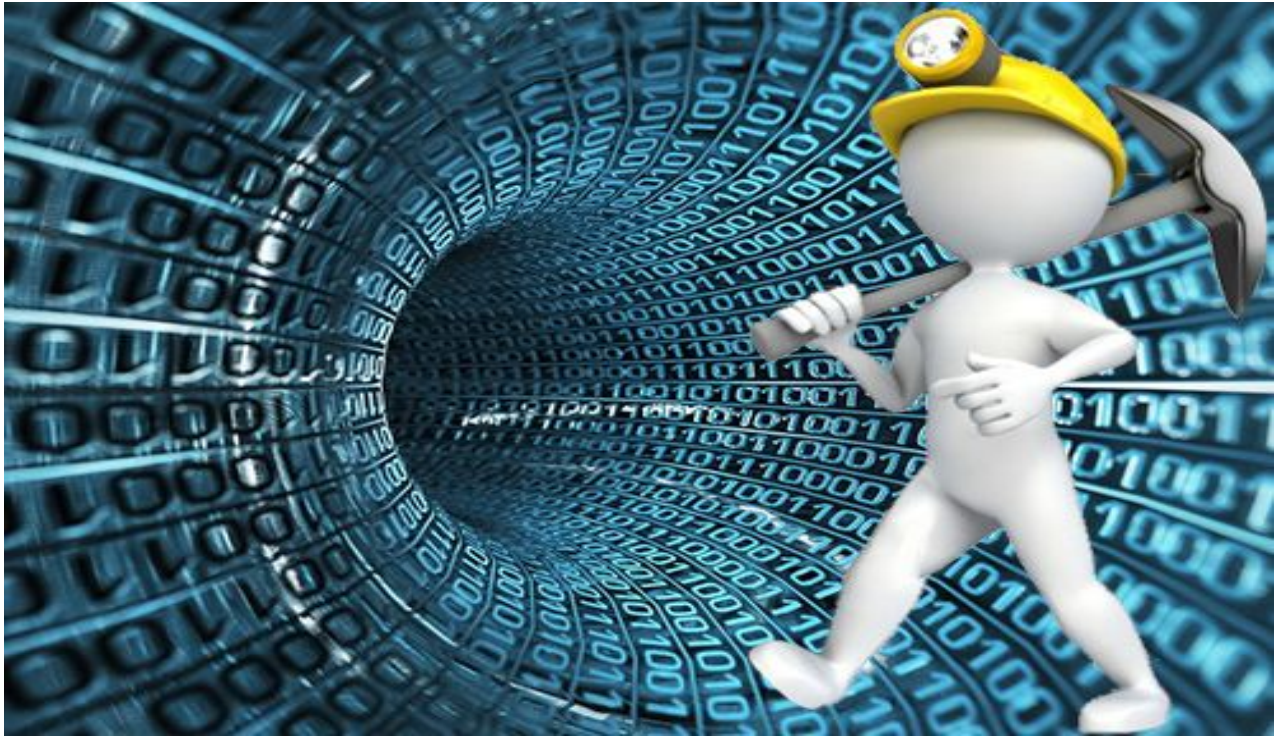
Data mining for pattern discovery: How income and education levels affect crime rates



- Aashay Mokadam
- Mitwa Palkhiwala
- Liubov Tovbin
- Rohit Sapkal

Project Definition and significance to the real world

- Crime is an age-old societal problem which impedes progress and adversely affects human societies. There is much value in predicting the indicators of future criminal behavior based on past records.

- In this project, we look towards analysing mean income levels of households and average education levels, county-by-county, and their respective contributions to crime rates in the United States.

- Additionally, we analyse the relationships between violent crime and property crime, high-school dropout levels and rate of higher-education, and income distribution with respect to mean income levels.

Data Collection

Veracity: Source of Data

We focused our attention on the publicly available data sets from the US Census Bureau and criminal records from the FBI from the years 2005-2009.

United States 2005 - 2009

- US Census Bureau (income, education)
- FBI criminal records

Data Collection

Variety: Types of Data used

- We are focusing on a variety of features and sub-features on a county level:

  Education Levels: No Degree, High School Dropouts and Graduate

  Income Levels: Household Mean Income, Household Median Income, Standard Deviation in Income, Per Capita Income

  Crimes: Property Crime, Violent Crime, Total Crime

    Property Crime: Burglary, Larcency Theft, Motor-Vehicle Theft, Arson Theft

    Violent Crime: Murder and nonnegligent manslaughter, Forcible rape, Robbery, Aggravated Assault

Data Collection

Volume: Amount of Data

Total Entries = Number of Counties = 2665

Data Collection

Velocity: Streaming Data in Real Life

The project is focused on crime rate analysis based on past records. Hence, this section doesn’t apply to us.

Data Extraction

Format of Raw Data

ID	Attribute Name
EDU600209D	Persons 25 years and over, total 2005-2009
EDU610209D	Educational attainment - persons 25 years and over completing less than 9th grade 2005-2009
EDU620209D	Educational attainment - persons 25 years and over completing 9th to 12th grade, no diploma 2005-2009
EDU635209D	Educational attainment - persons 25 years and over - percent high school graduate or higher 2005-2009
EDU640209D	Educational attainment - persons 25 years and over - high school graduate (includes equivalency) 2005-2009
EDU660209D	Educational attainment - persons 25 years and over - some college, no degree 2005-2009
EDU670209D	Educational attainment - persons 25 years and over - associate degree 2005-2009
EDU685209D	Educational attainment - persons 25 years and over - percent bachelor's degree or higher 2005-2009
EDU690209D	Educational attainment - persons 25 years and over - bachelor's degree 2005-2009
EDU695209D	Educational attainment - persons 25 years and over - graduate or professional degree 2005-2009

Data Extraction

Format of Raw Data

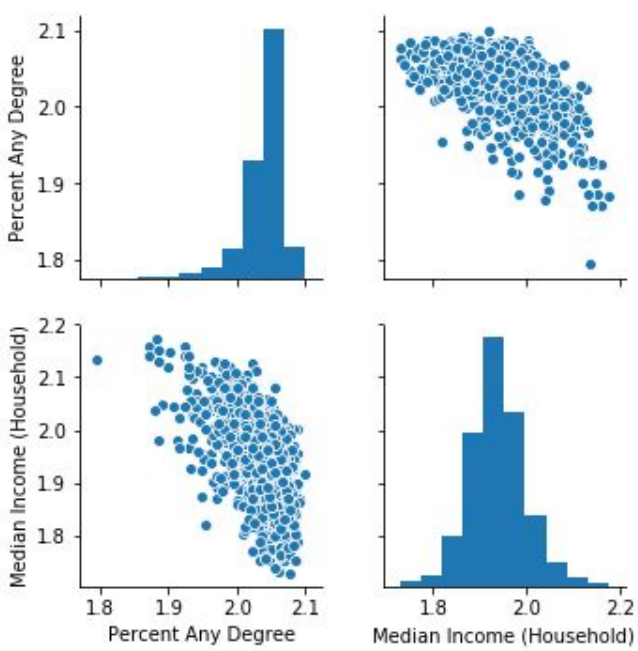
ID	Attribute Name
INC110209D	Median household income in the past 12 months (in 2009 inflation-adjusted dollars) in 2005-2009
INC120209D	Mean household income in the past 12 months (in 2009 inflation-adjusted dollars) in 2005-2009
INC140209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars), total 2005-2009
INC150209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) less than $10,000 in 2005-2009
INC170209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $10,000 to $14,999 in 2005-2009
INC180209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $15,000 to $19,999 in 2005-2009
INC190209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $20,000 to $24,999 in 2005-2009
INC200209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $25,000 to $29,999 in 2005-2009
INC210209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $30,000 to $34,999 in 2005-2009
INC220209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $35,000 to $39,999 in 2005-2009
INC240209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $40,000 to $44,999 in 2005-2009
INC250209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $45,000 to $49,999 in 2005-2009
INC270209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $50,000 to $59,999 in 2005-2009
INC280209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $60,000 to $74,999 in 2005-2009
INC300209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $75,000 to $99,999 in 2005-2009
INC310209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $100,000 to $124,999 in 2005-2009
INC320209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $125,000 to $149,999 in 2005-2009
INC340209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $150,000 to $199,999 in 2005-2009
INC350209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $200,000 or more in 2005-2009
INC910209D	Per capita income in the past 12 months (in 2009 inflation-adjusted dollars) in 2005-2009

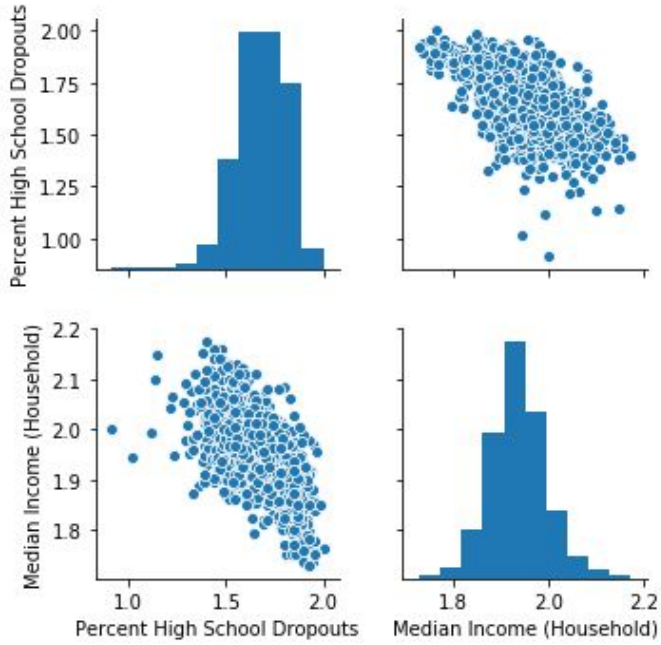
Preprocessing

- Income scaled to units of $1000
- Education levels treated as percentages
- 2nd and 4th Roots of the Logarithm of Crime Levels considered during evaluation
- Sklearn’s Standard Scaler applied prior to fitting the model

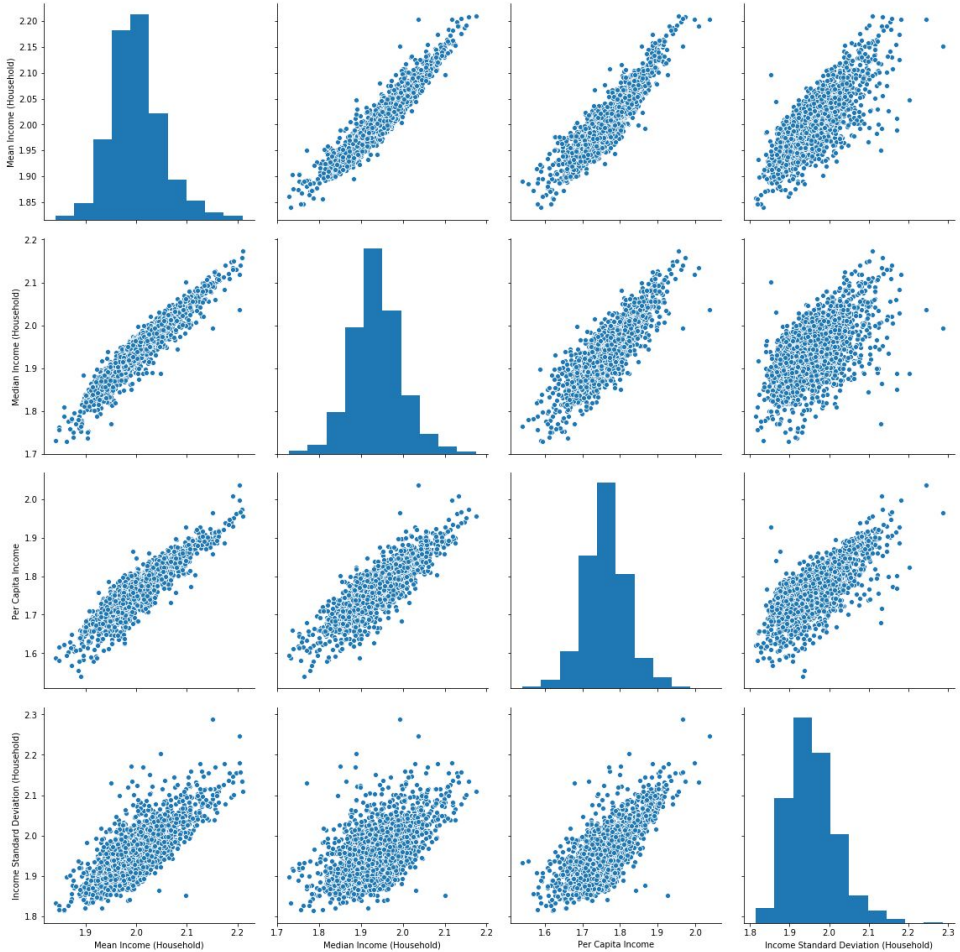
Data Visualization

We compared the plots of various parameters to get more clarity about the trends in each category

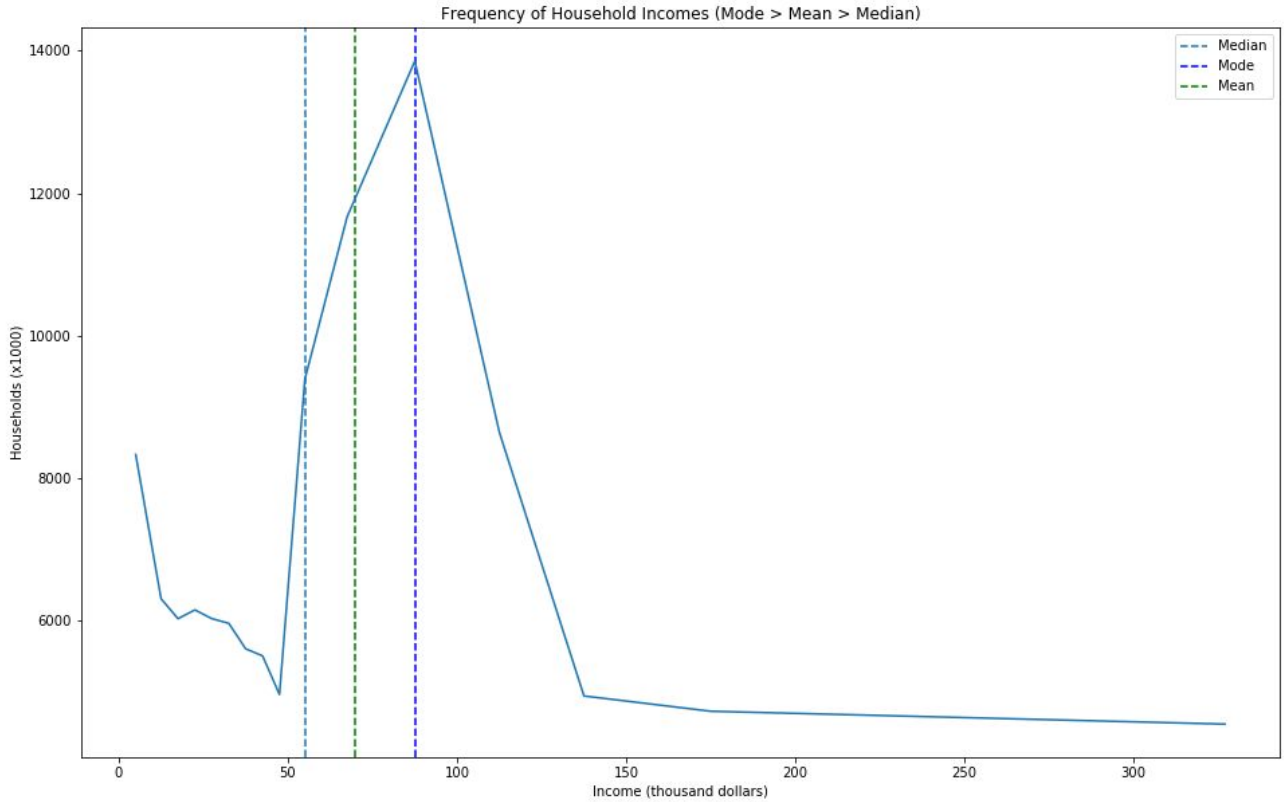




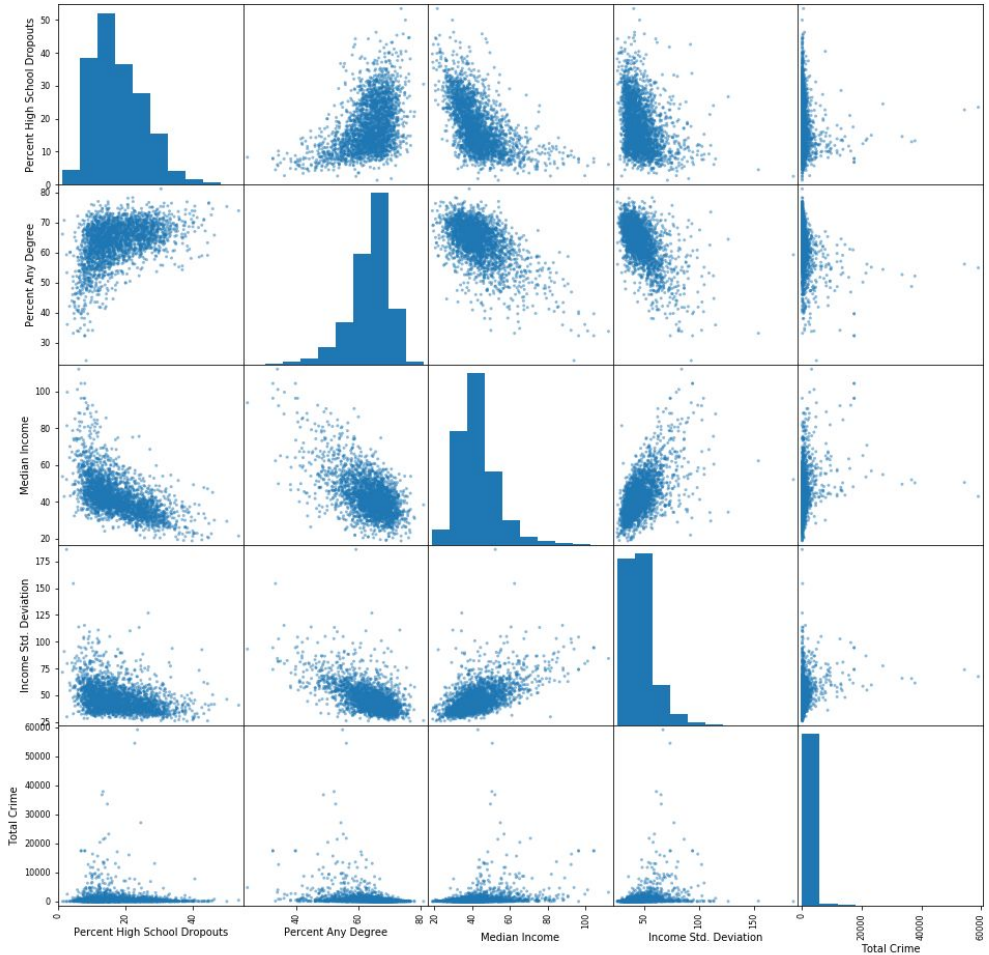
Comparison of Pairplot of percentage of people with and without any degree degree with their respective median income



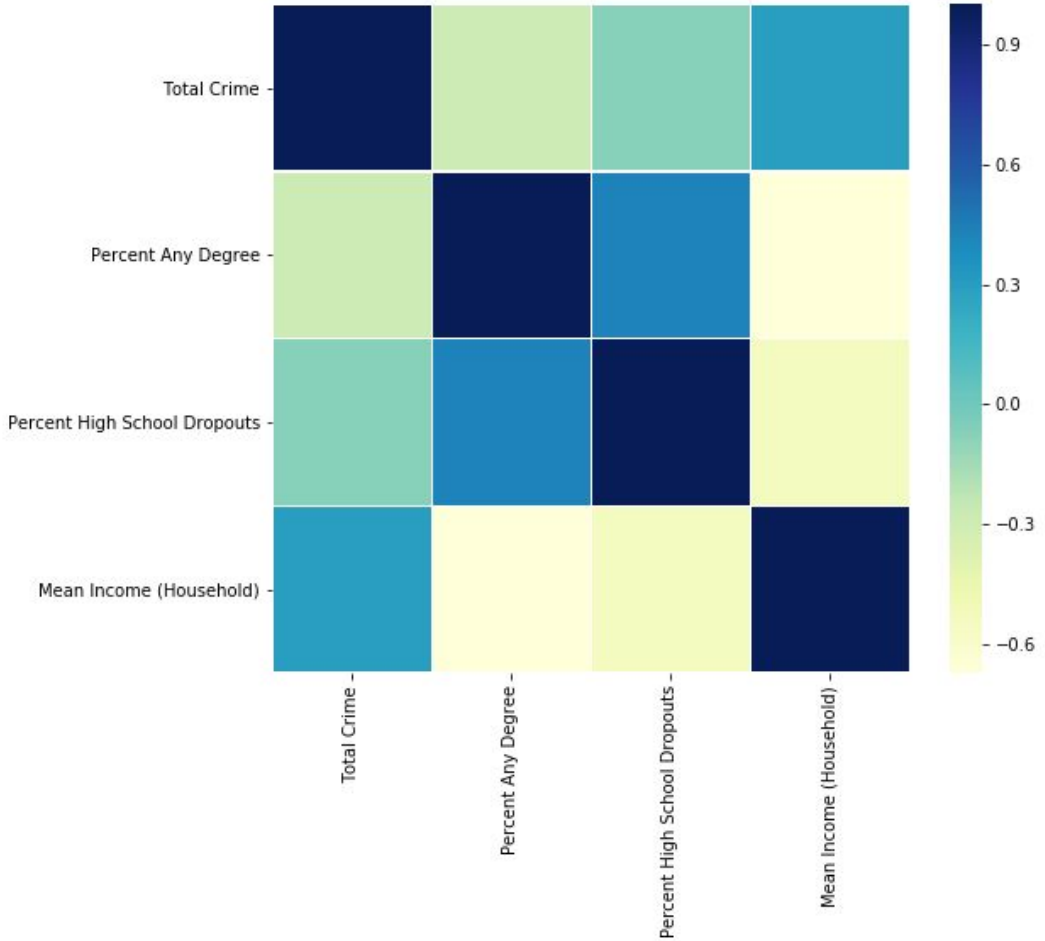
Income Pairplots: Correlation between mean, median, per capita and standard deviation Income



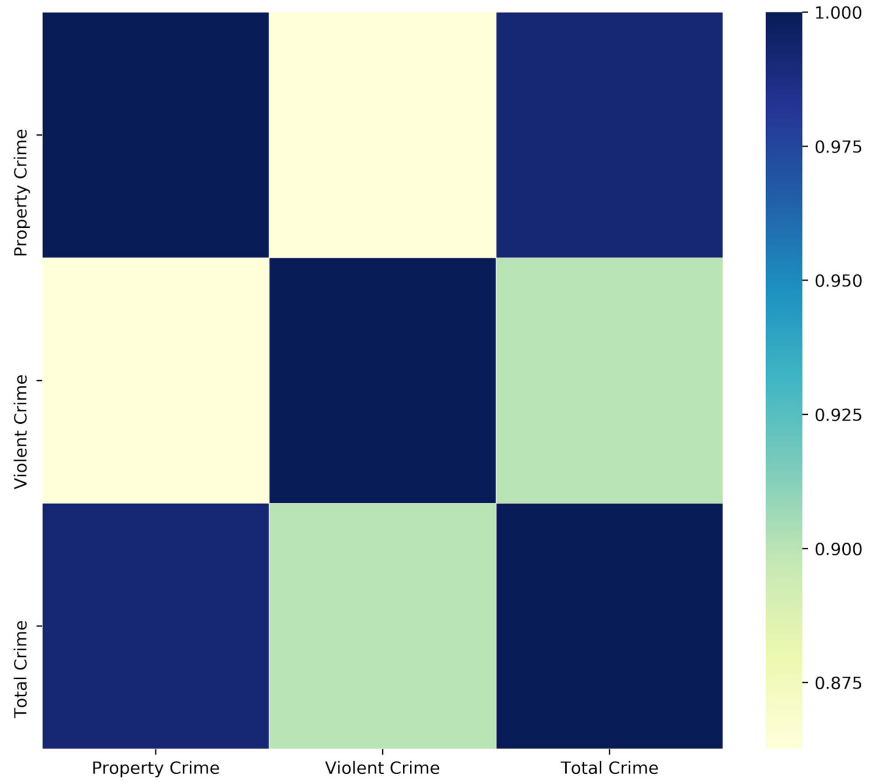
Mean, Median, and Mode Incomes on the Income Distribution (Nationwide)



Correlation between all the features - Crime, Education and Income



Heatmap of correlation between Crime, Education and Mean income



Heatmap of correlation between types of crime

Crime Prediction Models

- Linear Regression
- Gradient Boosting Regressor
- Decision Tree
- K-neighbour Regressor
- Random Forest Regressor
- Ridge Regression
- **Support Vector Regression**

Technical Difficulties

Finding the Standard Deviation from Uncertain Ranges

INC350209D	Households with income in the past 12 months (in 2009 inflation-adjusted dollars) of $200,000 or more in 2005-2009

$$I_{high\_bucket} = I_{mean} * F_{mean} - \sum_{s \epsilon S} I_s * F_s$$

Lessons learnt

- Data often doesn’t follow intuition.
- Correlation analysis through plots is an accurate predictor of effectiveness of features.
- SVR eager to overfit in this case.
- Simple Linear Models perform well due to low dimensionality of data.
- Logarithmic smoothing of target variable led to monumental improvement of results.

Results

	BoostingMSE	DecisionTreeMSE	KNeighbourMSE	LinearRegMSE	RandomForestMSE	RidgeMSE	SVR_MSE	crime_regularization	education_type	income_type
0	0.135707	0.142844	0.187286	0.135341	0.135707	0.135341	0.133981	Square Root of Log	High School Dropout Percent	Median Income
1	0.143092	0.150008	0.196002	0.145034	0.143092	0.145033	0.140423	Square Root of Log	High School Dropout Percent	Income Standard Deviation
2	0.139998	0.141787	0.188993	0.134803	0.139998	0.134802	0.142382	Square Root of Log	Percent with Any Degree	Median Income
3	0.138230	0.140847	0.189016	0.137201	0.138230	0.137202	0.136058	Square Root of Log	Percent with Any Degree	Income Standard Deviation
4	0.020830	0.022114	0.029812	0.020799	0.020830	0.020799	0.020618	Fourth Root of Log	High School Dropout Percent	Median Income
5	0.021630	0.021925	0.030625	0.022015	0.021630	0.022014	0.021241	Fourth Root of Log	High School Dropout Percent	Income Standard Deviation
6	0.021219	0.021721	0.029206	0.020920	0.021219	0.020920	0.021868	Fourth Root of Log	Percent with Any Degree	Median Income
7	0.021558	0.021700	0.030082	0.021203	0.021558	0.021203	0.021044	Fourth Root of Log	Percent with Any Degree	Income Standard Deviation

Support Vector Regressor gave the least Mean Squared Error in all the education level and income categories.

Conclusion

- Positive correlation between high school dropout rates and crime.
- Positive correlation between income levels and crime.
- Negative correlation between rates of higher studies and crime.
- Rate of higher studies are also negatively correlated with income levels.

Conclusion

- Amount of correlation is weak.
- Data has patterns amongst itself. Distributions should be encapsulated in features.
- Potentially could get better results by treating as a classification problem.
- PCA may not be helpful in capturing irregular correlation between income and education

Github Repository

https://github.com/LubaTovbin/CMPE-255-02-Data-Mining/

Data Reduction

Statewise Data

Q & As


Thank You !