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

Name: Sakshi D. Lonare UID: 2021300069

Experiment no 4

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 Example

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.

Dataset:

https://www.kaggle.com/datasets/AnalyzeBoston/crimes-in-boston?select=crime.csv

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:

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

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

```
install.packages("ggplot2")
install.packages("dplyr")
library(ggplot2)
library(dplyr)
```

```
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)


Attaching package: 'dplyr'


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

Load the Dataset:

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

Data Preprocessing:

- Inspect and clean the data if necessary (handle missing values, filter relevant columns, etc.).

```
[2] df <- read.csv("crime.csv")
    df <- df %>% na.omit()

    str(df)
```

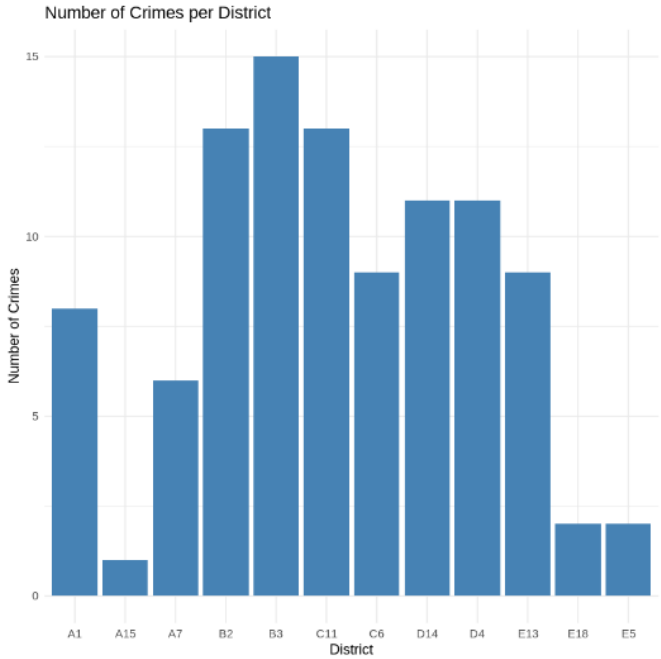
```
'data.frame':   296292 obs. of  17 variables:
 $ INCIDENT_NUMBER    : chr  "I182070945" "I182070943" "I182070941" "I1820
 $ OFFENSE_CODE       : int  619 1402 3410 3114 3114 3820 724 3301 301 330
 $ OFFENSE_CODE_GROUP : chr  "Larceny" "Vandalism" "Towed" "Investigate Pr
 $ OFFENSE_DESCRIPTION: chr  "LARCENY ALL OTHERS" "VANDALISM" "TOWED MOTOR
 $ DISTRICT           : chr  "D14" "C11" "D4" "D4" ...
 $ REPORTING_AREA     : int  808 347 151 272 421 398 330 584 177 364 ...
 $ SHOOTING           : chr  "" "" "" "" ...
 $ OCCURRED_ON_DATE   : chr  "2018-09-02 13:00:00" "2018-08-21 00:00:00" "
 $ YEAR               : int  2018 2018 2018 2018 2018 2018 2018 2018 2018
 $ MONTH              : int  9 8 9 9 9 9 9 9 9 9 ...
 $ DAY_OF_WEEK        : chr  "Sunday" "Tuesday" "Monday" "Monday"
```

```
[3]  #sample 100 entries from dataset for cleaner chart
     df_sample <- df %>%
       sample_n(100)
```

2. **Create Visualizations:**

Bar Chart:

```
# Bar chart showing the number of crimes per district
df_sample %>%
  group_by(DISTRICT) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = DISTRICT, y = count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Number of Crimes per District", x = "District", y = "Number of Crimes") +
  theme_minimal()
```

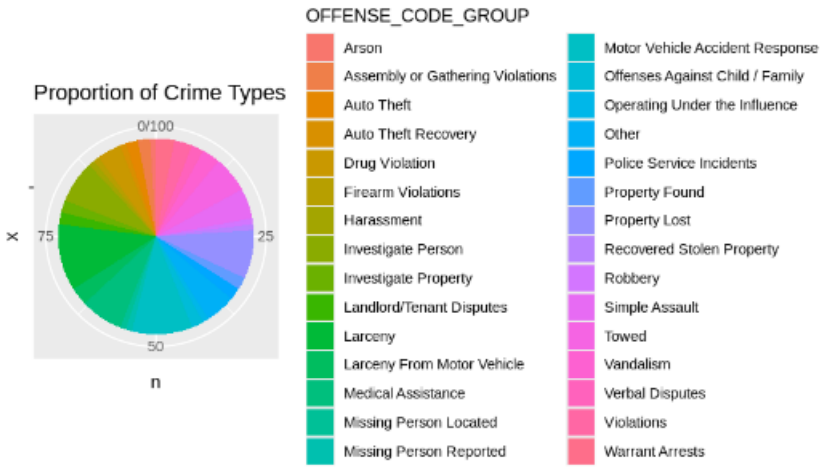


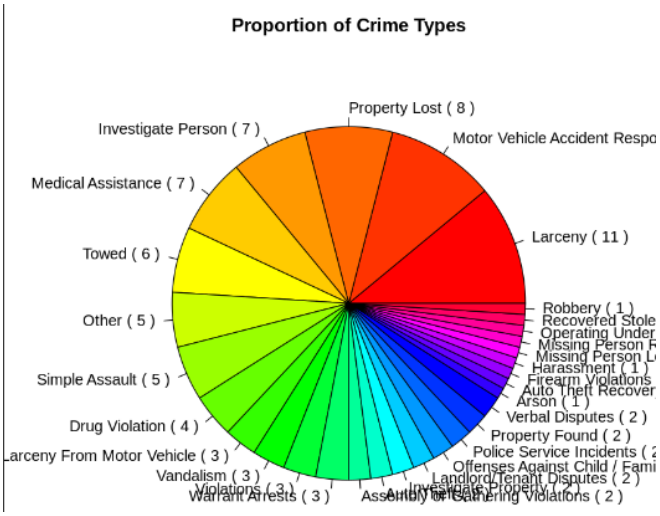
- **Observations:** The bar chart shows a significant variation in the number of crimes across different districts. Some districts, such as A1 and A15, have relatively low crime rates, while others, like B2 and B3, experience significantly higher numbers of crimes.

Pie Chart:

```
# Pie chart showing the proportion of crime types
```

```
df_sample %>%
  count(OFFENSE_CODE_GROUP) %>%
  ggplot(aes(x = "", y = n, fill = OFFENSE_CODE_GROUP)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Proportion of Crime Types")
```



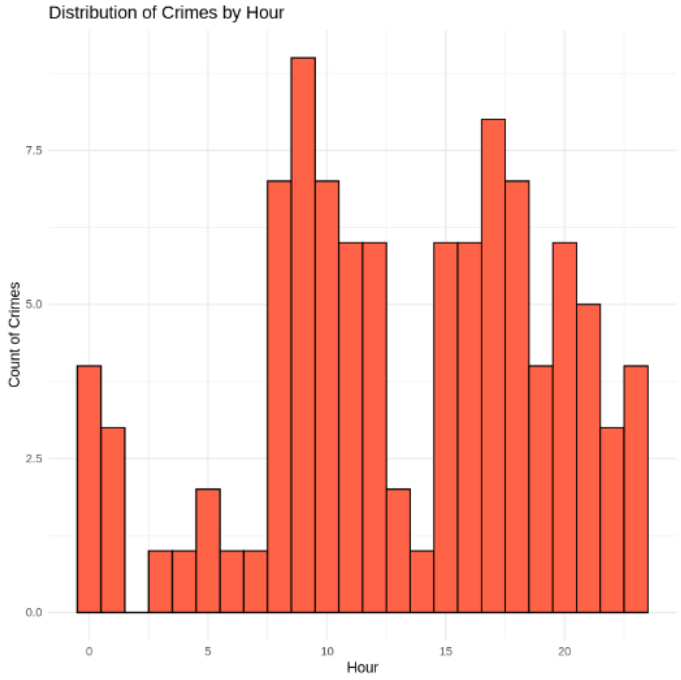


- **Observation:**

- The pie chart shows that a significant portion of the crime types are related to property, such as Larceny, Property Lost, Recovered Stolen Property, and Robbery.
- While property crimes are prevalent, violent crimes like Arson, Assault, and Drug Violations constitute a smaller proportion of the total crime types.

Histogram:

```
# Histogram showing the distribution of crimes by hour of the day
ggplot(df_sample, aes(x = HOUR)) +
  geom_histogram(binwidth = 1, fill = "tomato", color = "black") +
  labs(title = "Distribution of Crimes by Hour", x = "Hour", y = "Count of Crimes") +
  theme_minimal()
```

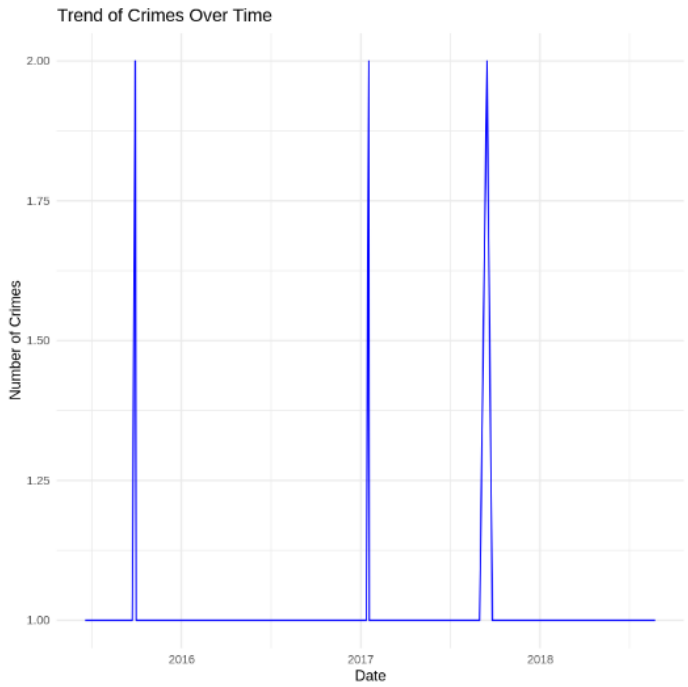


- **Observation:** The histogram shows a bimodal distribution. One of the peaks appears to be in the late evening hours (around 20-22 hours). This might indicate that a significant portion of crimes occur during this time, potentially due to factors such as reduced visibility, decreased police presence, or increased social activity..

Timeline Chart:

```
# Ensure the date column is in date format
df_sample$OCCURRED_ON_DATE <- as.Date(df_sample$OCCURRED_ON_DATE)

crime_trend <- df_sample %>%
  group_by(OCCURRED_ON_DATE) %>%
  summarise(Crime_Count = n())  # Count the number of crimes per day
ggplot(crime_trend, aes(x = OCCURRED_ON_DATE, y = Crime_Count)) +
  geom_line(color = "blue") +  # Line plot to show trend over time
  ggtitle("Trend of Crimes Over Time") +
  xlab("Date") +
  ylab("Number of Crimes") +
  theme_minimal()  # Use a clean, minimal theme
```

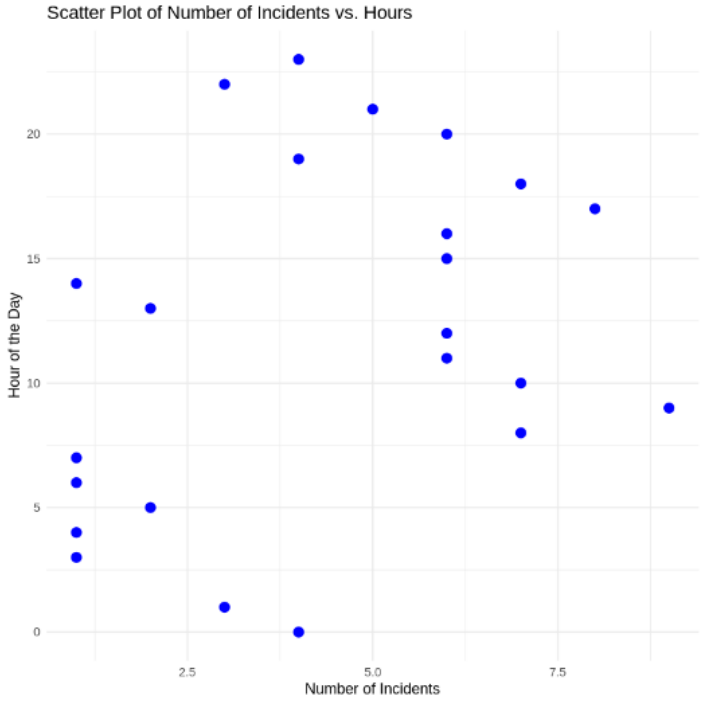


- **Observation:**

- The line chart shows dramatic spikes in the number of crimes during certain years. Notably, there are significant peaks in 2016 and 2017.
- While there are clear spikes, there doesn't seem to be a consistent upward or downward trend in crime rates over time.

Scatter Plot:

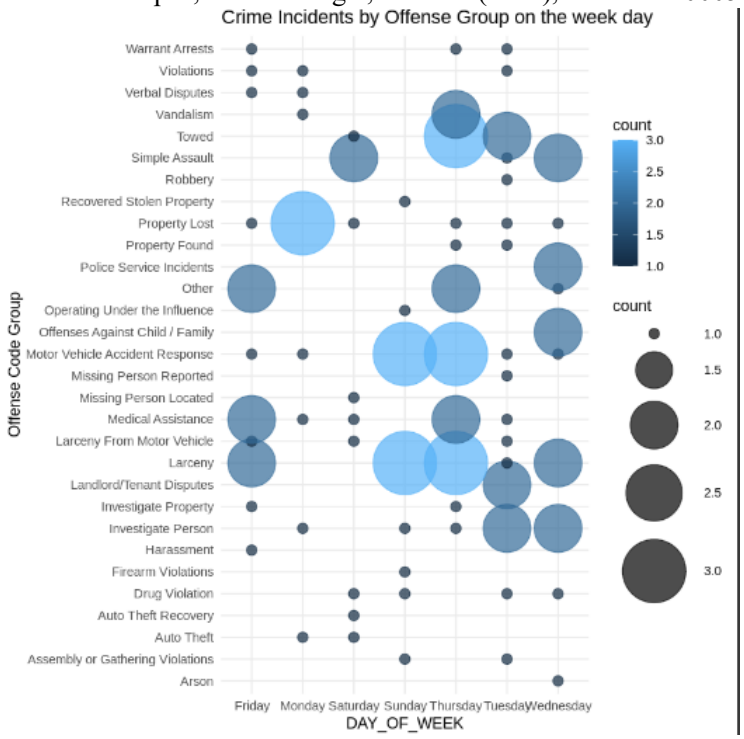
```
# Group the sampled data by hour and count the number of incidents
hourly_incidents <- df_sample %>%
  group_by(HOUR) %>%
  summarise(incident_count = n())  # Count the number of incidents per hour
ggplot(hourly_incidents, aes(x = incident_count, y = HOUR)) +
  geom_point(color = "blue", size = 3) +  # Scatter plot points
  labs(title = "Scatter Plot of Number of Incidents vs. Hours",
       x = "Number of Incidents",
       y = "Hour of the Day") +
  theme_minimal()
```



○ **Observation:** While there's no overall correlation, we can observe that there are clusters of data points around specific hours of the day. For instance, there seems to be a concentration of incidents between 10 and 15 hours, and another around 20 hours. These clusters might indicate patterns or trends related to specific types of incidents or factors influencing their occurrence at these times.

Bubble Plot:

```
# Bubble plot showing the number of crimes per offense group and district
df_sample %>%
  group_by(OFFENSE_CODE_GROUP, DAY_OF_WEEK) %>%
  summarise(count = n()) %>%
  ggplot(aes(x =  DAY_OF_WEEK, y = OFFENSE_CODE_GROUP, size = count, color = count)) +
  geom_point(alpha = 0.7) +
  scale_size(range = c(3, 20)) +
  labs(title = "Crime Incidents by Offense Group on the week day", x = "DAY_OF_WEEK", y = "Offense Code
Group") +
  theme_minimal()
```



- **Observation:**

- The bubble chart shows that the number of Warrant Arrests and Violations remains relatively constant across all days of the week.
- Crimes, such as Larceny, Property Lost, and Property Found, seem to be more prevalent on weekends (Friday, Saturday, and Sunday). The bubbles representing these categories are generally larger on these days, indicating higher incident counts compared to weekdays.

Outcomes:

- Created various charts in R to effectively visualize and analyze crime data of Boston.
- Derived key insights into the distribution, frequency, and correlations within the dataset.
- Developed proficiency in using different chart types to explore and present data comprehensively.

Conclusion:

From this experiment Ive showcased the effectiveness of data visualization in revealing patterns and trends within a crime dataset of Boston. Using R, we efficiently generated visual representations that enabled us to examine the data from multiple angles, leading to more informed insights and conclusions.