



**Sardar Patel Institute of Technology, Mumbai**  
**Department of Electronics and Telecommunication Engineering**  
**B.E. Sem-VII- PE-IV (2024-2025)**

**Name :- Manthan Ayalwar**

**Batch :- ADV I**

**Uid :- 2021700003**

**Branch :- Cse-Ds**

## **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**

**Database:**

<https://www.kaggle.com/datasets/adoumtaiga/crime-data-set>

**R Script :-**

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

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

```
library(ggplot2)
```

```
library(dplyr)
```

```
# Check for missing values
```

```
summary(Crime_Data)
```

```
Crime_Data$Occurred.Date <- as.Date(Crime_Data$Occurred.Date, format = "%m/%d/%Y")
```

```
Crime_Data$Reported.Date <- as.Date(Crime_Data$Reported.Date, format = "%m/%d/%Y")
```

```
# Bar Chart
```

```
# Summarize and sort the data to get the top 10 categories
```

```
top_10_crime <- Crime_Data %>%
```

```
  group_by(Crime.Subcategory) %>%
```

```
  summarise(Count = n()) %>%
```

```
  arrange(desc(Count)) %>%
```

```
  slice_head(n = 10) # Select the top 10 categories
```

```
# Create a bar chart for the top 10 categories
ggplot(top_10_crime, aes(x = reorder(Crime.Subcategory, -Count), y = Count)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  theme_minimal(base_size = 15) +
  labs(title = "Top 10 Crime Subcategories", x = "Crime Subcategory", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotate x-axis labels for readability
```

#### #Pie Chart

```
# Create a frequency table for the Precinct column
pie_data_precinct <- table(Crime_Data$Precinct)
```

```
# Create a pie chart for Precincts
pie(pie_data_precinct,
    main = "Pie Chart of Precincts",
    col = rainbow(length(pie_data_precinct)))
```

```
# Optional: Add percentages to the pie chart for better clarity
percent_labels <- round(100 * pie_data_precinct / sum(pie_data_precinct), 1)
labels <- paste(names(pie_data_precinct), "(", percent_labels, "%)", sep="")
pie(pie_data_precinct, labels = labels, main = "Pie Chart of Precincts", col =
rainbow(length(pie_data_precinct)))
```

#### #Histogram

```
ggplot(Crime_Data, aes(x = Reported.Time)) +
  geom_histogram(binwidth = 100, fill = "orange", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Reported Times", x = "Reported Time", y = "Frequency")
```

```
ggplot(Crime_Data, aes(x = Occurred.Time)) +
  geom_histogram(binwidth = 100, fill = "yellow", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Occurred Times", x = "Occurred Time", y = "Frequency")
```

#### # Time-Line Chart

```
ggplot(Crime_Data, aes(x = Occurred.Date)) +
  geom_histogram(binwidth = 365, fill = "purple", color = "black") +
  theme_minimal(base_size = 15) +
  labs(title = "Timeline of Crime Occurrences", x = "Occurred Date", y = "Count") +
  scale_x_date(limits = as.Date(c("2006-01-01", "2020-01-01")),
```

```
date_breaks = "5 years",  
date_labels = "%Y") # Setting breaks and labels on the x-axis
```

```
# Scatter PLOT
```

```
# Filter data between 2006 and 2020
```

```
filtered_data <- Crime_Data %>%  
  filter(Occurred.Date >= as.Date("2006-01-01") & Occurred.Date <= as.Date("2020-12-31"))
```

```
# Plot the filtered data
```

```
ggplot(filtered_data, aes(x = Occurred.Date, y = Reported.Date)) +  
  geom_point(color = "darkgreen") +  
  theme_minimal(base_size = 15) +  
  labs(title = "Scatter Plot of Occurred Date vs Reported Date (2006-2020)",  
       x = "Occurred Date", y = "Reported Date")
```

```
# Bubble Plot
```

```
# Create a frequency table to count the number of occurrences for each crime subcategory
```

```
crime_count <- Crime_Data %>%  
  group_by(Crime.Subcategory) %>%  
  summarise(count = n())
```

```
# Join the frequency count back to the original data
```

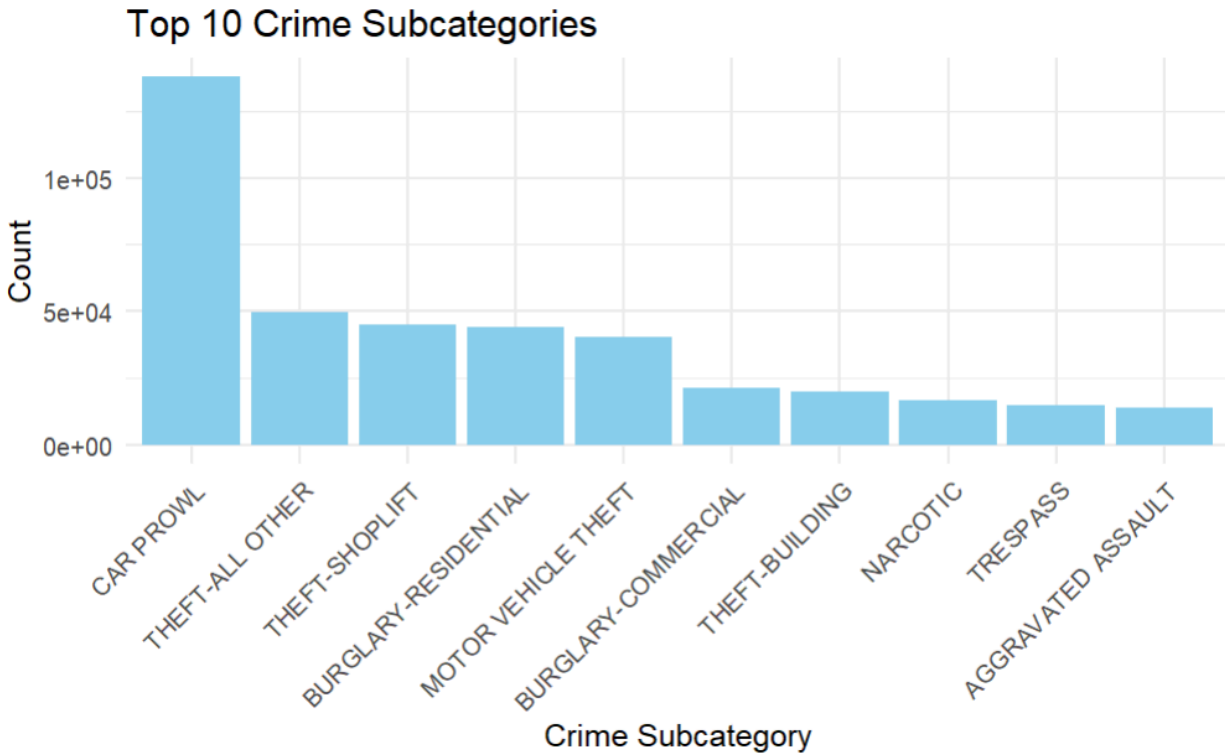
```
Crime_Data <- Crime_Data %>%  
  left_join(crime_count, by = "Crime.Subcategory")
```

```
# Plot the bubble plot with count as the size
```

```
ggplot(Crime_Data, aes(x = Occurred.Date, y = Reported.Date, size = count, color = Precinct))  
+  
  geom_point(alpha = 0.6) +  
  theme_minimal(base_size = 15) +  
  labs(title = "Bubble Plot of Occurred Date vs Reported Date",  
       x = "Occurred Date", y = "Reported Date", size = "Crime Count") +  
  scale_size_continuous(range = c(3, 10)) # Adjust the size range for bubbles
```

## **Visualization -**

### **Bar Chart -**

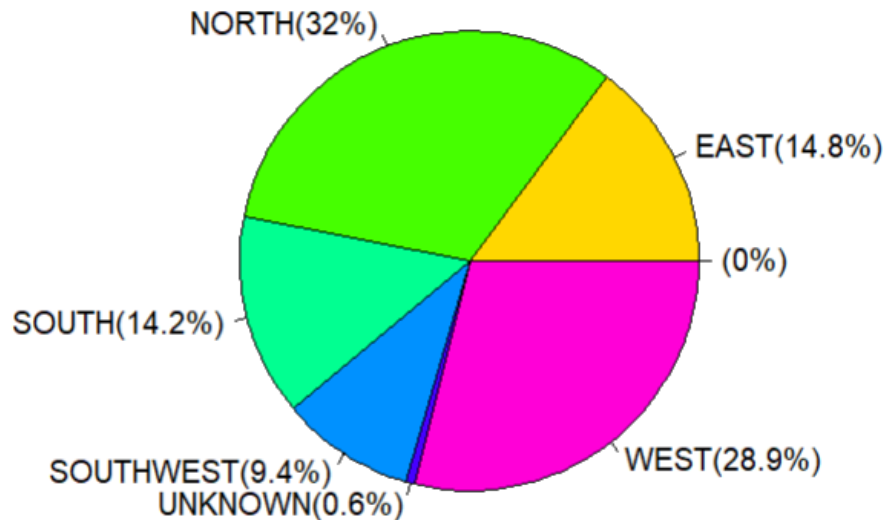


#### Observation :-

- Car prowl is by far the most common crime subcategory, with over 150,000 incidents recorded. This stands out significantly compared to other categories.
- The next four most common subcategories (theft-all other, theft-shoplift, burglary-residential, and motor vehicle theft) all have similar frequencies, ranging between approximately 40,000 to 50,000 incidents each.
- There's a noticeable drop in frequency after the top 5 categories. The remaining categories (burglary-commercial, theft-building, narcotic, trespass, and aggravated assault) have much lower incident counts.
- Property crimes dominate the top of the list. The top 7 categories all involve theft or burglary of some kind.
- Violent crime (aggravated assault) appears only at the bottom of this top 10 list, suggesting it's less frequent than property crimes in this dataset.
- Trespass and narcotic offenses are the only non-theft related crimes in the middle of the list.
- The distribution of crime types is quite uneven, with a large gap between the most common (car prowl) and the least common (aggravated assault) in this top 10 list.
- The y-axis scale suggests that even the least frequent crime in this top 10 list (aggravated assault) still occurs thousands of times.

#### Pie Chart -

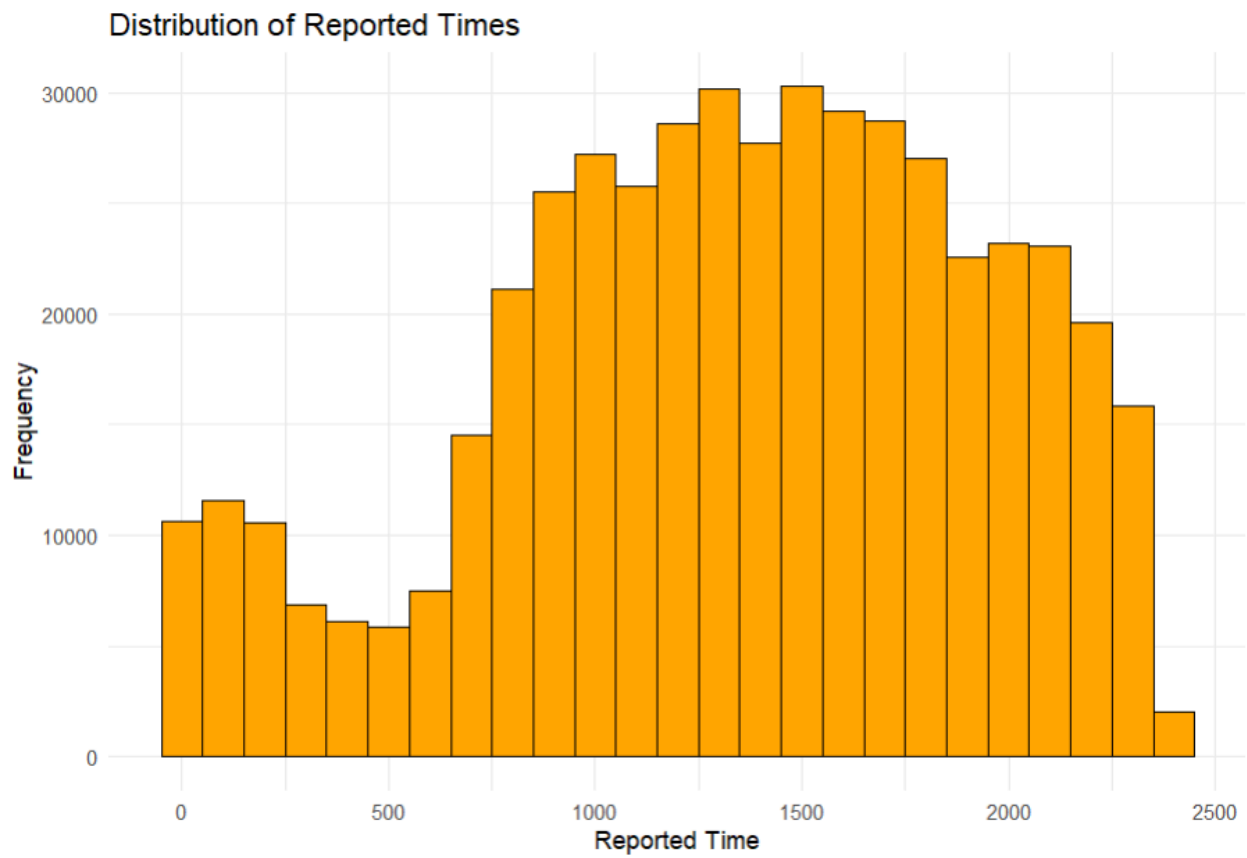
## Pie Chart of Precincts

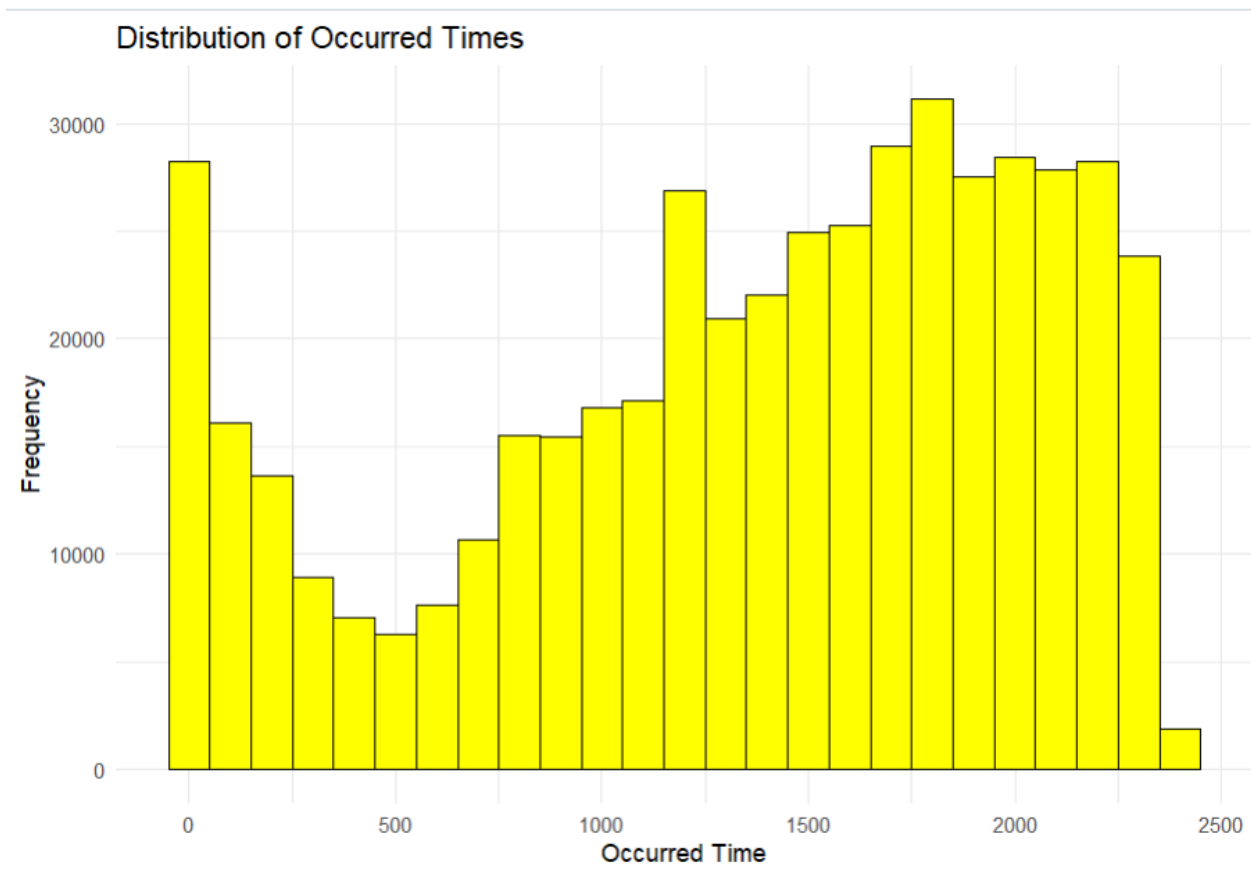


### Observation :-

- Largest precinct: The North precinct accounts for the largest portion of the data, representing 32% of all incidents.
- Second largest: The West precinct is the second most represented, with 28.9% of incidents.
- Similar mid-range precincts: The East and South precincts have similar representations, with 14.8% and 14.2% respectively.
- Smaller precinct: The Southwest precinct accounts for a smaller portion, at 9.4% of incidents.
- Minimal unknown data: There's a very small percentage (0.6%) of incidents with an unknown precinct.
- Unexplained slice: There's a 0% slice in the chart, which may be a visualization error or represent an extremely small category.
- Coverage distribution: The North and West precincts combined account for over 60% of all incidents, suggesting a potentially higher concentration of reported crimes or policing activity in these areas.
- Geographic insights: Without knowing the exact geography, this distribution suggests that crime reports or police activity are not evenly spread across the city/region, with some areas seeing significantly more incidents than others.
- Data completeness: The very low percentage of unknown precincts (0.6%) indicates good data quality in terms of location recording.

### Histogram -





#### Observation :-

##### Reported Times :

- The distribution is unimodal, peaking around 1500 (3:00 PM).
- There's a gradual increase from about 800 (8:00 AM) to the peak.
- After the peak, there's a gradual decline until about 2300 (11:00 PM).
- Very few crimes are reported between midnight and 6:00 AM.
- The distribution is roughly bell-shaped, suggesting most crime reports happen during daytime and early evening hours.

##### Occurred Times :

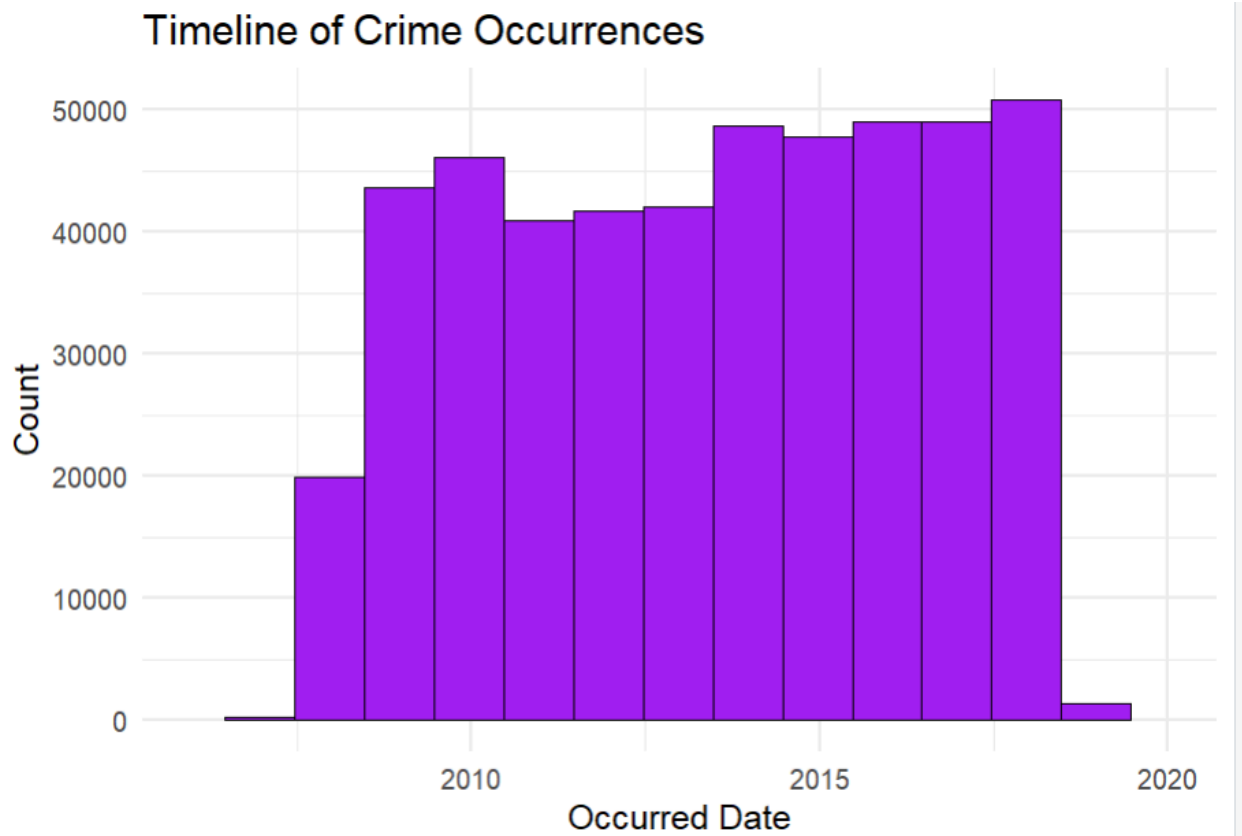
- This distribution is bimodal, with two distinct peaks.
- The first major peak is at 0 (midnight), suggesting many crimes occur or are discovered then.
- There's a second, smaller peak around 1800 (6:00 PM).
- The frequency is generally higher from noon to midnight compared to early morning hours.
- There's a noticeable dip in occurrences around 500-700 (5:00-7:00 AM).

##### Comparing the two:

- The occurred times show more variability and distinct patterns compared to reported times.

- There's a significant mismatch between when crimes occur (often at night) and when they're reported (mostly during the day).
- The high frequency of occurrences at midnight in Image 2 isn't reflected in the reported times, suggesting a delay in reporting.
- The daytime peak in reported crimes doesn't correspond to a similar peak in occurred crimes, indicating that many nighttime crimes are likely reported the next day.

#### Time Line -



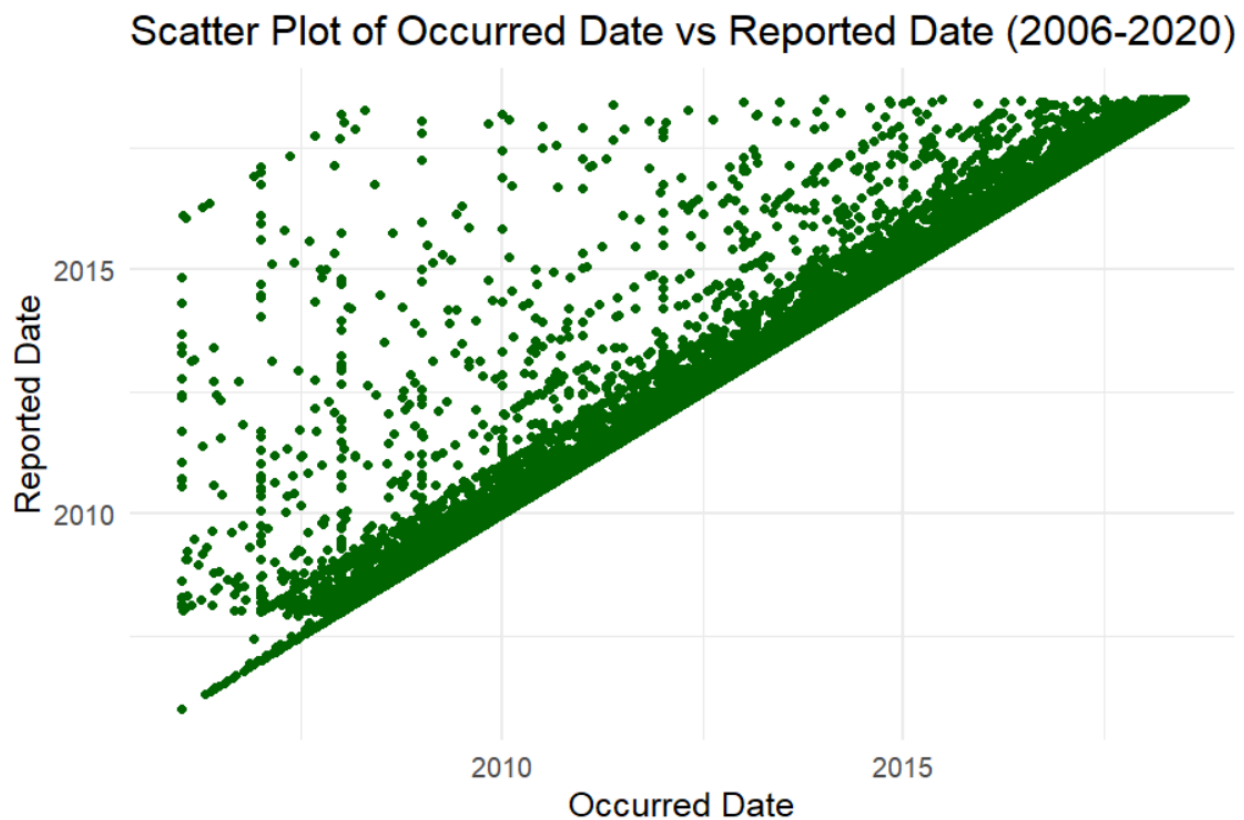
#### Observation :-

- Time range: The graph covers crime data from approximately 2008 to 2020.
- Overall trend: There's a general upward trend in crime occurrences over the years, with some fluctuations.
- Initial spike: There's a sharp increase in crime occurrences from 2008 to 2010, jumping from very low numbers to around 45,000 annually.
- Plateau and slight decline: From 2010 to 2013, there's a relatively stable period with a slight decline in crime occurrences.
- Steady increase: Starting from around 2013, there's a consistent upward trend in crime occurrences until 2019.
- Peak: The highest number of crime occurrences appears to be in 2019, reaching slightly over 50,000 incidents.



- Recent drop: There's a sharp decline in 2020, likely only representing partial data for that year or possibly influenced by external factors (e.g., COVID-19 pandemic).
- Data completeness: The very low numbers before 2008 suggest that the dataset might not have complete records for earlier years.
- Yearly variations: While there's an overall increasing trend, there are noticeable year-to-year variations throughout the timeline.
- Consistent reporting: The relatively smooth progression of the graph suggests consistent crime reporting practices over the years, with no major gaps or anomalies (except for the beginning and end of the timeline).

#### Scatter Plot -

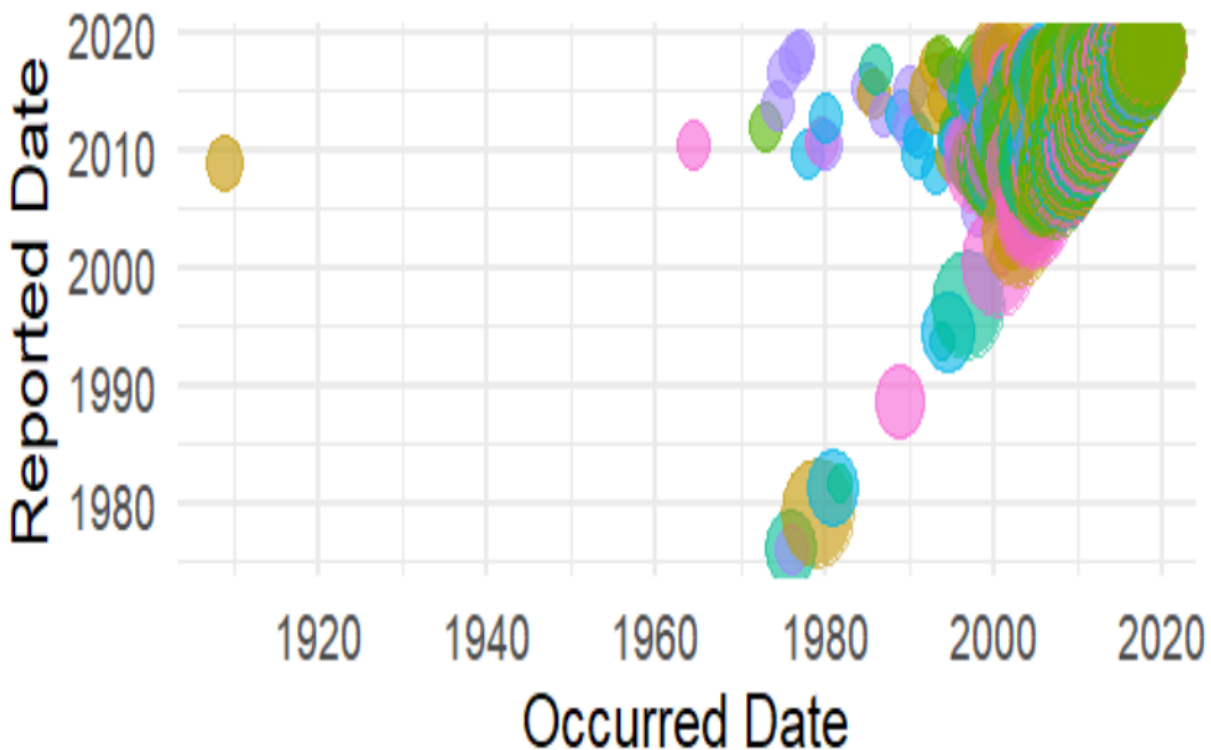


#### Observation :-

- Strong correlation: There's a clear positive correlation between occurred and reported dates, shown by the dense line along the diagonal. This indicates that most crimes are reported on or soon after they occur.
- Timely reporting: The majority of points fall on or very close to the diagonal line, suggesting that many crimes are reported on the same day they occur.
- Delayed reporting: There's a significant number of points above the diagonal line, representing crimes reported after their occurrence date. This spread increases for older occurrences, indicating longer reporting delays for some past events.

- Reporting lag: The vertical spread of points above the diagonal line shows varying delays in reporting, from a few days to several years in some cases.
- No points below diagonal: As expected, there are no points below the diagonal, which would represent impossible cases of crimes being reported before they occurred.
- Long-term delayed reporting: Some points appear far above the diagonal, indicating crimes reported years after they occurred. This could represent cold cases or delayed discoveries of crimes.
- Density variation: The density of points is highest along the diagonal and decreases as you move away, showing that immediate or near-immediate reporting is most common.
- Time range: The plot covers data from 2006 to 2020, providing a comprehensive view of reporting patterns over a 14-year period.
- Consistent pattern: The overall pattern remains relatively consistent across the time range, suggesting stable reporting behaviors over the years.

**Bubble Plot -**



**Observation :-**

- Wide time range: The plot covers a surprisingly large timespan, from around 1920 to 2020 for occurred dates, suggesting some very old cases are included.
- Diagonal concentration: The majority of data points fall along or near the diagonal line, indicating that most crimes are reported close to when they occurred.

- Vertical clusters: There are several distinct vertical clusters of points, particularly noticeable around 1980 and 2000 on the x-axis. This suggests batches of crimes being reported at the same time, possibly due to administrative processes or discovery of historical cases.
- Historical reporting: Some crimes that occurred decades ago (as far back as the 1920s) were reported much more recently, shown by points in the upper-left quadrant of the plot.
- Bubble size variation: The varying sizes of the bubbles indicate different frequencies of crime types, with larger bubbles representing more common crime subcategories.
- Color distribution: The mix of colors throughout the plot suggests that the patterns of crime occurrence and reporting are generally consistent across different precincts.
- Recent density: The density of points increases significantly for more recent years (post-2000), likely due to better record-keeping and more immediate reporting in recent times.

**Conclusion :-** From this experiment, I learned about R language and how to use r studio and how to import dataset and plot visualization in R studio.