# CRIME & COMMUNITIES : PREDICT THE TOTAL NUMBER OF VIOLENT CRIMES PER 100K POPULATION

Md Jahid HASSAN

# CRIME DATASET

The Crime Dataset contains **128 socio-economic features** from the US 1990 Census. Describing U.S. communities in terms of demographics, income, employment, housing, law enforcement and more.

**Dataset Characteristics:**

- **Instances**: 1,994 US communities
- **Variables:** 128 (122 predictive, 5 non-predictive, 1 target)
- **Predictable Variables:** medIncome, agePct16t24, PctUnemployed, PolicPerPop, and so on
- **Non - Predictable Attributes:** 5 ( state, county, communityname, ...)
- **Target Variable:** ViolentCrimesPerPop (number of violent crimes per 100K population)

**Project Goal:**

- **Prediction**: Estimate *violent crimes per 100K population*

- **Optimization**: meaning to understand what social and law enforcement conditions lead to lower crime rates.

# DATA CHALLENGES AND LIMITATIONS

•**Missing data:** from certain communities (especially LEMAS data).

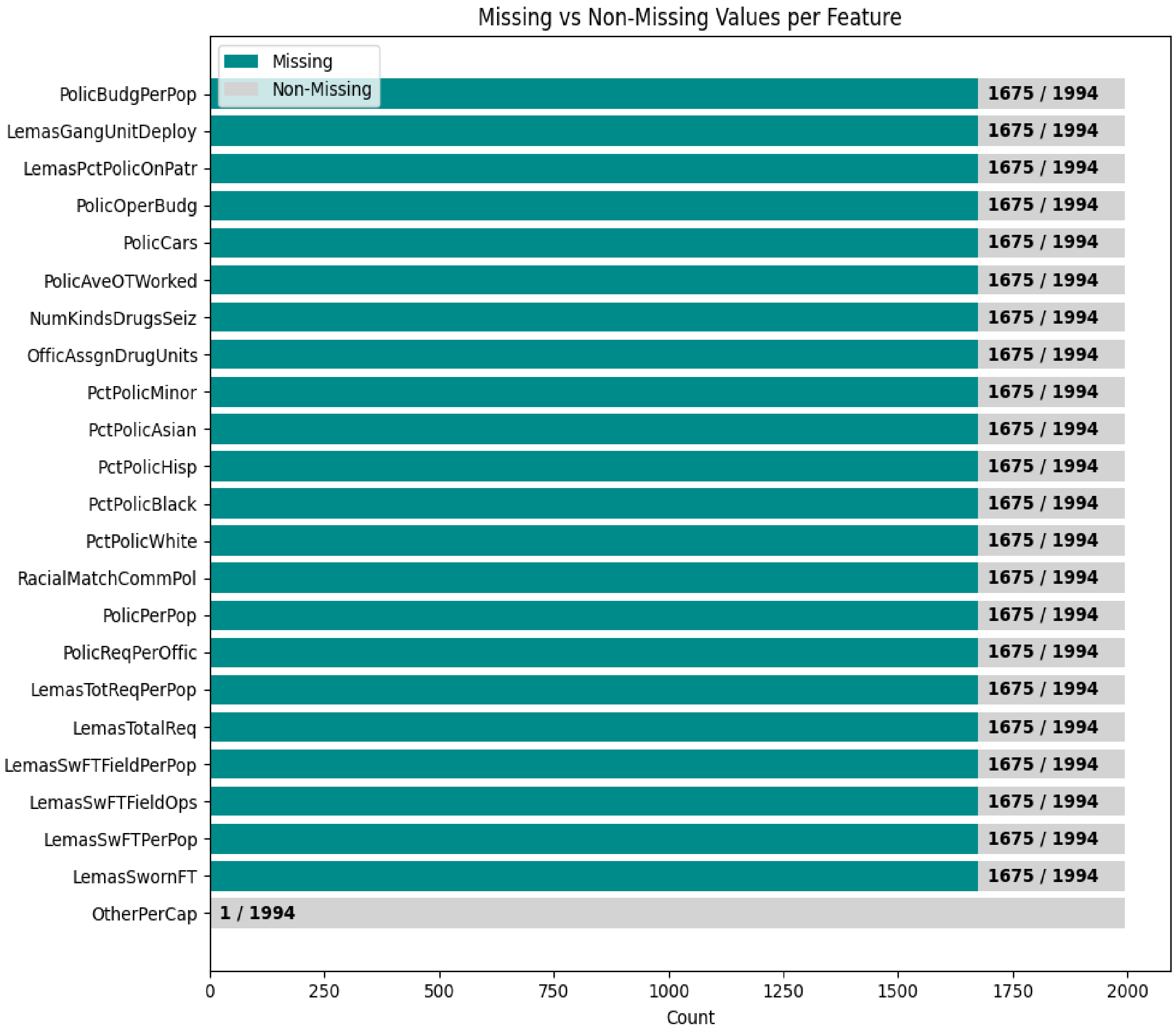
•**Variable normalization:** all feature values were scaled between 0 and 1 using an equal-interval binning method.

•Inconsistent relationships **between Variables** due to normalization.

•In some Midwestern U.S. states, rape reporting was inconsistent, which led to missing or unreliable violent crime totals.

**Additional Considerations:**
•Communities not found in both census and crime datasets were omitted.

# DATASET PRE-PROCESSING:



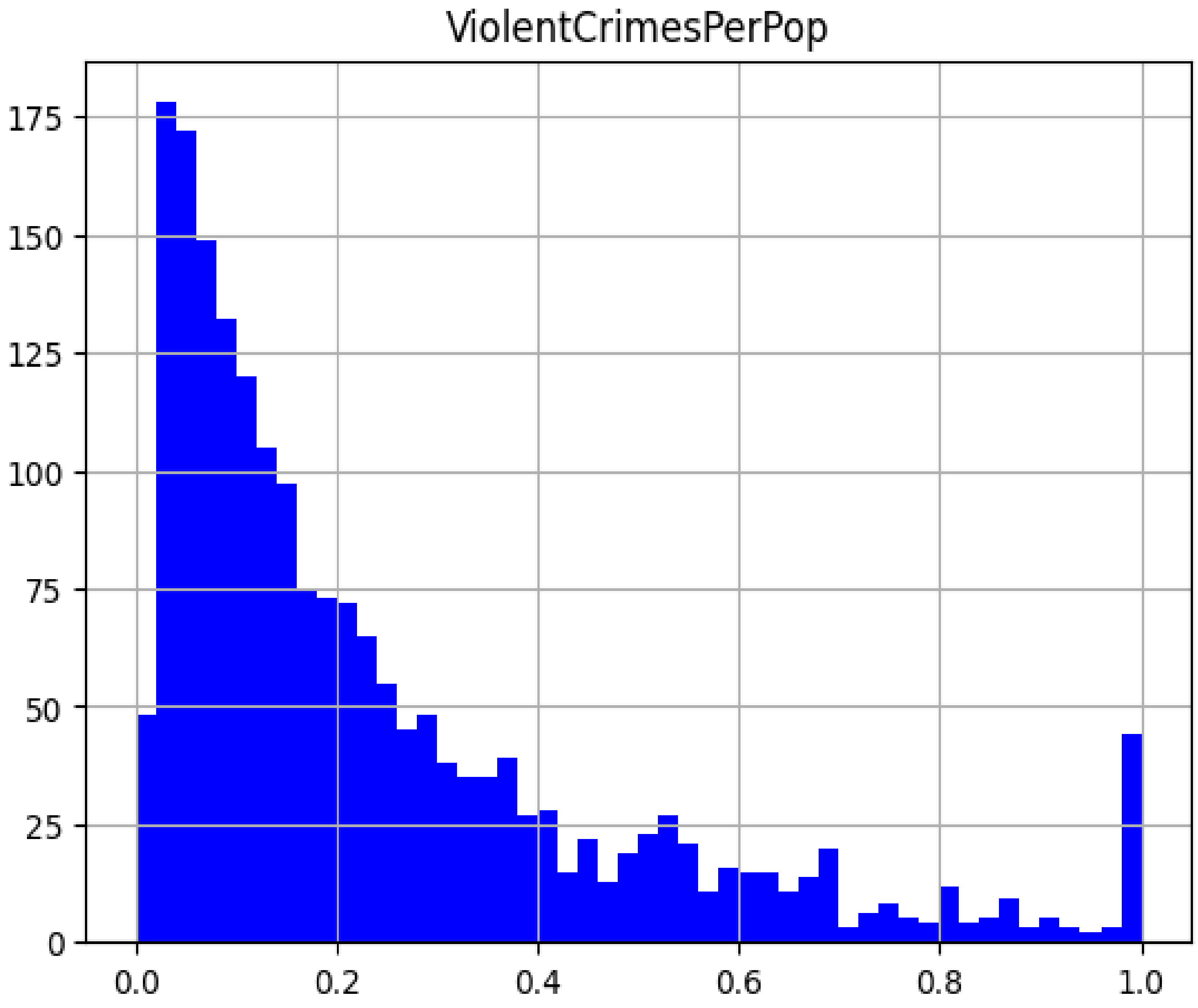

- From 122 predictive features 23 contain missing values. There are **22** variables missing **84%** of data. **Dropped**

- Most of these variables came from the 1990 Law Enforcement Management and Admin Stats survey (LEMAS).

- **OtherPerCap** has only one missing value. filled by mean value using Imputer from **sklearn.preprocessing.**

- No presence of any degenerate Columns.

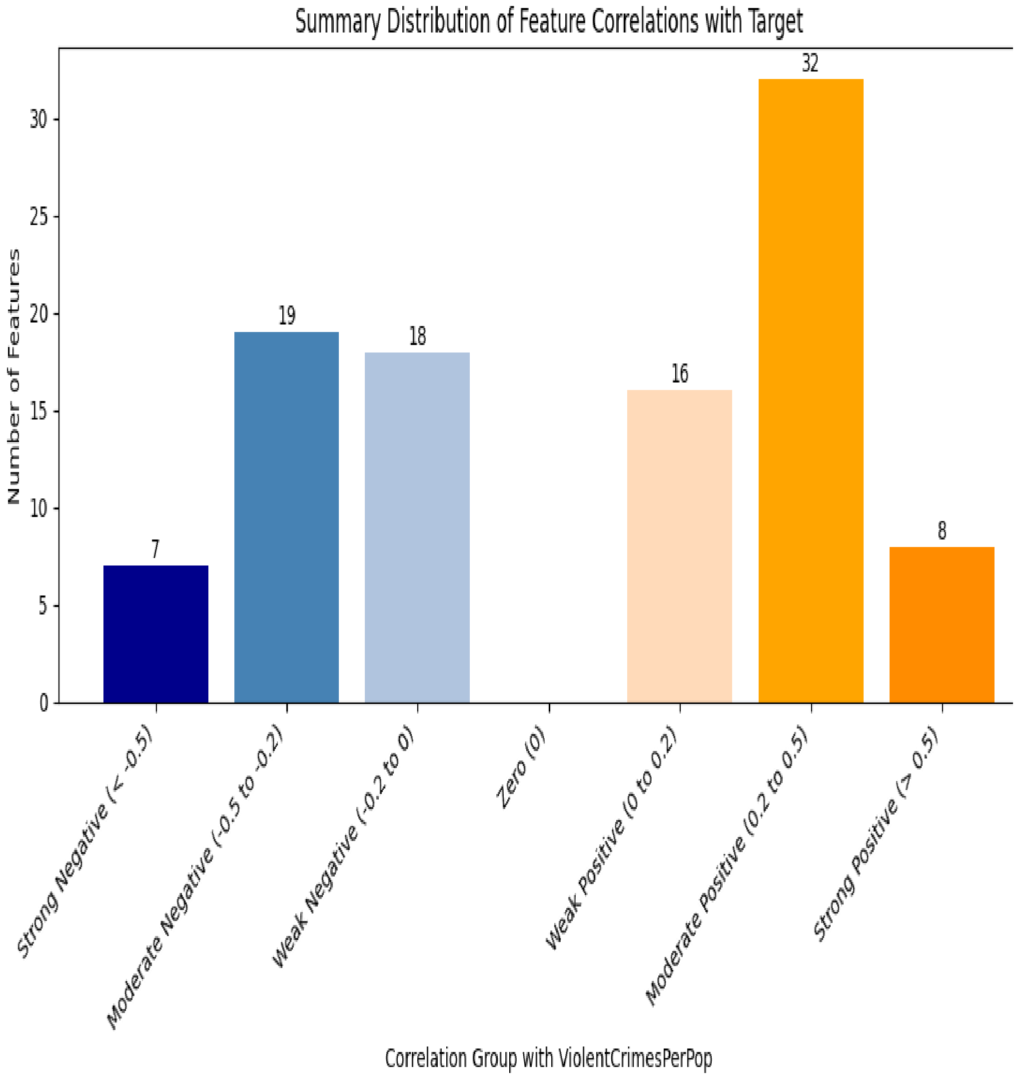
# TARGET DISTRIBUTION



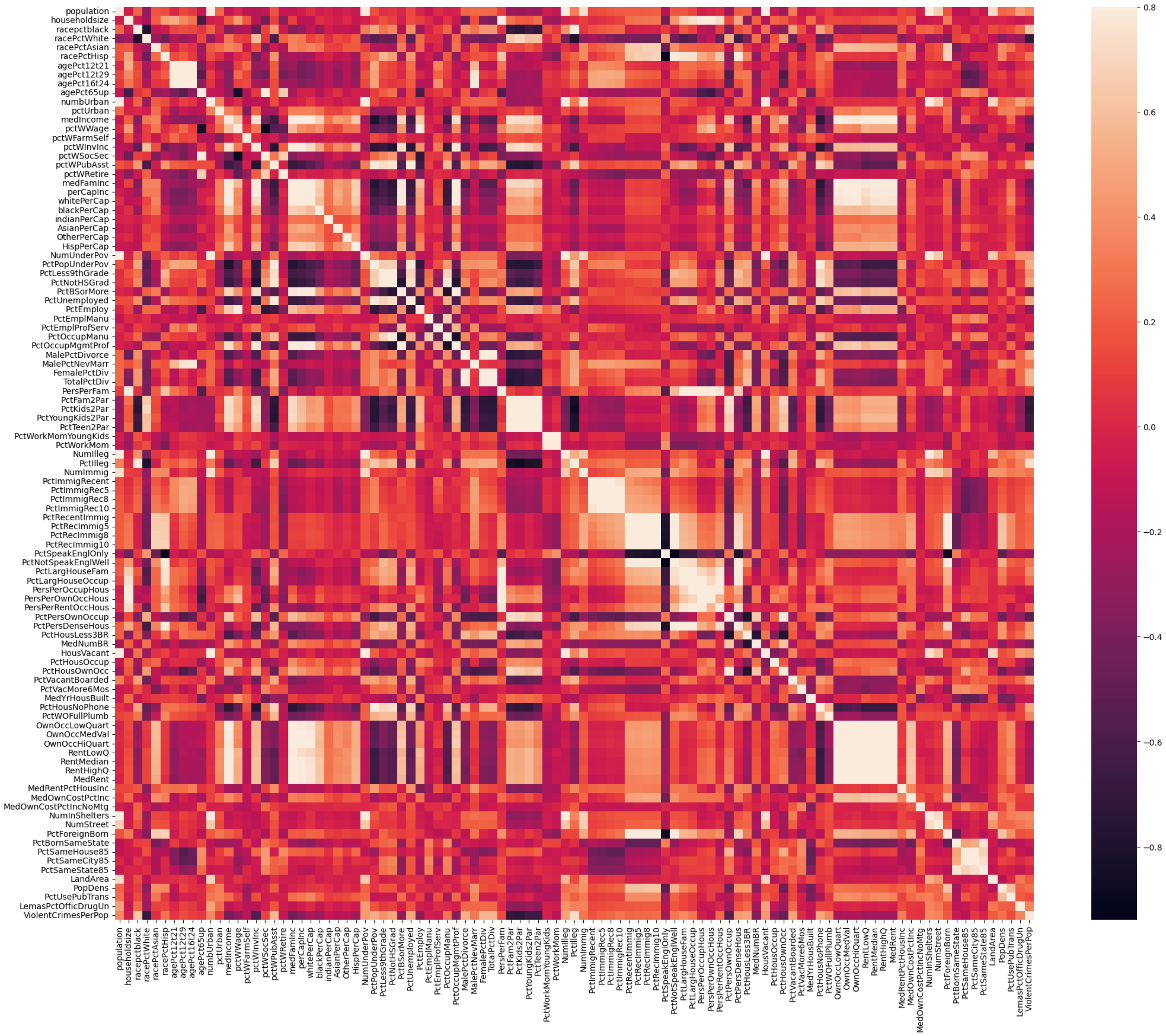

- Distribution of the ViolentCrimesPerPop variable shows , how violent crime rates (per 100,000 people) are spread across all 1,994 communities.

- Majority of communities having lower crime rates and a smaller number of communities with significantly higher rates.

# Co-relation:

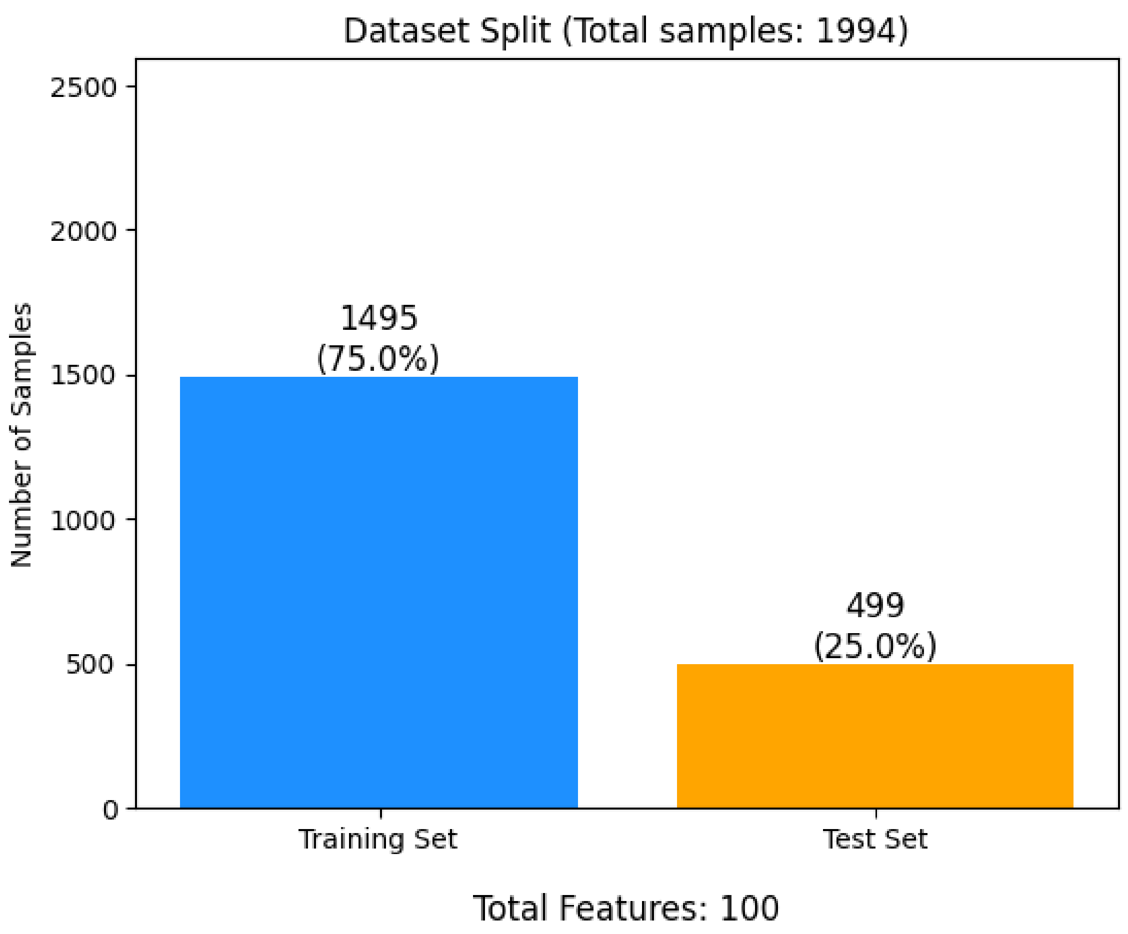


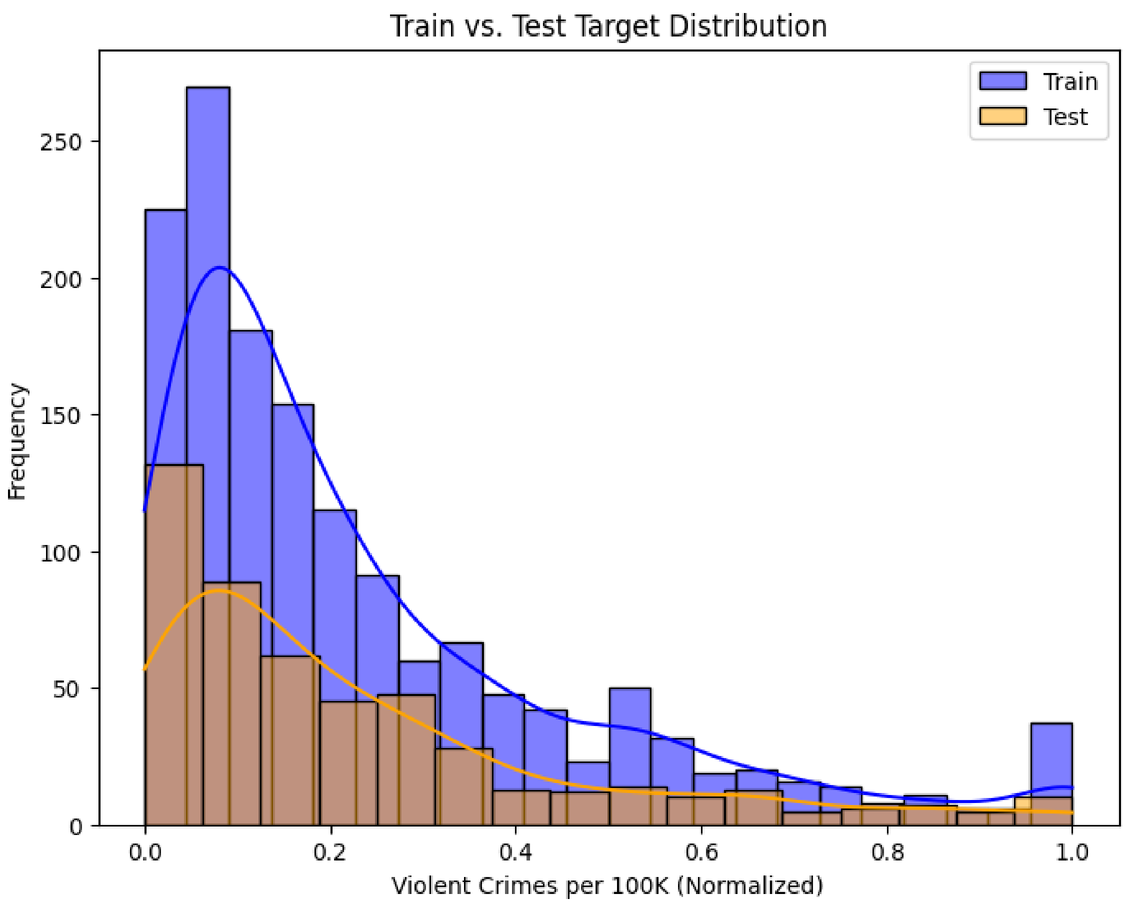
Summary Distribution of Feature Correlations with Target
Number of Features
0
5
10
15
20
25
30
7
19
18
16
32
8
Strong Negative (< -0.5)
Moderate Negative (-0.5 to -0.2)
Weak Negative (-0.2 to 0)
Zero (0)
Weak Positive (0 to 0.2)
Moderate Positive (0.2 to 0.5)
Strong Positive (> 0.5)
Correlation Group with ViolentCrimesPerPop

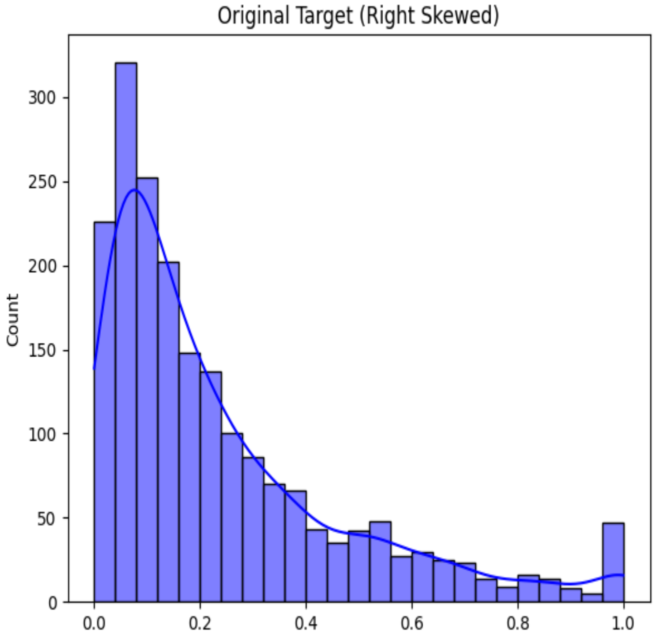
population
householdsize
racepctblack
racePctWhite
racePctAsian
racePctHisp
agePct12t21
agePct12t29
agePct16t24
agePct65up
numbUrban
pctUrban
medIncome
pctWWage
pctWFarmSelf
pctWInvInc
pctWSocSec
pctWPubAsst
pctWRetire
medFamInc
perCapInc
whitePerCap
blackPerCap
indianPerCap
AsianPerCap
OtherPerCap
HispPerCap
NumUnderPov
PctPopUnderPov
PctLess9thGrade
PctNotHSGrad
PctBSorMore
PctUnemployed
PctEmploy
PctEmplManu
PctEmplProfServ
PctOccupManu
PctOccupMgmtProf
MalePctDivorce
MalePctNevMarr
FemalePctDiv
TotalPctDiv
PersPerFam
PctFam2Par
PctKids2Par
PctYoungKids2Par
PctTeen2Par
PctWorkMomYoungKids
PctWorkMom
NumIlleg
PctIlleg
NumImmig
PctImmigRecent
PctImmigRec5
PctImmigRec8
PctImmigRec10
PctRecentImmig
PctRecImmig5
PctRecImmig8
PctRecImmig10
PctSpeakEnglOnly
PctNotSpeakEnglWell
PctLargHouseFam
PctLargHouseOccup
PersPerOccupHous
PersPerOwnOccHous
PersPerRentOccHous
PctPersOwnOccup
PctPersDenseHous
PctHousLess3BR
MedNumBR
HousVacant
PctHousOccup
PctHousOwnOcc
PctVacantBoarded
PctVacMore6Mos
MedYrHousBuilt
PctHousNoPhone
PctWOFullPlumb
OwnOccLowQuart
OwnOccMedVal
OwnOccHiQuart
RentLowQ
RentMedian
RentHighQ
MedRent
MedRentPctHousInc
MedOwnCostPctInc
MedOwnCostPctIncNoMtg
NumInShelters
NumStreet
PctForeignBorn
PctBornSameState
PctSameHouse85
PctSameCity85
PctSameState85
LandArea
PopDens
PctUsePubTrans
LemasPctOfficDrugUn
ViolentCrimesPerPop
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8

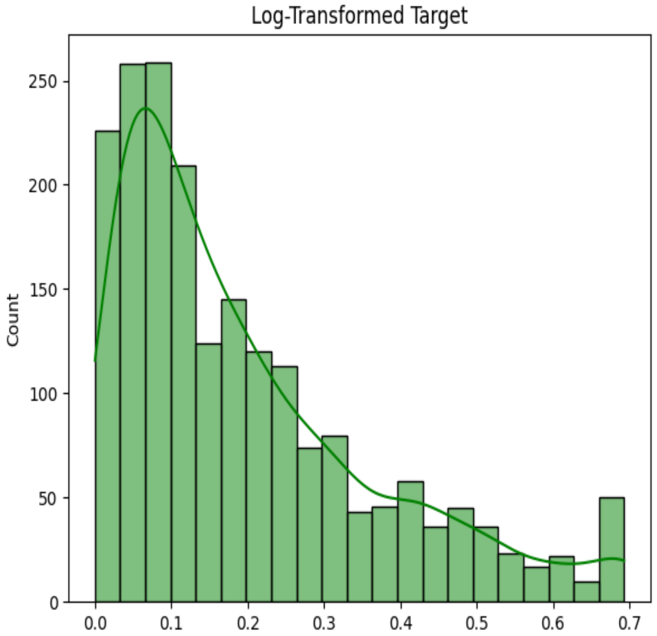
# Data Splitting :

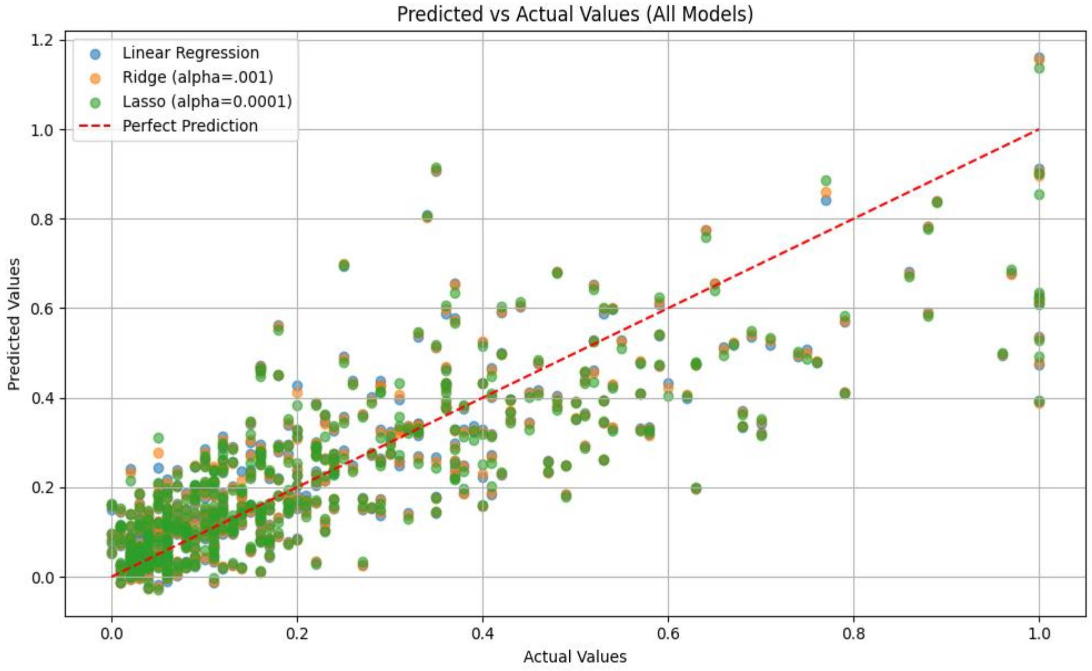


# Linear Regression Family:




**Input shape : 1495 X 100**

**Cross validation : 5**

Lasso alpha: 0.0001
Ridge alpha: .0001

| Model | Train $R^2$ | Test $R^2$ | MAE | RMSE |
|---|---|---|---|---|
| **Linear Regression** | 0.698 | 0.652 | 0.0937 | 0.1301 |
| **Ridge** | 0.693 | 0.655 | 0.0927 | 0.1296 |
| **Lasso** | 0.689 | 0.653 | 0.0924 | 0.1299 |

## Dimensionality Reduction:

- **Objective**: Perform PCA to reduce dimensionality.
- Identifies orthogonal components capturing maximum variance

**Condition:**

•**Standardization**: Z-score scaling (StandardScaler) to zero mean, unit variance.

**Pseudocode :**

- Compute covariance matrix. (100 x 100 )
- Perform eigenvalue decomposition ( extract most variances)
- Project data onto principal components
- Use projected data as a training data.

# Pseudocode : PCA

```
Compute Covariance Matrix:
    Input: Standardized data matrix X (n x p)
    Center X: X_centered = X - mean(X, axis=0)
    Compute C = (1/(n-1)) * X_centered^T * X_centered
    Return C
```

```
Eigenvalue Decomposition:
    Input: Covariance matrix C
    Compute eigenvalues λ_i and eigenvectors w_i of C
    Sort eigenvalues in descending order: λ_1 ≥ λ_2 ≥ ... ≥ λ_p
    Sort eigenvectors accordingly: W = [w_1, w_2, ..., w_p]
    Return W, λ_i
```

# Pseudocode : PCA

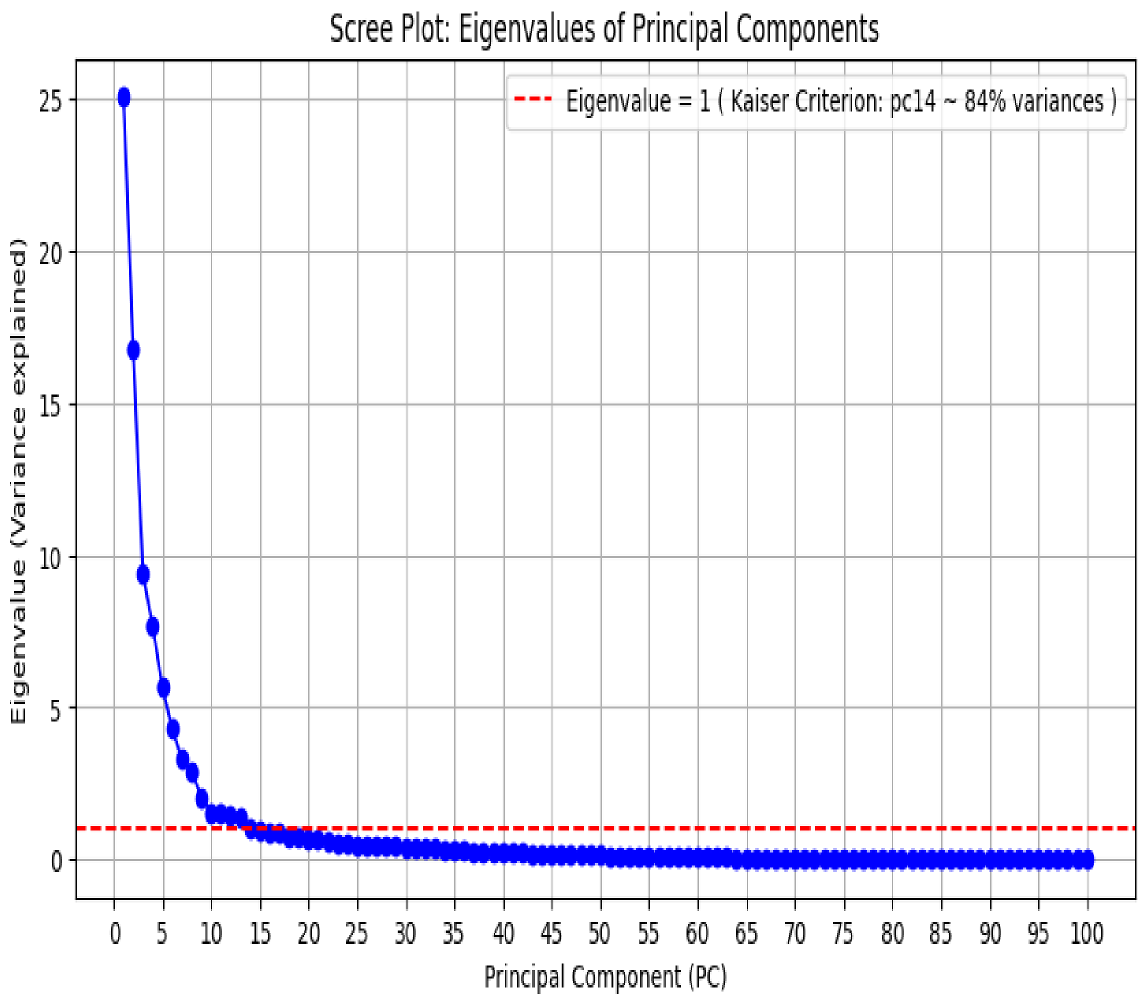
```
Compute Explained Variance Ratio:
    Input: Eigenvalues, number of components k
    Compute sum of top k eigenvalues: sum_top_k = sum(λ_i for i=1 to k)
    Compute total sum of eigenvalues: sum_total = sum(λ_i for i=1 to p)
    Compute ratio = sum_top_k / sum_total
    Return ratio
```

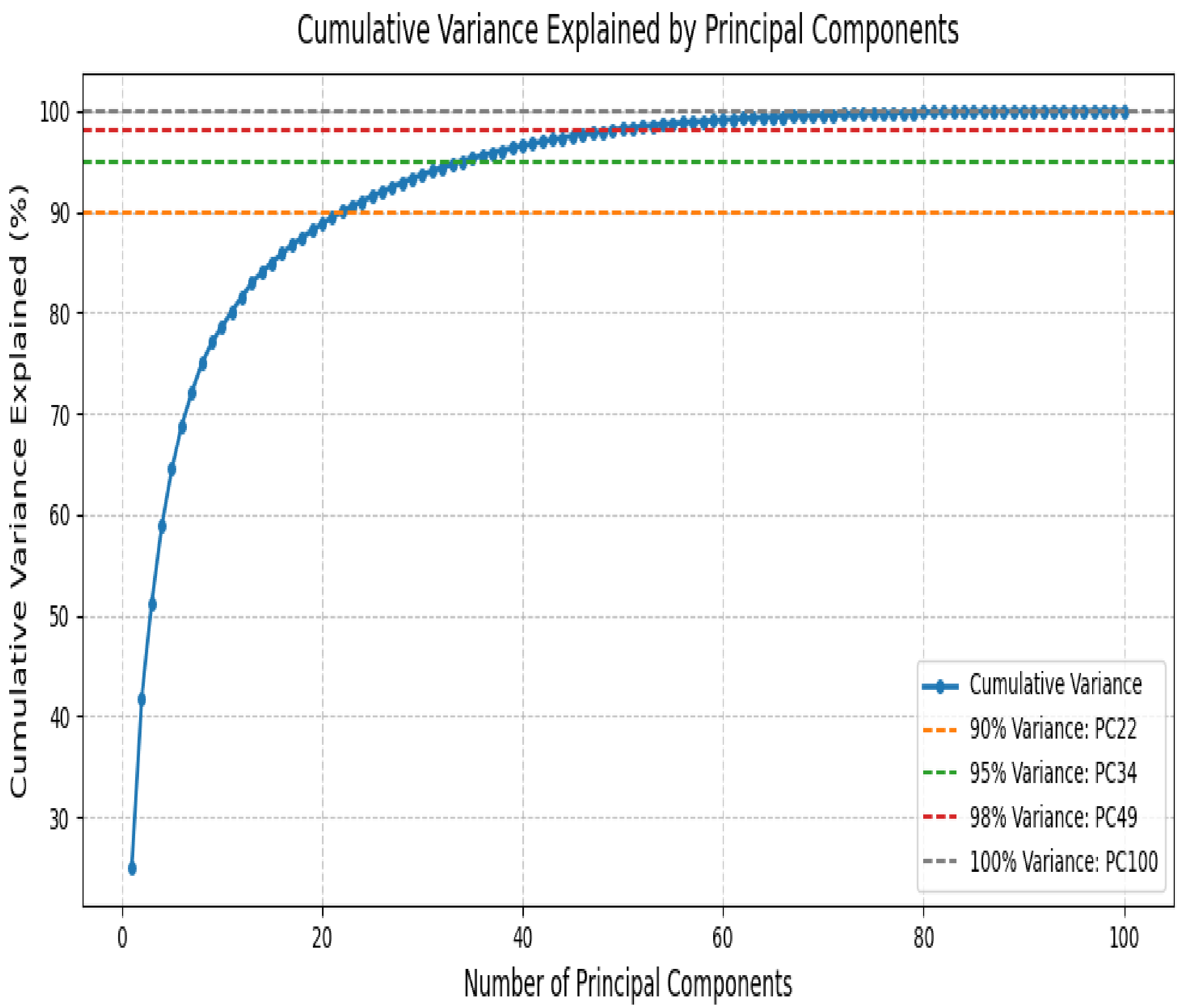
```
Project Data:
    Input: Centered data X, eigenvectors W (p x k)
    Compute scores: T = X * W
    Return T (n x k reduced data)
```

**Input shape : 1495 X 100**

- **X = n x p :** n=1495 , P=100
- **w=p x k** : n=100 , K= components.
- New Representation = **n x K**
- For pca5 X= 1495 x 5
- For pca10 X= 1495 x 10
- For pca34 X= 1495 x 34

# Dimensionality Reduction with PCA:



PCA Results and comparison:

**Results with 14 PCA Components**

| **Model** | **MAE** | **RMSE** | **$R^2$** |
|---|---|---|---|
| Linear Regression | 0.016377 | **0.028228** | **0.5759** |
| Ridge | **0.016212** | 0.028301 | 0.5737 |
| Lasso | **0.016124** | 0.028414 | 0.5703 |

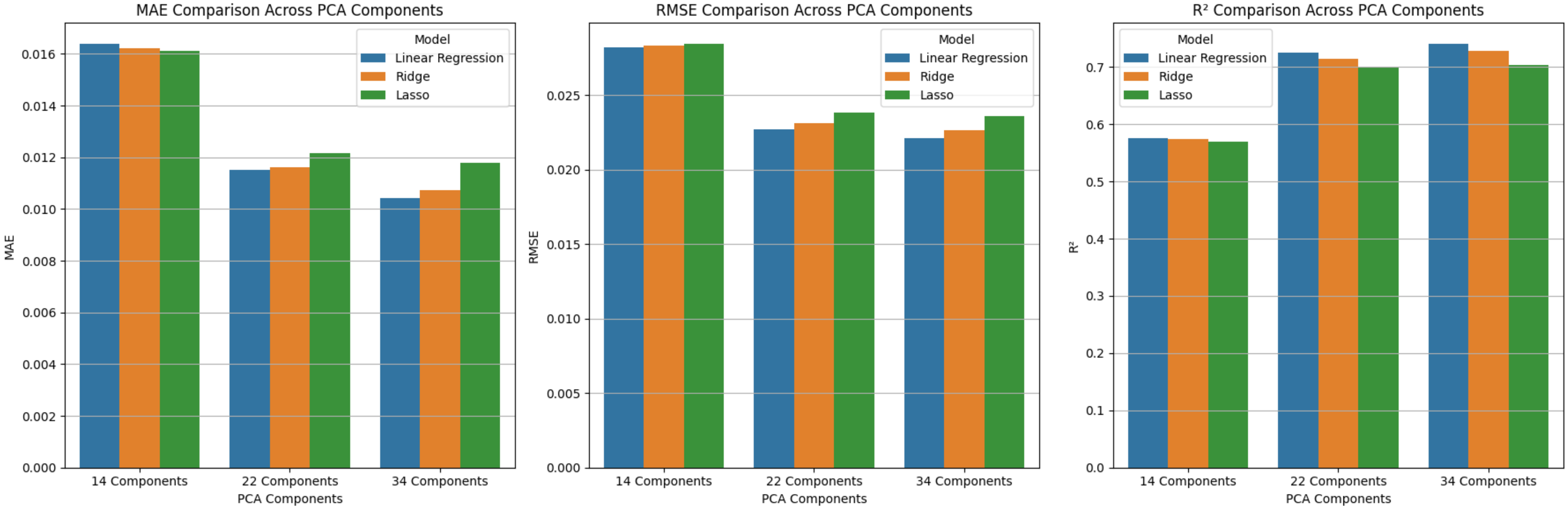
**Results with 22 PCA Components**

| Model | MAE | RMSE | $R^2$ |
|---|---|---|---|
| Linear Regression | 0.011505 | 0.022706 | 0.7256 |
| Ridge | 0.011627 | 0.023136 | 0.7151 |
| Lasso | 0.012158 | 0.023802 | 0.6985 |

**Results with 34 PCA Components**

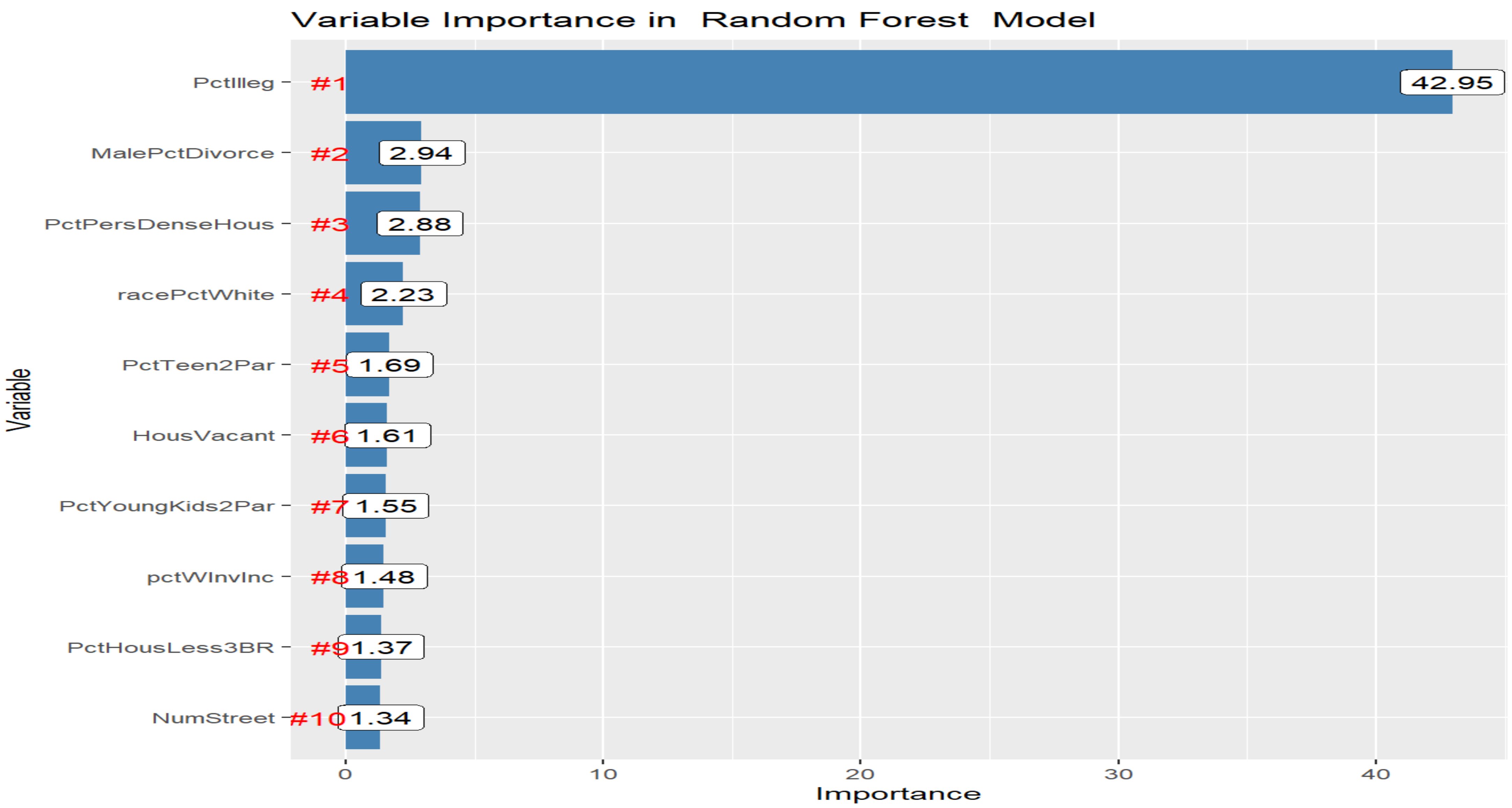
| **Model** | **MAE** | **RMSE** | **$R^2$** |
|---|---|---|---|
| Linear Regression | **0.010427** | **0.022100** | **0.7400** |
| Ridge | 0.010731 | 0.022612 | 0.7279 |
| Lasso | 0.011801 | 0.023611 | 0.7033 |

# PCA Results and comparison:




Linear Regression | **0.010427** | **0.022100** | **0.7400**

## **Optimal conditions minimizing the crime rate.**



# Conclusion

- A data-driven approach was used to predict violent crime rates using socio-economic and law enforcement data from 1,994 U.S. communities.
- Tree-based models (Random Forest) identified **family structure**, **education**, and **housing conditions** as key predictors of violent crime.
- The **percentage of children born to unmarried parents (`PctIlleg`)** consistently ranked as the most important factor influencing crime.
- While some racial features appeared important, they likely reflect **underlying economic inequalities** rather than race itself.

# Thank you for your attention

Any questions?