Bharatiya Vidya Bhavan's
Sardar Patel Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai)
Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India

Experiment no 4 **Name-Divya Suvarna** **UID-2021200114** **BATCH-A**

Aim :
Create basic charts using R programming language on dataset Crime or Police / Law and Order

- Basic - Bar chart, Pie chart, Histogram, Time line chart, Scatter plot, Bubble plot
- Write observations from each chart

Link: https://www.kaggle.com/datasets/mrayushagrawal/us-crime-dataset/data

Objectives:

- To understand and apply basic data visualization techniques in R.
- To create various types of charts (Bar chart, Pie chart, Histogram, Timeline chart, Scatter plot, Bubble plot) using a crime-related dataset.
- To interpret and analyze the data through visual representations.

Theory:

Data visualization is an essential skill in data analysis that helps in understanding trends, patterns, and relationships within a dataset. R, a powerful statistical programming language, provides a wide range of tools for creating visually appealing and informative charts. In this experiment, we will use basic chart types to analyze crime data and derive insights.

Chart Types:

1. **Bar Chart:** A bar chart is used to display categorical data with rectangular bars representing the frequency or count of each category.
2. **Pie Chart:** A pie chart shows the proportion of categories as slices of a pie, useful for comparing parts of a whole.
3. **Histogram:** A histogram is used to represent the distribution of numerical data by grouping it into bins.
4. **Timeline Chart:** A timeline chart visualizes data points in chronological order, often used to show trends over time.
5. **Scatter Plot:** A scatter plot displays the relationship between two numerical variables using points in a Cartesian plane.
6. **Bubble Plot:** A bubble plot is an extension of a scatter plot where the size of the points (bubbles) represents an additional variable.

Steps to Perform in R:

1. **Set Up the Environment:**
    - Install and load necessary libraries.

R
Copy code

```
# Install necessary packages (if not already installed)
install.packages(c("ggplot2", "dplyr", "tidyr", "lubridate"))

# Load the libraries
library(ggplot2)
library(dplyr)
library(tidyr)
library(lubridate)
```

2. **Load the Dataset:**
    - Load the crime dataset (replace `crime_data.csv` with your dataset's file name).

```
# Load the CSV file
crime_data <- read.csv("/content/US_Crime_DataSet.csv", header = TRUE)
```

3. **Data Preprocessing:**
    - Inspect and clean the data if necessary (handle missing values, filter relevant columns, etc.).

R
Copy code

```
crime_data[crime_data == "Unknown"] <- NA
# Convert 'Month' to a factor with ordered levels
crime_data$Month <- factor(crime_data$Month, levels = month.name)

# Convert 'Year' and 'Month' into a Date column for easier plotting
crime_data$Date <- as.Date(paste0(crime_data$Year, "-", match(crime_data$Month,
month.name), "-01"))

# Handle missing values (e.g., replace NA in numeric columns with 0)
crime_data$Victim.Age[is.na(crime_data$Victim.Age)] <- 0
```

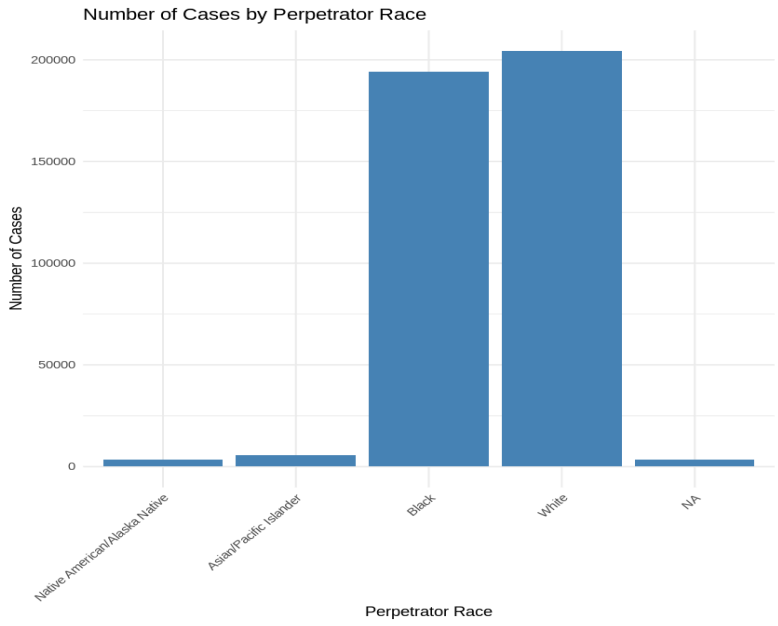
```r
crime_data$Perpetrator.Age[is.na(crime_data$Perpetrator.Age)] <- 0

# Convert 'Crime.Solved' to a factor
crime_data$Crime.Solved <- as.factor(crime_data$Crime.Solved)
```

```
  Record.ID      Agency.Code        Agency.Name        Agency.Type
Min.   :     1   Length:638454      Length:638454      Length:638454
1st Qu.:159614   Class :character   Class :character   Class :character
Median :319228   Mode  :character   Mode  :character   Mode  :character
Mean   :319228
3rd Qu.:478841
Max.   :638454

    City              State               Year         Month
Length:638454      Length:638454      Min.   :1980   Length:638454
Class :character   Class :character   1st Qu.:1987   Class :character
Mode  :character   Mode  :character   Median :1995   Mode  :character
                                      Mean   :1996
                                      3rd Qu.:2004
                                      Max.   :2014

   Incident       Crime.Type        Crime.Solved        Victim.Sex
Min.   :  0.00   Length:638454      Length:638454      Length:638454
1st Qu.:  1.00   Class :character   Class :character   Class :character
Median :  2.00   Mode  :character   Mode  :character   Mode  :character
Mean   : 22.97
3rd Qu.: 10.00
Max.   :999.00

  Victim.Age     Victim.Race        Victim.Ethnicity   Perpetrator.Sex
Min.   :  0.00   Length:638454      Length:638454      Length:638454
1st Qu.: 22.00   Class :character   Class :character   Class :character
Median : 30.00   Mode  :character   Mode  :character   Mode  :character
Mean   : 35.03
3rd Qu.: 42.00
Max.   :998.00

Perpetrator.Age Perpetrator.Race   Perpetrator.Ethnicity Relationship
Min.   : 0.00   Length:638454      Length:638454         Length:638454
1st Qu.: 0.00   Class :character   Class :character      Class :character
Median :21.00   Mode  :character   Mode  :character      Mode  :character
Mean   :20.32
3rd Qu.:31.00
Max.   :99.00
NA's   :1
   Weapon           Victim.Count     Perpetrator.Count Record.Source
Length:638454      Min.   : 0.0000   Min.   : 0.0000   Length:638454
Class :character   1st Qu.: 0.0000   1st Qu.: 0.0000   Class :character
Mode  :character   Median : 0.0000   Median : 0.0000   Mode  :character
                   Mean   : 0.1233   Mean   : 0.1852
                   3rd Qu.: 0.0000   3rd Qu.: 0.0000
                   Max.   :10.0000   Max.   :10.0000
```

4. **Create Visualizations:**

Bar Chart:

```
# Count the number of cases by Perpetrator.Race
perpetrator_race_counts <- crime_data %>%
  count(Perpetrator.Race)

# Create a bar chart
ggplot(perpetrator_race_counts, aes(x = reorder(Perpetrator.Race, n), y = n))
+
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Number of Cases by Perpetrator Race",
       x = "Perpetrator Race",
       y = "Number of Cases") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

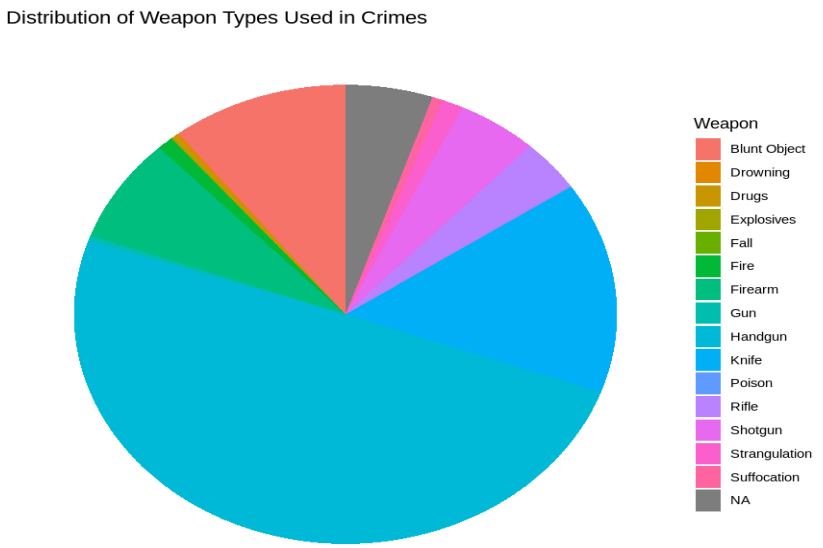


- **Observation:** Number of Cases by Perpetrator Race:

This bar chart shows the number of cases categorized by the race of the perpetrator. The chart indicates that Whites and Blacks are the most common races associated with the recorded crimes, with relatively similar numbers, followed by a significantly smaller number of Asians and an even smaller count labeled as "NA".

Pie Chart:

```
# Create a summary table for Weapon types
weapon_summary <- crime_data %>%
  count(Weapon) %>%
  mutate(Percentage = n / sum(n) * 100)

# Generate the pie chart
ggplot(weapon_summary, aes(x = "", y = Percentage, fill = Weapon)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  labs(title = "Distribution of Weapon Types Used in Crimes") +
  theme_void()
```

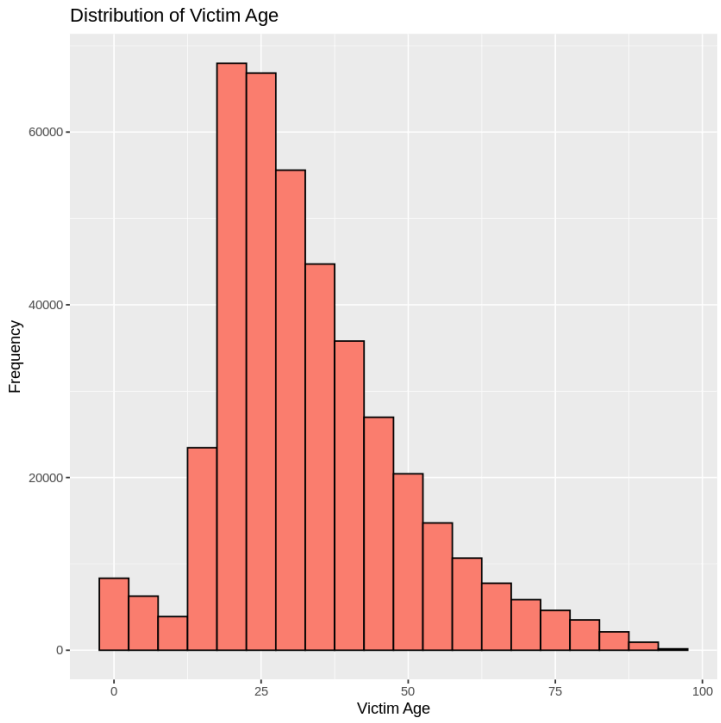


- **Observation:** Distribution of Weapon Types Used in Crimes:

The pie chart illustrates the proportions of various weapon types used in crimes. The most commonly used weapon is a gun, followed by a knife and then blunt objects. Other types like explosives, fire, and poison represent smaller fractions.

Histogram:

```
ggplot(crime_data, aes(x = Victim.Age)) +
  geom_histogram(binwidth = 5, fill = "salmon", color = "black") +
  labs(title = "Distribution of Victim Age", x = "Victim Age", y =
"Frequency")
```



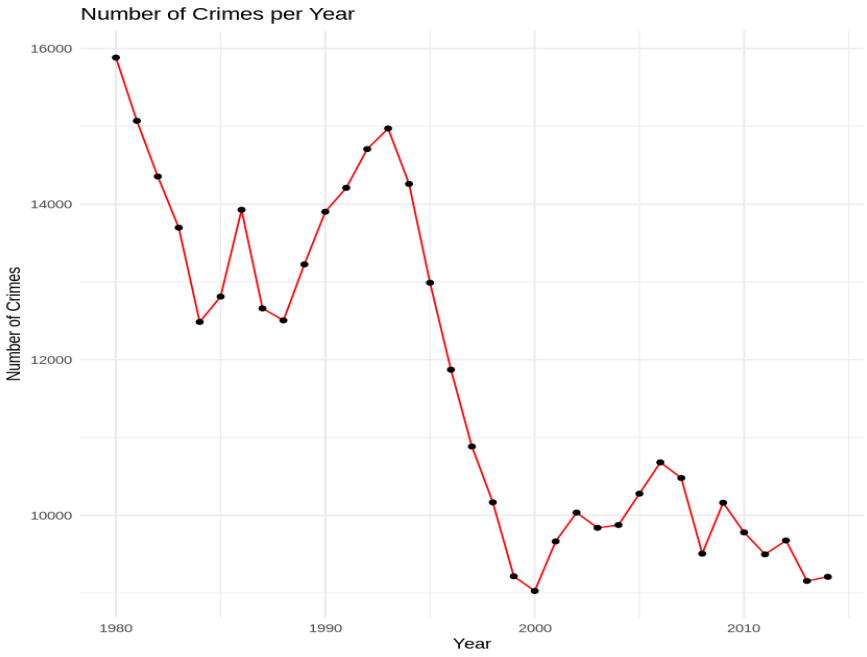
- **Observation:** Distribution of Victim Age:

This histogram shows the distribution of victim ages, which appears right-skewed, indicating that most victims are younger, with a peak around the age group of 20-25 years

Timeline Chart:

```
# Timeline Chart: Number of Crimes per Year
crime_by_year <- crime_data %>%
  group_by(Year) %>%
  summarise(crime_count = n())

ggplot(crime_by_year, aes(x = Year, y = crime_count)) +
  geom_line(color = "red") +
  geom_point() +
```

```
  labs(title = "Number of Crimes per Year", x = "Year", y = "Number of
Crimes") +
  theme_minimal()
```

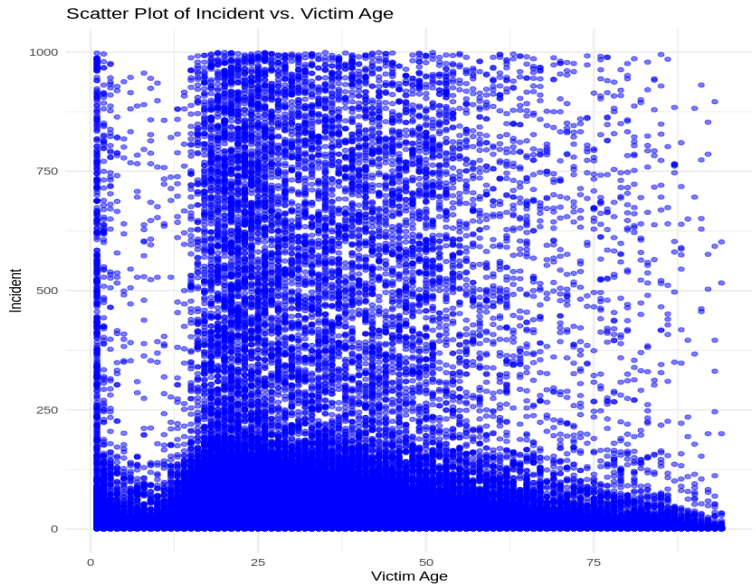


- **Observation:** Number of Crimes per Year:

The Time line chart tracks the number of crimes over years, showing fluctuations in crime rates over time with peaks and troughs. There is a notable peak around the year 1993, followed by a general decline with some variation.

Scatter Plot:

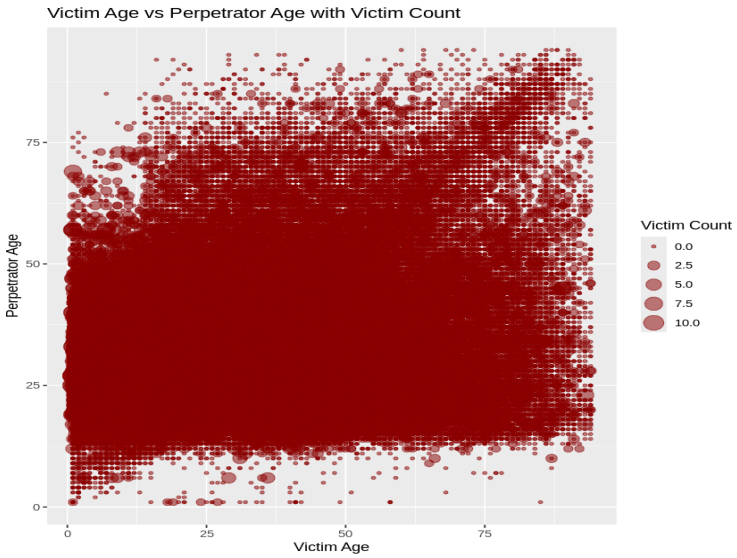
```
# Create a scatter plot of Incident vs. Victim Age
ggplot(crime_data, aes(y = Incident, x = Victim.Age)) +
  geom_point(alpha = 0.5, color = "blue") +  # Add points with
      transparency and color
  labs(title = "Scatter Plot of Incident vs. Victim Age",
      y = "Incident",
      x = "Victim Age") +
  theme_minimal()
```



Observation: The scatter plot shows that most incidents involve victims aged between 15 and 50, with a particularly dense concentration of data points in this age range. As the victim age increases beyond 50, the number of incidents decreases, with very few involving victims over 75. The distribution of incidents is fairly even across the range of incident numbers, suggesting that no specific incident count is disproportionately linked to a particular age group.

Bubble Plot:

```
ggplot(crime_data, aes(x = Victim.Age, y = Perpetrator.Age, size
= Victim.Count)) +
  geom_point(alpha = 0.5, color = "darkred") +
  labs(title = "Victim Age vs Perpetrator Age with Victim
Count", x = "Victim Age", y = "Perpetrator Age", size = "Victim
Count")
```



○ **Observation:** The bubble plot adds another dimension by showing the relationship between victim age and perpetrator age, with the size of each point representing the number of victims. Most incidents involve perpetrators aged between 15 and 50, with a similar age range for victims. The larger points are concentrated among younger age groups, indicating higher victim counts for younger victims and perpetrators. As both ages increase beyond 50, the frequency of incidents declines, and the victim counts are generally lower for older age groups.

Outcomes:

- **Created Multiple Chart Types**: Successfully generated various visualizations such as bar charts, scatter plots, pie charts, histograms, and bubble plots using R. Each chart helped to display different aspects of the crime data, allowing for a deeper understanding of the dataset.
- **Analyzed Crime Distribution**: The charts revealed significant patterns in the distribution of crime by factors such as victim age, perpetrator age, and crime type. For example, the victim age distribution showed that most crimes involved individuals between the ages of 15 and 50, while older individuals were less frequently victims.
- **Explored Relationships Between Variables**: Scatter plots and bubble plots effectively illustrated relationships between victim and perpetrator age, highlighting trends such as the correlation between younger victims and perpetrators. The use of bubble plots allowed us to examine incidents with multiple victims, further enriching the insights derived from the data.
- **Comparison of Solved and Unsolved Crimes**: The pie chart visualization allowed us to see the proportion of crimes solved versus unsolved, offering insights into law enforcement effectiveness and possible areas of improvement.
- **Developed Visualization Skills**: This experiment enhanced the understanding of how various chart types can be applied to analyze specific elements of a dataset. Each chart type provided a unique perspective, showing the power of visual representations in summarizing large datasets and making the information more accessible.

Conclusion:

This experiment demonstrated the immense value of data visualization in exploring and analyzing crime datasets. By utilizing R, we were able to efficiently create a variety of charts that offered different perspectives on the data, revealing patterns and trends that might have otherwise been overlooked. The visualizations enabled a comprehensive exploration of key aspects of the data, including the age distribution of victims and perpetrators, the frequency of different crime types, and the success rate of solved versus unsolved cases.Furthermore, the experiment highlighted how different chart types can be strategically used to address specific analytical questions. Bar charts provided a quick overview of crime type frequency, while scatter plots revealed relationships between variables. Bubble plots offered insights into victim counts, and pie charts clarified the breakdown of solved crimes.In conclusion, this exercise showcased the power of data visualization in transforming raw data into meaningful insights. By enabling the visualization of complex datasets, R empowers analysts to identify trends, correlations, and patterns that inform better decision-making and understanding. This experiment underscored the importance of selecting the right visualization for the data at hand, which can significantly impact the quality of the analysis and the conclusions drawn.