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()
```

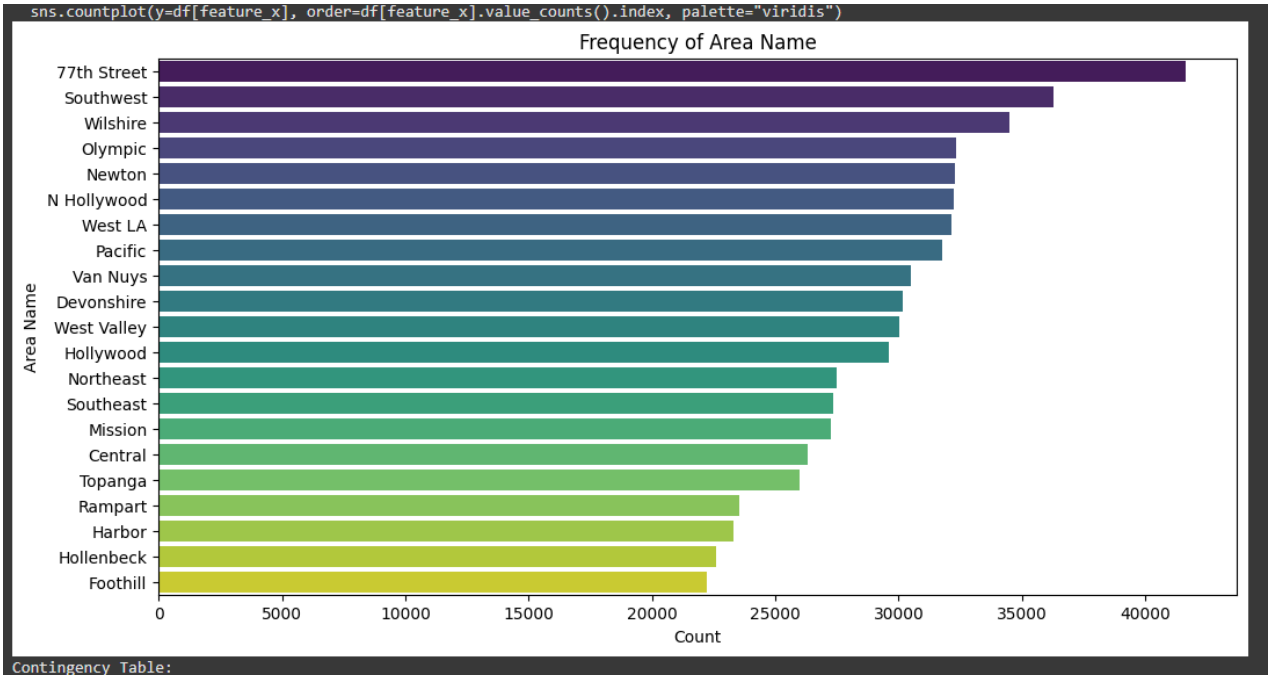
```
df.head()
```

	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 above bar graph represents the frequency of vehicle collisions across different areas. The x-axis shows the number of collisions, while the y-axis lists the area names. 77th Street has the highest number of reported collisions, followed by Southwest and Wilshire. This might mean that certain areas experience significantly more traffic collisions, which could indicate high traffic density, accident-prone roads, or other factors.

Contingency Table:

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 contingency table displays the number of traffic collisions across different areas. Each row represents an area name, while the corresponding value indicates the number of reported traffic collisions in that area. 77th Street has the highest number of collisions, followed by Wilshire and Southwest, indicating higher traffic incidents in these regions. While areas like Foothill, Harbor, and Rampart have relatively fewer collisions. This distribution suggests variations in traffic density, road conditions, or reporting frequency across different areas.

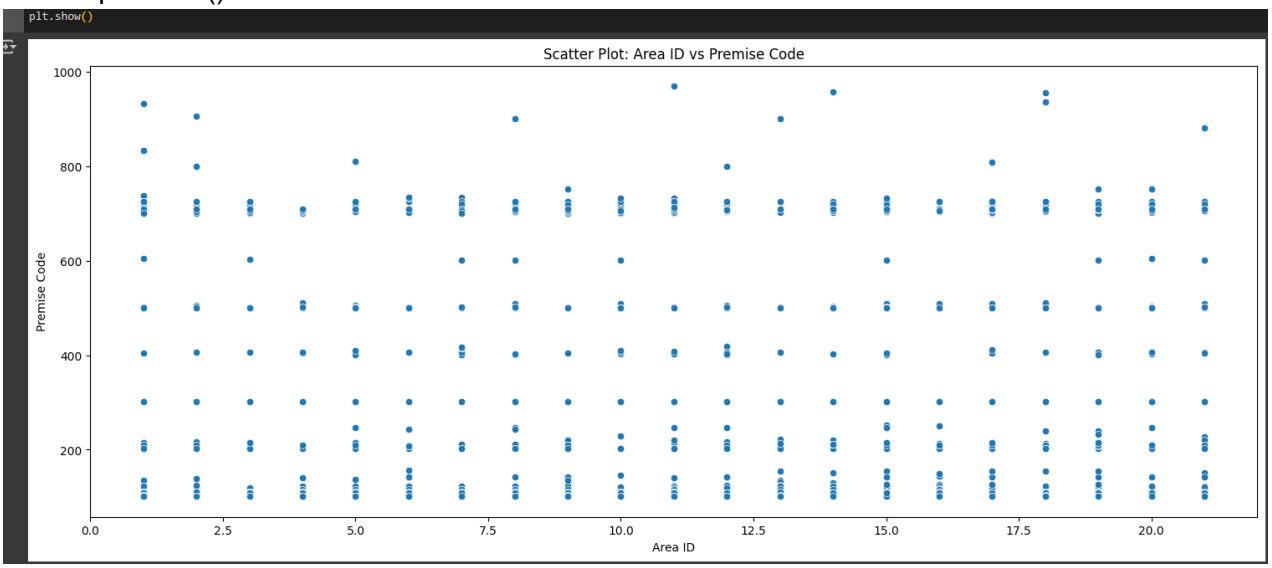
ADITYA LALWANI D15C ROLL NO.30

- 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()
```

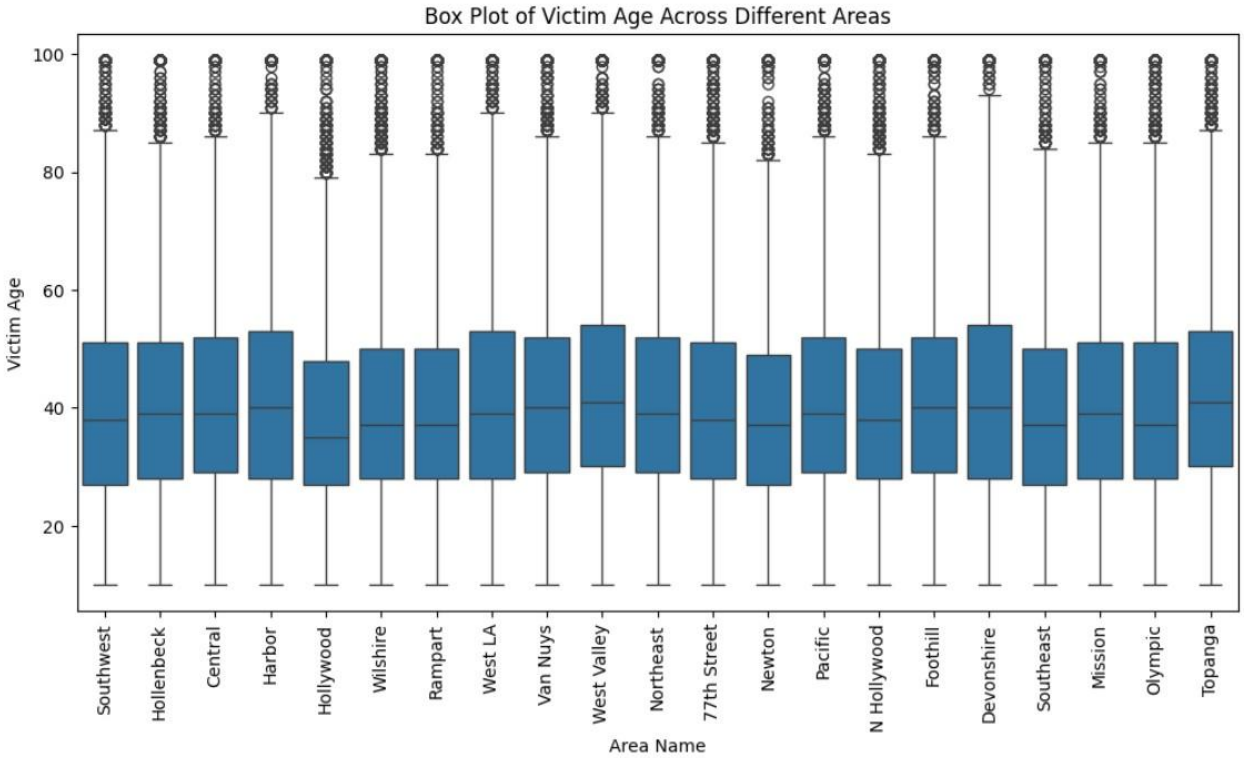


The scatter plot visualizes the relationship between Area ID and Premise Code in vehicle collision data. The x-axis represents different area IDs, while the y-axis represents premise codes, which categorizes the type of location where the collision occurred. The scattered points suggest that collisions happen across various premises in all areas, with some areas showing higher concentrations at specific premise codes. There are a few outliers, indicating locations where collisions are significantly more or less frequent.

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()
```

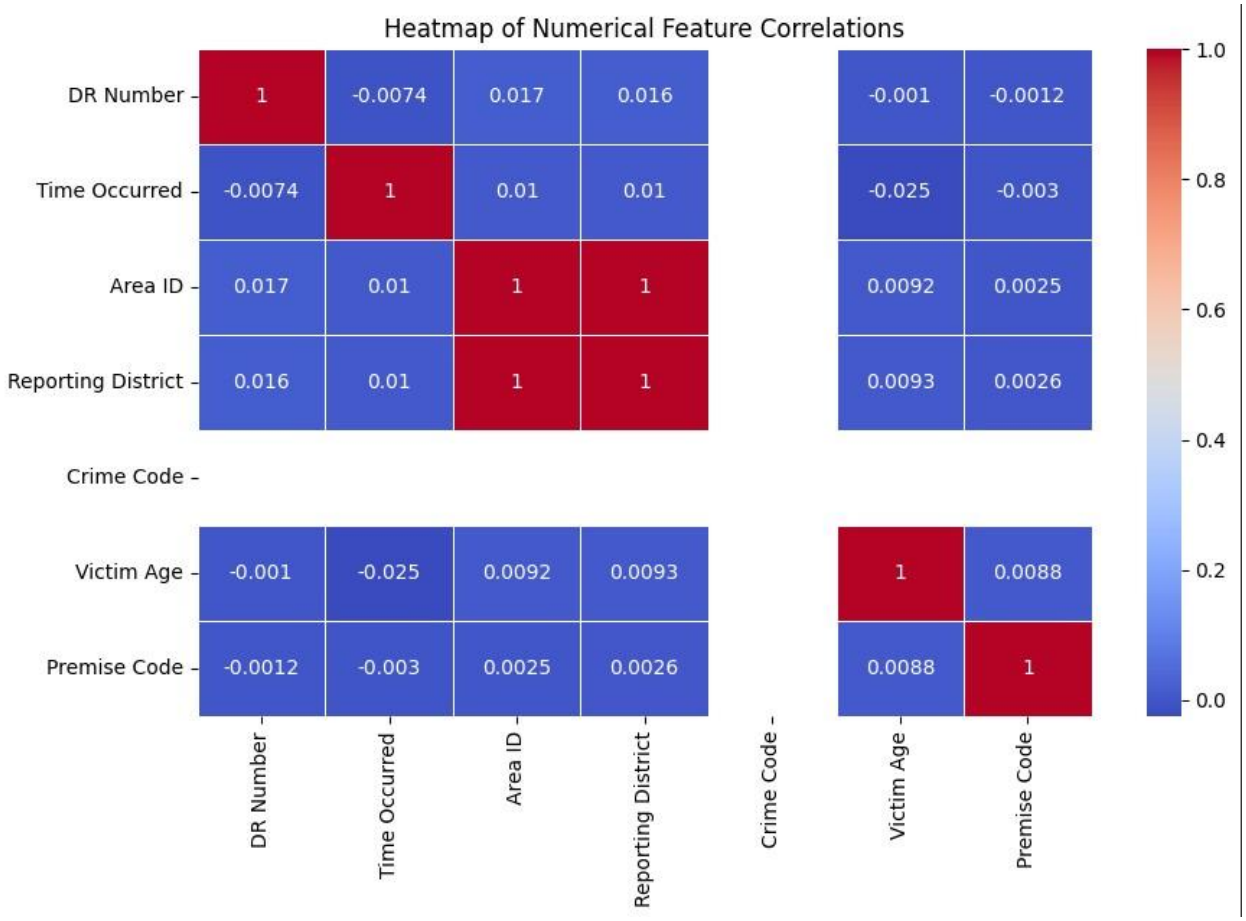


The box plot visualizes the distribution of victim ages across different areas, highlighting variations in age demographics. The median victim age appears to be around 35-45 years in most areas, with interquartile ranges spanning from approximately 25 to 55 years. The whiskers extend towards younger and older victims, with numerous outliers above 80 years, indicating some elderly victims involved in incidents. The overall distribution remains fairly consistent across areas, suggesting similar age patterns in reported cases regardless of location.

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 visualizes the correlation matrix of numerical features, where values range from -1 to 1. A strong correlation (value = 1) is observed between Area ID and Reporting District, indicating they are closely related. Most other features exhibit weak or near-zero correlations, suggesting minimal linear relationships. Victim Age shows little correlation with Time Occurred and Premise Code, implying that age does not significantly influence when or where incidents occur. Similarly, DR Number and Crime Code have no meaningful correlation with other variables, indicating they function as independent identifiers. Overall, the heatmap suggests that most numerical features are weakly correlated, except for geographical identifiers, which show a strong relationship.

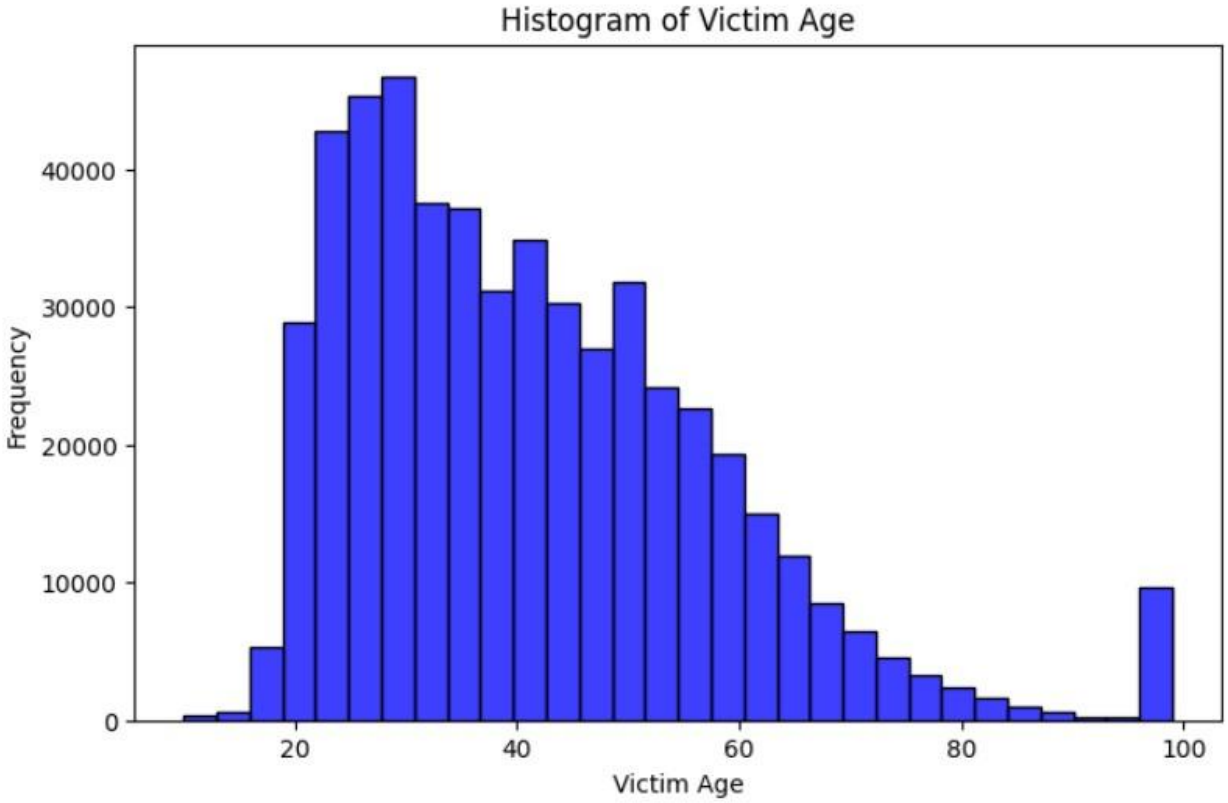
ADITYA LALWANI D15C ROLL NO.30

- Create histogram and normalized Histogram:-

1. Histogram:

Command:

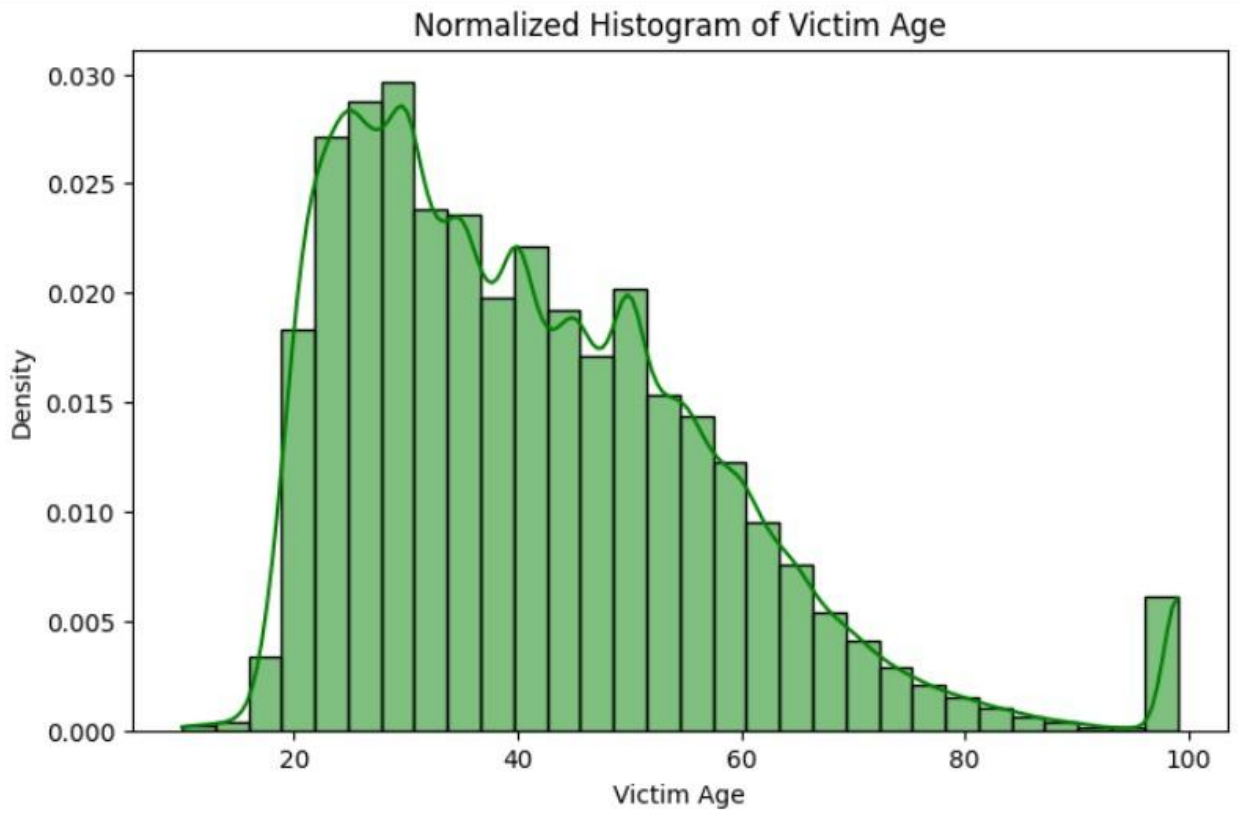
```
plt.figure(figsize=(8, 5))
sns.histplot(df["Victim Age"], bins=30, kde=False, color="blue")
plt.title("Histogram of Victim Age")
plt.xlabel("Victim Age")
plt.ylabel("Frequency")
plt.show()
```



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")
```

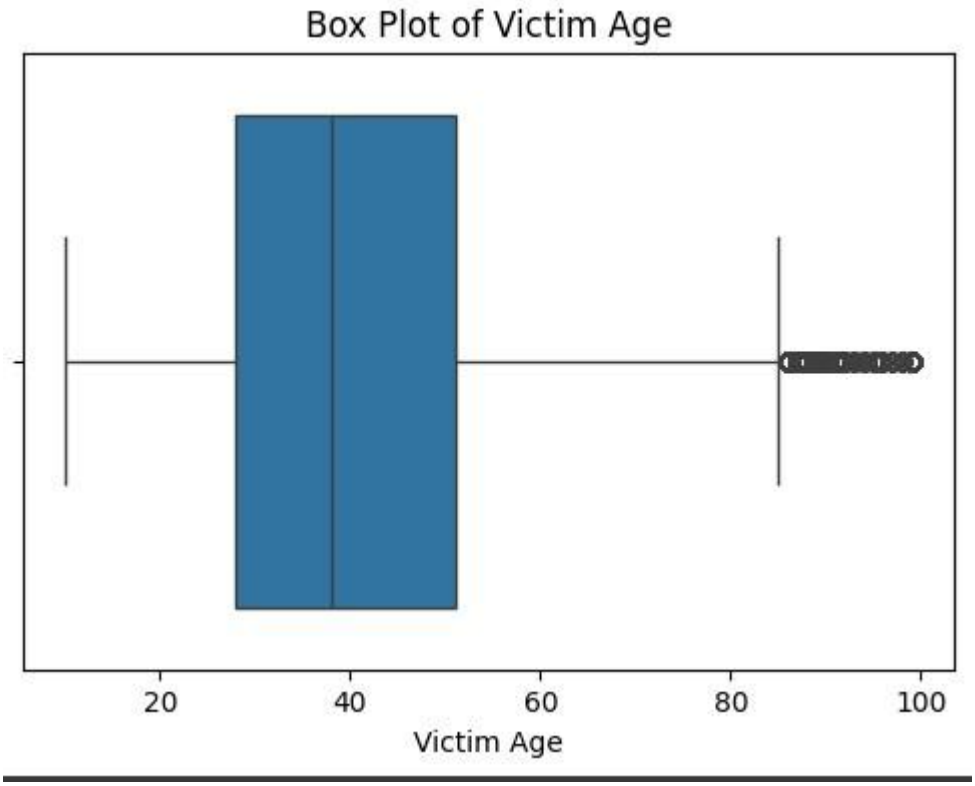


- Handle outlier using box plot and Inter quartile range:

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: