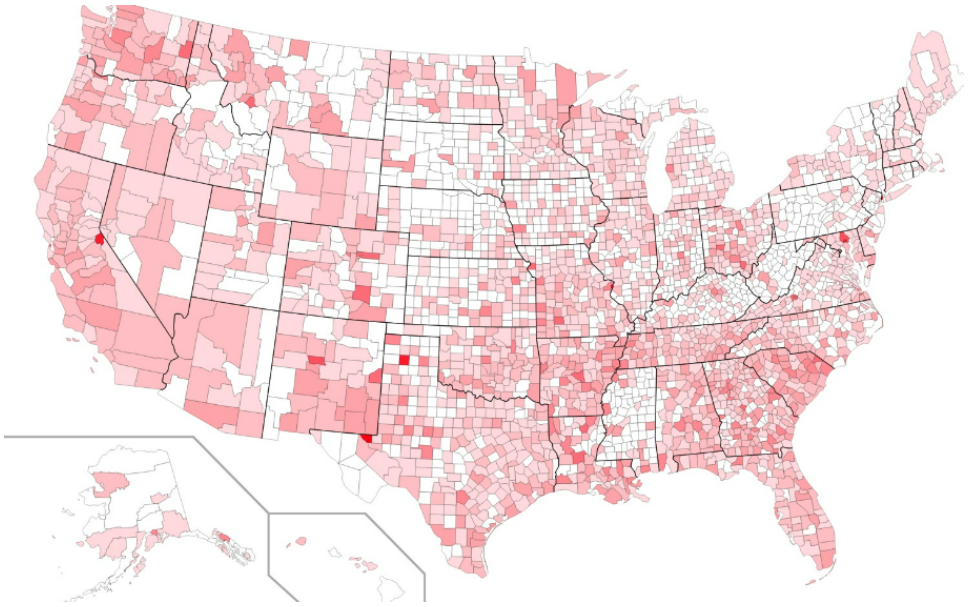
DSCI 5330

ENTERPRISE APPLICATIONS OF BI

Final Project Report

CRIME RATE IN U.S. STATES/COMMUNITIES



Group 5330

AkashVasanthan(11441194)

Fabiha Amir Lodhi (11511873)

Shyla Sree Konda (11506574)

Archana Chilumula (11609268)

Venkata Sai Challa (11629015)

Research Problem:

Any activity that violates either the law or community protocols is a crime. If violated, the person must repent and follow the legal repercussions. Most of the crimes take place due to connections with drugs, people with bad network, personal grudges, for purpose of money, socio-economic conditions, and unemployment etc. Hacking websites, networks, and manipulating software/hardware is also considered a crime.

The definition of crime can differ from state to state as every state has its own set of principles and guidelines. Based on the act of violence, the criminal justice system in Unites States has divided into two major crimes i.e., violent, and non-violent. So, our focus is to predict violent crime rates in communities. The basic difference between violent and non-violent crime is the motive of the criminal, whether he/she wants to hurt the victim or not. Violent crimes include assault, robbery, manslaughter, homicide etc. While the non-violent is opposite of it, it includes property offenses, gambling, bribery, fraud etc.

To know the crime rates and imbalances that occurred due to violence in communities, an associate professor, Michael Redmond at "La Salle University" has conducted a US LEMAS survey and collected real time data.

The major imbalance that is in correlation with the crime rate is inflation. As inflation rises, the crime rate also tends to increase. The inflation effect is high on low-income employees who later try to leverage their economic situations in wrongdoing. Since unemployment seems to be a prominent factor in the crime rate, our future correlation analysis will be performed between them which comes under Exploratory Data Analysis (EDA). Also, with the help of geographical charts, we will forecast violent and non-violent crime rates as per state. To increase the granularity of crimes in location, we will geographically represent as per communities as well.

Research Objective:

With a survey conducted in US communities it was identified that, on average there has been a 12% increase in crime rate specifically violent crimes since 2010. Whereas crimes which including murder, rape and assault have been increased by 25% in us communities.

The research objective revolves around understanding the various categories of crimes experienced in U.S (United States). communities, interpreting the reasons for the high crime rates to find safest places for residents to stay.

- With this research survey, we found that most violent crimes are being committed by people around age 20-30years old. Violence also remained concentrated among youthful individuals.
- Aggravated attack is the most well-known brutal crime which incorporates criminal way of behaving that includes an assault on somebody with the purpose to cause injury. It could incorporate the utilization of a weapon.

Communication and immediate response are the basic requirements that were most highlighted for controlling crime scenes. In response to that one of the strategies was public surveillance cameras, that are in widespread usage and have already helped to reduce crimes and for better understanding the crime structures and the need of police for immediate action to have a helping hand for the victims. This structured system in USA communities consists of less crime rate make the national security a level head. Eventually this will increase tourism and create opportunities for financial and modern technology exposures with wide connects as the people would feel this as the safest place.

Research Questions:

Our study will allow us to answer the following questions

Q1) What are the top 2 violent crimes experienced in US communities?

Q2) What are the factors that influence people in the US to commit violent crimes?

Q3) What are the top 3 states with the highest crime rates?

Q4) What are the top 10 safest states in the US for people to reside in.

Q5) What preventive measures can be taken to reduce crimes?

This research will help us in understanding the bigger picture of the types of crimes experienced in the different states of America, which states have the highest rate of crimes and most importantly what are the reasons for increased crimes in the different states. Moreover, with *feature selection techniques we can provide people with an insight on what will be the top 10 safest places in the U.S. to reside.*

Research Design:

There are 147 independent variables and one target variable (ViolentCrimePerPop). The dependent variable relies on predictors, depending on how they correlate. We used different correlation methods below to understand the predictors.

DATA USED:

The dataset for this research has been taken from the UCI Machine Learning Repository. The datasets available for our research study is about crime rate at different community among different states at USA, we have a training dataset that contains 147 features/columns and 2215 observation/ rows of which **ViolentCrimesPerPop (Total number of violent crimes per 100K population)** the target variable leaving 146 features behind of which 116 are floating point variables, 29 are integer variables and rest are categorical variable.

Data Description

-- state	US state (by number) - not counted as predictive above, but if considered, should be considerd
-- county	numeric code for county - not predictive, and many missing values (numeric)
-- community	numeric code for community - not predictive and many missing values (numeric)
-- communityname	community name - not predictive - for information only (string)
-- fold	fold number for non-random 10-fold cross validation, potentially useful for debugging, paired tests - not predictive (numeric)
-- population	population for community
-- householdsize	mean people per household (numeric - decimal)
-- racepctblack	percentage of population that is African American (numeric - decimal)
-- racePctWhite	percentage of population that is Caucasian (numeric - decimal)
-- racePctAsian	percentage of population that is of Asian heritage (numeric - decimal)
-- racePctHisp	percentage of population that is of Hispanic heritage (numeric - decimal)
-- agePct12t21	percentage of population that is 12-21 in age (numeric - decimal)
-- agePct12t29	percentage of population that is 12-29 in age (numeric - decimal)
-- agePct16t24	percentage of population that is 16-24 in age (numeric - decimal)
-- agePct65up	percentage of population that is 65 and over in age (numeric - decimal)
-- numbUrban	number of people living in areas classified as urban (numeric - decimal)
-- pctUrban	percentage of people living in areas classified as urban (numeric - decimal)
-- medIncome	median household income (numeric - decimal)
-- pctWWage	percentage of households with wage or salary income in 1989 (numeric - decimal)
-- pctWFarmSelf	percentage of households with farm or self-employment income in 1989 (numeric -
decimal)	

-- pctWInvInc	percentage of households with investment / rent income in 1989 (numeric - decimal)
-- pctWSocSec	percentage of households with social security income in 1989 (numeric - decimal)
-- pctWPubAsst	percentage of households with public assistance income in 1989 (numeric - decimal)
-- pctWRetire	percentage of households with retirement income in 1989 (numeric - decimal)
-- medFamInc	median family income (differs from household income for non-family households)
-- perCapInc	per capita income (numeric - decimal)
-- whitePerCap	per capita income for Caucasians (numeric - decimal)
-- blackPerCap	per capita income for African Americans (numeric - decimal)
-- indianPerCap	per capita income for native Americans (numeric - decimal)
-- AsianPerCap	per capita income for people with Asian heritage (numeric - decimal)
-- OtherPerCap	per capita income for people with 'other' heritage (numeric - decimal)
-- HispPerCap	per capita income for people with Hispanic heritage (numeric - decimal)
-- NumUnderPov	number of people under the poverty level (numeric - decimal)
-- PctPopUnderPov	percentage of people under the poverty level (numeric - decimal)
-- PctLess9thGrade	percentage of people 25 and over with less than a 9th grade education (numeric -
-- PctNotHSGrad	percentage of people 25 and over that are not high school graduates (numeric -
-- PctBSorMore	percentage of people 25 and over with a bachelor's degree or higher education (numeric)
-- PctUnemployed	percentage of people 16 and over, in the labor force, and unemployed (numeric – decimal)
-- PctEmploy	percentage of people 16 and over who are employed (numeric - decimal)
-- PctEmplManu	percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
-- PctEmplProfServ	percentage of people 16 and over who are employed in professional services

-- PctOccupManu	percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
-- PctOccupMgmtProf	percentage of people 16 and over who are employed in management or professional occupations (numeric - decimal)
-- MalePctDivorce	percentage of males who are divorced (numeric - decimal)
-- MalePctNevMarr	percentage of males who have never married (numeric - decimal)
-- FemalePctDiv	percentage of females who are divorced (numeric - decimal)
-- TotalPctDiv	percentage of population who are divorced (numeric - decimal)
-- PersPerFam	mean number of people per family (numeric - decimal)
-- PctFam2Par	percentage of families (with kids) that are headed by two parents (numeric - decimal)
-- PctKids2Par	percentage of kids in family housing with two parents (numeric - decimal)
-- PctYoungKids2Par	percent of kids 4 and under in two parent households (numeric - decimal)
-- PctTeen2Par	percent of kids aged 12-17 in two parent households (numeric - decimal)
-- PctWorkMomYoungKids	percentage of moms of kids 6 and under in labor force (numeric - decimal)
-- PctWorkMom	percentage of moms of kids under 18 in labor force (numeric - decimal)
-- NumIlleg	number of kids born to never married (numeric - decimal)
-- PctIlleg	percentage of kids born to never married (numeric - decimal)
-- NumImmig	total number of people known to be foreign born (numeric - decimal)
-- PctImmigRecent	percentage of _immigrants_ who immigrated within last 3 years (numeric - decimal)
-- PctImmigRec5	percentage of _immigrants_ who immigrated within last 5 years (numeric - decimal)
-- PctImmigRec8	percentage of _immigrants_ who immigrated within last 8 years (numeric - decimal)
-- PctImmigRec10	percentage of _immigrants_ who immigrated within last 10 years (numeric - decimal)
-- PctRecentImmig	percent of _population_ who have immigrated within the last 3 years (numeric - decimal)

-- PctRecImmig5	percent of _population_ who have immigrated within the last 5 years (numeric - decimal)
-- PctRecImmig8	percent of _population_ who have immigrated within the last 8 years (numeric - decimal)
-- PctRecImmig10	percent of _population_ who have immigrated within the last 10 years (numeric - decimal)
-- PctSpeakEnglOnly	percent of people who speak only English (numeric - decimal)
-- PctNotSpeakEnglWell	percent of people who do not speak English well (numeric - decimal)
-- PctLargHouseFam	percent of family households that are large (6 or more) (numeric - decimal)
-- PctLargHouseOccup	percent of all occupied households that are large (6 or more people) (numeric - decimal)
-- PersPerOccupHous	mean persons per household (numeric - decimal)
-- PersPerOwnOccHous	mean persons per owner occupied household (numeric - decimal)
-- PersPerRentOccHous	mean persons per rental household (numeric - decimal)
-- PctPersOwnOccup	percent of people in owner occupied households (numeric - decimal)
-- PctPersDenseHous	percent of persons in dense housing (more than 1 person per room) (numeric - decimal)
-- PctHousLess3BR	percent of housing units with less than 3 bedrooms (numeric - decimal)
-- MedNumBR	median number of bedrooms (numeric - decimal)
-- HousVacant	number of vacant households (numeric - decimal)
-- PctHousOccup	percent of housing occupied (numeric - decimal)
-- PctHousOwnOcc	percent of households owner occupied (numeric - decimal)
-- PctVacantBoarded	percent of vacant housing that is boarded up (numeric - decimal)
-- PctVacMore6Mos	decimal)
-- MedYrHousBuilt	median year housing units built (numeric - decimal)

-- PctHousNoPhone	percent of occupied housing units without phone (in 1990, this was rare!) (Numeric - decimal)
-- PctWOFullPlumb	percent of housing without complete plumbing facilities (numeric - decimal)
-- OwnOccLowQuart	owner occupied housing - lower quartile value (numeric - decimal)
-- OwnOccMedVal	owner occupied housing - median value (numeric - decimal)
-- OwnOccHiQuart	owner occupied housing - upper quartile value (numeric - decimal)
-- RentLowQ	rental housing - lower quartile rent (numeric - decimal)
-- RentMedian	rental housing - median rent (Census variable H32B from file STF1A) (numeric - decimal)
-- RentHighQ	rental housing - upper quartile rent (numeric - decimal)
-- MedRent	median gross rent (Census variable H43A from file STF3A - includes utilities) (numeric - decimal)
-- MedRentPctHousInc	median gross rent as a percentage of household income (numeric - decimal)
-- MedOwnCostPctInc	median owners cost as a percentage of household income - for owners with a mortgage (numeric - decimal)
-- MedOwnCostPctIncNoMtg	without a mortgage (numeric - decimal)
-- NumInShelters	number of people in homeless shelters (numeric - decimal)
-- NumStreet	number of homeless people counted in the street (numeric - decimal)
-- PctForeignBorn	percent of foreign people born (numeric - decimal)
-- PctBornSameState	percent of people born in the same state as currently living (numeric - decimal)
-- PctSameHouse85	percent of people living in the same house as in 1985 (5 years before) (numeric - decimal)
-- PctSameCity85	percent of people living in the same city as in 1985 (5 years before) (numeric - decimal)
-- PctSameState85	percent of people living in the same state as in 1985 (5 years before) (numeric - decimal)
-- LemasSwornFT	number of sworn full time police officers (numeric - decimal)
-- LemasSwFTPerPop	sworn full time police officers per 100K population (numeric - decimal)

-- LemasSwFTFieldOps	number of sworn full time police officers in field operations (on the street as opposed to administrative etc) (numeric - decimal)
-- LemasSwFTFieldPerPop	sworn full time police officers in field operations (on the street as opposed to administrative etc) per 100K population (numeric - decimal)
-- LemasTotalReq	total requests for police (numeric - decimal)
-- LemasTotReqPerPop	total requests for police per 100K population (numeric - decimal)
-- PolicReqPerOffic	total requests for police per police officer (numeric - decimal)
-- PolicPerPop	police officers per 100K population (numeric - decimal)
-- RacialMatchCommPol	a measure of the racial match between the community and the police force. High values indicate proportions in community and police force are similar (numeric - decimal)
-- PctPolicWhite	percent of police that are Caucasian (numeric - decimal)
-- PctPolicBlack	percent of police that are African American (numeric - decimal)
-- PctPolicHisp	percent of police that are Hispanic (numeric - decimal)
-- PctPolicAsian	percent of police that are Asian (numeric - decimal)
-- PctPolicMinor	percent of police that are minority of any kind (numeric - decimal)
-- OfficAssgnDrugUnits	number of officers assigned to special drug units (numeric - decimal)
-- NumKindsDrugsSeiz	number of various kinds of drugs seized (numeric - decimal)
-- PolicAveOTWorked	police average overtime worked (numeric - decimal)
-- LandArea	land area in square miles (numeric - decimal)
-- PopDens	population density in persons per square mile (numeric - decimal)
-- PctUsePubTrans	percent of people using public transit for commuting (numeric - decimal)
-- PolicCars	number of police cars (numeric - decimal)
-- PolicOperBudg	police operating budget (numeric - decimal)
-- LemasPctPolicOnPatr	percent of sworn full time police officers on patrol (numeric - decimal)
-- LemasGangUnitDeploy	means YES, 0.5 means Part Time)

-- LemasPctOfficDrugUn	percent of officers assigned to drug units (numeric - decimal)
-- PolicBudgPerPop	police operating budget per population (numeric - decimal)
-- ViolentCrimesPerPop	total number of violent crimes per 100K population (numeric - decimal) GOAL attribute (to be predicted)

Descriptive Analysis (EDA):

From the crime rate dataset, we found that there are missing values for 1/4th of the columns. To get rid of columns with the highest percentage of missing values, we performed "Missing value analysis" and plotted its histogram.

- Missing Value Analysis:
  - Out of 147 columns, there are 41 columns with missing values.
  - Depending upon the total percentage of missing values in the column, it is removed. We removed columns with more than 60% missing values and left with 122 columns.
  - Here are the dropped columns:

'PolicCars', 'LemasGangUnitDeploy', 'PolicOperBudg', 'PolicAveOTWorked',

'NumKindsDrugsSeiz', 'OfficAssgnDrugUnits', 'PctPolicMinor',

'PctPolicAsian', 'PctPolicHisp', 'PctPolicBlack', 'PctPolicWhite',

'RacialMatchCommPol', 'PolicPerPop', 'PolicReqPerOffic',

'LemasTotReqPerPop', 'LemasTotalReq', 'LemasSwFTFieldPerPop',

'LemasSwFTFieldOps', 'LemasSwFTPerPop', 'LemasSwornFT',

'PolicBudgPerPop', 'LemasPctPolicOnPatr','communityCode','countyCode'

  - The column with highest missing values percentage (84.5%) are mostly related to "police" and zero percentage missing values are from "robberies," "robbeperpop."
  - With help of histogram plot, the columns with less than 60% missing values are visualized

For more information about columns, their unique values are shown.

After extracting the dataset with less missing values, the next step is to perform imputation.

Imputation:

We choose KNN (k-nearest neighbor) algorithm, here missing data is marked as 'NaN' and replaced with its nearest neighbor estimated values.

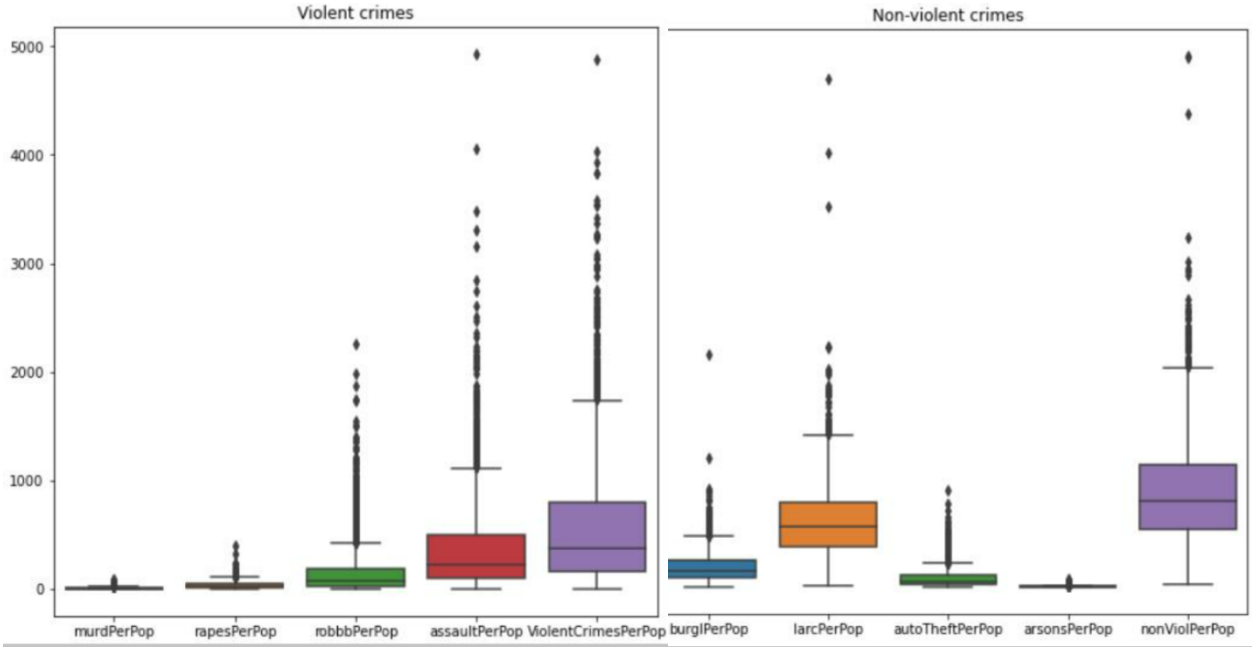
To get better performance of the model, before the imputation, categorical conversion is done on "state" column using "category_encoders" package.

KNNImputer is a class from scikit-learn machine learning library used to predict the missing value based on distance parameter. KNNImputer uses 'k' nearest neighbors to estimate the value to be imputed. In our dataset, we choose k (n_neighbors) value to be 5 which is a default value.

'Imputer' object is created from KNNImputer class, and fit is applied to the encoded dataset to achieve a cleaned dataset.

Once the data is preprocessed and cleaned, the next step is to check whether asymmetry exists within the dataset. Both positive and negative skewness is observed in the data.

After Imputation, we performed explanatory data analysis on key features which account for violent and non-violent crimes. Below are the attached snippets of the boxplots for both the violent and non-violent crimes.



Data Analysis: Tables or Graphs

Data processing:

The below table indicates some of the columns that the dataset consists of regarding the communities, houses, people in communities that are extracted as part of selected dataframe for analyzing the crimes data.

	CommunityName	state	countyCode	communityCode	fold	population	householdsize
0	BerkeleyHeightstownship	NJ	39.0	5320.0	1	11980	3.10
1	Marpletownship	PA	45.0	47616.0	1	23123	2.82
2	Tigardcity	OR	NaN	NaN	1	29344	2.43
3	Gloversvillecity	NY	35.0	29443.0	1	16656	2.40
4	Bemidjicity	MN	7.0	5068.0	1	11245	2.76

agePct12t21	agePct12t29	agePct16t24	agePct65up	numbUrban	pctUrban	medIncome	pctWWage	pctWFarmSelf
12.47	21.44	10.93	11.33	11980	100.0	75122	89.24	1.55
11.01	21.30	10.48	17.18	23123	100.0	47917	78.99	1.11
11.36	25.88	11.01	10.28	29344	100.0	35669	82.00	1.15
12.55	25.20	12.19	17.57	0	0.0	20580	68.15	0.24
24.46	40.53	28.69	12.65	0	0.0	17390	69.33	0.55

Missing value analysis:

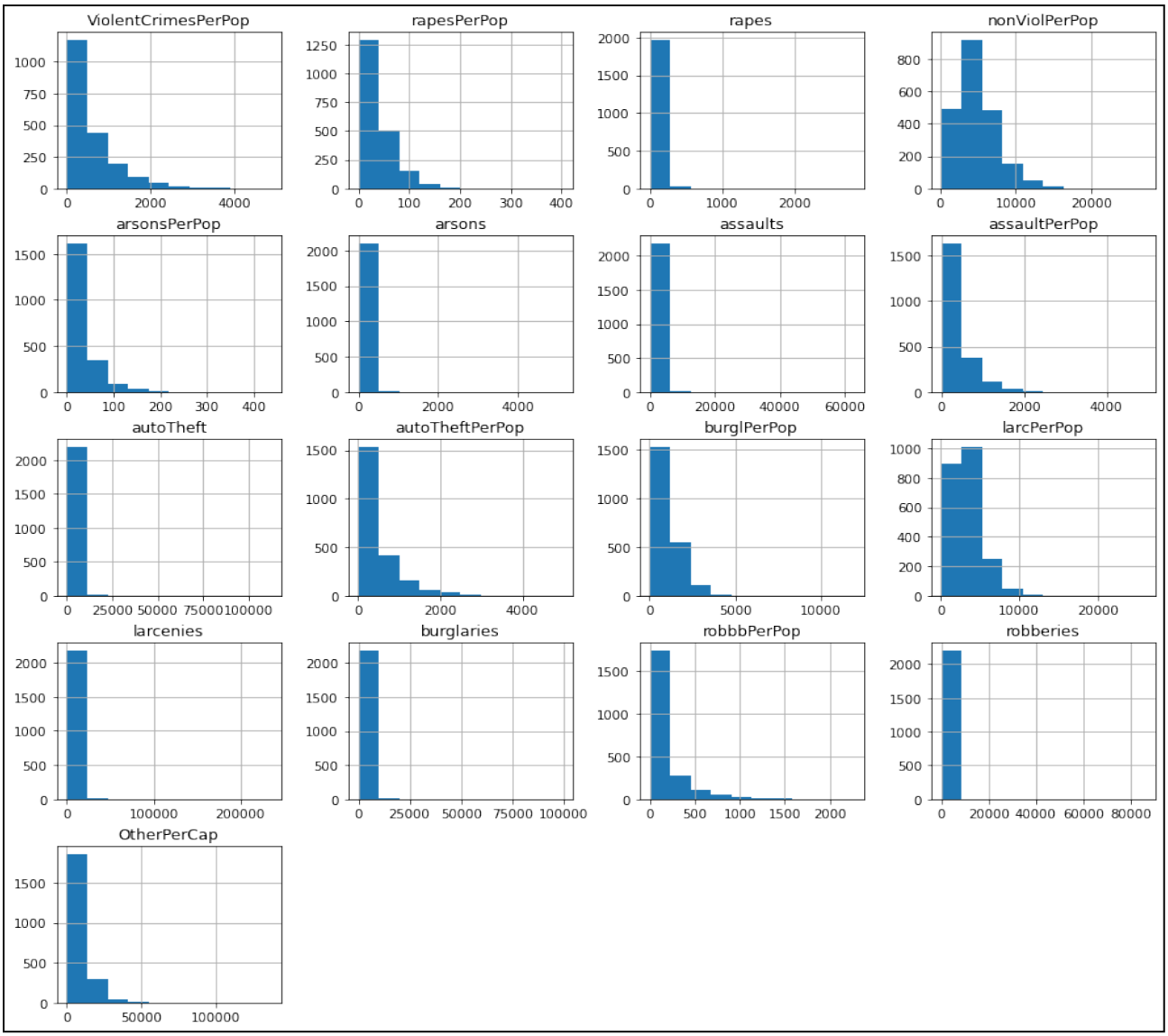
In our Dataset out of 147 columns, there are 41 columns with missing values. Our selected data frame has 17 columns whose missing values percentage is less than 60 to the Total Values. These missing values data has been shown in the table below along with their respective percentage to the total value.

Different Crimes	Missing Values	% of Total Values
ViolentCrimesPerPop	221	10.0
Rapes	208	9.4
rapesPerPop	208	9.4
nonViolPerPop	97	4.4
arsonsPerPop	91	4.1
arsons	91	4.1
assaults	13	0.6

assaultPerPop	13	0.6
arcPerPop	3	0.1
burglaries	3	0.1
larcenies	3	0.1
autoTheft	3	0.1
burglPerPop	3	0.1
autoTheftPerPop	3	0.1
robbbPerPop	1	0.0
obberies	1	0.0
OtherPerCap	1	0.0

Histogram of missing value data:

Their graphical representation is given, each graph plot demonstrating different crimes taken place in communities with the missing values under 60 percent to the total values.



Observing unique values:

Below table data is a sample output as part of data processing, consisting of unique values out of the total values from the data frame.

	DataTypes	Count of Unique Values	Unique Values
CommunityName	object	2018	[BerkeleyHeightstownship, Marpletownship, Tiga...
state	object	48	[NJ, PA, OR, NY, MN, MO, MA, IN, ND, TX, CA, K...
fold	int64	10	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
population	int64	2154	[11980, 23123, 29344, 16656, 11245, 140494, 28...
householdsize	float64	198	[3.1, 2.82, 2.43, 2.4, 2.76, 2.45, 2.6, 2.46, ...
racepctblack	float64	1172	[1.37, 0.8, 0.74, 1.7, 0.53, 2.51, 1.6, 14.2, ...
racePctWhite	float64	1609	[91.78, 95.57, 94.33, 97.35, 89.16, 95.65, 96....
racePctAsian	float64	667	[6.5, 3.44, 3.43, 0.5, 1.17, 0.9, 1.47, 0.4, 1...
racePctHisp	float64	1026	[1.88, 0.85, 2.35, 0.7, 0.52, 0.95, 1.1, 0.63,...
agePct12t21	float64	950	[12.47, 11.01, 11.36, 12.55, 24.46, 18.09, 11....
agePct12t29	float64	1184	[21.44, 21.3, 25.88, 25.2, 40.53, 32.89, 27.41...
agePct16t24	float64	947	[10.93, 10.48, 11.01, 12.19, 28.69, 20.04, 12....
agePct65up	float64	1221	[11.33, 17.18, 10.28, 17.57, 12.65, 13.26, 14....
numbUrban	int64	1600	[11980, 23123, 29344, 0, 140494, 28700, 59449,...
pctUrban	float64	293	[100.0, 0.0, 96.51, 33.1, 74.94, 73.6, 82.94, ...
medIncome	int64	2141	[75122, 47917, 35669, 20580, 17390, 21577, 428...
pctWWage	float64	1536	[89.24, 78.99, 82.0, 68.15, 69.33, 75.78, 79.4...
pctWFarmSelf	float64	290	[1.55, 1.11, 1.15, 0.24, 0.55, 1.0, 0.39, 0.67...
pctWInvInc	float64	1774	[70.2, 64.11, 55.73, 38.95, 42.82, 41.15, 47.7...
pctWSocSec	float64	1548	[23.62, 35.5, 22.25, 39.48, 32.16, 29.31, 30.2...
pctWPubAsst	float64	1125	[1.03, 2.75, 2.94, 11.71, 11.21, 7.12, 5.41, 8...
pctWRetire	float64	1258	[18.39, 22.85, 14.56, 18.33, 14.43, 14.09, 17....
medFamInc	int64	2150	[79584, 55323, 42112, 26501, 24018, 27705, 503...
perCapInc	int64	2069	[29711, 20148, 16946, 10810, 8483, 11878, 1819...

Imputing and scaling data for fitting using Standard scaler

We are scaling categorical columns with Binary Encoding and the sample output is shown below.

	state_0	state_1	state_2	state_3	state_4	state_5	population	householdsize	racepctblack	racePctWhite	racePctAsian	racePctHisp
0	0.0	0.0	0.0	0.0	0.0	1.0	0.000270	0.407609	0.014172	0.919030	0.112659	0.018493
1	0.0	0.0	0.0	0.0	1.0	0.0	0.001794	0.331522	0.008276	0.958123	0.059377	0.007670
2	0.0	0.0	0.0	0.0	1.0	1.0	0.002645	0.225543	0.007655	0.945333	0.059203	0.023432
3	0.0	0.0	0.0	1.0	0.0	0.0	0.000910	0.217391	0.017586	0.976483	0.008184	0.006094
4	0.0	0.0	0.0	1.0	0.0	1.0	0.000170	0.315217	0.005483	0.892006	0.019850	0.004203

Here is the code snippet for Standard scaler:

```
from sklearn.impute import KNNImputer
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
data_encoded = pd.DataFrame(scaler.fit_transform(data_encoded), columns = data_encoded.columns)
data_encoded.head()
```

Observing Skewness

Skewness is a measure of the asymmetry of a dataset or distribution, which helps us in understanding the shape of a distribution. Value can be either positive negative . The output for the selected dataframe skewness is shown below.

1--population --> 24.32061336600579
2--numbUrban --> 24.04630102925205
3--indianPerCap --> 15.822751795082514
4--NumUnderPov --> 23.81123970524638
5--NumKidsBornNeverMar --> 25.209632522300325
6--NumImmig --> 30.513238374922757
7--HousVacant --> 15.032465434484834
8--NumInShelters --> 33.19662489874914
9--NumStreet --> 36.19467694645503
10--LandArea --> 23.84973298935827
11--murders --> 22.333077417858267
12--rapes --> 14.306767566711596
13--robberies --> 29.24494594413932
14--assaults --> 22.501676907796067
15--burglaries --> 18.907390129535088
16--larcenies --> 17.885487908952758
17--autoTheft --> 23.100543324645944
18--arsons --> 20.186216997512563

Feature Engineering:

Feature Engineering is the technique of improving the performance on a dataset by transforming its feature space, and it is the practice of constructing suitable features from given features of the dataset, which leads to improving the performance of the prediction model.

We use 3 types of techniques in Feature selection process

1) Using Pearson correlation

2) VIF Analysis

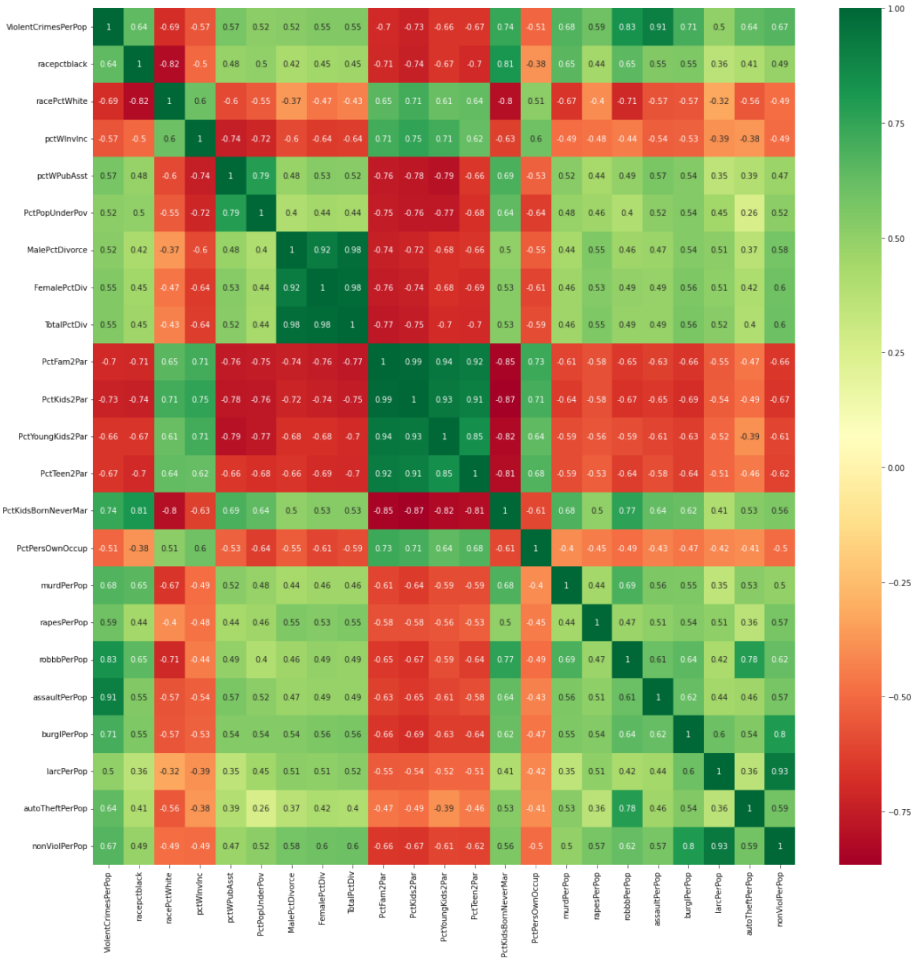
3) Feature Importance from Machine learning algorithm

We used “Pearson’s Correlation,” a feature selection method tells how much dependent variable (ViolentCrimesPerPop) is related to predictor variables. ‘.Corr()’ does pairwise positive/negative correlation on scale of 1 to 10.

We kept all the variables which have a correlation greater than 0.5 to the target variable.

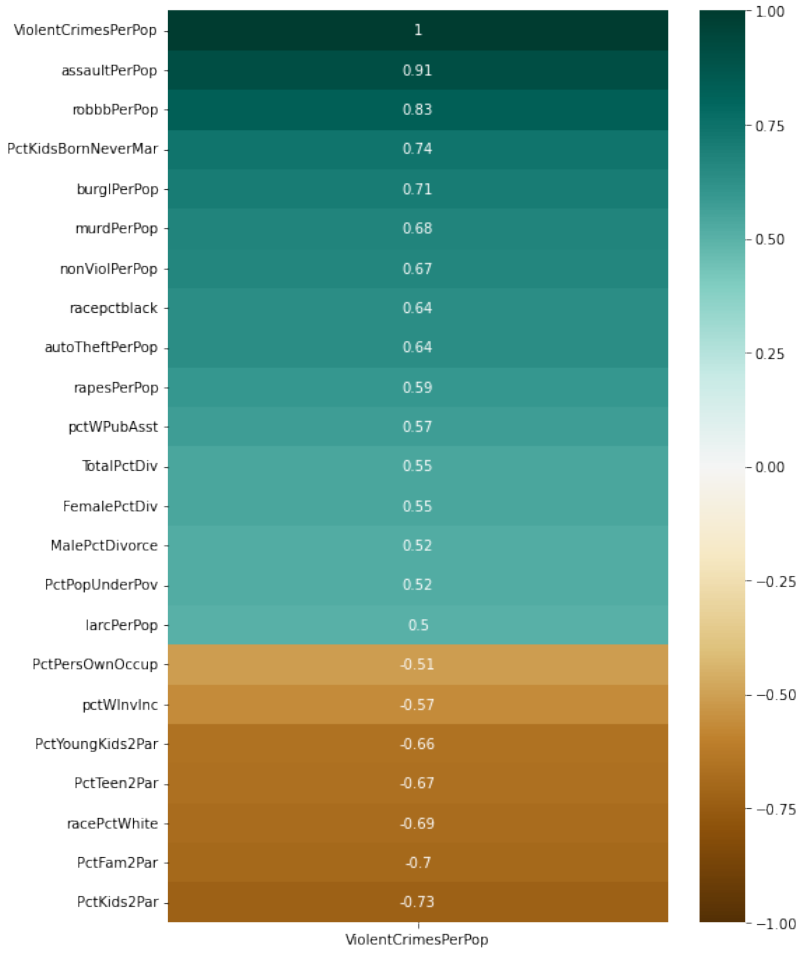
Pearson’s correlation Analysis:

Using heat map, correlation between different variables is shown.



Correlation:

Using “Pearson’s Correlation,” a feature selection method tells how much dependent variable (ViolentCrimesPerPop) is related to predictor variables. ‘.Corr()’ does pairwise positive/negative correlation on scale of 1 to 10. In crime dataset, variables are highly correlated with each other.



Variance Inflation Factor (VIF) :

VIF is calculated by taking the ratio of the variance of all a given model's betas divided by the variance of a single beta if it were fit alone. It is calculated by importing python package variance_inflation_factor . VIF is a measure of collinearity among predictor variables within a multiple regression.

Below is the sample table outcome for the selected data frame as part of VIF extraction for the existing crime data.

	feature	VIF
92	RentQrange	inf
91	RentHighQ	inf
85	OwnOccLowQuart	inf
89	RentLowQ	inf
87	OwnOccHiQuart	inf
88	OwnOccQrange	inf
6	population	3007.572807
16	numbUrban	2928.417330
47	TotalPctDiv	2886.603696
73	PctPersOwnOccup	926.010435
46	FemalePctDiv	913.793279
79	PctHousOwnOcc	892.472643
44	MalePctDivorce	611.385601
64	PctRecImmig8	606.329665
63	PctRecImmig5	403.707457
65	PctRecImmig10	353.924507
70	PersPerOccupHous	284.159989
86	OwnOccMedVal	255.036065
55	NumKidsBornNeverMar	249.236280
68	PctLargHouseFam	238.805725
33	NumUnderPov	218.563093
69	PctLargHouseOccup	200.304636
14	agePct16t24	193.060584
112	robberies	184.664131

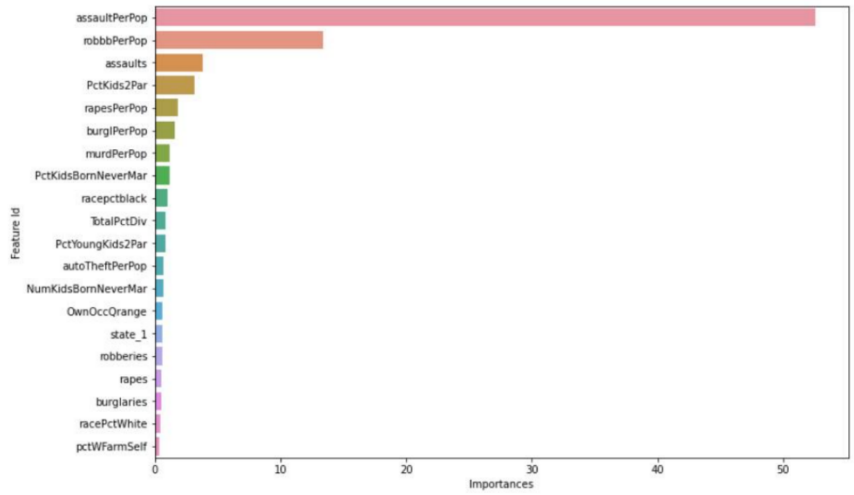
Below is the table data consisting of 21 records that is having VIF >= 5

	feature	VIF
8	TotalPctDiv	11994.140764
7	FemalePctDiv	4119.854548
6	MalePctDivorce	2214.100397
9	PctFam2Par	1692.189793
10	PctKids2Par	1322.026428
11	PctYoungKids2Par	380.602719
22	nonViolPerPop	222.465197
12	PctTeen2Par	210.127897
2	racePctWhite	125.727515
20	larcPerPop	110.730530
14	PctPersOwnOccup	54.037975
0	ViolentCrimesPerPop	42.147976
3	pctWInvInc	35.530951
19	burglPerPop	18.857282
18	assaultPerPop	17.885618
17	robbbPerPop	14.305160
4	pctWPubAsst	14.091265
13	PctKidsBornNeverMar	13.858658
5	PctPopUnderPov	11.542066
21	autoTheftPerPop	8.823914
1	racepctblack	8.327565

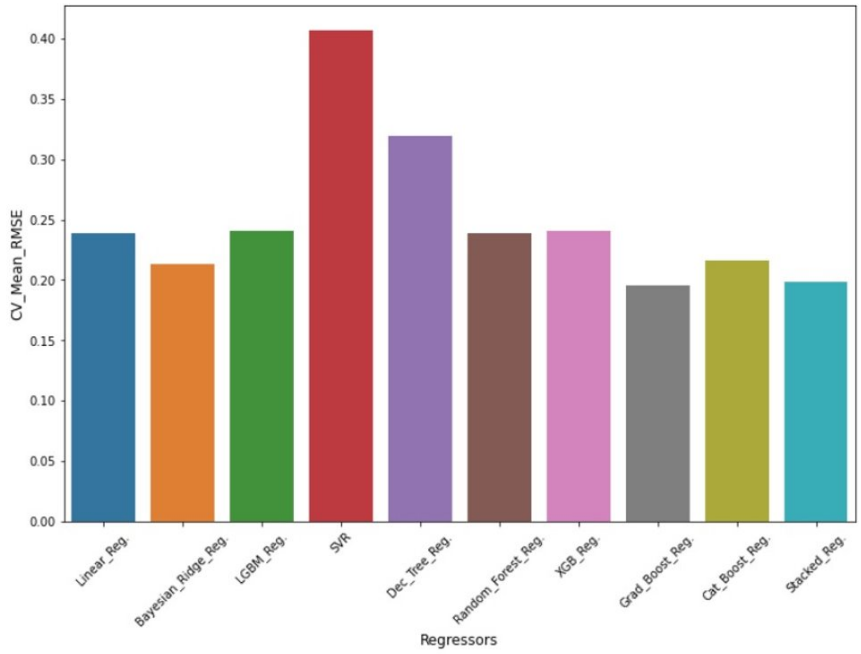
Feature Importance:

Key features are being selected from the output of the model using feature importance algorithm like Shaply or Algorithm_featureImportance(). Feature important selects feature which have the highest impact on the target variable.

Below is the attachment of the Feature Importance with the highest importance given to assaultPerPop, robberperpop, assaults, pctkids2par.



Model and Algorithms:



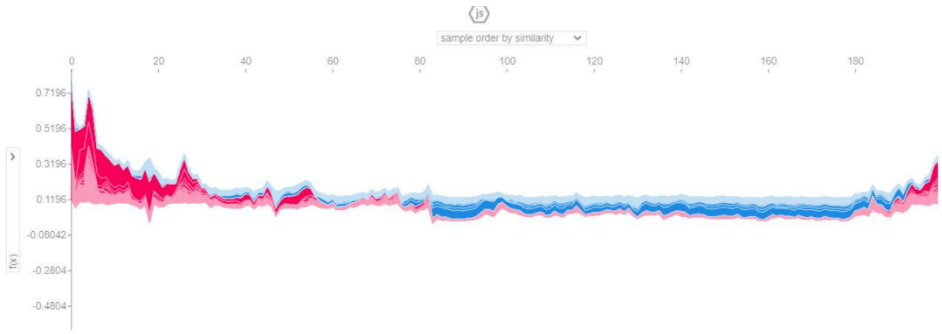
This is a regression problem since the target variable is numerical in nature (violentCrimeRate).

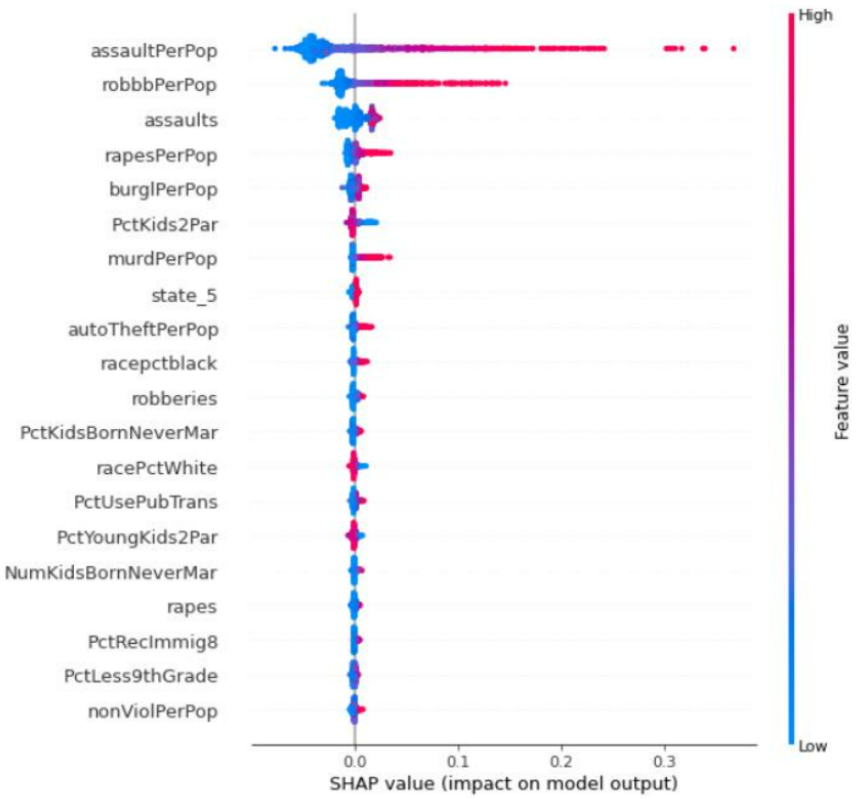
To create a model, we need to first split the data into training and validation set. In our analysis we have taken 33% of data for validation purposes. We performed a stacked algorithm, to observe the performance of each model. Considering the Random Forest Regressor provides the least score we decided to conclude it as our base model and start our research from there.

We first used Random Forest Regressor as our baseline model. A random forest is a meta estimator that employs averaging to increase predicted accuracy and reduce overfitting after fitting numerous classification decision trees to different dataset subsamples. Our random forest regression model will be trained using the sklearn package, more especially the RandomForestRegressor function.

Shaply Feature Importance:

By calculating the contribution of each feature to the prediction, SHAP seeks to explain the prediction of an instance x. Shapley values are calculated using the SHAP explanation approach using coalitional game theory. A data instance's feature values participate in a coalition as players. We can equitably distribute the "payout" (i.e., the prediction) among the characteristics by using Shapley values. A player could be a single feature value, for example in tabular data. A collection of feature values can also constitute a player. Pixels can be clustered into superpixels and distributed with the prediction, for instance, to describe an image. The Shapley value explanation is portrayed as an additive feature attribution approach in SHAP.

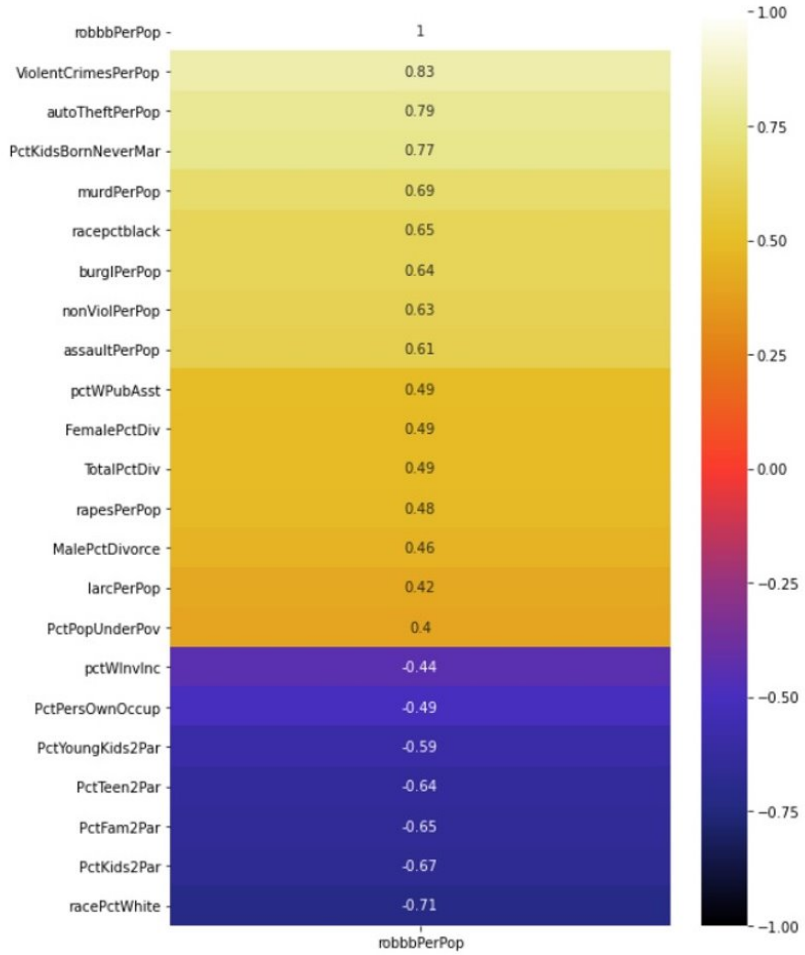




Evaluation Metrics:

- We created a lazy regressor which has all the possible machine learning models for a regression problem.
- We used RMSE as our score to select the best model.
- From our analysis Gradient boosting regressor gave us the best result with an RMSE of 0.19.

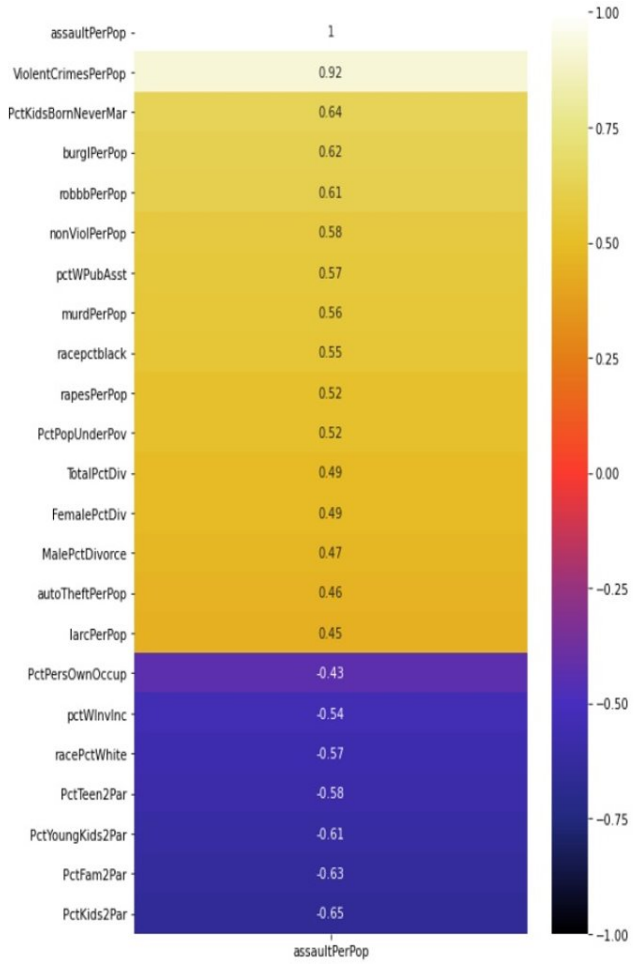
Root cause analysis:



Since Assault per population and Robbery per Population accounts for 95% of violent crimes we further dig deeper to find the reasons which lead to assaults and robberies.

After running correlation on assault and Robbery we found the following results:

- People who have been born to never married couples, committed non-violent crimes like burglary, and auto theft are more likely to commit robbery.



People who are born to never married, engaged in non-violent crimes, African-American race and people under poverty are most likely to commit an assault.

Solutions to Research Questions:

Q1) What are the top 2 violent crimes experienced in US communities?

- Stealing things by hurting/threatening people, Murders and Sexual Assault.

Q2) What are the factors that influence people in the US to commit violent crimes?

- There are many factors that make people commit violent crimes. However, most important ones are as follows:
    - Unemployment: Due to inflation in prices, an average salaried person commits crime to fulfill their comforts.
    - Income Inequality: Wealthy people becoming even wealthier and same case with poor people.

- Lack of Education: Children with no proper ethics, morals, values, and bad experiences.
- Orphans: Due to lack of supervision by elders, most of the kids end up pathetic and develop illegal activities.
- Homeless people, kids born to never married people, poverty.

Q3) What are the top 3 states with the highest crime rates?

- District of Columbia (Washington D.C)
- Los Angeles
- South Carolina

Q4) What are the top 10 safest states in the US for people to reside in.

- North Dakota
- Vermont
- Maine
- Wisconsin
- New Hampshire
- South Dakota
- Utah
- Connecticut
- Idaho
- Wyoming

Q5) What preventive measures can be taken to reduce crimes?

- Social/Community Services help reduce vulnerability in society.
- Supporting poverty and orphans with more educational opportunities.
- Vanishing Income Inequality.
- Counseling and creating awareness among people about repercussions that happen after committing crime, so that people understand and stop it.
- Spend government funds to improve the socioeconomic status of lower salary earners.

Conclusions:

Research Findings:

- Most of the time, people with financial problems tend to involve in illegal activities.
- From our analysis we can observe people who tend to commit violent crimes are the ones who have committed non-violent crimes in some part of their lives.
- People who run big households also tend to do violent crimes.
- These people are either born to unmarried couples who have abandoned them at the time of their birth and lack proper education, love, and guidance from their parents.
- Children who are homeless are more likely to be influenced by negative factors.
- Most of these people belong to Afro-American race or face poverty.

Recommendations/Suggestions:

- ACCESS TO FREE PRIMARY EDUCATION FOR ALL THOSE FACING POVERTY. (The cost of violent crimes is far more than free accessible education)
- ACCESS TO FOOD as deprivation of basic needs leads to criminal acts.
- Adoption of children should be highly encouraged.
- Certain schemes which are durable must be taken up by the government to reduce unemployment.
- Government should start schemes which improve the socioeconomic status of society.
- If all the precautions are taken, people in communities will feel safe to live-in.

Articulation of Response:

The most essential factor when working with any dataset is checking for co-relation between response variable and every predictor variable. Because if there is high correlation, there is valuable information left to analyze in the dataset. Correlation is analyzed using Heat Map. Here, in crime dataset, we have strong correlations with socioeconomic variables such as percentage of households with public assistance income and percentage of households with investment/rent income. Additionally, sometimes it may be encouraging for immigrants who struggle financially to meet necessities of living. From above analysis, cities with higher immigrants have high number of violent crimes.

From the models (Linear Regression, Random Forest, Decision Tree Regression, Bayesian Ridge, etc), we could see Gradient Booster Regressor has least error, so we conclude it is our base model to do further deductions, the topmost variables that we need to consider is percentage of kids born to never-couples, percentage of kids with two parents and percentage of families headed by two parents.

However, the results prove that most of the violent crimes are assault, robberies, rapes, and murders.