Statistical Inferencing of Violent Crimes in US from Socio-economic Factors

1. Motivation & Goals: Statistical analysis of violent crime rate is a topic of heavy debate and highly funded research in the field of statistics and economics. Our main endeavour is to understand the underlying factors contributing to the violent crimes in the US. Violent crime is an umbrella term that covers rage crimes, hate crimes, racist crimes, mafia crimes and felonies such as hit-and-run, murder, kidnapping and extortion among many others. Our dataset, taken from the UCI Machine Learning Repository, consists of roughly 2000 data vectors, each representing a city. Each data vector has 128 attributes, ranging from violent crime (target attribute) to various other fields such as percent white, income, number of illegitimate children, high school dropouts percentage. Each field is normalized to [0,1], using a "unsupervised equal frequency binning" method that retains skew and mutual correlation between the attributes of the data.

This report is structured into two broad sections. The first section, tackles three individual factors could contribute to violent crime − education, race and income. In the second section, we take a multifactor approach involving more than one of the factors. Finally, we propose a conclusion about the prediction of the target variable and the potential shortcomings of the model, along with suggestions on how to improve the model.

2. Education : ***It is a general notion that the more educated the society is; the less violent it will be.*** We assess this statement first by finding the correlation between the education distribution of the different cities, and the violent crime rates in the last 10 years. The data attribute we choose are the percentage of people who studied less than 9th grade (PL9), percentage who are not high school grads(PNHS), and the percentage of people who earned a Bachelor's degree(PBSM). We expect that the former two correlates positively with violent crime (VCPP) and the latter, negatively.

Figure 1(a) shows the distribution of PL9, which is skewed to the left. This indicates that in most cities 50% or less of the population have studied less than 9th grade. Figure 1(b) displays the distribution of PNHS. We can see that the graph is uniform in the middle and is tailed on each side. When compared with figure 1(a) we see it is more spread out, indicating that high school degree attainment is more varied by city. ***Unsurprisingly, correlation is high (0.9444) between the two sets: cities with more high school dropouts also have more 9th grade dropouts. Hence, we can drop an attribute while performing linear regression.*** Figure 1(c) shows the histogram of people who are college graduates, which again skews left indicating most cities are not college educated. This along with PNHS implies that many cities have a large percentage of people dropping out after high school. 1(d) shows the histogram of violent crimes and it is clear that a large number of cities have near zero violent crime rates and it decreases exponentially sporadic peaks.

In the scatterplots in Figure 2, we can verify the correlation between the variables described and VCPP. With PL9, we see that there is a strong positive correlation and it is the same with PNHS, with the correlation coefficients being ***0.8945*** and ***0.9217***. A strong negative correlation is seen in communities with higher PBSM, which is again expected. The correlation coefficient between the violent crimes and the proportion of bachelor's holders is ***-0.6191.***

Our second analysis is by comparing the education level of the cities that are most crime prone, and the least crime prone. We split the data set into 2, where the crime rate is greater than the median, and where it is lesser. Then we test the null hypothesis that PL9, PHNS are greater and PBSM is lesser in cities with more crime. We choose a ***two sample T-Test*** instead of a paired t-test since we have independent groups of data for each sample. We would expect the hypothesis to be rejected in all 3 cases. The results are tabulated in Table 1. To verify our intuition, we now split the data into 2 sets, one with cities less than 10% quantile, and another with 90% quantile of the violent crime data. Here, we expect that our results be further reinforced and the p-values of the hypothesis tests to be lower. As p-value is lesser it is more easier to reject null hypothesis now at the 5% significance level, since p-value is a measure of confidence. In conclusion, as we saw from the correlation analysis, variables PL9, PNHS and PBSM are highly correlated within themselves, and contribute only about 14% of the total variance (R^2) as we would see later. ***Hence, the VCPP prediction ability by the education attributes are 'counter-intuitively' lesser than expected.***

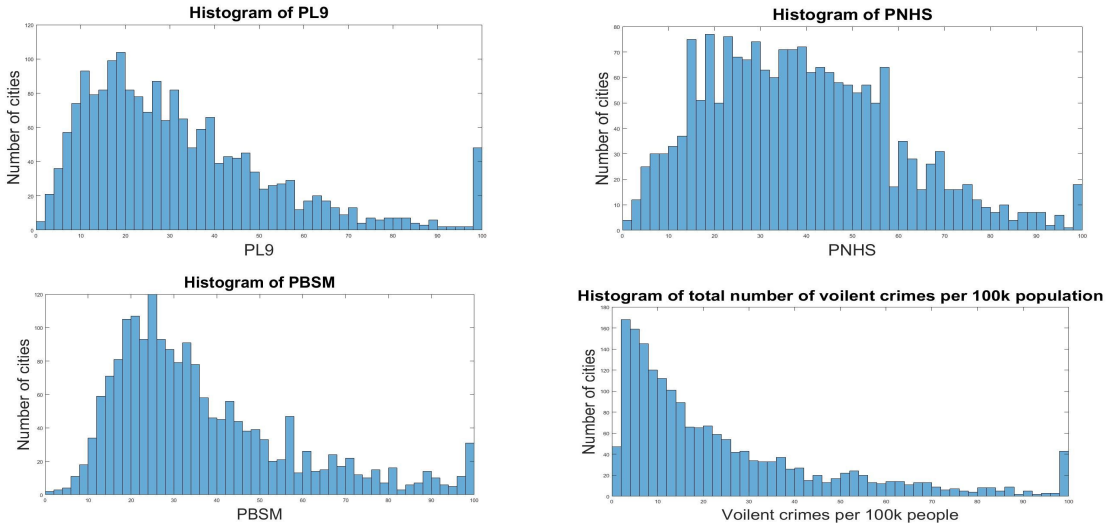


Figure 1(a-d) : Histograms of PL9, PNHS, PBSM and VCPP

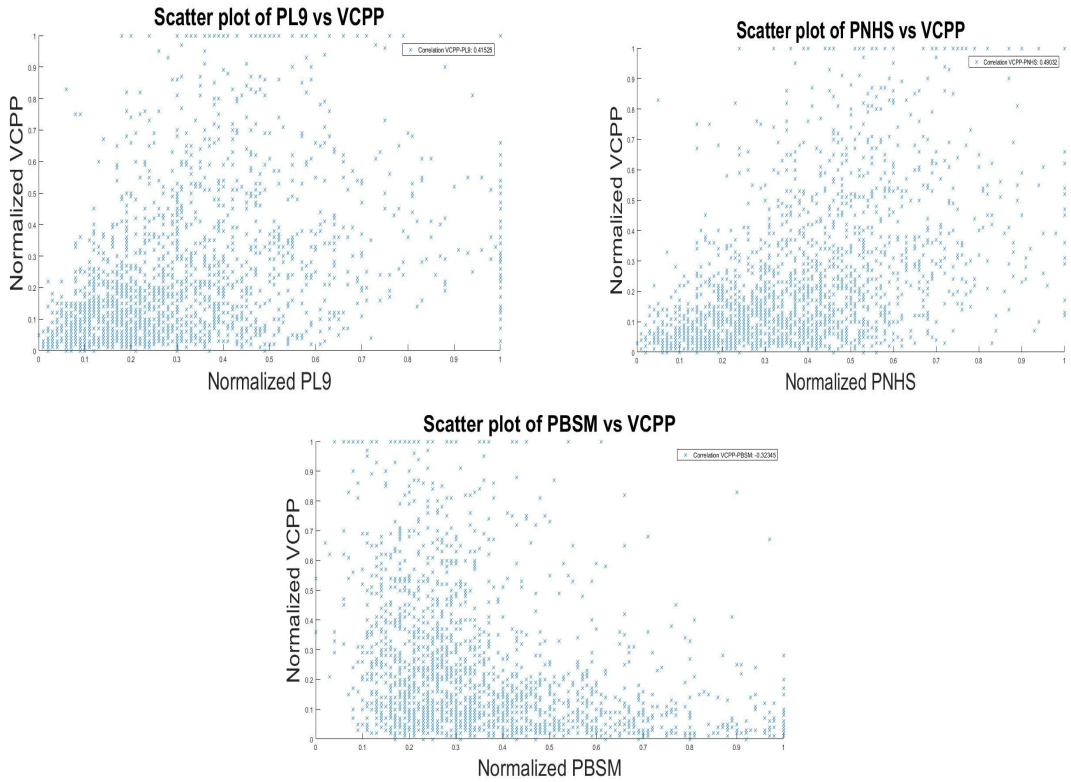


Figure 2(a-c):Scatterplots of people with different education background vs violent crimes per population

Split	Data	h	H1	p	95% CI for Δmean
Less than and greater than Median (M)	PC9	1	μ(<M) < μ(>M)	5.5097e-77	[0.0149, 0.3395]
Less than and greater than Median (M)	PHNS	1	μ(<M) < μ(>M)	8.8163e-100	[0.0138, 0.3640]
Less than and greater than Median (M)	PBSM	1	μ(>M) < μ(<M)	3.0893e-31	[-0.2755, -0.0151]
Less than 10% quantile and greater than 90% quantile	PBSM	1	μ(>M) < μ(<M)	4.7738e-53	[-0.0323,-0.4683]

3. Race:

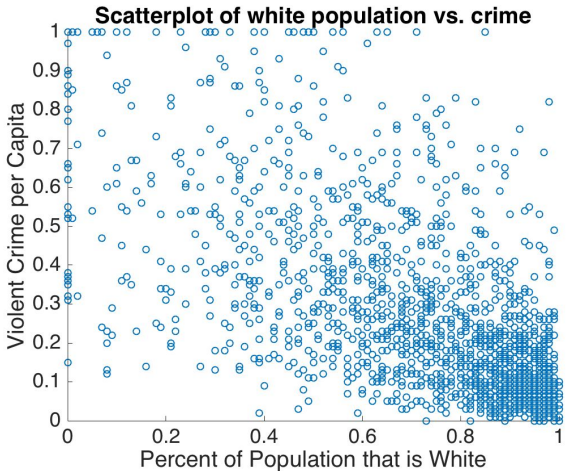


Figure 3.1: A scatter plot of white population vs. crime

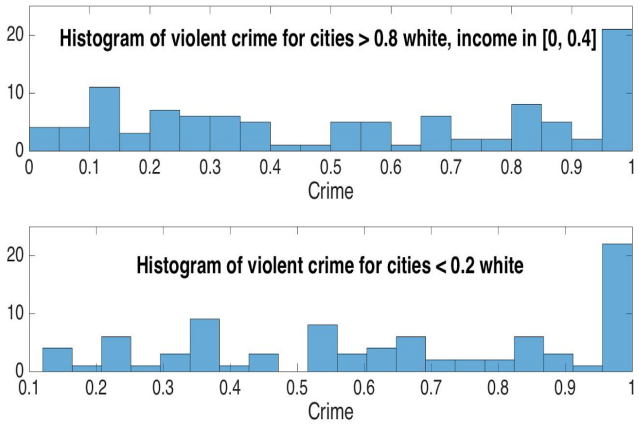


Figure 3.2: A plot comparing violent crime after adjusting for income.

Our first step was to understand the relationship between race and violent crime. It is widely held that the higher the ratio of Caucasians in a city, the lower the violent crime. In Figure 3.1, we can indeed see that this is the case. We calculate the coefficient of correlation between violent crime and white population and find a value of -0.68. That is a fairly high magnitude. However, there is a natural question with regards to this: are there any lurking factors? We suspected that income may account for this entire correlation, and so we set out to test this hypothesis.

Our goal, therefore, was to attempt to normalize for income and then compare races. We looked at the incomes of predominantly minority areas (defined as ranging from [0, 0.2] on the @percentWhite attribute) and predominantly white areas (defined as ranging from [0.8, 1] on the @percentWhite attribute) and compared their incomes. We found that the minority areas lay on the scale of [0, 0.4]. Consequently, we took all the predominantly white cities that fell into the range of [0, 0.4] (call this **D1**) and sought to compare their crime rates to the predominantly minority areas (call this **D2**). This is plotted in Figure 3.2, and it is clear that crime rates are more similarly distributed. Can we measure it?

We first tested to see if the crime rates of **D1** and **D2** were distributed similarly, by running a Kolmogorov-Smirnov test. This allows us to compare if two datasets come from the same distribution. We found they didn't at a 0.05-significance level (p=0.04). So we had reason to reject the notion that after cutting off the wealthier white areas, the two populations were distributed similarly. Therefore, we looked to compare their actual means. The mean crime of the minority area is 0.65, whereas the white area is 0.55. This seems to be a significant difference. We ran a Bayesian comparison of the means − Theorem 9.8.2 covers the exact procedure. This turned into a t-test with 190 degrees of freedom, with a t-score of -21. For this configuration, the p-value is less than 0.001, and so we rejected the null hypothesis.

However, there is one result to note. Running a Kolmogorov-Smirnov test on the *incomes* (instead of violent crime, as done above) of **D1** and **D2** gave us a p-value of 0.04 (rejected at a 0.05-significance level). What this meant was that, even after removing the wealthier predominantly-white areas, the incomes were not distributed equally. Therefore, we may say that while there is evidence to suggest that income does not fully account for the racial correlation with violent crime, it is not close to conclusive. We now take a look at income in particular.

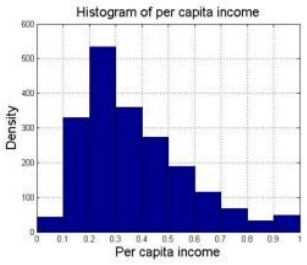
4. Income - To analyze how economic factors affect the crime rate in communities across the USA, we pick the per capita income of communities as a random variable to evaluate its effect on violent crime rate in the communities. The data was normalized by unsupervised binning method with equal frequencies to preserve the skew and the nature. We see from the histogram that this measure can be roughly estimated to be normally distributed with a ***mean of 0.37*** and ***standard deviation of 0.1983***.

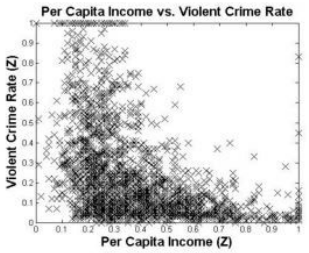
We plot scatter plots and find correlation coefficients. From the scatter plot, we see that the violent crime rate is clearly

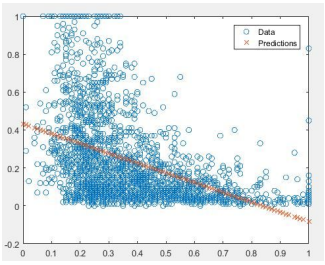
higher in communities with lower per capita income while the crime rate significantly decreases as the per capita income increases. Intuitively, this is to be expected as economic status generally correlates negatively with crime rate. The ***correlation coefficient*** between per capita income and violent crime rate in communities turns out to be ***-0.4391*** (p-value<0.001) showing a strong negative correlation.

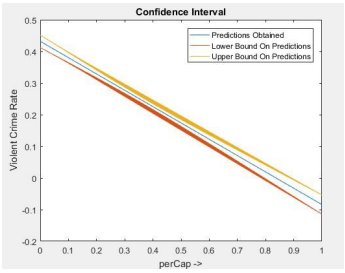
Now, with ***perCapitaIncome*** as ***predictorVariable (X)*** and ***violentCrimeRate*** to be ***responseVariable (y)***, linear-regression is applied to find out that ***beta0=0.43184***, ***beta1=-0.51603***, ***R2 statistic = 0.193***. Negative correlation obtained by scatter plot is supported by the negative slope for beta1.

Further, a study was made by sorting violent crime rate based on per-capita income. Now, the sorted crime rate samples were equally divided into set of two. The first set corresponds to crime-rates of lower end of per-capita income and the other is the crime-rates associated with higher end of per-capita income. Each of which were tested for null hypothesis having same values for mean. ***P-values*** thus obtained for ***2-sample t-test*** (mean) was ***8.1462e-86***, clearly ***rejecting null hypothesis***. Thus, it is certain that as per-capita income changes, crime rate is bound to change. This alludes to the inferences made from the aforementioned scatter plot, regression and making it more certain.









5. Multiple Linear Regression - After analyzing how racial factors, economic factors and education contribute to violent crime rate, we wanted to try to fit a more generic multi factor linear model to explain the rate of violent crime in different communities across the USA.

We first computed correlation coefficients between each of the predictive factors and the violent crime rate. We then ordered the factors based on which was most correlated with violent crime rate. The order was based on the absolute values of the correlation coefficients. We obtained the top ten features by following this method. As we were specifically interested in the education, racial and financial features, we included one of each of these features with the ten obtained. To ensure that we did not use features that were correlated among themselves in the regression, we removed a feature each from pairs that had a correlation coefficient of 0.6 or higher. In this manner, we arrived at the features shown in Table 5.1 to use to perform multiple linear regression. The second column lists the magnitude of the correlation coefficients. All correlation values were found to be statistically significant with p-values much lower than 0.001.

Table 5.1 : Features selected for multiple linear regression and their correlation coefficient with the violent crime rate

Factor Description	$\lvert \rho_{xy} \rvert$
Illegitimate kids	0.74
Kids in family housing with 2 parents	0.74
Population that is caucasian	0.68
Females who are divorced	0.56
Per capita income	0.44

We then performed multiple linear regression using the formula : $B = (Z'Z)^{-1}Z'y$, where B is the matrix of coefficients, Z is the design matrix with the feature vectors and y is the vector of values of the violent crime rate. The B values obtained and the results of the regression are discussed in section 5.2.

5.1 Analysis of Residuals : It is important to verify that the observed data appear to satisfy the assumptions on which the analysis is based.One of the restriction is that the sum of error residuals should be zero and in our case it is -1.1019e-12.The error residuals are assumed to be normally distributed with mean 0 and variance σ^2. The histogram plot of the residuals resembles a normal distribution with mean zero and is shown in figure 5.1 (a). The normal quantile plot compares quantiles of a normal distribution with the ordered values of the residuals. It is seen that the points roughly lie on the line, and it is common to see a certain curvature from in the plot at the ends. One typically pays attention to the middle of the plot and it is fairly linear as shown in Figure 5.1 (b)

For the 1994 sample points, the plot of a couple of predictors and residuals (xi , ei), for i = 1, . . . , 1994 are shown alongside.If the crime rate is truly a linear function of the predictor and its observations are independent, the positive and negative residuals are scattered randomly among the points .Otherwise, the residuals tend to be concentrated at the centre and tend to exhibit a regular pattern.Figure 5.1 (c) shows that crime rate is not a linear function of the percentage of households with social security income while figure 5.1(d) illustrates that violent crime rate varies with percentage of population that is caucasian, in a linear fashion. Similar plots for the other predictors used for the above multiple regression were also analysed and were found to satisfy the assumptions.

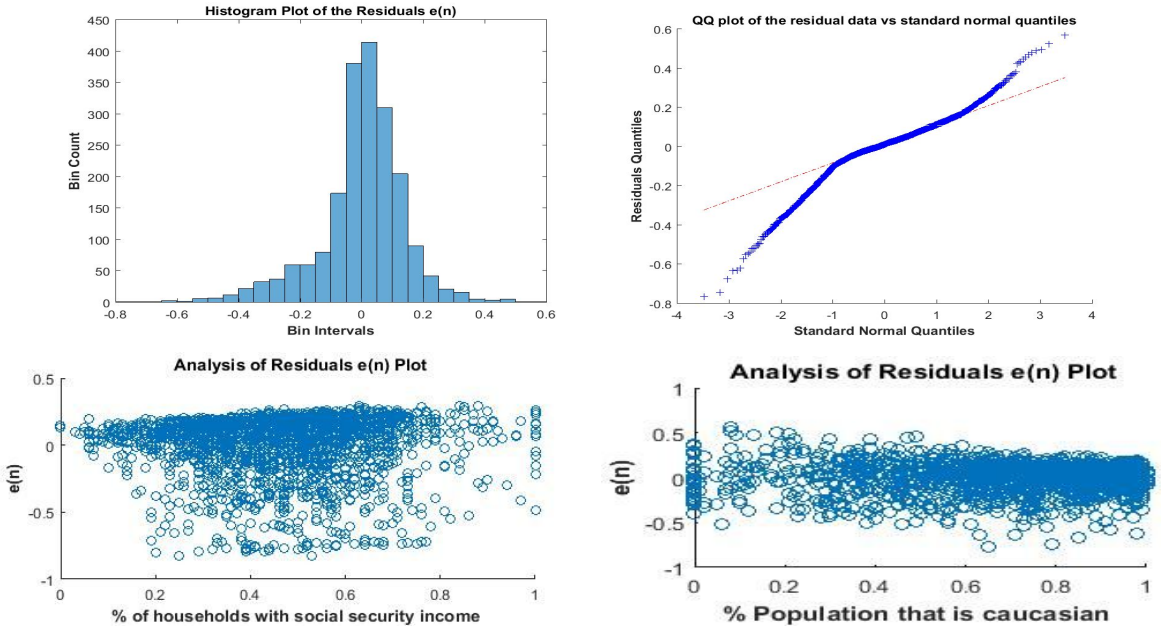


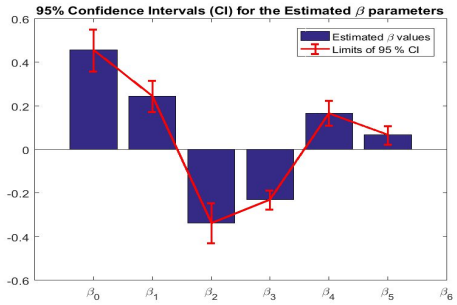
Figure 5.1 : **a)** Histogram plot of the residuals e(n). **b)** QQ plot of the residual data vs the standard normal quantiles. **c)** Analysis of residuals e(n) plot for a 'bad' linear predictor % of household with social security income. **d)** Analysis of residuals e(n) plot for a 'good' linear predictor % of population that is caucasian.

5.2 Results of the regression : From Table 5.2, we see that the B values obtained roughly follow the same trend as the correlation coefficients. Since we are using the pseudoinverse, the coefficients for correlated features are not dependent on their position in the design matrix. Interestingly, we observe that as the percentage of illegitimate kids in the society and the percentage of divorced females in the society increases, the violent crime rate tends to increase while it decreases with increase in the percentage of population that is Caucasian and with increase in the percentage of kids living in family housing with two parents. The R^2 statistic obtained is 0.6152 which means that our general linear model is able to explain 61.52% of the variance in the data with just 5 features. The mean squared error is 0.0209 which also suggests that the model fits the data very well.

To analyze how well our model predicts, we pick two samples randomly without replacement from the dataset. We use our model to compute the predicted value of the violent crime rate in that community. We also compute the prediction interval. We see that the predicted values are close to the actual values and they lie in the prediction interval.

Figure 5.2: 95% Confidence Interval plot for the estimated beta parameters. (right up)
Table 5.2: The estimated Parameters for the multiple linear regression model. (left)

Feature	B value
Bias term	0.46
Illegitimate kids	0.25
Kids in family housing with 2 parents	-0.34
Population that is Caucasian	-0.23
Females who are divorced	0.17
Per capita income	0.07



Actual violent crime rate	0.06	0.05
Predicted violent crime rate	0.043	0.28
Prediction interval	(-0.2414, 0.3274)	(-0.0073, 0.5610)

Table 5.3 Actual, predicted and prediction interval of the percentage of violent crime rate. (Right bottom) (----------↕)

5.3) A Few Interesting Hypothesis Tests :-

1. Null hypothesis - Violent crime rate in a community is independent of the percentage of females who are divorced ($\beta_4 = 0$).We get a P value = 1.0571e-08, Conclusion - We reject the null hypothesis at 5% significance level.

2. Null hypothesis - Violent crime rate in a community increases with increase in the percentage of illegitimate kids in the community ($\beta_1 > 0$).We get a P value = 1, Conclusion - We do not reject the null hypothesis at 5% significance level.

3. Null hypothesis - Violent crime rate in a community decreases with increase in the percentage of caucasians in the community ($\beta_3 < 0$).We get a P value = 1, Conclusion - We do not reject the null hypothesis at 5% significance level.

6. Conclusions and Future Work :

Our goals were to learn about the underlying factors behind crime to more efficiently address this problem. By intuition, we suspected that education, racial composition, and income would all be important factors. In terms of education, we found that the most correlated educational component is the percentage of high school graduates, versus college graduates or 9th grade dropouts. But we found that all of the attributes were counter-intuitively not well correlated with the crime rate. In terms of racial composition, we found that income does not account for the racial correlation with violent crime, but it does explain a good portion of the association. In terms of income, we found the unsurprising result that higher income leads to lower crime. We then performed a general linear regression on the data using more factors. We found that our initial three factors were not as predictive as we originally thought. Rather, social factors such as the percentage of illegitimate kids and the percentage of women who are divorced seem to explain more of the data.

Inference : Of course, these social factors correlate with our initial three factors. However, it suggests a more prudent use of money and time to fight violent crime would be to address underlying social factors. Incentivizing adoption or researching ways that local communities can promote healthy marriages might be a better avenue to fix.

Future Work : As a continuation to the current analysis, if we can get more data (which in turn gives us more features), we can learn more accurately about the underlying factors by using better regression methods and nonlinear models. We could also perform more intuitive hypothesis tests, and better model fitting to glean better information from the data. We mainly believe that the lack of the data-points (around 2000), could be a major reason for the incapability to predict the crime rate well.

Appendix:

A] Education :

```
clc;
close all;
load('crime_prog.mat')
cities = crime(:,4);
PctLess9thGrade = table2array(crime(:,35));
PctNotHSGrad = table2array(crime(:,36));
PctBSorMore = table2array(crime(:,37));
ViolentCrimesPerPop = table2array(crime(:,128));
ViolentCrimesPerPop(end)=median(ViolentCrimesPerPop(1:end-1));

PL9=PctLess9thGrade;
PNHS=PctNotHSGrad;
PBSM=PctBSorMore;
VCPP=ViolentCrimesPerPop;

[H,P,CI,STATS]=ttest(PctLess9thGrade);

c9th=corrcoef(PL9,ViolentCrimesPerPop,'rows','complete');
chs=corrcoef(PNHS,ViolentCrimesPerPop,'rows','complete');
cbs=corrcoef(PBSM,ViolentCrimesPerPop,'rows','complete');

figure;
scatter(PBSM,ViolentCrimesPerPop,'x');
hold on;
scatter(PBSM,ViolentCrimesPerPop,'x');
title('Scatter plot of PBSM vs VCPP','Fontsize',30);
xlabel('Normalized PBSM','Fontsize',30);
ylabel('Normalized VCPP','Fontsize',30);
legend('Correlation VCPP-PBSM: '+string(cbs(1,2)),'Location','northeast');

figure;
scatter(PL9,ViolentCrimesPerPop,'x');
title('Scatter plot of PL9 vs VCPP');
xlabel('Normalized PL9');
ylabel('Normalized VCPP');
% legend('Correlation : '+string(c9th(1,2)),'Location','southoutside');

% figure;
% histogram(PctLess9thGrade*100,50);
% hold on;
% histogram(PctNotHSGrad*100,50);
% hold on;
% histogram(PctBSorMore*100,50);
% title('Histogram of PL9, PNHS and PBSM');
% ylabel('Number of cities');
% xlabel('Percentage of people');
% legend('PL9','PNHS','PBSM');

title('Histogram of percentage of people who are not high school graduates');
ylabel('Number of cities');
xlabel('Percentage of people who are not high school graduates');
title('Histogram of percentage of people who hold a Bachelors degree');
ylabel('Number of cities');
xlabel('Percentage of people who hold a Bachelors degree');
```

```
figure;
histogram(PBSM*100,50);
title('Histogram of PBSM','Fontsize',30);
ylabel('Number of cities','Fontsize',30);
xlabel('PBSM','Fontsize',30);

% Finding indices of cities with crime rates > or <= median
index_gr_median=find(VCPP>median(VCPP));
index_lessEqual_median=find(VCPP<=median(VCPP));
% Testing the hypothesis that (%<9th) > for more violent cities
[h,p,ci,stats]=ttest2(PL9(index_gr_median),PL9(index_lessEqual_median),'Tail','right','Vartype','unequal')
% Testing the hypothesis that (%<HS) > for more violent cities
[h1,p1,ci1,stats1]=ttest2(PNHS(index_gr_median),PNHS(index_lessEqual_median),'Tail','right','Vartype','unequal')
% Testing the hypothesis that (%<HS) > for more violent cities
[h2,p2,ci2,stats2]=ttest2(PBSM(index_gr_median),PBSM(index_lessEqual_median),'Tail','left','Vartype','unequal')

% Finding indices of cities with crime rates > or <= median
quantile10=quantile(VCPP,0.1);
quantile90=quantile(VCPP,0.9);
index_lessEqual_quantile10=find(VCPP<=quantile10);
index_gr_quantile90=find(VCPP>=quantile90);
% Testing the hypothesis that (%<9th) > for more violent cities
[h,p,ci,stats]=ttest2(PL9(index_gr_quantile90),PL9(index_lessEqual_quantile10),'Tail','right','Vartype','unequal')
% Testing the hypothesis that (%<HS) > for more violent cities
[h1,p1,ci1,stats1]=ttest2(PNHS(index_gr_quantile90),PNHS(index_lessEqual_quantile10),'Tail','right','Vartype','unequal')
% Testing the hypothesis that (%<HS) > for more violent cities
[h2,p2,ci2,stats2]=ttest2(PBSM(index_gr_quantile90),PBSM(index_lessEqual_quantile10),'Tail','left','Vartype','unequal')
```

B] Racial Crime:

```
percentWhite = communities(:, 9);
percentWhite = [percentWhite{:}];
violentCrime = communities(:,128);
violentCrime = [violentCrime{:}];
figure;
scatter(percentWhite, violentCrime)

title('Scatterplot of white population vs. crime', 'FontSize', 18)
xlabel('Percent of Population that is White', 'FontSize', 18)
ylabel('Violent Crime per Capita', 'FontSize', 18)

R = corrcoef(percentWhite, violentCrime)

figure;
subplot(2, 1, 1);
histogram(percentWhite, 50)
title('Histogram of white population, 50 bins')
xlabel('Percentage white')
ylabel('Count')
subplot(2, 1, 2);
histogram(violentCrime, 50)
title('Histogram of crime rates in cities, 50 bins')
xlabel('Crime Rate')
ylabel('Count')

whiteCrime = violentCrime(percentWhite>0.80);
minorityCrime = violentCrime(percentWhite<0.20);
```

```
figure;
subplot(3, 1, 1);
histogram(whiteCrime, 20);
title('Histogram of crime rates for cities > 0.8 white')
xlabel('Crime Rate')
subplot(3, 1, 2);
histogram(minorityCrime, 20);
title('Histogram of crime rates for cities < 0.2 white')
xlabel('Crime Rate')
subplot(3, 1, 3);

histogram(minorityCrime(minorityCrime<1), 10);
title('Histogram of crime rates for cities < 0.2 white, excluding crime rate = 1')
xlabel('Crime Rate')

%----------%
income = communities(:,26);
income = [income{:}];

whiteIncome = income(percentWhite>0.80);
minorityIncome = income(percentWhite<0.20);

mean(income(percentWhite>0.80))

figure;
subplot(2, 1, 1);
histogram(whiteIncome, 20);
title('Histogram of income for cities > 0.8 white')
xlabel('Income')
subplot(2, 1, 2);
histogram(minorityIncome, 20);
title('Histogram of income for cities < 0.2 white')
xlabel('Income')

combined = violentCrime(percentWhite>0.80 * income < 0.40);
figure;
subplot(2, 1, 1);
histogram(combined, 20);
title('Histogram of violent crime for cities > 0.8 white, income in [0, 0.4]')
xlabel('Crime')
subplot(2, 1, 2);
histogram(minorityCrime, 20);
title('Histogram of violent crime for cities < 0.2 white')
xlabel('Crime')

[h,p] =kstest2(combined, minorityCrime)

%bayesian test
m = size(minorityCrime, 2);
n = size(combined, 2);
t = sqrt(m+n-2) * (0 - 0 - (mean(minorityCrime) - mean(combined))) / (1/m+1/n)^0.5 / (var(minorityCrime)+var(combined))^0.5;

% Compare distributions of income
poorWhiteIncome = income(percentWhite>0.80 * income < 0.40);
poorMinorityIncome = income(percentWhite<0.20 * income < 0.40);
[h,p] =kstest2(combined, minorityCrime)
```

C] Multiple Linear Regression

```
%% COGS 243 Project
% Title - Multiple Linear Regression
% By - Girish Bathala and Shreyas Udupa Balekudru
% Predictors -
% Percent of Illegitimate kids
% Percent of Kids in family housing with 2 parents
% Percent of Population that is caucasian
% Percent of Females who are divorced
% Percent of Per Capita Income
% Response Variable
% Percent of violent Crime Rate
%% Clean Up
clc;
clear;
close all;
%% Load     Dataset
load('communities_data.mat');
load('meanvarcorr.mat');
communities = communities(:,6:end);
crime_rate = communities(:,123);
X = communities(:,1:122);
Y = crime_rate;
meanvarcorr1 = meanvarcorr;
corr_ind = 6;
meanvarcorr(:,corr_ind) = abs(meanvarcorr(:,corr_ind));
sorted_mevaco = flipud(sortrows(meanvarcorr,corr_ind));
start = 99;
stop = 101;
selected = sorted_mevaco(start:stop)
selected = [51    45    44     4    41 21];
sel = selected;
n=length(crime_rate);
Z = ones(n,1);
for i=1:length(selected)
   Z = [Z, communities(:,selected(i))];
end
beta = pinv(Z'*Z)*Z'*crime_rate
crime_pred = Z*beta;
residual = Z*beta - crime_rate;
mse = mean((residual).^2)
y_bar = mean(crime_rate);
den = sum((crime_rate- y_bar).^2);
num = sum((crime_rate-crime_pred).*(crime_rate-crime_pred));
R2 = 1 - (num/den)
%% Main Corr between feture and drime Rate
cnt = 1;
 for i = 1:size(X, 2)
    [R, P] = corrcoef(Y, X(:, i));
    p_value = P(1,2);
    % Eliminate the Xi's that do not have a significant correlation with Y (p>0.05).
    if p_value > 0.05
       fprintf('Between Y and X%d: Correlation Coefficient = %f, p-value = %f  -  ELIMINATED\n', i, R(1,2), P(1,2));

    elseif p_value <= 0.05
       fprintf('Between Y and X%d: Correlation Coefficient = %f, p-value = %f  -  ACCEPTED\n', i, R(1,2), P(1,2));
        sel(cnt,1) = i;
       cnt = cnt + 1;
```

```
    end
end
cnt


%% Corr Matrix
p_val_mat = eye(length(sel));
corr_mat = p_val_mat;
for i = 1:length(sel)
   for j = 1:length(sel)
      [R, P] = corrcoef(communities(:, sel(i)), communities(:, sel(j)));
      p_value = P(1,2);
      p_val_mat(i, j) = p_value;
      corr_mat(i, j) = R(1,2);
  % For Xi and Xj that show a significant correlation (p?0.05), eliminate one of the two ables.
      %fprintf('Between X%d and X%d: Correlation Coefficient = %f, p-value = %f\n', i, j, R(1,2), P(1,2));
   end
end

%% Analysis of Residuals Plots
figure;
for i=1:length(selected)
   figure;
   %subplot(3,2,i);
   scatter(communities(:,selected(i)),residual);
   title(' Analysis of Residuals e(n) Plot');
   ylabel('e(n)','FontWeight','bold');
end
saveas(gcf,'analysisofresidual.png');
%% Histogram Plots
figure;
histogram(residual);
title('Histogram Plot of the Residuals e(n)');
ylabel('Bin Count','FontWeight','bold');
xlabel('Bin Intervals','FontWeight','bold');
saveas(gcf,'hist_resid.png');
%% Predicted vs Actual
figure;
scatter(crime_pred,residual)

%% QQplot
figure;
qqplot(residual);
title('QQ plot of the residual data vs standard normal quantiles');
ylabel('Residuals Quantiles','FontWeight','bold');
xlabel('Standard Normal Quantiles','FontWeight','bold');
saveas(gcf,'qq_resid.png');
%% R2 Statistic
y_bar = mean(crime_rate);
den = sum((crime_rate- y_bar).^2);
num = sum((crime_rate-crime_pred).*(crime_rate-crime_pred));
R2 = 1 - (num/den)

%% Hypothesis Testing - Feature 4 - Females who are divorced
p = length(selected)+1;
beta_test = 5;
beta_j = 0;
cov_mat = pinv(Z'*Z);
s2 =  sum((crime_rate-crime_pred).*(crime_rate-crime_pred));
sigma_1 = sqrt( s2 / (n-p));
```

```
u1 = (beta(beta_test) - beta_j )/(sqrt( cov_mat(beta_test,beta_test) ) * (sigma_1) );
deg_fre = n - p;
p_val1 = 2*(1 - tcdf(u1,deg_fre))

%% Hypothesis Testing  - Feature  - Caucasian
p = length(selected)+1;
beta_test = 4;
beta_j = 0;
cov_mat = pinv(Z'*Z);
s2 =  sum((crime_rate-crime_pred).*(crime_rate-crime_pred));
sigma_1 = sqrt( s2 / (n-p));
u1 = (beta(beta_test) - beta_j )/(sqrt( cov_mat(beta_test,beta_test) ) * (sigma_1) );
            deg_fre = n - p;
p_val2 = (1 - tcdf(u1,deg_fre))

%% Hypothesis Testing - Feature  - Illegitimate Kids
p = length(selected)+1;
beta_test = 2;
beta_j = 0;
cov_mat = pinv(Z'*Z);
s2 =  sum((crime_rate-crime_pred).*(crime_rate-crime_pred));
sigma_1 = sqrt( s2 / (n-p));
u1 = (beta(beta_test) - beta_j )/(sqrt( cov_mat(beta_test,beta_test) ) * (sigma_1) );
deg_fre = n - p;
p_val1 = (1 - tcdf(-u1,deg_fre))
%% Prediction Interval - Example row number 1222

z = [ones(1,1)];
for i=1:length(selected)
    z = [z; mean(communities(:,selected(i)))];
end
sel = 1222;
z = Z(sel,:)';
y_a = crime_rate(sel);
y_p = z'*beta;
p = length(selected)+1;
s2 =  sum((crime_rate-crime_pred).*(crime_rate-crime_pred));
sigma_1 = sqrt( s2 / (n-p));
alpha = 0.05;
deg = n-p;
pred_int = tinv(1-(alpha/2),deg)*sigma_1*(sqrt(1+z'*pinv(Z'*Z)*z));
pred_interval_lower = y_p - pred_int;
pred_interval_upper = y_p + pred_int;

%% Prediction Interval - Example row number 800
z = [ones(1,1)];
for i=1:length(selected)
    z = [z; mean(communities(:,selected(i)))];
end
sel = 800;
z = Z(sel,:)';
y_a = crime_rate(sel);
y_p = z'*beta;
p = length(selected)+1;
s2 =  sum((crime_rate-crime_pred).*(crime_rate-crime_pred));
sigma_1 = sqrt( s2 / (n-p));
alpha = 0.05;
deg = n-p;
pred_int = tinv(1-(alpha/2),deg)*sigma_1*(sqrt(1+z'*pinv(Z'*Z)*z));
pred_interval_lower = y_p - pred_int;
```

```
pred_interval_upper = y_p + pred_int;

%% Confidence Intervals for beta
[b,bint]=regress(crime_rate,Z);
e = b - bint(:,1);
figure;
bar(b');
hold on;
errorbar(b',e','LineWidth',1.8,'color','r');
title('95% Confidence Intervals (CI) for the Estimated \beta parameters');
legend('Estimated \beta values','Limits of 95 % CI');
xticks(1:11);
xticklabels({'\beta_0','\beta_1','\beta_2','\beta_3','\beta_4','\beta_5','\beta_6','\beta_7','\beta_8','\beta_9','\beta_{10}'});
saveas(gcf,'betaCI.png');
```

D] Income:

```
clc;
close all;
clear

load('finance.mat');
violentCrime = communities(:,end);


[R,P] = corrcoef(perCap,violentCrime)


model = fitlm(perCap,violentCrime, 'Linear')


ci = coefCI(model)


[yPred, yCI] = predict(model,perCap);



array = horzcat(perCap,violentCrime);
sortedResult = sortrows(array,1);

sortedCrimeRate = sortedResult(:,2);

sortedCrimeRateFirstHalf = sortedCrimeRate(1:997);
sortedCrimeRateSecondHalf = sortedCrimeRate(998:1994);

[h_mean,pValueMean] = ttest2(sortedCrimeRateFirstHalf, sortedCrimeRateSecondHalf)
```