**GITHUB:**

## Crime Data Analysis on Women

**1. Statistical Data**

Below are the important findings from the statistical moments and analysis of the crime data on three kinds of crimes, namely: Rape, Kidnapping & Abduction, and Dowry Deaths. Meaning from the column "Rape," the average is 727.86, which is way lower than its median at 348.50, representing extreme values based on a huge standard deviation of 977.02. Indeed, these values support a positively skewed distribution with heavy tails in terms of a high level of skewness at 2.28 and kurtosis at 6.65.

Similarly, the "K&A" column has a high mean of 1134.54 and a median of 290.00 with a very large standard deviation of 1993.54. The skewness of 3.13 and kurtosis of 12.51 indicate a long-tailed, positively skewed distribution. The "DD" column indicates the lowest mean at 215.69 and a median of 29.00; however, it has the highest skewness of 3.37 and kurtosis of 12.52, indicating rare extreme cases.

The correlation matrix reveals a strong relationship of 0.70 between Rape and K&A, and a moderate correlation of 0.55 between Rape and DD. From the descriptive statistics, cases of Rape range from 0 to 6337; K&A cases are from 0 to 15381; and finally, DD ranges between 0 and 2524 with large variability for all measures.

**1. Categorical Graph: Total Rapes by State (Bar Chart)**

The following bar chart aggregates the number of rape cases reported from different states, allowing categorical comparison. States like Madhya Pradesh and Rajasthan have very high bars, thus clearly outliers with higher crime numbers. These may reflect variations in population size, socioeconomic conditions, and cultural attitudes toward the reporting of these crimes. On the other hand, the smaller states and union territories have reported lower numbers, which may not be indicative of the reality on the ground but might reflect underreporting or a lack of reporting mechanisms. This chart is invaluable for targeting areas that need immediate intervention and resource allocation. The x-axis lists all the states in alphabetical order for easy reference, and the y-axis quantifies the total reported cases. This visualization will enable regional analysis, helping stakeholders to focus on critical areas. Furthermore, this variation raises very important questions on how systemic barriers like stigma or poor law enforcement would then impact data accuracy and effectiveness of policies.

A graph of a number of people

Description automatically generated with medium confidence

**2. Relational Graph: Total Cases of Rape Across the Years (Line Chart)**

The line chart visualizes the trend of total reported rape cases from 2000 to 2020, showing trends over time. The periodic rises and falls in the crime graph most likely mirror changes in public awareness, legal reforms, and reporting mechanisms. Peaks in the timeline may correspond with periods of heightened media attention or governmental campaigns encouraging reporting, while troughs may reflect underreporting or diminished public attention. This visualization is good at showing how societal aspects of crime reporting change over time. For instance, one might find a sudden drop or a gradual decline in the cases of crimes after the institution of strict laws or police reforms. The x-axis denotes the year, and the y-axis shows the total cases, with each point interconnected by lines to show their relationship over time. This analysis will help policymakers to evaluate past interventions, identify the lagging periods, and propose new strategies geared toward crime reduction or improving data accuracy.

A graph with a line going up

Description automatically generated

3. **Heat map:**

The heat map illustrates the relationship among three types of crimes: Rape, Kidnapping & Abduction (K&A), and Dowry Deaths (DD), to understand their interrelationships. High correlations, such as 0.70 between Rape and K&A, indicate that these crimes often co-occur, which may be driven by some common underlying factors like gender inequality or social norms. Moderate correlations, such as 0.55 between Rape and DD, indicate less direct but still significant relationships. This analysis provides a much richer view of the crime ecosystems and aids policymakers in designing holistic intervention strategies. The color gradients used here, from red for strong correlations to blue for weak, make complex numerical relationships visually intuitive. The red square at the junction of Rape and K&A points out their strong interdependence. This heatmap becomes a very important analytical tool for researchers and authorities toward understanding and eventually solving issues of domestic violence and other gender-based crimes in their inter connectedness.

**4. Elbow Plot for Optimal Clusters (Line Chart)**

An elbow plot visualizes the relation of the number of clusters with the distortion, or inertia, in k-means clustering, used to determine the optimal number of clusters.

A red and blue squares with white text

Description automatically generated

The "elbow" point, where the distortion sharply decreases before flattening out, indicates the most efficient cluster count—three clusters in this case. This becomes important for the good segmentation of data without overfitting. Each cluster represents a group of regions or data points sharing similar crime patterns, such as high-crime versus low-crime areas. The identification of such groups enables the implementation of targeted interventions tailored to specific needs by stakeholders. The x-axis is the number of clusters, and the y-axis quantifies the distortion. After the elbow point, the addition of more clusters yields very little in terms of separation of data in meaningful ways. This plot ensures that clustering really captures the essence of the data without adding unnecessary complexity and is, therefore, a critical step in exploratory data analysis and policy formulation.

A graph with a line

Description automatically generated

**5. Fitting Plot: Linear Regression Predictions with Uncertainty (Scatter Plot)**

This is a scatter plot of observed data points with a fitted regression line and confidence intervals, showing the relationship between Rape cases and Kidnapping & Abduction (K&A) cases. The red line shows the best linear fit, and the grey area shows a 95% confidence interval, demonstrating the uncertainty of predictions within this range. This visualization is very important to understanding how an increase in cases of rape might predict a rise in abduction cases and provide insight for policymakers. For example, if there are increased cases of rape, then that gives the police the opportunity to anticipate such crimes and prepare for them. Also, error bars are added to the predicted values for the plot for making the analysis robust. This combination of observed data, a fitted line, and quantifying uncertainty guarantees reliable interpretations.

A graph with a red line and blue dots

Description automatically generated

Such a plot becomes a practical tool for forward-looking crime-prevention strategies and helps to allocate resources effectively in tackling related crimes by adding

Predictive modelling.

**6. Cluster Visualization Prediction Plot (Scatter Plot)**

A diagram of a cluster

Description automatically generated with medium confidenceThis plot provides a clear understanding of the pattern of crimes based on the clustering of regions when the features are Rape, Kidnapping & Abduction (K&A), and Dowry Deaths (DD). Cluster 0 (Blue) represents a region with high to medium K&A but relatively low Rape cases, indicating abduction-related crimes. Cluster 1 (Orange) consists of regions having low to moderate levels of Rape and K&A, reflecting a certain balance in crime level that could be the consequence of effective control measures. Cluster 2 (Green) represents regions with high levels of crime for both Rape and K&A, categorizing these areas as highly criminalized and thus in need of urgent intervention and resource allocation. Centroids (red X markers) represent the average values in the clusters, giving a bird's view of the crime pattern across regions. Further, new point predictions indicate that a region with moderate crimes such as Rape=800, K&A=600, and DD=100 and a region with Rape=500, K&A=400, and DD=50 fall into Cluster 1, while a region with high crimes of Rape=2000, K&A=1500, and DD=300 falls into Cluster 2, confirming consistency in the model in selecting patterns.