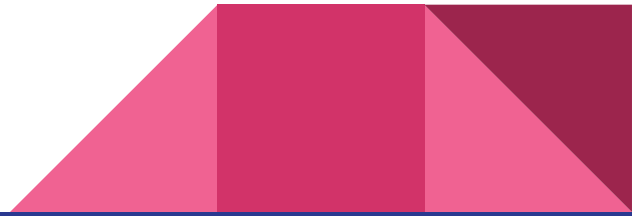
Crime/Murder in California Cities

Tigran Manukyan, Aigerim Toleukhanova

Problem

Predicting the number of murders in California cities, based on the different crimes committed in those cities.

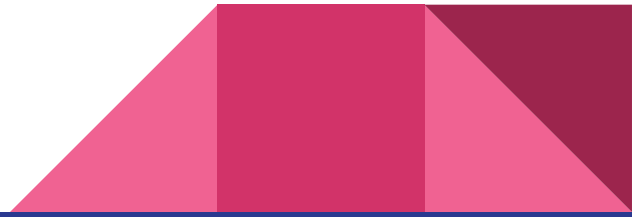


Task Distribution So Far

Web Scraper - Tigran

Data Merging/Cleaning - Aigerim

Data Visualization - Both



Data

From 2005-2019

Each year has around 460 tuples

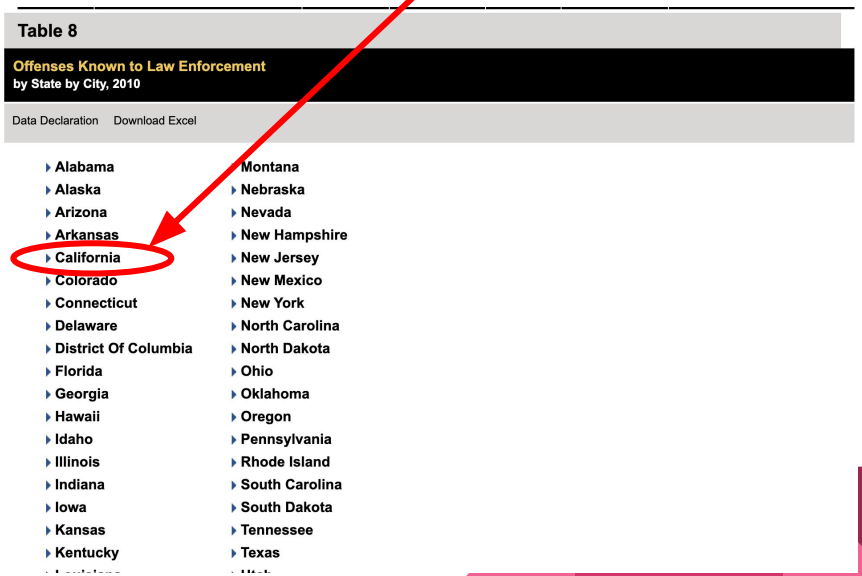
Around 6900 after merging

Scrapper

```
offensesKnownLink <- paste0("https://ucr.fbi.gov/crime-in-the-u.s/", year,
                            "/crime-in-the-u.s.-", year, "/offenses-known-to-law-enforcement")
```


```
read_html(offensesKnownLink) %>%
  html_nodes("#col_b") %>%
  html_nodes("a") %>%
  html_attr("href")->stateLinkz
```

```
for(tableLink in tableLinkz){
  if(((grepl("ca", tableLink, fixed = TRUE) | grepl("california", tableLink, fixed = TRUE) | grepl("California", tableLink, fixed = TRUE)) &
    break
  }
```



Scrapper

```
read_html(tableLink) %>%
  html_nodes("#table-data-container")%>%
  html_table() %>%
  as.data.frame() %>%
  write.csv(paste0("FBITables/",year,"table.csv"))

Sys.sleep(5)
```

ucr.fbi.gov/robots.txt

```
Sitemap: https://ucr.fbi.gov/sitemap.xml

User-agent: Googlebot
Disallow: /*.js$
Disallow: /*.css$

User-agent: *
Crawl-delay: 5
Disallow: /cgi-bin/
Disallow: /includes/
Disallow: /Includes/
Disallow: /images/
Disallow: /scripts/
Disallow: /test/
Disallow: /login
Disallow: /help/
Disallow: /javascript/
Disallow: /css/
Disallow: /redirects/
Disallow: /cgi/
Disallow: /error/
Disallow: /esi/
Disallow: /graphics/
Disallow: /js/
Disallow: /swf/
Disallow: /news/
Disallow: /xml/
```

U.S. Department of Justice
Federal Bureau of Investigation

CRIME IN THE UNITED STATES 2010

Criminal Justice Information Services Division | Feedback | Contact Us | Data Quality Guidelines | UCR Home

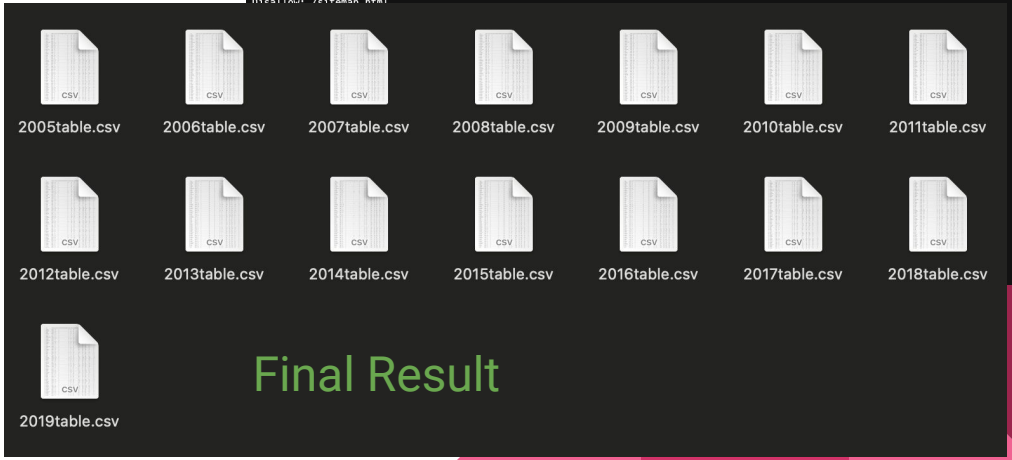
CIUS Home | Offenses Known to Law Enforcement | Violent Crime | Property Crime | Clearances | Persons Arrested | Police Employee Data

Table 8

CALIFORNIA
Offenses Known to Law Enforcement
by State by City, 2010

Data Declaration Download Excel Table 8 State Listing

City	Population	Violent crime	Murder and nonnegligent manslaughter	Forcible rape	Robbery	Aggravated assault	Property crime	Burglary	Larceny-theft	Motor vehicle theft	Arson
Adelanto	29,697	242	4	9	39	190	726	359	273	94	13
Agoura Hills	22,318	18	0	2	1	15	240	59	173	8	0
Alameda	71,016	173	1	15	74	83	1,964	389	1,360	215	10
Albany	16,096	37	0	5	24	8	582	95	392	95	4
Alhambra	84,987	185	3	8	82	92	2,075	345	1,460	270	18
Aliso Viejo	41,989	37	0	4	8	25	439	87	324	28	3
Alturas	2,733	5	0	2	0	3	63	28	30	5	1
American Canyon	17,507	38	0	2	20	16	515	100	375	40	1
Anaheim	338,492	1,161	7	88	492	574	8,473	1,594	5,869	1,010	47

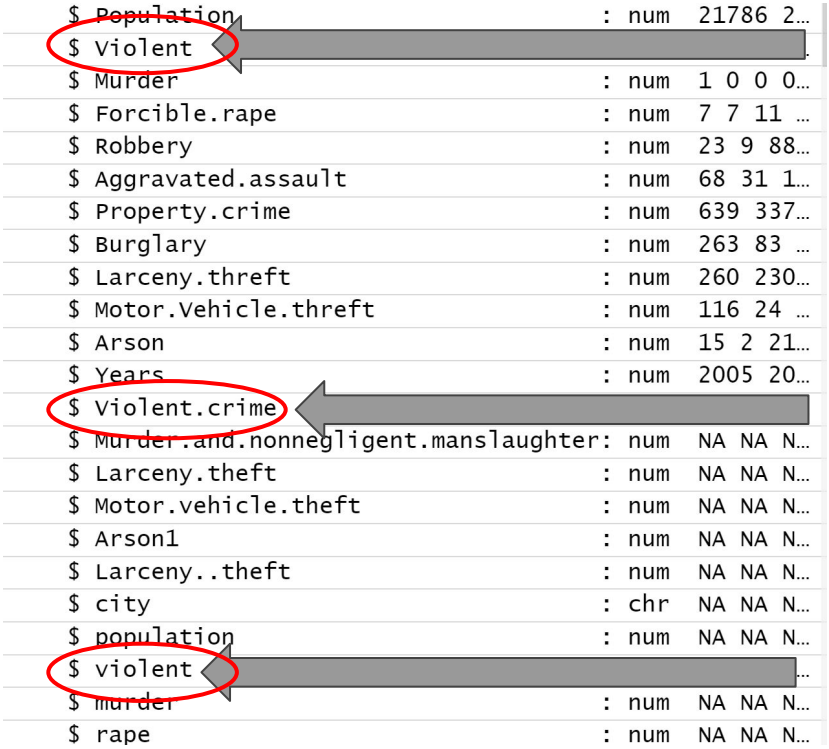


Merging

```
folder_path <-"C:\\Users\\aiger\\OneDrive\\Desktop\\ComputerScience\\CS_DM_541\\ProjectDM\\CSV"
#folder_path <-"C:\\Users\\aiger\\OneDrive\\Desktop\\ComputerScience\\CS_DM_541\\ProjectDM\\CSV"
file.exists(folder_path)
file_paths <- list.files(folder_path, pattern = "\\.csv$", full.names = TRUE)

#create empty list and call allSamples
allSamples <- list()

#combine all cvs files into list
for(file_path in file_paths) {
  table_name <- tools::file_path_sans_ext(basename(file_path))
  allSamples[[table_name]] <- read.csv(file_path)
}
```

```
combinedAllSamples <- bind_rows(allSamples, .id = "Dataset")
```

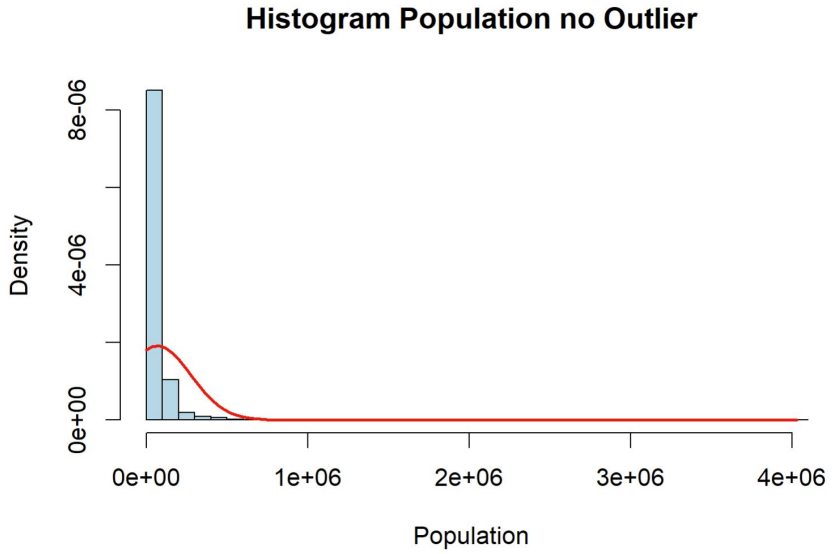
Cleaning

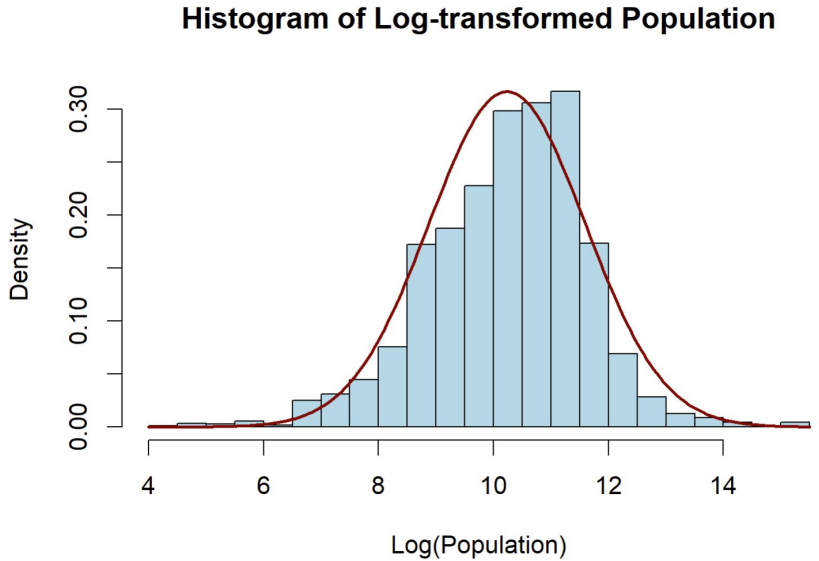


- Fix redundancy

```
combinedAllSamples <-combinedAllSamples%>%
  mutate(Violent = case_when(
    !is.na(violent)~violent,
    !is.na(Violent.crime)~Violent.crime,
    !is.na(Violent)~Violent,
    TRUE~NA_real_
  ))
combinedAllSamples$violent<-NULL
combinedAllSamples$Violent1<-NULL
combinedAllSamples$Violent.crime<-NULL
```

Population Transformation (Handling NA)





Cleaning (Handling NA's)

```
   Population         Robbery         Burglary          Years          Arson
Min.   :     89   Min.   :    0   Min.   :    0.0   Min.   :2005   Min.   :   0.00
1st Qu.:  11982   1st Qu.:    5   1st Qu.:   63.0   1st Qu.:2008   1st Qu.:   1.00
Median :  31234   Median :   22   Median :  162.0   Median :2012   Median :   4.00
Mean   :  68054   Mean   :  114   Mean   :  380.3   Mean   :2012   Mean   :  15.67
3rd Qu.:  70389   3rd Qu.:   75   3rd Qu.:  372.0   3rd Qu.:2016   3rd Qu.:  12.00
Max.   :4029741   Max.   :14353   Max.   :22592.0   Max.   :2019   Max.   :2356.00
                                  NA's   :3                        NA's   :38

   NameCity            Violent           Murder             Rape          Assault
Length:6896        Min.   :    0.0   Min.   :  0.000   Min.   :   0.00   Min.   :    0.0
Class :character   1st Qu.:   29.0   1st Qu.:  0.000   1st Qu.:   2.00   1st Qu.:   18.0
Mode  :character   Median :   86.0   Median :  1.000   Median :   6.00   Median :   53.0
                   Mean   :  310.9   Mean   :  3.396   Mean   :  18.88   Mean   :  174.5
                   3rd Qu.:  227.0   3rd Qu.:  2.000   3rd Qu.:  16.00   3rd Qu.:  135.0
                   Max.   :31767.0   Max.   :489.000   Max.   :2528.00   Max.   :17216.0
                   NA's   :5                                             NA's   :5
PropertyCrime          Theft
Min.   :     0.0   Min.   :    0
1st Qu.:   273.0   1st Qu.:  168
Median :   721.5   Median :  463
Mean   :  1897.3   Mean   : 1189
3rd Qu.:  1876.0   3rd Qu.: 1193
Max.   :117285.0   Max.   :67963
NA's   :6
>
```

Area Data

	X	X1	X2	
1	1	Rank	Land Area ▼	City / Population
2	2	1.	468.67 sq mi	Los Angeles, CA / 3,862,210
3	3	2.	325.19 sq mi	San Diego, CA / 1,341,510
4	4	3.	203.52 sq mi	California City, CA / 13,243
5	5	4.	176.53 sq mi	San Jose, CA / 986,320
6	6	5.	142.16 sq mi	Bakersfield, CA / 358,700
7	7	6.	111.96 sq mi	Fresno, CA / 506,132
8	8	7.	105.96 sq mi	Palmdale, CA / 155,810
9	9	8.	105.59 sq mi	Lucerne Valley, CA / 5,767

Cleaning

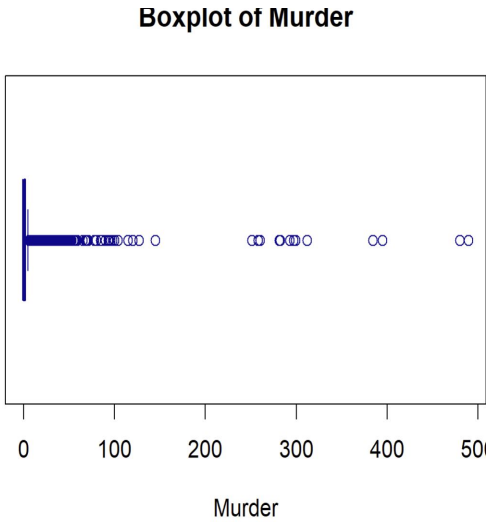
```
area$X1<-NULL
area$City<-area$X3
area$SQmiles<- area$X2

area$X2<-NULL
area$X3<-NULL
#delete row 1
area<-area[-c(1), ]
#get ridNULL#get rid of outliers


#delete everything after comma ,
area$City<-gsub(",.*", "", area$City)
#delete sq.miles for each entry
area$SQmiles<-gsub("s.*", "", area$SQmiles

#convert into numeric values
area$SQmiles<-as.numeric(area$SQmiles)
```

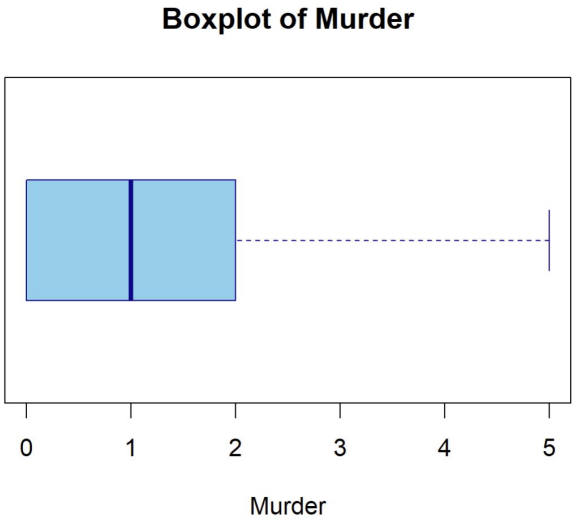
Result

	City	SQmiles
2	Los Angeles	468.67
3	San Diego	325.19
4	California City	203.52
5	San Jose	176.53
6	Bakersfield	142.16
7	Fresno	111.96
8	Palmdale	105.96
9	Lucerne Valley	105.59

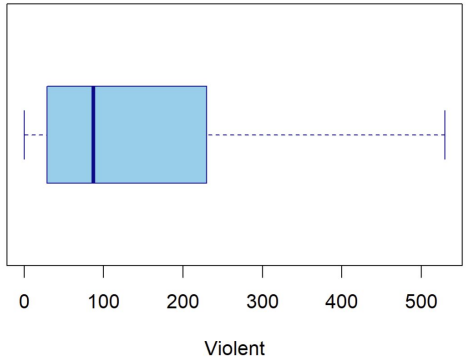
Visualization -BoxPLots

```
boxplot(mergedMurder$Murder,
        main = "Boxplot of Mu
        xlab = "Murder",
        ylab = "Murder ",
        outline = TRUE,   # wi
        col = "skyblue",   # C
        border = "darkblue",
        horizontal = TRUE)


####    boxplot of  Murder wit
boxplot(mergedMurder$Murder,
        main = "Boxplot of Mu
        xlab = "Murder",
        ylab = "Murder ",
        outline = FALSE,   # w
        col = "skyblue",   # C
        border = "darkblue",
        horizontal = TRUE)
```

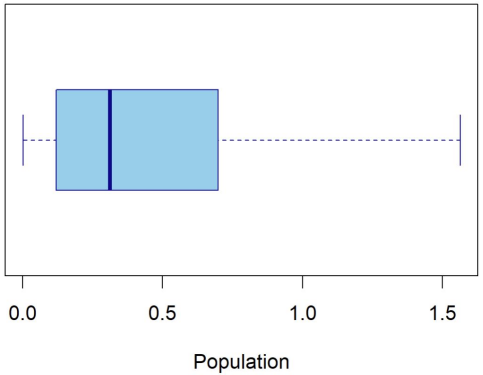




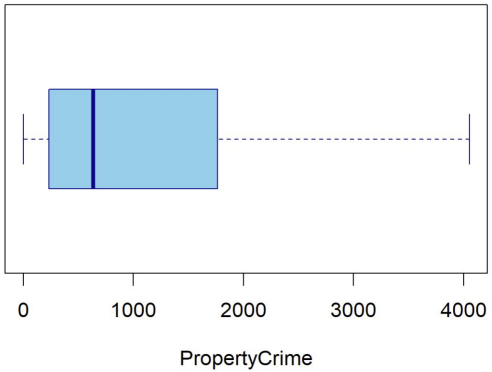
Boxplot of Violent with No outlier



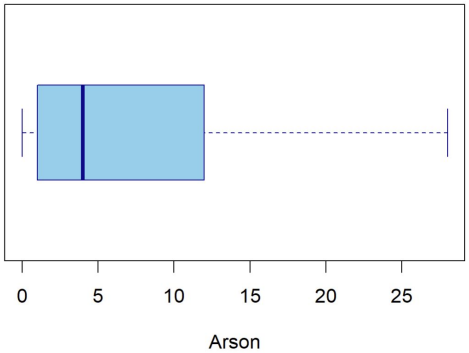
Boxplot of Population with No outlier(in 100K)



Boxplot of Property Crime with No outlier



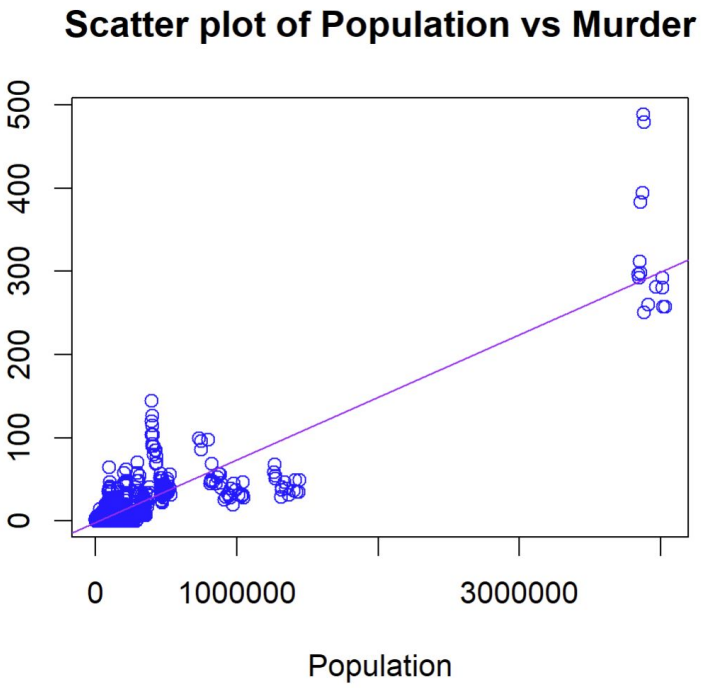
Boxplot of Arson with No outlier

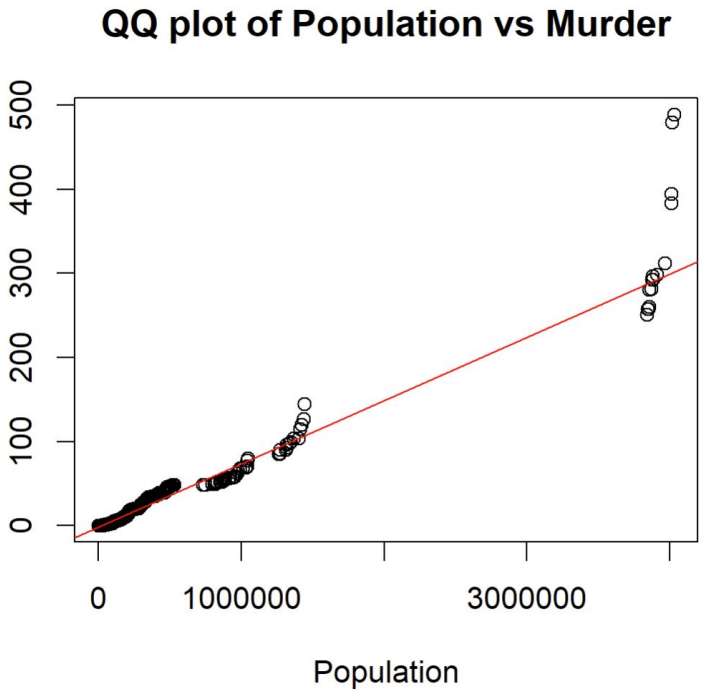


Deleting outlier from data

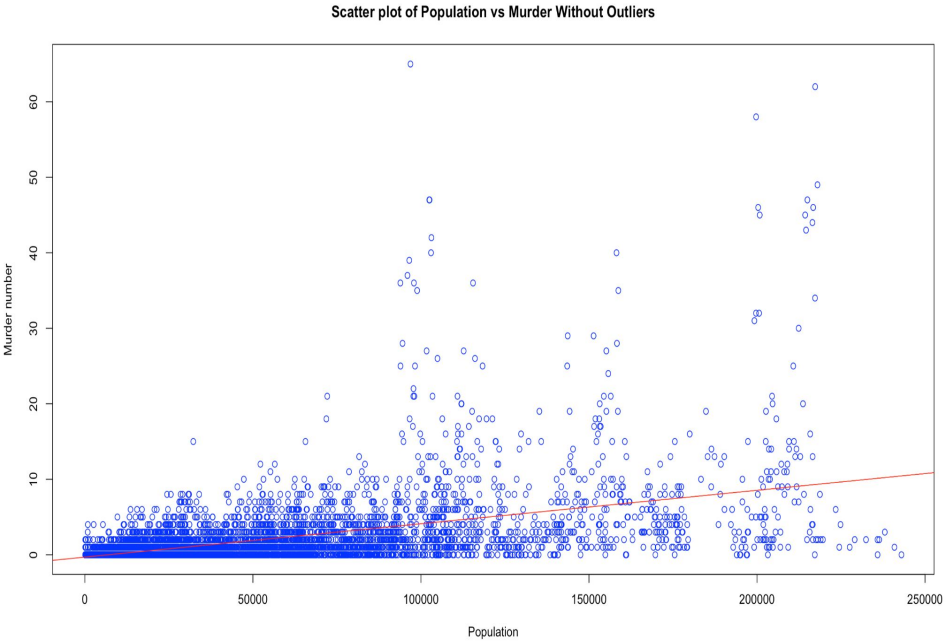
```
# Calculate the first quartile, third quartile
Q1 <- quantile(mergedMurder$Population, 0.25)
Q3 <- quantile(mergedMurder$Population, 0.75)

# interquartile range (IQR)
IQR <- Q3 - Q1

#the lower and upper bounds
lower <- Q1 -3* IQR
upper <- Q3 + 3 * IQR
#3 is multiplier for changing the bounds

cleanedMurder<-mergedMurder[mergedMurder$Population>=lower &
                            mergedMurder$Population<=upper,]
```

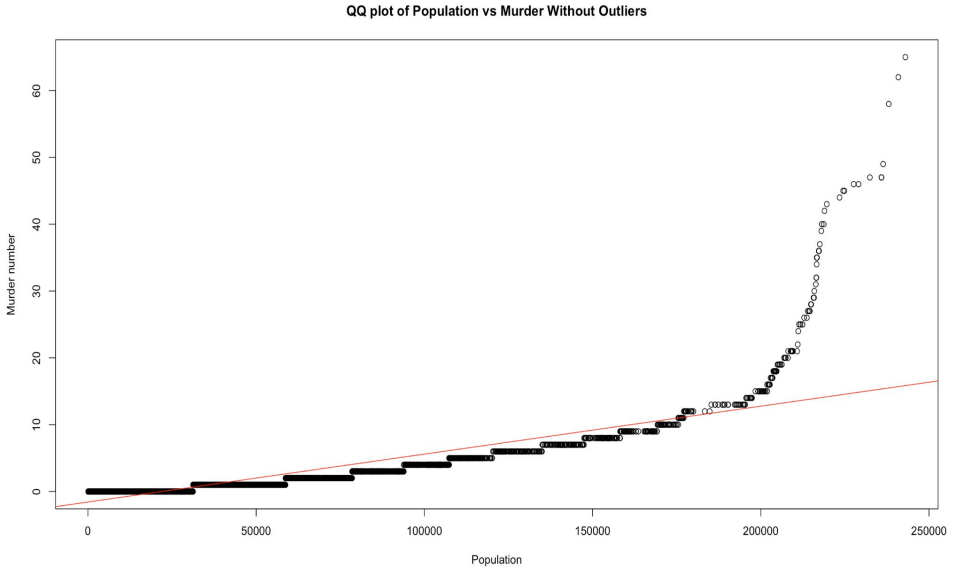
Visualization -Scatter Plot and QQ-Plot





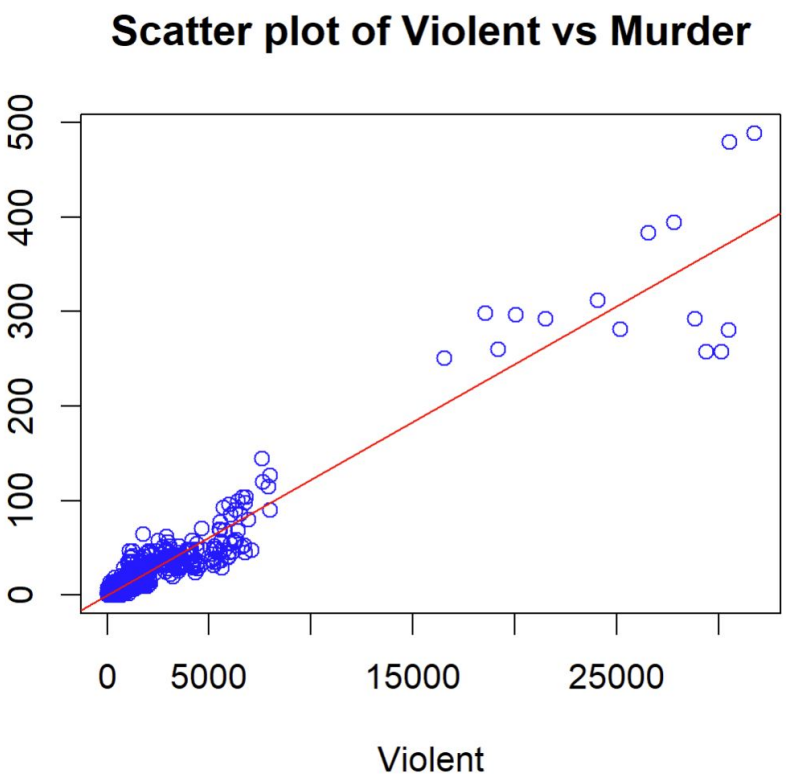
Visualization -Scatter Plot and QQ-Plot

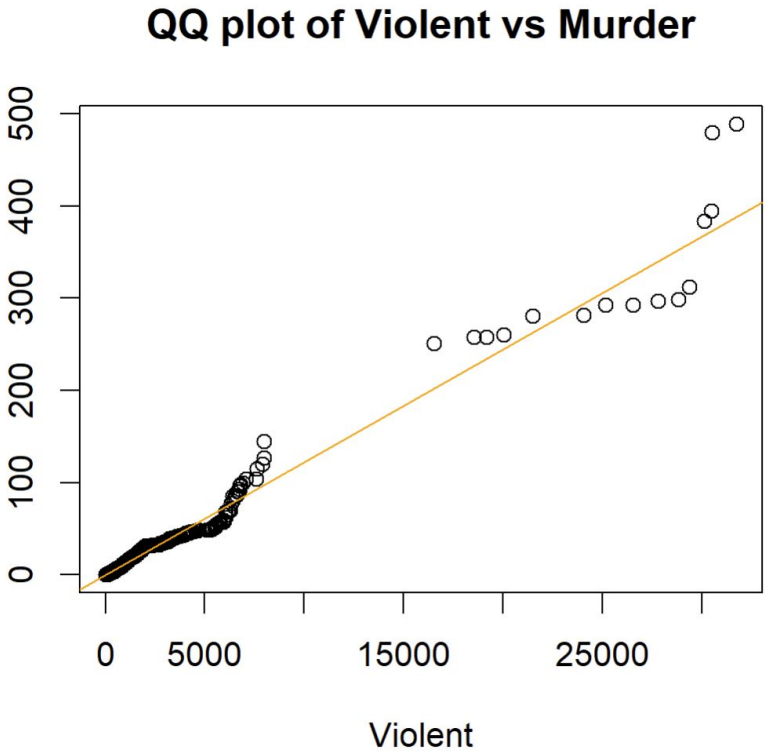




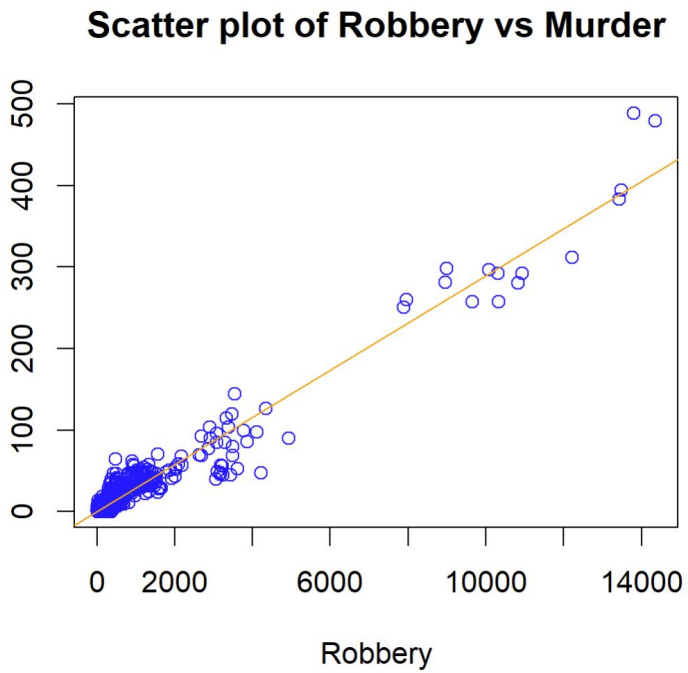
Correlation: 0.5077996

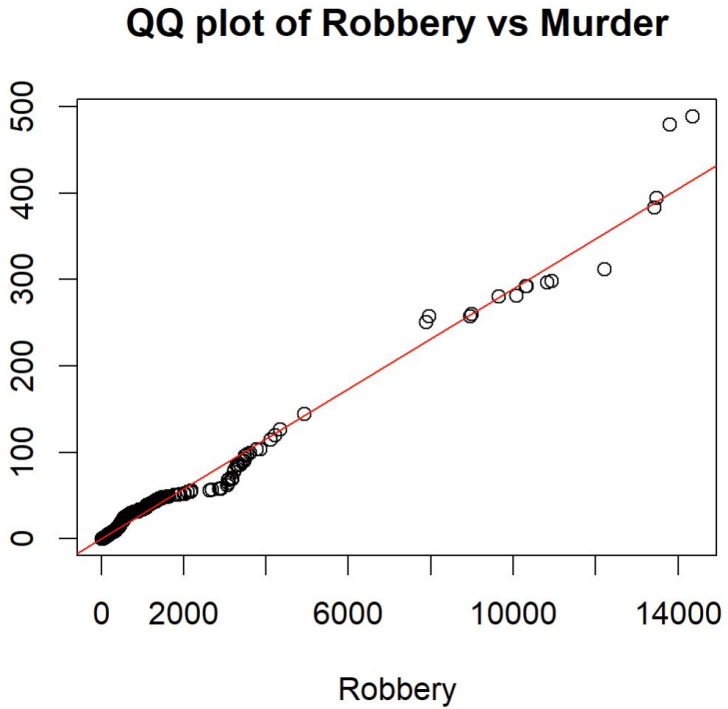
Visualization -Scatter Plot and QQ-Plot



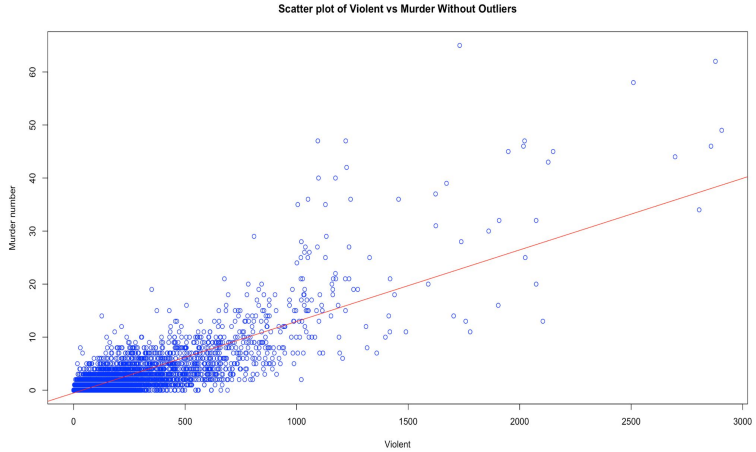


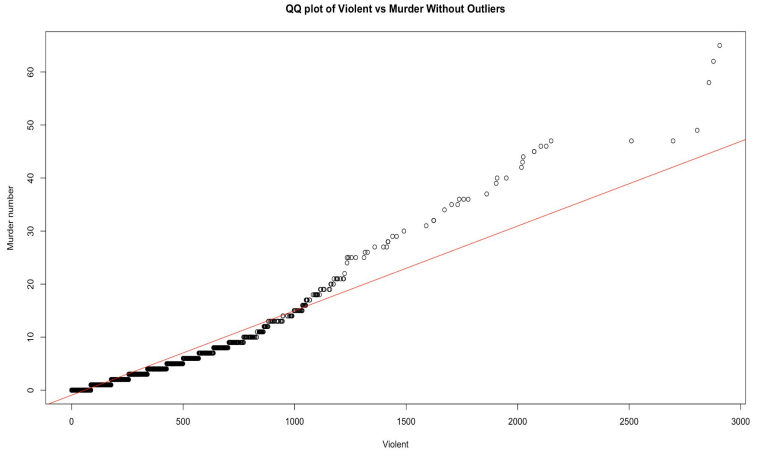
Visualization -Scatter Plot and QQ-Plot

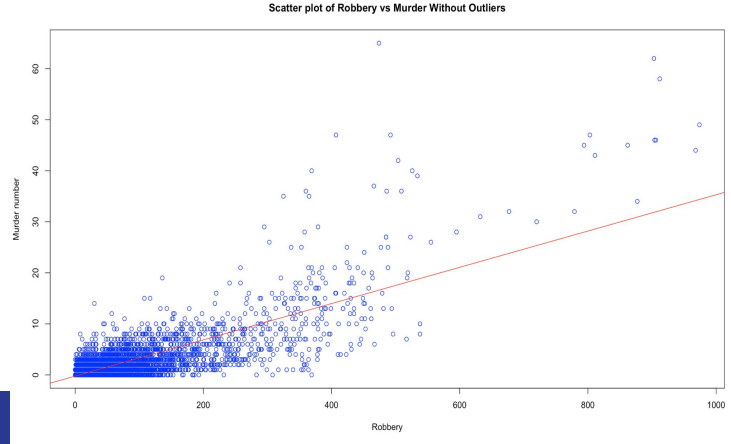


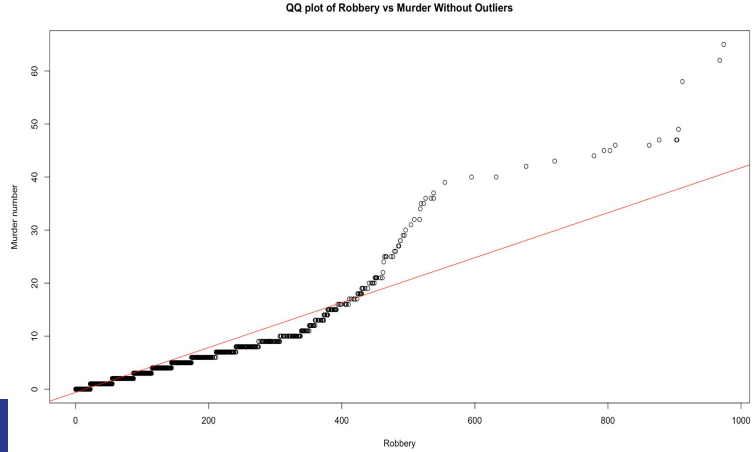


Visualization -Scatter Plot and QQ-Plot

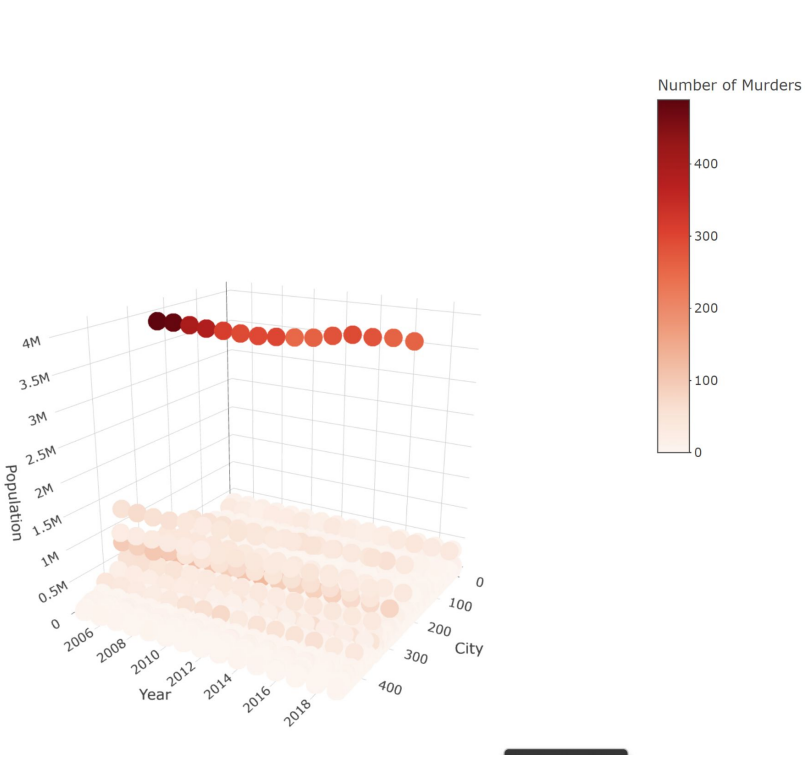




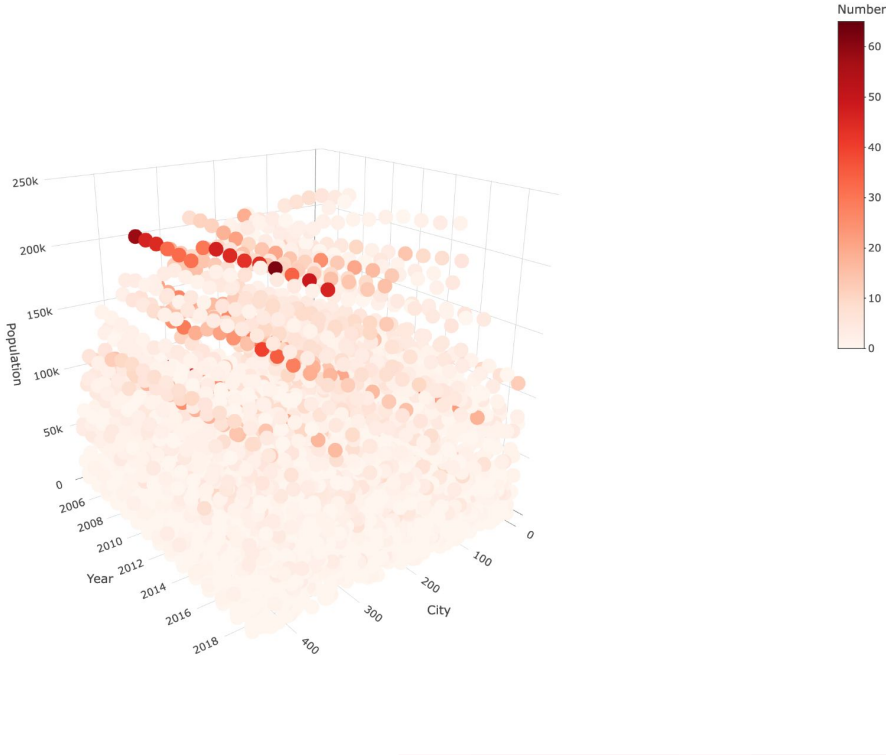




Population vs Year vs City 3d Graph

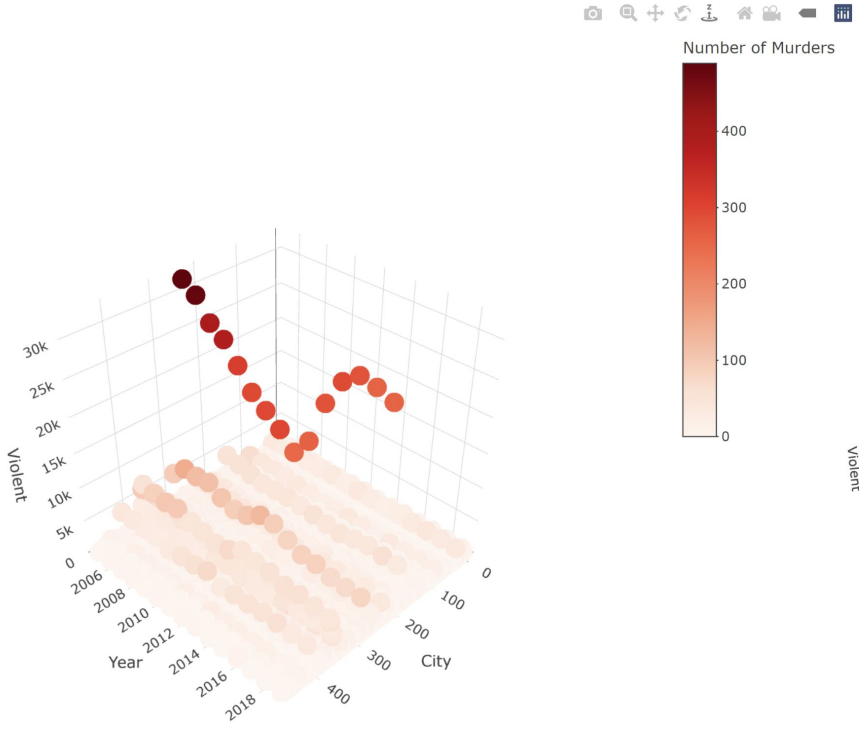


With Outliers

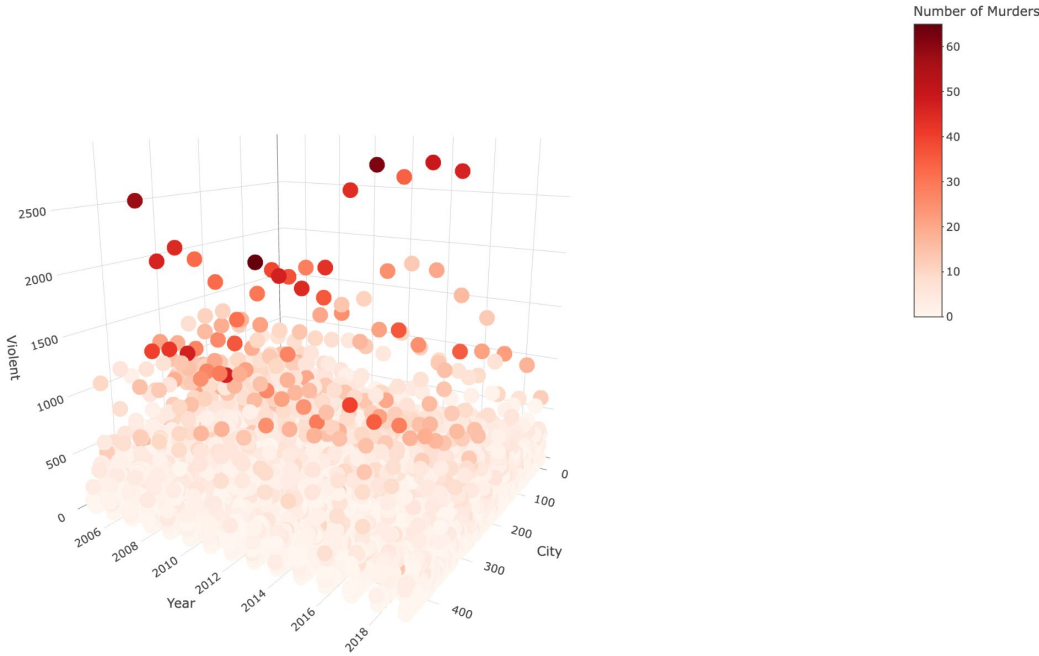


Without Outliers

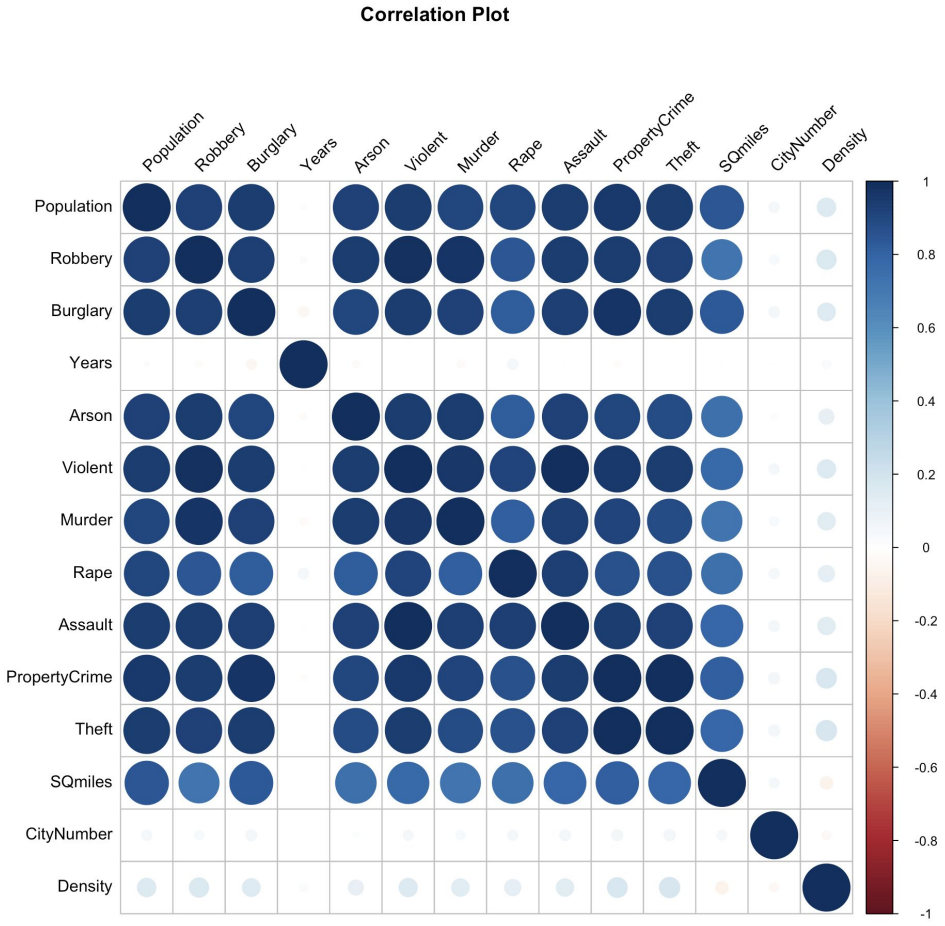
Violent Crime vs Year vs City 3d Graph



With Outliers

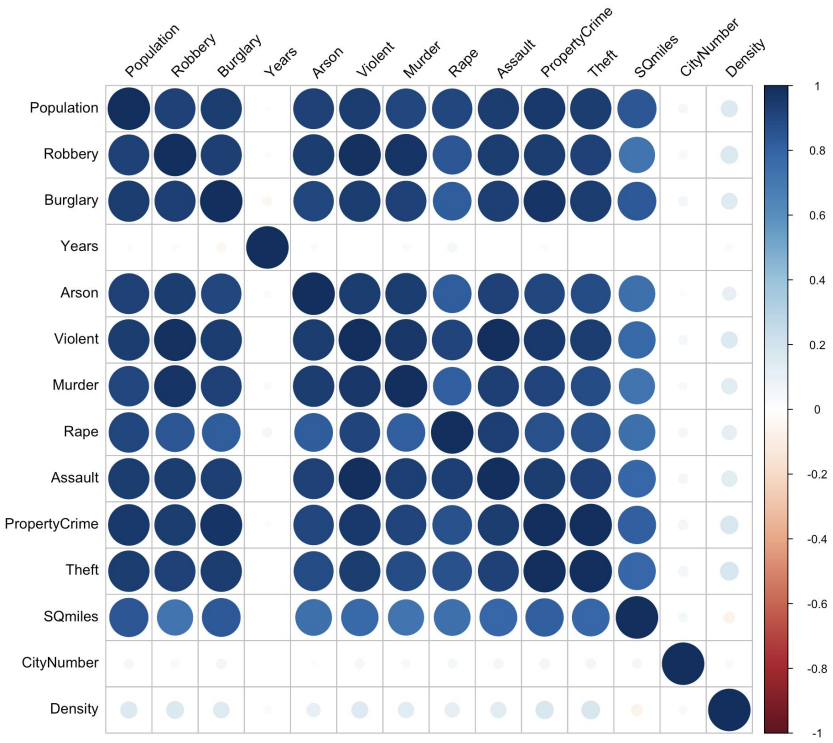


Without Outliers

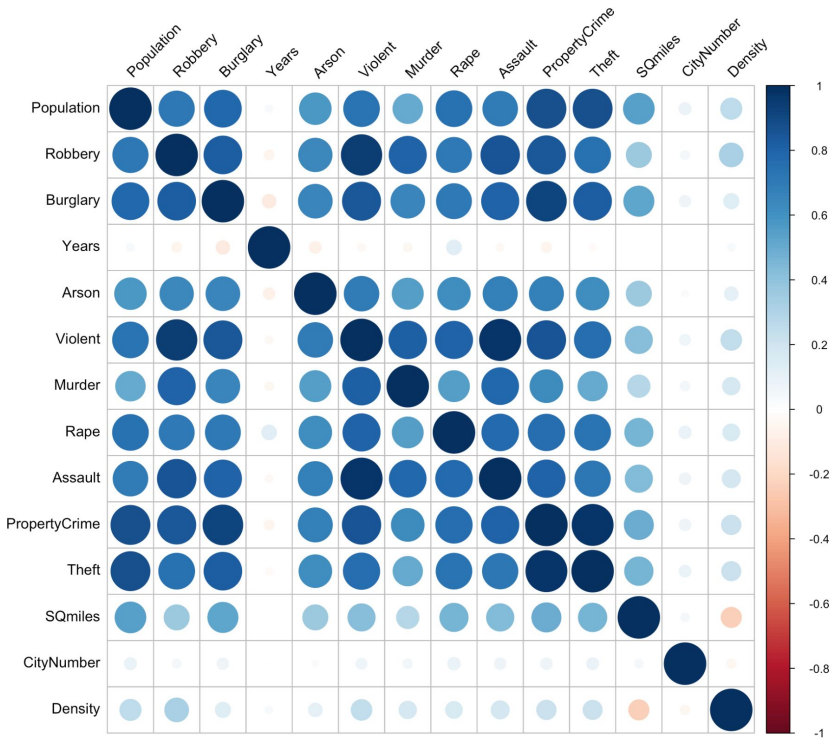


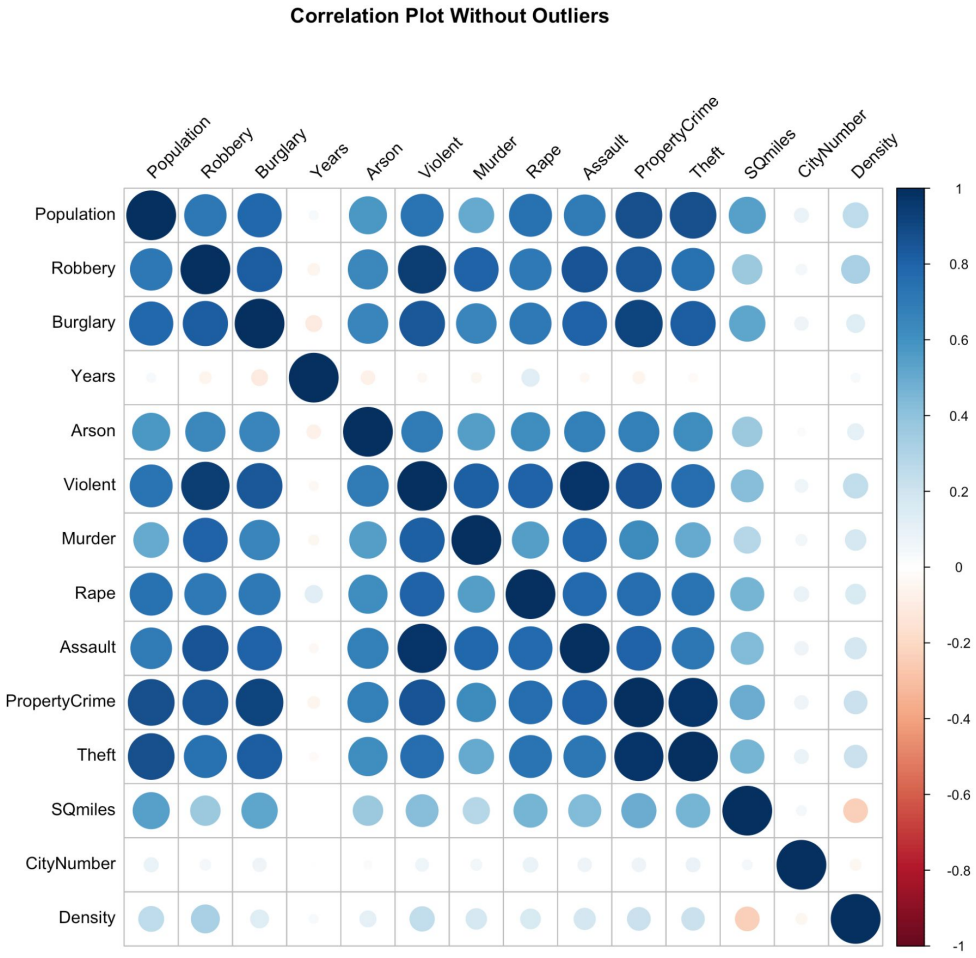

```
correlation_matrix<-cor(numericMergedMurder)
correlation_matrix

# Visualize the correlation matrix
corrplot(correlation_matrix, method = "square",
         # type = "lower",
         tl.col = "black",
         tl.srt = 45,
         diag = TRUE,
         #addCoef.col = "yellow",
         number.cex = 0.7,
         title = "Correlation Plot for Normalized Data",
         tl.pos = "lt",
         mar = c(0,0,1,0),
         addgrid.col = "gray"  # Add grid lines for better
readability
         )
```

Correlation Plot



Correlation Plot Without Outliers





Interesting Pointers

Density and Robbery has correlation

Violent Crime is pretty much correlated with everything

Number of murders isn't highly correlated with Population

Library

Visualization

- EnvStats
- scatterplot3d
- ggplot2
- plotly
- corrplot

Cleaning and Merging

- tidyverse
- vroom
- readr