ABSTRACT

This study delves into the intricate dynamics of murder rates spanning the years 2015 to 2021, focusing on discerning trends, spatial patterns, and their correlation with socioeconomic factors. Utilizing a dataset encompassing these years, our analysis employs statistical methodologies to unearth meaningful insights.

We first scrutinize temporal trends, identifying fluctuations and patterns in murder rates over the specified period. By comparing these findings with historical data from the preceding decade, we offer a comprehensive understanding of the evolving landscape of lethal crimes.

Spatial analysis plays a pivotal role in our study, as we investigate geographical variations in murder rates. Through advanced mapping techniques, we delineate hotspots and coldspots, shedding light on the spatial distribution of violent incidents.

Furthermore, our research extends beyond mere descriptive analysis to uncover underlying correlations between murder rates and socioeconomic factors. By integrating data on poverty, education levels, and other pertinent metrics, we scrutinize the interplay between social conditions and crime prevalence. Through regression analysis and correlation studies, we aim to identify significant associations, providing valuable insights for policymakers and stakeholders.

Introduction

Crime is as old as human history itself. It has evolved into different forms, saw variations in it through cultural, temporal, and socio-economic changes. Through all this, the action of crime still persists in our society. Among all the crimes human commits, the crime of murder seems to be the most arrogant. Life is a precious gift to behold, regardless of its struggles, so there is little reasoning behind murder. It is a devastating act that not only ends a life but also leaves a lasting impact on families, communities, and society as a whole. Understanding the motives behind murders is crucial in addressing this serious issue and working towards creating safer communities.

This project delves into the motives behind murders in India over the past 7 years. By analyzing available data on murder cases and their motives, we aim to uncover insights that can inform strategies for crime prevention and intervention. Murders often stem from a variety of factors, including personal disputes, socio-economic inequalities, and systemic issues. By studying these motives, we hope to gain a deeper understanding of the root causes of violent crime and identify ways to address them effectively.

Our analysis will explore patterns and trends in murder motives, examining how they have evolved over time and vary across different regions and demographic groups. By examining the underlying factors contributing to homicides, we aim to provide actionable insights for law enforcement agencies, policymakers, and community organizations. Ultimately, our goal is to contribute to efforts aimed at reducing violent crime and promoting safety and well-being in Indian society.

Analysis on the murder counts in Years (2015-2021)

We have the data of total murder counts of the year 2015-2021, so we look to plot them taking the years in the abscissa and the murder counts in the ordinate axis. We obtain the plot in Fig 1.1

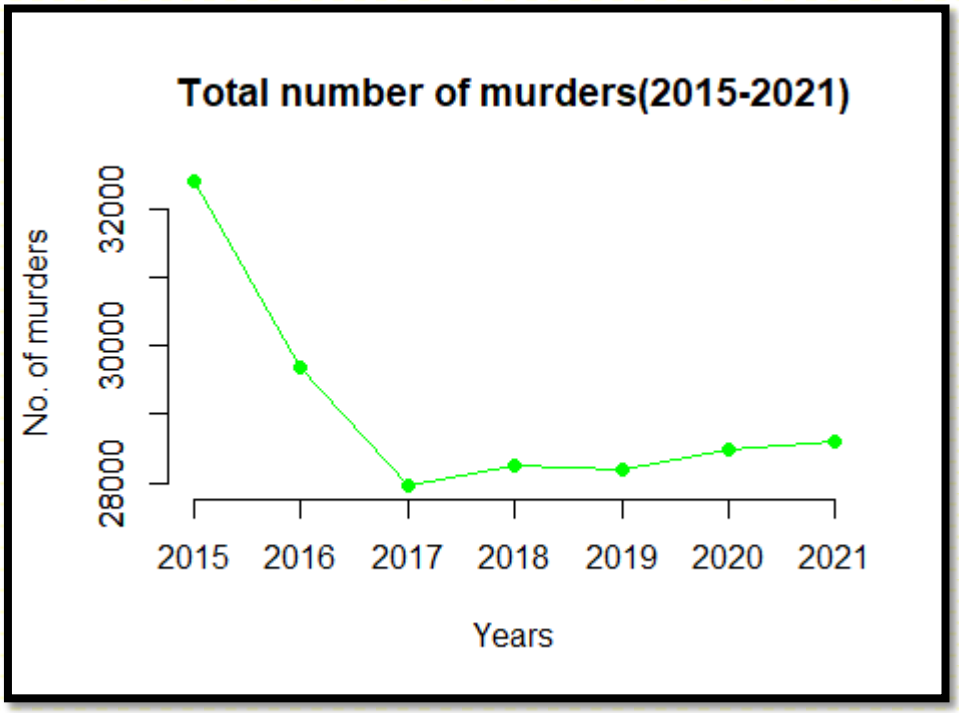

Fig 1.0 The time series plot of the murder counts

Through a cursory glance we see that the murder counts have seen a sharp decline in the period of 2015-2017 and saw its minimum at 2017. After 2017, the murder counts picked up but at a more humble rate than its previous decline in the former mentioned period. Now, we look to find a trend line to this time series plot.

1.1 Fitting a trend line

The most common trend line fitting is the linear one but the graph hints us to model it with a less common trend model but still a famous one. It would be really intuitive to model the trend of the murder counts of the period 2017-2021 with an exponential trend model to be specific the inverse exponential trend

model. The trend model fitting of the linear type, the exponential type are shown in Fig 1.1.1.

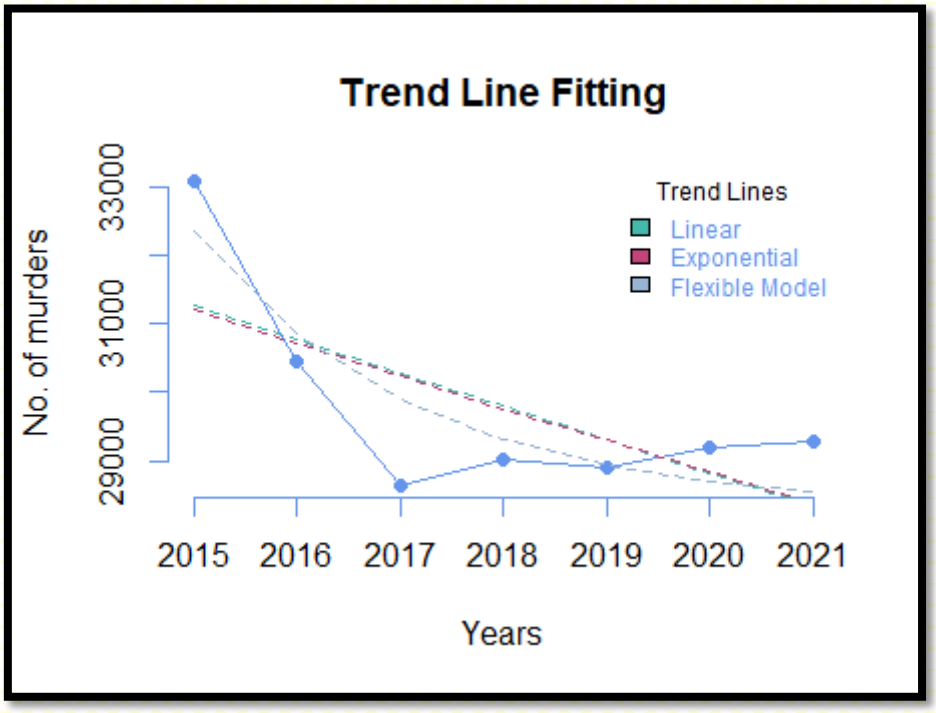

Fig 1.1.1 Fitting the linear, exponential and the inverse exponential trend models

We see that the linear trend line tries to give a representation of the trend but is not that useful, the exponential doesn't improve the situation either. Now a more flexible model is used to capture the underlying trend. The form of the fitted model is provided below:

$$murder\ counts = \alpha + \beta e^{-years}$$

The estimates of α and β are calculated using the least squares method. The estimates along with the actual fitted equation is given below:

$$\alpha = 27580 \text{ and } \beta = 1012$$

So the model becomes,

$$murder\ counts = 27580 + 1012 * e^{-years}$$

Here the years variable used is the standardized version so as to obtain a non-zero estimate of β. This model, however, quite efficient in capturing the trend has a huge downside to it, it is not interpretable. The coefficients in the model α

and β doesn't have an interpretable meaning. This is an example of the trade-off between prediction accuracy and model interpretability.

The linear trend equation is

$$murder\ counts = 1.015 * 10^{-6} - 488.2143 * years$$

And the exponential trend equation is

$$murder\ counts = e^{42.1510 - 0.0158 * years}$$

Keeping in mind the previously mentioned trade-off, we use the linear trend to interpret the trend of the murder counts. The linear fit gives us a decrease of about 488 murder counts as we go 1 year into the future from 2015. The p-value for the β test is 0.09, so we can accept the alternate hypothesis of β being non-zero at 10% level of significance based on the data provided. So murder counts have seen a yearly decline of 488 from 2015-2021 and we conclude this with sufficient statistical evidence.

1.2 Spatial Analysis

As we move deeper into the study, we look to bring out some spatial patterns or occurrences of murder in the period. Now we look to get a perspective of the murders in terms of their spatial occurrences. We take data on the state wise murder counts and look to dissect some information from it. We start by looking at a bar chart of the top 6 states in terms of murder counts in the period 2015-2021. The bar chart is provided in the Fig 1.2.1.

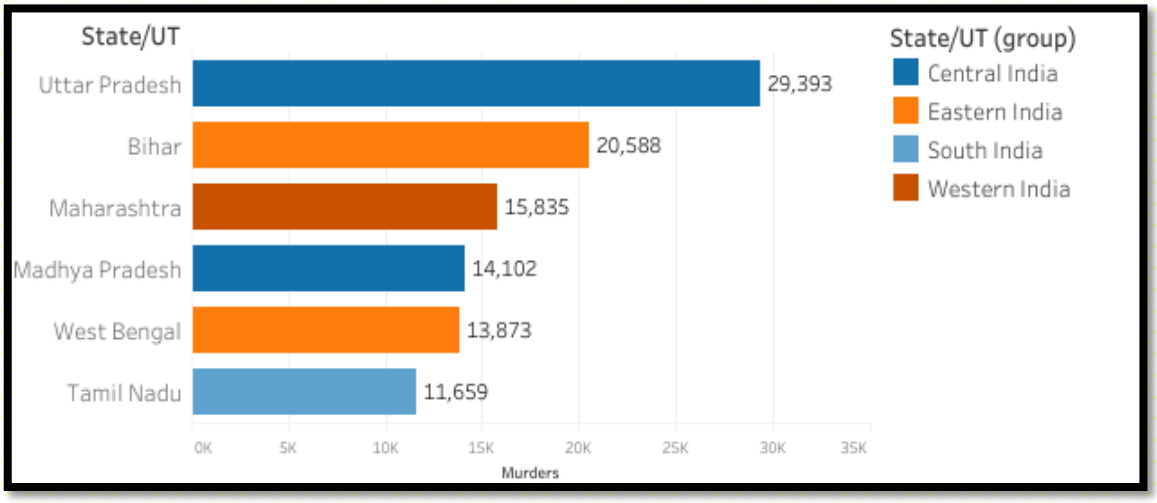

Fig 1.2.1 Top 6 States in terms of murder counts in 2015-2021

We also divided the states into six administrative blocks to get an idea of the location and culture of that state. The blocks are namely – Central, Eastern, Southern, Northern, North-Eastern and Western. We see that Uttar Pradesh has the most counts of murders in the period. There are two representatives of Central and Eastern India, and none from Northern and North-Eastern. The map with state wise murder counts is provided in Fig 1.2.2.

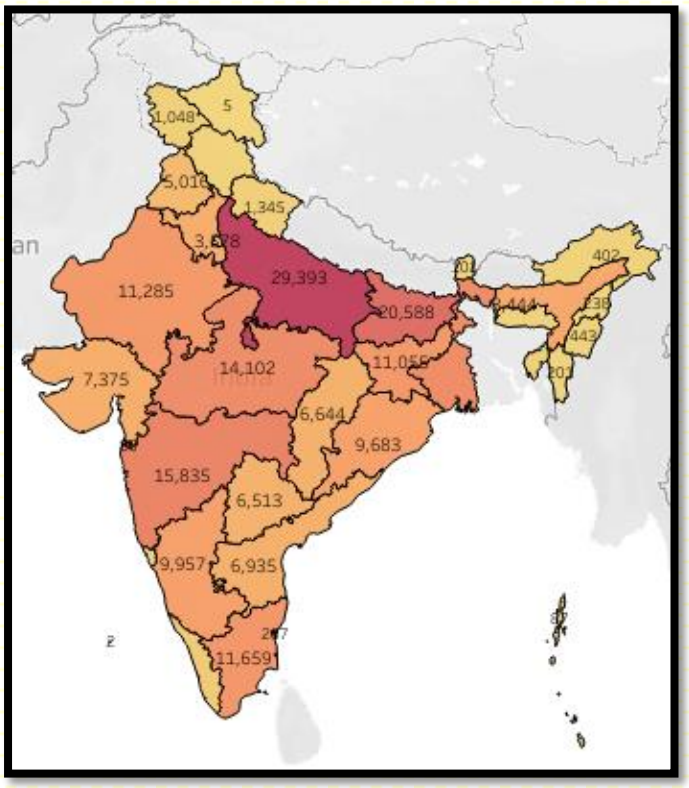

Fig 1.2.2 Map showing the state wise murder counts in 2015-2021

Through a cursory glance it might seem that states in Central and Eastern India are the most prone to murders in the period. But we are taking into account a big information, which can benefit our analysis. We must take into consideration the population of each state so as to obtain a much fairer perspective on the murder counts. As it is very logical that a state with high population is quite likely to have more murders. It doesn't imply that it is a dangerous state. We have adjusted the murder counts with the population and created a new calculated field, namely,

$$Murders\ in\ a\ state\ per\ 1000\ person = \frac{Murder\ count\ of\ state}{Population\ of\ the\ state} * 1000$$

The data on the population is collected from the crime report on India 2021. The plot so obtained is given in Fig 1.2.3.

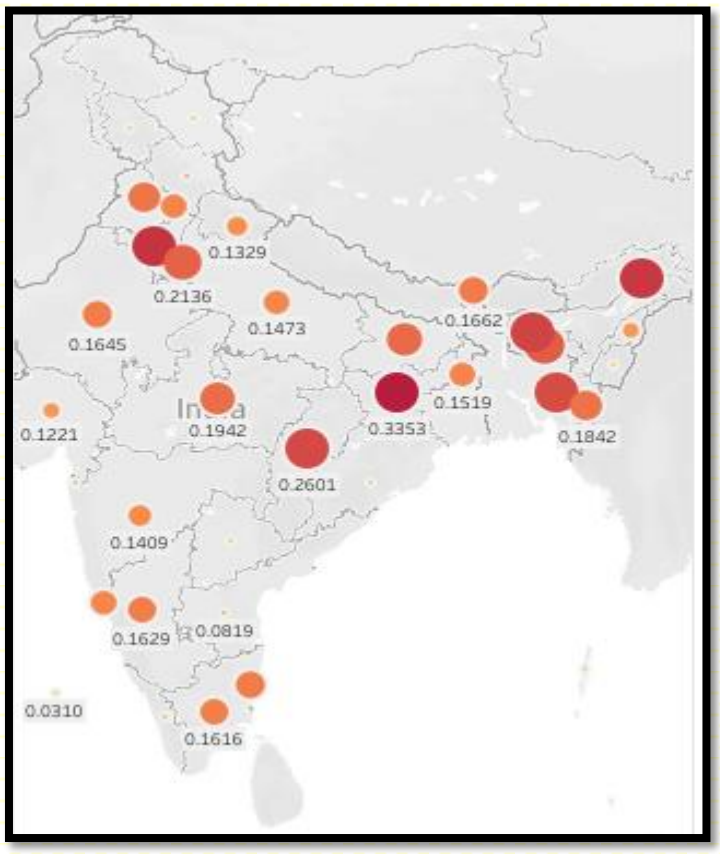

Fig 1.2.3 Map showing murders per 1000 people in 2015-2021

We see a very different scenario here after adjusting for population in our murder totals. Through visual inspection we see that there is a higher incidence of murder in the north eastern side of our country along with a few states in Central and Northern India. Let's look closer at the top 6 states with worst murder incidence per population. The bar graph is given in Fig 1.2.4. At a closer look we see that there are 3 states from the North Eastern section of our country and two from Northern and one from Central and Eastern India each. This is quite a stark contrast from the aggregate murder counts. Now a question will naturally arise, is the murder incidence higher in the North Eastern area than in the rest of India.

So to solve this query we make use of statistical inference to find an answer with adequate statistical backing to the question.

CLAIM: Is murder per population is higher in North East India than in the rest of India?

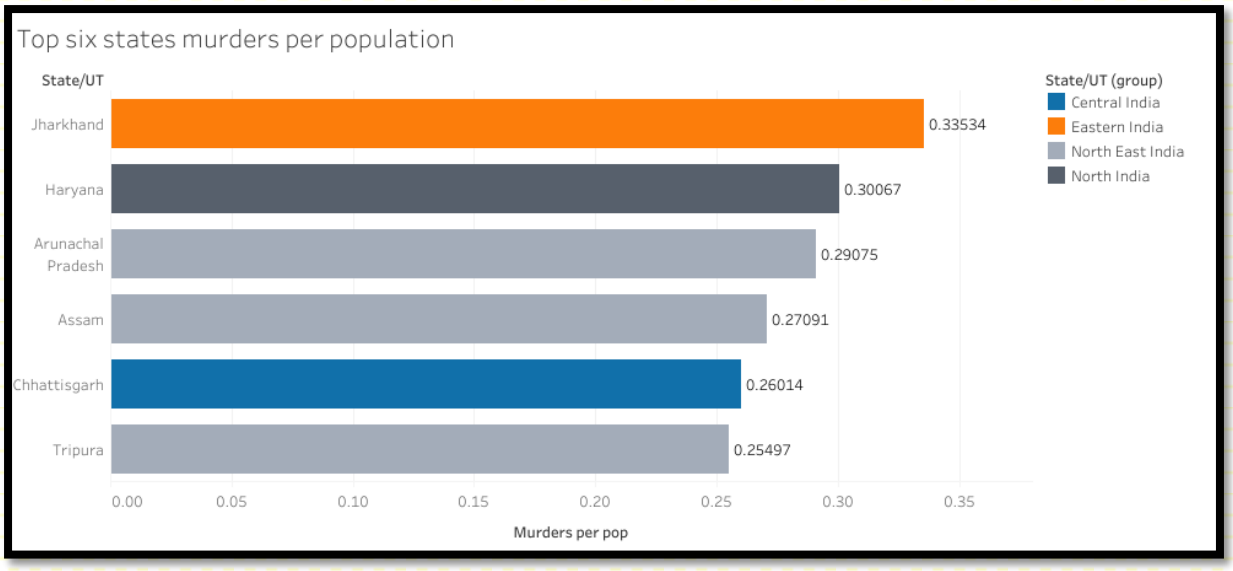

Fig 1.2.4 Top 6 states with highest murders per 1000 person

FORMULATION:

Now first we have to define properly by what we mean by "North East India". Fig 1.2.5 gives the proper partition of the states of India into the two groups.

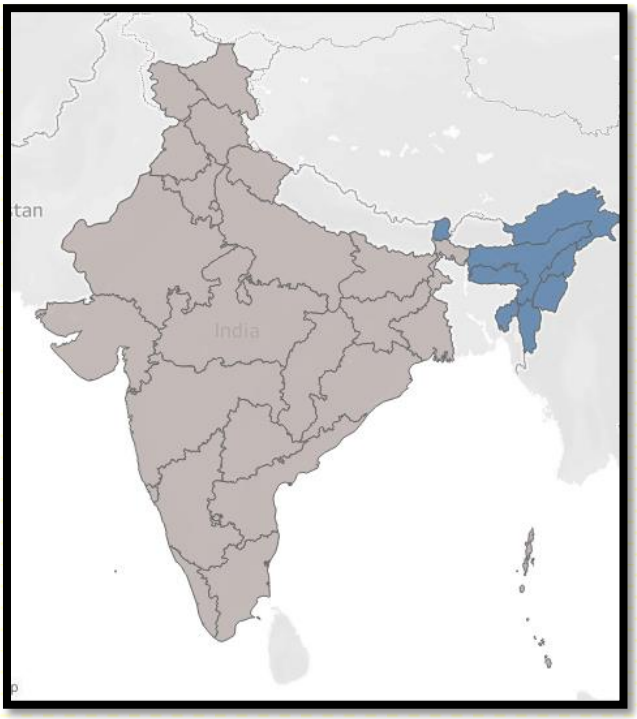

Fig 1.2.5 "North East and its neighbors" is in blue

Let us call this region NEI (North East India). Now we need to test whether the proportion of murder is higher in NEN than the Rest of India (ROI).

We do a binomial test for testing this claim. Let the murder proportion in NEI be defined by P_{NEI} and the proportion for ROI is P_{ROI}. Let's define our null hypothesis and the alternate.

Null Hypothesis :

$$H_0 : P_{NEI} = P_{ROI}$$

and the alternate is

$$H_1 : P_{NEI} > P_{ROI}$$

So this is a right tailed test. Now we apply two sample binomial test.

The test statistic for this test is

$$Z = \frac{p1-p2}{\sqrt{p(1-p)(\frac{1}{n1}+\frac{1}{n2})}},$$

Here n1 = population of NEI region, n2 = population of ROI region.

where p is the proportion of successes for the combined sample and which under null follows standard normal.

We test this at 0.05 level of significance and calculate the p-value.

$$\text{p-value} = P(Z > Z_0),$$

where Z_0 is the observed value of the test statistic.

We reject null hypothesis if p-value < 0.05 in favour of the alternate.

The observed z-statistic value is 44.23 and the p-value of this test is approximately 0 . So we conclude that based on the data we have the murder proportion in the NEI region is greater than that of ROI region at 2.5% level of significance. This null hypothesis will be accepted at any value of α in (0,100).

Through visual inspection it might seem that the NEI region has a higher incidence of murder than ROI, and there is overwhelming statistical evidence to back that claim.

Taking the same route of action, we now try to divide the States and Union Territories of India into six administrative blocks, namely, Central, Western, Eastern, Northern, Southern and North Eastern. This classification is done on cultural and geographical factors. The given classification is visually represented in the Fig 1.2.6.

Now we look to find some patterns in the various regions of India on murder counts and rates.

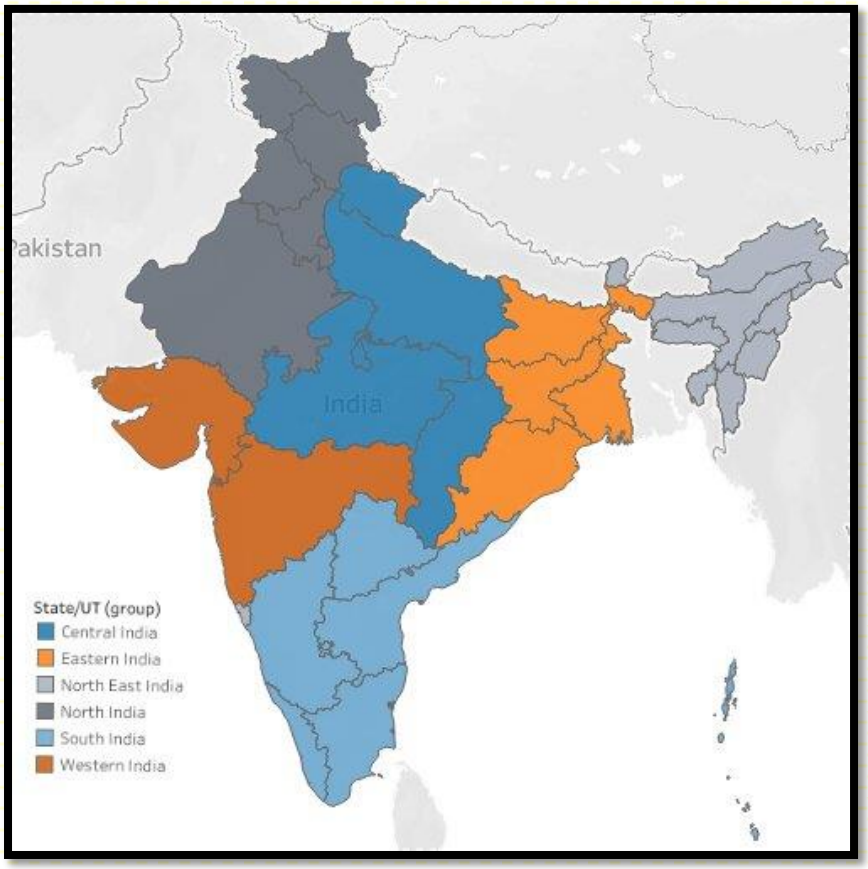

Fig 1.2.6 Six Administrative Blocks of India

The division is done based on the already made six blocks of administration. The states and UTs are divided in the following way:

Central : Chhattisgarh, Madhya Pradesh, Uttarakhand, Uttar Pradesh.

Eastern : West Bengal, Jharkhand, Odisha, and Bihar.

North Eastern : Assam, Arunachal Pradesh, Manipur, Meghalaya, Mizoram, Nagaland, Tripura, Sikkim

North : Jammu & Kashmir, Ladakh, Chandigarh, Delhi, Haryana, Punjab, Rajasthan, Himachal Pradesh.

South : Puducherry, Andhra Pradesh, Karnataka, Kerala, Tamil Nadu, Telangana.

Western : Maharashtra, Dadra and Nagar Haveli and Daman and Diu, Goa, Gujarat.

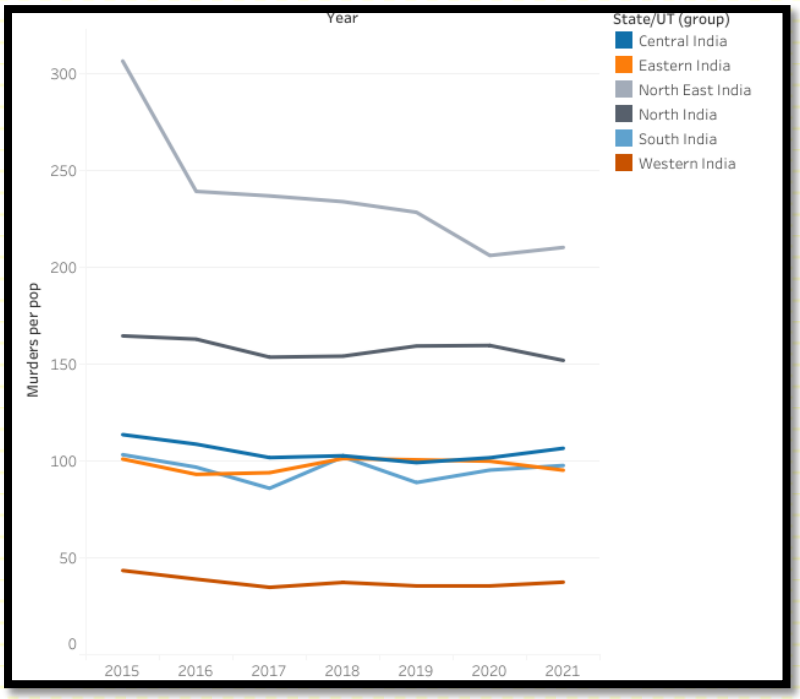

Fig 1.2.7 Plot of murders per million in blocks of India

Now the first task is to plot the murder counts for years 2015-2021 for these six administrative blocks of India.

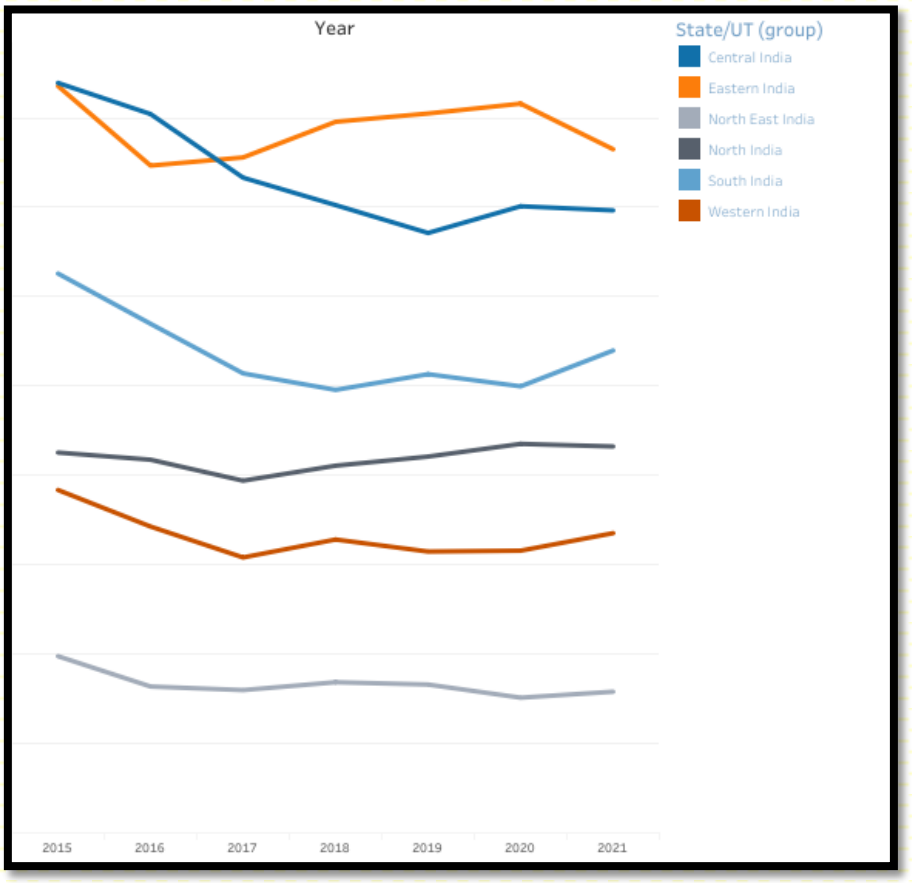

Fig 1.2.8 Murder Counts in the six regions in 2015-2021

Eastern India seems to have seen a sudden surge in murder counts from 2016, which has seen some decline after 2020. Central India has seen a gradual decrease in murder counts over the period. South India shows similar pattern as does central. North India has seen a steady increase in murder counts. Eastern India also has seen a similar steady increase.

We pose the following questions from this plot :

1. Is the murder counts in Central and Southern India correlated?
2. Is the increase of murder counts significant for Northern India?

CLAIM: There is a correlation between murder counts in Central and Southern India.

FORMULATION:

We look to test the correlation between crimes in Central and Southern India in the years 2015-2021. Here we take an assumption that the occurrences of murders in different years are independent to each other.

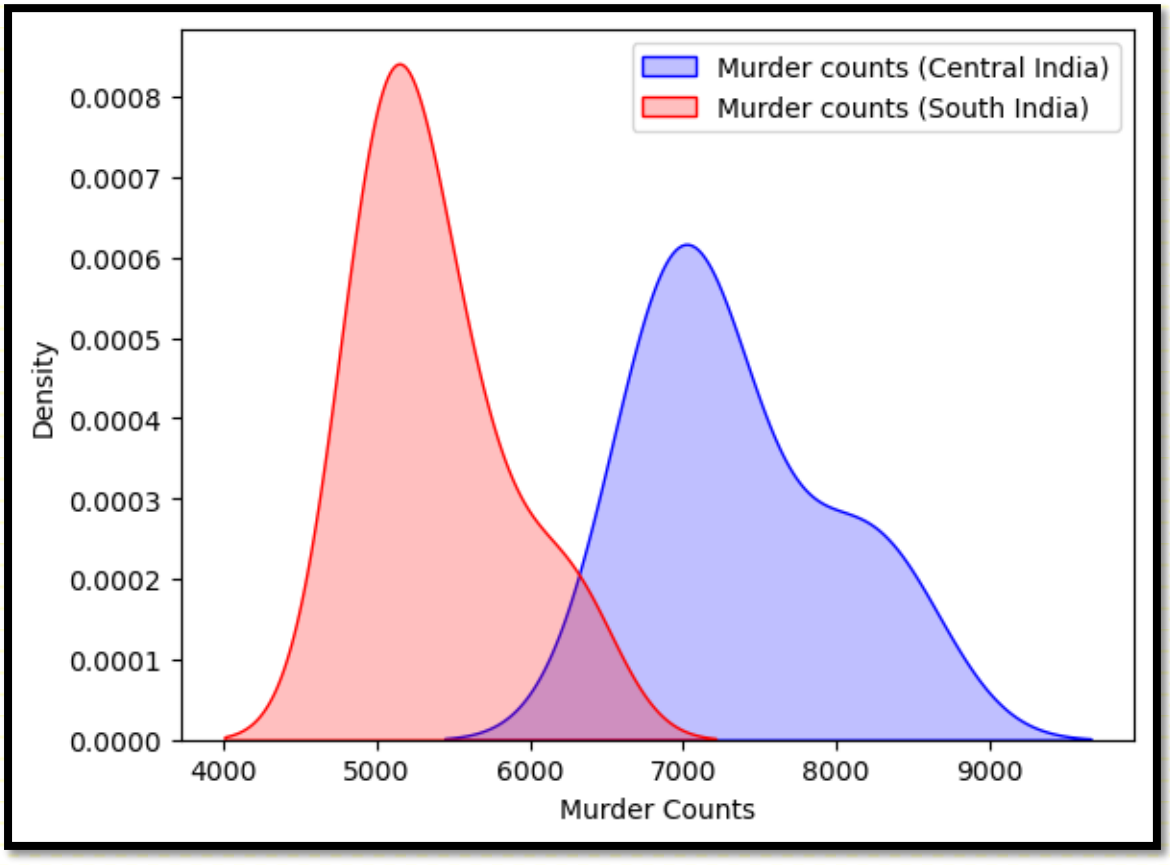

Fig 1.2.9 KDE Plot of murder counts in Central and Southern India

Let's take a look at the kernel density plot of the two regions to get an idea of the underlying distribution. If the distribution is normal or close to normal we will apply the parametric Pearson correlation coefficient test otherwise we will apply the non-parametric Spearman rank order correlation test. The plot is given in Fig 1.2.8.

From this plot we see that the distributions are quite close to normal, we will not lose much differentiating power of the Pearson Correlation Coefficient test.

We calculate the correlation coefficient of X (murder counts in Central India) and Y (murder counts in Southern India) using the formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Let the correlation coefficient between the two regions be r then our null and alternative test are

$$H_0 : r = 0$$

and

$$H_1 : r \neq 0$$

respectively. This is a two-tailed test and out test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$ which under null

follows t distribution with n-2 degrees of freedom. We take the level of significance to be 5%. Here the value of the test statistic is 0.8877 and the p-value is 0.007 < 0.05. So we reject the null hypothesis.

So we can conclude that there is a positive correlation between crimes in Central and Southern India.

CLAIM: There is a significant increase in murders in Northern India.

We look to first plot the percentage crimes in various states of North India. Some of the states with very low contributions to the total were excluded for generating graph with proper scaling. The graph is provided in Fig 1.2.9.

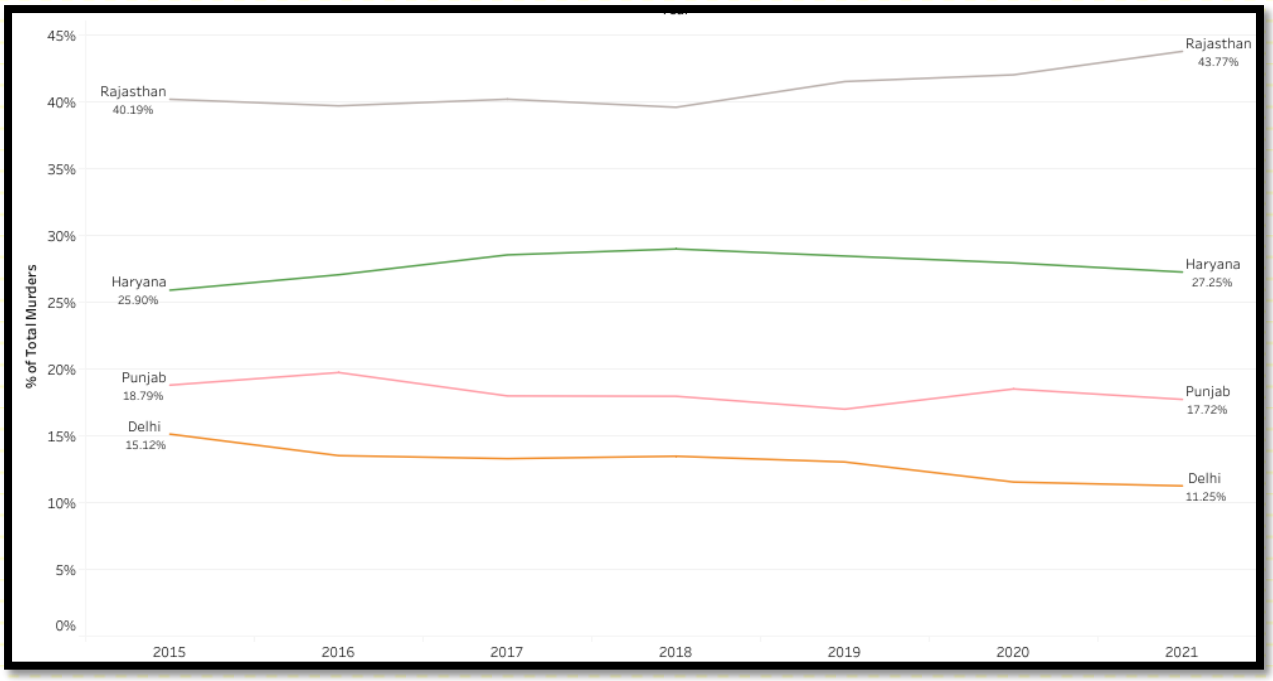

Fig 1.2.10 Percentage murders in states of North India

From the plot we see that Rajasthan is by far the most violent in terms of total murders and it has been increasing for few the past few years. The remaining states has remained almost constant. Now we check whether the increase in numbers has any statistical significance we fit a linear model to the data and test for the slope parameter. After fitting a simple linear model on the data we get the following summary:

```
Residuals:
      1       2       3       4       5       6       7
 149.00   39.43 -225.14  -86.71  -14.29   98.14   39.57

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -55479.43   52113.76  -1.065    0.336
x               29.57      25.82   1.145    0.304

Residual standard error: 136.7 on 5 degrees of freedom
Multiple R-squared:  0.2078,    Adjusted R-squared:  0.04932
F-statistic: 1.311 on 1 and 5 DF,  p-value: 0.304
```

Now as the p-value is quite high (0.304) we cannot conclude that there is significant increase in murder counts in Northern India. The trend fitting plot is given below :

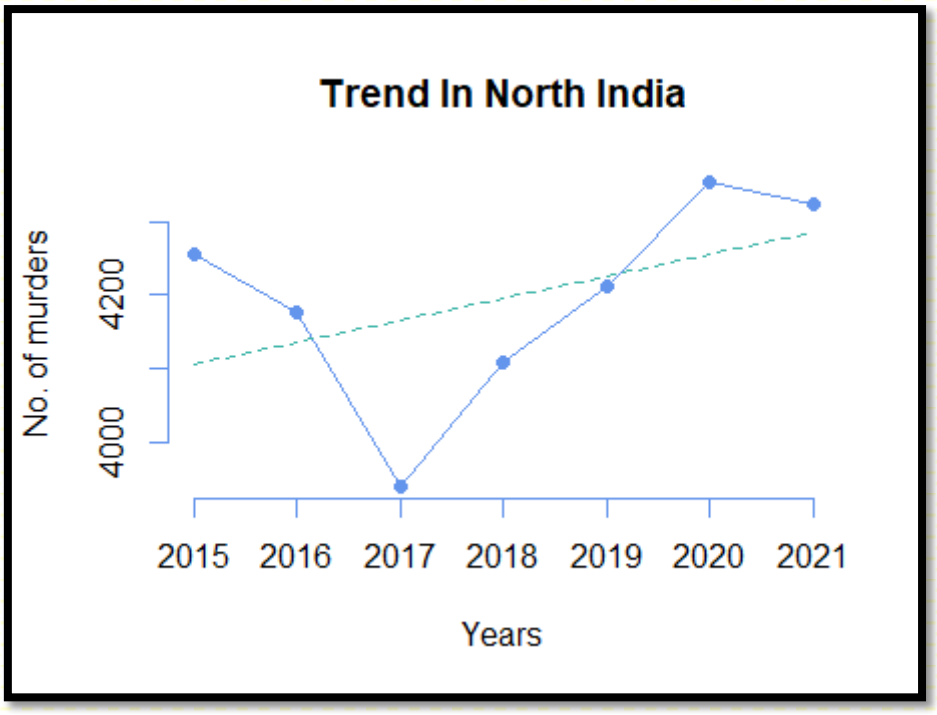

Fig 1.2.11 Trend line for North India

We see that although the hypothesis of non-zero increment is falsified, the murder counts of North India correlates not so heavily with the total India murder counts. To get a better idea of the correlations we plot a heat-map of the correlation matrix.

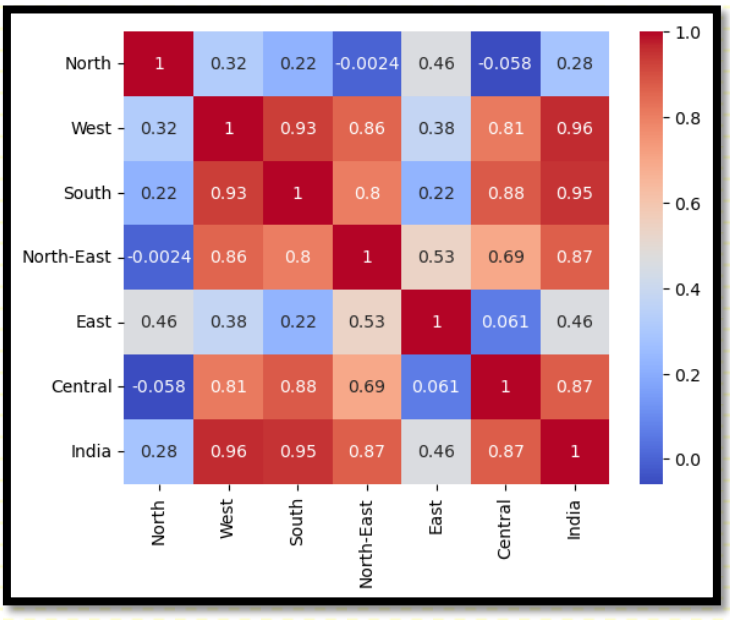

Fig 1.2.12 Heatmap of Correlation matrix

So we already have tested for the significance of correlation of crimes in Southern and Central India. Here we get a more detailed understanding of the correlations between various blocks of India. North India is the least correlated in terms of murder counts with Indian murder counts. We got a good idea about the correlations. Now let's get an idea of the variance of the murder counts. For that reason we plot the box plots.

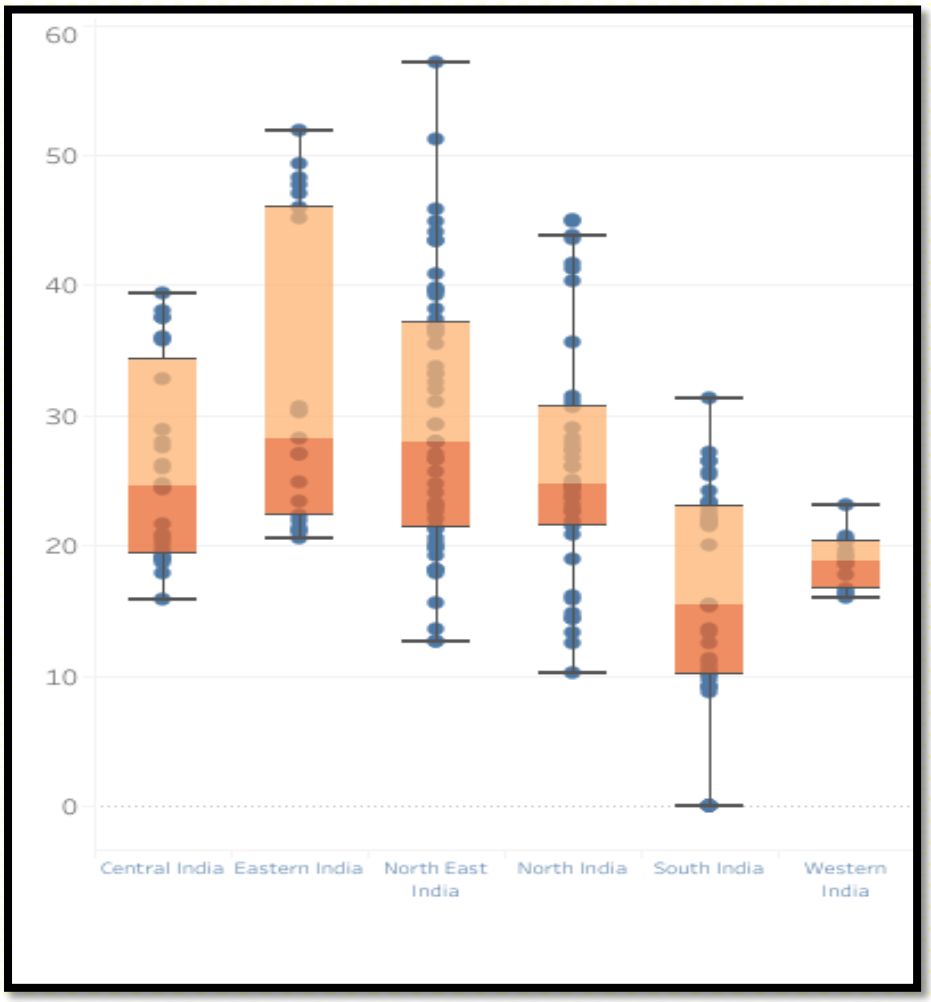

Fig 1.2.13 Region-wise box plot for murders per million

The murders per million in all the region is positively skewed. We see that North East India has the most scattered data, while western India has the least scattered data. So we can say there is more homogeneity in murder rates in the states of West India. South India has the least value of median. So South India on average is the safest region of India. We can also see that we cannot apply ANOVA testing here to check the differences in means as the groups are not homoscedastic. So we use the non-parametric equivalent of ANOVA i.e. Kruskal

Wallis test. This is a test of equality of median of the populations. The test result is as follows:

```
> kruskal.test(data,region)

        Kruskal-Wallis rank sum test

data:  data and region
Kruskal-Wallis chi-squared = 48.31, df = 5, p-value =
3.07e-09
```

We see that the p-value here is very close to 0. So we reject the null hypothesis that the median of the distributions of the different groups are same. Now for post hoc test we apply Pair-wise Wilcoxon Rank Sum Tests to see the pair wise differences and their statistical significance. The result is as follows:

```
> pairwise.wilcox.test(data, region, p.adjust.method = "bonferro
ni")

        Pairwise comparisons using Wilcoxon rank sum exact test

data:  data and region

   C     E       N     NE    S
E  0.095 -       -     -     -
N  1.000 4.4e-05 -     -     -
NE 0.048 3.7e-07 1.000 -     -
S  0.047 1.8e-10 0.255 1.000 -
W  1.000 0.025   1.000 1.000 1.000

P value adjustment method: bonferroni
```

As we see here NE is different from C, E. N is different from E and S is different from C. W is different from E. The parts, which show 1 as p-value, are those columns with ties so the p-value cannot be computed exactly.

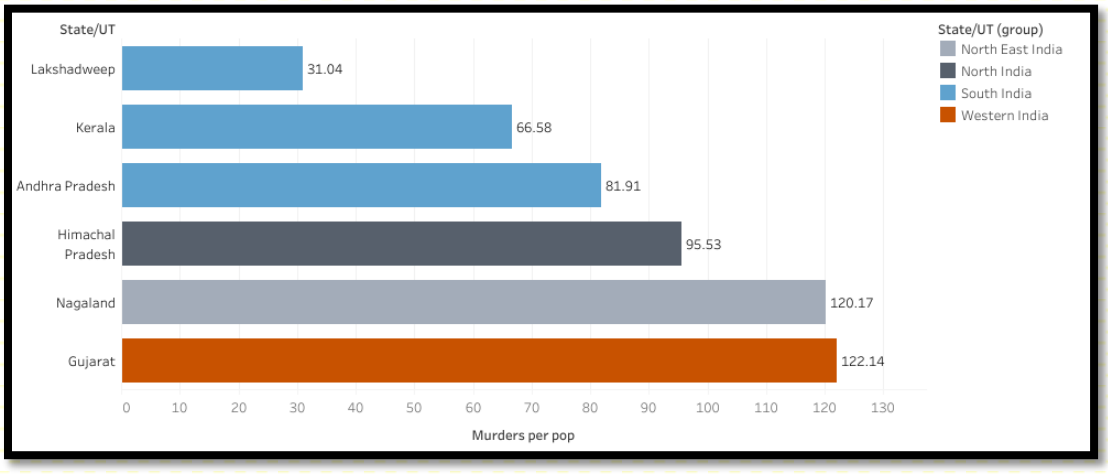

Fig 1.2.14 Top 6 states with least murders per million population

Now we just look at the top 6 states with least murders per million. We see that there are three states of South. So it also makes our claim that South India is the safest region in terms of murder incidence even stronger.

1.3 Critic

There are two concerns regarding areal spatial type of data. These issues are the sensitivity of spatial estimates to the areal units that are employed by the researcher and the effects of omitted areal units on spatial diagnosis and estimation. The first concern is known as the modifiable areal unit problem (MAUP) and the second is known as the boundary value problem. We are mainly concerned with MAUP here.

The MAUP comprises two distinct problems. The scale problem refers to the dependence of spatial correlation findings on the number of areal units into which a spatial plane is divided. A given plane may be divided into 5, 50, 500, or some other arbitrary number of polygons, with spatial autocorrelation results differing fundamentally depending on the n that is chosen. The aggregation problem refers to the dependence of spatial autocorrelation findings on the way that the spatial plane is divided into a particular set number of polygons.

Here we have taken administrative blocks divided in terms of cultural homogeneity and administrative benefits. This line of grouping might not be very effective in finding correlation between spatial units.

Motive Analysis
(2016-2021)

Motives are a major factor in murders, providing insight into the underlying causes and societal issues that lead to violent acts. Understanding these motives allows us to grasp the complex dynamics of human behavior and societal tensions. Motives can vary widely, from personal conflicts and financial desperation to ideological beliefs and psychological factors. By analyzing motives, we can identify patterns and trends in murder cases, shedding light on prevalent issues such as gang-related crime, and hate crimes.

It enables us to develop targeted interventions and preventive measures to address these underlying issues and reduce the occurrence of violent crime. Ultimately, analyzing murder motives provides valuable insights into our society's challenges and helps us devise strategies to create safer and more harmonious communities.

2.1 Feature Engineering

We have data of years 2016 to 2021 based on motives. On each of these datasets the motives have been uniquely divided into roughly 20 classes. The analysis based on all these classes will be really tedious and tough. So we have broadly classified the motives into 3 categories, namely, Gain or Greed, Love/Lust and Loath. We have excluded unintentional murders and murders due to other causes from the study as the data based on those motives were incomplete and the actual underlying motive is slippery. We look to analyze the data based on these grouped motives and find patterns and trends from them.

It was a huge task to group these motives into the following classes. The grouping was done by the judgement of the analyst. The idea of grouping these motives into the particular groups is supported by the paper given by Peter Morrell. This makes our analysis much easier and lighter. We might have lost some information in grouping the smaller groups into larger ones but that was one risk we had to take to move forward with our analysis. As the stage of grouping the motives is successfully dealt with, now we move onto the actual work of analyzing the data to discern some patterns in it.

2.2 Analysis

We first look to plot the total number of murders through different motives in the time line 2016-2021. Here the three primary motives are Greed, Loath and Love/ Lust. The plot is provided in Fig 2.2.1.

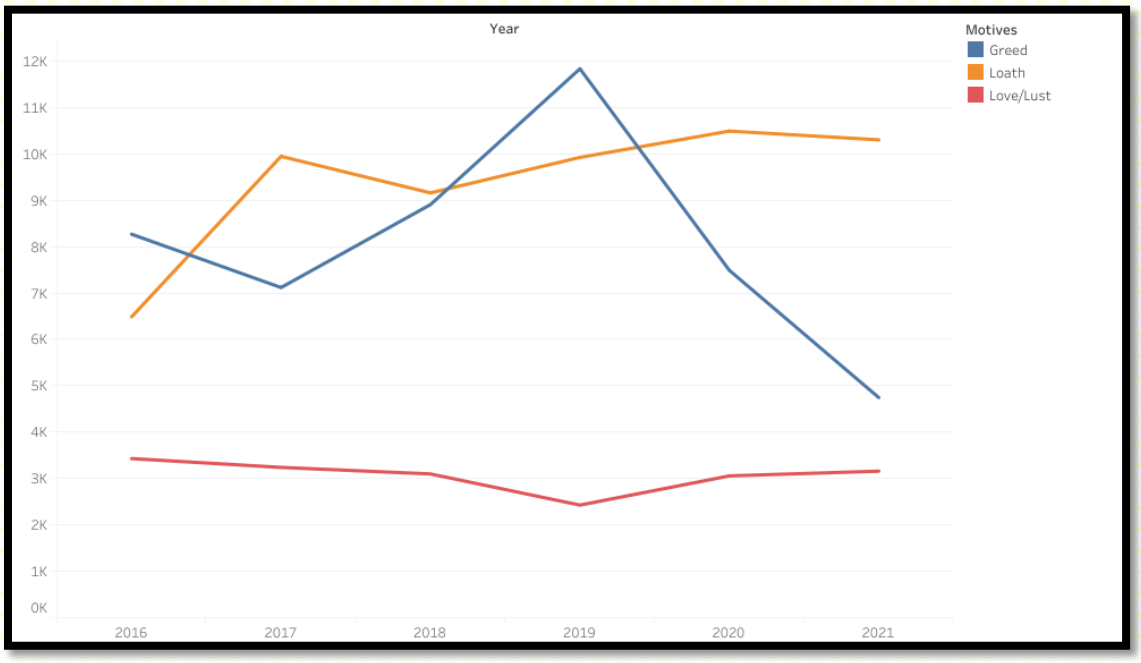

Fig 2.2.1 Time series plot of the total murders due to different motives

From this plot we see a concerning increase in the total number of murders due to loath. To give more context about these motives we provide the list of motives under the respective three primary motives in the table below.

Greed	Loath	Love/Lust
Political Causes	Disputes	Illicit Relationship
Gain	Honour Killing	Love Affairs
Robbery and Dacoity	Extremism	Rape
Sale of body parts	Casteism	
Dowry	Class Conflicts	
Property Dispute	Religious	
	Gang Rivalry	
	Communalism	

Now here we have excluded causes like Witchcraft, Lunacy, Blind Murder etc. as the murder counts in them are very low and not of much importance in our study. After making a good idea about what the respective motives are comprised of we move on with our analysis.

Next we plot a stacked bar chart of the total murders and their distribution in percentage among the three motives. The figure is provided in Fig 2.2.2

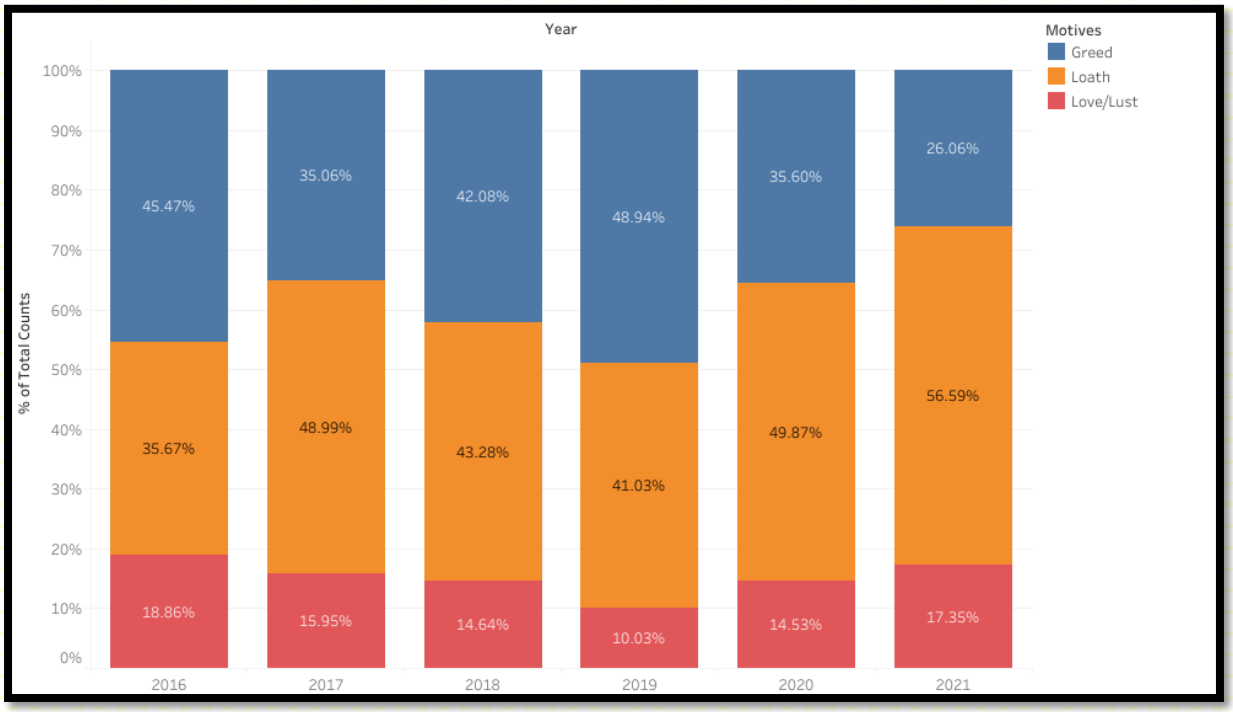

Fig 2.2.2 Stacked bar chart of murder percentage for three motives

We see that murder due to greed has declined by around 19% in the 6 year period. In 2016, majority of the murders have been due to greed at 45.47%. But after six years the landscape of Indian murder motives has changed drastically. In 2021, the majority of crimes has been attributed to Loath at 56.59%, which is a 21% increase from that of 2016. This gives us a clear indication that murders due to loath has increased in the past six years.

Now we dig deeper into the administrative blocks of India where murders due to loath has occurred and their respective percentages to the total murder counts. The plot is provided in Fig 2.2.3.

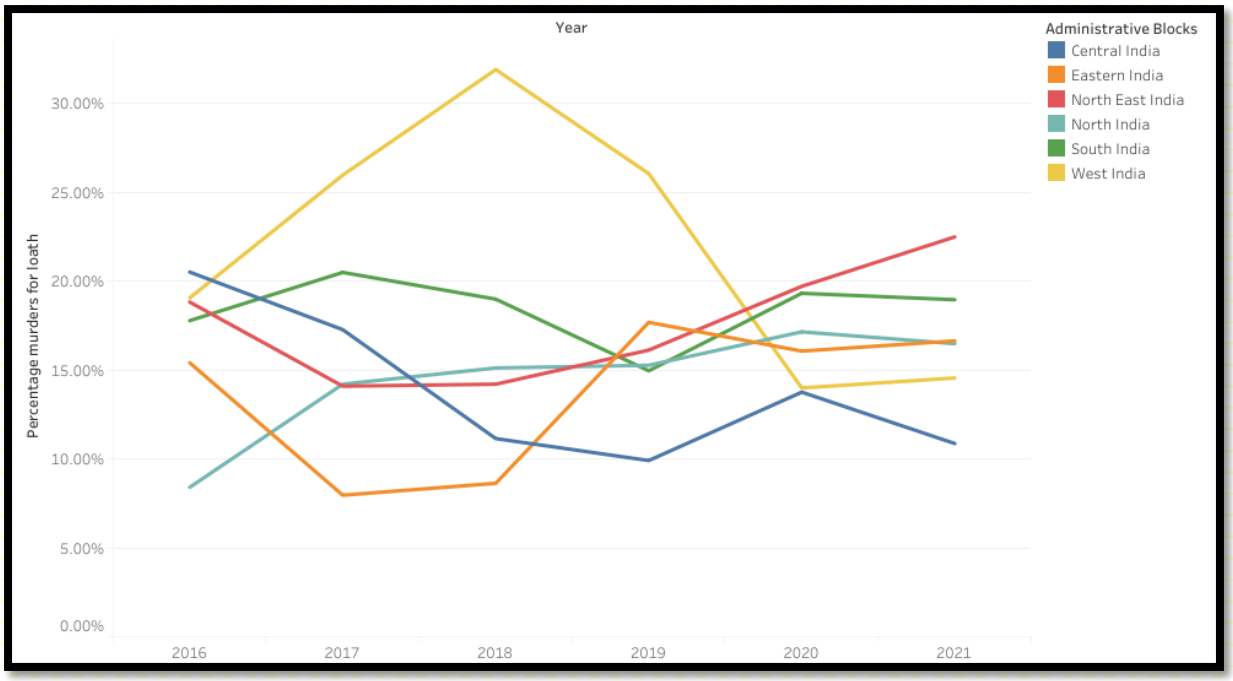

Fig 2.2.3 Percentage Murders due to Loath in the administrative blocks

From this plot we see that there has been a steady increase in the percentage of murders due to loath in North East India. While West India had an exhorbitant percentage of murders to loath in the years from 2016-2019, it has seen a decrease in percentage from the year 2019. East India has also seen a rise in percentage of murders over the years. Now we look at the distribution of motives in the six administrative blocks of India and find some interesting findings.

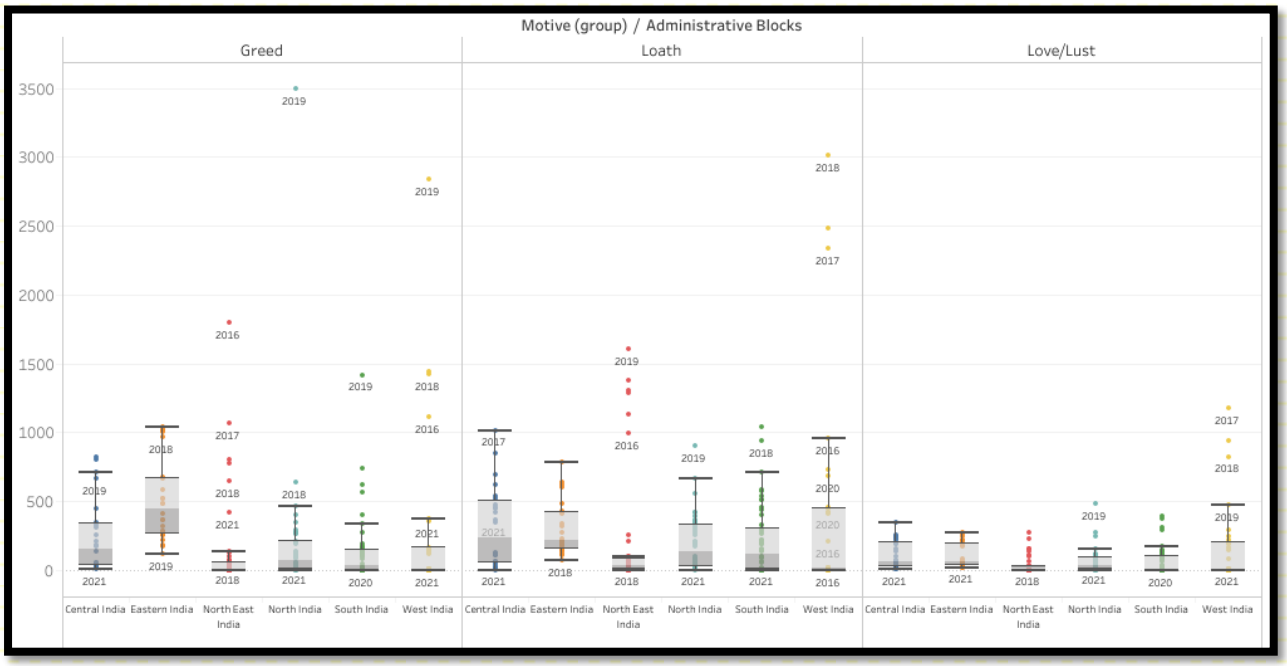

Fig 2.2.4 Box Plots of murder counts in administrative blocks for different motives

In the motive greed, West India has a lot of outliers. The years of those outlier points are also given in the plot. These are the points of Maharashtra. This similar pattern is repeated for the other motives as well. So we can safely say that Maharashtra itself is an outlier state in West India in terms of murder counts. The same can be said for Assam in North East, Rajasthan in North India and Telengana in South India. Special care must be taken for these four states as they are the standouts in their particular regions.

Eastern India has the highest median for Murder related to greed while West India has the lowest. Eastern India has high variability as well while North East India has least variability. This variability can be because of sudden changes in numbers in total which can be verified by the time series plot given above. North East has low variability in all the three motives this imply there is not much change in the murder counts over years and over states in the block, except for Assam. The high variability can also be attributed to changes in socio-economic factor as well, which will be covered later in the project.

2.3 Correlation Analysis

Now we try to look for any correlation among the various motives as this can give us some crucial information about the dynamics of the murder totals at hand. The correlation matrix between the motives is given below:

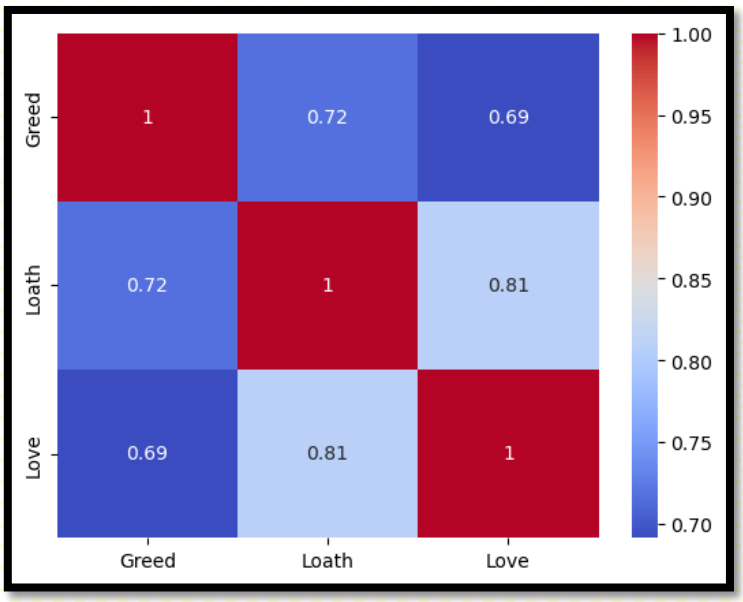

Fig 2.3.1 Correlation Matrix in a form of a heatmap

Here we see high positive correlation between all the motives. This might mean a few things. This might suggest that similar underlying factors or environments contribute to both motives. For example, areas with high economic disparity might see increases in both loath- and greed-related murders.

Shared Underlying Causes:

Economic Factors: High correlation between Greed and Loath might indicate that economic downturns lead to both types of crimes, driven by financial stress and social tensions.

Social Instability: High correlation between Loath and Love/Lust might suggest that social unrest or breakdown in family structures contributes to both motives.

Policy Implications:

Targeted Interventions: If high correlation is found between certain motives, policy measures can be designed to address multiple issues simultaneously. For instance, improving economic conditions could reduce both greed- and loath-related murders.

Resource Allocation: Understanding correlated motives can help law enforcement agencies allocate resources more efficiently. If certain areas or times are prone to multiple correlated motives, focused interventions can be implemented.

Now we take a look at how this correlation changes with time and space.

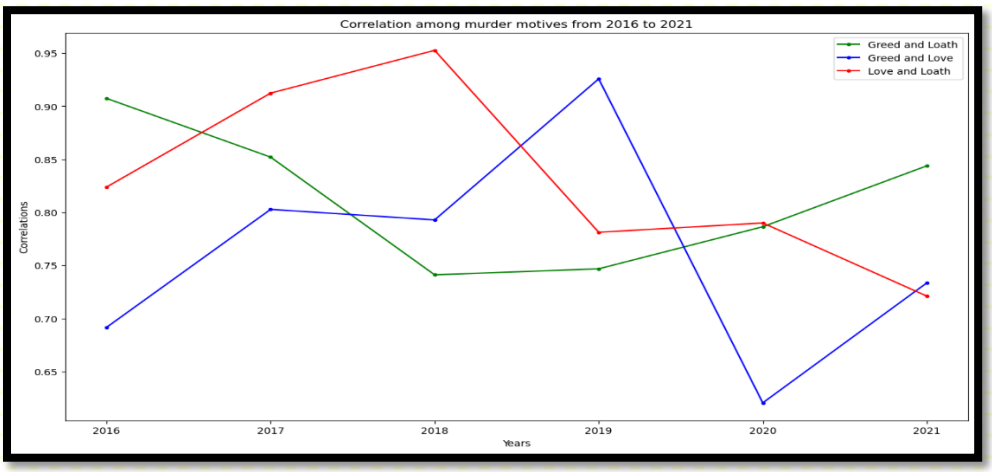

Fig 2.3.2 Correlation among motives over 2016-2021

Greed and Loath (Green Line):

The correlation starts high in 2016 (~0.90), dips to its lowest in 2018 (~0.80), and then rises again to a high point in 2021 (~0.85).

Interpretation: This indicates that in most years, there is a strong positive correlation between murders motivated by greed and those motivated by loath. This could suggest common underlying factors influencing both types of murders during these years.

Greed and Love (Blue Line):

This correlation starts relatively lower (~0.70 in 2016), peaks in 2019 (~0.93), dips significantly in 2020 (~0.65), and rises again in 2021 (~0.70).

Interpretation: The variability here might suggest different social or economic conditions affecting the correlation between these motives. The sharp dip in 2020 might reflect disruptions due to the COVID-19 pandemic, altering typical patterns of crime.

Love and Loath (Red Line):

This correlation is high in 2016 (~0.83), increases to 2018 (~0.95), then decreases steadily to its lowest point in 2021 (~0.70).

Interpretation: The initial high correlation suggests that factors leading to crimes of passion (Love) might often be associated with intense personal conflicts or hatred (Loath). The decreasing trend could indicate changing social dynamics or improved interventions in later years.

2020 Dip: Notably, Greed and Love have the lowest correlation, possibly due to the pandemic's disruption on social interactions and economic activities, leading to more isolated and motive-specific murders.

For murders due to Love/Lust, the victims are mainly women. The correlation between love and loath has decreased from 2018. Also, In 2018 Indian government introduced Criminal Law (Amendment) Act 2018, which in turn introduced punishments for rape, including the death penalty for the rape of minors under 12 years. Moreover One Stop Centres (OSCs) were established to provide integrated services to women affected by violence, such as medical aid, police assistance, and legal counseling. Universalisation of Women Helpline (181) was implemented a 24-hour helpline for women in distress across various states and union territories.

Now we take a look at the correlations among murder motives for each region. The figure is provided in Fig 2.2.6.

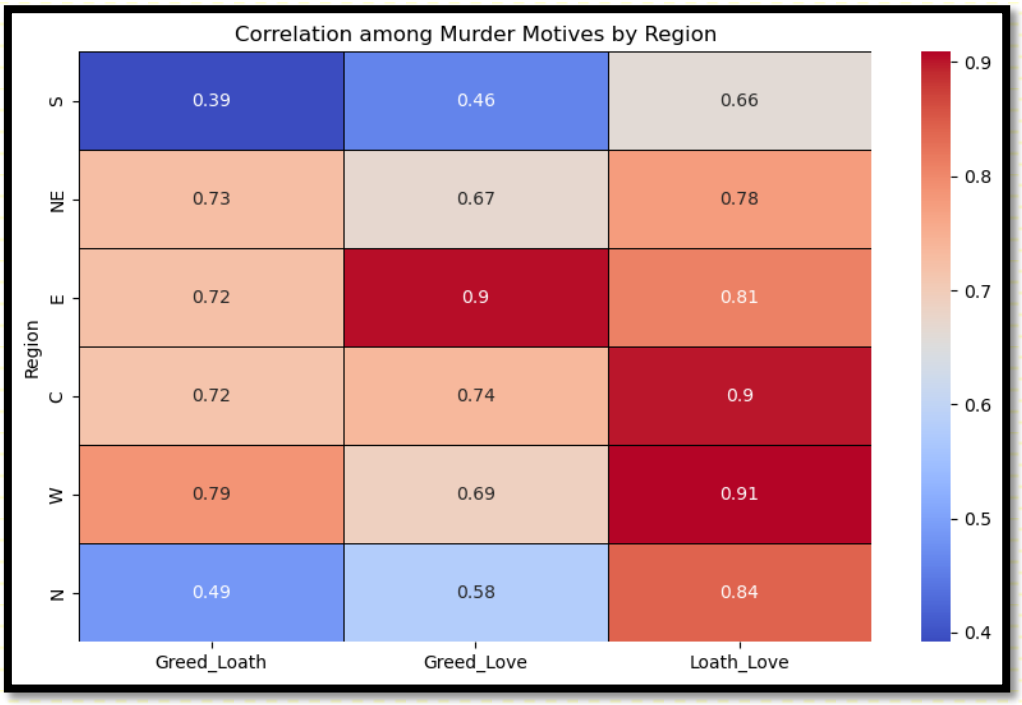

Fig 2.3.3 Correlation among murder motives by region

South (S):

Greed and Loath: Low correlation (~0.4).

Greed and Love: Low correlation (~0.4).

Loath and Love: Moderate correlation (~0.66).

Interpretation: In the South, murders driven by loath and love are more likely to be influenced by similar factors compared to murders driven by greed.

North-East (NE):

Greed and Loath: High correlation (~0.73).

Greed and Love: Moderate correlation (~0.67).

Loath and Love: High correlation (~0.78).

Interpretation: The North-East shows strong correlations across all motives, indicating that similar underlying factors might be affecting all types of murders.

East (E):

Greed and Loath: Moderate correlation (~0.72).

Greed and Love: High correlation (~0.9).

Loath and Love: High correlation (~0.81).

Interpretation: The East has high correlations for love-related motives, suggesting that economic or social conditions affecting greed might also strongly influence crimes of passion.

Central (C):

Greed and Loath: Moderate correlation (~0.72).

Greed and Love: Moderate correlation (~0.74).

Loath and Love: High correlation (~0.9).

Interpretation: Central India shows balanced correlations across all motives, suggesting a uniform influence of various factors.

West (W):

Greed and Loath: Moderate correlation (~0.79).

Greed and Love: Moderate correlation (~0.69).

Loath and Love: High correlation (~0.91).

Interpretation: Similar to Central India, the West has high correlations particularly for loath and love, indicating intertwined socio-economic factors.

North (N):

Greed and Loath: Low correlation (~0.49).

Greed and Love: Low correlation (~0.58).

Loath and Love: High correlation (~0.84).

Interpretation: In the North, loath and love correlations are strong, but greed shows lower correlations with other motives, indicating more distinct influencing factors.

Insights:

High Correlations: Loath and Love correlations are consistently high across all regions, indicating a strong connection between these motives.

Moderate Correlations: Greed and Loath show moderate correlations, suggesting some common underlying factors but also distinct differences.

Regional Variations: Different regions show varying degrees of correlation among motives, reflecting unique socio-economic and cultural factors influencing crime.

Also we observe that South India has the least correlations among all the murder motives.

2.4 Cluster Analysis

As the administrative blocks don't give a correct grouping of states, we use cluster analysis to get an idea of the correct grouping of states in terms of murder counts so that the variation within the cluster is less. We do the clustering only on the year 2021 as this was the most recent data we have. We use K-Means Clustering for this process. This is known as unsupervised learning as we are trying to find patterns from the data and we don't have a target variable for prediction.

Now the major task in K-Means Clustering is to pre define the number of clusters. This can be done through two ways, namely, using sum of squared distance and silhouette analysis. First we do the sum of squares method and look for a formation of an elbow shape in the graph.

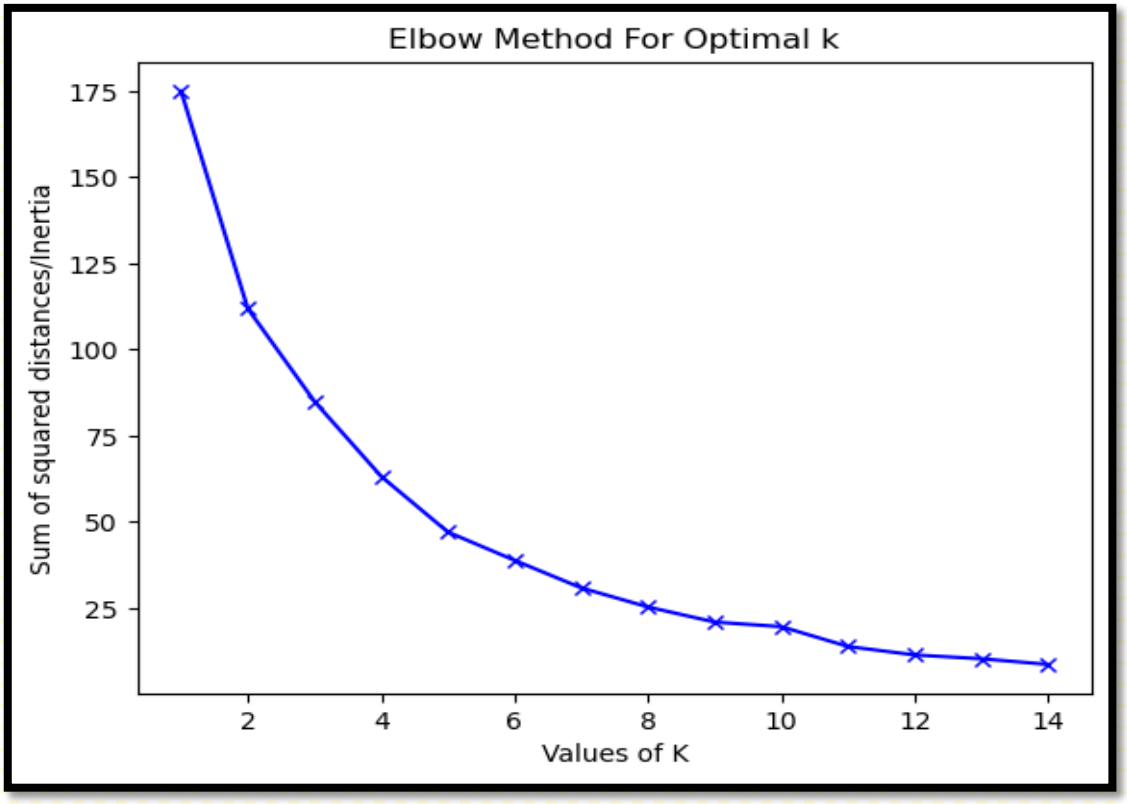

Fig 2.4.1 The sum of squared distance vs no. of clusters

Here we see that the sum of squared distance is a monotonically decreasing curve. Our goal is to find K such that it minimizes the function. But here we see

that the function is minimized at k = 14 and that many clusters are not a good idea. So we move onto the next step that is the silhouette analysis.

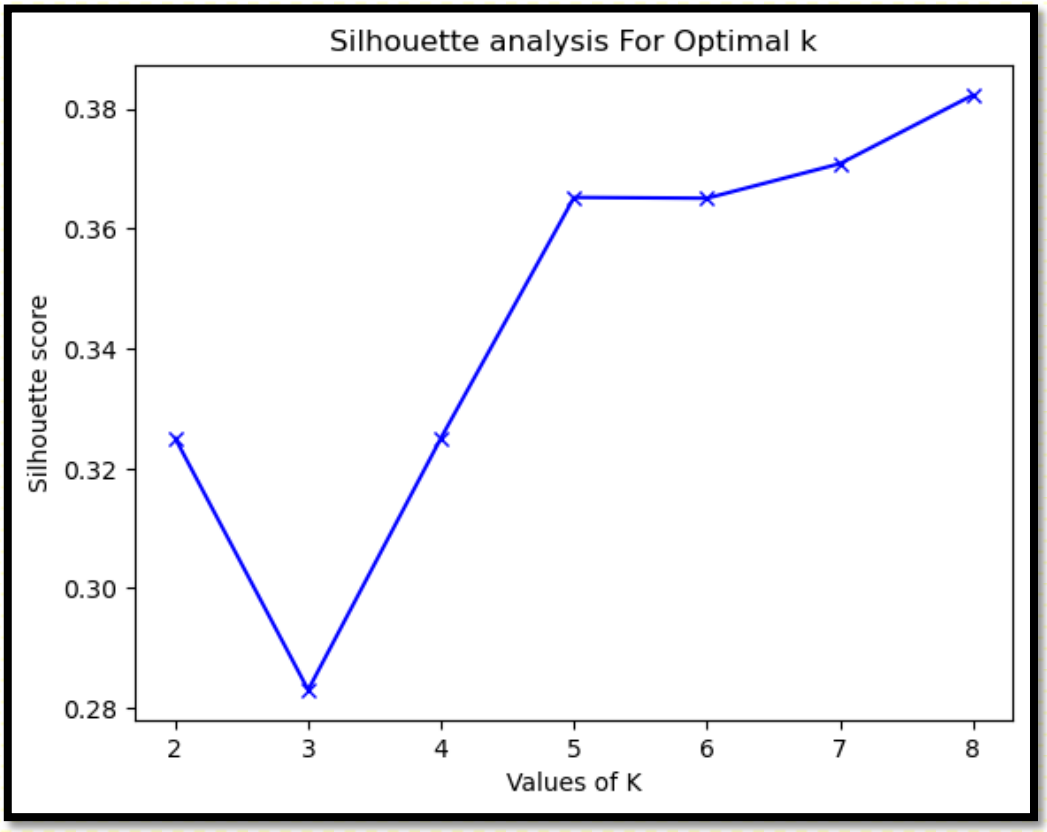

Fig 2.4.2 Silhoutte Score vs Values of K

Now here we have to pick K such that the silhouette score is maximum, but here we see that the score starts to increase monotonically after K = 5. So here we have to make a trade-off so as to pick a less accurate model to get the class number or cluster number down. So for that reason we pick K = 5, which is also close to the initial number of administrative blocks we pick.

So we pick the K-Means Clustering model and fit the data for 2021 for K = 5 clusters and get the following clusters

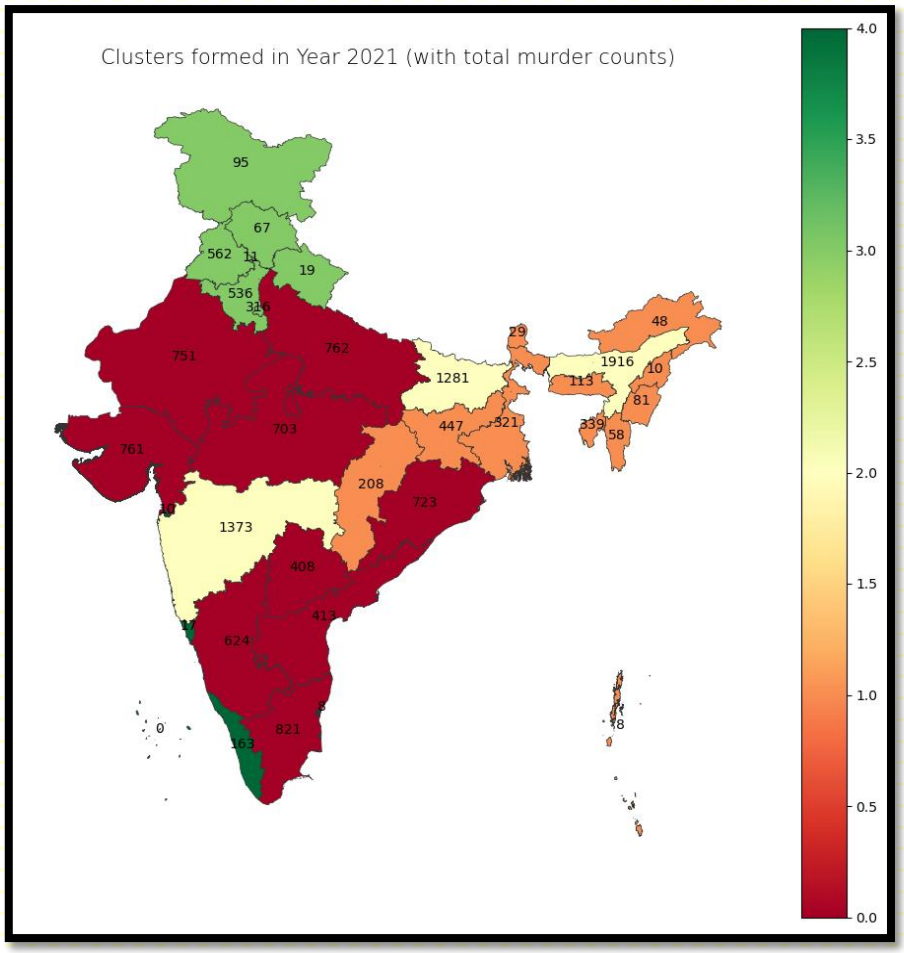

Fig 2.4.3 Clusters formed in year 2021

The map is color-coded to represent different clusters formed based on the proportions of murders motivated by Love, Loath, and Greed.

Different colors indicate different clusters, suggesting regions with similar patterns in murder motives.

States like Uttar Pradesh (762), Bihar (1281), and Maharashtra (1373) have high murder counts. They are grouped into different clusters based on the motives, indicating that high murder rates do not necessarily correlate with similar motive distributions.

Northern and Central states like Uttar Pradesh, Madhya Pradesh appear to be in one cluster, indicating similar distributions of murder motives in these areas.

Southern states like Kerala and Karnataka are grouped differently, suggesting a distinct pattern in murder motives.

States in the Northeast and some in the South have lower murder counts and are grouped into different clusters, indicating diverse motives even with fewer incidents.

Green & Orange Clusters (Low Proportion Cluster): States like Kerala and some Northeastern states show low proportions in all three motives. This could imply effective law enforcement or unique socio-cultural factors.

Red & Yellow Clusters (High Proportion Cluster): Central and Northern states with high total counts show a high proportion of murders across motives. This could suggest a broader issue with violence or less effective crime prevention strategies.

Socio-economic factors

Murders can often be understood through the lens of socio-economic factors, which influence the conditions that lead to violent behaviors. Poverty and unemployment can drive individuals to commit crimes out of desperation or as a means of survival. High population density and urbanization can increase stress and anonymity, reducing social cohesion and making it easier for crimes to occur unnoticed. Homelessness exacerbates vulnerability and instability, potentially leading to higher crime rates. Economic disparities, measured by metrics like net state domestic product, can highlight regions where socio-economic inequality fuels social tension and violence. By analyzing these factors, we can better understand the root causes of murder and design targeted interventions to mitigate them.

3.1 Identifying the factors

Screening through various articles provided in the internet, we narrowed down some socio-economic factors, they can be mildly correlated with each other. The socio-economic factors so found are as follows: income levels, unemployment rate, poverty rate, population density, age distribution, urbanization, literacy rates, education attainments, migration patterns, mental health problems, rates of alcohol and drug use, judicial efficiency, religious composition and governance quality. Now the process of data collection was performed, due to various constraints we were unable to collect data on age distribution, education attainment, migration patterns, mental health problems, judicial efficiency, religious composition and governance quality. We now look to see how these socio-economic factors have a causal relationship to murders, their motives and also how the patterns vary in the various blocks of India.

3.2 Analysis

First we see the distribution of the various socio-economic factors. The plot is given below:

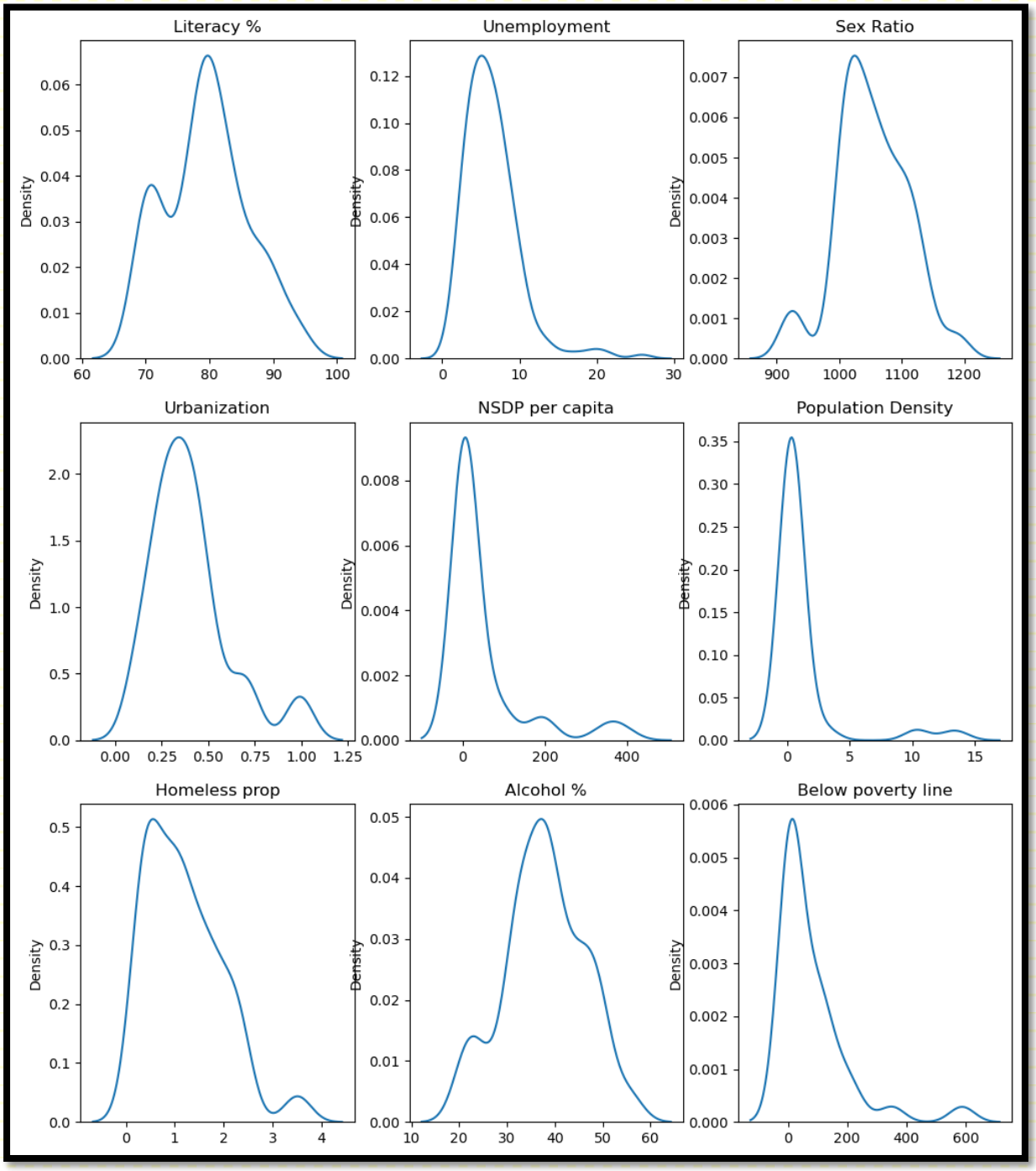

Fig 3.2.1 Distribution of the various socio-economic factors

Literacy Percentage: Most states have a literacy rate clustered around 70-80%, with fewer states having very high or very low literacy rates.

Unemployment: The distribution shows a peak at low unemployment rates between 5 and 10, indicating that most states have relatively low unemployment, with a long tail towards higher unemployment rates.

Sex Ratio: This shows a peak around the 950-1050 range, indicating most states have a sex ratio close to this value, with some outliers at both higher and lower ends.

Urbanization: The urbanization rate distribution shows a concentration around the lower percentages, indicating many states have low to moderate levels of urbanization.

NSDP per Capita: The distribution of net state domestic product per capita shows most states have a low to moderate NSDP per capita, with a long tail indicating a few states with significantly higher values.

Population Density: This shows a sharp peak at very low values, suggesting most states have low population density, with a few states experiencing very high density.

Homeless Population Proportion: Most states have a low proportion of the homeless population, with the distribution decreasing as the proportion increases.

Alcohol Consumption Percentage: The majority of states have alcohol consumption percentages around 30-40%, with fewer states at the extremes.

Below Poverty Line Population: The distribution peaks at low values, indicating most states have a lower proportion of the population below the poverty line, with a long tail toward higher proportions.

The socio-economic factors generally show that most states fall within moderate ranges for these metrics, with a few states experiencing extremes.

Literacy rates, unemployment, and urbanization tend to be concentrated around central values, suggesting similar socio-economic conditions across many states.

High skewness in factors like population density, NSDP per capita, and below poverty line indicates substantial disparities between states.

Understanding these distributions is crucial for analyzing how these factors may correlate with murder rates and motives across different regions in India.

This analysis provides a foundation for exploring correlations between socio-economic factors and crime, aiding in identifying areas that might require targeted interventions.

Also we see that there is no socio-economic factor that closely resembles the normal distribution. Now we look to test the significance of correlations among the factors and the murder counts.

We provide below the heatmap which shows the correlations between the row and column variable after testing for their significance, the zeroes signify that it failed the correlation test. This gives us an idea about how various factors and murder counts are related.

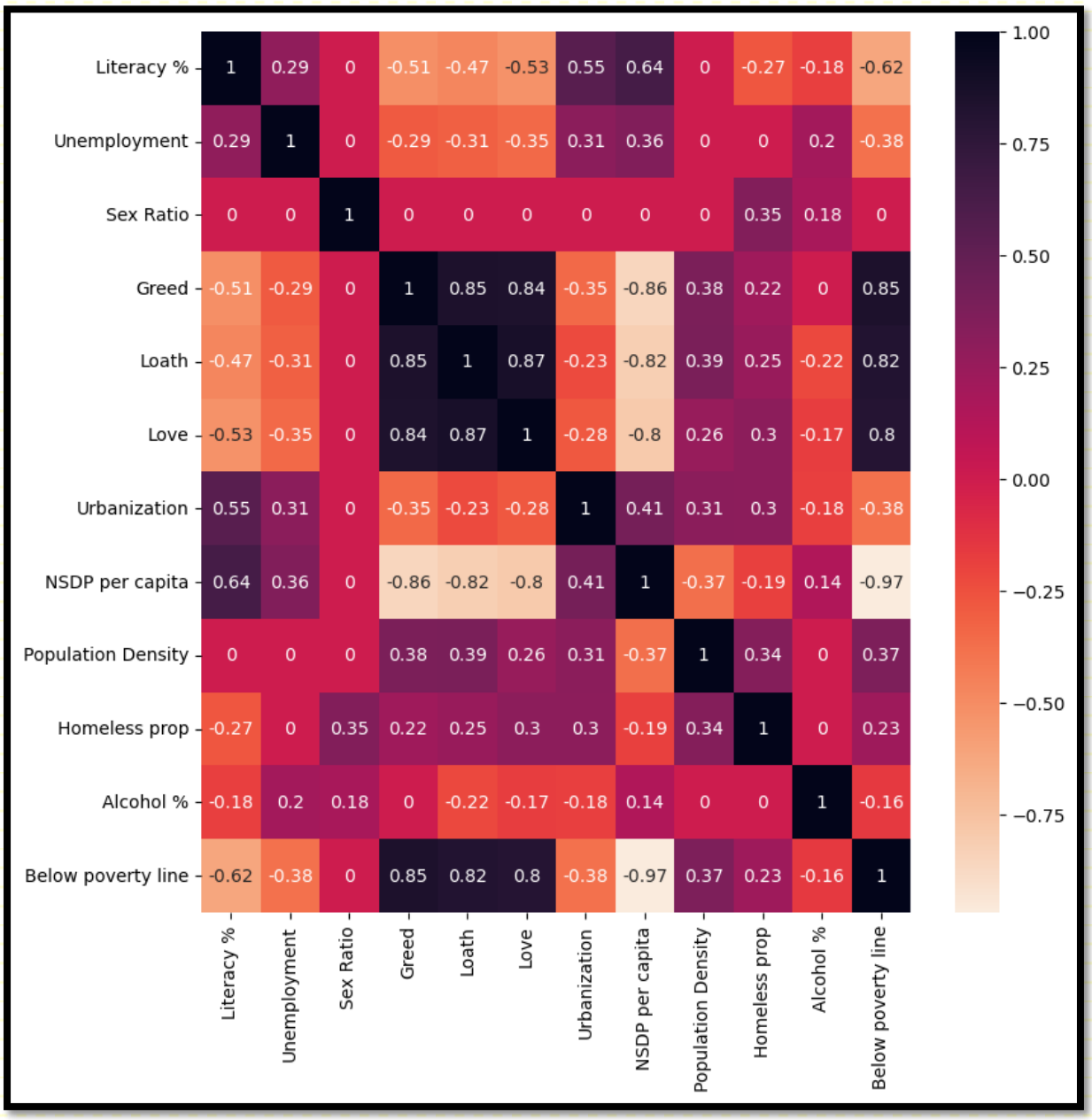

Fig 3.2.2 Correlation matrix after testing for significance

Here we see that barring sex ratio every other socio-economic factor has a significant correlation with murder counts of different motives. Now we see that there is significant correlation among the various factors as well so this also pushes the concern for multicollinearity. Now we look specifically how different factors affect the different motives of murders. First we need to check for multicollinearity.

For checking multicollinearity we use VIF for each variable. The interpretation of VIF is as follows:

VIF = 1 No correlation between the predictor and other predictors.

1 < VIF< 5 Moderate correlation, generally acceptable.

VIF > 5 High correlation, potential multicollinearity issue.

VIF > 10 Very high correlation, serious multicollinearity issue.

By examining the VIF values, you can decide if you need to address multicollinearity in your regression model.

Now the VIF for the variables are given below:

	Feature	VIF
0	const	1203.639753
1	Literacy %	2.656863
2	Unemployment	1.400329
3	Sex Ratio	2.300315
4	Urbanization	5.873914
5	NSDP per capita	1.927570
6	Population Density	3.400235
7	Homeless prop	2.917024
8	Alcohol %	1.416049
9	Below poverty line	1.516056

Fig 3.2.3 VIF scores

We see that Urbanization has a VIF of 5.87 i.e. close to 6 so we need to investigate the dependence of the variable with the other variables. For that we employ a linear model with all the variables and a model without the Urbanization variable and to a F test whether both the models are similar or different.

We use regularization technique to deal with this problem. We use lasso regression and see which factors are responsible for which types of murders and to what extent. We plot the lasso regression estimates of the various factors for the three motives and find the insights provided by them.

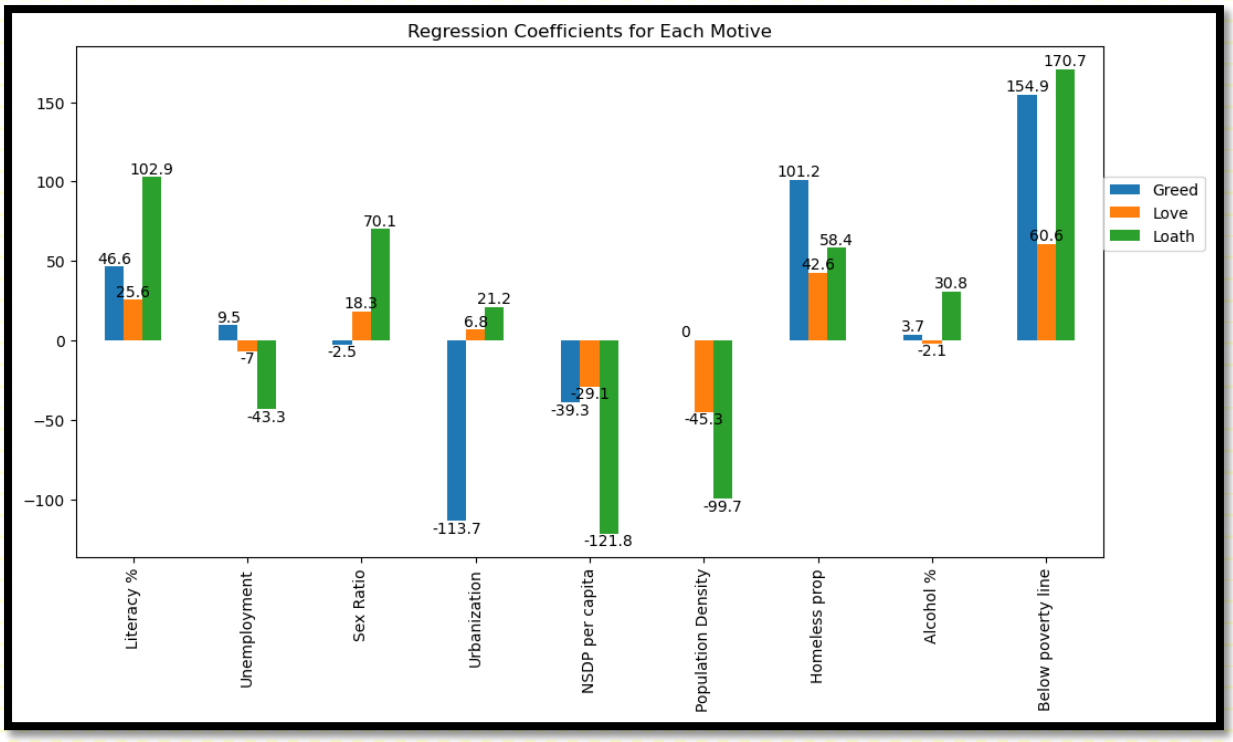

Fig 3.2.4 Lasso Regression estimates

Literacy %:

Greed: A significant positive coefficient (~46.6) suggests that higher literacy rates are associated with more murders due to greed.

Love: A moderate positive coefficient (~25.6) suggests a similar trend, but weaker compared to greed.

Loath: A very high positive coefficient (~102.9) suggests that literacy significantly increases murders due to loath.

Unemployment:

Greed: A small positive coefficient (~9.5) suggests a slight increase in murders due to greed with higher unemployment rates.

Love: A negligible coefficient (~-7) suggests a slight decrease in murders due to greed with higher unemployment rates.

Loath: A negative coefficient (~-43.3) indicates a moderate decrease in murders due to loath with higher unemployment rates.

Sex Ratio:

Greed: A negative coefficient (~-2.5) suggests a slight decrease in murders due to greed with a higher sex ratio.

Love: A small positive coefficient (~18.3) suggests a slight increase in murders due to love with a higher sex ratio.

Loath: A very high positive coefficient (~70.1) suggests that a higher sex ratio significantly increases murders due to loath.

Urbanization:

Greed: A very large negative coefficient (~-113.7) suggests that higher urbanization significantly decreases murders due to greed.

Love: A small negative coefficient (~6.8) suggests a slight increase in murders due to love with higher urbanization.

Loath: A small positive coefficient (~21.2) suggests a slight increase in murders due to loath with higher urbanization.

NSDP per capita:

Greed: A moderate negative coefficient (~-39.3) suggests that higher per capita NSDP is associated with fewer murders due to greed.

Love: A negligible coefficient (~-29.1) suggests that higher per capita NSDP is associated with fewer murders due to love.

Loath: A strong negative coefficient (~-121.8) suggests that higher per capita NSDP significantly reduces murders due to loath.

Population Density:

Greed: A negligible coefficient (~0) suggests no significant relationship.

Love: A negative coefficient (~-45.3) suggests that higher population density reduces murders due to love.

Loath: A strong negative coefficient (~-99.7) suggests that higher population density significantly reduces murders due to loath.

Homeless proportion:

Greed: A significant positive coefficient (~101.2) suggests that a higher proportion of homeless people is associated with more murders due to greed.

Love: A moderate positive coefficient (~42.6) suggests a similar trend for murders due to love.

Loath: A strong positive coefficient (~58.4) suggests that a higher proportion of homeless people significantly increases murders due to loath.

Alcohol %:

Greed: A small positive coefficient (~3.7) suggests a slight increase in murders due to greed with higher alcohol consumption.

Love: A negligible negative coefficient (~-2.1) suggests no significant relationship.

Loath: A small positive coefficient (~30.8) suggests a slight increase in murders due to loath with higher alcohol consumption.

Below poverty line:

Greed: A very strong positive coefficient (~154.9) suggests that a higher percentage of the population below the poverty line is associated with significantly more murders due to greed.

Love: A strong positive coefficient (~60.6) suggests a similar trend for murders due to love.

Loath: An extremely high positive coefficient (~170.7) suggests that poverty is a very strong predictor of murders due to loath.

So all in all, poverty (Below poverty line) is a strong predictor of murders across all motives, particularly for loath. Urbanization shows a strong negative relationship with murders due to greed but a positive relationship with loath. Homeless proportion and Literacy % are also significant predictors for murders due to all motives, with varying degrees of influence. NSDP per capita and Population Density generally have a negative relationship with murders, particularly for greed and loath. Unemployment, Sex Ratio, and Alcohol % show smaller and more varied impacts across different motives.

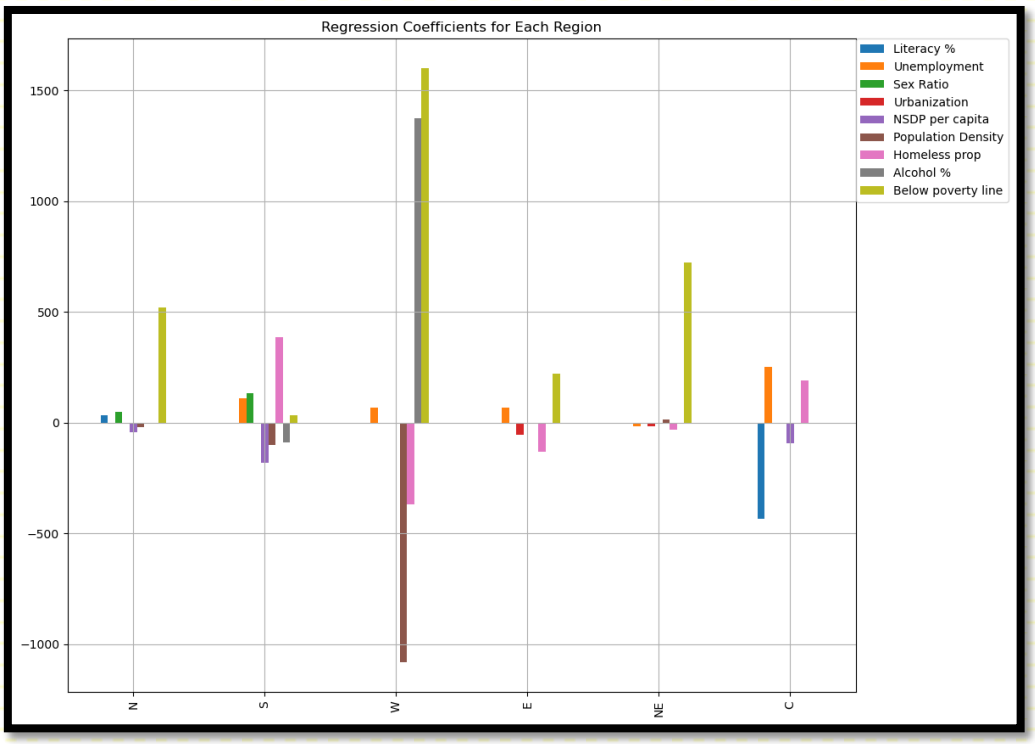

Fig 3.2.5 Lasso Regression estimates for regions

Now we look at the murder rates over the different regions and their relations with the various socio economic variables. We plot the estimated coefficients corresponding to each variable in the six blocks of India. We also provide the table for extra readability

	N	S	W	E	NE	C
Literacy %	33.7	-0.0	0.0	-2.7	-0.0	-434.4
Unemployment	0.0	111.9	68.9	68.9	-15.5	252.5
Sex Ratio	50.4	131.7	-0.0	0.0	-0.0	0.0
Urbanization	0.0	0.0	0.0	-55.8	-14.5	-0.0
NSDP per capita	-41.2	-181.3	-0.0	-0.0	-0.0	-94.0
Population Density	-21.6	-100.2	-1081.8	0.0	13.8	0.0
Homeless prop	-0.0	385.1	-369.3	-129.8	-32.0	192.1
Alcohol %	-2.0	-90.0	1374.8	0.0	0.0	-0.0
Below poverty line	518.2	32.1	1599.7	221.4	721.3	0.0

Below Poverty Line: This variable shows strong positive coefficients across most regions, particularly in the West (1597.9), Central (721.3), and North (518.2), indicating that higher poverty levels are strongly associated with higher murder rates in these regions.

Unemployment: In the Central region, unemployment has a very high positive coefficient (252.5), suggesting a significant relationship with murder rates. The South also shows a notable positive relationship (111.9).

Population Density: This variable shows a strong negative coefficient in the West (-1081.8), indicating that higher population density is associated with lower murder rates in this region.

Homeless Proportion: In the South, the homeless proportion has a high positive coefficient (385.1), suggesting a significant relationship with murder rates.

Urbanization: Has a significant negative coefficient in the East (-55.8), suggesting higher urbanization is associated with lower murder rates in this region.

NSDP per Capita: Shows strong negative coefficients in the North (-41.2) and South (-181.3), suggesting higher economic development is associated with lower murder rates.