Name:	Prashil Deepak Kadam
UID:	2021600031
Experiment No:	4
Aim:	To perform data preprocessing and EDA a Law and Order / Crime dataset in RStudio using R.
Dataset link:	https://www.kaggle.com/datasets/mayase/crime-data-from-2020-to-present

Code:	library(dplyr)
	library(lubridate)
	library(tidyr)
	
	Crime_data_from_2020
	cleaned_crime_df <- Crime_Data_from_2020_to_Present
	
	cleaned_crime_df$`TIME OCC` <- format(strptime(cleaned_crime_df$`TIME OCC`, "%H%M"), "%H:%M")
	
	cleaned_crime_df <- cleaned_crime_df %>%
	  mutate(
	  `Weapon Desc` = ifelse(is.na(`Weapon Desc`), "UNKNOWN WEAPON/OTHER WEAPON", `Weapon Desc`),
	  `Vict Sex` = ifelse(is.na(`Vict Sex`), "X", `Vict Sex`),
	  `Vict Descent` = ifelse(is.na(`Vict Descent`), "Unknown", `Vict Descent`)
	  )
	
	descent_dict <- c(
	  'A' = 'Other Asian', 'B' = 'Black',
	  'C' = 'Chinese', 'D' = 'Cambodian',
	  'F' = 'Filipino', 'G' = 'Guamanian',
	  'H' = 'Hispanic/LATin/Mexican', 'I' = 'American Indian/Alaskan Native',
	  'J' = 'Japanese', 'K' = 'Korean', 'L' = 'Laotian',
	  'O' = 'Other', 'P' = 'Pacific Islander', 'S' = 'Samoan',
	  'U' = 'Hawaiian', 'V' = 'Vietnamese', 'W' = 'White',
	  'X' = 'Unknown', 'Z' = 'Asian Indian'
	)
	cleaned_crime_df$`Vict Descent` <- recode(cleaned_crime_df$`Vict Descent`, !!!descent_dict)
	
	cleaned_crime_df$`DATE OCC` <- as.Date(cleaned_crime_df$`DATE OCC`, format = "%Y-%m-%dT%H:%M:%S")
	cleaned_crime_df$`Date Rptd` <- as.Date(cleaned_crime_df$`Date Rptd`, format = "%Y-%m-%dT%H:%M:%S")

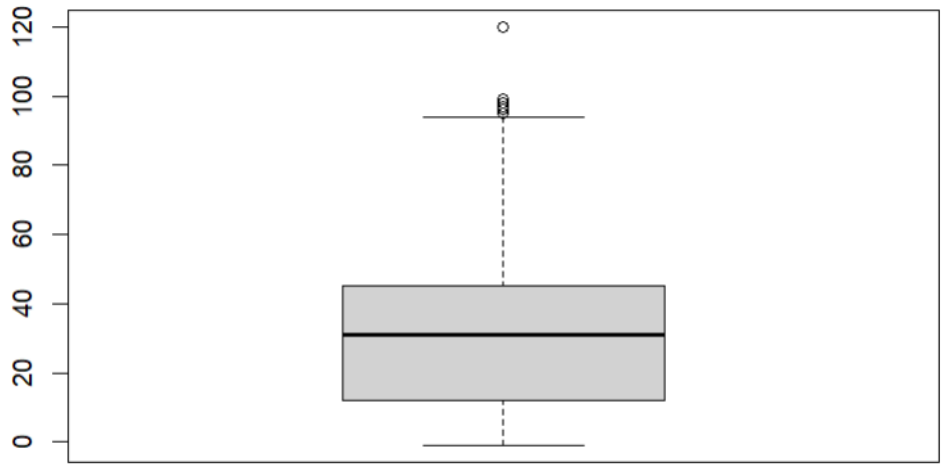
```
unique(cleaned_crime_df$`Vict Age`)

sum(duplicated(cleaned_crime_df))

colSums(is.na(cleaned_crime_df))

boxplot(cleaned_crime_df$`Vict Age`)

str(cleaned_crime_df)

df <- cleaned_crime_df %>% group_by(`Vict Sex`) %>% summarise(count =
n())

df


library(dplyr)
library(lubridate)
# library(leaflet)    # Equivalent of folium in R
library(plotly)     # For interactive plots
library(tidyr)
library(janitor)    # For additional cleaning utilities

dim(cleaned_crime_df)

str(cleaned_crime_df)

cleaned_crime_df <- cleaned_crime_df %>%
  mutate(
    `Vict Age` = as.integer(`Vict Age`),
    `Crm Cd` = as.integer(`Crm Cd`),
    `AREA` = as.integer(`AREA`),
    `Rpt Dist No` = as.integer(`Rpt Dist No`),
    `DR_NO` = as.integer(`DR_NO`),
    `LON` = as.numeric(`LON`),
    `LAT` = as.numeric(`LAT`)
  )


cleaned_crime_df <- cleaned_crime_df %>%
  mutate(
    `DATE OCC` = as.POSIXct(`DATE OCC`, format =
"%Y-%m-%dT%H:%M:%S"),
    `Date Rptd` = as.POSIXct(`Date Rptd`, format =
"%Y-%m-%dT%H:%M:%S")
  )

cleaned_crime_df <- cleaned_crime_df %>%
  mutate(
    month = month(`DATE OCC`, label = TRUE, abbr = FALSE),
    month_num = month(`DATE OCC`),
```

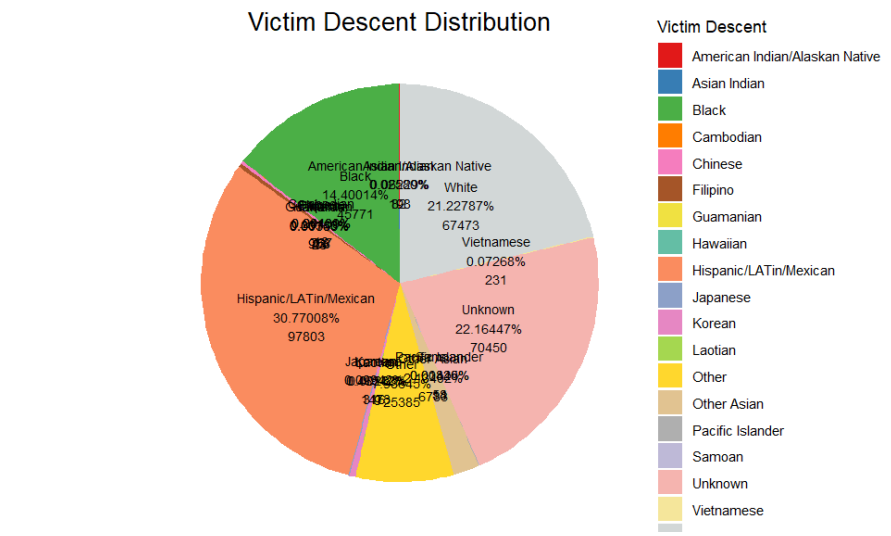
```
    year = year(`DATE OCC`)
  )


cleaned_crime_df <- cleaned_crime_df %>%
  mutate(`Crm Cd Desc` = tools::toTitleCase(`Crm Cd Desc`))

clean_military_time <- function(time_int) {
  # Check if the input is NA
  if (is.na(time_int)) {
    return(NA)
  }

  time_str <- as.character(time_int)

  if (nchar(time_str) == 3) {
    time_mod <- paste0("0", substr(time_str, 1, 1), ":", substr(time_str, 2, 3))
  } else if (nchar(time_str) == 4) {
    time_mod <- paste0(substr(time_str, 1, 2), ":", substr(time_str, 3, 4))
  } else if (nchar(time_str) == 1) {
    time_mod <- paste0("00:0", time_str)
  } else if (nchar(time_str) == 2 && as.integer(time_str) <= 59) {
    time_mod <- paste0("00:", time_str)
  } else if (nchar(time_str) == 2 && as.integer(time_str) > 59) {
    time_mod <- paste0("0", substr(time_str, 1, 1), ":", substr(time_str, 2, 2),
"0")
  } else {
    time_mod <- NA
  }

  return(format(strptime(time_mod, "%H:%M"), "%H:%M"))
}

cleaned_crime_df <- cleaned_crime_df %>%
  mutate(`TIME OCC` = sapply(`TIME OCC`, clean_military_time),
         `TIME OCC` = as.POSIXct(`TIME OCC`, format = "%H:%M"))

cleaned_crime_df <- cleaned_crime_df %>%
  mutate(
    `Weapon Desc` = ifelse(is.na(`Weapon Desc`), 'UNKNOWN
WEAPON/OTHER WEAPON', `Weapon Desc`),
    `Vict Sex` = ifelse(is.na(`Vict Sex`), 'Unknown', `Vict Sex`),
    `Vict Descent` = ifelse(is.na(`Vict Descent`), 'Unknown', `Vict Descent`),
    `Premis Cd` = ifelse(is.na(`Premis Cd`), 256, `Premis Cd`),
    `Premis Desc` = ifelse(is.na(`Premis Desc`), 'Unknown', `Premis Desc`)
  )

colSums(is.na(cleaned_crime_df))
sum(duplicated(cleaned_crime_df))

cleaned_crime_df <- cleaned_crime_df %>%
```

```
  distinct()

cleaned_crime_df %>%
  filter(`Vict Age` > 100 | `Vict Age` < 0)

cleaned_crime_df <- cleaned_crime_df %>%
  filter(`Vict Age` <= 100, `Vict Age` >= 0)

mean_age <- mean(cleaned_crime_df$`Vict Age`[cleaned_crime_df$`Vict
Age` != 0])
median_age <- median(cleaned_crime_df$`Vict Age`[cleaned_crime_df$`Vict
Age` != 0])

unique(cleaned_crime_df$`Vict Age`)

cleaned_crime_df <- cleaned_crime_df %>%
  mutate(`Vict Sex` = ifelse(`Vict Sex` %in% c('M', 'F', 'Unknown'), `Vict Sex`,
'Unknown'))



cleaned_crime_df <- cleaned_crime_df %>%
  filter(`Vict Descent` != '-')

descent_dict <- c('A' = 'Other Asian', 'B' = 'Black', 'C' = 'Chinese', 'D' =
'Cambodian',
                  'F' = 'Filipino', 'G' = 'Guamanian', 'H' = 'Hispanic/Latin/Mexican',
                  'I' = 'American Indian/Alaskan Native', 'J' = 'Japanese', 'K' =
'Korean',
                  'L' = 'Laotian', 'O' = 'Other', 'P' = 'Pacific Islander', 'S' = 'Samoan',
                  'U' = 'Hawaiian', 'V' = 'Vietnamese', 'W' = 'White', 'X' = 'Unknown',
                  'Z' = 'Asian Indian')
cleaned_crime_df <- cleaned_crime_df %>%
  mutate(`Vict Descent` = recode(`Vict Descent`, !!!descent_dict))

cleaned_crime_df %>%
  filter(`DATE OCC` > `Date Rptd`) %>%
  nrow()

cor(cleaned_crime_df %>% select_if(is.numeric))

library(dplyr)
library(ggplot2)
library(scales)

vict_descent_df <- cleaned_crime_df %>%
  group_by(`Vict Descent`) %>%
  summarise(`Count` = n(), .groups = 'drop')  # Ensure .groups = 'drop' to
ungroup

color_palette <- c("#E41A1C", "#377EB8", "#4DAF4A", "#FF7F00",
"#F781BF",
```

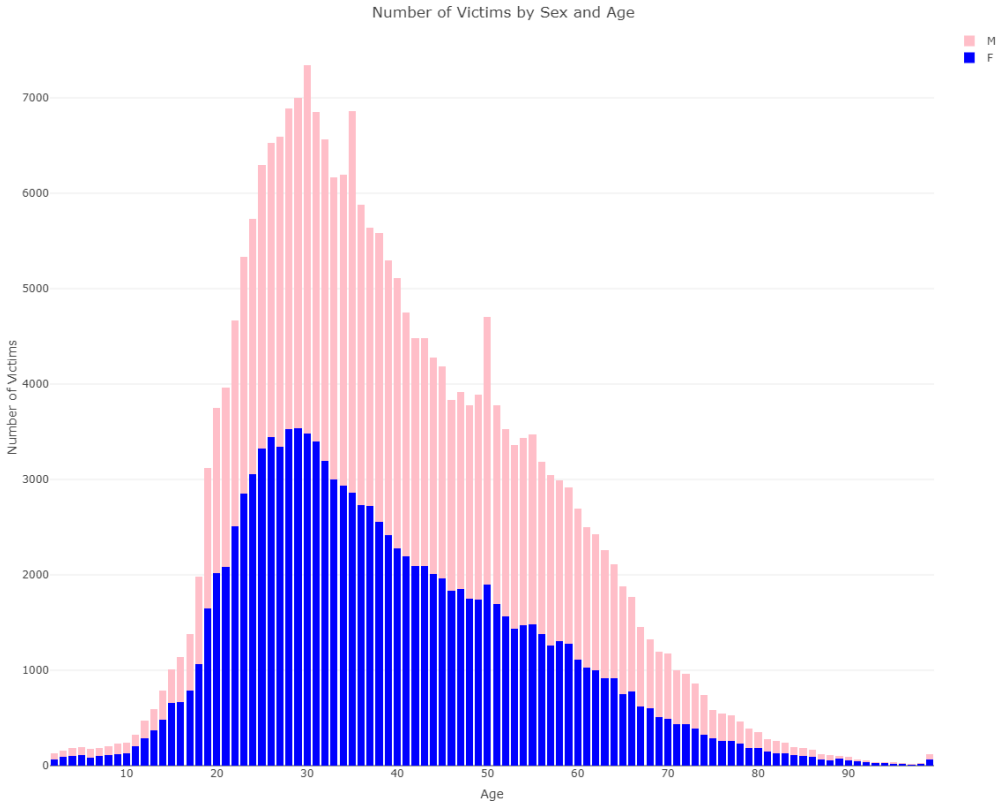
```
              "#A65628", "#F0E442", "#66C2A5", "#FC8D62", "#8DA0CB",
              "#E78AC3", "#A6D854", "#FFD92F", "#E5C494", "#B3B3B3",
              "#BEBADA", "#F5B7B1", "#F9E79F", "#D5DBDB",
"#C0392B")

ggplot(vict_descent_df, aes(x = "", y = `Count`, fill = `Vict Descent`)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  scale_fill_manual(values = color_palette) +
  labs(
    title = "Victim Descent Distribution",
    fill = "Victim Descent",
    y = "Number of Victims"
  ) +
  theme_void() + # Remove axis lines and labels
  theme(
    plot.title = element_text(hjust = 0.5, size = 16),
    legend.position = "right"
  ) +
  geom_text(aes(label = paste0(`Vict Descent`, "\n", percent(`Count` /
sum(`Count`)), "\n", `Count`)),
            position = position_stack(vjust = 0.5), size = 3)


######################################################################

victim_sex_age_df <- cleaned_crime_df %>%
  filter(`Vict Age` > 0, !is.na(`Vict Sex`) & `Vict Sex` != "Unknown") %>%
  group_by(Vict_Sex = `Vict Sex`, Vict_Age = `Vict Age`) %>%
  summarize(Number_of_Victims = n(), .groups = 'drop')

fig <- plot_ly(
  data = victim_sex_age_df,
  x = ~Vict_Age,
  y = ~Number_of_Victims,
  color = ~Vict_Sex,
  type = 'bar',
  colors = c('blue', 'pink'),
  height = 900
) %>%
  layout(
    title = 'Number of Victims by Sex and Age',
    xaxis = list(title = 'Age'),
    yaxis = list(title = 'Number of Victims'),
    barmode = 'stack'
  )

fig
######################################################################

weapon_crime_df <- cleaned_crime_df %>%
```

```
  group_by(Weapon_Desc = `Weapon Desc`) %>%
  summarize(Number_of_Crimes = n(), .groups = 'drop')

# Create the bar plot
fig <- plot_ly(
  data = weapon_crime_df,
  x = ~Weapon_Desc,
  y = ~Number_of_Crimes,
  type = 'bar',
  height = 900
) %>%
  layout(
    title = 'Number of Crimes by Weapon Type',
    xaxis = list(title = 'Weapon Type', tickangle = -45),  # Rotate x-axis labels
for better readability
    yaxis = list(title = 'Number of Crimes')
  )

fig

######################################################################

crime_type_df <- cleaned_crime_df %>%
  group_by(Crime_Type = `Crm Cd Desc`) %>%
  summarize(Number_of_Crimes = n(), .groups = 'drop')


fig <- plot_ly(
  data = crime_type_df,
  x = ~Crime_Type,
  y = ~Number_of_Crimes,
  type = 'bar',
  height = 900,
  colors = 'red'
) %>%
  layout(
    title = 'Number of Crimes by Crime Type',
    xaxis = list(title = 'Crime Type', tickangle = -45),  # Rotate x-axis labels for
better readability
    yaxis = list(title = 'Number of Crimes')
  )

fig

######################################################################

area_crime_age_df <- cleaned_crime_df %>%
  filter(`Vict Age` > 0) %>%
  group_by(AREA = `AREA`) %>%
  summarize(
    Number_of_Crimes = n(),
```

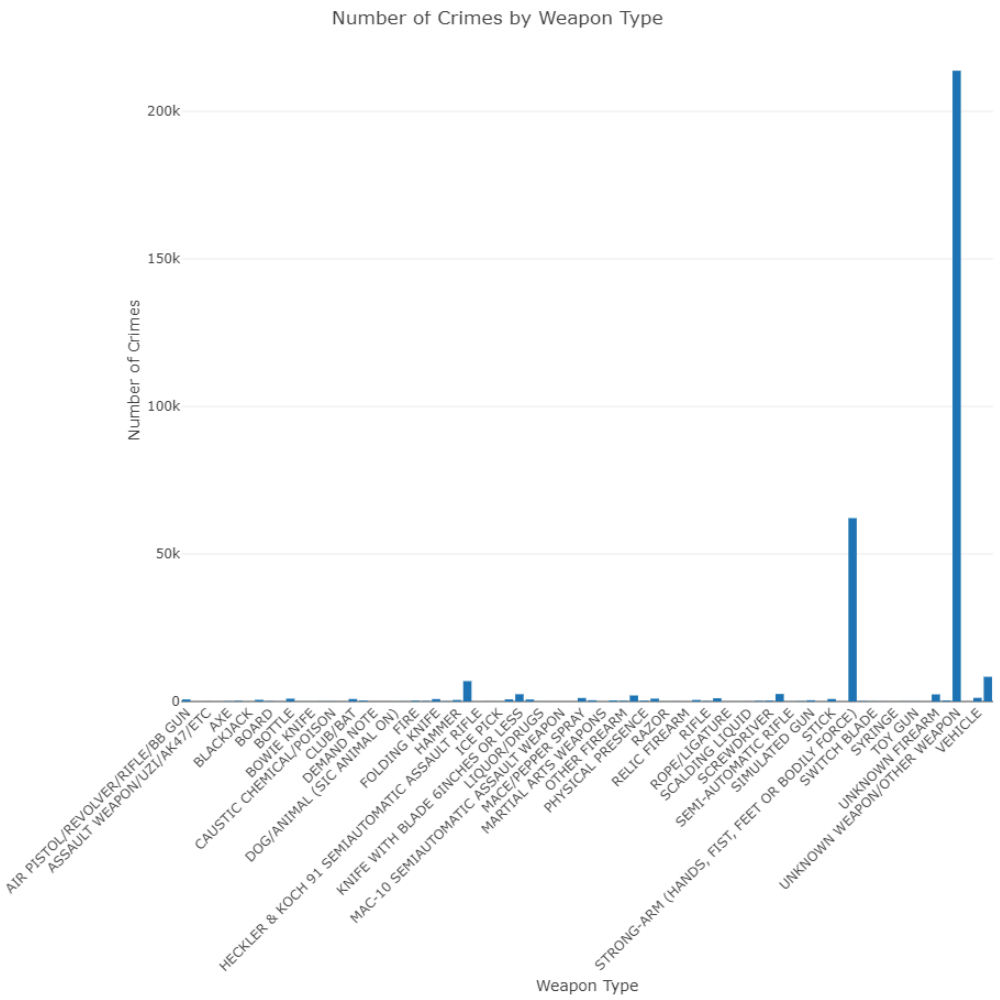
	Avg_Vict_Age = mean(`Vict Age`, na.rm = TRUE), .groups = 'drop') lm_model <- lm(Number_of_Crimes ~ Avg_Vict_Age, data = area_crime_age_df) fig <- plot_ly(data = area_crime_age_df, x = ~Avg_Vict_Age, y = ~Number_of_Crimes, type = 'scatter', mode = 'markers', marker = list(size = 10), height = 900) %>% add_lines(x = area_crime_age_df$Avg_Vict_Age, y = predict(lm_model, area_crime_age_df), line = list(color = 'red', width = 2), name = 'Linear Regression Line') %>% layout(title = 'Linear Regression of Number of Crimes by Average Victim Age', xaxis = list(title = 'Average Victim Age'), yaxis = list(title = 'Number of Crimes')) fig
Results / Outputs:	Box Plot of Victim Ages  Pie Chart of Victim Descent



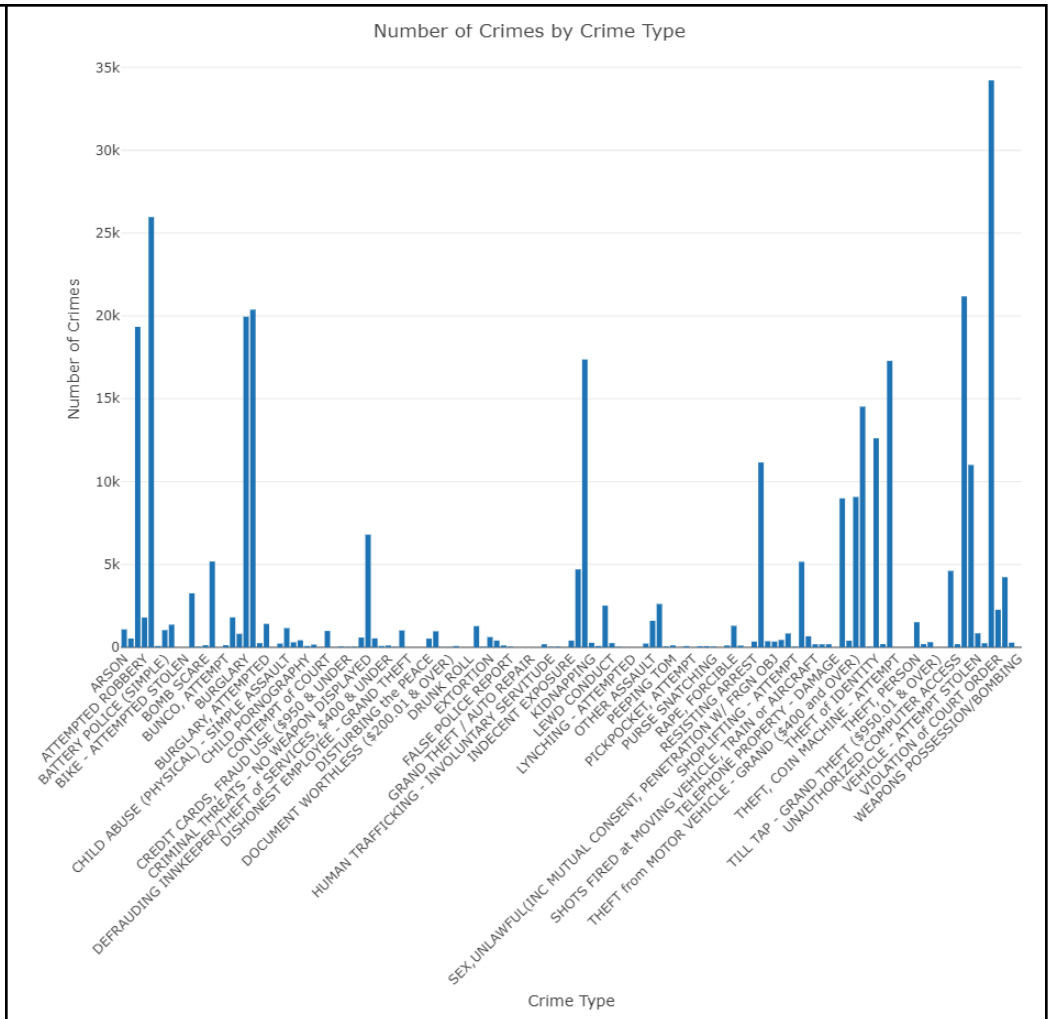
Distribution of number of victims by sex and age

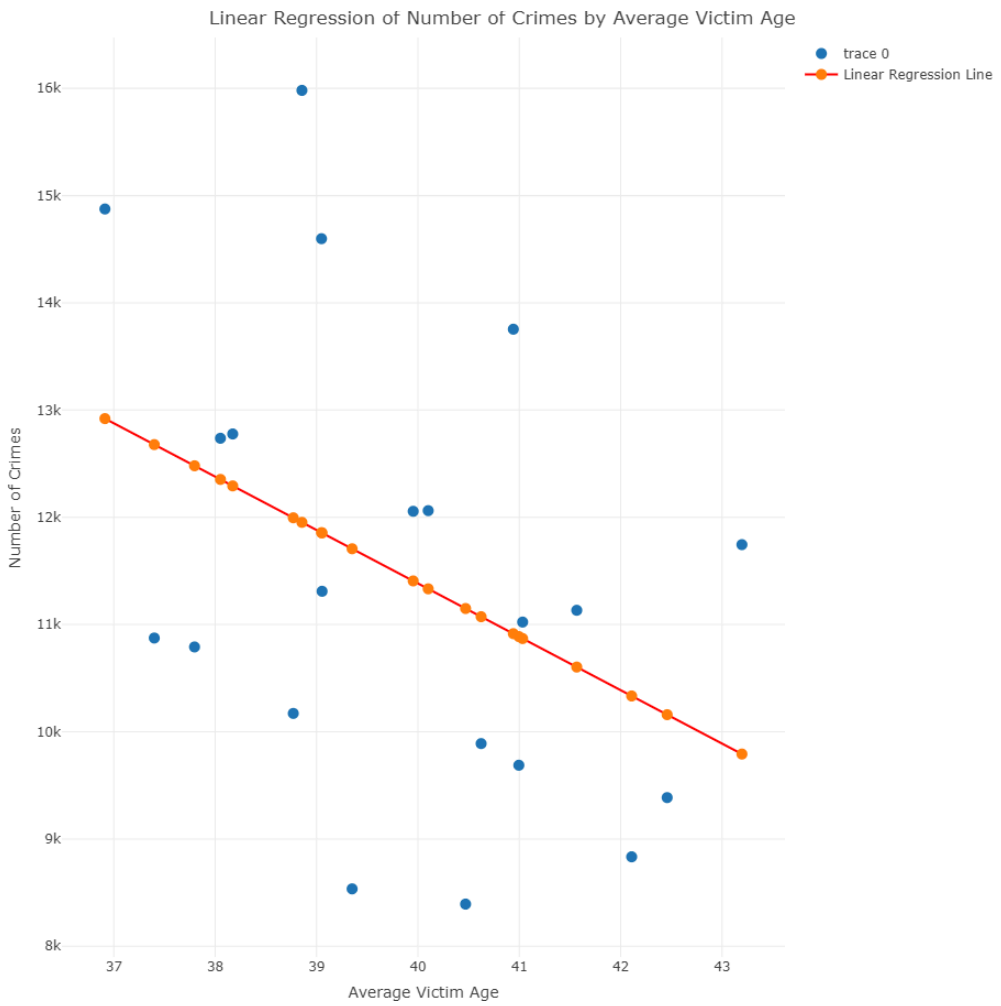


Plot of number of crimes by weapon used



Plot of number of crimes by the crime type


Number of Crimes by Crime Type
Number of Crimes
35k
30k
25k
20k
15k
10k
5k
0
ARSON
ATTEMPTED ROBBERY
BATTERY POLICE (SIMPLE)
BIKE - ATTEMPTED STOLEN
BOMB SCARE
BUNCO, ATTEMPT
BURGLARY
BURGLARY, ATTEMPTED
CHILD ABUSE (PHYSICAL) - SIMPLE ASSAULT
CHILD PORNOGRAPHY
CONTEMPT of COURT
CREDIT CARDS, FRAUD USE ($950 & UNDER
CRIMINAL THREATS - NO WEAPON DISPLAYED
DEFRAUDING INNKEEPER/THEFT of SERVICES, $400 & UNDER
DISHONEST EMPLOYEE - GRAND THEFT
DISTURBING the PEACE
DOCUMENT WORTHLESS ($200.01 & OVER)
DRUNK ROLL
EXTORTION
FALSE POLICE REPORT
GRAND THEFT / AUTO REPAIR
HUMAN TRAFFICKING - INVOLUNTARY SERVITUDE
INDECENT EXPOSURE
KIDNAPPING
LEWD CONDUCT
LYNCHING - ATTEMPTED
OTHER ASSAULT
PEEPING TOM
PICKPOCKET, ATTEMPT
PURSE SNATCHING
RAPE, FORCIBLE
RESISTING ARREST
SEX,UNLAWFUL(INC MUTUAL CONSENT, PENETRATION W/ FRGN OBJ
SHOPLIFTING - ATTEMPT
SHOTS FIRED at MOVING VEHICLE, TRAIN or AIRCRAFT
TELEPHONE PROPERTY - DAMAGE
THEFT from MOTOR VEHICLE - GRAND ($400 and OVER)
THEFT of IDENTITY
THEFT, COIN MACHINE - ATTEMPT
THEFT, PERSON
TILL TAP - GRAND THEFT ($950.01 & OVER)
UNAUTHORIZED COMPUTER ACCESS
VEHICLE - ATTEMPT STOLEN
VIOLATION of COURT ORDER
WEAPONS POSSESSION/BOMBING
Crime Type



Linear regression plot of number of crimes by average victim age

Conclusion

Box Plot of Victim Ages

- The median victim age is around 30-35 years old.
- The interquartile range (middle 50% of victims) is approximately between 15-45 years old.
- There are outliers on the upper end, with some victims over 80 years old.
- The distribution is slightly right-skewed.
- Most victims tend to be young to middle-aged adults.
- There's a wide range of victim ages, suggesting crimes affect people across different life stages.
- Elderly individuals (outliers) are less frequently victimized, but still vulnerable.

Victim Descent Distribution

- Hispanic/Latino/Mexican victims comprise the largest group at about 36.77%.
- White victims are the second largest group at around 21.27%.
- Black victims are the third largest group, represented by a green slice.
- There's a significant "Unknown" category, suggesting data collection issues.
- The crime victim demographics reflect a diverse population.
- Hispanic/Latino communities may be disproportionately affected by crime in this area.
- The large "Unknown" category indicates a need for improved data collection on victim ethnicity.

	Number of Crimes by Crime Type
	- Theft-related crimes (e.g., "THEFT-GRAND" and "BURGLARY") are among the most common.
	- Violent crimes like assault and battery also show high frequencies.
	- There's a wide variety of crime types recorded.
	- Property crimes appear to be more prevalent than violent crimes.
	- Law enforcement may need to focus resources on preventing theft and burglary.
	- The diverse range of crimes suggests a complex criminal landscape requiring varied prevention strategies.
	Number of Victims by Sex and Age
	- There are more male victims (pink) than female victims (blue) across all age groups.
	- The peak for both sexes is around 25-35 years old.
	- The number of victims decreases sharply after age 60 for both sexes.
	- Men are more likely to be victims of reported crimes than women.
	- Young adults are at the highest risk of becoming crime victims.
	- Elderly individuals are less likely to be victims, possibly due to lifestyle factors or underreporting.
	Number of Crimes by Weapon Type
	- The vast majority of crimes involve no weapon or an unknown weapon type.
	- Among identified weapons, firearms appear to be the most common.
	- There's a wide variety of weapon types used in crimes, but most occur at very low frequencies.
	- Most reported crimes do not involve weapons, suggesting a prevalence of non-violent property crimes.
	- When weapons are used, firearms pose the greatest threat.
	- The large "unknown" category suggests challenges in weapon identification or reporting.
	Linear Regression plot for Number of crimes vs Victim Age
	- Strong negative correlation between average victim age and number of crimes is observed.
	- As the average victim age increases from 37 to 43, crime numbers generally decrease.
	- Significant data scatter suggests factors beyond age influence crime rates
	- Younger average victim age associated with higher crime frequency, potentially due to:
	  - Higher victimization rates among younger populations
	  - Greater likelihood of younger victims reporting crimes
	  - Younger individuals' presence in higher-risk areas or situations
	- Linear relationship supports using victim age as a predictor for crime rates
	- Findings could inform targeted crime prevention and resource allocation strategies
	- Correlation doesn't necessarily imply that this is the only factor affecting crime rate directly, other variables may influence both age and crime frequency