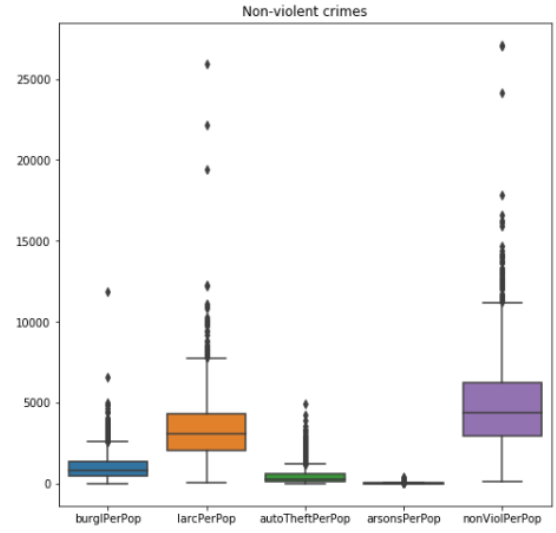
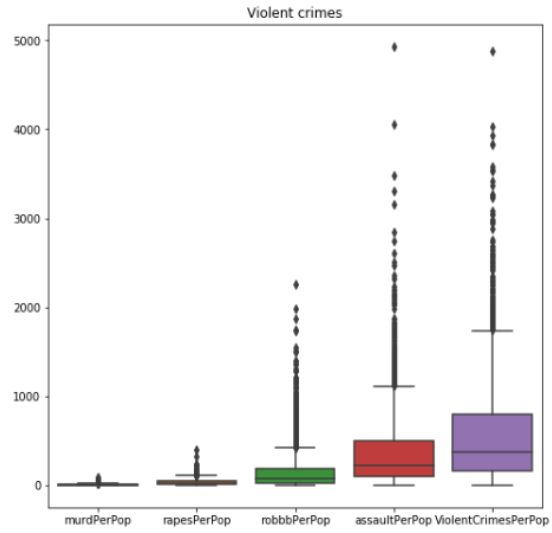
1. ***Objective***

A data of crimes that happened in the early and mid-90's in US has been provided for both violent and non-violent categories. The objective is to identify the drivers that can cause an increase in any of these crime categories and if there are any social, economic, political backgrounds to any such incidents.

2. ***Exploratory Data Analysis***

a. Feelers for the violent and non-violent crime categories





b. Geographic Observation

Top 10 states in both crime categories

	statecode	ViolentCrimesPerPop
7	DC	3048.380000
17	LA	1312.713636
37	SC	1233.455714
19	MD	1217.699167
9	FL	1159.046889
1	AL	1030.699070
10	GA	973.413514
25	NC	973.036087
8	DE	887.290000
15	KS	874.690000

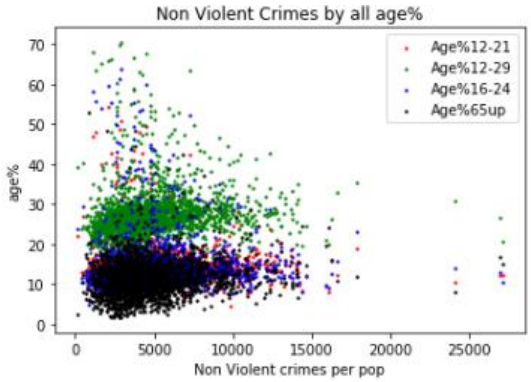
	statecode	ViolentCrimesPerPop	nonViolPerPop
7	DC	3048.380000	9252.350000
9	FL	1159.046889	8011.110222
10	GA	973.413514	7804.054595
25	NC	973.036087	7628.824565
44	WA	547.444000	7464.229744
17	LA	1312.713636	7458.676316
8	DE	887.290000	7276.460000
19	MD	1217.699167	6813.866667
24	MS	663.550526	6734.207000
11	IA	406.452500	6674.236667

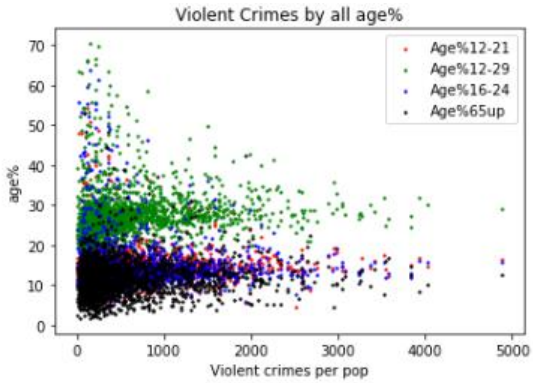
If we observe the above 2 outputs, we will realize that some states like District of Colombia, Florida, Los Angeles, North Carolina, Georgia, Delaware, Maryland rank in top 10 for both violent and non-violent categories of crimes. Note that all of them being coastal areas.

Studies show that during early 80's and 90's, crimes increased a lot in US due to cocaine epidemic and homicides and the easy access to them is through east coast. Tourism is also a driver for such crimes.

c. Demographic Observation

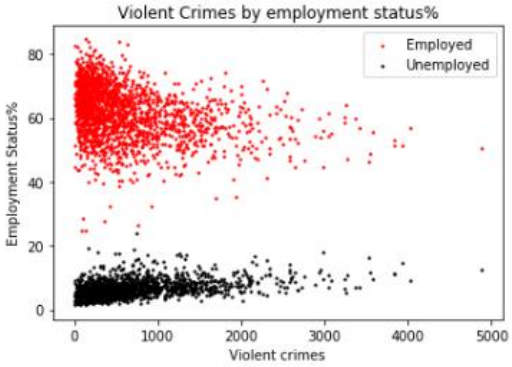
Age brackets and their behavior

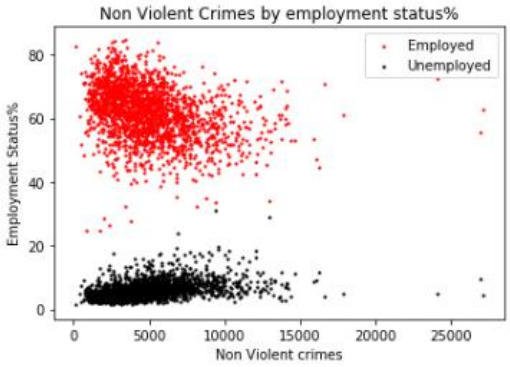




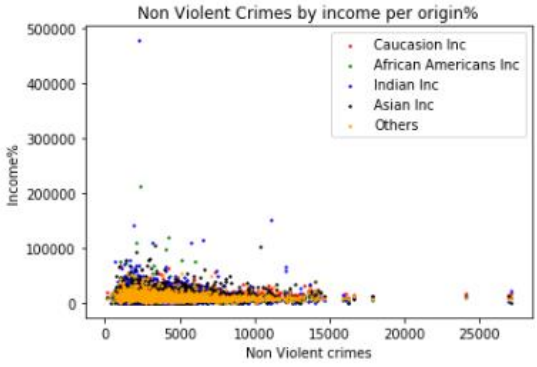
A clear indicator that the unemployed age bracket (who could have been college drop outs or rehab patients) and the retired bracket (who no longer are fit to work but also need to run their daily activities) indulge in crimes.

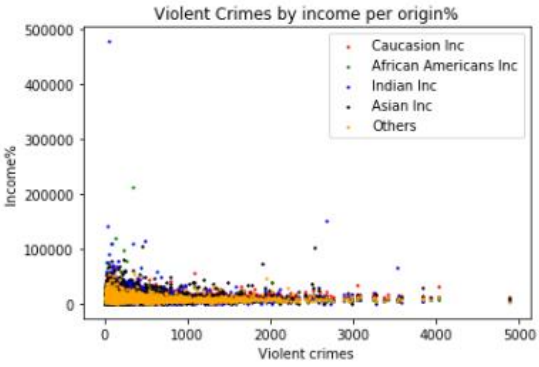
Crime Pattern per Employment Status





Crime Pattern with origin

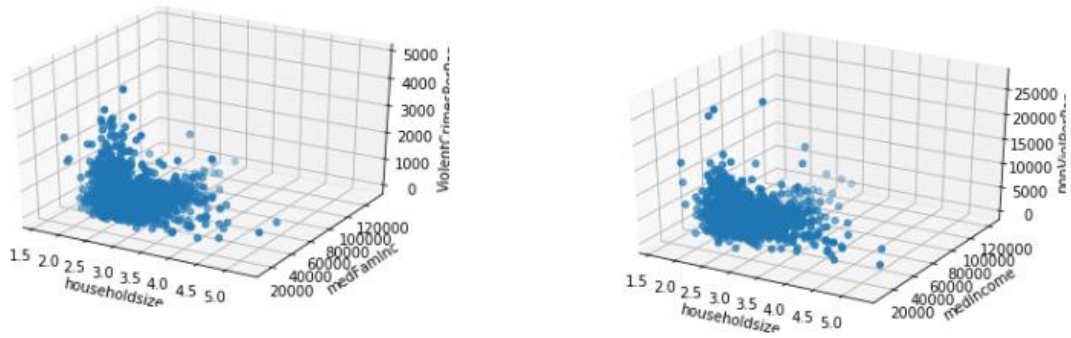




An elaborate data for “Others” race might be needed to explore this further. But apart from others, the crime pattern also shows a bit more of Indian, Asian and African American race.

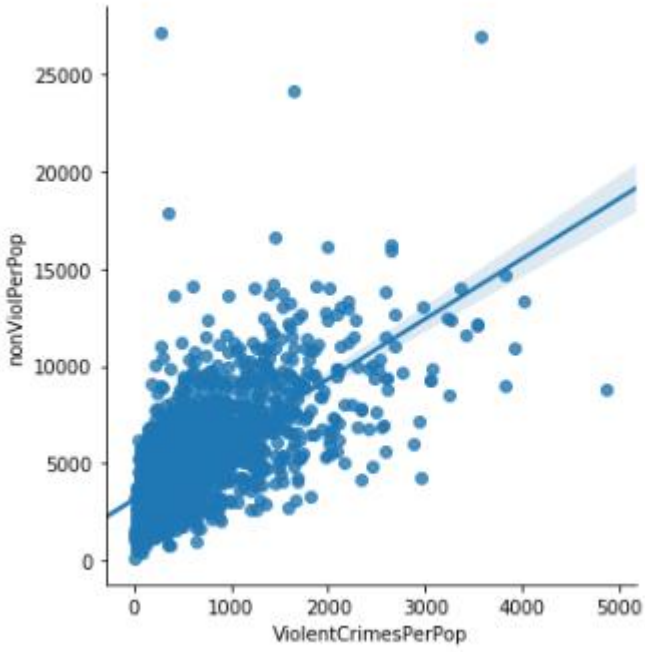
d. Economic Observation

What a household size and income level can do?



As the number of people increase in a family, but the income stays meagre, crimes tend to happen. We see a large concentrate of crime occurrences at lower income groups with more than 3 people in a family.

e. A general mash up of Violent v/s non-violent crimes



3. ***Data Preprocessing activities conducted***

a. The provided text file did not contain any headers, but it only had all the relevant data. Hence an approach was taken to create a data frame with the header information fed manually.

b. When we see the initial data, we will realize that there are a lot of "?" in most of the columns which would mislead us in case of checking NULLs. Hence such occurrences of "?" were replaced with NULLs.

c. A regular check for NULLs was later conducted in which around 24 features have more than 50% NULLs in them. Since the number is high, considering that we have just 2215 observations recorded, such columns have been dropped.

4. ***Predictive Modelling Methodology adopted***

a. We have 2 dependent variables in our scenario – nonViolPerPop for Non-violent crimes and ViolentCrimesPerPop for violent crimes.
b. The approach taken would be to consider each dependent variable at a time, meaning create two different models for two dependent variables.
c. We observe a lot of multicollinearity in the data set and PCA is done to handle such multi collinearity issues and to help in dimensionality reduction. The outcome of PCA says that the top 31 features help us arrive at a cumulative variance of 90%.
d. With the dimensions suggested by PCA, we will proceed with our modelling activities after the usual test train split and scaling applied.
e. The first model was chosen to be Linear Regression. Metrics used to evaluate them include R2, MAE, MSE and RMSE.
f. After this Random Forest Regression is done and MAE and RMSE is used for evaluation.
g. This is followed by XGBoost Regression with evaluation metric of MAE and RMSE.

5. ***Metrics chosen per model for Violent Crimes***

Violent crimes being more grievous and a huge loss to the society where a lot of potential young talent is caught up with mischiefs, the onus is on us to focus on as many predictors as we can to avoid crimes from occurring.

Below is a break up of the metrics evaluated for 3 models shortlisted.

Violent Crimes		
Linear	MAE	194
	RMSE	270
RF	MAE	217
	RMSE	295
XGBoost	MAE	210
	RMSE	397

Finding the predictors but also keeping the modelling strategy simple, even the evaluation metric chosen is very simple. MAE and not RMSE is chosen because RMSE assigns weights on larger errors in the data and squares it up. We have not gotten an opportunity to address such errors efficiently yet and hence MAE is a safer bet. Going by this, Linear Regression seems to be simpler.

6. ***Plans to increase the model efficiency***

a. Although we have done hyper parameter tuning in Random Forest, it still doesn't satisfy us with the outcome. Hyper parameter tuning is one of the ways to increase the efficiency.
b. Can try other modelling methods to identify the predictors and see if there is any other modelling technique that would help us closer to the expectation.

7. ***Relevant variables from Linear Regression***

Violent	Non-violent
% of households with only rental or investment income	People under poverty level
% of people who are employed in professional occupations	Unemployed
Kids born to single parents	Over 25 but not high school graduates
Households that are owner occupied	Kids born to single parents
	Male who are divorced

Non-violent crimes are categorized as: burglary, thefts, arsons. This is clearly for the sake of more money. Our predictors indicate to such characteristics where the crimes are conducted by harmless population of people only to satisfy their day-to-day needs.

Violent Crimes are categorized as: murder, rape, robbery and assault. EDA suggested coastal areas and our model suggests criminals who have more a disturbing upbringing (like kids to single parents) or someone who has enough income but is disturbed mentally (like people who are professionals which means they have a good education system). A deeper analysis can indicate as to what category of crime has been conducted by different mind sets and how it can be tackled.

8. Conclusion

a. Measures can be taken for more vigilance in areas that may have easy access to illegal imports, tourists, suburbs with more population of unemployed residents (mostly retirement villages).
b. Awareness campaigns must be conducted on a regular basis in schools and other public institutions to mold children in the right direction. It is important to spread the message of being empathetic.
c. Appropriate measures must also be taken to address unemployment and quality education which will in turn be a solution to the deep poverty situation in the country.
d. Crime rates and mental health of young population also influence the economy, and it is important to stop crimes from occurring, but also more important to create an awareness and a holistic inclusive society where everyone has opportunities for better quality of life and well-being.
e. Inclusive policies rehabilitation convicts into mainstream society can be created to reduce repeat offenders' cases.