

STAT467 Project

Communities and Crime

Group 28

Seyyed Mohammad Tavallaiyan

Hermann Diderot Ndamen Nounang

Variables

state	US state (by number) - not counted as predictive
urban	if the area is classified as urban: 1 if sub-urban: 0
householdsize	mean people per household
perCapInc	per capita income
PctPopUnderPov	percentage of people under the poverty level
PctNotSpeakEnglWell	percent of people who do not speak English well
PctPersOwnOccup	percent of people in owner occupied households
MedRentPctHousInc	median gross rent as a percentage of household income
PctForeignBorn	percent of people foreign born
PolicPerPop	police officers per 100K population
PolicBudgPerPop	police operating budget per population
ViolentCrimesPerPop	total number of violent crimes per 100K population (Response)

Variables

- 2 nominal independent variables
- 9 continuous independent variables
- 1 continuous response variable

- All continuous variables are normalized (from 0 to 1)

- Normalization preserves rough ratios of values WITHIN an attribute
- Normalization does not preserve relationships between values BETWEEN attributes

- Values 3 SD above mean set to 1
- Values 3 SD below mean set to 0

Research Question: Socioeconomic Factors and Crime Rate

- How the socioeconomic factors can contribute to the crime rate of communities?
- What factors can effectively predict the crime rate?

Summary of variables

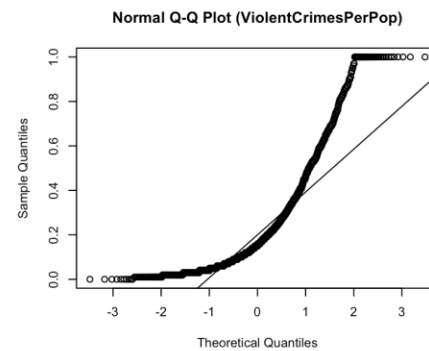
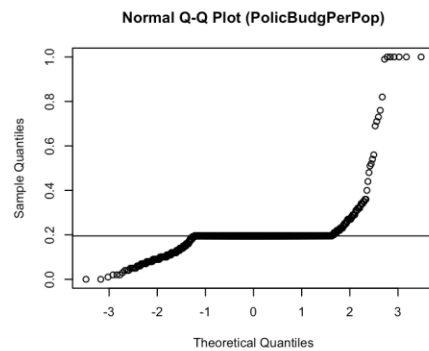
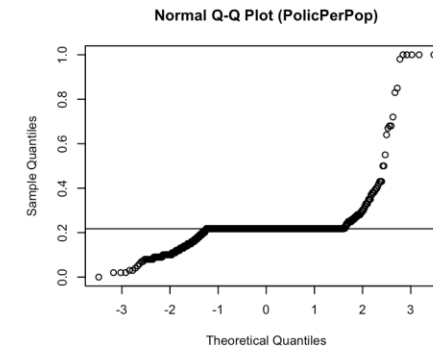
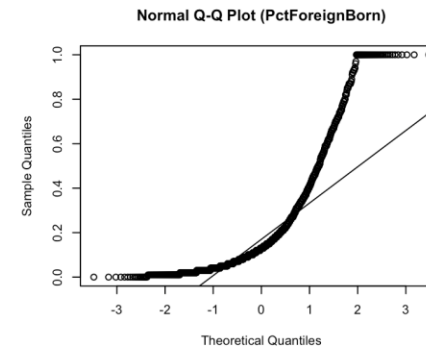
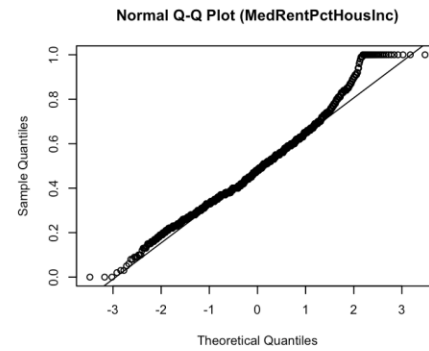
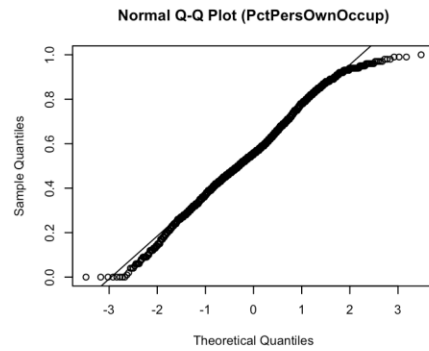
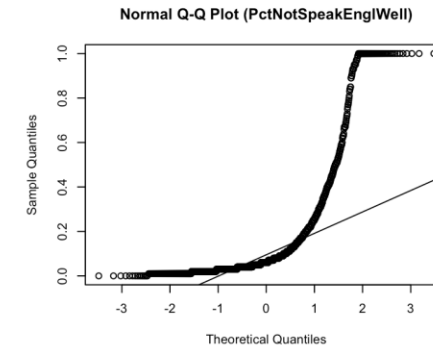
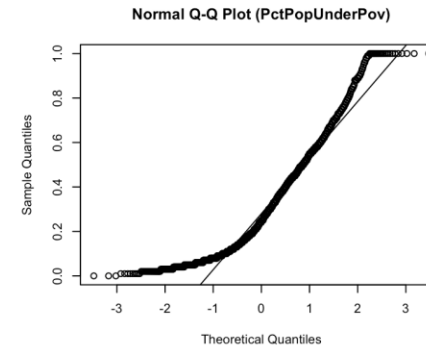
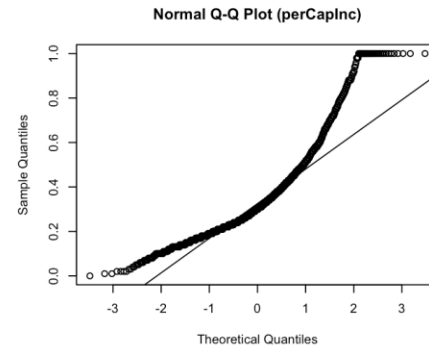
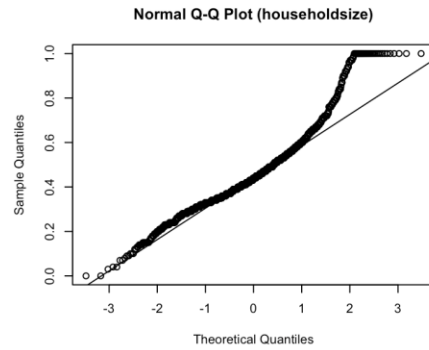
```
> summary(CC)
```

state	urban	householdsize	perCapInc	PctPopUnderPov	PctNotSpeakEnglWell
Length:1994	Length:1994	Min. :0.0000	Min. :0.0000	Min. :0.000	Min. :0.0000
Class :character	Class :character	1st Qu.:0.3500	1st Qu.:0.2200	1st Qu.:0.110	1st Qu.:0.0300
Mode :character	Mode :character	Median :0.4400	Median :0.3000	Median :0.250	Median :0.0600
		Mean :0.4634	Mean :0.3503	Mean :0.303	Mean :0.1506
		3rd Qu.:0.5400	3rd Qu.:0.4300	3rd Qu.:0.450	3rd Qu.:0.1600
		Max. :1.0000	Max. :1.0000	Max. :1.000	Max. :1.0000

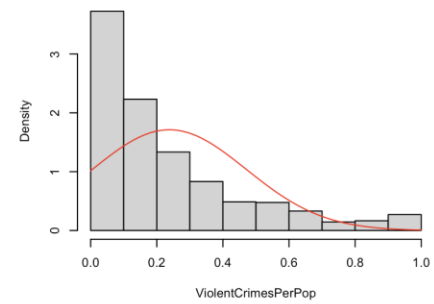
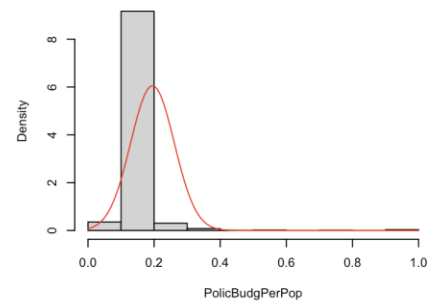
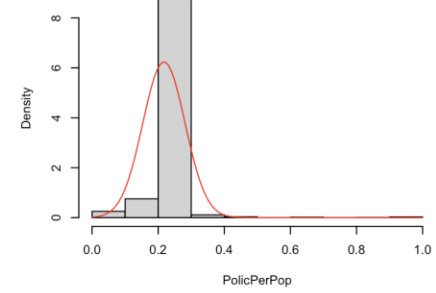
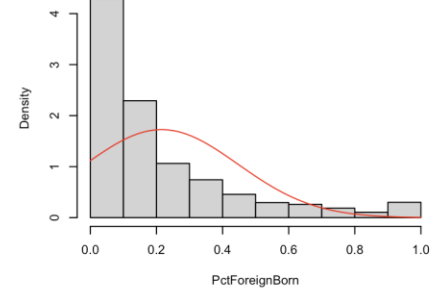
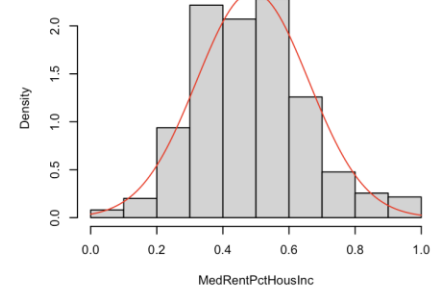
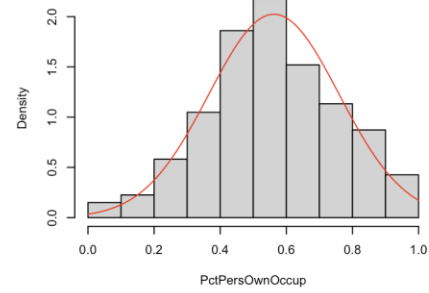
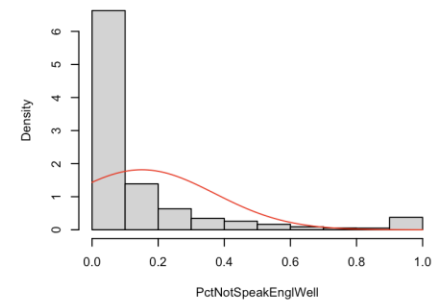
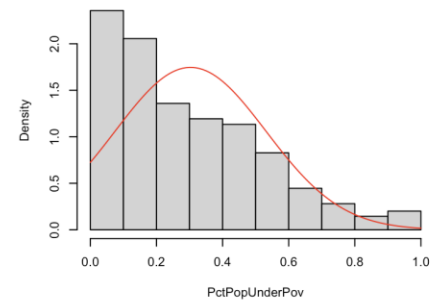
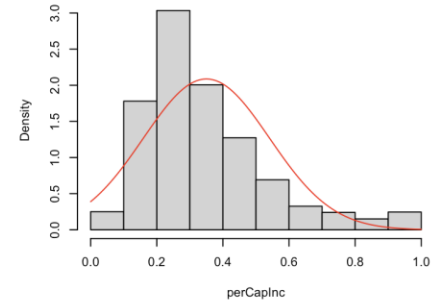
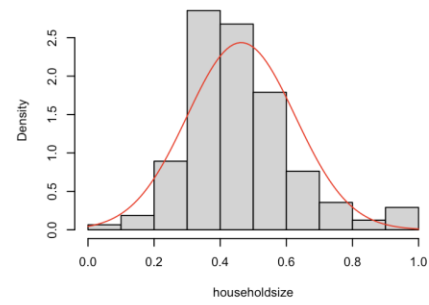
PctPersOwnOccup	MedRentPctHousInc	PctForeignBorn	PolicPerPop	PolicBudgPerPop	ViolentCrimesPerPop
Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.000
1st Qu.:0.4400	1st Qu.:0.3700	1st Qu.:0.0600	1st Qu.:0.1300	1st Qu.:0.1100	1st Qu.:0.070
Median :0.5600	Median :0.4800	Median :0.1300	Median :0.1800	Median :0.1500	Median :0.150
Mean :0.5626	Mean :0.4901	Mean :0.2156	Mean :0.2175	Mean :0.1951	Mean :0.238
3rd Qu.:0.7000	3rd Qu.:0.5900	3rd Qu.:0.2800	3rd Qu.:0.2500	3rd Qu.:0.2200	3rd Qu.:0.330
Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.000

NA's :1675	NA's :1675
------------	------------

The QQ plots for MVN

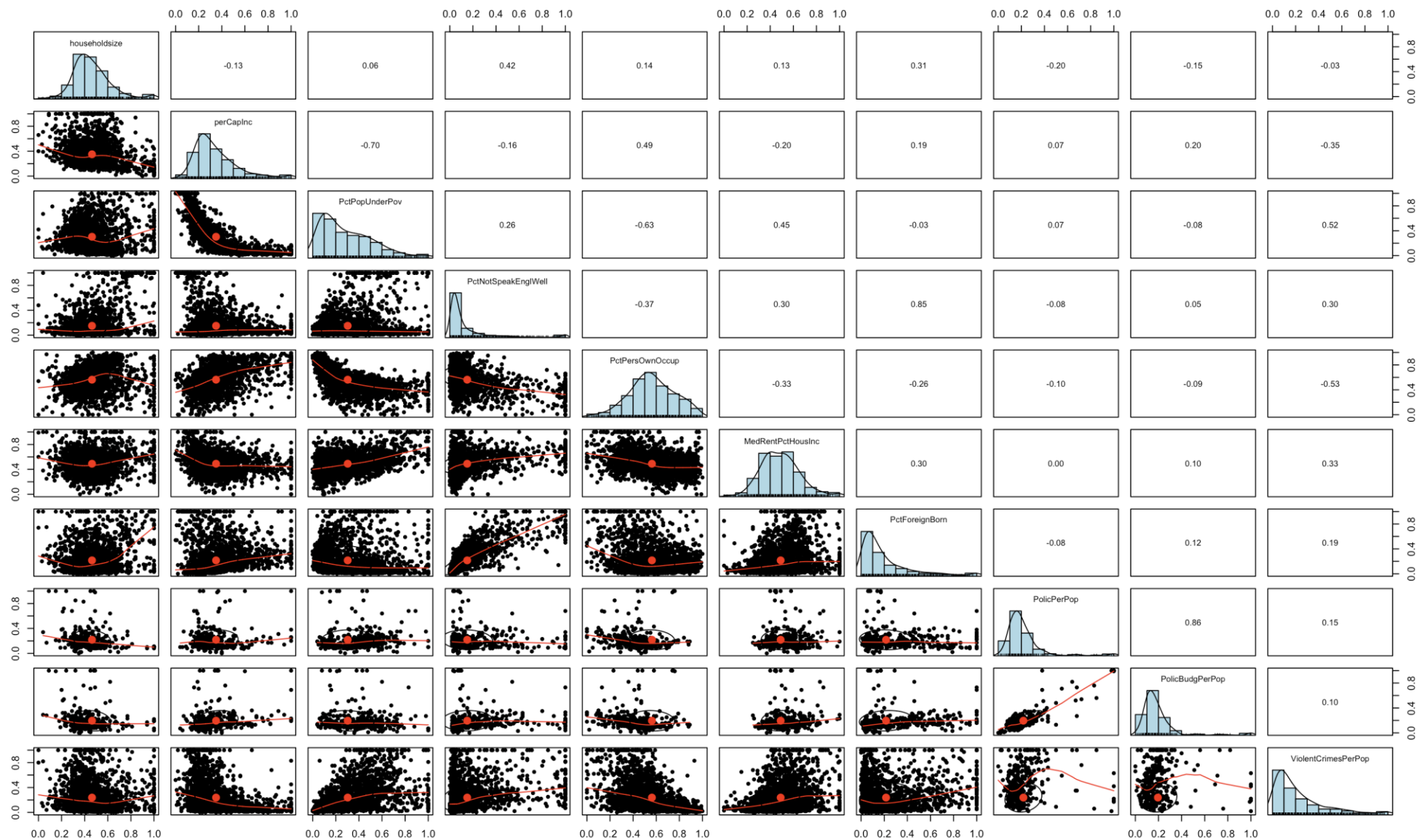


Histogram plots

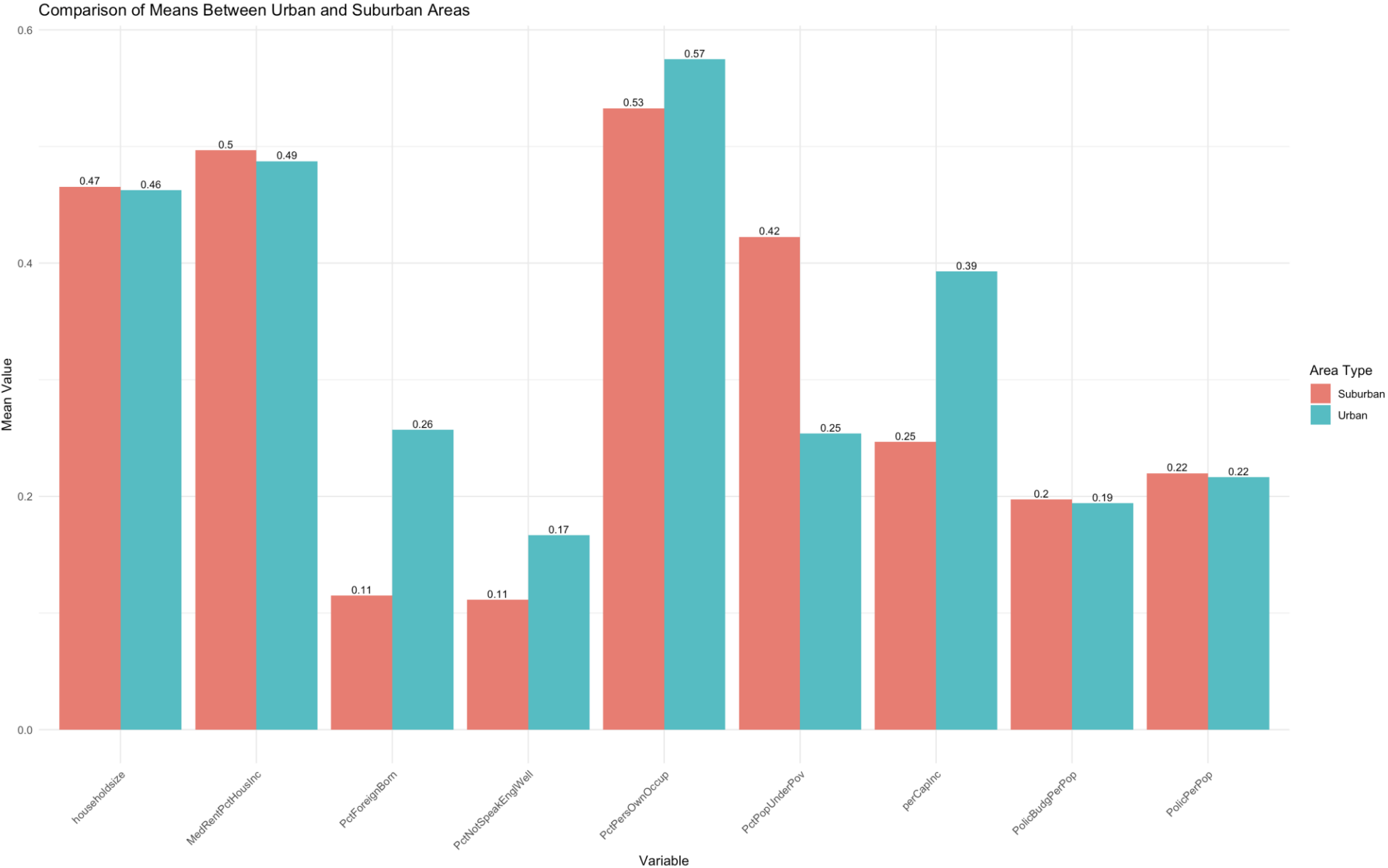


Scatterplot Matrix of Predictors and Response

Scatterplot Matrix of Predictors And Response



Urban vs Suburban Comparison



Hotelling's T2-test Results

Test stat: 533.77

Numerator df: 9

Denominator df: 1984

P-value: 0

|

Individual tests Results

Significant (P-val < 0.05) ▾		Insignificant (P-val > 0.05) ▾	
perCapInc		householdsize	
PctPopUnderPov		MedRentPctHousInc	
PctNotSpeakEnglWell		PolicPerPop	
PctPersOwnOccup		PolicBudgPerPop	
PctForeignBorn			

MANOVA Results

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> # Check the p-values for each dependent variable

> summary(manova_result, test = "Wilks") # Use Wilks's lambda for robustness

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
urban	1	0.80936	117.12	4	1989	< 2.2e-16 ***

Residuals 1992

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Individual MANOVA tests Results

Significant (P-val < 0.05) ▼		Insignificant (P-val > 0.05) ▼	
perCapInc		householdsize	
PctPopUnderPov		MedRentPctHousInc	
PctNotSpeakEnglWell		PolicPerPop	
PctPersOwnOccup		PolicBudgPerPop	
PctForeignBorn			

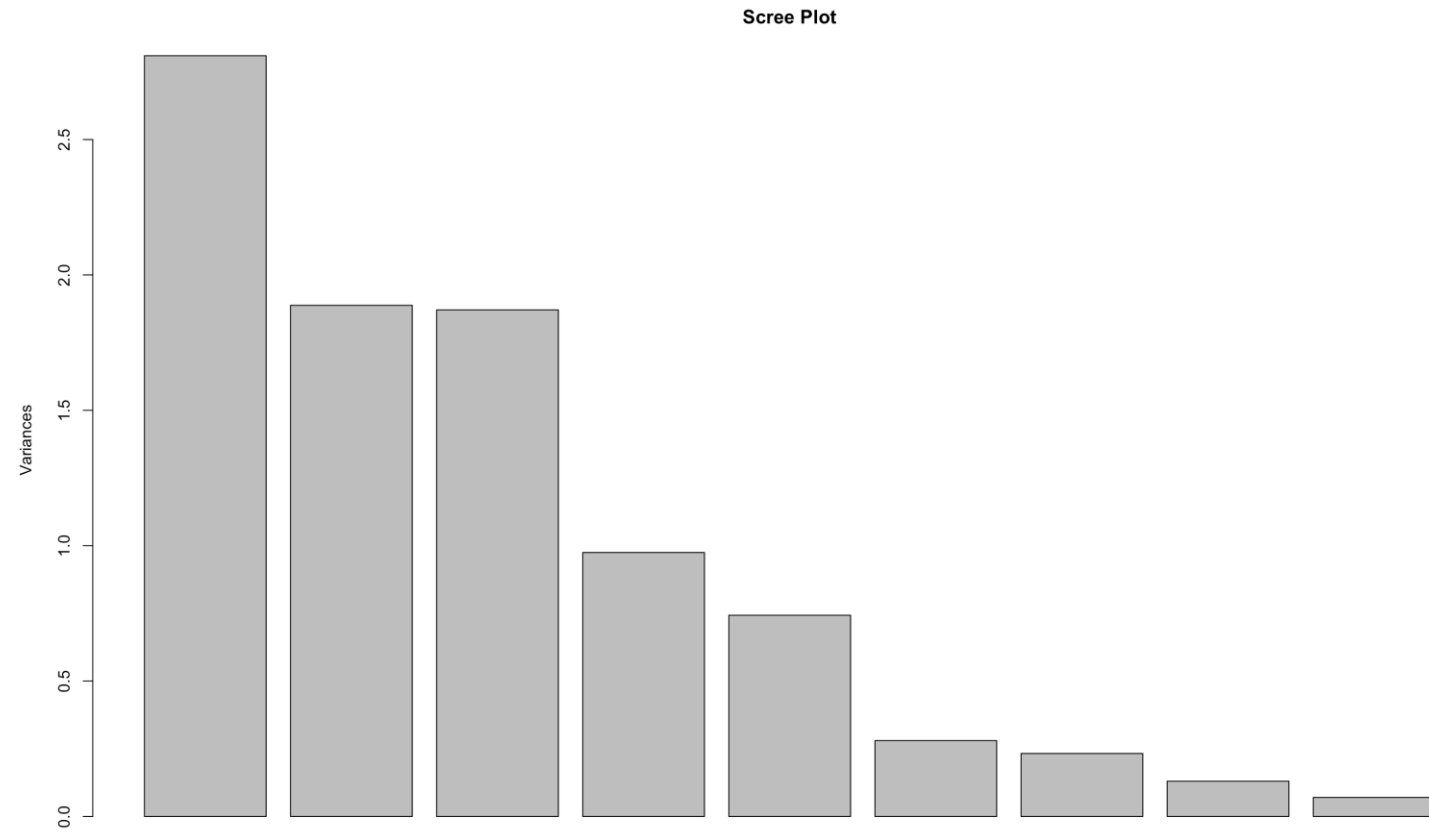
PCA Summary

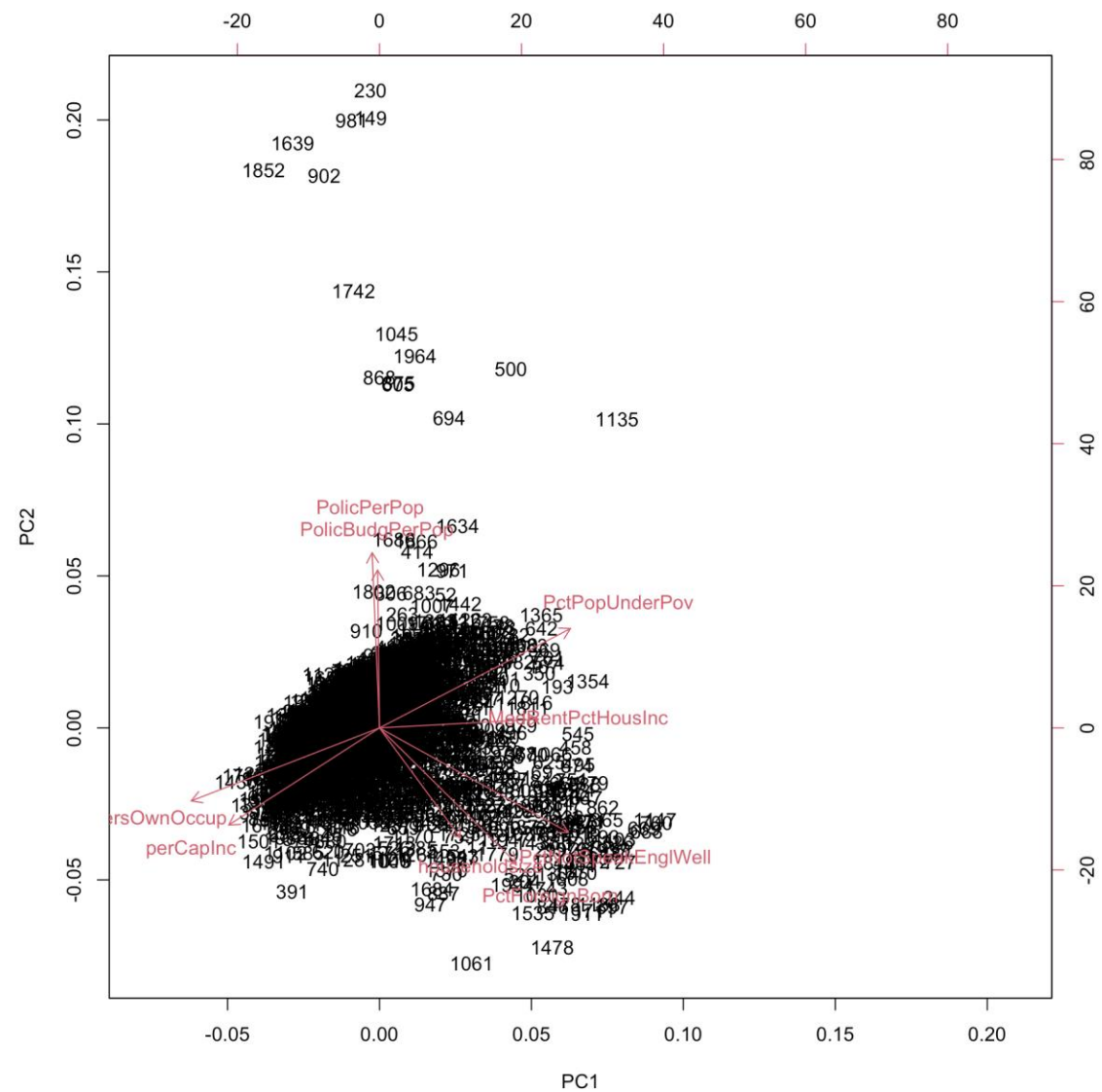
```
> summary(pca_result)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1.6762	1.3740	1.3678	0.9874	0.86211	0.52937	0.48237	0.36095	0.26535
Proportion of Variance	0.3122	0.2098	0.2079	0.1083	0.08258	0.03114	0.02585	0.01448	0.00782
Cumulative Proportion	0.3122	0.5219	0.7298	0.8381	0.92071	0.95185	0.97770	0.99218	1.00000

Scree plot of Principal Components



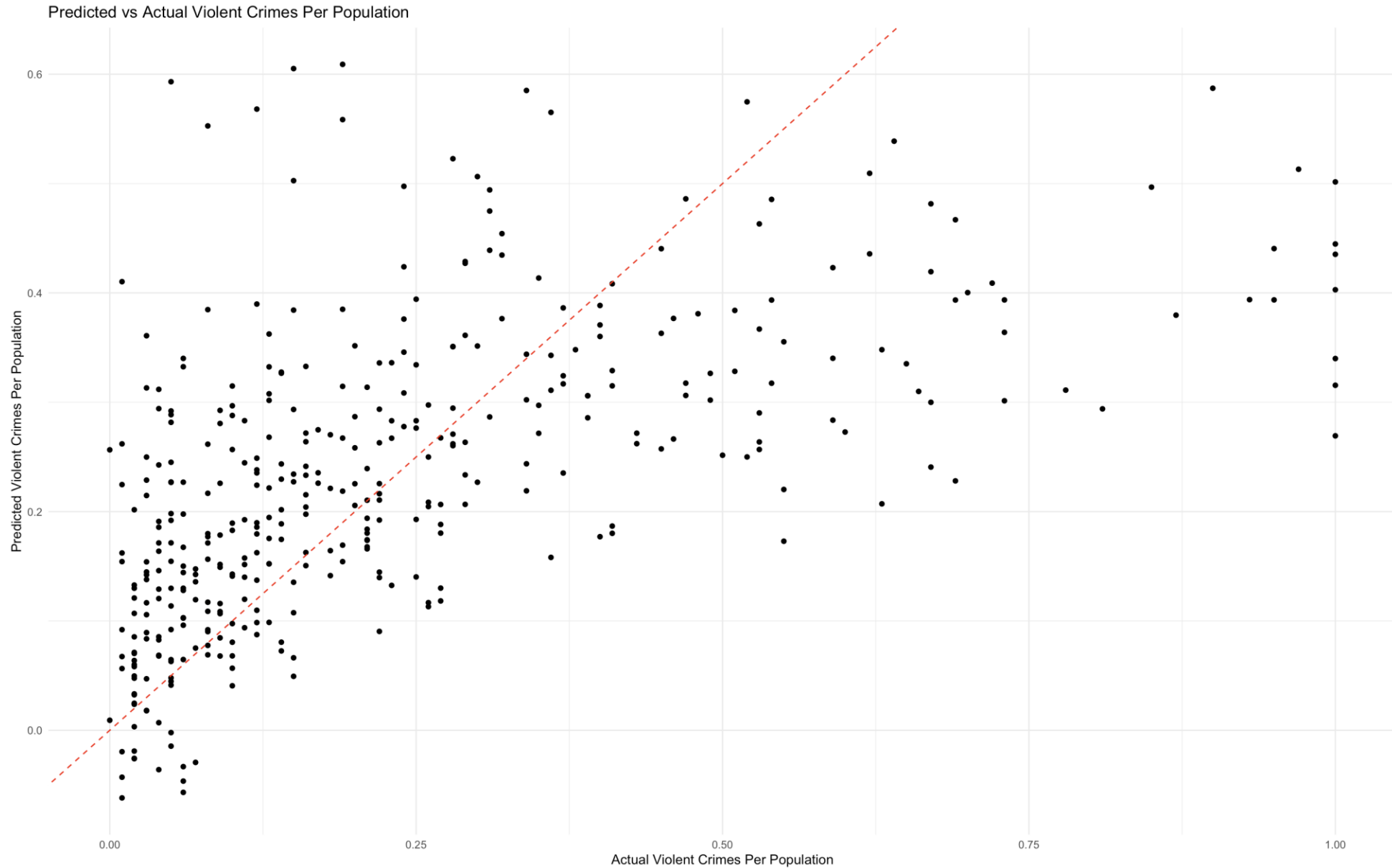


```
> print(loadings)
```

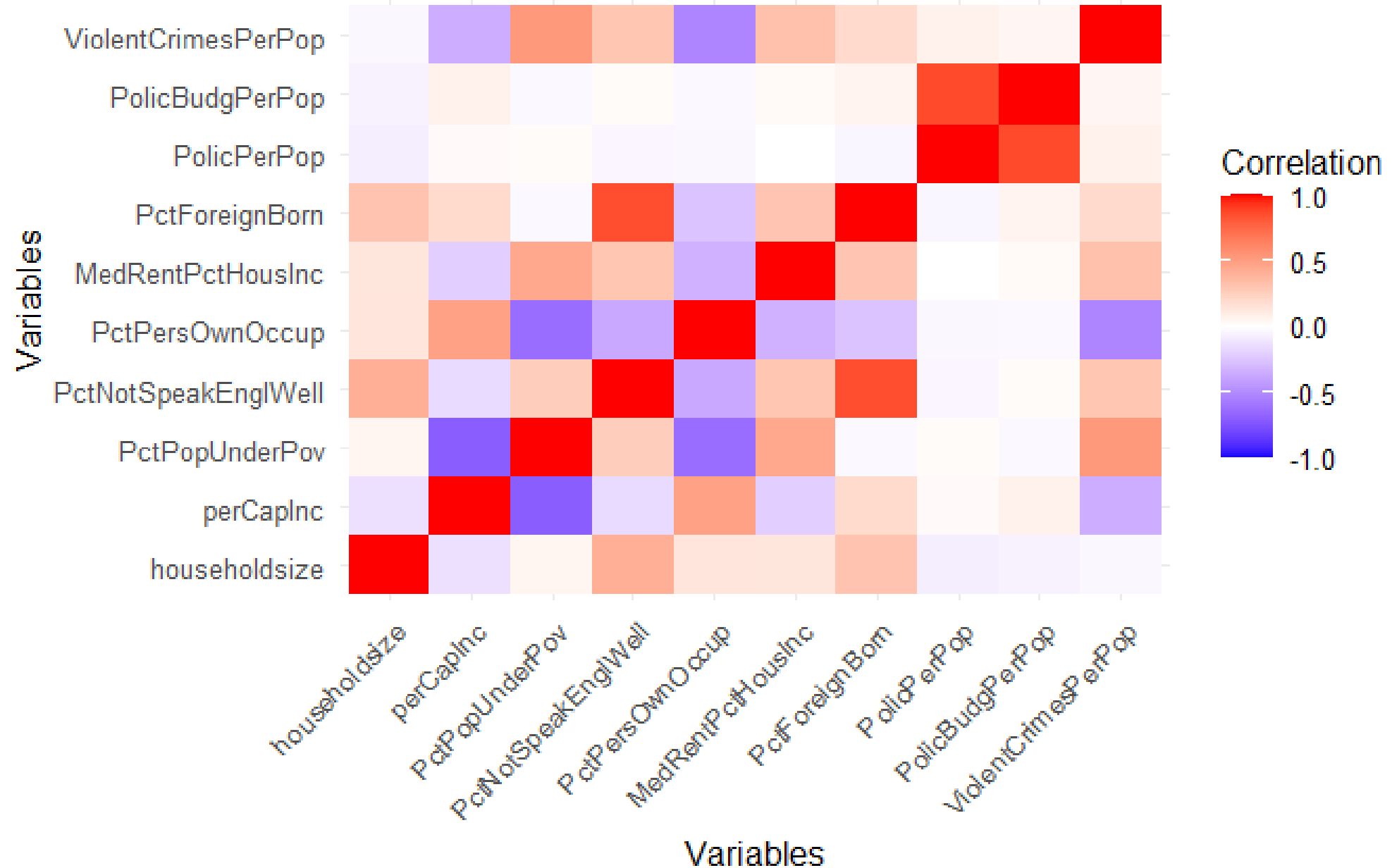
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
householdsize	0.190652951	-0.31568288	-0.16302610	-0.78099342	-0.06179181	-0.47135288	-0.03277928	-0.01531563	0.01291190
perCapInc	-0.354312832	-0.27847074	-0.33804090	0.36261038	-0.23275363	-0.44925423	0.50393104	-0.04361612	-0.19825753
PctPopUnderPov	0.449108694	0.28486884	0.26275586	-0.09115639	-0.04771316	0.01321306	0.76615287	-0.12817862	0.18420066
PctNotSpeakEnglWell	0.443904686	-0.29937854	-0.28663976	0.03432352	0.26758546	0.35967504	0.12494234	0.04009195	-0.64171224
PctPersOwnOccup	-0.442571221	-0.20937035	-0.11486182	-0.38177472	-0.28747554	0.65646884	0.26932215	-0.02065576	0.12163467
MedRentPctHousInc	0.373806280	0.02477265	-0.02572945	0.12749615	-0.87974584	0.05852900	-0.22504993	0.05274696	-0.10922365
PctForeignBorn	0.321630674	-0.38780204	-0.40889395	0.26159800	0.10552663	0.10459033	-0.03223552	0.02241303	0.69779584
PolicPerPop	-0.017215085	0.50245464	-0.48706707	-0.11543841	0.01936575	-0.01442233	0.08861628	0.69826490	0.02593984
PolicBudgPerPop	-0.004561843	0.45245981	-0.53820443	-0.07286051	0.00481249	0.03053289	-0.10151030	-0.69895696	-0.02204669

RMSE value: 0.18

Predicted vs Actual Violent Crimes per Population

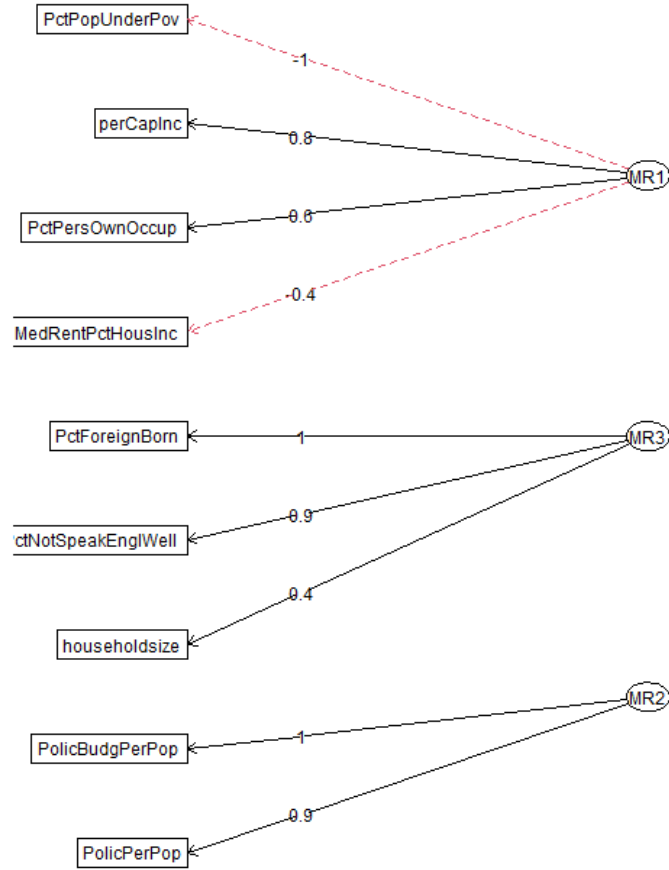


Correlation Matrix

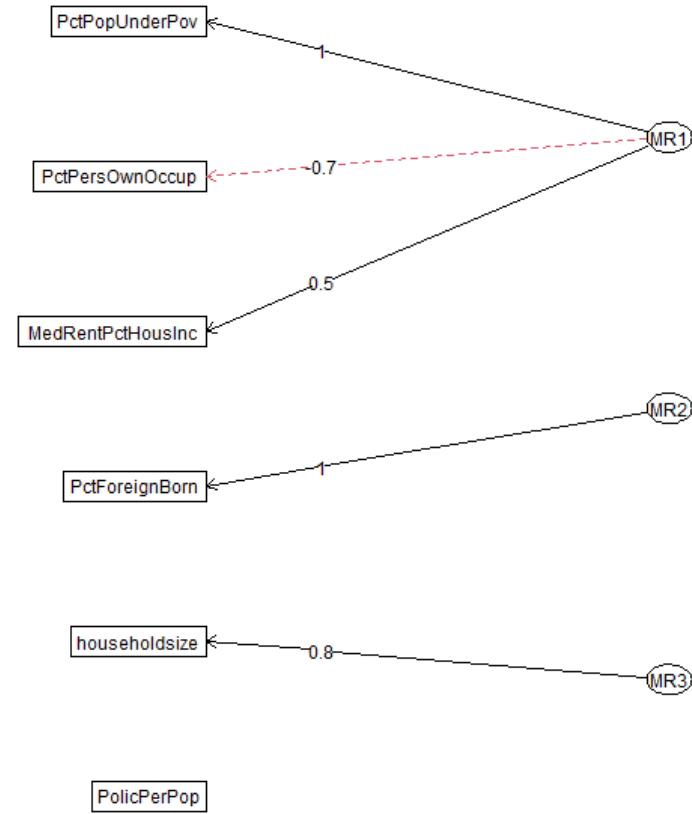


Factor Analysis

Factor Analysis



Factor Analysis



Factor Analysis

- **First model (9 vars)**

Rotation: promax

Cumulative Proportion var: 0.66

RMSR: 0.05 (good fit)

Tucker Lewis Index (TLI): 0.653 (moderate fit)

Mean item complexity = 1.1

- **Second model (6 vars)**

Rotation: varimax

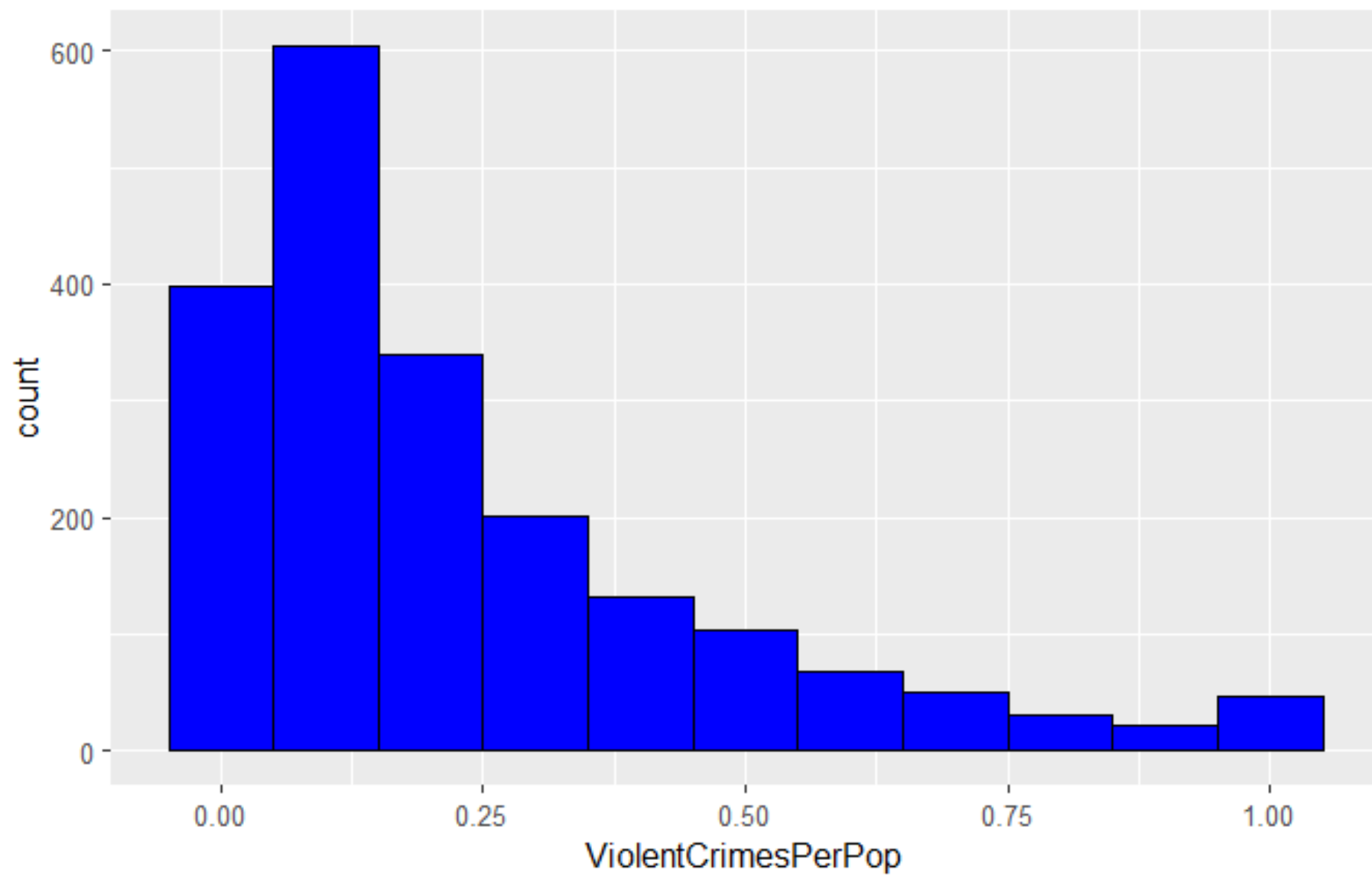
Cumulative Proportion var: 0.58

RMSR: 0.01 (excellent fit)

TLI: -Inf (fitting issues)

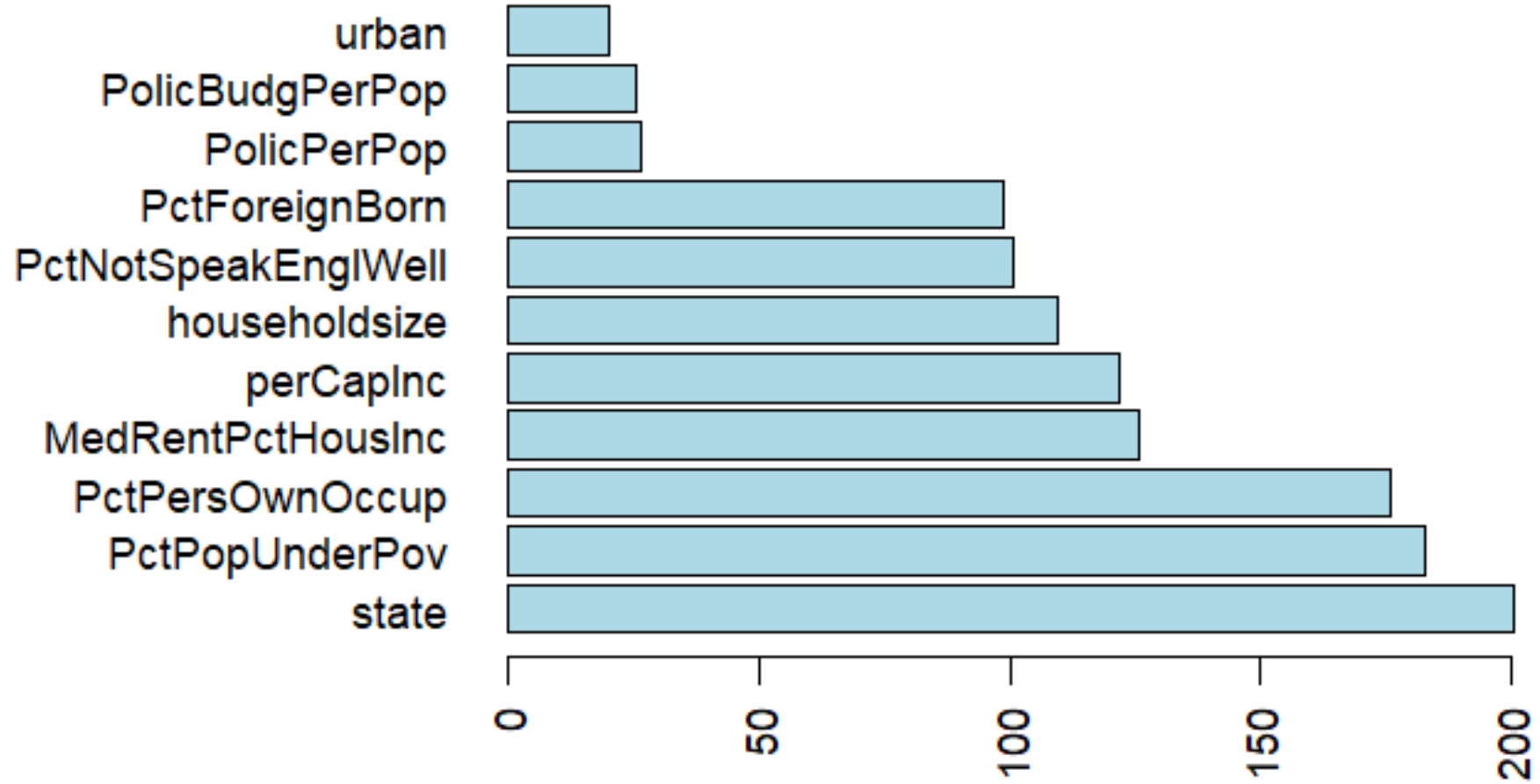
Mean item complexity = 1.3

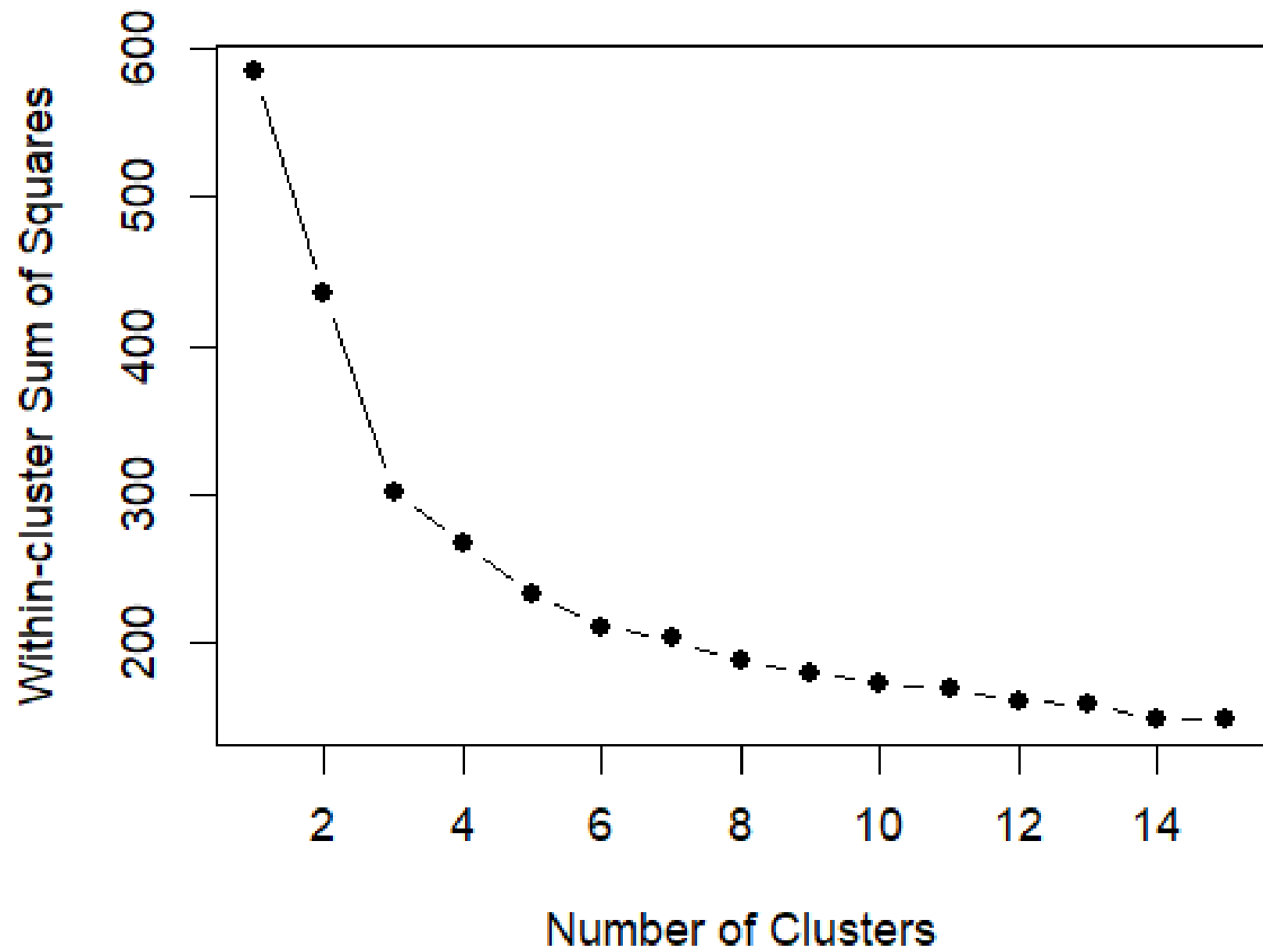
Distribution of Violent Crimes per Population



- **Continuous Response Variable (Regression):**
 - Mean Absolute Error (MAE): 0.0986
 - R-squared: 0.6184
- **Categorical Response Variable (Classification-Random Forests method):**
 - Accuracy: 0.5995
 - Sensitivity: 0.6942
 - Specificity: 0.9094
 - Balanced Accuracy: 0.8018

Feature Importance







Contributing factors to 2nd cluster

- **Household Size Mean: 62.30%** – Larger average household size.
- **Per Capita Income Mean: 28.10%** – Lower income level.
- **Percentage of Population Under Poverty: 42.67%** – High poverty rate.
- **Percentage Not Speaking English Well: 67.97%** – High percentage of non-English speakers.
- **Percentage of Owner-Occupied Housing: 36.86%** – Lower homeownership rate.
- **Median Rent as Percent of Household Income: 62.60%** – High rent burden.
- **Percentage Foreign Born: 73.17%** – High proportion of foreign-born residents.
- **Police per Population: 21.24%** – Average police presence.
- **Police Budget per Population: 20.46%** – Average police budget.

- **Violent Crimes per Population: 42.11%** – Higher crime rate.

Main Conclusions

- Urban: causes significant difference among variables, not a good predictor.
- Police forces and budget: Not a strong predictor (noticeable missing values).
- Foreign born and non-fluent English: moderate predictor of high crime rate.
- Economical factors: the strongest predictor of high crime rate.

High household size

Low income

High poverty rate

Low rate of house ownership

High rent as a percentage of household income

Data Source

- Original dataset downloaded from UC Irvine Machine Learning Repository Archive
- Link to dataset:
<https://archive.ics.uci.edu/dataset/183/communities+and+crime>