

**CRIMES AGAINST  
WOMEN IN INDIA:  
PREDICTIVE  
MODELING**





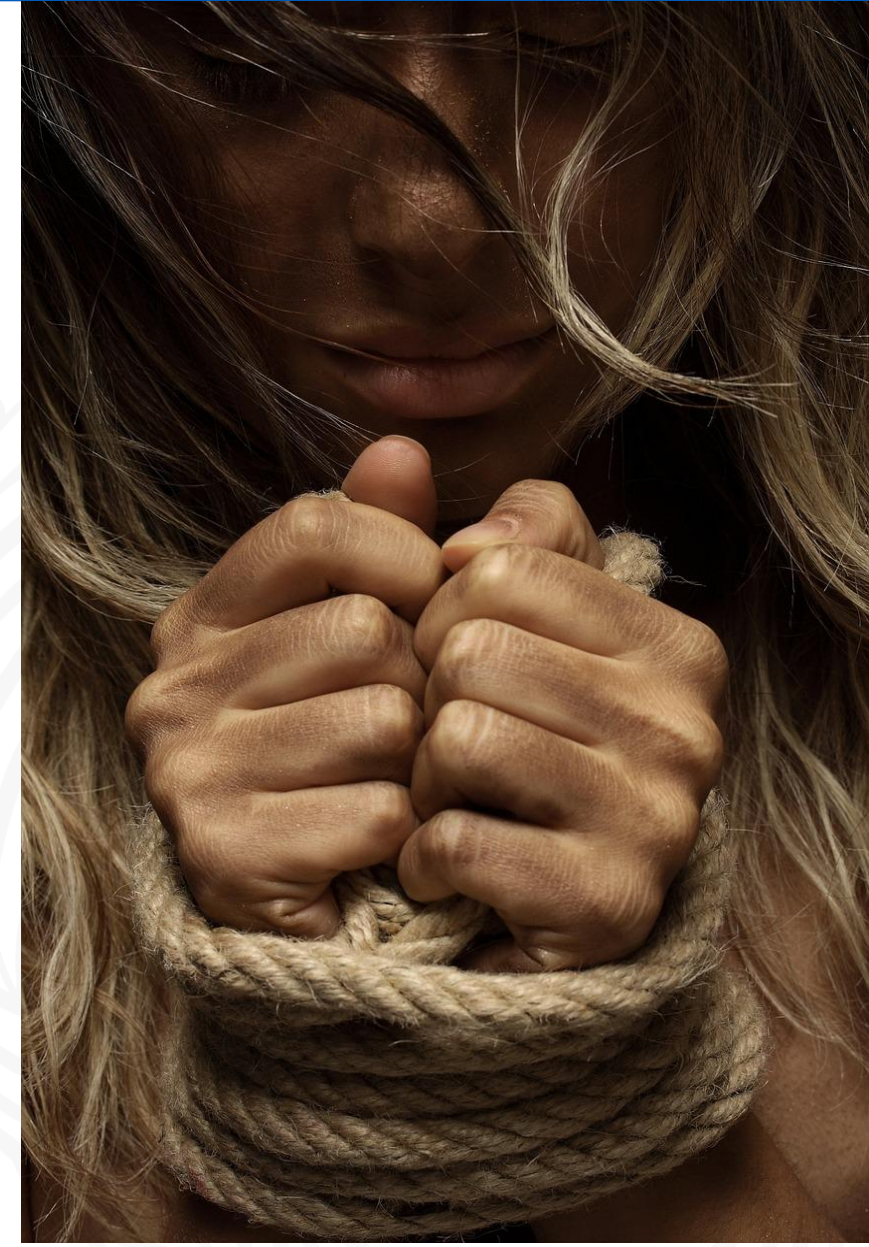
## INTRODUCTION

### Problem Statement:

- Crime Rate against Women in India has been a major issue
- It has seen a rise in recent years as ~30k women have been raped across the states in India in 2020

### Impact

- Women's participation in socio-economic field is hindered
- It further restricts their opportunities and empowerment



## OBJECTIVE



Explore different demographic variables to see how they are associated with crime rate



The final objective of the project is to predict future crime rate in India against women



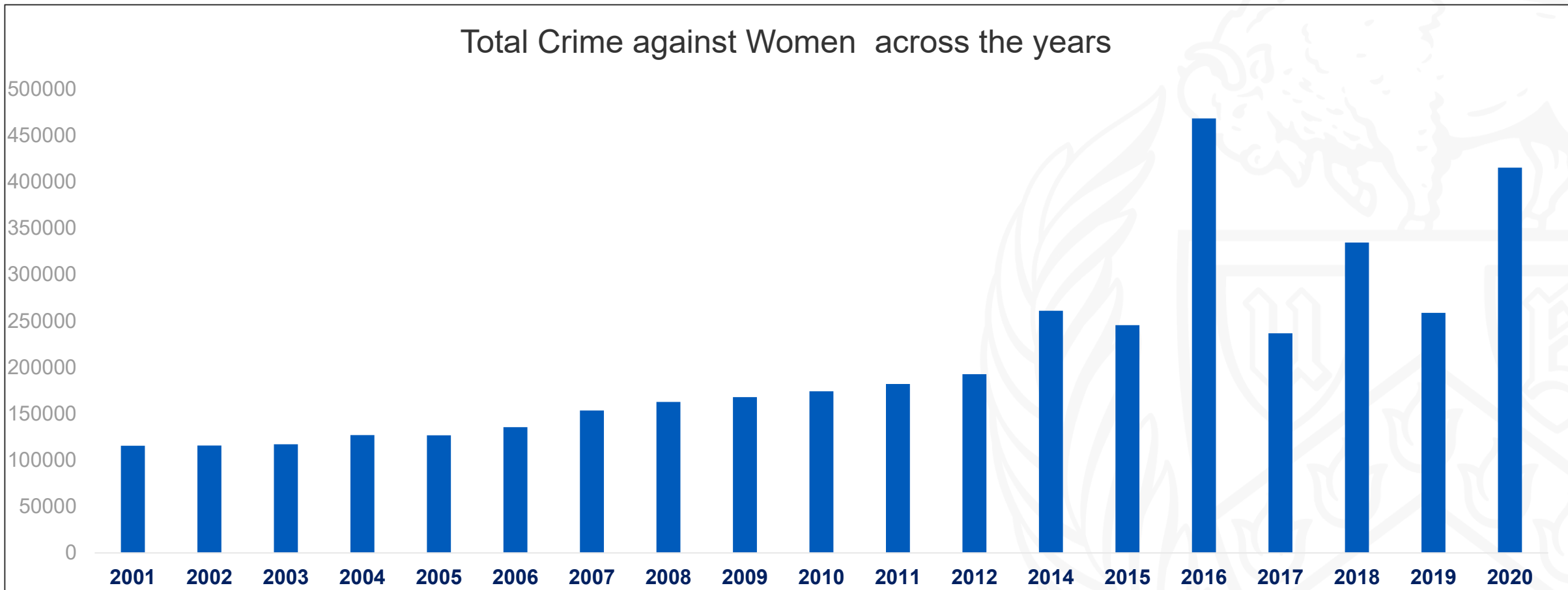
The study has been done as a country-level analysis



It has considered different types of crime against women in India such as rape, assault, dowry deaths etc.,



Total Crime against women in India has drastically increased in 2016 (~46k) which again stabilized in the subsequent years and plunged to a ~42k in 2020



# DATA SOURCES

## CRIME DATA

- Statistics of Crime Report from NCRB (National Crime Record Bureau) – 2001-2020



## DEMOGRAPHIC DATA

- India Handbook of statistics (Census data)- RBI



## DATA DESCRIPTION

Predictors Names	Type	Category	Frequency
Year	Int	Time	Annual
State	Int	Location	Annual
No of Factories	Int	Demographic	Annual
Birth Rate	Float	Demographic	Annual
Death Rate	Float	Demographic	Annual
Working Capital	Float	Demographic	Annual
No of Workers	Float	Demographic	Annual
Gross Value Added	Float	Demographic	Annual
Net Value Added	Float	Demographic	Annual
Availability of Power	Float	Demographic	Annual
Unemployment Rate	Float	Demographic	Annual
Poverty Rate	Float	Demographic	Annual

Every data has been collected across states of India and yearly frequency

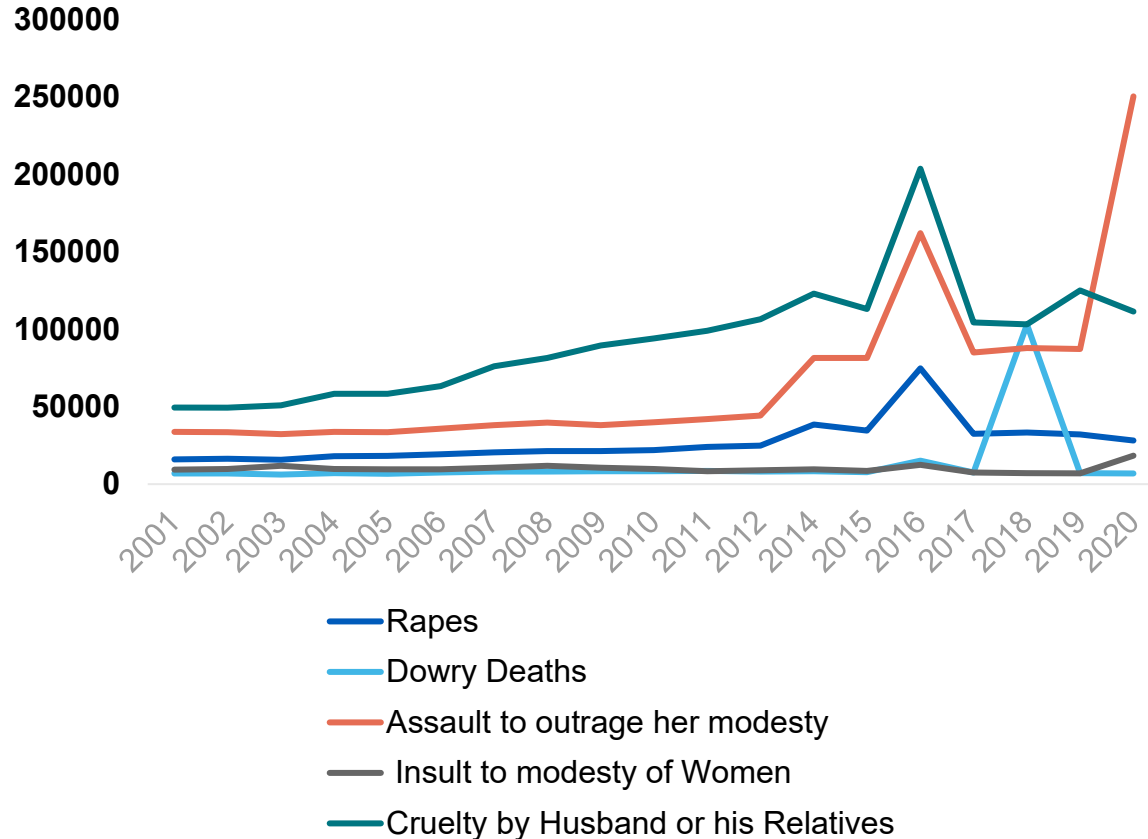
Predictors Names	Type	Category	Frequency
cpi	Float	Demographic	Annual
Literacy rate	Float	Demographic	Annual

Response Variables	Type	Category	Frequency
Rape	Float	Crime Data	Annual
Dowry Deaths	Float	Crime Data	Annual
Assault on women with intent to outrage her modesty	Float	Crime Data	Annual
Insult to modesty of Women	Float	Crime Data	Annual
Cruelty by Husband or his Relatives	Float	Crime Data	Annual

# Exploratory Data Analysis

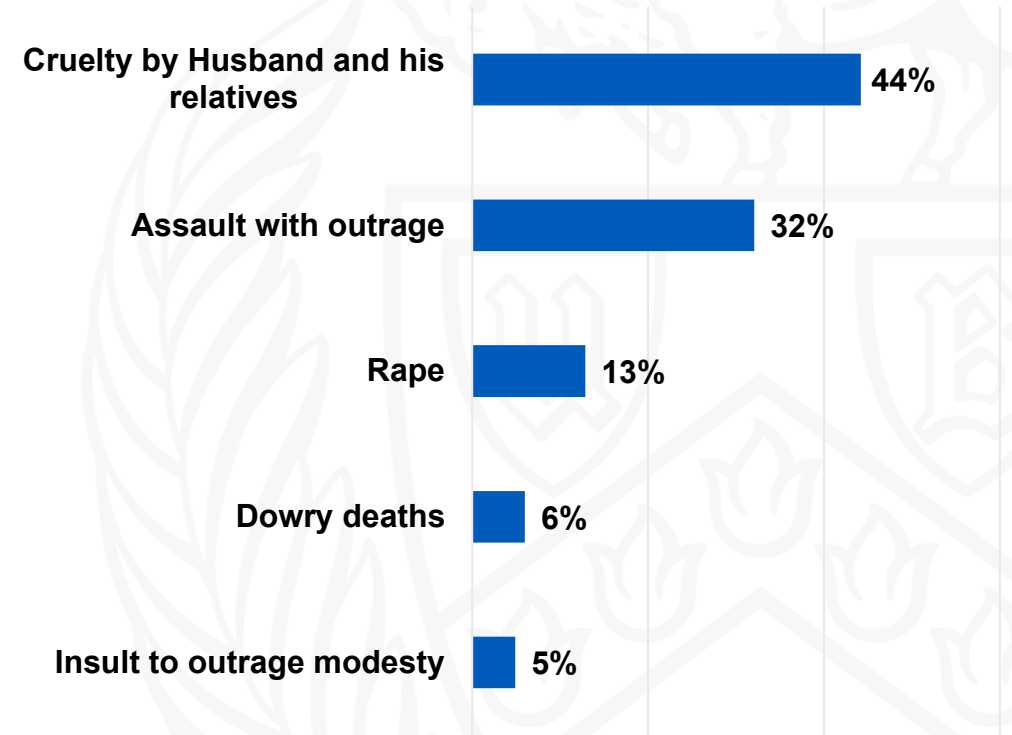
Crime against women has seen the spike in 2016 and then show an overall downward trend except assault cases

**Crime Prevalence over time**

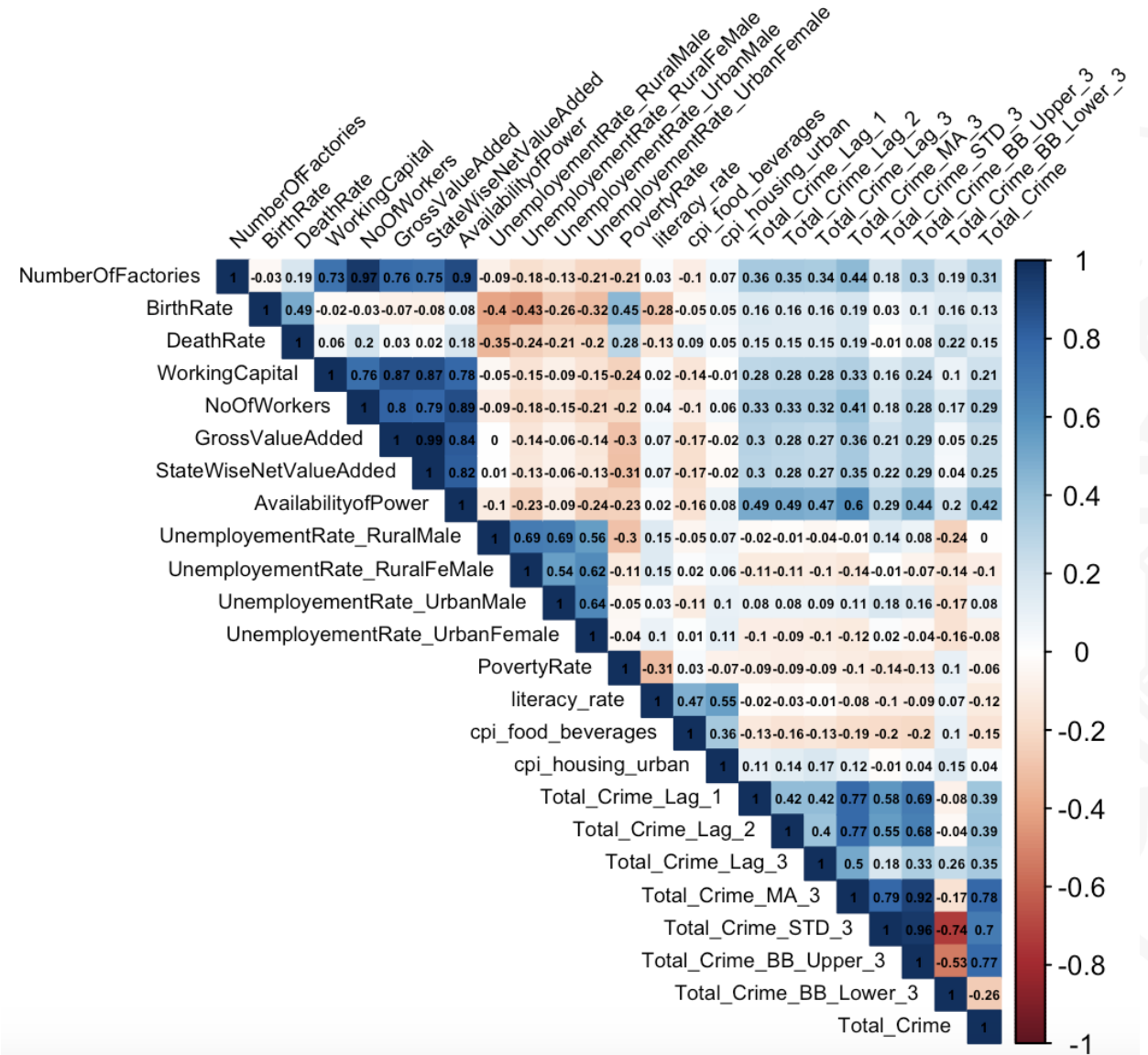


Almost half of the crime against women are contributed by husbands or their relatives

**Prevalent Crime Against Women in India**

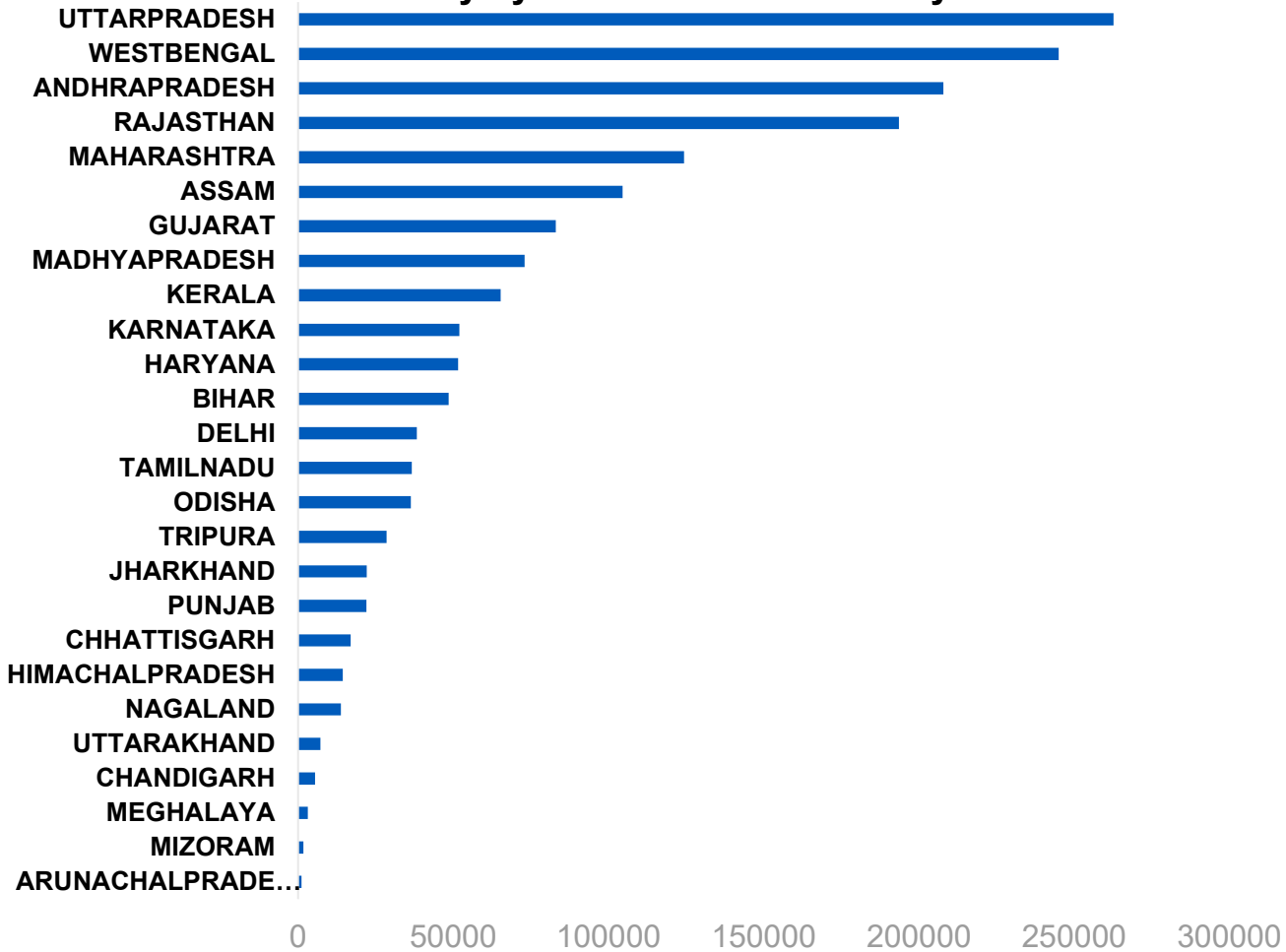


# Exploratory Data Analysis

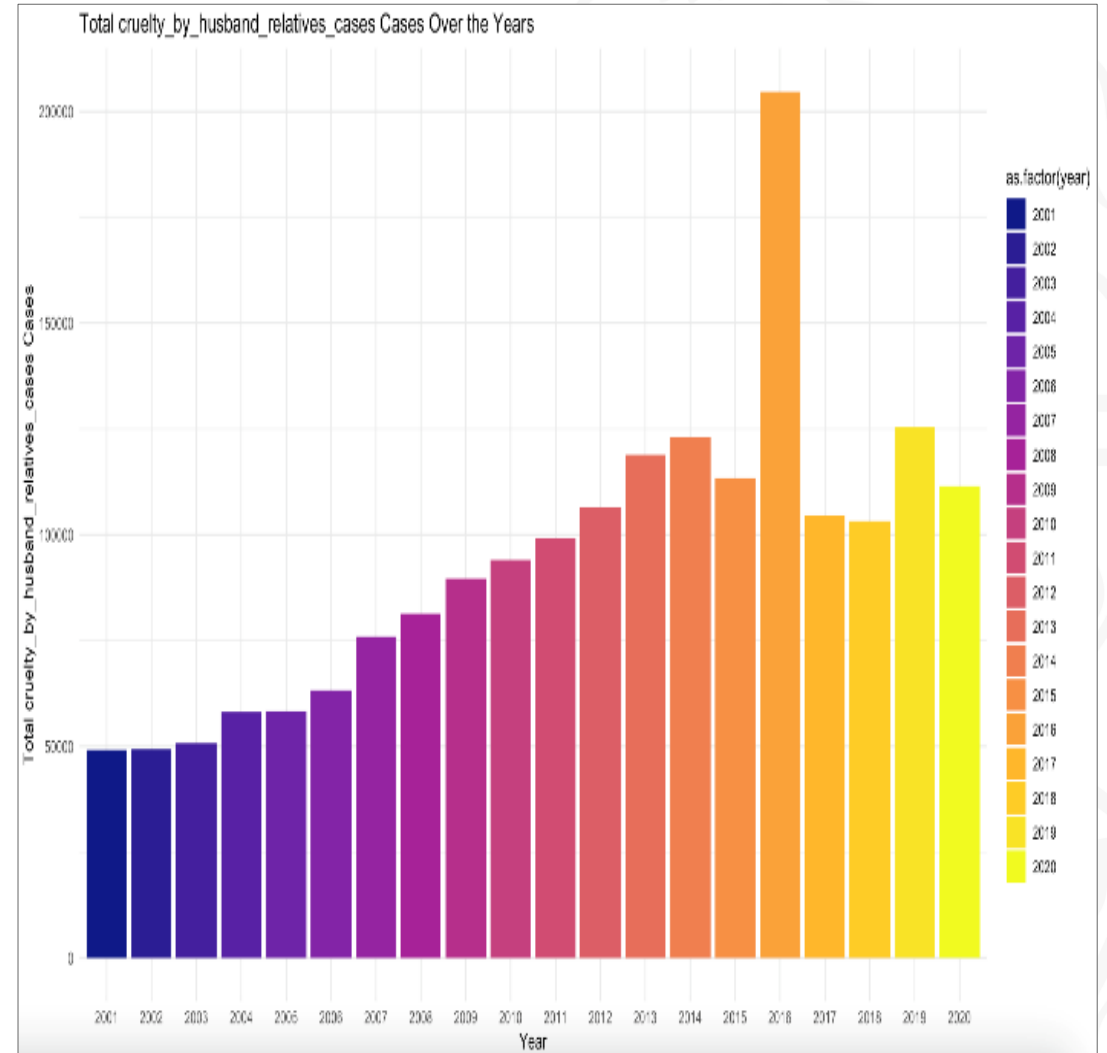


# Exploratory Data Analysis: Cruelty By husbands and his relatives

**Cruelty by husband and his family**

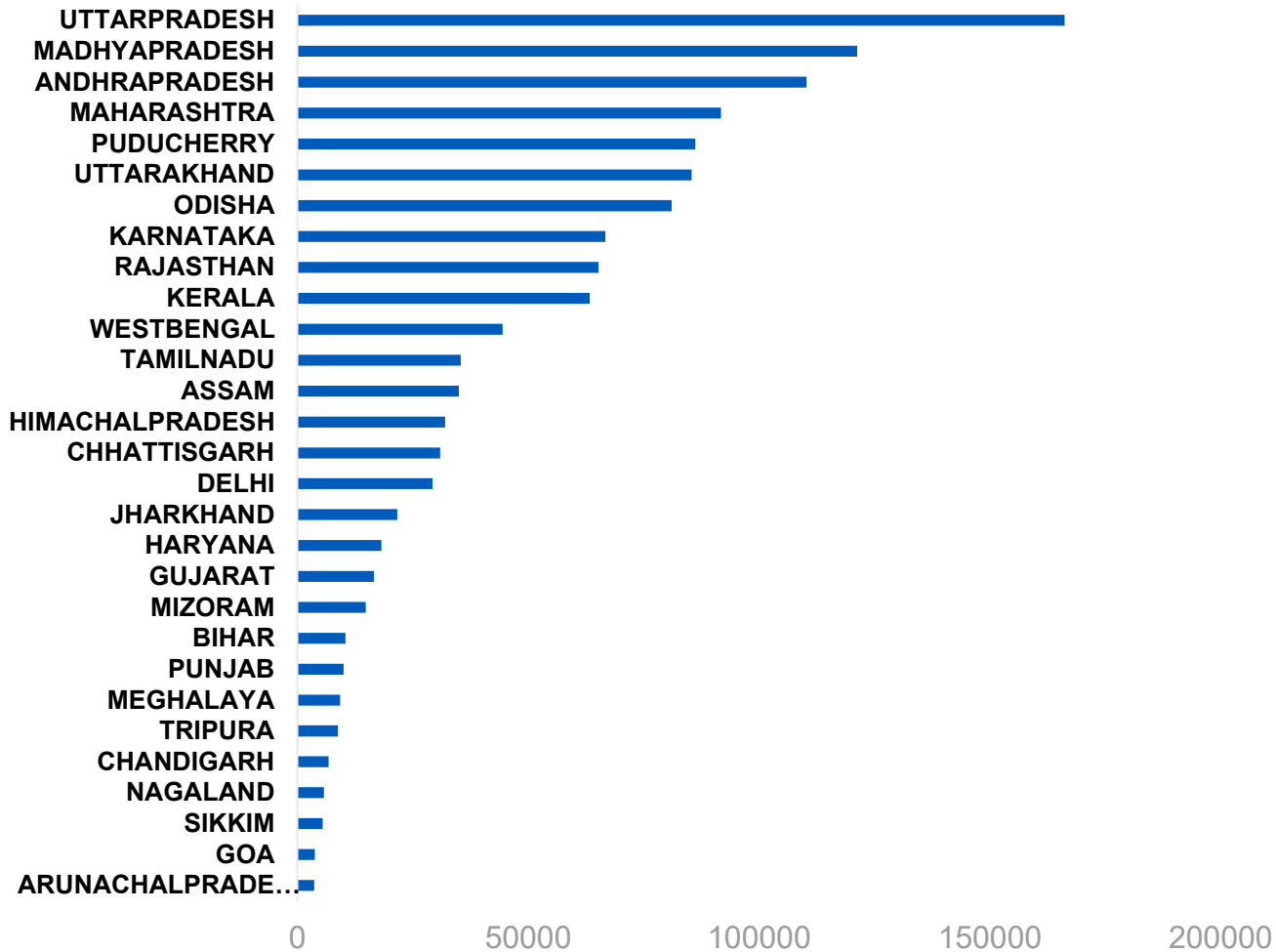


**Total cruelty\_by\_husband\_relatives\_cases Cases Over the Years**

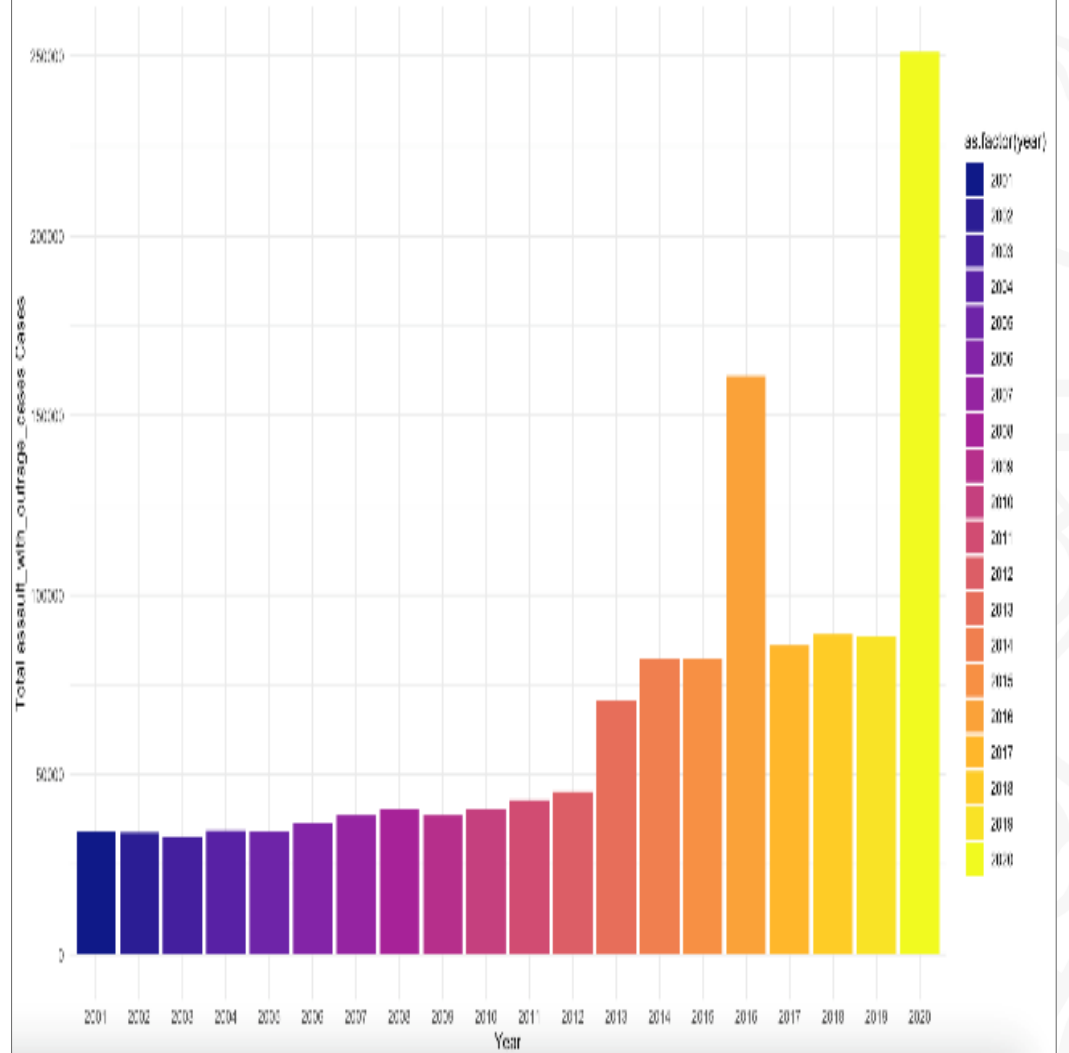


# Exploratory Data Analysis: Assault to outrage modesty

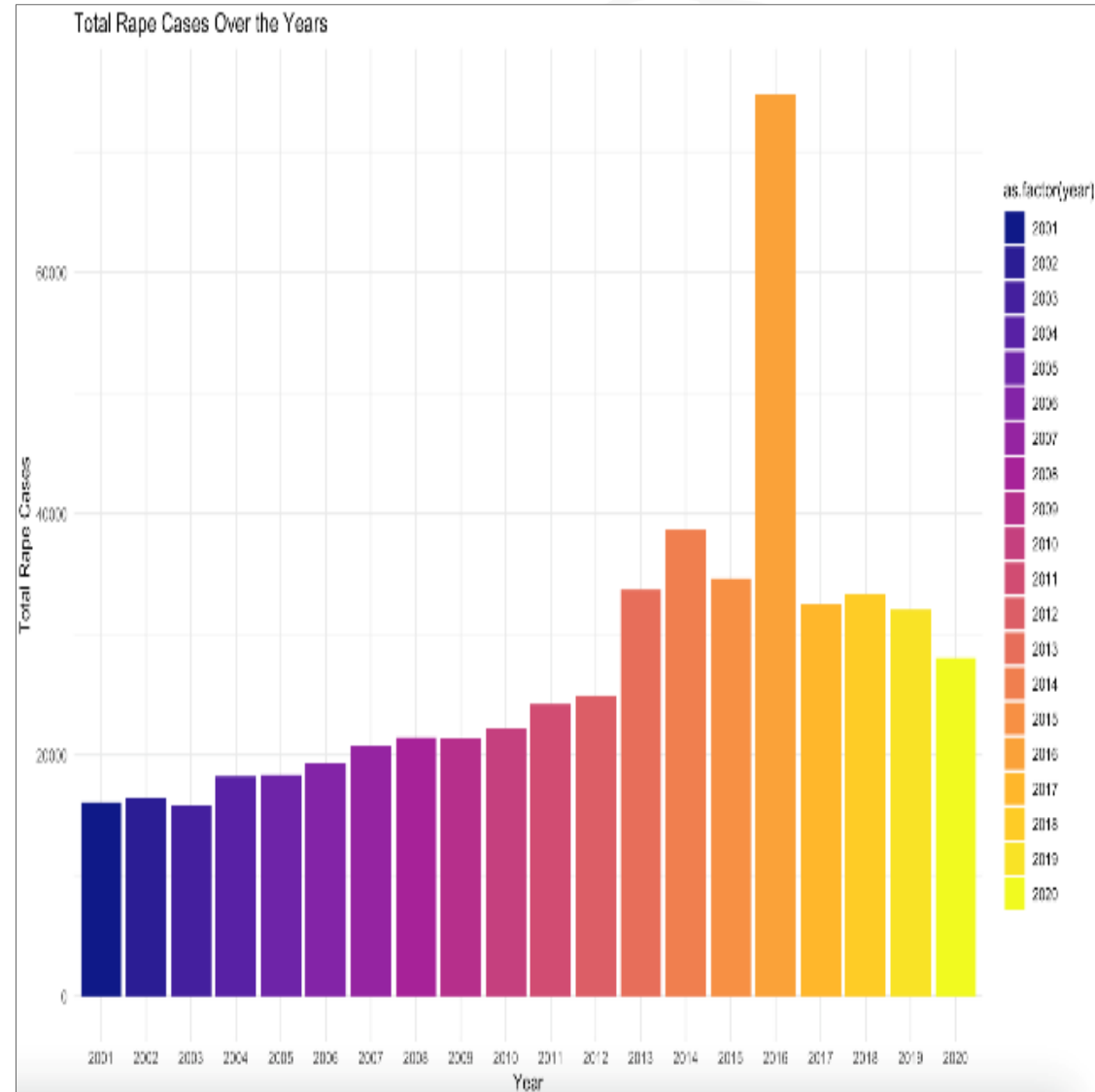
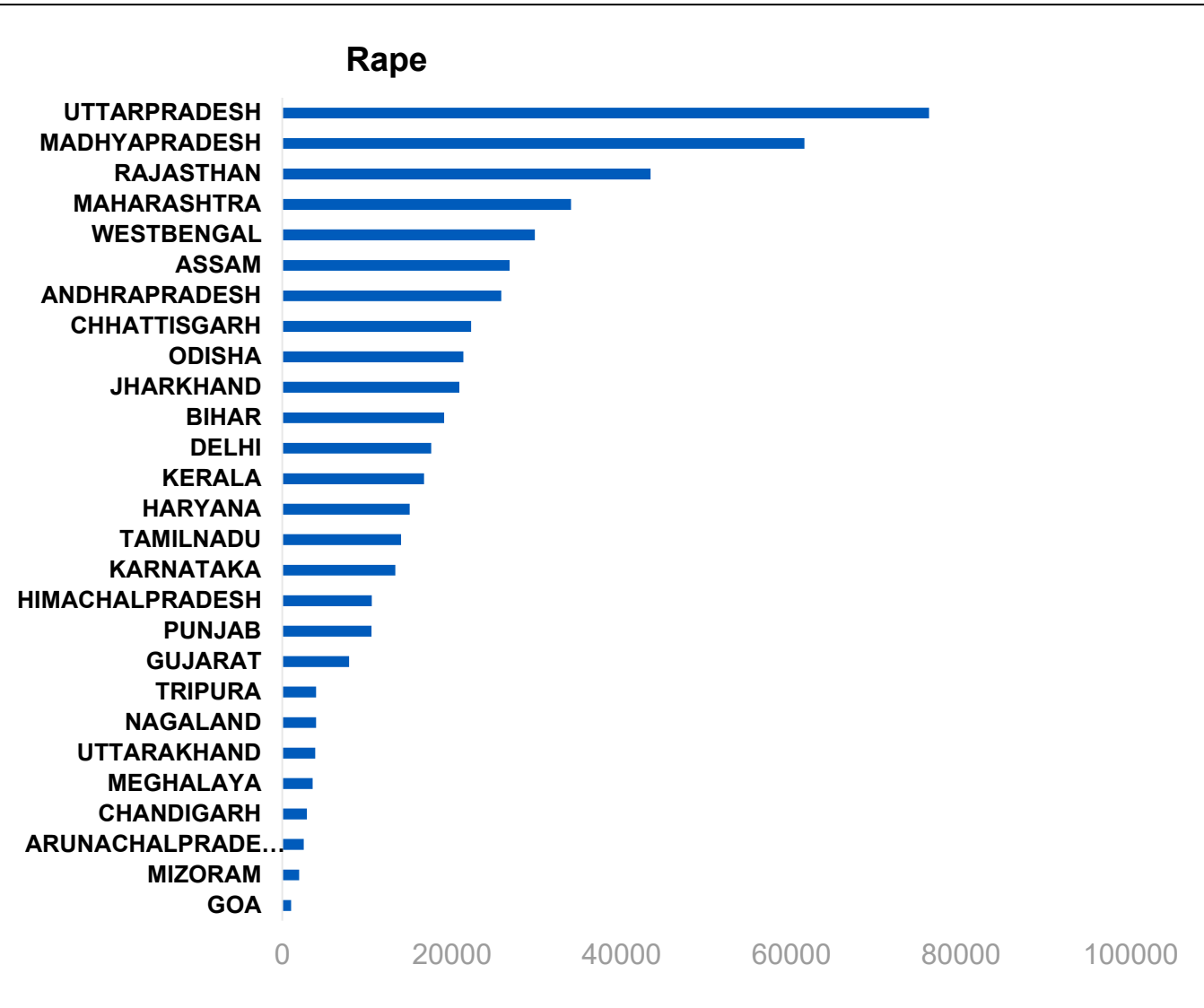
Assault to outrage modesty



Total assault\_with\_outrage\_cases Cases Over the Years

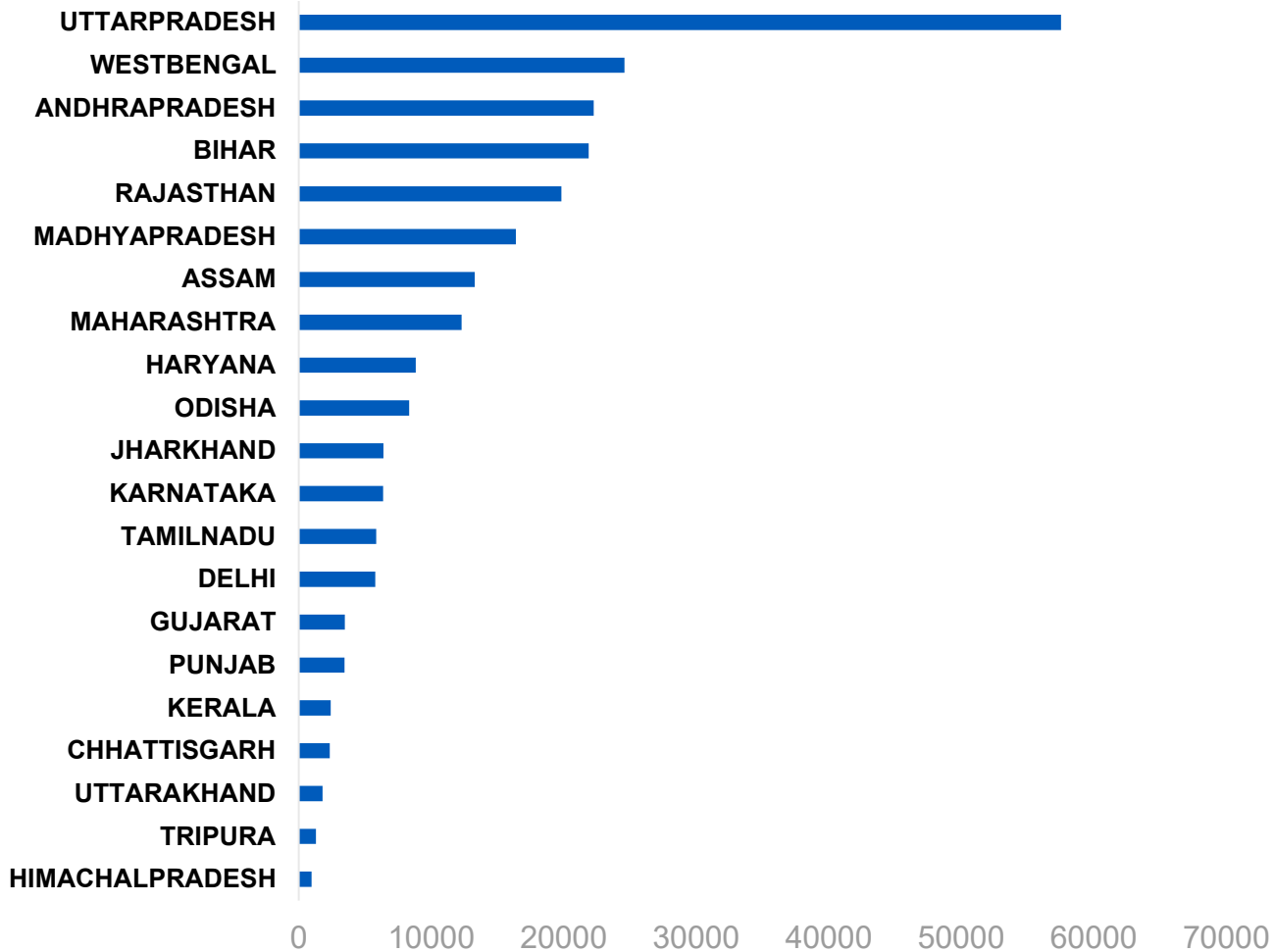


# Exploratory Data Analysis: Rape

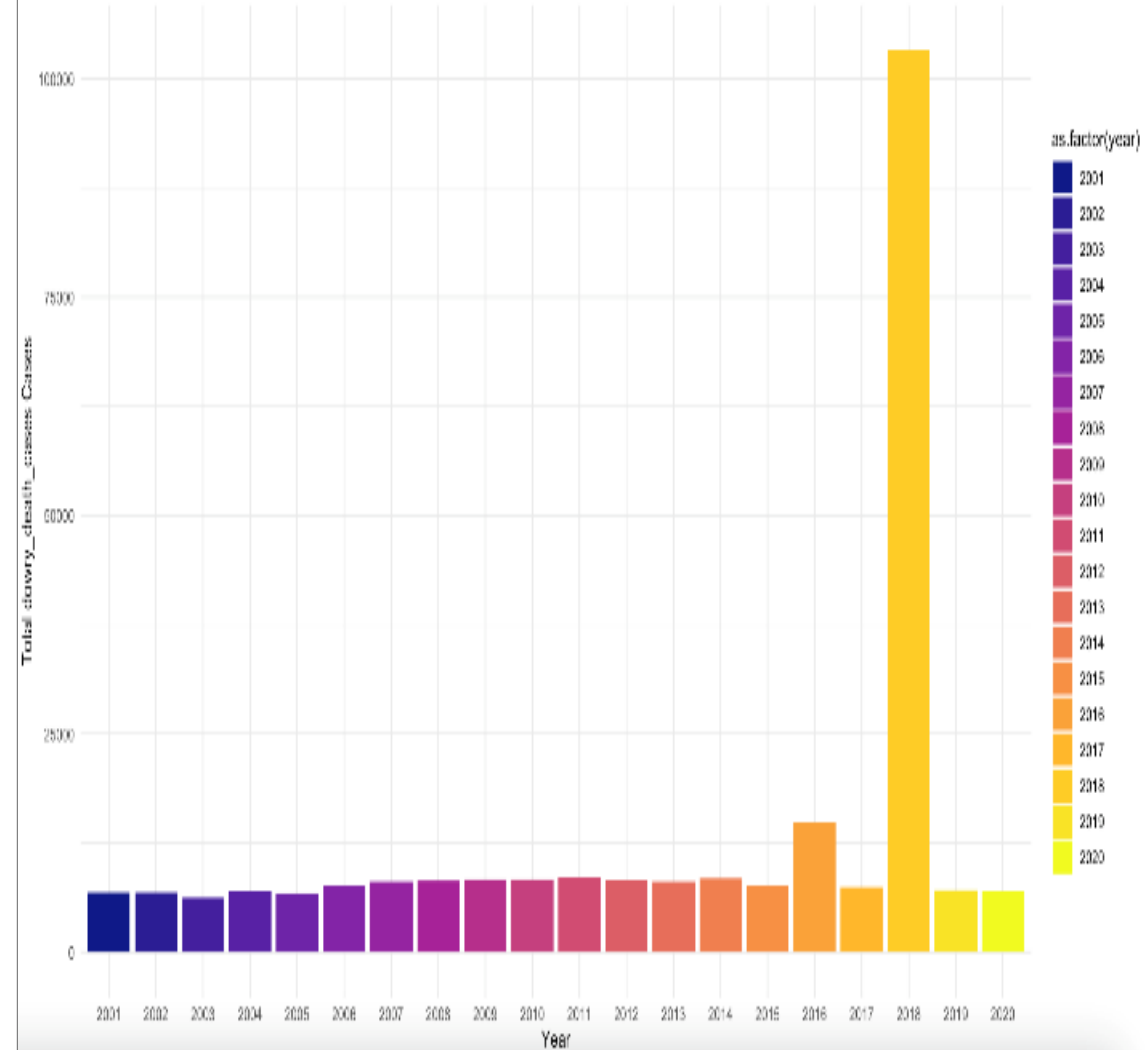


# Exploratory Data Analysis: Dowry deaths

Dowry deaths

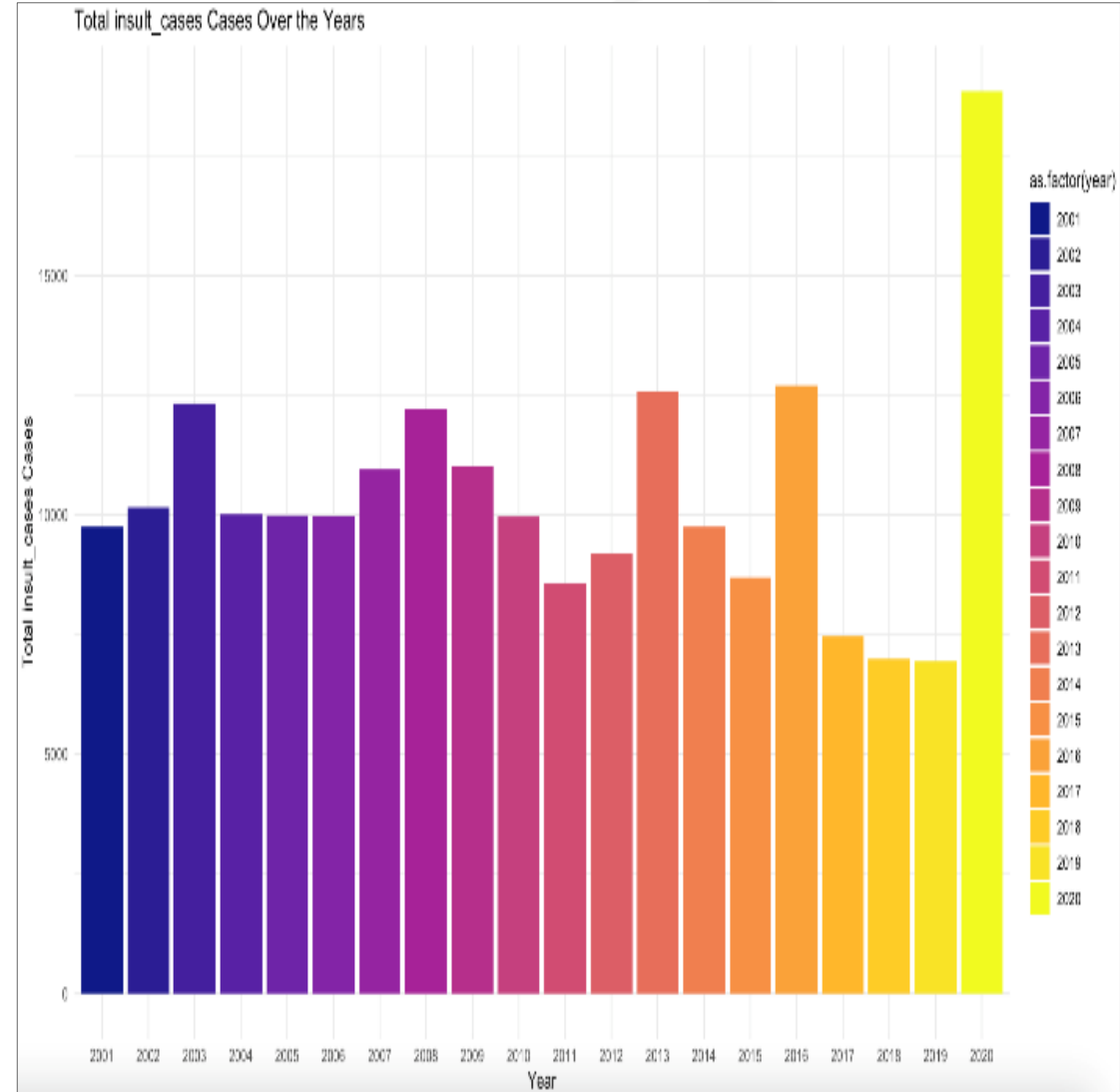
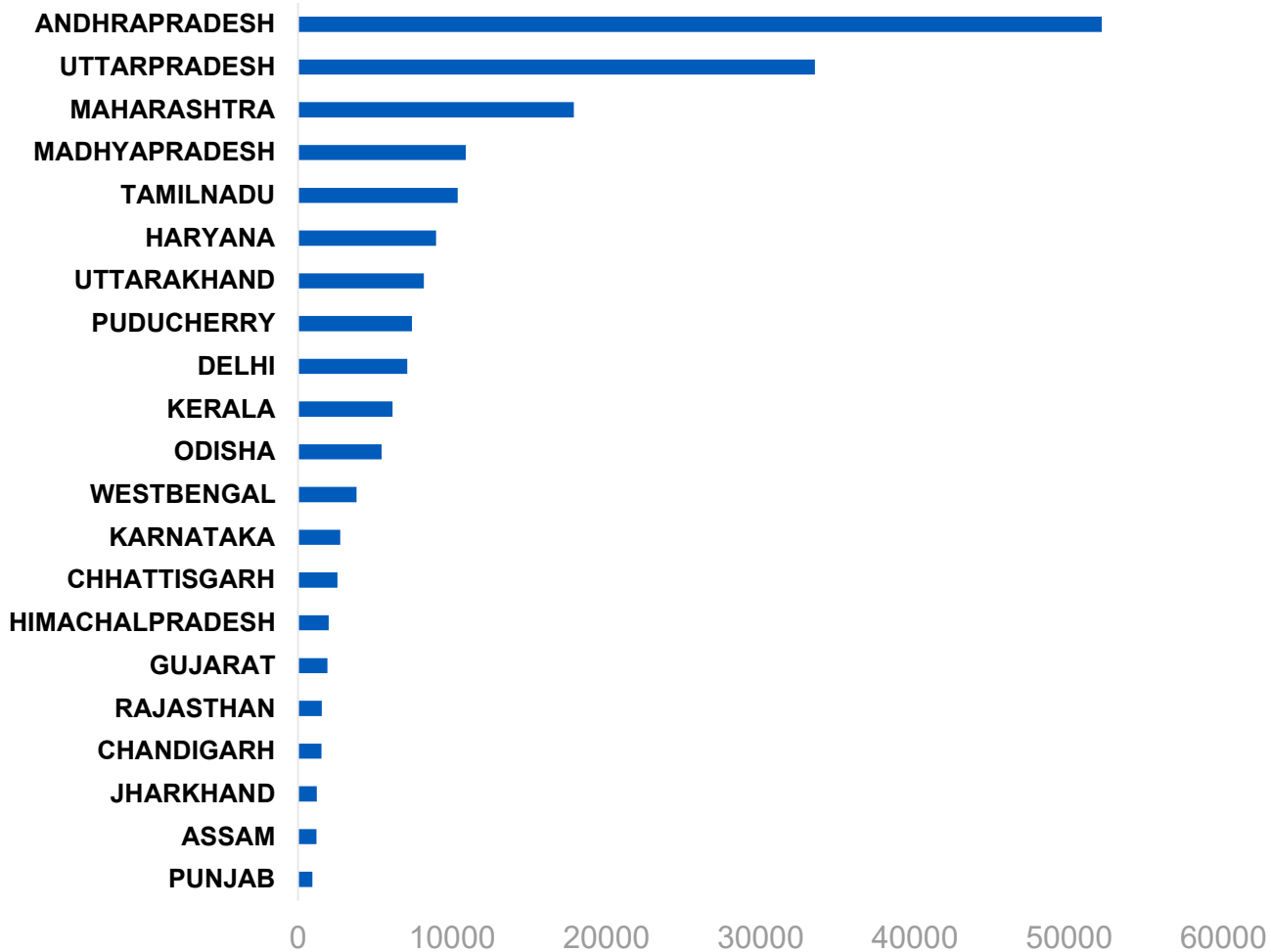


Total dowry\_death\_cases Cases Over the Years

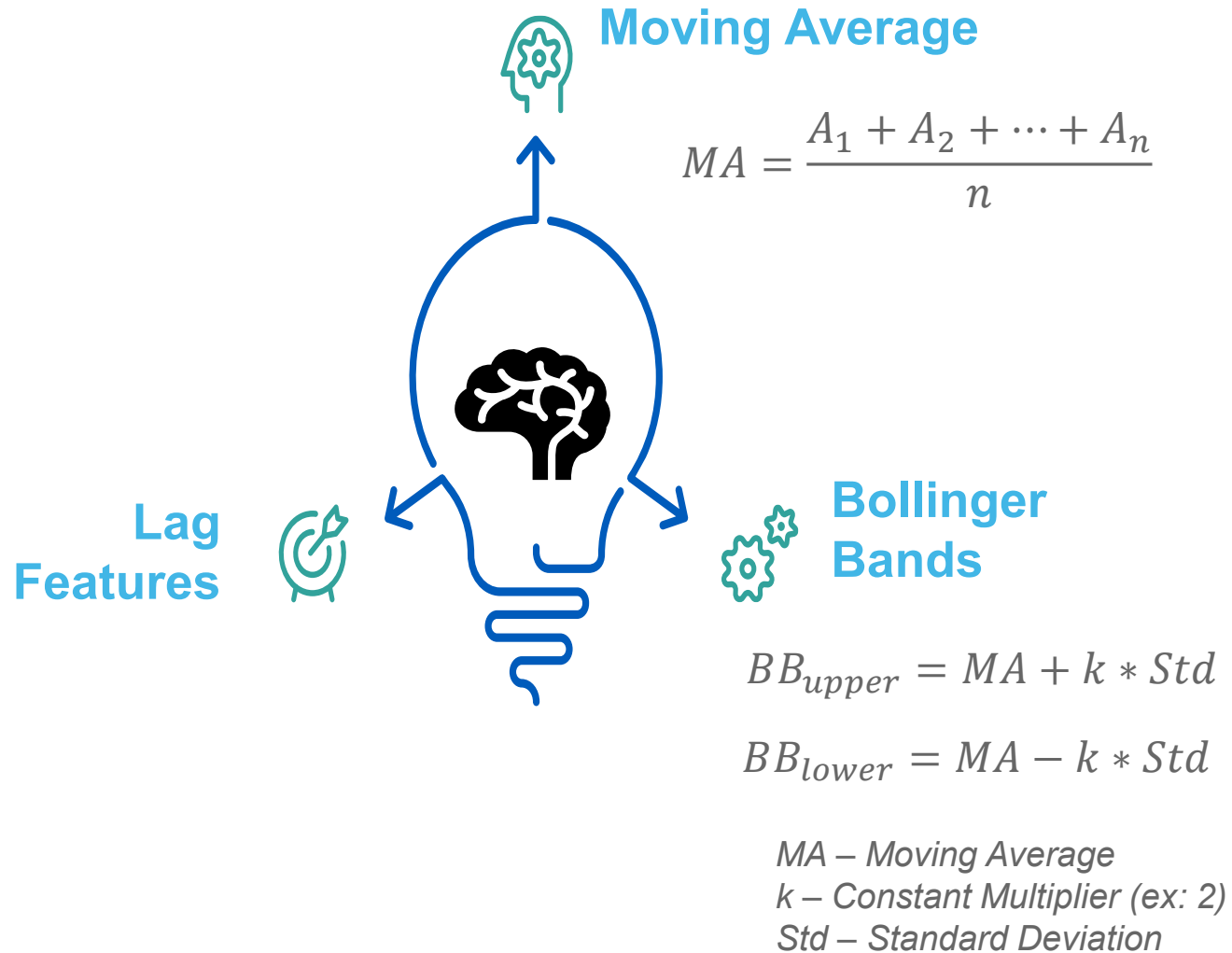


## EDA: Insult to Women with the intention of outraging modesty

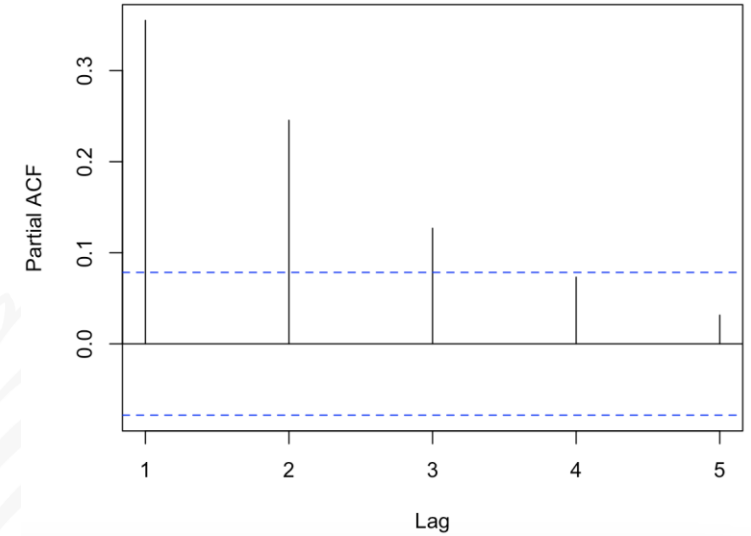
Insult to Women with the intention of outraging modesty



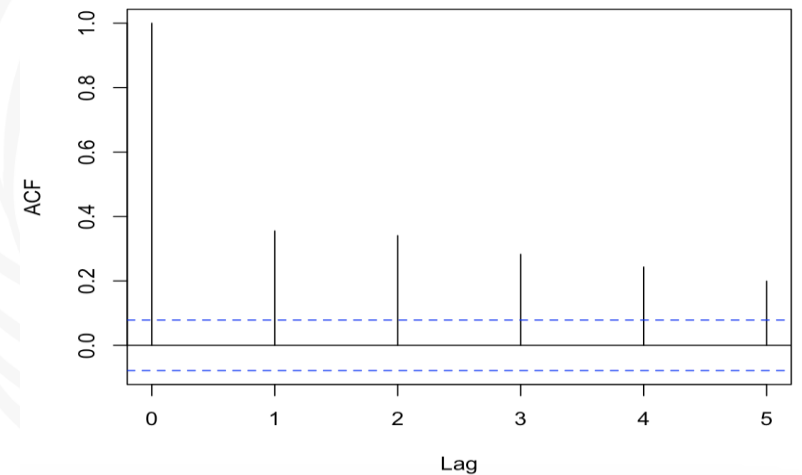
# FEATURE ENGINEERING



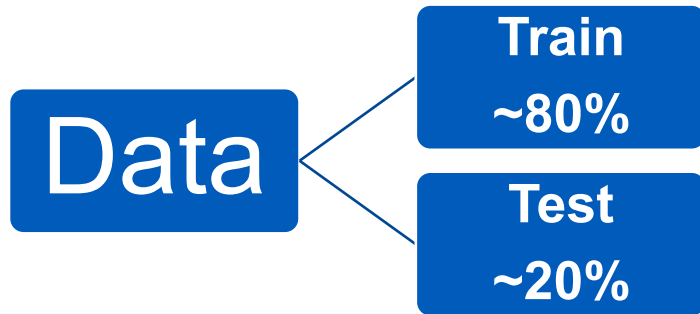
Partial Autocorrelation of Total Crime



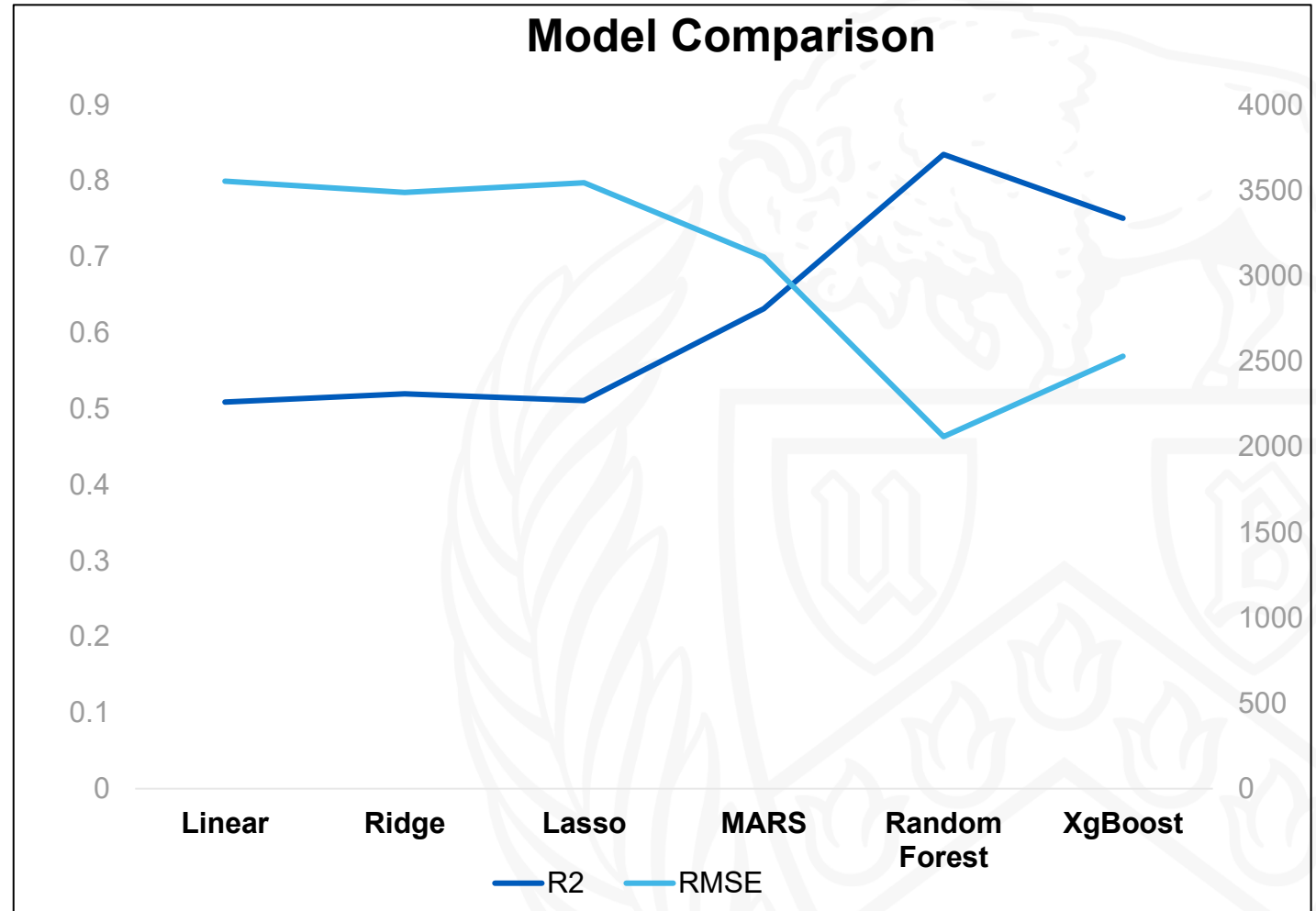
Autocorrelation of Total Crime



## Model Dev: Without Additional Features



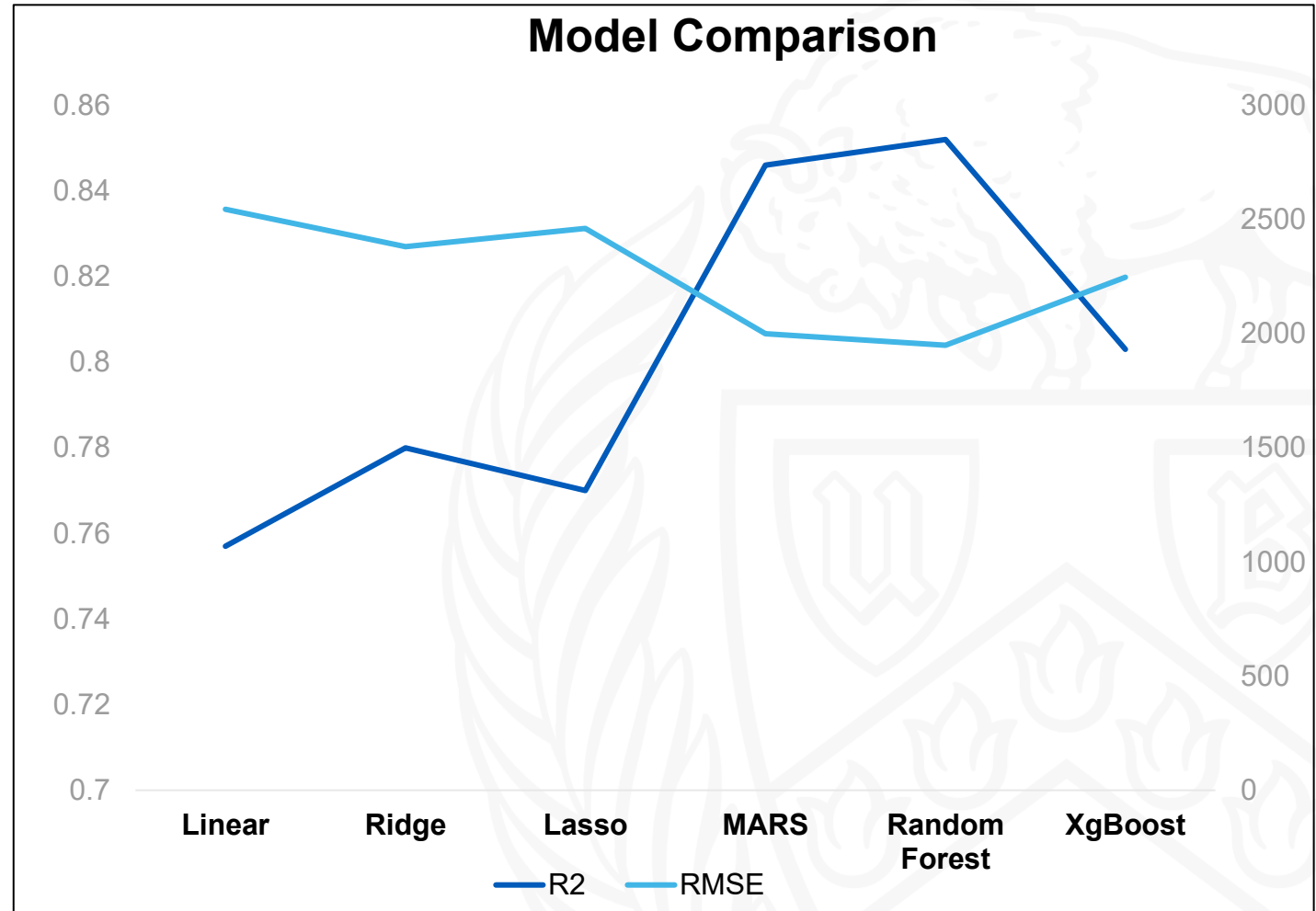
Model	R2	RMSE
Linear	0.509	3554.667
Ridge	0.52	3488.712
Lasso	0.511	3545.335
MARS	0.632	3110.663
Random Forest	0.835	2060.137
XgBoost	0.751	2530.862



## Model Selection: Random Forest works best for the dataset

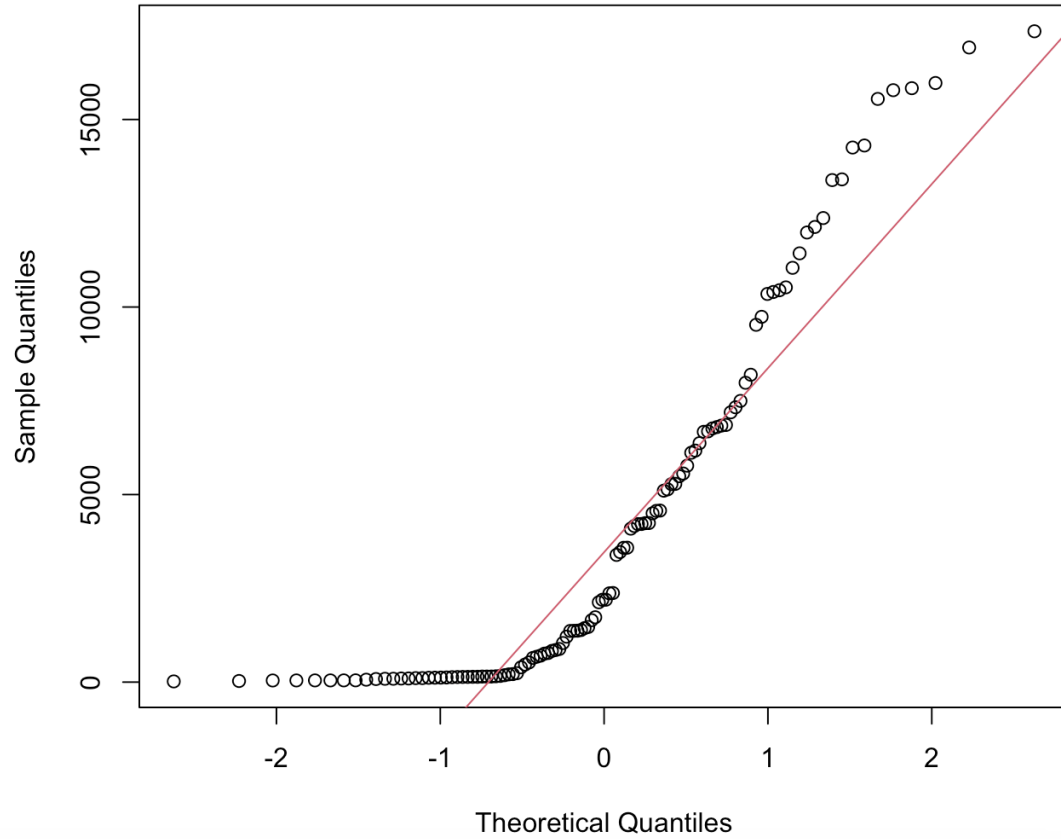
### Including Additional Features

Model	R2	RMSE
Linear	0.757	2544.579
Ridge	0.78	2380.809
Lasso	0.77	2460.53
MARS	0.846	1999.457
Random Forest	0.852	1949.075
XgBoost	0.803	2246.598

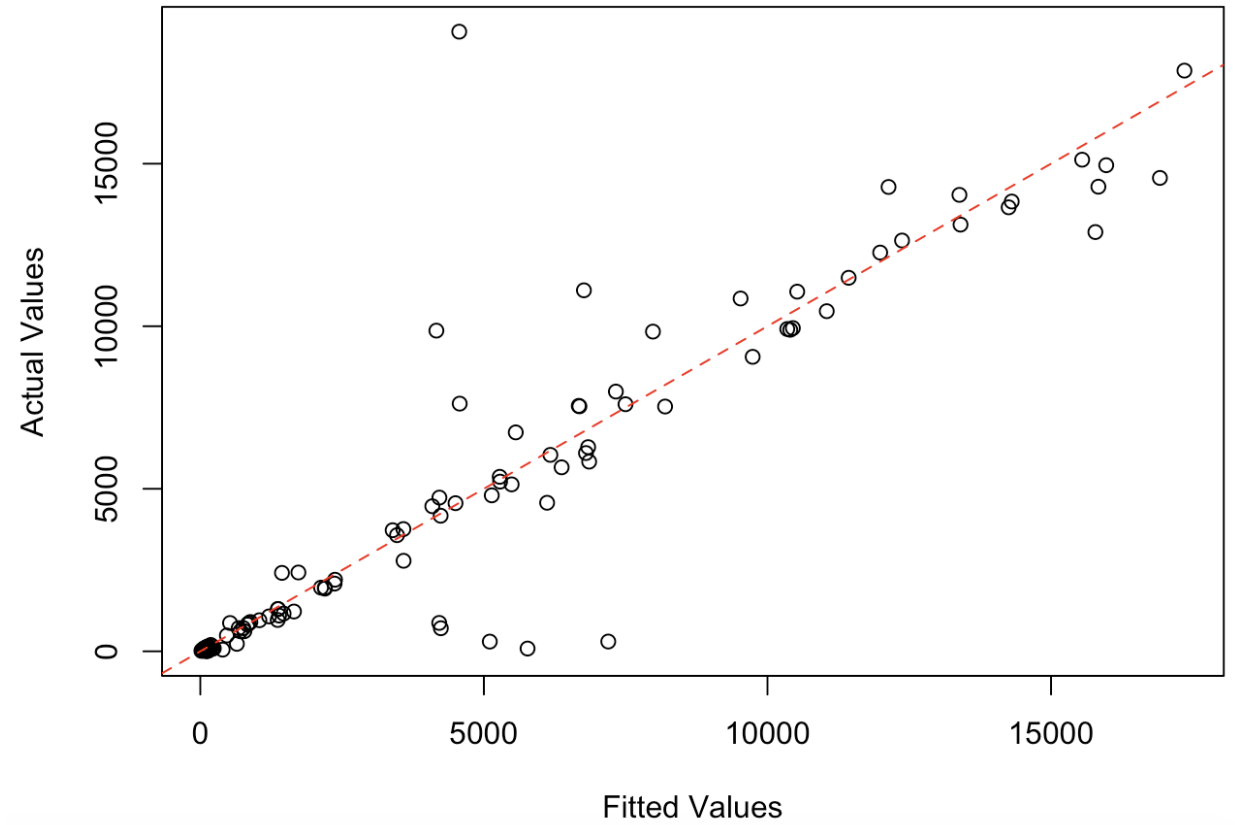


# Random Forest: This validates the good fit of model further

Normal Q-Q Plot



Actual vs. Fitted Plot - Random Forest

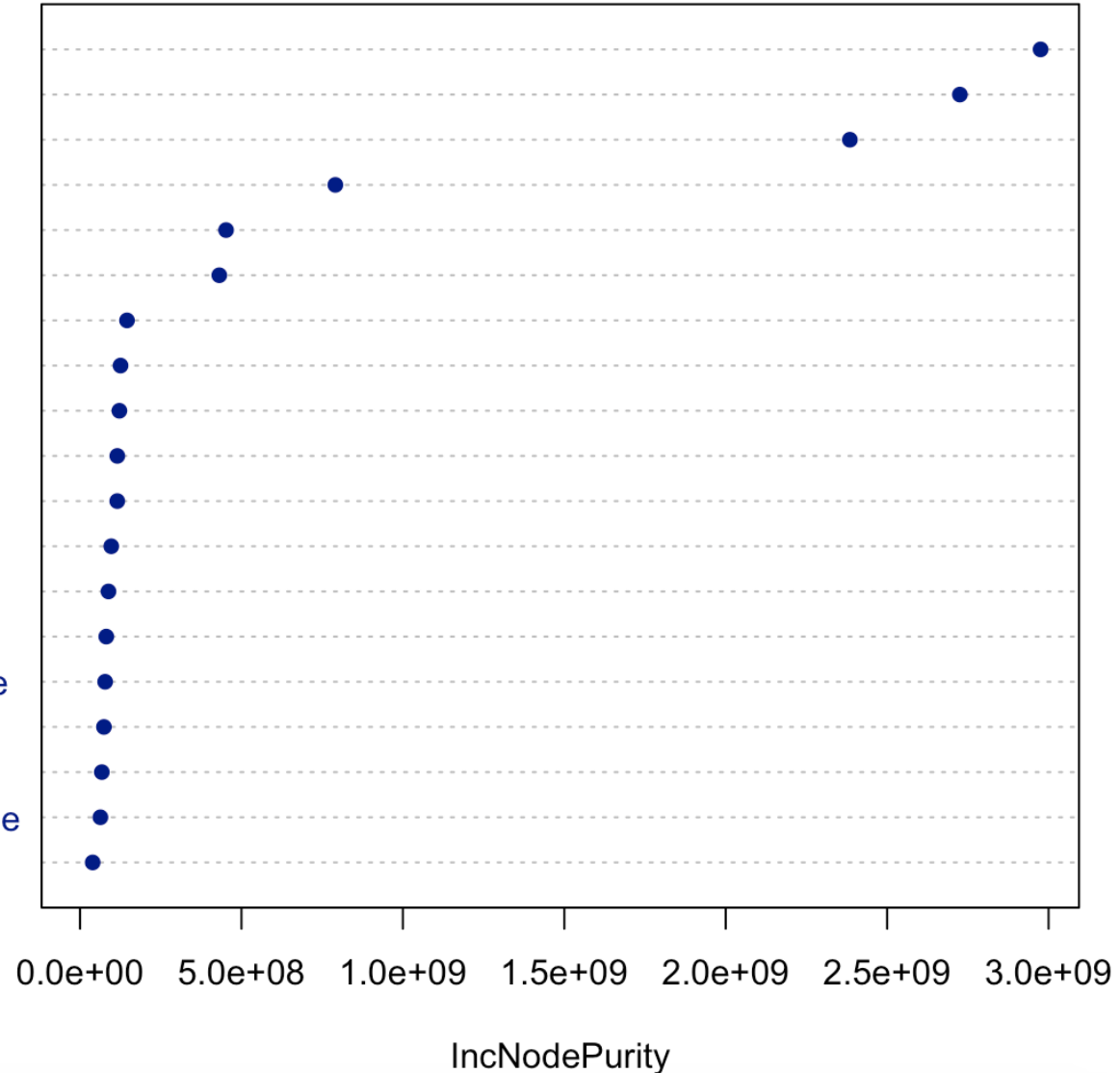


## Variable Importance:

- Total\_Crime\_MA\_3
- Total\_Crime\_BB\_Lower\_3
- Total\_Crime\_BB\_Upper\_3
- NumberOfFactories
- NoOfWorkers
- AvailabilityofPower
- cpi\_housing\_urban
- WorkingCapital
- BirthRate
- GrossValueAdded
- literacy\_rate
- cpi\_food\_beverages
- PovertyRate
- UnemploymentRate\_RuralMale
- UnemploymentRate\_RuralFeMale
- DeathRate
- StateWiseNetValueAdded
- UnemploymentRate\_UrbanFemale
- UnemploymentRate\_UrbanMale

**From the plot it is noticeable that the created features are important for predicting crime, also the number of factories, and availability of power is important**

## Feature Importance



# CONCLUSION

## Summary

- Random Forest predicts the best for our dataset having ~85% R2
- No of Factories and availability of power are major predictors



## Future Work

- Analyzing the crime in a more granular way (district or city level in quarterly frequency), crime hotspots can be identified
- Analyze the data punishment of the accused

# PRESENTED BY

Lohith Prasanna Teja Kakumanu

Poulomee Roy

Amila Viraj



# THANK YOU

## Q&A

