

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

Uniform Crime Reporting Program
75 YEARS OF PROVIDING CRIME DATA TO THE NATION

CRIME in the United States 2005

U.S. Department of Justice
Federal Bureau of Investigation

[Crime in the United States Home](#) | [Violent Crime](#) | [Property Crime](#) | [Clearances](#) | [Persons Arrested](#) | [Police Employees](#) | [About CIUS](#)

[Return to Previous Page](#)

Table 8

CALIFORNIA
Offenses Known to Law Enforcement
by State by City, 2005

U.S. DEPARTMENT OF JUSTICE • FEDERAL BUREAU OF INVESTIGATION • CRIMINAL JUSTICE INFORMATION SERVICES DIVISION



2019 CRIME in the UNITED STATES

[Criminal Justice Information Services Division](#) [Feedback | Contact Us](#) | [Data Quality Guidelines](#) | [UCR Home](#)

[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 City, 2019

Scrapper

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

The screenshot shows the homepage of the 'CRIME IN THE UNITED STATES 2010' website. At the top, there's a navigation bar with links for 'Criminal Justice Information Services Division', 'Feedback | Contact Us', 'Data Quality Guidelines', and 'UCR Home'. Below this, a main menu includes 'CIUS Home', 'Offenses Known to Law Enforcement', 'Violent Crime', 'Property Crime', 'Clearances', 'Persons Arrested', and 'Police Employee Data'. A sub-menu for 'Offenses Known to Law Enforcement' is expanded, showing options like 'Violent Crime', 'Property Crime', and 'Clearances', each with a list of specific offense types. A red arrow points from the R code to the 'Violent Crime' section of this menu.

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

justifiable homicides.

Violent Crime

Violent crime is composed of four offenses: murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault. Violent crimes are defined in the UCR Program as those offenses which involve force or threat of force.

```
stateLinkz %in% stateLinkz2{  
  grep("table_8", stateLinkz, fixed = TRUE) | grep("table_8", stateLinkz2, fixed = TRUE)  
}  
break
```

Property crime includes the offenses of burglary, larceny-theft, motor vehicle theft, and arson. The object of the theft-type offenses is the taking of money or property, but there is no force or threat of force against the victims. More information about property crime and an overview of property crime data for 2010 is provided in the Property Crime section of this report.

Clearances

This screenshot shows the 'Data Tables' section of the website. It features a sidebar with 'Browse by' categories: 'Violent Crime', 'Property Crime', and 'Clearances', each with a list of offense types. Below this is a 'Browse by Links' sidebar with categories: 'National data', 'Region', 'State totals', 'County agency', 'City agency', 'Universities and colleges', 'State, tribal, and other agencies'. The main content area is titled 'Data Tables' and contains a list of 22 tables, numbered 1 through 22. A red box highlights 'Table 8', and another red box highlights 'Expanded Offense Data' under the 'Clearances' category in the sidebar. A red arrow points from the R code to the 'Clearances' sidebar.

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

Table 8

Offenses Known to Law Enforcement by State by City, 2010

Data Declaration Download Excel

- ▶ Alabama
- ▶ Alaska
- ▶ Arizona
- ▶ Arkansas
- ▶ California
- ▶ Colorado
- ▶ Connecticut
- ▶ Delaware
- ▶ District Of Columbia
- ▶ Florida
- ▶ Georgia
- ▶ Hawaii
- ▶ Idaho
- ▶ Illinois
- ▶ Indiana
- ▶ Iowa
- ▶ Kansas
- ▶ Kentucky
- ▶ Louisiana
- ▶ Maine
- ▶ Maryland
- ▶ Massachusetts
- ▶ Michigan
- ▶ Minnesota
- ▶ Mississippi
- ▶ Missouri
- ▶ Nebraska
- ▶ Nevada
- ▶ New Hampshire
- ▶ New Jersey
- ▶ New Mexico
- ▶ New York
- ▶ North Carolina
- ▶ North Dakota
- ▶ Ohio
- ▶ Oklahoma
- ▶ Oregon
- ▶ Pennsylvania
- ▶ Rhode Island
- ▶ South Carolina
- ▶ South Dakota
- ▶ Tennessee
- ▶ Texas
- ▶ Utah
- ▶ Vermont
- ▶ Washington
- ▶ Wisconsin
- ▶ Wyoming

Scrapper

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

Sys.sleep(5)
```

HOME | Crime in the U.S. | 2010 | Crime in the U.S. | 2010 | Radios | Radios | Contact | California

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	64,987	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

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: /admin/

Disallow: /cgi-bin/

Disallow: /includes/

Disallow: /Includes/

Disallow: /images/

Disallow: /Images/

Disallow: /scripts/

Disallow: /test/

Disallow: /login

Disallow: /help/

Disallow: /javascript/

Disallow: /css/

Disallow: /cssDirectives/

Disallow: /cgi/

Disallow: /error/

Disallow: /esi/

Disallow: /graphics/

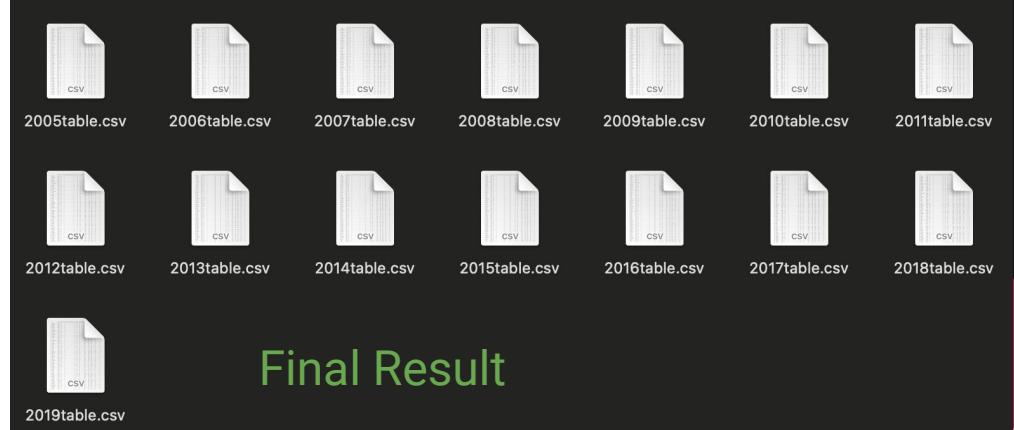
Disallow: /java/

Disallow: /swf/

Disallow: /news/

Disallow: /xml/

Disallow: /ttfman.html



Final Result

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 csv 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

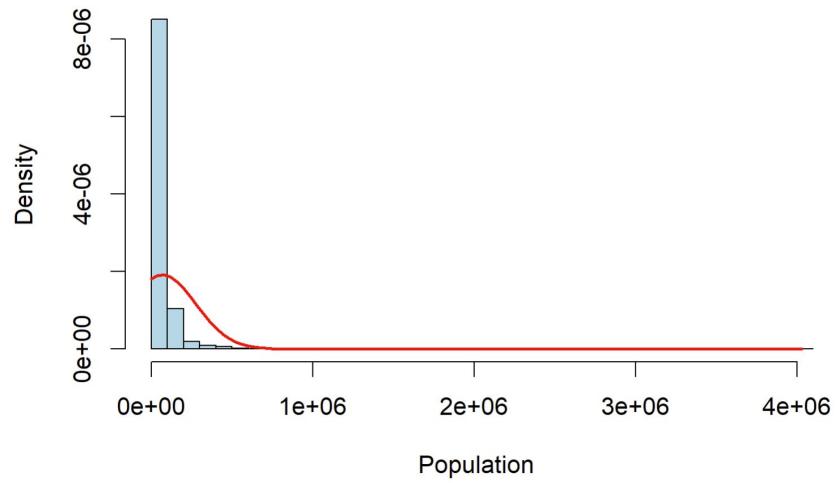
\$ Population	:	num	21786	2...		
\$ Violent	→					
\$ Murder	:	num	1	0	0	0...
\$ Forcible.rape	:	num	7	7	11	...
\$ Robbery	:	num	23	9	88...	
\$ Aggravated.assault	:	num	68	31	1...	
\$ Property.crime	:	num	639	337...		
\$ Burglary	:	num	263	83	...	
\$ Larceny.threft	:	num	260	230...		
\$ Motor.Vehicle.threft	:	num	116	24	...	
\$ Arson	:	num	15	2	21...	
\$ Years	:	num	2005	20	...	
\$ Violent.crime	→					
\$ Murder.and.nonnegligent.manslaughter	:	num	NA	NA	N...	
\$ Larceny.theft	:	num	NA	NA	N...	
\$ Motor.vehicle.theft	:	num	NA	NA	N...	
\$ Arson1	:	num	NA	NA	N...	
\$ Larceny..theft	:	num	NA	NA	N...	
\$ city	:	chr	NA	NA	N...	
\$ population	:	num	NA	NA	N...	
\$ violent	→					
\$ murder	:	num	NA	NA	N...	
\$ rape	:	num	NA	NA	N...	

- Fix redundancy

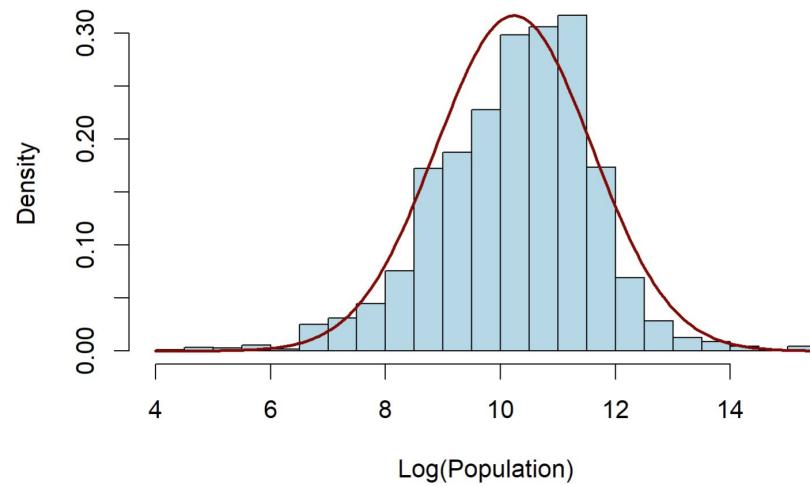
```
combinedAllSamples <- combinedAllSamples %>%  
  mutate(Violent = case_when(  
    !is.na(violent) ~ violent,  
    !is.na(Violent.crime) ~ Violent.crime,  
    !is.na(violent) ~ violent,  
    TRUE ~ NA_real_  
  ))  
combinedAllSamples$violent <- NULL  
combinedAllSamples$Violent1 <- NULL  
combinedAllSamples$Violent.crime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

Population Transformation (Handling NA)

Histogram Population no Outlier



Histogram of Log-transformed Population



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

Cleaning

Result

	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

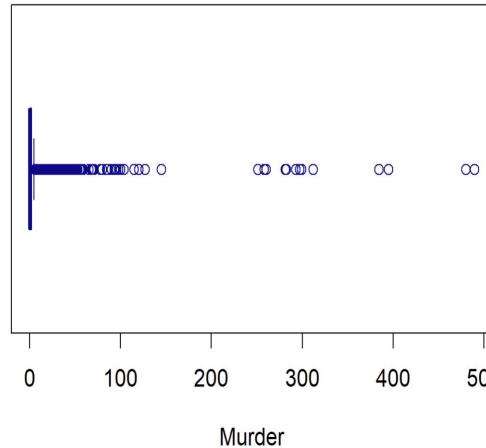
```
area$X1<-NULL  
area$City<-area$X3  
area$SQmiles<- area$X2  
  
area$X2<-NULL  
area$X3<-NULL  
#delete row 1  
area<-area[-c(1), ]  
#get ridNULL#get rid of outliers  
  
#delete everything after comma ,  
area$City<-gsub(",.*", "", area$city)  
#delete sq.miles for each entry  
area$SQmiles<-gsub("s.*", "", area$SQmiles)  
  
#convert into numeric values  
area$SQmiles<-as.numeric(area$SQmiles)
```

	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

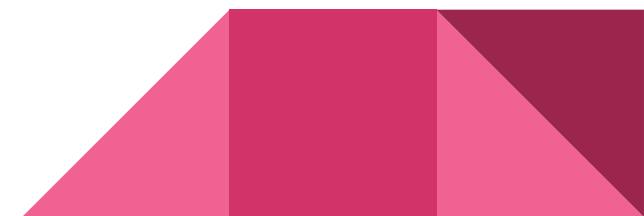
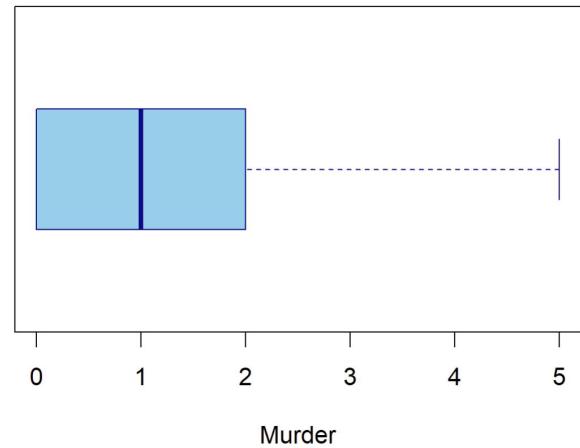
```
boxplot(mergedMurder$Murder,
        main = "Boxplot of Mu
        xlab = "Murder",
        ylab = "Murder ",
        outline = TRUE, # wi
        col = "skyblue", # C
        border = "darkblue",
        horizontal = TRUE)
```

Boxplot of Murder

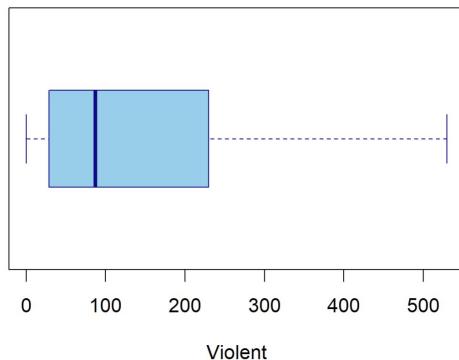


```
#### boxplot of Murder wit
boxplot(mergedMurder$Murder,
        main = "Boxplot of Mu
        xlab = "Murder",
        ylab = "Murder ",
        outline = FALSE, # w
        col = "skyblue", # C
        border = "darkblue",
        horizontal = TRUE)
```

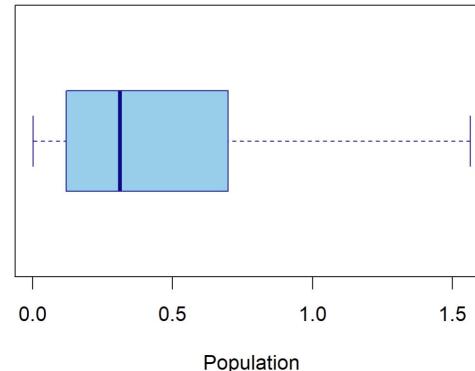
Boxplot of Murder



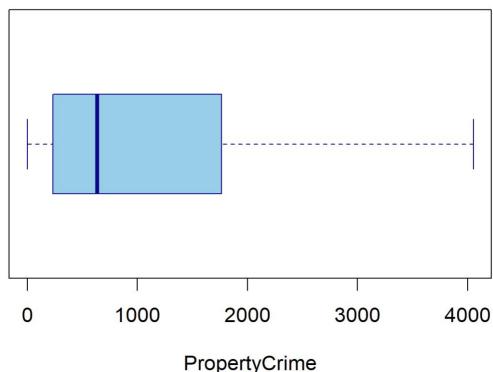
Boxplot of Violent with No outlier



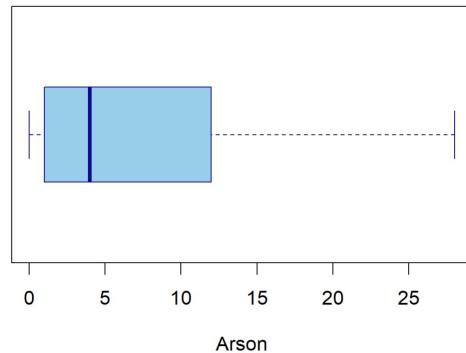
Boxplot of Population with No outlier(in 100K)



Boxplot of Property Crime with No outlier



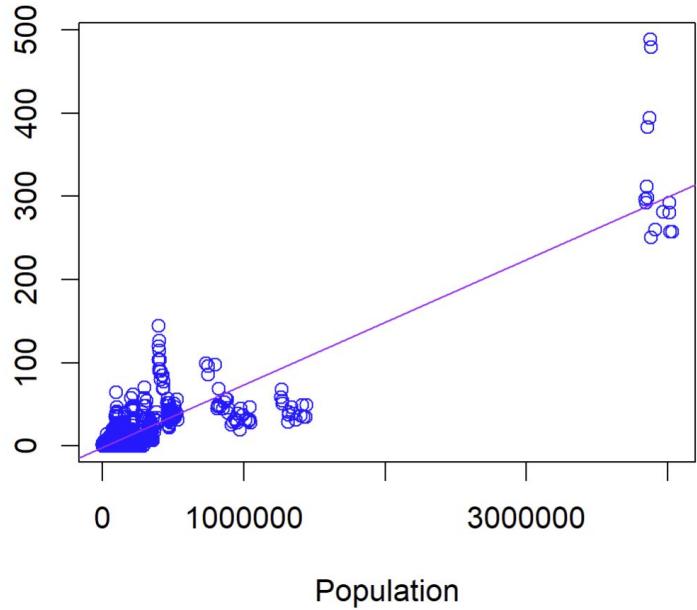
Boxplot of Arson with No outlier



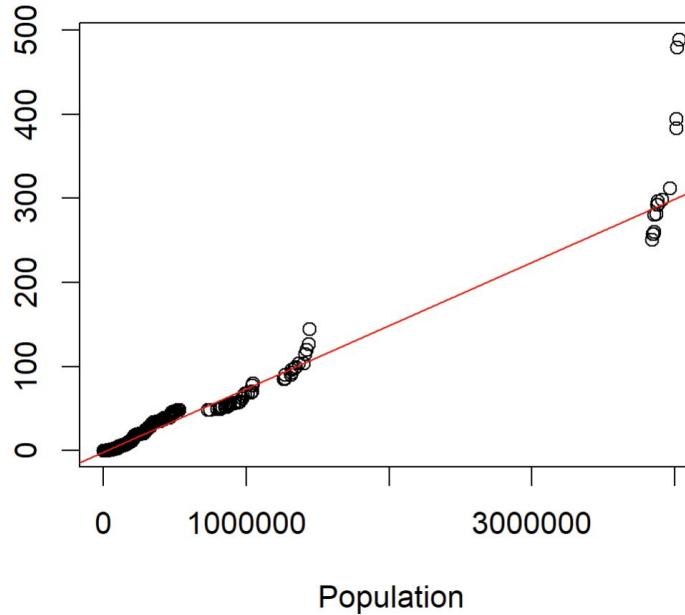
Deleting outlier from data

Visualization -Scatter Plot and QQ-Plot

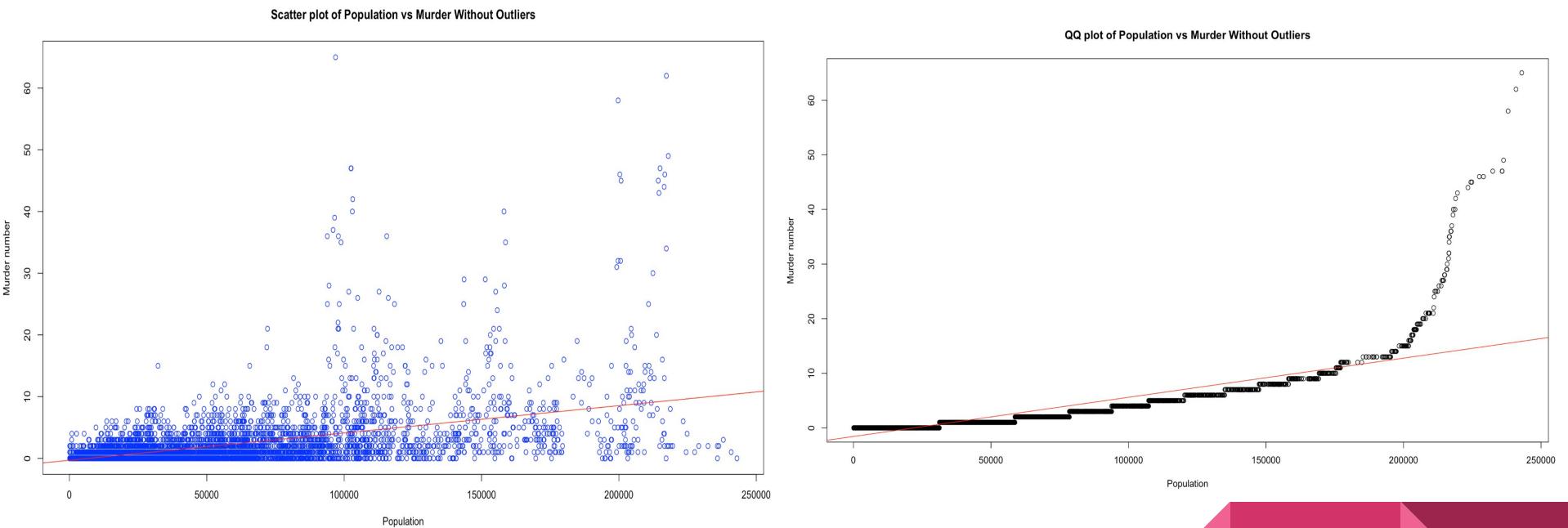
Scatter plot of Population vs Murder



QQ plot of Population vs Murder



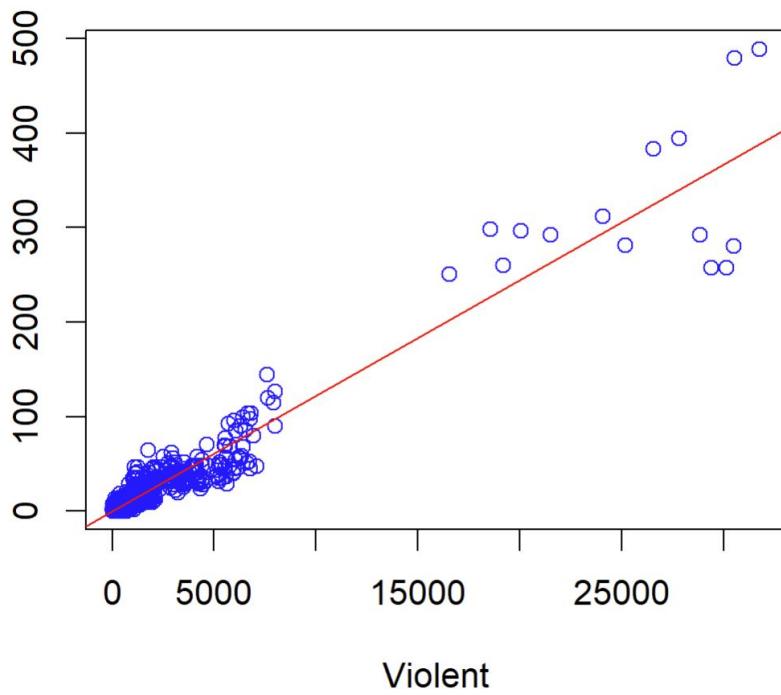
Visualization -Scatter Plot and QQ-Plot



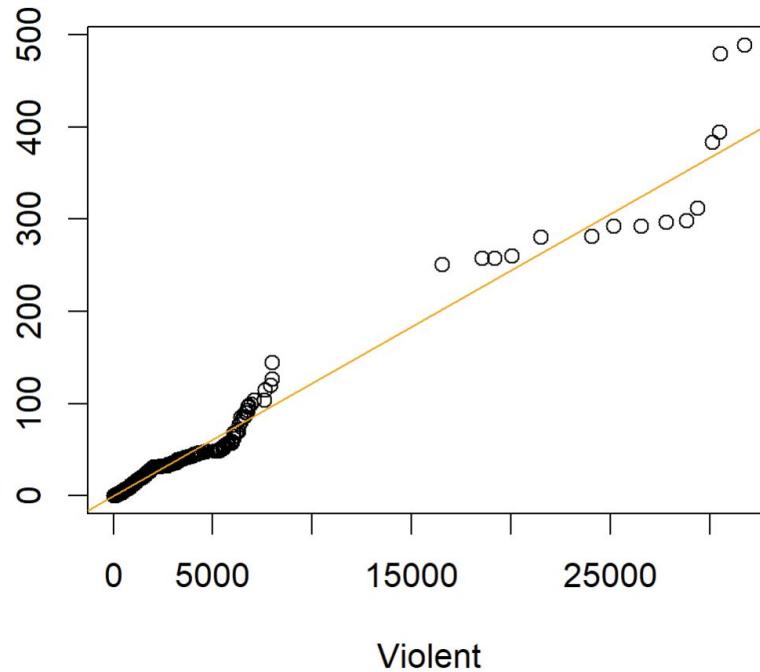
Correlation: 0.5077996

Visualization -Scatter Plot and QQ-Plot

Scatter plot of Violent vs Murder

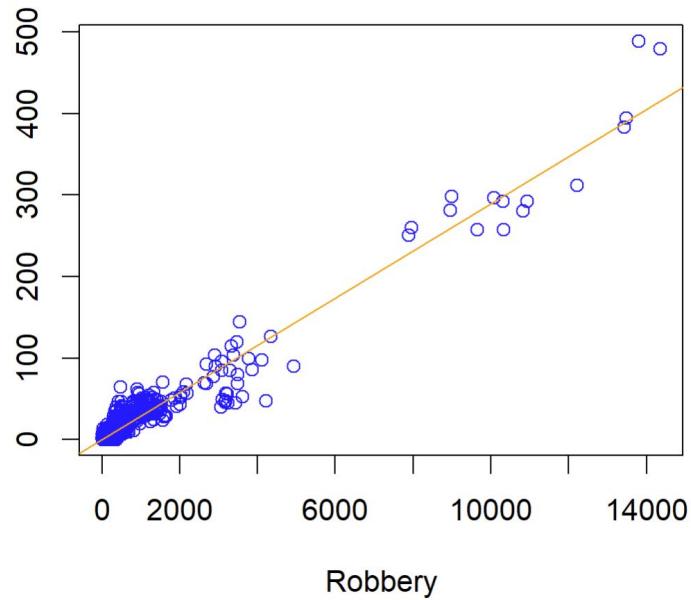


QQ plot of Violent vs Murder

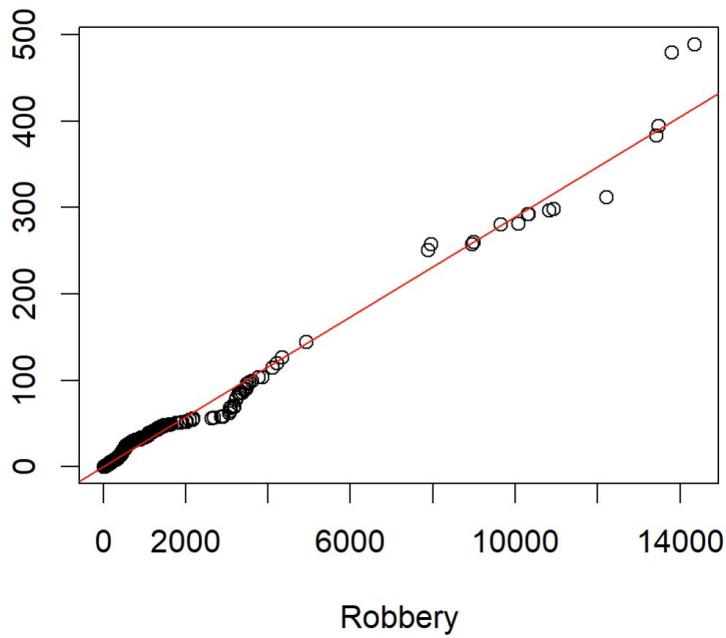


Visualization -Scatter Plot and QQ-Plot

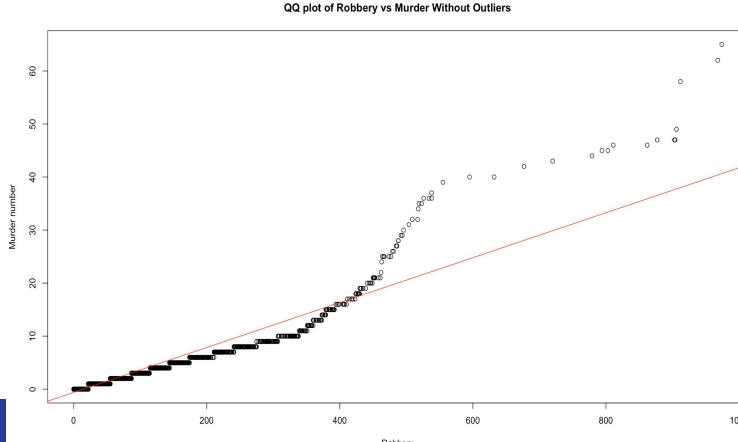
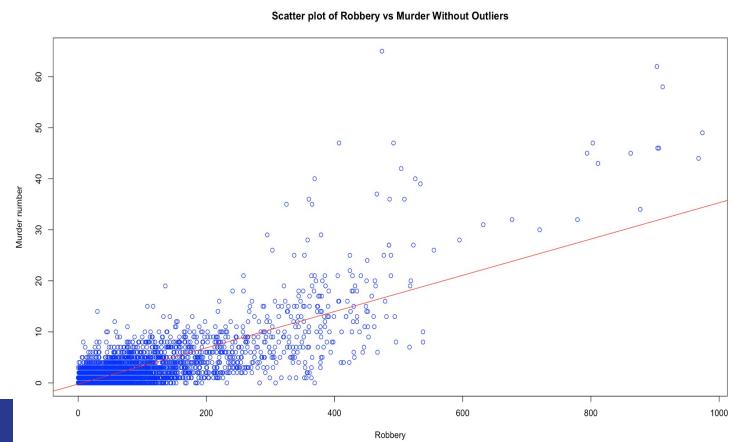
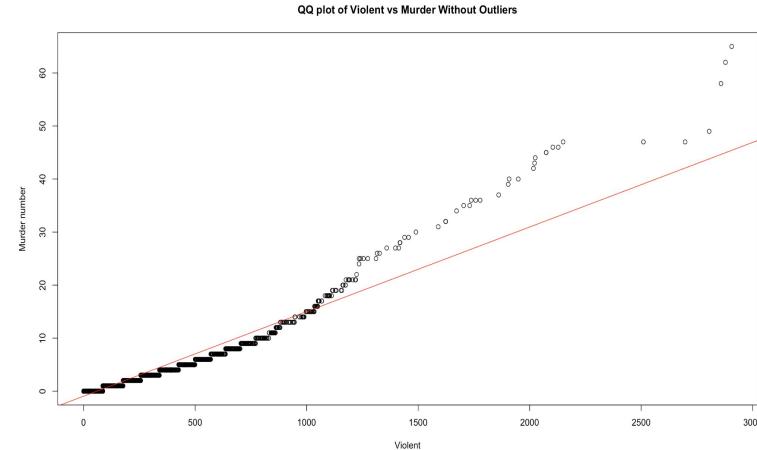
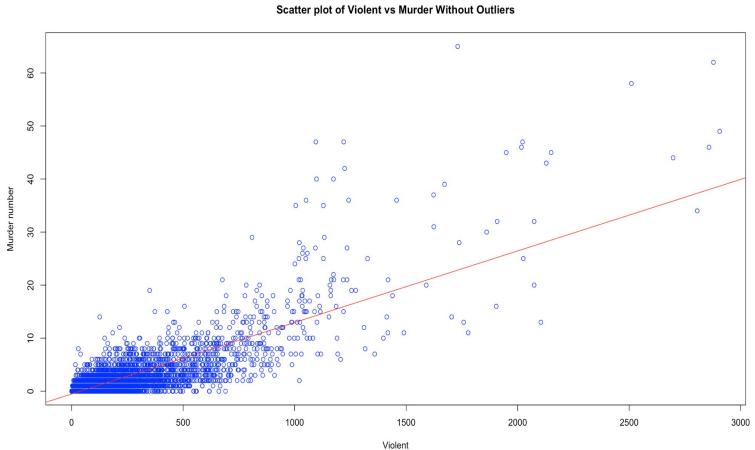
Scatter plot of Robbery vs Murder



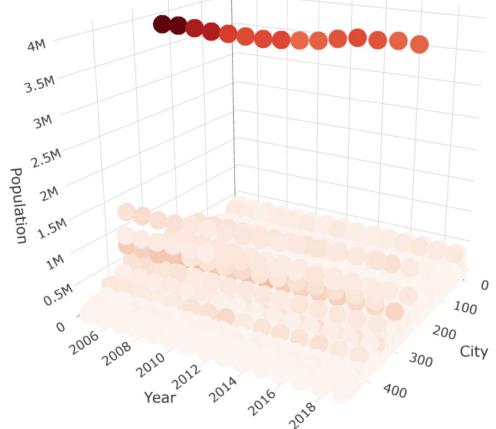
QQ plot of Robbery vs Murder



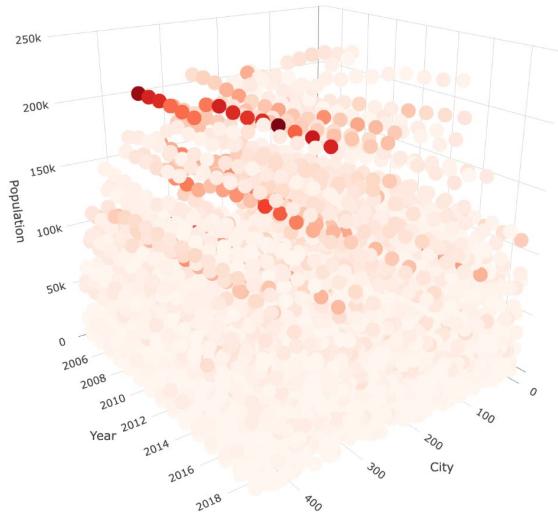
Visualization -Scatter Plot and QQ-Plot



Population vs Year vs City 3d Graph

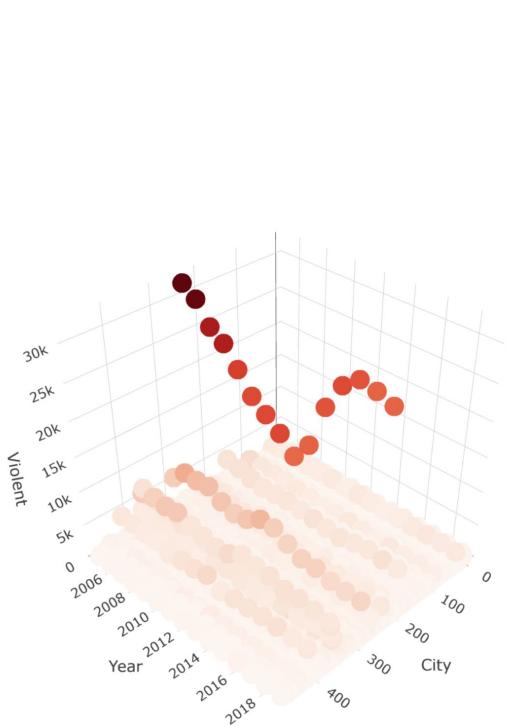


With Outliers

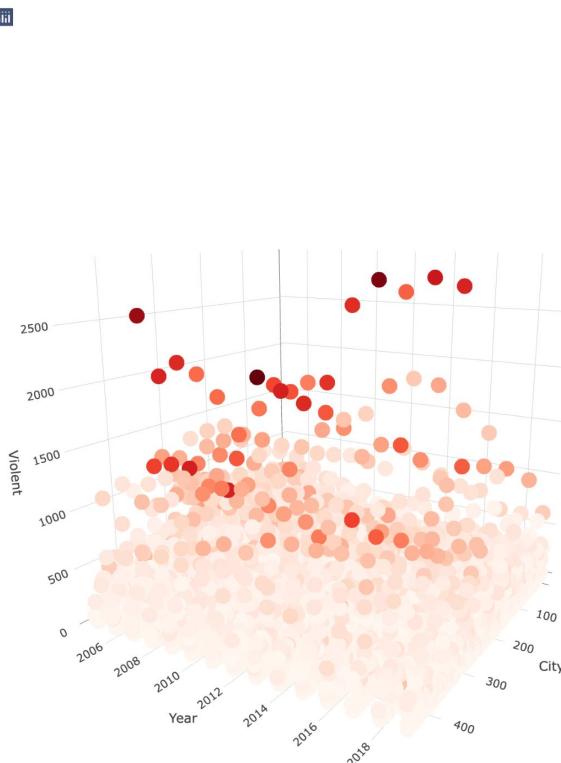


Without Outliers

Violent Crime vs Year vs City 3d Graph



With Outliers



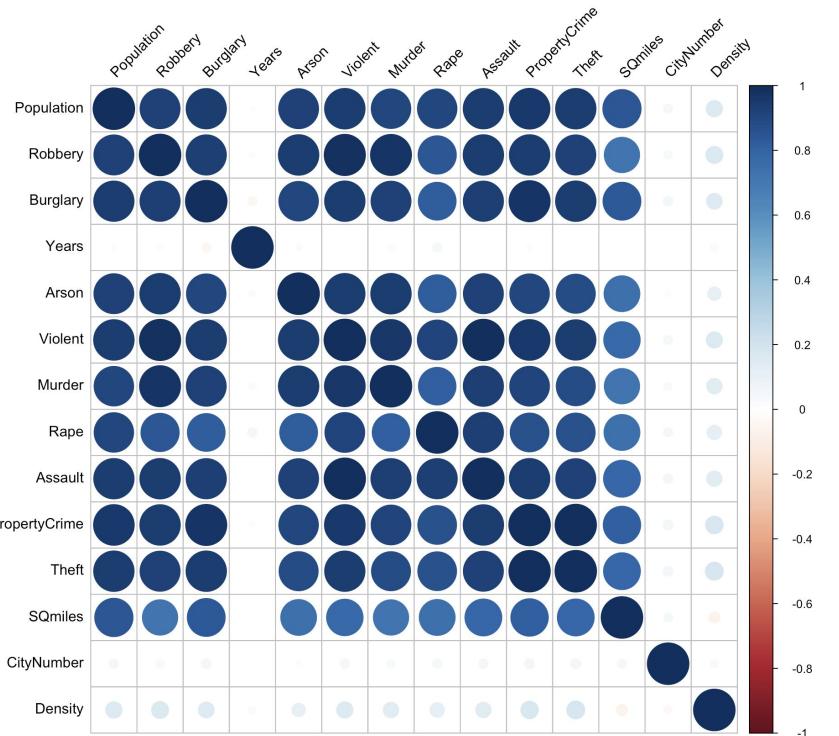
Without Outliers



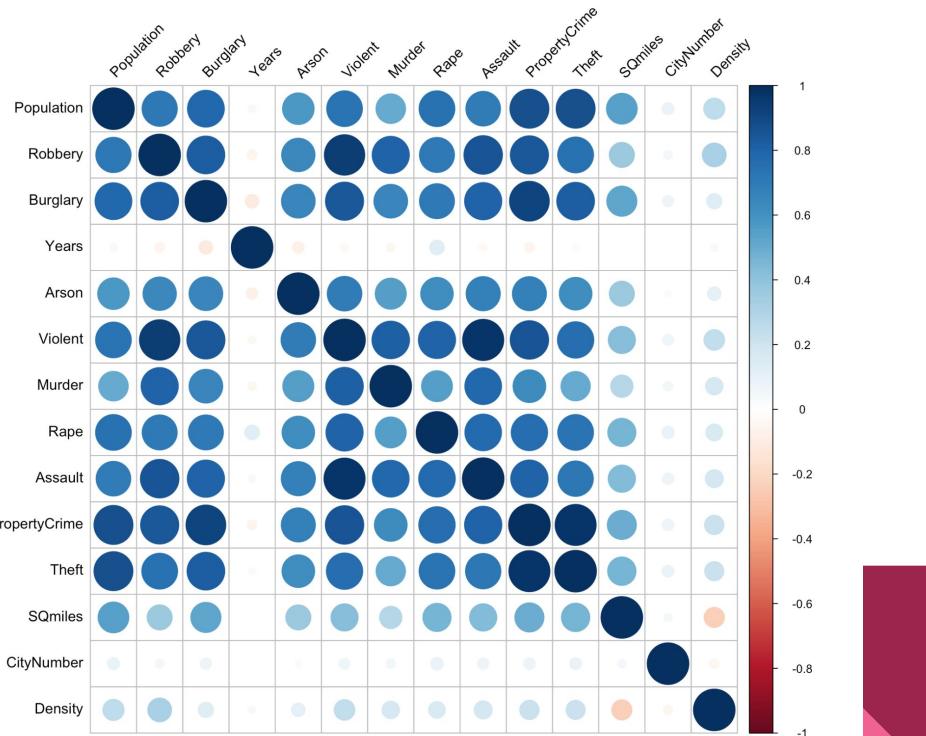
```
correlation_matrix<-cor(numericMergedMurder)
correlation_matrix
```

```
# Visualize the correlation matrix
corplot(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
      )
```

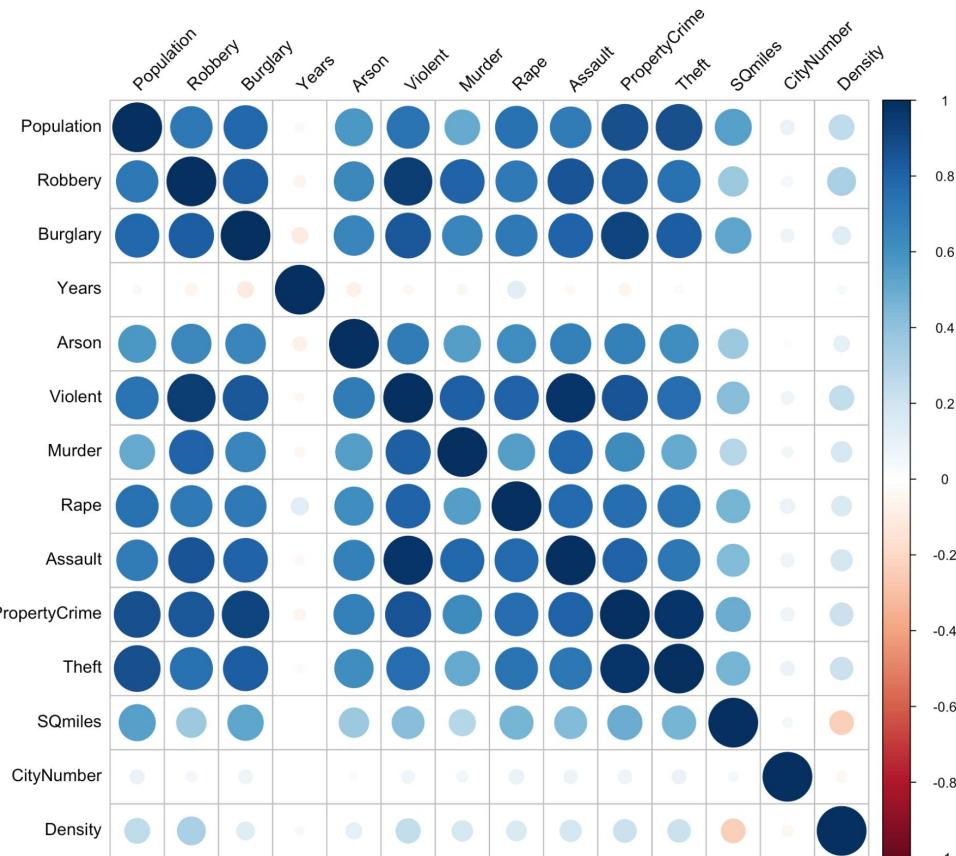
Correlation Plot



Correlation Plot Without Outliers



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