ADITYA LALWANI D15C ROLL NO.30

# Experiment 2

**Aim**: Data Visualization/ Exploratory data Analysis using Matplotlib and Seaborn.
Perform following data visualization and exploration on your selected dataset:-

- Create bar graph, contingency table using any 2 features.
- Plot Scatter plot, box plot, Heatmap using seaborn.
- Create histogram and normalized Histogram. ● Describe what this graph and table indicates.
- Handle outlier using box plot and Inter quartile range.

## Performance:

- Prerequisite: Import all the required libraries (pandas for data manipulation, numpy for numerical computations, and data visualization using matplotlib for basic plotting and seaborn for enhanced statistical graphics) and load data into Pandas:

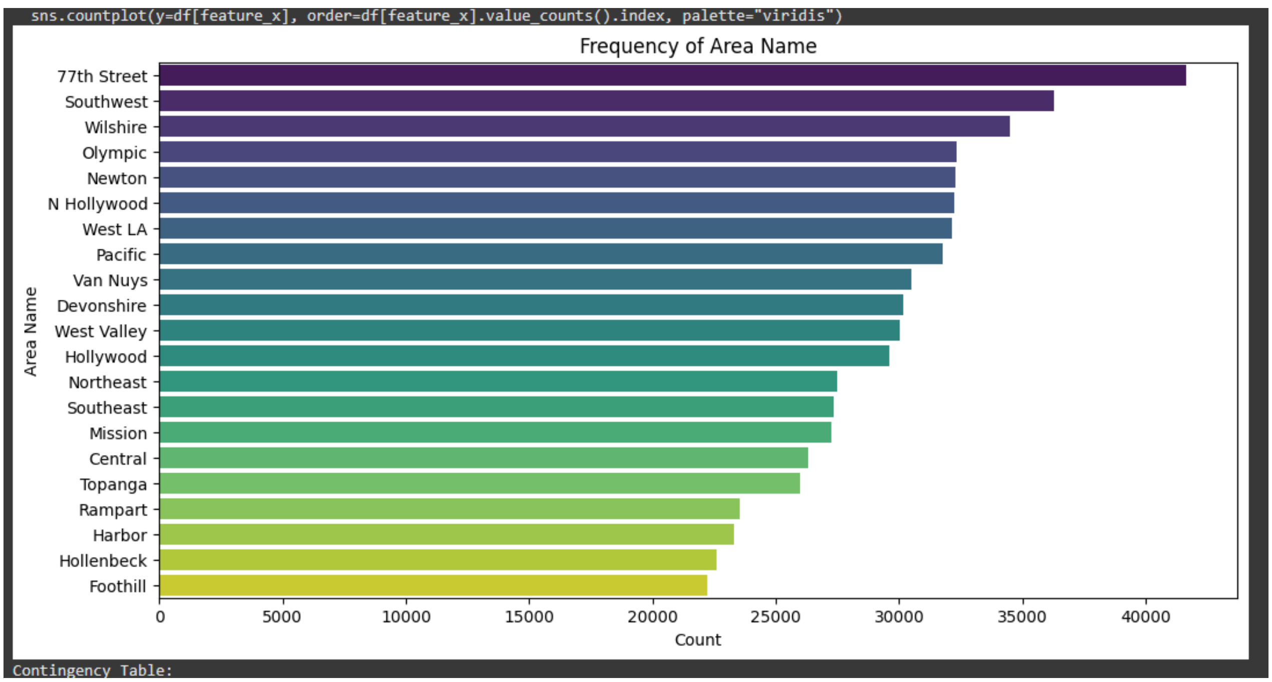
Command:
```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
df = pd.read_csv('Dataset.csv')
df.head()
```

```
df.head()
```

| | DR Number | Date Reported | Date Occurred | Time Occurred | Area ID | Area Name | Reporting District | Crime Code | Crime Code Description | MO Codes | Victim Age | Victim Sex | Victim Descent | Premise Code | Premise Description | Address | Cross Street | Location |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 190319651 | 08/24/2019 | 08/24/2019 | 450 | 3 | Southwest | 356 | 997 | TRAFFIC COLLISION | 3036 3004 3026 3101 4003 | 22.0 | M | H | 101.0 | STREET | JEFFERSON BL | NORMANDIE AV | (34.0255, -118.3002) |
| 1 | 190319680 | 08/30/2019 | 08/30/2019 | 2320 | 3 | Southwest | 355 | 997 | TRAFFIC COLLISION | 3037 3006 3028 3030 3039 3101 4003 | 30.0 | F | H | 101.0 | STREET | JEFFERSON BL | W WESTERN | (34.0256, -118.3089) |
| 2 | 190413769 | 08/25/2019 | 08/25/2019 | 545 | 4 | Hollenbeck | 422 | 997 | TRAFFIC COLLISION | 3101 3401 3701 3006 3030 | NaN | M | X | 101.0 | STREET | N BROADWAY | W EASTLAKE AV | (34.0738, -118.2078) |
| 3 | 190127578 | 11/20/2019 | 11/20/2019 | 350 | 1 | Central | 128 | 997 | TRAFFIC COLLISION | 0605 3101 3401 3701 3011 3034 | 21.0 | M | H | 101.0 | STREET | 1ST | CENTRAL | (34.0492, -118.2391) |
| 4 | 190319695 | 08/30/2019 | 08/30/2019 | 2100 | 3 | Southwest | 374 | 997 | TRAFFIC COLLISION | 0605 4025 3037 3004 3025 3101 | 49.0 | M | B | 101.0 | STREET | MARTIN LUTHER KING JR | ARLINGTON AV | (34.0108, -118.3182) |

- Create bar graph, contingency table using any 2 features:

Command:
feature_x = "Area Name"
feature_y = "Crime Code Description"
plt.figure(figsize=(12, 6))
sns.countplot(y=df[feature_x], order=df[feature_x].value_counts().index,
palette="viridis")
plt.title(f"Frequency of {feature_x}")
plt.xlabel("Count")
plt.ylabel(feature_x)
plt.show()
contingency_table = pd.crosstab(df[feature_x], df[feature_y])
print("Contingency Table:")
print(contingency_table)




The bar graph shows how many car crashes happen in different areas. The bottom line (x-axis) shows the number of crashes, and the side line (y-axis) lists the areas. 77th Street has the most crashes, then Southwest and Wilshire. This means some areas have a lot more crashes, possibly because they have more traffic, dangerous roads, or other reasons.

ADITYA LALWANI D15C ROLL NO.30

Contingency Table:

| Crime Code Description / Area Name | TRAFFIC COLLISION |
|---|---|
| 77th Street | 41631 |
| Central | 26309 |
| Devonshire | 30191 |
| Foothill | 22215 |
| Harbor | 23307 |
| Hollenbeck | 22594 |
| Hollywood | 29601 |
| Mission | 27235 |
| N Hollywood | 32259 |
| Newton | 32282 |
| Northeast | 27508 |
| Olympic | 32316 |
| Pacific | 31787 |
| Rampart | 23541 |
| Southeast | 27351 |
| Southwest | 36285 |
| Topanga | 25979 |
| Van Nuys | 30518 |
| West LA | 32129 |
| West Valley | 30047 |
| Wilshire | 34510 |

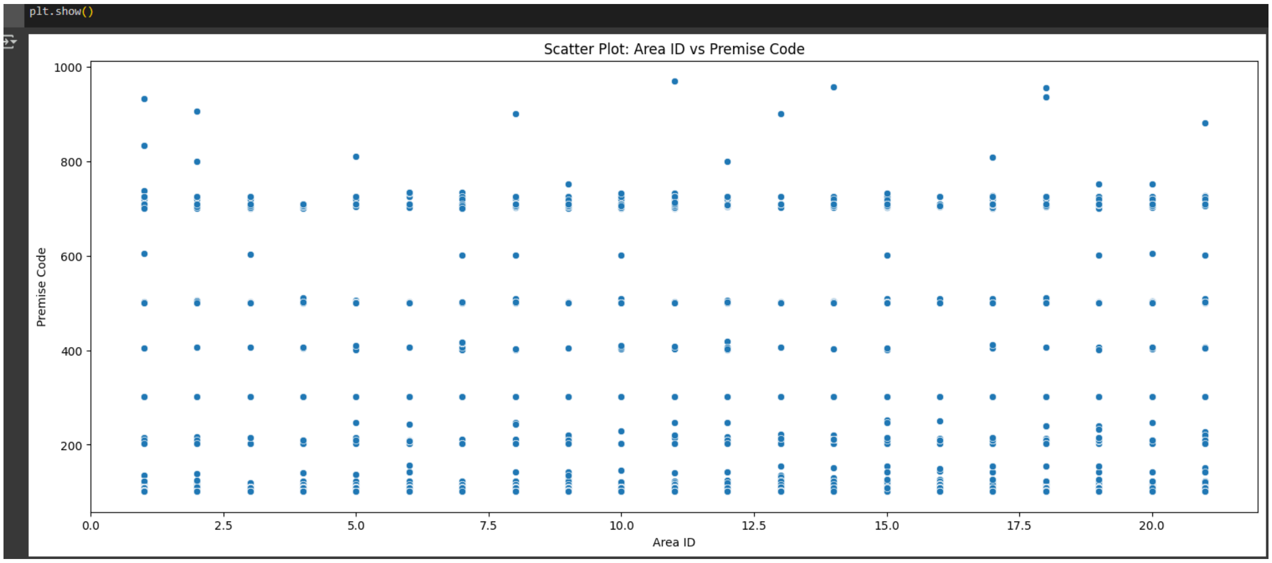
The table shows car crashes in different areas. Each row is an area, and the number shows how many crashes happened there. 77th Street has the most crashes, then Wilshire and Southwest. Areas like Foothill, Harbor, and Rampart have fewer crashes. This means some areas have more traffic, worse roads, or different reporting.

ADITYA LALWANI D15C ROLL NO.30

- Plot Scatter plot, box plot, Heatmap using seaborn:

1. Scatter plot:-

Command:
plt.figure(figsize=(18, 7))
sns.scatterplot(x=df["Area ID"], y=df["Premise Code"])
plt.title("Scatter Plot: Area ID vs Premise Code")
plt.xlabel("Area ID") plt.ylabel("Premise Code")
plt.show()



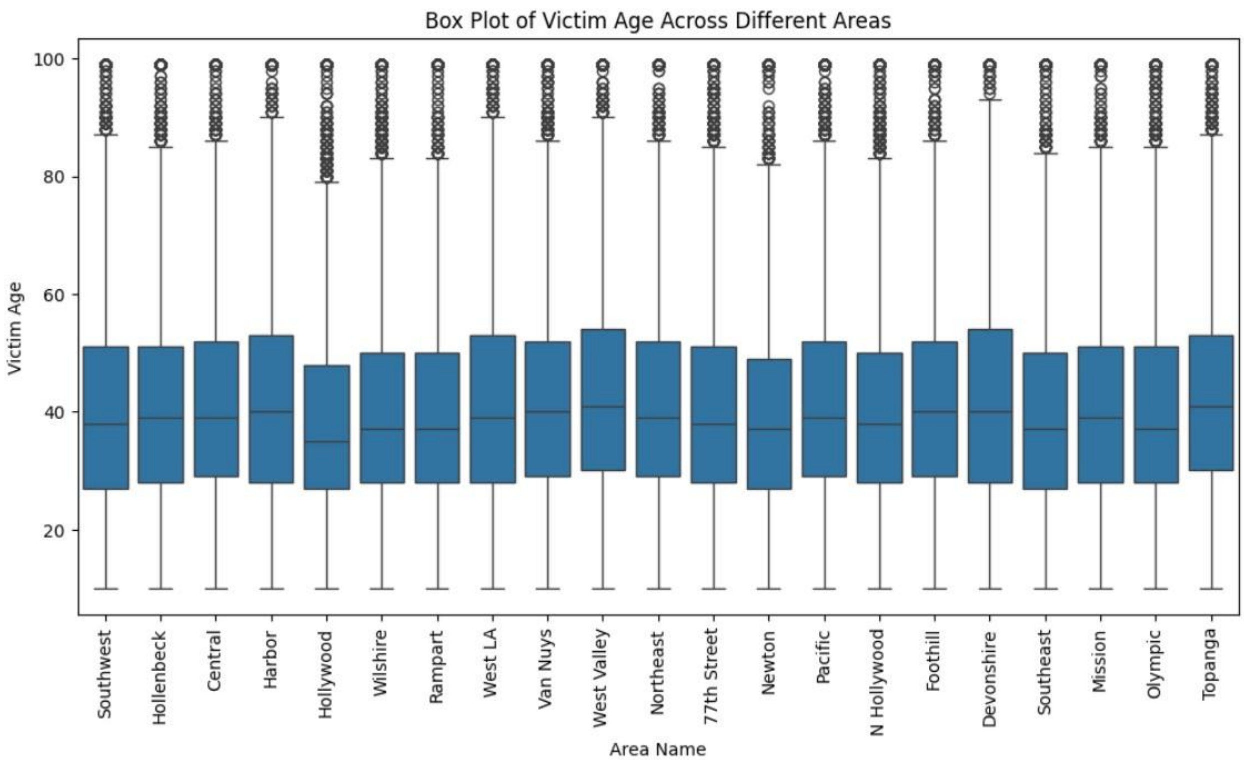

The scatter plot shows how Area ID and Premise Code relate in crash data. The x-axis is Area ID, and the y-axis is Premise Code, which shows where crashes happened. The dots mean crashes occur in many places across all areas, with some areas having more crashes at certain locations. A few dots are far from the rest, meaning those places have way more or fewer crashes.

ADITYA LALWANI D15C ROLL NO.30

2. Box Plot:-

Command:

```
plt.figure(figsize=(12, 6))
sns.boxplot(x=df["Area Name"], y=df["Victim Age"])
plt.xticks(rotation=90)
plt.title("Box Plot of Victim Age Across Different Areas")
plt.xlabel("Area Name")
plt.ylabel("Victim Age")
plt.show()
```

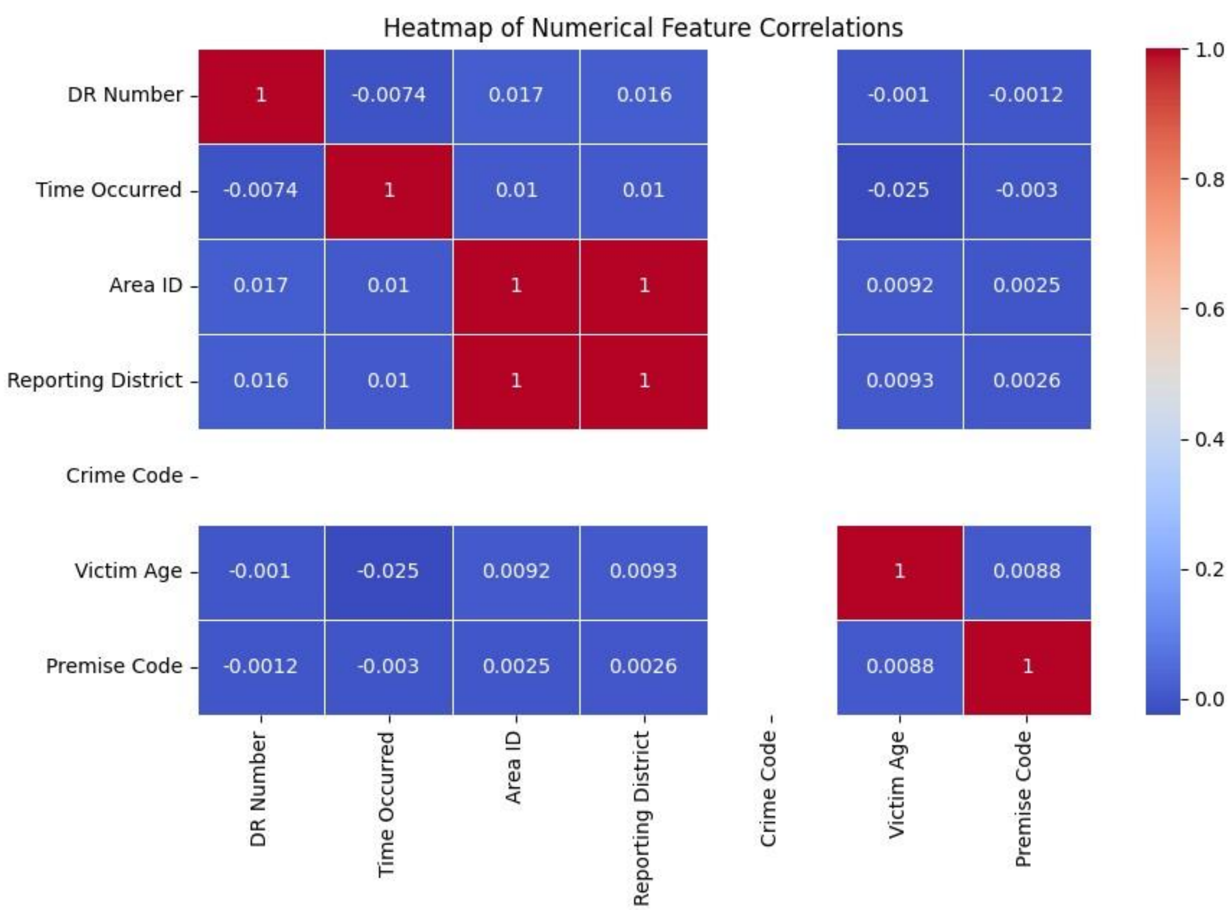



The box plot shows the ages of victims in different areas. Most victims are around 35-45 years old, with ages typically ranging from 25 to 55. Some victims are much younger or older, with a few over 80 years old. The age patterns are similar in all areas.

3. Heatmap:

Command:

```
plt.figure(figsize=(10, 6))
sns.heatmap(df.select_dtypes(include=np.number).corr(), annot=True, cmap="coolwarm",
linewidths=0.5)
plt.title("Heatmap of Numerical Feature Correlations")
plt.show()
```



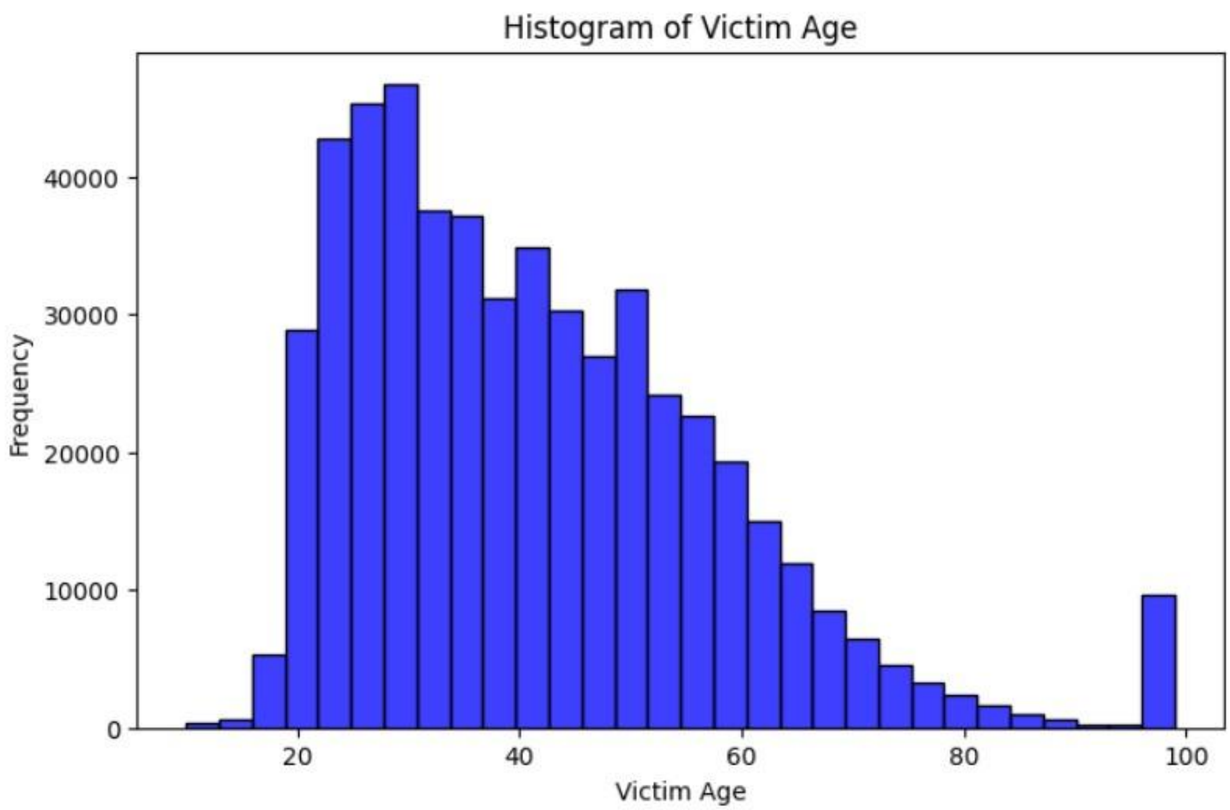

The heatmap shows how number-based features are related, with values from -1 to 1. Area ID and Reporting District are strongly linked (value = 1). Most other features have weak or no connections. Victim Age doesn't affect when or where crashes happen. DR Number and Crime Code don't relate to other features and act as unique IDs. Overall, most features aren't strongly connected, except for geographical ones.

- Create histogram and normalized Histogram:-

1. Histogram:

Command:

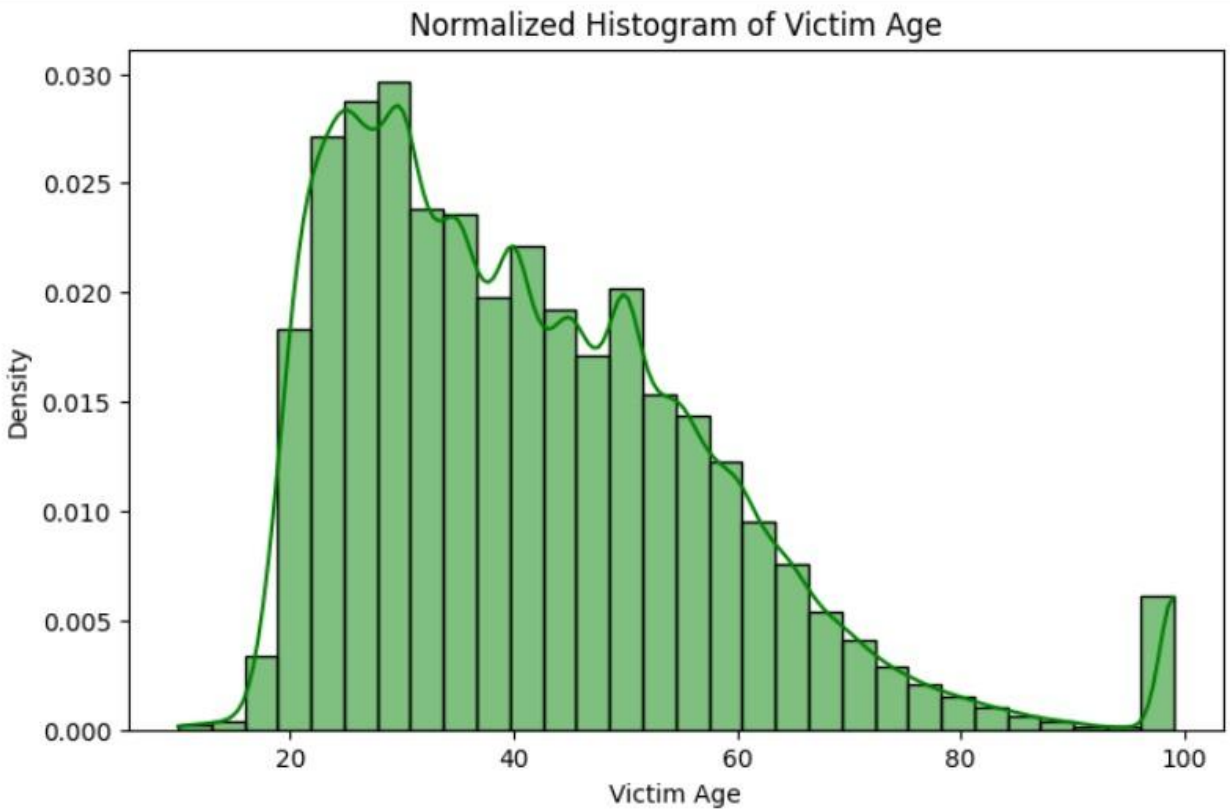
```
plt.figure(figsize=(8, 5))
sns.histplot(df["Victim Age"], bins=30, kde=False, color="blue")
plt.title("Histogram of Victim Age")
plt.xlabel("Victim Age")
plt.ylabel("Frequency")
plt.show()
```



2. Normalized Histogram:

Command:

```
plt.figure(figsize=(8, 5))
sns.histplot(df["Victim Age"], bins=30, kde=True, color="green", stat="density")
plt.title("Normalized Histogram of Victim Age")
plt.xlabel("Victim Age")
plt.ylabel("Density")
```

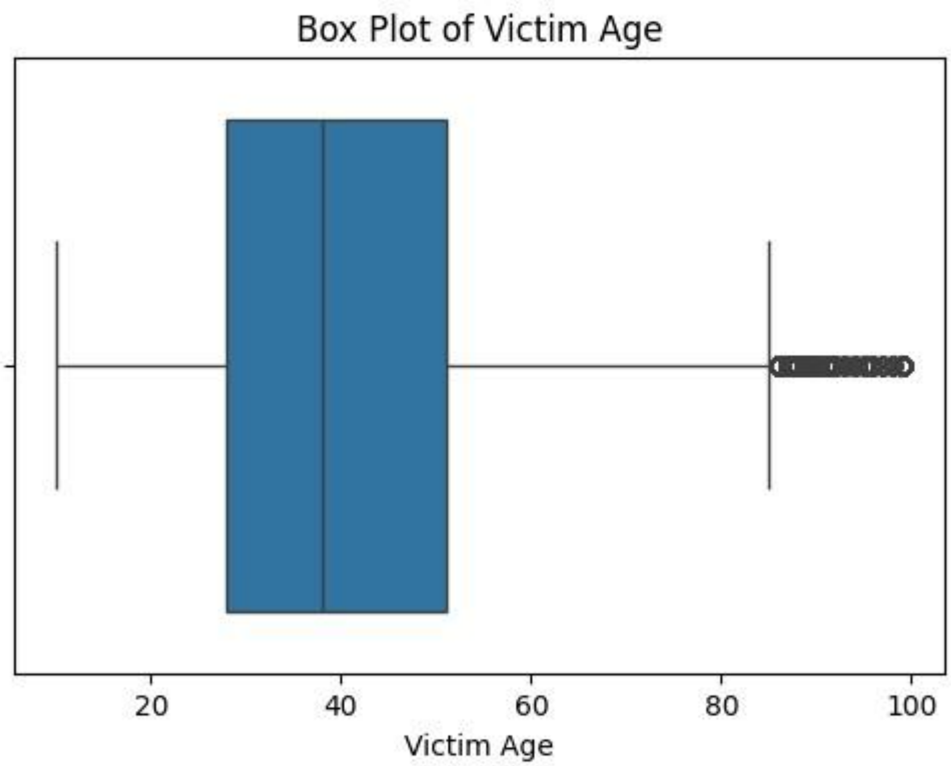


- Handle outlier using box plot and Inter quartile range:

1. Using box plot:-

Command:

```
plt.figure(figsize=(6, 4))
sns.boxplot(x=df["Victim Age"])
plt.title("Box Plot of Victim Age")
plt.show()
```



2. Using Interquartile range:-

Command:

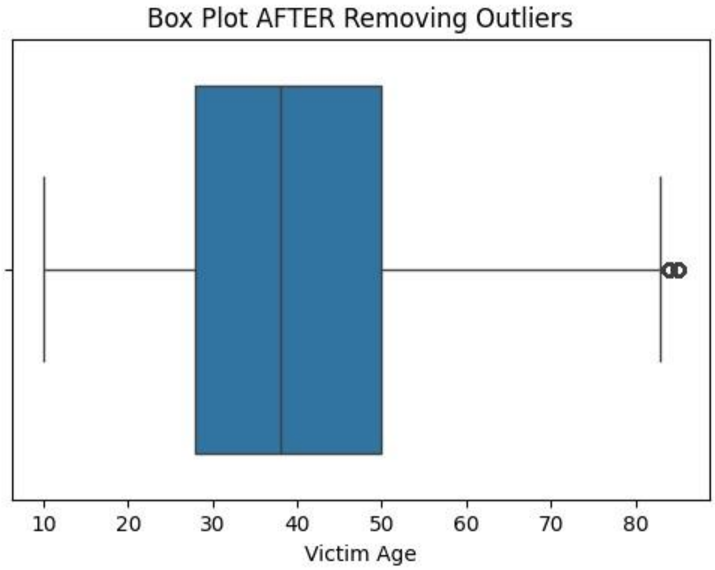
```
Q1 = df["Victim Age"].quantile(0.25)
Q3 = df["Victim Age"].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers = df[(df["Victim Age"] < lower_bound) | (df["Victim Age"] >
upper_bound)]
print("Outliers in Victim Age Column:\n", outliers)
df_cleaned = df[(df["Victim Age"] >= lower_bound) & (df["Victim Age"] <=
upper_bound)]
print(f"Original dataset size: {df.shape[0]} rows")
print(f"Dataset size after removing outliers: {df_cleaned.shape[0]} rows")
```

```
Outliers in Victim Age Column:
        DR Number Date Reported Date Occurred  Time Occurred  Area ID  \
101     190814470    08/21/2019    08/21/2019           1220        8
141     190915755    08/24/2019    08/24/2019           1655        9
146     190916045    08/30/2019    08/30/2019             10        9
152     191008351    04/11/2019    04/11/2019            540       10
250     191418726    08/25/2019    08/19/2019           1230       14
...           ...           ...           ...            ...      ...
619427  240713209    12/12/2024    12/11/2024           1230        7
619432  241415646    12/07/2024    12/07/2024              5       14
619530  240613544    11/25/2024    11/24/2024           1600        6
619546  240812440    12/09/2024    12/09/2024            335        8
619578  241714453    11/24/2024    11/24/2024             45       17
```

```
[11396 rows x 18 columns]
Original dataset size: 619595 rows
Dataset size after removing outliers: 520295 rows
```

ADITYA LALWANI D15C ROLL NO.30

Box plot after removing outliers:




The box plot without outliers shows victim ages more clearly, as extreme values are removed. The whiskers now cover only the normal range (1.5*IQR). A few mild outliers might remain, but the data looks more balanced. This makes the analysis more accurate.

Conclusion:

We learned about Data Visualization and Exploratory Data Analysis using Matplotlib and Seaborn.The bar graph showed 77th Street and Wilshire have the most crashes, likely due to heavy traffic or bad roads.The scatter plot showed some areas have more crashes at specific places, with a few unusual spots having different patterns.The contingency table confirmed 77th Street and Wilshire report the most crashes.The box plot showed most victims are 25-55 years old, with a few over 80.The heatmap showed most number-based features are weakly linked, except for strong ties between Area ID and Reporting District.The histogram showed victim ages are mostly in the mid-20s to early 30s, with fewer older victims.Removing outliers using the IQR method made the data more accurate.