DATE: 1st November,2023

NAME: P. SATYA PREM
PRN: 22070126073
AIML-A3

NAME: M. SRINIVAS
PRN: 22070126073
AIML-A3

BLOG ON EDA AND DPL PROJECT

CRIME TRENDS AND PATTERNS IN INDIA AGAINST WOMEN: A COMPARATIVE STUDY

INTRODUCTION:

Exploring the landscape of crimes against women is a multifaceted endeavor, delving into the intricate patterns and occurrences of various offenses. This analysis constitutes an extensive examination of a sensitive societal concern, demanding meticulous attention and a comprehensive approach. The fundamental aim is to dissect the temporal and geographic trends of these crimes, highlighting their nuances and evolving patterns. Understanding these dynamics is crucial for effective intervention strategies and social policymaking. This research is carried out through detailed data analysis and visualization techniques, leveraging comprehensive datasets that encompass an array of crimes perpetrated against women. By examining shifts in the frequency, types, and geographic hotspots of these offenses, the goal is to paint a comprehensive picture of their changing nature. The investigation goes beyond mere enumeration and seeks to uncover the underlying factors contributing to these crime trends. The significance of this analysis extends beyond statistics, aiming to provide a deeper understanding of the societal, cultural, and systemic elements shaping these crimes.

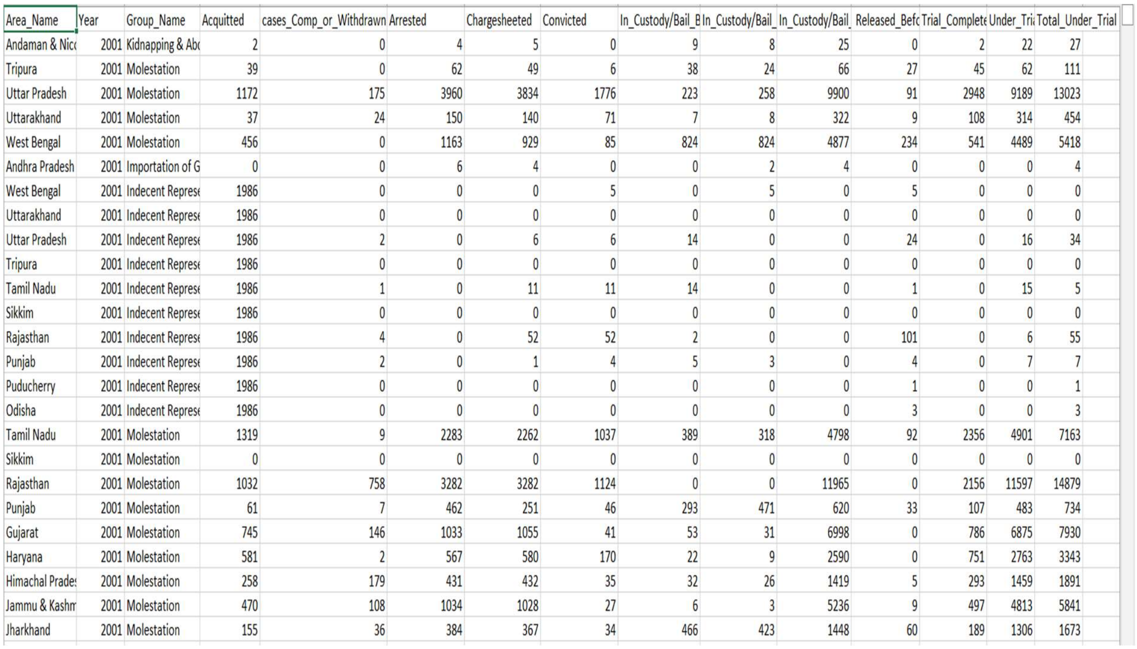
The context of this exploration lies in the intersection of societal, legal, and ethical frameworks, wherein crimes against women are not merely statistical entries but represent grave transgressions against human rights. This analysis isn't just about data; it's about understanding the experiences, vulnerabilities, and the evolving nature of these crimes, aiming to drive impactful change. It emphasizes a nuanced understanding of how different types of crimes against women have altered over time and how these variations manifest across diverse geographical landscapes. The goal here isn't solely to compile numbers and figures but to comprehend the underlying patterns, disparities, and influences that drive these offenses. This extensive endeavour serves as a critical tool in informing policymakers, law enforcement agencies, and advocacy groups, ultimately aiming to build safer, more informed communities and fostering a better understanding of this complex societal issue.

PROBLEM STATEMENT:

The objective of this study is to comprehensively explore the landscape of crimes against women in India over the years, using sophisticated data preprocessing and exploratory data analysis (EDA) techniques. The primary focus is to examine and decipher trends and patterns associated with various crimes perpetrated against women in India, encompassing offenses such as domestic violence, sexual assault, dowry-related crimes, and other relevant categories.

Utilizing Python programming language and Data Pre-Processing methods, this project aims to conduct an extensive Exploratory Data Analysis (EDA) on crime data related to incidents against women in India. The primary objective is to systematically investigate and represent state-wise and crime-wise variations, trends, patterns, and potential correlations within the dataset.

DATASET DESCRIPTION:



Area_Name	Year	Group_Name	Acquitted	cases_Comp_or_Withdrawn	Arrested	Chargesheeted	Convicted	In_Custody/Bail_B	In_Custody/Bail_	In_Custody/Bail_	Released_Befc	Trial_Complete	Under_Tri	Total_Under_Trial
Andaman & Nic	2001	Kidnapping & Abc	2	0	4	5	0	9	8	25	0	2	22	27
Tripura	2001	Molestation	39	0	62	49	6	38	24	66	27	45	62	111
Uttar Pradesh	2001	Molestation	1172	175	3960	3834	1776	223	258	9900	91	2948	9189	13023
Uttarakhand	2001	Molestation	37	24	150	140	71	7	8	322	9	108	314	454
West Bengal	2001	Molestation	456	0	1163	929	85	824	824	4877	234	541	4489	5418
Andhra Pradesh	2001	Importation of G	0	0	6	4	0	0	2	4	0	0	0	4
West Bengal	2001	Indecent Represe	1986	0	0	0	5	0	5	0	5	0	0	0
Uttarakhand	2001	Indecent Represe	1986	0	0	0	0	0	0	0	0	0	0	0
Uttar Pradesh	2001	Indecent Represe	1986	2	0	6	6	14	0	0	24	0	16	34
Tripura	2001	Indecent Represe	1986	0	0	0	0	0	0	0	0	0	0	0
Tamil Nadu	2001	Indecent Represe	1986	1	0	11	11	14	0	0	1	0	15	5
Sikkim	2001	Indecent Represe	1986	0	0	0	0	0	0	0	0	0	0	0
Rajasthan	2001	Indecent Represe	1986	4	0	52	52	2	0	0	101	0	6	55
Punjab	2001	Indecent Represe	1986	2	0	1	4	5	3	0	4	0	7	7
Puducherry	2001	Indecent Represe	1986	0	0	0	0	0	0	0	1	0	0	1
Odisha	2001	Indecent Represe	1986	0	0	0	0	0	0	0	3	0	0	3
Tamil Nadu	2001	Molestation	1319	9	2283	2262	1037	389	318	4798	92	2356	4901	7163
Sikkim	2001	Molestation	0	0	0	0	0	0	0	0	0	0	0	0
Rajasthan	2001	Molestation	1032	758	3282	3282	1124	0	0	11965	0	2156	11597	14879
Punjab	2001	Molestation	61	7	462	251	46	293	471	620	33	107	483	734
Gujarat	2001	Molestation	745	146	1033	1055	41	53	31	6998	0	786	6875	7930
Haryana	2001	Molestation	581	2	567	580	170	22	9	2590	0	751	2763	3343
Himachal Prade:	2001	Molestation	258	179	431	432	35	32	26	1419	5	293	1459	1891
Jammu & Kashm	2001	Molestation	470	108	1034	1028	27	6	3	5236	9	497	4813	5841
Jharkhand	2001	Molestation	155	36	384	367	34	466	423	1448	60	189	1306	1673

- The dataset was taken from NCRB(National Crime Records Bureau) which is widely used by policy makers, researchers etc.
- The initial dataset contains of about 3800 rows and 16 columns.

The columns are:

- **Area_Name**: This column indicates the geographical area or region within India where the reported crimes against women occurred.
- **Year:** This column represents the year in which the crime data was recorded, providing a time reference for the reported incidents.
- **Crime_Name:** This column categorizes the data into specific groups or types of crimes such as, Kidnapping & Abduction - Women & Girls, molestation, cruelty of husband and relatives, importation of girls etc, indicating the nature of the offenses under consideration.
- **Acquitted:** This column likely contains the count of individuals who were accused of crimes against women but were subsequently acquitted, meaning they were found not guilty in the legal proceedings.
- **cases_Comp_or_Withdrawn:** This column records the number of cases where charges were either compounded (settled or resolved with an agreement)
- **Arrested:** This column provides the count of individuals who were arrested in connection with crimes against women.
- **Chargesheeted:** It likely represents the number of individuals against whom formal charges were filed, indicating the commencement of legal proceedings.
- **Convicted:** This column indicates the count of individuals who were found guilty and convicted of the reported crimes.

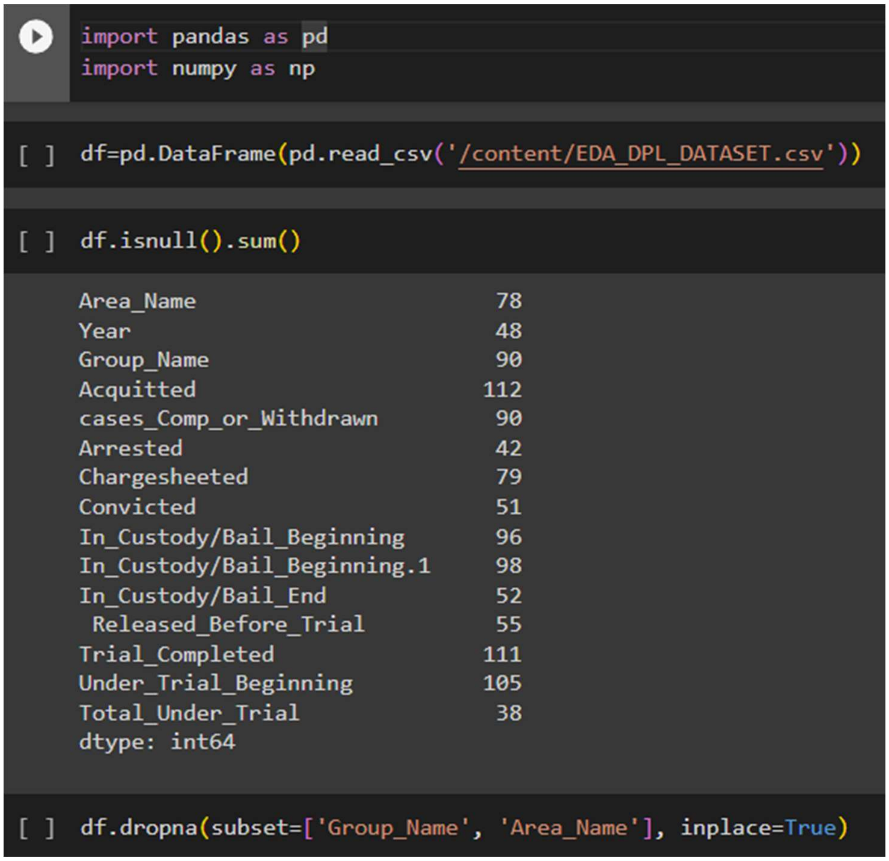
- **In_Custody/Bail_Beginning:** This column likely refers to the number of individuals who were either in police custody or on bail during the initial stages of the investigation
- **In_Custody/Bail_End::** It probably records the number of individuals who were in custody or on bail during the investigation at the end of the year.
- **In_Custody/Bail_End(TRIAL):** This column may represent the count of individuals who were in custody or on bail during the trial phase at the end of the year.
- **Released_Before_Trial:** This column likely contains the number of individuals who were released or freed by the magistrate before the trial due to reasons such as lack of evidence.
- **Trial_Completed:** This column indicates the count of individuals for whom the trial proceedings were completed during the year.
- **Under_Trial_Beginning:** It probably represents the number of individuals who were in the process of trial at the beginning of the year.
- **Total_Under_Trial:** This column likely sums up the total count of individuals who were under trial for crimes against women during the year.

DATA PRE-PROCESSING:

To maintain the integrity of our analysis, a rigorous **data cleaning** process was initiated. This encompassed a thorough examination and treatment of the dataset to eliminate inconsistencies and inaccuracies. The primary steps involved the identification and removal of non-numeric entries, missing values, and outlier rows within the dataset. Non-numeric entries and incomplete data points were excluded to ensure the uniformity and reliability of the information under scrutiny.

Moreover, in addressing numerical outliers, a systematic approach was employed to replace these anomalous data points with statistically appropriate measures. Numerical outliers and missing values were imputed using statistical measures such as the median. This method aims to mitigate the impact of outliers on subsequent analysis and maintain a consistent dataset, free from irregular values that might skew the findings.

The significance of this data cleaning process cannot be overstated. It serves as the foundational groundwork, ensuring that the subsequent analysis is based on a high-quality dataset. This meticulous data preparation is akin to the meticulous crafting of a clean canvas before an artist begins to paint. It sets the stage for reliable and robust analysis, providing a clear and accurate picture of the underlying patterns within the data.



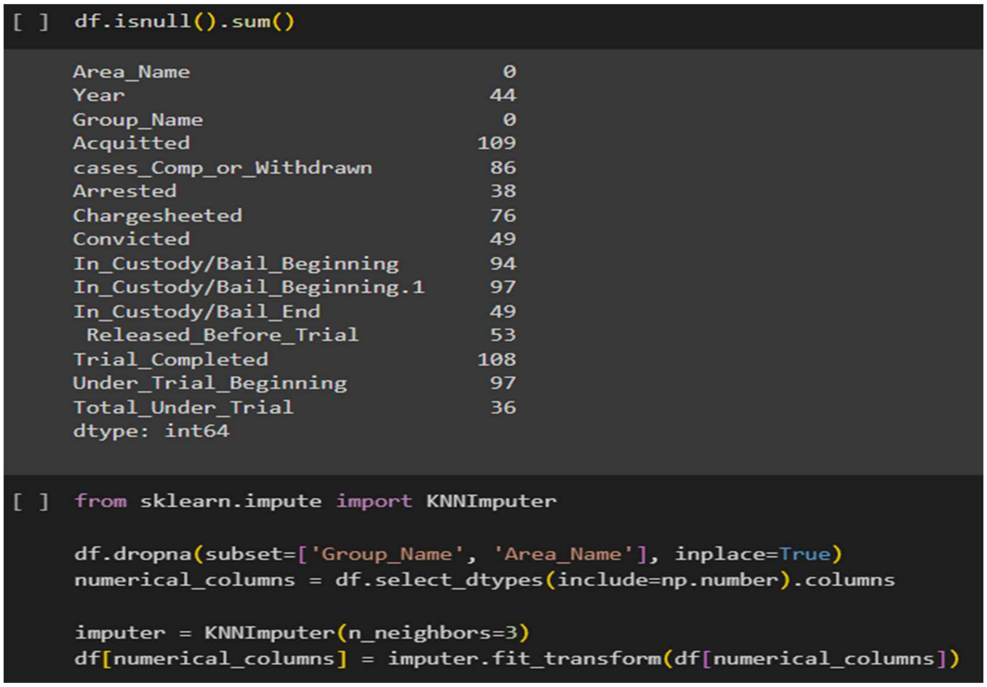
```
import pandas as pd
import numpy as np
```

```
df=pd.DataFrame(pd.read_csv('/content/EDA_DPL_DATASET.csv'))
```

```
df.isnull().sum()
```

```
Area_Name                        78
Year                             48
Group_Name                       90
Acquitted                       112
cases_Comp_or_Withdrawn          90
Arrested                         42
Chargesheeted                    79
Convicted                        51
In_Custody/Bail_Beginning        96
In_Custody/Bail_Beginning.1      98
In_Custody/Bail_End              52
 Released_Before_Trial           55
Trial_Completed                 111
Under_Trial_Beginning           105
Total_Under_Trial                38
dtype: int64
```

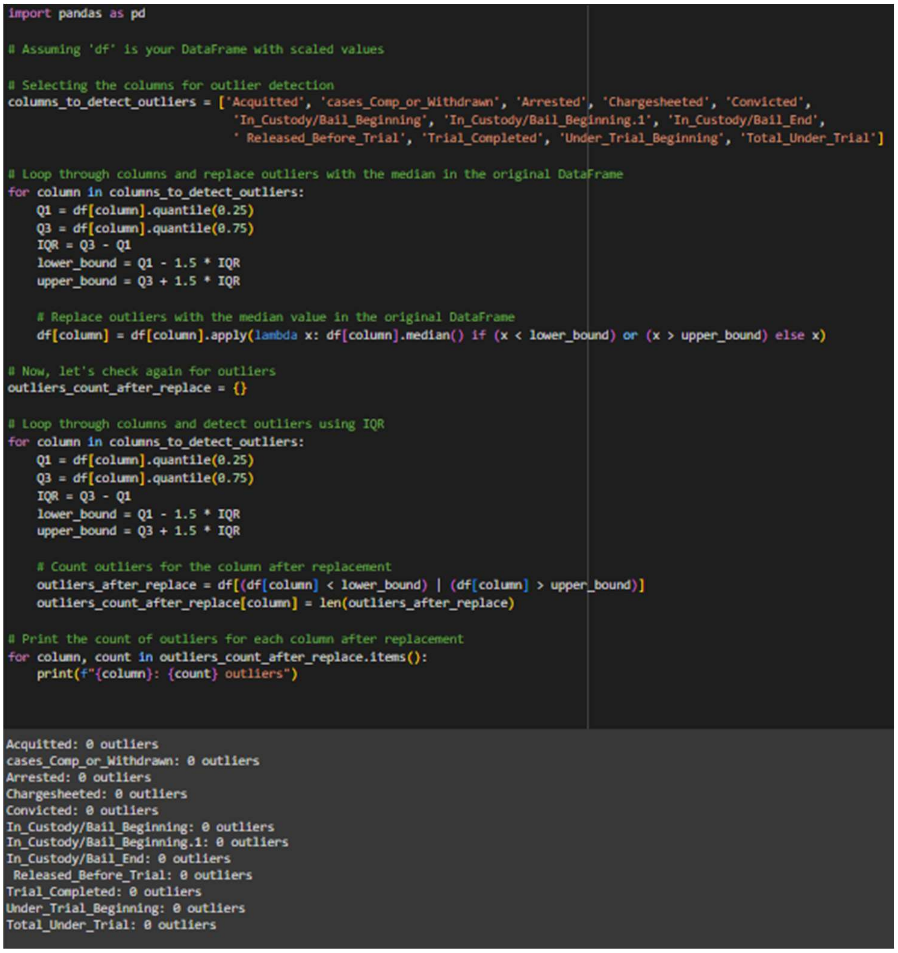
```
df.dropna(subset=['Group_Name', 'Area_Name'], inplace=True)
```



```
df.isnull().sum()
```

```
Area_Name                        0
Year                            44
Group_Name                       0
Acquitted                      109
cases_Comp_or_Withdrawn         86
Arrested                        38
Chargesheeted                   76
Convicted                       49
In_Custody/Bail_Beginning       94
In_Custody/Bail_Beginning.1     97
In_Custody/Bail_End             49
 Released_Before_Trial          53
Trial_Completed                108
Under_Trial_Beginning           97
Total_Under_Trial               36
dtype: int64
```

```
from sklearn.impute import KNNImputer

df.dropna(subset=['Group_Name', 'Area_Name'], inplace=True)
numerical_columns = df.select_dtypes(include=np.number).columns

imputer = KNNImputer(n_neighbors=3)
df[numerical_columns] = imputer.fit_transform(df[numerical_columns])
```



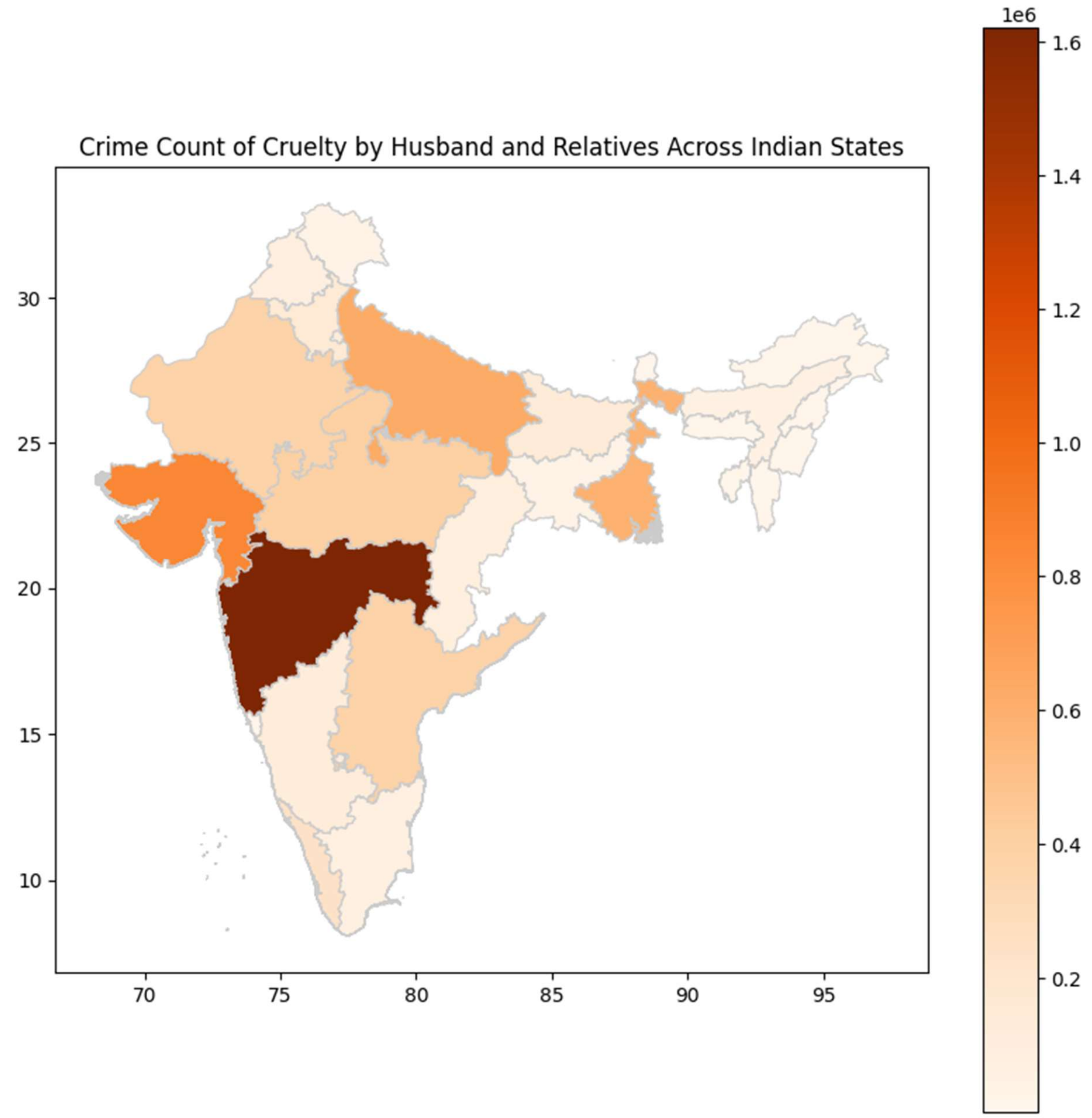
```
import pandas as pd

# Assuming 'df' is your DataFrame with scaled values

# Selecting the columns for outlier detection
columns_to_detect_outliers = ['Acquitted', 'cases_Comp_or_Withdrawn', 'Arrested', 'Chargesheeted', 'Convicted',
                              'In_Custody/Bail_Beginning', 'In_Custody/Bail_Beginning.1', 'In_Custody/Bail_End',
                              ' Released_Before_Trial', 'Trial_Completed', 'Under_Trial_Beginning', 'Total_Under_Trial']

# Loop through columns and replace outliers with the median in the original DataFrame
for column in columns_to_detect_outliers:
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Replace outliers with the median value in the original DataFrame
    df[column] = df[column].apply(lambda x: df[column].median() if (x < lower_bound) or (x > upper_bound) else x)

# Now, let's check again for outliers
outliers_count_after_replace = {}

# Loop through columns and detect outliers using IQR
for column in columns_to_detect_outliers:
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Count outliers for the column after replacement
    outliers_after_replace = df[(df[column] < lower_bound) | (df[column] > upper_bound)]
    outliers_count_after_replace[column] = len(outliers_after_replace)

# Print the count of outliers for each column after replacement
for column, count in outliers_count_after_replace.items():
    print(f"{column}: {count} outliers")
```

```
Acquitted: 0 outliers
cases_Comp_or_Withdrawn: 0 outliers
Arrested: 0 outliers
Chargesheeted: 0 outliers
Convicted: 0 outliers
In_Custody/Bail_Beginning: 0 outliers
In_Custody/Bail_Beginning.1: 0 outliers
In_Custody/Bail_End: 0 outliers
 Released_Before_Trial: 0 outliers
Trial_Completed: 0 outliers
Under_Trial_Beginning: 0 outliers
Total_Under_Trial: 0 outliers
```

DATA VISUALIZATIONS:

When we look at violence against women, it's a big deal. To understand it better, I used **GeoPandas -** a neat tool that helps make maps interesting. It showed me where these things happen on a map of India. By seeing it on the map, it's easier to spot how these incidents are scattered and discover trends that you might miss if you just check out the numbers. The goal of **GeoPandas** is to make working with geospatial data in python easier. It combines the capabilities of pandas and shapely, providing geospatial operations in pandas and a high-level interface to multiple geometries to shapely.

Plotly express is a high-level data visualization package that allows you to create interactive plots with very little code. It is built on top of Plotly Graph Objects, which provides a lower-level interface for developing custom visualizations.


Crime Count of Cruelty by Husband and Relatives Across Indian States
30
25
20
15
10
70
75
80
85
90
95
1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2

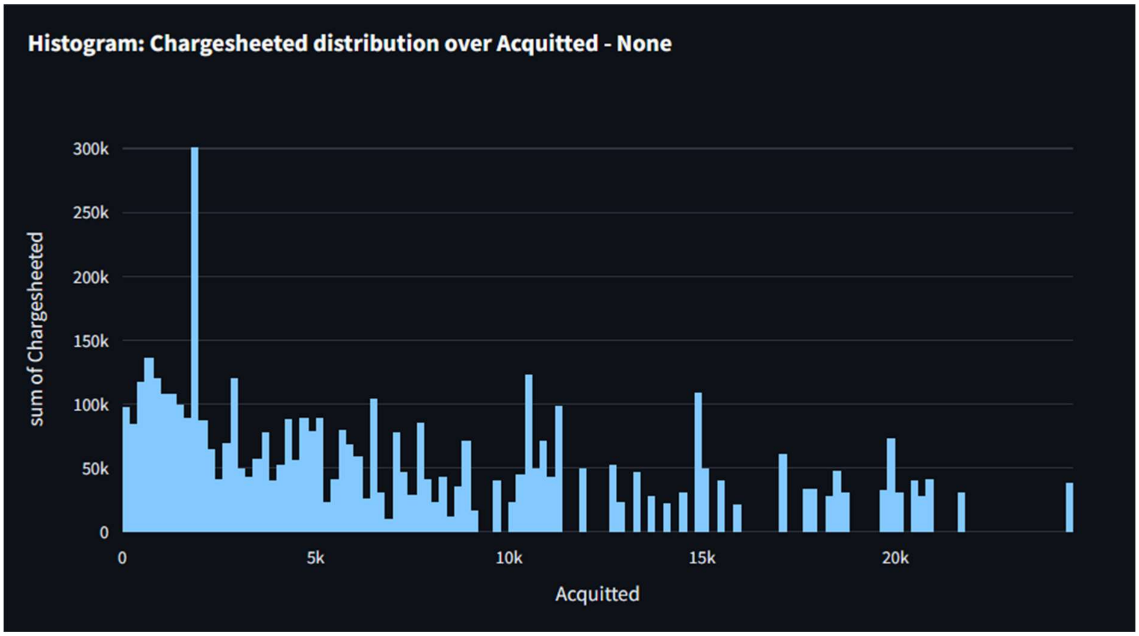
WHY EDA?

EDA isn't just about looking at the numbers; it's like being a detective trying to find connections between different things. For instance, you might look at things like how rich or poor an area is, or how many people live there. By digging into these details, you can find links between these factors and the crimes against women. It's like finding clues that help understand why these incidents happen in certain places.

DEPLOYMENT:

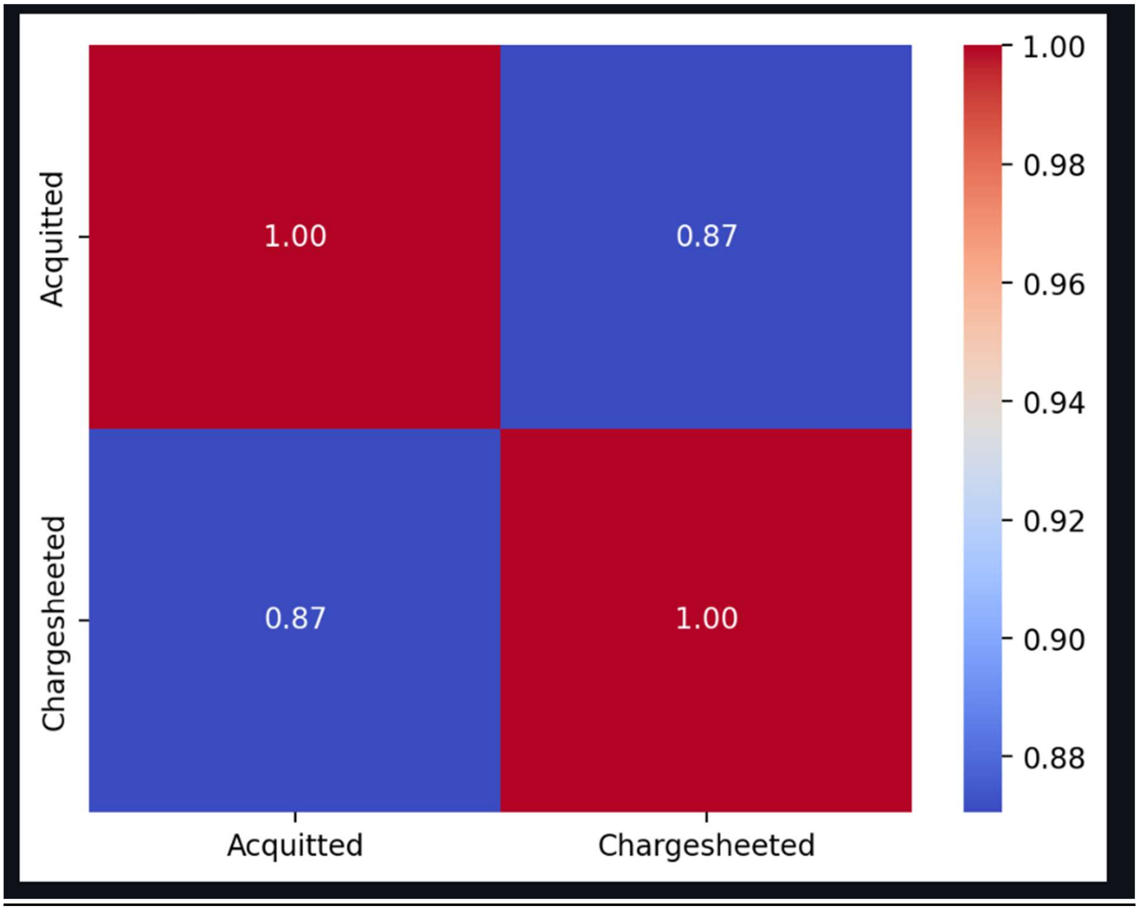
Conviction Rates App:

Our first stop is the Conviction Rates app, a powerful tool for understanding the legal outcomes of crimes against women. Users can select a specific crime type and a state, and the app unveils the corresponding conviction rate. Under the hood, the app loads and processes data from the "EDA_DPL_DATASET.csv," providing users with valuable insights into conviction trends.



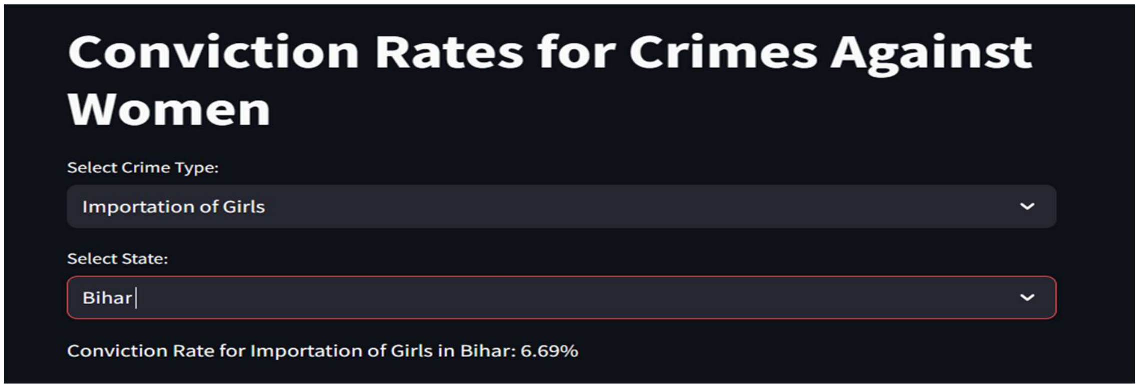
Correlation Analysis App:

Moving on, our Correlation Analysis app provides users with the ability to explore relationships between different types of crimes against women. By selecting two crime types, the app generates a heatmap of the correlation matrix, offering a visual representation of potential connections.



Data Visualization App:

Our final destination is the Data Visualization app, where we bring the data to life through interactive visualizations. Users can choose x-axis and y-axis columns, select a plot type, and filter by state for a more granular analysis. The app supports bar plots, histograms, scatter plots, and area plots.



Conviction Rates for Crimes Against Women

Select Crime Type:

Importation of Girls

Select State:

Bihar

Conviction Rate for Importation of Girls in Bihar: 6.69%

JUSTIFICATION:

The investigation into crime trends and patterns against women in India holds profound significance for several compelling reasons. Primarily, India grapples with a complex societal landscape where women face multifaceted challenges. Understanding crime trends against women provides invaluable insights into the dynamics of violence, exploitation, and discrimination, which are often underreported or misrepresented. By delving into this intricate web of data, the study aims to uncover critical insights, providing a comprehensive understanding of the nuances in crime patterns and their temporal or geographical variations. Such insights are vital for policy formulation, effective law enforcement, and the design of intervention strategies. Moreover, this project contributes to the broader societal goal of fostering a safer environment for women by identifying areas that demand immediate attention and proactive measures. Ultimately, the project's findings can serve as a foundational resource for policymakers, law enforcement agencies, and social activists, fostering a more informed and targeted approach towards reducing crimes against women in India.

CONCLUSION:

The exploration of crimes against women in India through comprehensive data analysis and visualization techniques unveils a layered narrative, revealing intricate patterns and temporal shifts within these offenses. This in-depth investigation delves beyond statistical representation, aiming to uncover the nuanced dynamics and underlying factors contributing to these crimes. By deciphering the evolving nature and geographic variations, this analysis provides not just statistics but a deeper understanding of societal, cultural, and systemic influences shaping these transgressions.

Moreover, this analysis is not just about numerical figures; it's akin to detective work. It involves uncovering connections between various factors such as socio-economic status, population density, and the prevalence of crimes against women. By scrutinizing these details, it offers insights, akin to uncovering clues that help comprehend the contextual reasons behind these incidents within specific geographic regions.

Future implementations include:

- **Pattern Recognition**: EDA and data preprocessing reveal distinct patterns in crimes against women like dowry attacks and sati, aiding targeted interventions.
- **Predictive Modeling**: Classification helps predict and prevent such crimes, enabling early intervention and proactive measures.
- **Holistic Policy**: Analysis uncovers societal factors, informing comprehensive policies for gender equality and safer communities.
- **Societal Impact**: Data-driven approaches contribute to broader societal change, empowering women and advancing gender equality.
- **Efficient Resource Allocation**: Clustering identifies crime hotspots, optimizing resource allocation for law enforcement and support services.

Link for GitHub repository : https://github.com/SrinivasMotepalli/EDA_DPL_PROJECT

Streamlit Links:

[1]. https://edadplmultiplegraphs-65-73.streamlit.app

[2]. https://edadplheatmapcorrelation-65-73.streamlit.app

[3]. https://edadplconvictionrate-65-73.streamlit.app