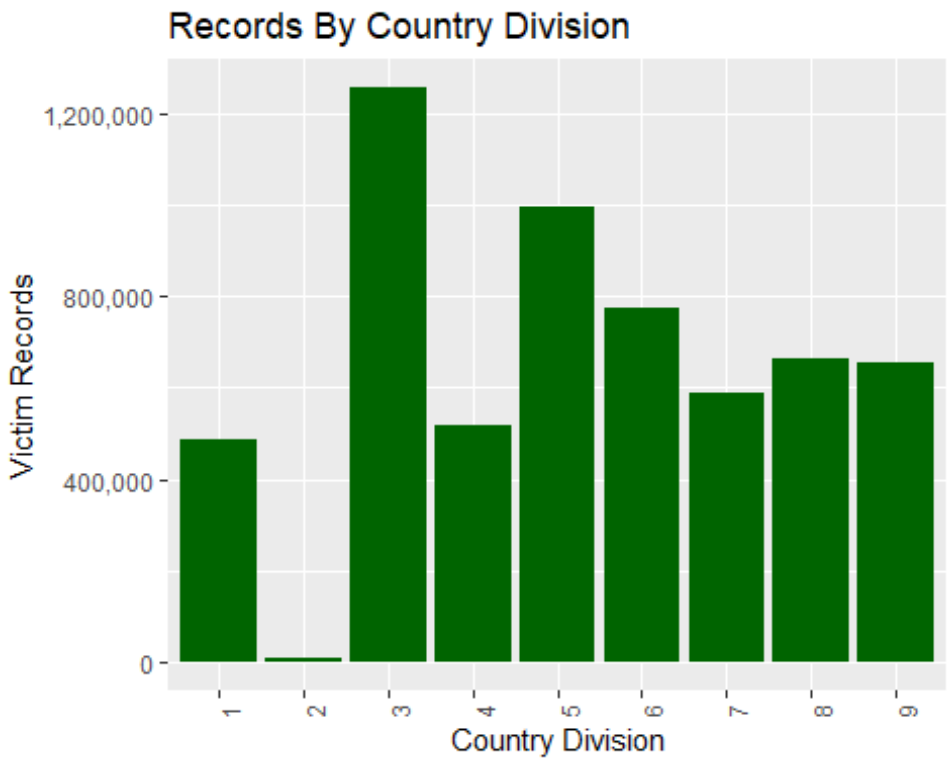
```
# Population Bin

p <- ggplot(clean_data, aes(x=POP_GROUP)) +
     geom_bar(fill="dark green") +
     labs(x="Population Bin", y="Victim Records", title="Records By
Population Bin") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



1B=Cities from 500,000 thru 999,999
1C=Cities from 250,000 thru 499,999
2=Cities from 100,000 thru 249,999
3=Cities from 50,000 thru 99,999
4=Cities from 25,000 thru 49,999
5=Cities from 10,000 thru 24,999
6=Cities from 2,500 thru 9,999
7=Cities under 2,500
8A=Non-MSA Counties 100,000 or over
8B=Non-MSA Counties from 25,000 thru 99,999
8C=Non-MSA Counties from 10,000 thru 24,999
8D=Non-MSA Counties under 10,000
8E=Non-MSA State Police
9A=MSA Counties 100,000 or over
9B=MSA Counties from 25,000 thru 99,999
9C=MSA Counties from 10,000 thru 24,999
9D=MSA Counties under 10,000
9E=MSA State Police

```
# Country Division

p <- ggplot(clean_data, aes(x=CTRY_DIVISION)) +
     geom_bar(fill="dark green") +
     labs(x="Country Division", y="Victim Records", title="Records By Country
```

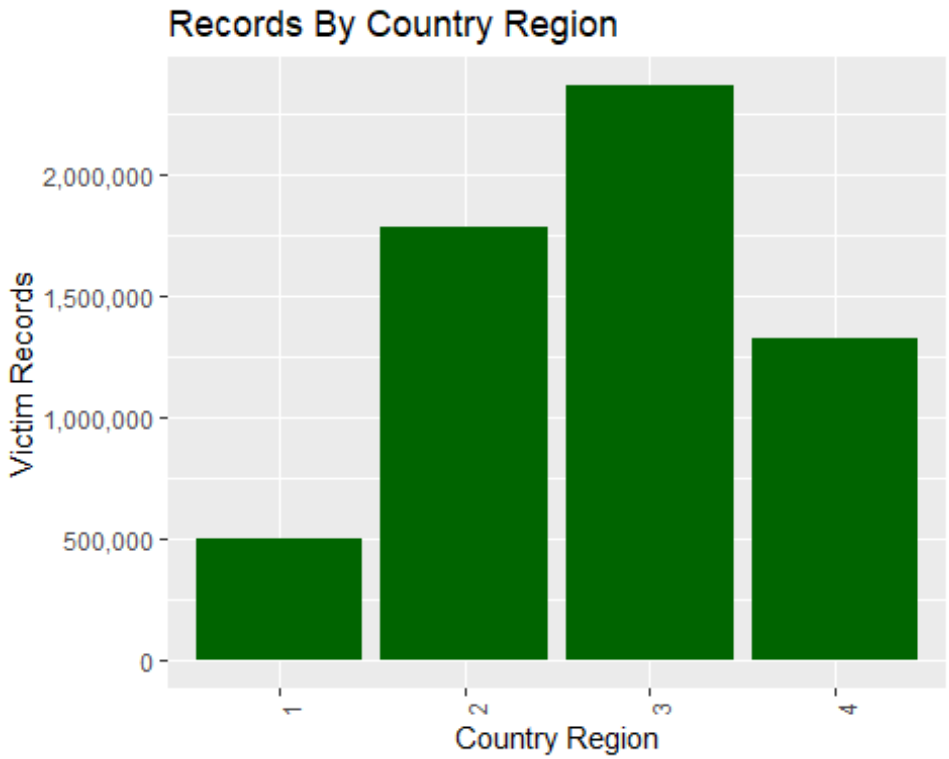
```
Division") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



1=New England
2=Middle Atlantic
3=East North Central
4=West North Central
5=South Atlantic
6=East South Central
7=West South Central
8=Mountain
9=Pacific

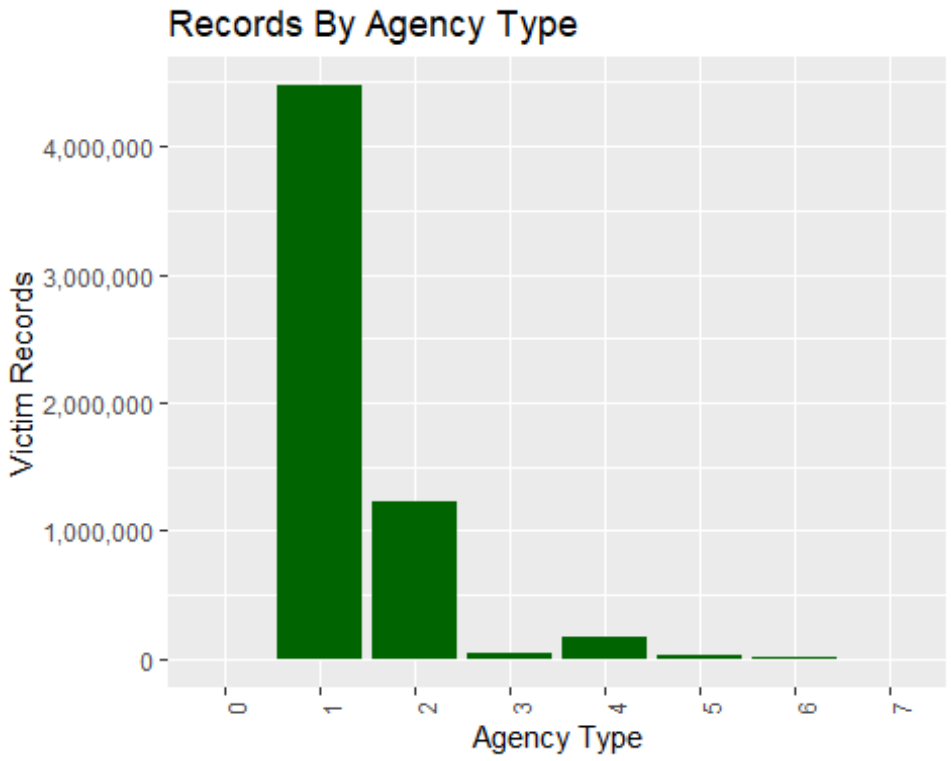
```
# Country Region

p <- ggplot(clean_data, aes(x=CTRY_REGION)) +
     geom_bar(fill="dark green") +
     labs(x="Country Region", y="Victim Records", title="Records By Country
Region") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



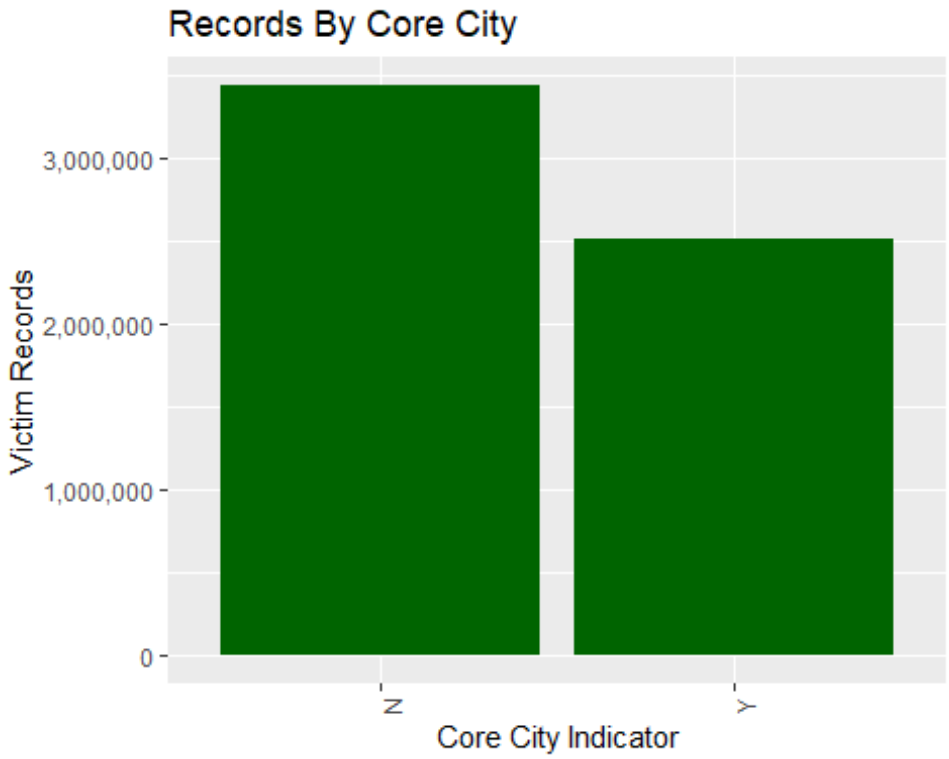
1=North East
2=North Central
3=South
4=West

```
# Agency Type

p <- ggplot(clean_data, aes(x=AGENCY_IND)) +
     geom_bar(fill="dark green") +
     labs(x="Agency Type", y="Victim Records", title="Records By Agency
Type") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```

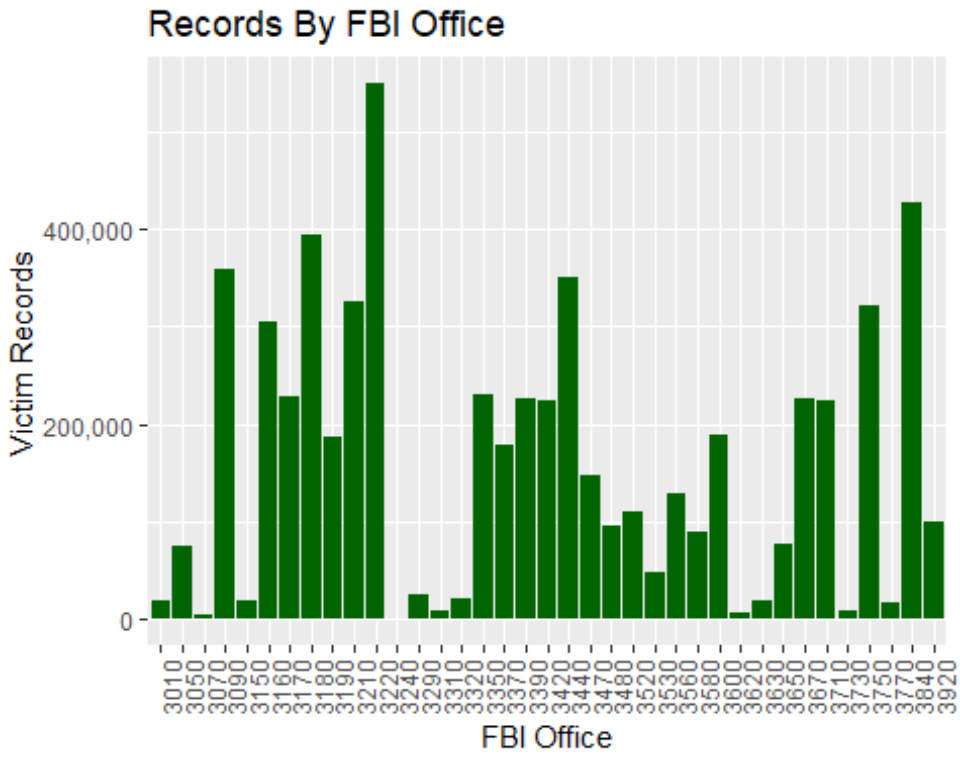


0=Covered-By Another Agency
1=City
2=County
3=University or College
4=State Police
5=Special Agency
6=Other State Agencies
7=Tribal Agencies

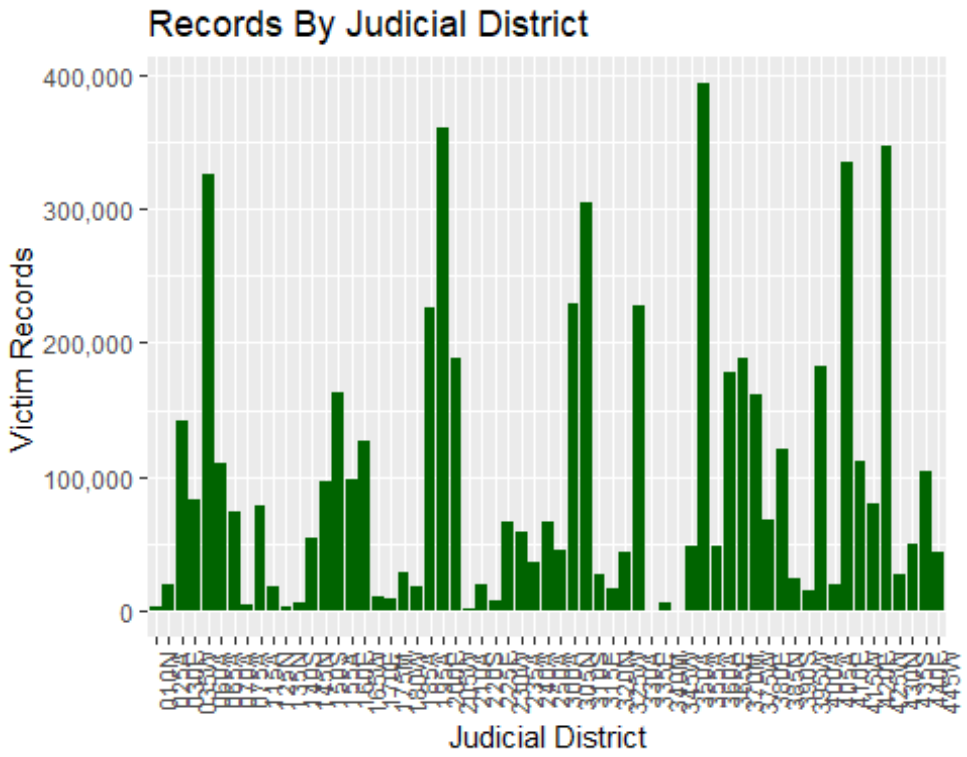
```
# Core City Indicator

p <- ggplot(clean_data, aes(x=CORE_CITY)) +
     geom_bar(fill="dark green") +
     labs(x="Core City Indicator", y="Victim Records", title="Records By Core
City") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



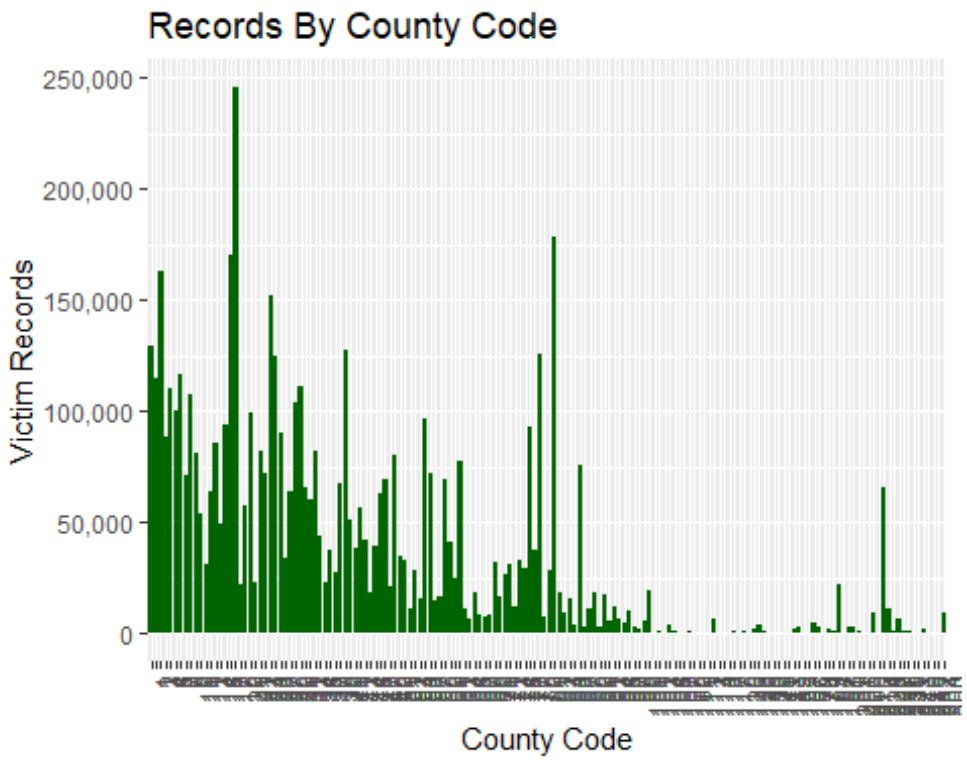

```r
# FBI Office

p <- ggplot(clean_data, aes(x=FBI_OFFICE)) +
     geom_bar(fill="dark green") +
     labs(x="FBI Office", y="Victim Records", title="Records By FBI Office")
+
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



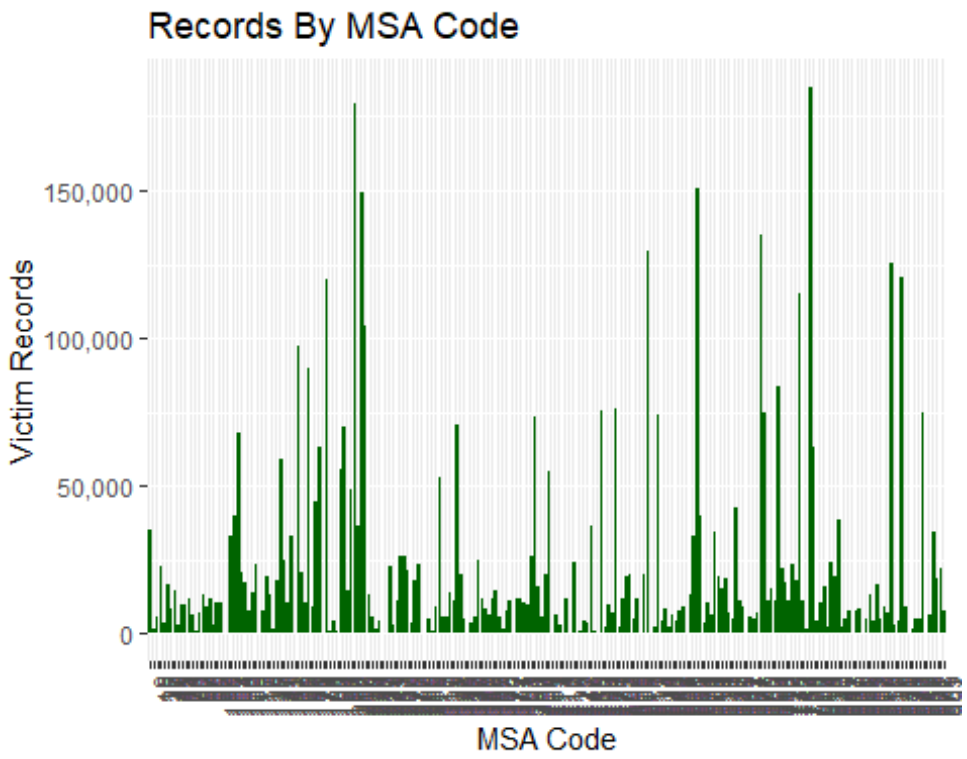

```r
# Judicial District

p <- ggplot(clean_data, aes(x=JUDICIAL_DIST)) +
     geom_bar(fill="dark green") +
     labs(x="Judicial District", y="Victim Records", title="Records By
Judicial District") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```




```
# County Code

# Most values are NA.  Removed NA to review remaining fields
p <- ggplot(data=subset(clean_data, !is.na(UCR_COUNTY_CD1)),
aes(x=UCR_COUNTY_CD1)) +
     geom_bar(fill="dark green") +
     labs(x="County Code", y="Victim Records", title="Records By County
Code") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```

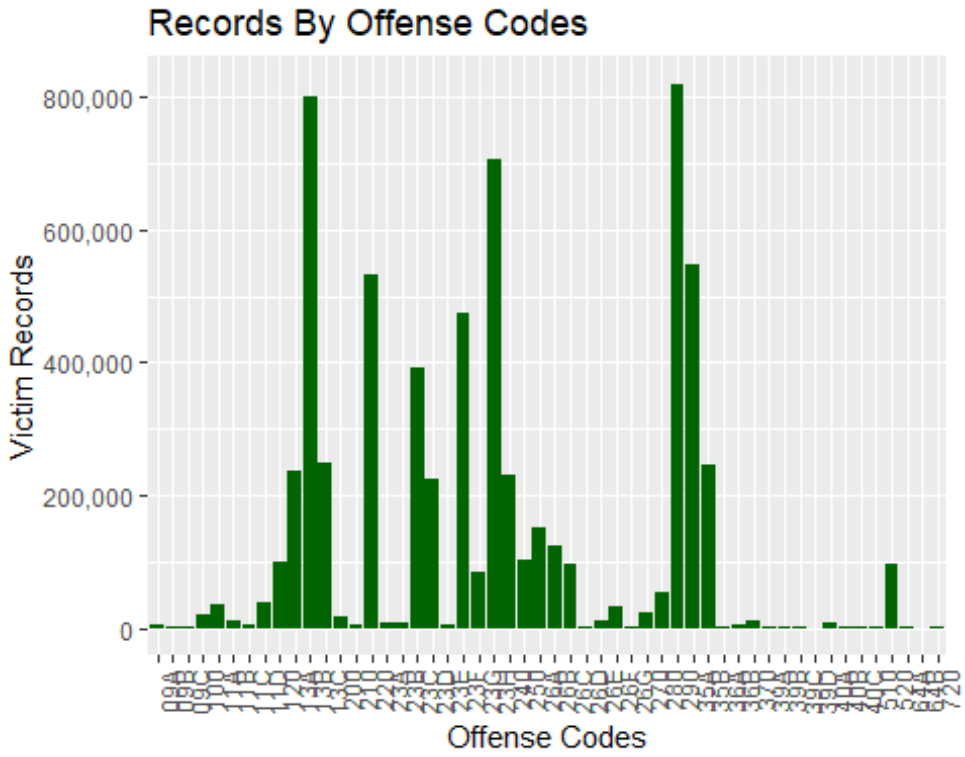



```
# MSA Code

# Most values are NA.  Removed NA to review remaining fields
p <- ggplot(data=subset(clean_data, !is.na(MSA_CD1)), aes(x=MSA_CD1)) +
     geom_bar(fill="dark green") +
     labs(x="MSA Code", y="Victim Records", title="Records By MSA Code") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



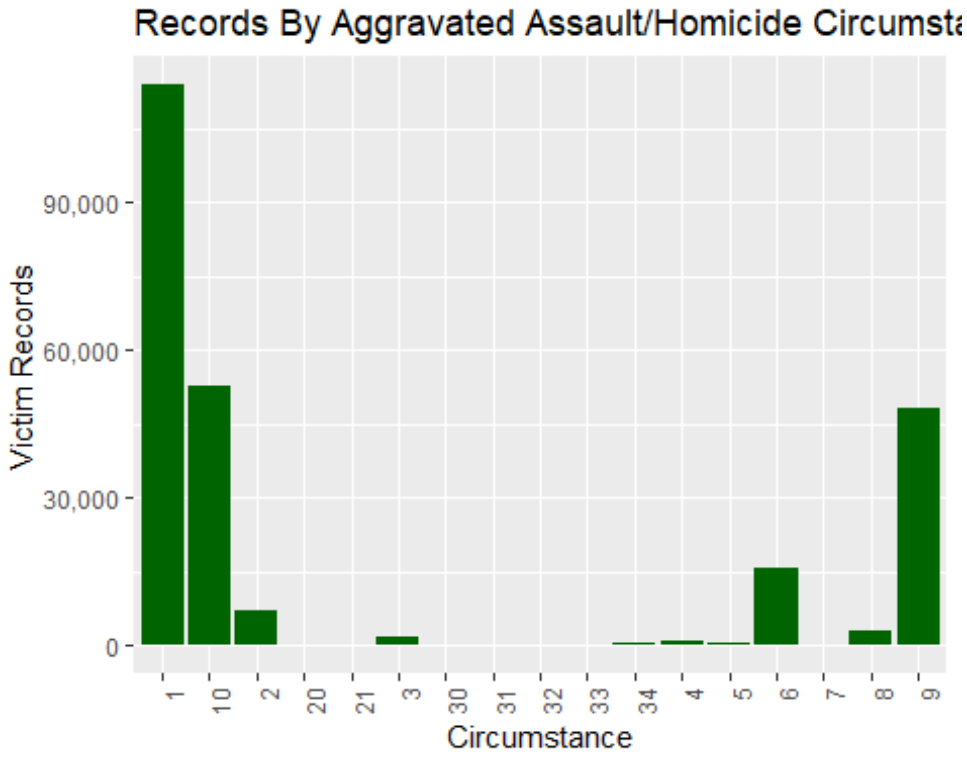

```
# Create new df to include all offenses and counts
off1_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$OFF_CODE01))
off2_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$OFF_CODE02))
off3_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$OFF_CODE03))
off4_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$OFF_CODE04))
off5_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$OFF_CODE05))
off6_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$OFF_CODE06))
off7_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$OFF_CODE07))

names(off1_df)[names(off1_df) == "clean_data.OFF_CODE01"] <- "OFF_CODE"
names(off2_df)[names(off2_df) == "clean_data.OFF_CODE02"] <- "OFF_CODE"
names(off3_df)[names(off3_df) == "clean_data.OFF_CODE03"] <- "OFF_CODE"
names(off4_df)[names(off4_df) == "clean_data.OFF_CODE04"] <- "OFF_CODE"
names(off5_df)[names(off5_df) == "clean_data.OFF_CODE05"] <- "OFF_CODE"
names(off6_df)[names(off6_df) == "clean_data.OFF_CODE06"] <- "OFF_CODE"
names(off7_df)[names(off7_df) == "clean_data.OFF_CODE07"] <- "OFF_CODE"

off_df <- rbind(off1_df, off2_df, off3_df, off4_df, off5_df, off6_df,
off7_df)
```

```r
#head(off_df)
#summary(off_df)

# Offense Codes

p <- ggplot(off_df, aes(x=OFF_CODE)) +
     geom_bar(fill="dark green") +
     labs(x="Offense Codes", y="Victim Records", title="Records By Offense
Codes") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```




```r
# Create new df to include all Aggravated Assault/Homicide Circumstances and
counts
asst1_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$ASSAULT_CIRC1))
asst2_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$ASSAULT_CIRC2))

names(asst1_df)[names(asst1_df) == "clean_data.ASSAULT_CIRC1"] <-
"ASSAULT_CIRC"
names(asst2_df)[names(asst2_df) == "clean_data.ASSAULT_CIRC2"] <-
"ASSAULT_CIRC"
```

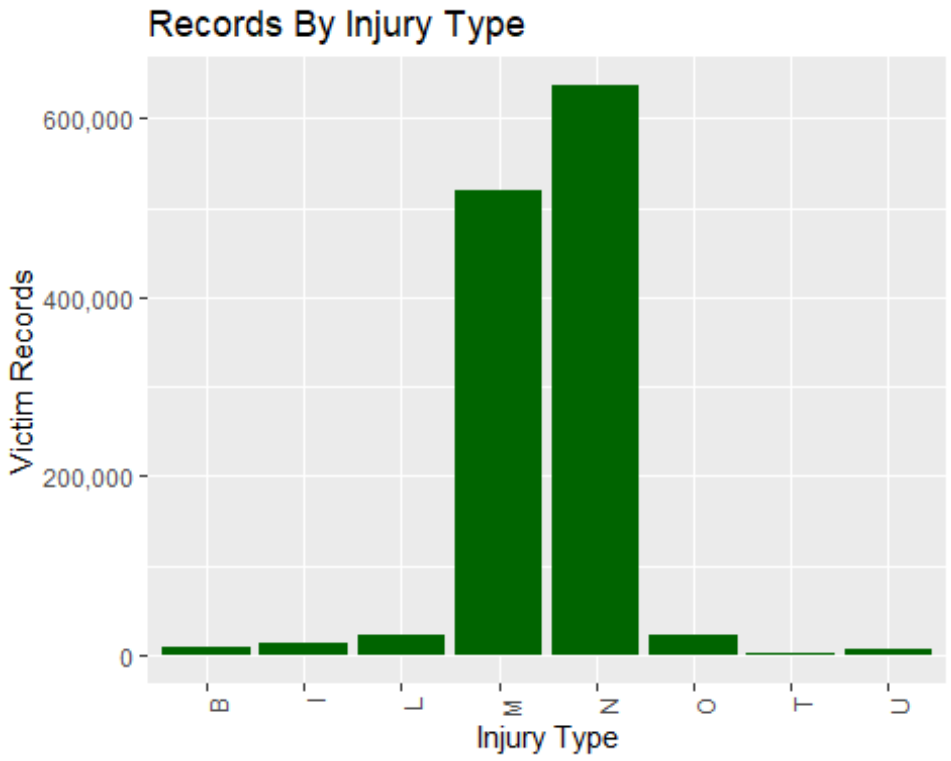
```
asst_df <- rbind(asst1_df, asst2_df)


#head(asst_df)

# Aggravated Assault/Homicide Circumstances

p <- ggplot(asst_df, aes(x=ASSAULT_CIRC)) +
     geom_bar(fill="dark green") +
     labs(x="Circumstance", y="Victim Records", title="Records By Aggravated
Assault/Homicide Circumstances") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



01=Argument
02=Assault on Law Enforcement Officer(s)
03=Drug Dealing
04=Gangland
05=Juvenile Gang
06=Lovers' Quarrel
07=Mercy Killing (Not applicable to Aggravated Assault)
08=Other Felony Involved
09=Other Circumstances

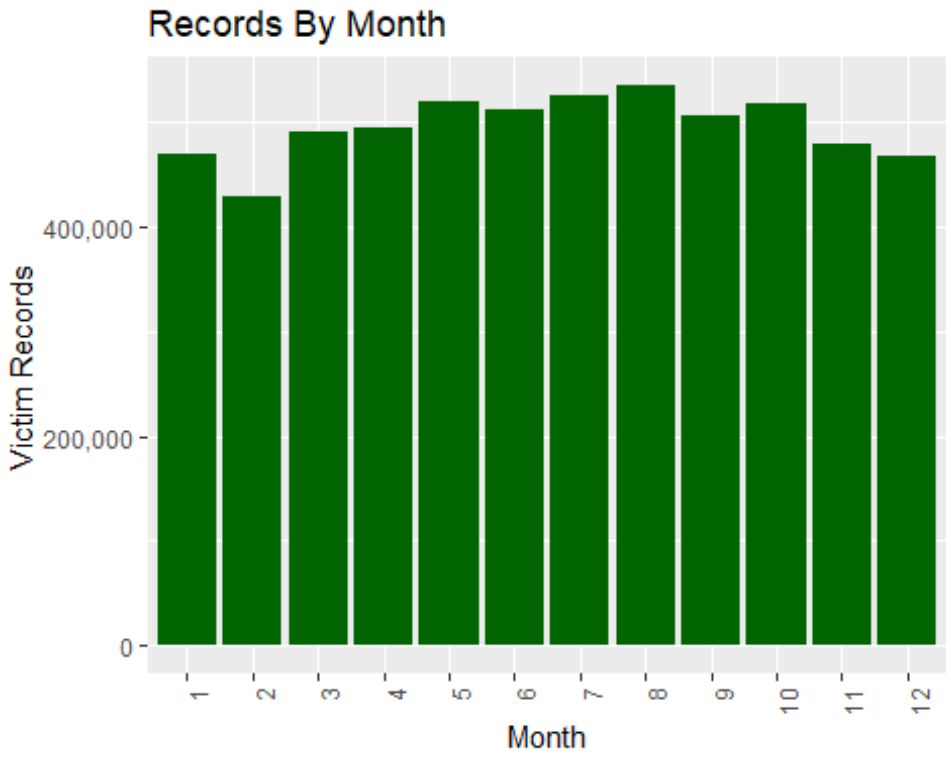
10=Unknown Circumstances
30=Child Playing With Weapon
31=Gun-Cleaning Accident
32=Hunting Accident
33=Other Negligent Weapon Handling
34=Other Negligent Killings
20=Criminal Killed by Private Citizen (JUSTIFIABLE HOMICIDE)
21=Criminal Killed by Police Officer (JUSTIFIABLE HOMICIDE)

```
# Create new df to include all offenses and counts
inj1_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$INJURY_TYPE1))
inj2_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$INJURY_TYPE2))
inj3_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$INJURY_TYPE3))
inj4_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$INJURY_TYPE4))
inj5_df <- na.omit(data.frame(clean_data$ORI, clean_data$INC_NUM,
clean_data$INJURY_TYPE5))

names(inj1_df)[names(inj1_df) == "clean_data.INJURY_TYPE1"] <- "INJURY_TYPE"
names(inj2_df)[names(inj2_df) == "clean_data.INJURY_TYPE2"] <- "INJURY_TYPE"
names(inj3_df)[names(inj3_df) == "clean_data.INJURY_TYPE3"] <- "INJURY_TYPE"
names(inj4_df)[names(inj4_df) == "clean_data.INJURY_TYPE4"] <- "INJURY_TYPE"
names(inj5_df)[names(inj5_df) == "clean_data.INJURY_TYPE5"] <- "INJURY_TYPE"

inj_df <- rbind(inj1_df, inj2_df, inj3_df, inj4_df, inj5_df)


#head(off_df)

# Injury Types

p <- ggplot(inj_df, aes(x=INJURY_TYPE)) +
     geom_bar(fill="dark green") +
     labs(x="Injury Type", y="Victim Records", title="Records By Injury
Type") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



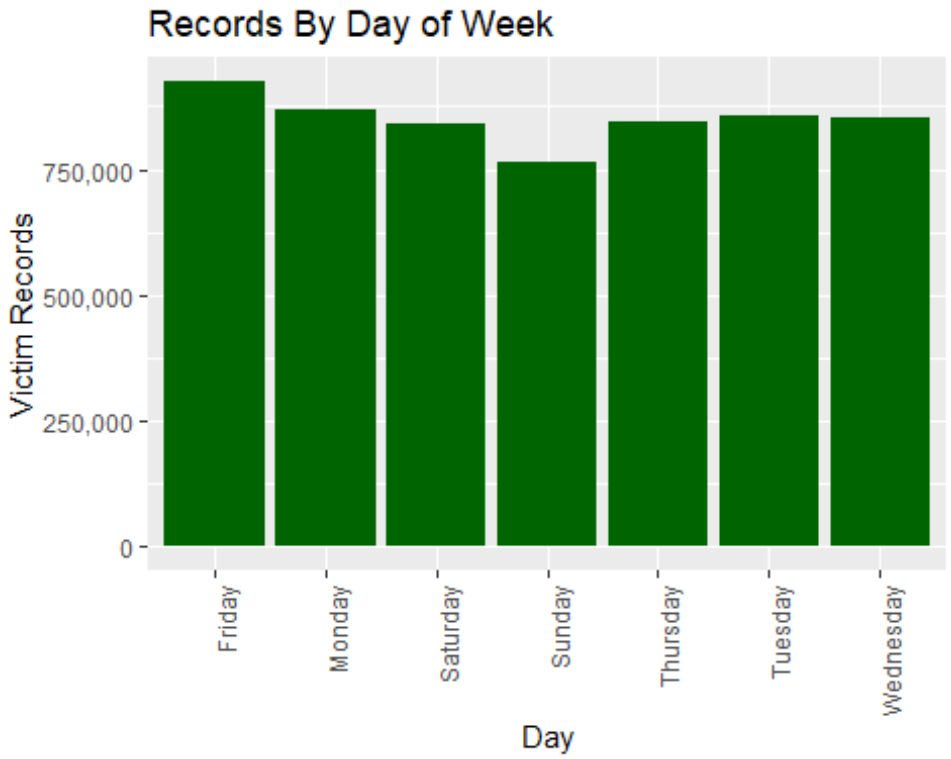
N=None
M=Apparent Minor Injury
B=Apparent Broken Bones
O=Other Major Injury
I=Possible Internal Injury
T=Loss of Teeth
L=Severe Laceration
U=Unconsciousness

```
# Month

p <- ggplot(clean_data, aes(x=VIC_INC_MONTH)) +
     geom_bar(fill="dark green") +
     labs(x="Month", y="Victim Records", title="Records By Month") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```



This is surprising. I would think crime might be lower in winter months and higher in summer. Although that is slightly true, the difference does not appear significant.

```
# Day of the Week

p <- ggplot(clean_data, aes(x=VIC_INC_DOW)) +
     geom_bar(fill="dark green") +
     labs(x="Day", y="Victim Records", title="Records By Day of Week") +
     theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Remove scientific notation
p + scale_y_continuous(labels = comma)
```

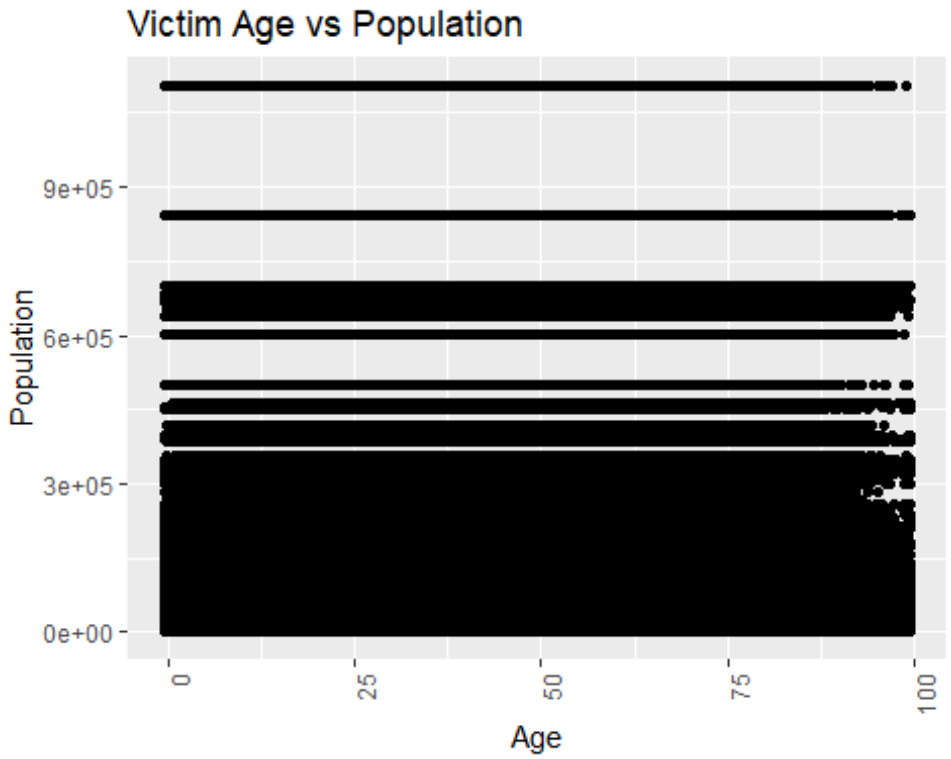


The dip on Sunday is expected, but I was surprised by the lower count on Saturday.

3. EDA - Look for correlation

a) Bivariate Plots

```
ggplot(clean_data, aes(x=AGE_OF_VICTIM, y=CURRENT_POP1)) +
    geom_point(position="jitter") +
    labs(x="Age", y="Population", title="Victim Age vs Population") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Possible candidate for min-max scaling.

b) Test for correlation

```
cor.test(clean_data$AGE_OF_VICTIM, clean_data$CURRENT_POP1, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  clean_data$AGE_OF_VICTIM and clean_data$CURRENT_POP1
## t = -74.715, df = 5951118, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03141576 -0.02981040
## sample estimates:
##        cor
## -0.0306131
```

This numeric value indicates there is no correlation between victim's age and population of reporting district.

4. Get Location Data & Export Final Dataset

a) Import the Data

```
# Load geodetic city data
geo_data <- read_csv("Data/Location_Data/uscities.csv")

#head(geo_data)
```

b) Re-Label fields

```
# Rename fields so they can be joined
names(clean_data)[names(clean_data) == "CITY_NAME"] <- "CITY"
names(clean_data)[names(clean_data) == "STATE_ABBR"] <- "STATE"
names(geo_data)[names(geo_data) == "city"] <- "CITY"
names(geo_data)[names(geo_data) == "state_id"] <- "STATE"

# Convert city names to uppercase for matching
geo_data$CITY <- toupper(geo_data$CITY)
```

c) Apply filters

```
# Limit only to high-reported states
# Filter both datasets to save memory
fil_clean_df <- filter(clean_data, STATE == "TN" | STATE == "MI" | STATE ==
"SC" |
                     STATE == "MA" | STATE == "OH" | STATE == "WA")

fil_geo_df <- filter(geo_data, STATE == "TN" | STATE == "MI" | STATE == "SC"
|
                     STATE == "MA" | STATE == "OH" | STATE == "WA")

#head(fil_clean_df)
#head(fil_geo_df)
```

d) Join Datasets

```
# Remove dataframes no longer needed to conserve memory
#rm("bat_data", "vic_data", "vic_new_data", "geo_data", "comb_df",
"clean_data", "vic_date_data", "just_hom_df",
#   "off1_df", "off2_df", "off3_df", "off4_df", "off5_df", "off6_df",
"off7_df",
#   "inj_df","inj1_df", "inj2_df", "inj3_df", "inj4_df", "inj5_df",
#   "asst_df","asst1_df", "asst2_df")

# Join cleaned and filtered data with city geodetic reference data
joint_df <- left_join(fil_clean_df, unique(fil_geo_df), c("CITY", "STATE"))

## Warning: Column `STATE` joining factor and character vector, coercing into
## character vector
```

```
#head(joint_df)
#summary(joint_df)
```

e) Clean and Output Data Frame

```
#rm("fil_clean_df", "fil_geo_df")

# Output clean and filtered file to save memory
# Can re-load for future use
write.csv(joint_df,"Data/crime_top6_states.csv")
```