Communities and Crime Rates

ANLY 506 COURSE PROJECT

GROUP #3

TONGTIAN FAN | YITING LI | FANGYA TAN | MENGYA TIAN | WENJUN XU | SHUQING YANG

Goal of the study

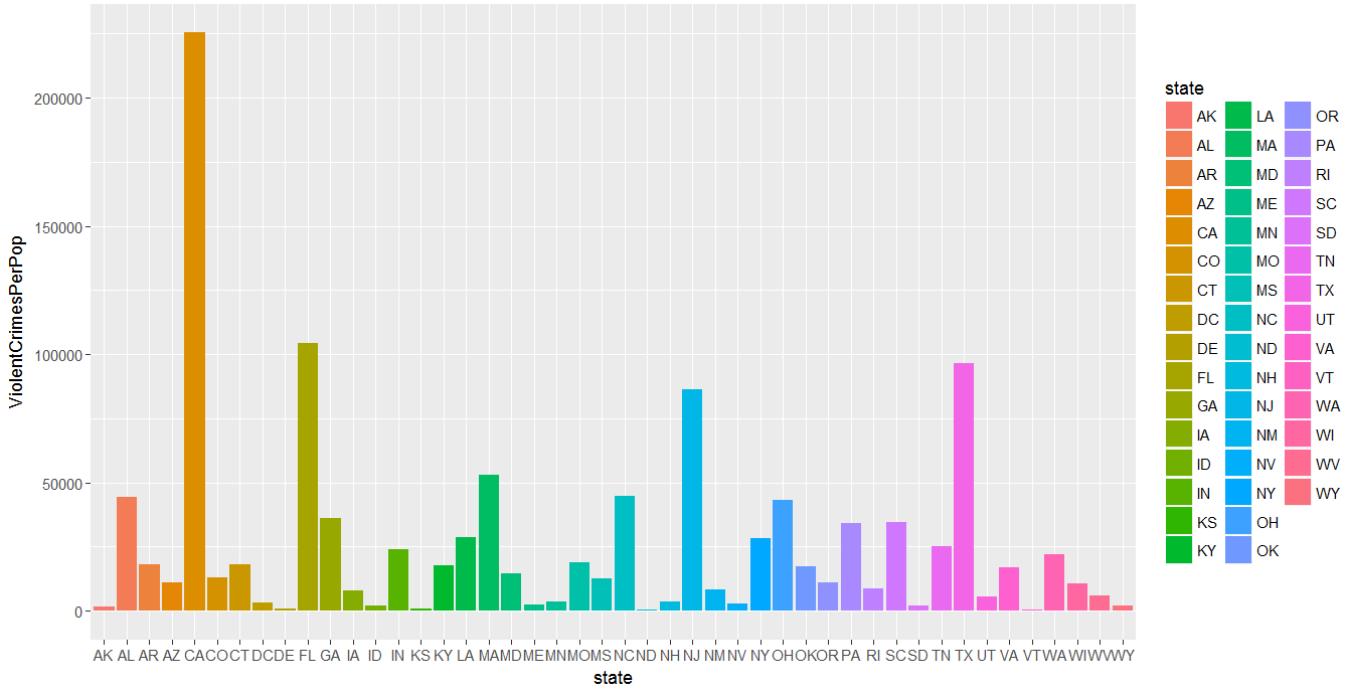
- Identify the correlations between crime rates and various risk factors
  - Population, ethnicity, age, income, education, marital status, housing, etc.
- Build a linear regression model to predict a community's violent and non-violent crime rate in the United States

Data Cleaning

- Communities and Crime Unnormalized Dataset
  - http://mlr.cs.umass.edu/ml/datasets/Communities+and+Crime+Unnormalized
- We were not able to update the data and obtain the national average crime rate
  - http://www.neighborhoodscout.com/
- 147 attributes (including 18 depending variables)
  - We selected ViolentCrimesPerPop and nonViolPerPop as our depending variables
  - Transferred character variables to numerical variable
- Raw data has 2215 observations, we cleaned the datasets by removing missing depending variables
  - Violent Crimes Dataset: 1994
  - Non Violent Crimes Dataset: 2118

EDA - Violent Crime Rate

- The Violent Crime Rate performs differently in different states. CA, FL, TX and NJ have the highest Violent Crime Rate.
- We found that crime rates differ by the type of community, such as metropolitan areas, cities outside of metropolitan areas, and rural areas.



EDA - Violent Crime Rate

Marital Status and race is positively correlated to Violent Crime Rate

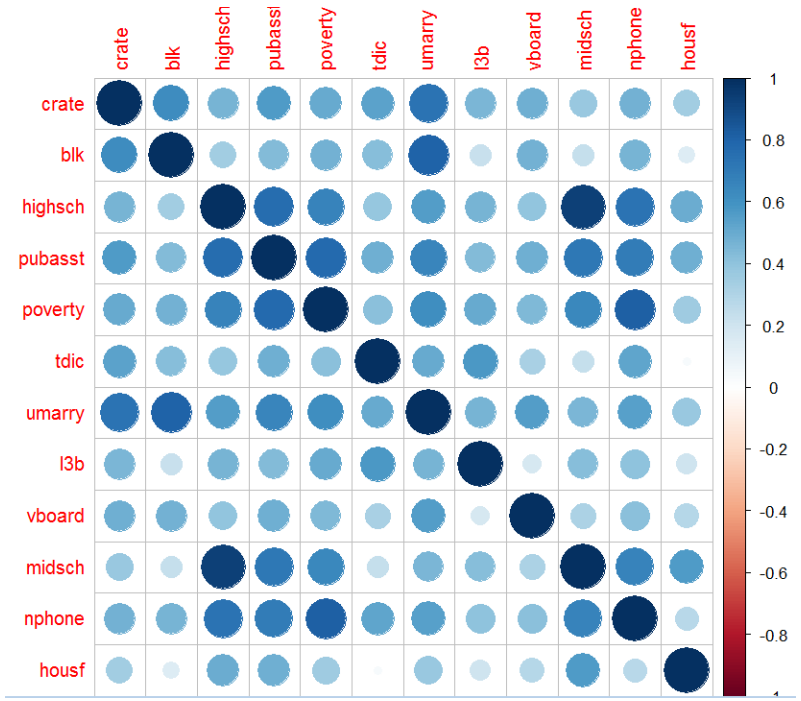
Variable	CORR
Percentage of kids born to never married	0.74
Total number of non-violent crimes per 100K population	0.68
Percentage of population that is African American	0.63
Percentage of households with public assistance income in 1989	0.56
Percentage of females who are divorced	0.54
Percentage of population who are divorced	0.54
Percent of police that are African America	0.51
Percentage of males who are divorced	0.51
Percentage of people under the poverty level	0.51
Percentage of people 16 and over, in the labor force, and unemployed	0.48
Percent of vacant housing that is boarded up	0.48

Income, Housing, Kids And Family are negatively correlated to Violent Crime Rate

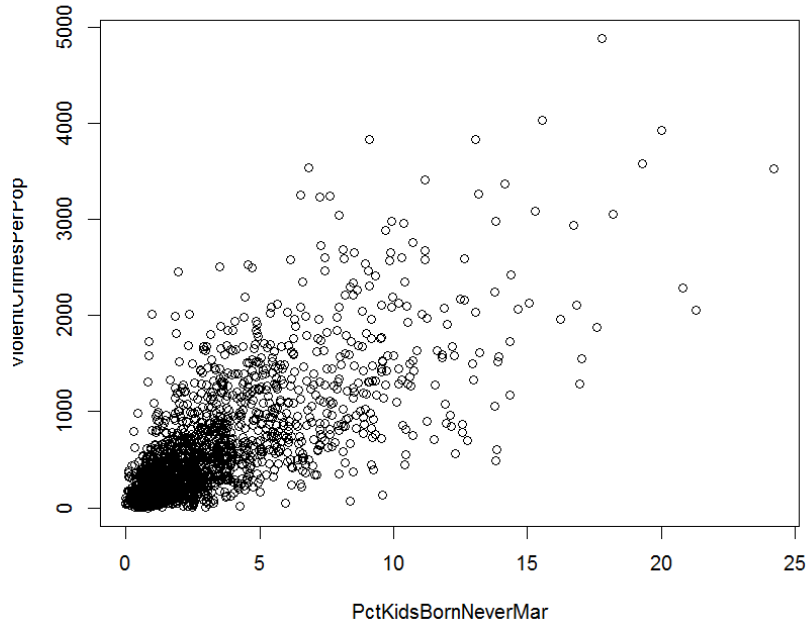
Variable	CORR
median household income	-0.40
median family income	-0.41
percent of households owner occupied	-0.46
a measure of the racial match between the community and the police force.	-0.47
percent of people in owner occupied households	-0.51
percentage of households with investment / rent income	-0.56
percent of kids age 12-17 in two parent households	-0.66
percent of kids 4 and under in two parent households	-0.66
percentage of population that is Caucasian	-0.68
percentage of families (with kids) that are headed by two parents	-0.70
percentage of kids in family housing with two parents	-0.73

EDA - Violent Crime Rate

Pearson Correlation Study

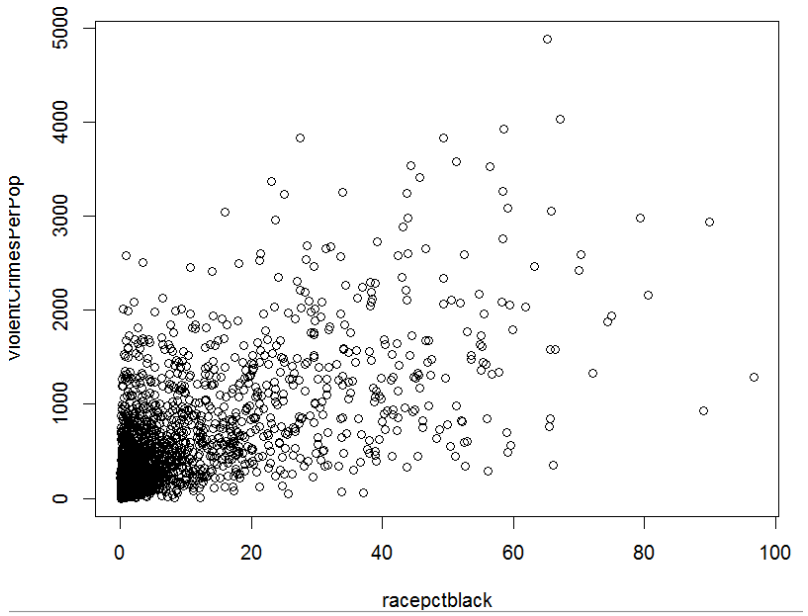


‘Percentage Of Kids Born To Never Married’ has the highest correlation with Violent Crime Rate

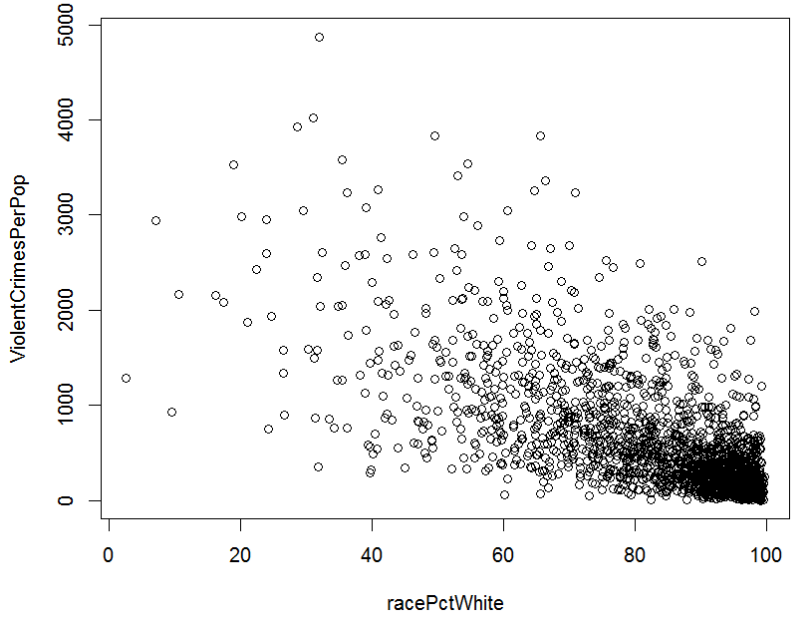


EDA - Violent Crime Rate

Correlation of Percentage Of Population That Is African America with Violent Crime Rate



Correlation of Percentage Of Population That Is Caucasian with Violent Crime Rate



Principal Component Analysis

- The first two components could explain more than 50% of the variance.
- We can explain about 90% of the variance by including 6 components.

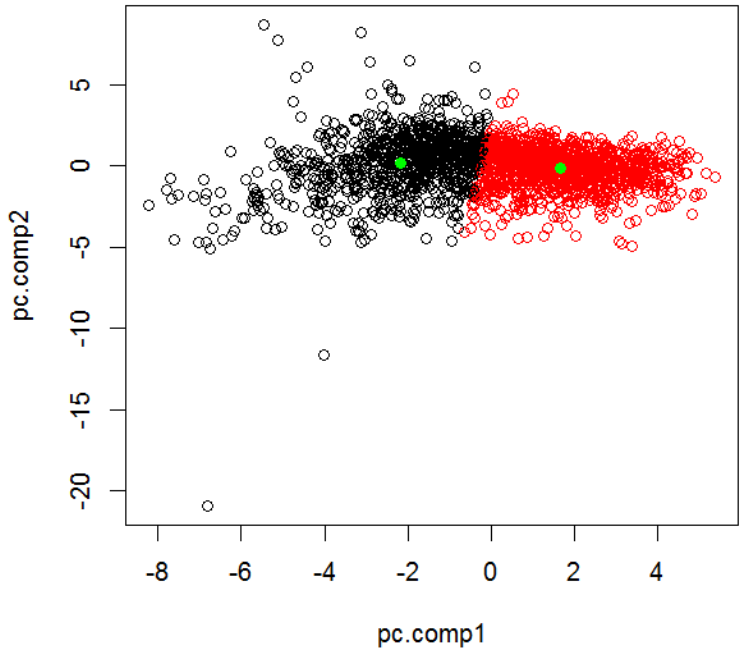
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	2.3498708	1.6146852	1.3818271	1.16392188	1.10062790	0.93312488	0.77930663
Proportion of Variance	0.3681262	0.1738139	0.1272964	0.09031428	0.08075879	0.05804814	0.04048792
Cumulative Proportion	0.3681262	0.5419401	0.6692365	0.75955077	0.84030956	0.89835769	0.93884561

	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13
Standard deviation	0.58090961	0.49914457	0.360134528	0.312521663	0.220583675	0.156644977
Proportion of Variance	0.02249706	0.01660969	0.008646459	0.006511319	0.003243811	0.001635843
Cumulative Proportion	0.96134268	0.97795237	0.986598825	0.993110144	0.996353955	0.997989798

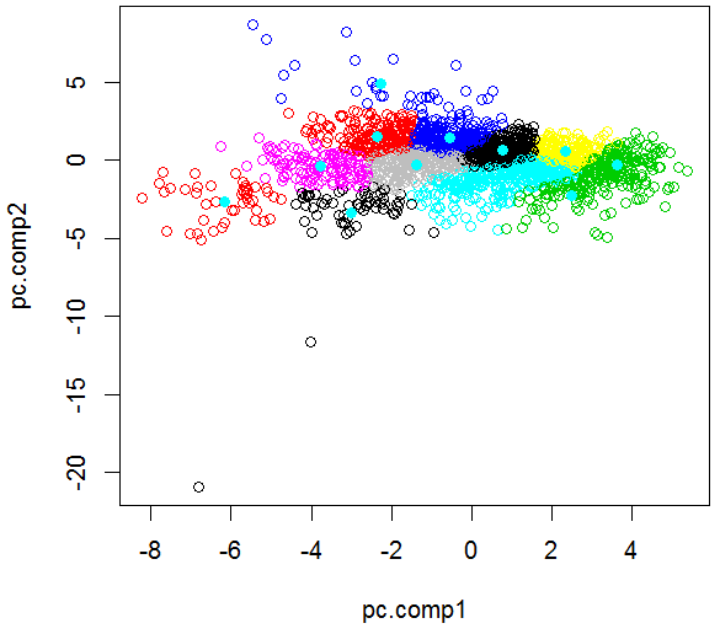
	Comp.14	Comp.15
Standard deviation	0.146437777	0.0933220819
Proportion of Variance	0.001429602	0.0005806007
Cumulative Proportion	0.999419399	1.0000000000

K-MEAN Clustering

K=2



K=13



Violent Crime Rate Regression Model

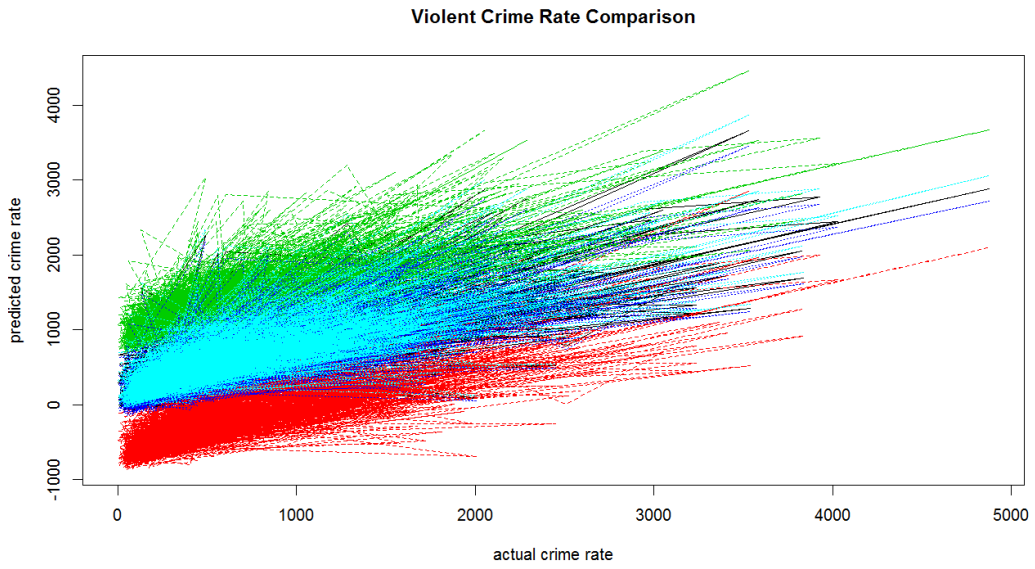
R^2= 0.59

```
lm(formula = crate ~ umarry + pubasst + blk + tdic + vboard,
    data = v2)

Residuals:
    Min      1Q  Median      3Q     Max
-1767.2  -205.7   -44.5   122.7  2249.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -293.428     34.557  -8.491  < 2e-16 ***
umarry        91.987      5.990  15.357  < 2e-16 ***
pubasst       12.197      2.789   4.374 1.28e-05 ***
blk            4.540      1.082   4.197 2.82e-05 ***
tdic          39.863      3.543  11.252  < 2e-16 ***
vboard        12.147      3.153   3.853 0.000121 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.4 on 1987 degrees of freedom
Multiple R-squared:  0.5915,    Adjusted R-squared:  0.5905
F-statistic: 575.5 on 5 and 1987 DF,  p-value: < 2.2e-16
```

Results Evaluation

PREDICTED CRIME RATE=

-293+91*UNMARRIED+12.2*PUBASST+4.5BLK+39.8TOTAL DIVORCE+12.1VBOARD



Example

we picked a random record, namely the 1983 in the crime dataset,

- Africa American=0.75,
- pubasst=7.11,
- vboard=1.85,
- total divorce=12.8.

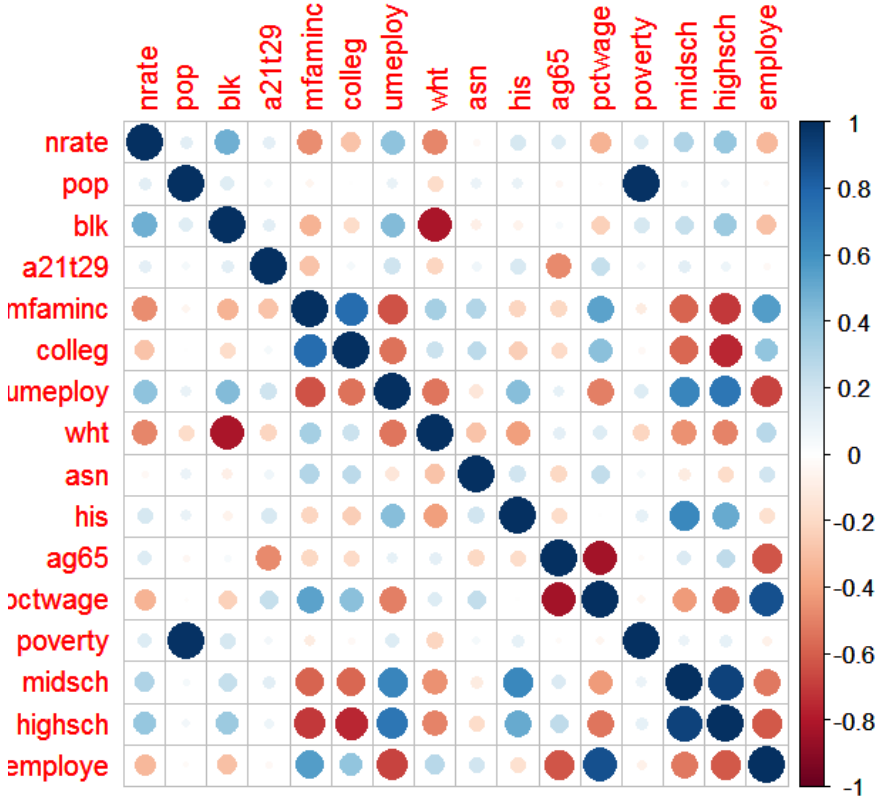
After plugging to the formula, we have predicting crime per Population=260, which is in the 85% confidence interval of the actual crime per population, 288.

Non Violent Crime Rate

	CORR
ViolentCrimesPerPop	0.68
TotalPctDiv	0.61
FemalePctDiv	0.60
MalePctDivorce	0.59
PctKidsBornNeverMar	0.56
PctPopUnderPov	0.52
PctHousNoPhone	0.50
racepctblack	0.48
PctHousLess3BR	0.48
pctWPubAsst	0.47
PctUnemployed	0.41

	CORR
medFamInc	-0.46
PctHousOwnOcc	-0.46
medIncome	-0.47
racePctWhite	-0.48
pctWInvInc	-0.49
PctPersOwnOccup	-0.50
PctYoungKids2Par	-0.61
PctTeen2Par	-0.62
PctFam2Par	-0.66
PctKids2Par	-0.66

Non Violent Crime Rate Regression Model



R^2 = 0.297

The model need to be further improved

```
Call:
lm(formula = nrate ~ blk + poverty + highsch + his, data = n1)

Residuals:
    Min      1Q  Median      3Q     Max
-7189.3 -1441.2  -317.9  1030.6 23393.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.999e+03  1.167e+02  25.686  < 2e-16 ***
blk         8.127e+01  4.041e+00  20.110  < 2e-16 ***
poverty     3.058e-03  1.277e-03   2.394   0.0168 *
highsch     4.393e+01  5.907e+00   7.437 1.49e-13 ***
his         1.978e+01  4.124e+00   4.797 1.72e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2299 on 2112 degrees of freedom
Multiple R-squared:  0.297,     Adjusted R-squared:  0.2956
F-statistic:   223 on 4 and 2112 DF,  p-value: < 2.2e-16
```

Conclusion

- We successfully applied the data analysis techniques we have learned in class, such as PCA and k-means clustering, in our research.
- We also performed Pearson Correlation analysis to explore relationships between variables.
- We built a linear model that models the relationship of factors with crime rates with good prediction capability: the predicted crime rates are very close to the actual crime rates.