Crime Dissipation Using Crime Data

Anna Damir

1) Introduction/Preface

Ding. Another email in my inbox, subject line: [dps-l]Crime Alert. I open the email and skim through, knowing that if my life was in immediate danger, the campus wouldn't choose email to notify me. Scanning past the case and alert details, my eyes are drawn to "Reported Offense", another strong-arm robbery. The date and time of occurrence, the first week of classes at university, in broad daylight, on a large straight, with probably dozens of witnesses, about a block from campus. The incident description offers a relatively cut and dry story of someone being shoved to the ground and their possessions stolen and the suspect flees. This specific email came into my inbox all of 2 days ago, nearly two years after I graduated; thanks to the amazing power of email forwarding, I am still in-the-know to what is happening on campus. But this type of email didn't start just recently. It became a norm for me while I was in school. Nearly every other day a new one would show up in my inbox; some were violent crime, and some were non-violent. I was lucky to grow up somewhere I felt safe walking alone at night, but when I started going to school in South Central Los Angeles, it really opened my eyes to how much crime was occurring around us, and how many, although cautious, had grown accustomed to it, and just expected it to happen. A lot of students are now afraid to go off campus alone at night, and their fears are justified: 3 students were killed in the last 6 years while just outside campus. I started looking into crime rates due to this reason. I had an idea of safety which was shattered by my time there, knowing that crime was happening around me near constantly, and no one should have to live in fear for their possessions or their lives. The overall goal of this is to, even on an extremely small scale, help in making our country a safer place to prosper in.

2) The Problem

The public perceives crime to be a problem, sometimes more so than the numbers would indicate (media sensationalism), yet based on a 2017 Gallup poll 30% of Americans are fearful of walking alone near their homes at night, and 22% of Americans say their household was victimized by crime in the past year. While over the last few decades the crime rate has been declining, it is still mildly disturbing that over 1 in 5 households have been recently victimized by crime.

Of course, the best-case situation is to eliminate all crime but it is not realistic. There are things that can be done to reduce crime, aside from the highly debated topic of gun control. Bloomberg compiled 4 things that could be done to reduce violent crime: lead abatement(replacing lead pipes and paint), drug decriminalization, prison education, and community policing. In order to deal with half of these strategies pointed about by Bloomberg, and the hundreds of others that help solve the problem, money needs to be injected into these projects, and crime tends not to be very high on the list for public servants to take care of. Thankfully, nonprofits do exist that target crime in urban areas. Sociologist Patrick Sharkey, in Uneasy Peace, points out that the crime rate decrease that we have seen in the last couple decades are "the result of the combined effort of multiple ordinary people and institutions working at the local level to reduce violence in their own communities. Local nonprofits have been essential to reduced crime rates." These nonprofits runs programs such as education, providing funds

for living area upgrades, and working with at-risk children to keep them from committing crime to begin with.

There is no doubt that crime has been decreasing over the years, but ideally it should be as close to zero as possible, and the key to solving this issue is getting to it on a local level, which this data allows us to do. This data set pertains to crime and socioeconomic and law enforcement data for thousands of cities and towns sample across the United States. Knowing the crime rates for a certain city during a certain year is nice when trying to decide where to live, establish a business, invest more money into local infrastructure, etc. These are all great reasons to come up with the best regression algorithm as a method of predicting the amount of violent and nonviolent crimes based on the area. Particularly, I wanted to frame this issue on what communities should be focused on when dividing up resources. The ultimate goal is to be able to reduce crimes across the nation, but focused on a city by city level, which leads to an explanation of potential clients.

3) The Client

I had briefly mentioned several possible clients that would benefit from this data and its analysis. Not only would this be helpful for someone who is trying to determine the best place to settle down or start a business, one of the best end users of this data are the people who are living in the communities that are mentioned. One of the tactics mentioned above, community policing, has the police getting to know the community on a personal level, rather than just being a service used after crime has been committed. If a city knows to dispatch police officers to a certain area, not to respond to crime, but to just meet the community, the community will feel safer, and criminals are also warier knowing that local law enforcement is looking after them.

The best potential client for these findings would most likely be national non-profits, with offices on the ground in many different cities whose goals it is to provide money or services to in need communities. In addition, cities having this information would be able to know if they need more personnel on the ground in the way of community policing. Having money to bolster law enforcement, aiding those most at risk to commit crime in the future, and otherwise increasing the overall standard of living will decrease crime as a whole in each area. Many nonprofits already exist, ranging from those who provide alternatives to juvenile delinquency, legal aid office and human services organizations that provide help people deal with the experiences, religious and secular groups that give emotional support to families of those that are incarcerated, and inmate rehabilitation groups.

The best approach to figuring out how funds should be distributed is to begin to predict which cities would have the highest crime rate for that time period then make a way down the list. Thus, for current data, I have decided to present the areas that need most work based on the available data set. However, a potential client would either have even more updated and inclusive data that what I was able to find, or also be interested in how to potentially gather the most useful data in the future. Therefore, for the client, I will also provide insights on how to maximize data collection in conjunction with the algorithm.

4) The Dataset

Information on the Dataset:

The dataset that I used for this project was compiled by the UCI Machine Learning Repository, and I decided to use the Communities and Crime Unnormalized Data Set. The dataset combines:

- data from the 1990 census
- law enforcement data from the 1990 Law Enforcement Management and Administrative Statistics Survey, originally investigated by the US Department of Justice and compiled by University of Michigan's ICPSR, and
- data from the 1995 FBI Uniform Crime Report

The dataset mostly contains independent variables, hundreds in fact, most of which had to be eliminated or cleaned up in steps detailed shortly. There were also some categorical data that could not be included during regression but were important for reincorporating back into a user-friendly output. Thankfully, the dataset was already organized in such a way that the independent and dependent variables were easy to separate. However, despite this data being collected and organized well, the data contained within the dataset was far from perfect.

Cleaning and Wrangling:

- Merging

The format of column names needed to be manipulated first before being able to merged. I grabbed the list, placed it in excel, then separated out the data using Excels "Text to Columns" function, which allowed me to pull just the column names, and separate out other extraneous characters including the definitions of said columns, which I know that I'll need when I do the final data analysis and wrap up, but is not necessary for data analysis right now, and therefore was excluded. With a little bit of massaging, I was able to create a CSV of column names, which I then uploaded to the Python notebook and saved as the original dataframe's column names.

- Dealing with Missing Values

I noticed when I was first importing the data that there were missing values in the data which seemed to be denoted by "?"s. After in fact checking that they were strings of question marks, I then wrote a function to count the number of ?s(missing data in this case) to count the number of columns with missing data. I noted after running this function that there were a few data sets that jumped out at me, with large amounts of missing data. One of these was a per capita column, where my intention is to fill that in with a possibly weighted average of other per capitas in that community area using the average of other races in the area to weight them. One other staggering thing I noticed was that of the 2215 rows of information I had, 1872 of them were missing police data, a whopping 84.5% of total police data was denoted as "?". With that much missing data, I did not feel comfortable wanting to keep that in my data set, and instead eliminated the columns entirely. Trying to make predictions based on such a small pool would most likely lead to inaccurate results. I decided to drop the 22 columns that had the vast majority of the data missing, most of which had to do with police census type data, and decided to just predict with the other 100+ columns. I used a manual drop, though I could have used a function to find where it said police or polic, since the spelling used was inconsistent, I decided to do it manually as to not accidentally keep or drop wrong columns.

After looking through the list of what are zeros and missing values in the dataset, it makes sense that the zeros in the columns are truly 0s and are not accidentally missing data. For instance, murders

and different percentages of races make sense, if the city is particularly safe or doesn't have any representation. Therefore, we will leave those, but all of our missing data that we turned into numpy NANs will now need to be filled. I am using the median to fill in due to the large spread in city size, and the 2 or 3 extremely large cities will skew the mean, so using the median of the entire column will make more sense for estimating what the missing data should be filled with.

Although this was already denoted when I first retrieved the data set, it wasn't actually clear what the irrelevant data was going to be until I successfully put the column names back into their associated columns. At that point, it was then I saw that there were going to be identifying variables, which would help in the end stages, but most likely hinder early data analysis. I knew that they would not need to be included in the analysis, but I did not believe they would need to be removed entirely. These columns included the community name, county code, and community code. I ultimately created separate dataframes: one with all the columns intact, and one with just the predictors.

In addition, it was noted in the data gathering portion that the last 18 columns/variables would be the ones to be predicted, but I did see that they already had data in them. This could be useful to check how accurate the algorithm is at the end, but serve no purpose in the early data analysis, nor in the machine learning portion. While I plan on leaving them in the original dataframe for the time being, they will not be used in the final product, other than to compare accuracy.

- Checking Data Values

I began doing this by first grabbing the columns that would be the easiest to check. For me, this was all the columns containing a percentage of some sort, which all began with "pct" which made it very easy pull all relevant columns into a new dataframe. At this point I wanted to make sure that the values were all between 0-100. Anything outside of those boundaries would be incorrect data and would need to be omitted. I wrote a np.where to check where the percentage would've been outside of these boundaries and an empty array was returned to me, so all data within that percentage dataframe was within acceptable boundaries. Will be doing similar things with the other columns.

- Dealing with Inconsistent Data Types

Despite this all coming from one combined dataset that had already been compiled, earlier when I checked data types, I noticed an issue when early on I checked the types of each column. I noticed that a lot of the columns were coded as objects. I notice that the objects columns either fall into one of two categories. The first involves the object columns at the beginning of the dataset. These are just identifying columns which include the state and city the data is of, mostly just identifying data. The other object columns seem to be because of how the missing data is indicated in the dataset, which is using a character, and not a number such as 0 leading the columns to be made of objects. After dealing with the missing data, I converted all remaining object columns to floats so that they could be used correctly in analysis.

5) Exploratory Analysis

A) Visual EDA

For the exploratory data analysis portion, I wanted to see if I could pinpoint certain variables that would be more important in generating the final analysis. It was at this point that I decided that I

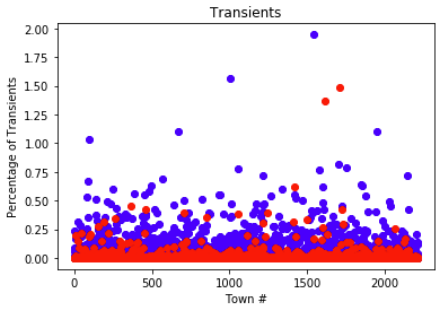
wanted to mainly look at the amount of violent crimes per 100k population, and non-violent crimes per 100k population. In this specific set of EDA, I posed a hypothesis that the financial stability of a certain area would greatly affect the crime rates, and with financial stability, I felt that I needed to look at not only poverty levels and incomes of certain people, but also the living situation as well.

In the first plot, I wanted to see what sizes the towns/cities we were going to deal with in case it affected analysis later on. I also wanted to make sure there were reasonable group sizes so I have split up the towns on population. Those with less than 25,000 residents are in one group, then 25,000-50,000, then 50,000-100,000, 100-500,000, and anything larger were all grouped separately. Later on during analysis, we can see if this is significant or not, but I wanted to have them grouped out in case.

In the follow plot, I wanted to see was the percentage of homeless people and those in shelters. This statistic was not included in the original data but I was able to calculate it relatively easily. Those who are able to stay shelters are more hidden from view of others and may not be subject as much to both violent and non-violent types of crime or commit as frequently.

```
In [22]: pctShelter = (df2.loc[:,'persEmergShelt']/df2.loc[:,'pop'])*100
         pctHomeless = (df2.loc[:,'persHomeless']/df2.loc[:,'pop'])*100
         plt.plot(pctShelter,marker='o',linewidth=0,color='blue')
         plt.plot(pctHomeless,marker='o',linewidth=0,color='red')
         plt.title("Transients")
         plt.xlabel("Town #")
         plt.ylabel("Percentage of Transients")
         #plotting the percentage of people in shelter and homeless in all the cities

Out[22]: Text(0,0.5,'Percentage of Transients')
```

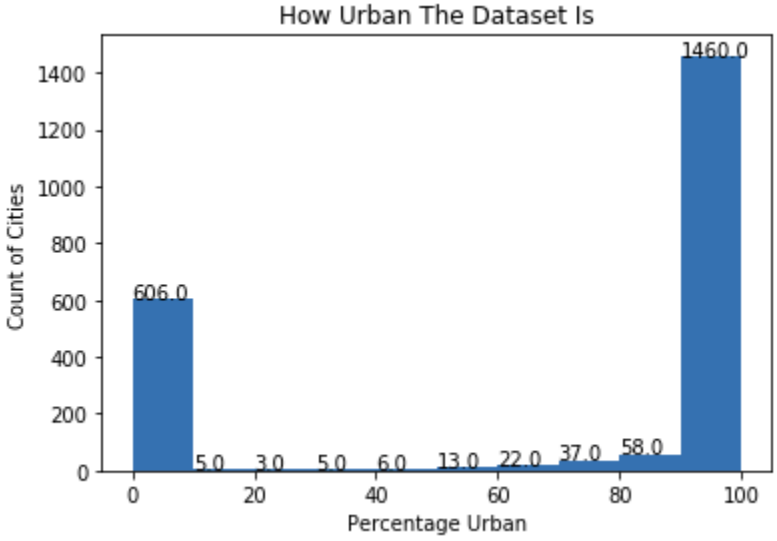


Immediately when looking at the graph we can see that the larges # of those living in shelters seems to be around 2% which is good since it means there isn't a high homeless pattern in any of the towns, leaving one less variable to worry about, since a significantly large homeless population when compared to other cities, may affect crime rates more than expected. I expected there to possibly include more higher amounts of homeless to create a better training set. However, even in New York City, the percentage seems to be extremely low. Generally, the percentage of those on the streets are lower than those in shelters, meaning the city has room to accommodate most of the people in the shelters, which could mean that since the city has enough money to combat the homeless situation, it would typically have a higher presence of officers present, leading to less crime, or a higher socio-economic level in the community, also possibly leading to lower crime rates.

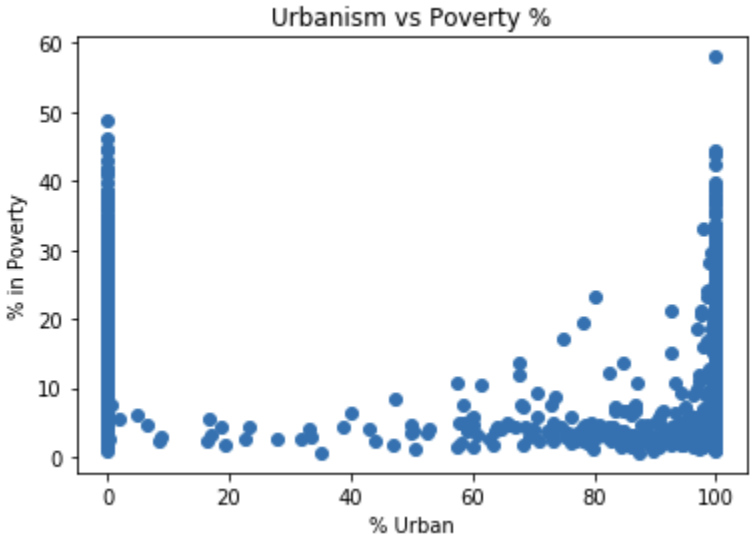
The next plot is them plotted against each other. Specifically, I wanted to do this to find outliers in the ratio of those living in the two separate places. These two types of cities would be particularly interesting to look at during the analysis: those with a higher percentage of those in shelters and

extremely low on the streets, as well as an extremely low percentage in shelters and a higher percentage on the street. The former would usually mean a higher QOL and lower crime rates while the latter could mean higher crime rates.

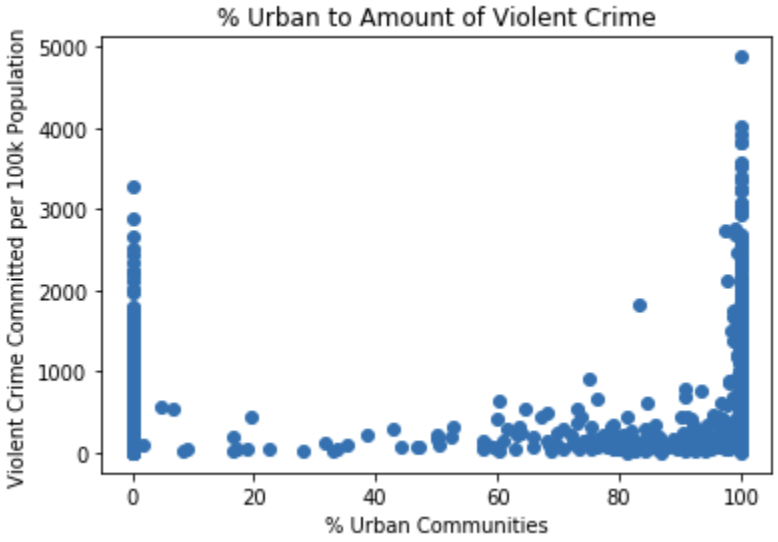
In the following plot, I wanted to see the distribution of cities in the data, and it seems like there is a rather unequal amount of urban to rural communities. There was not a baseline given for what was considered urban or not, but perhaps the best data set would have had a relatively equal number of urban and rural communities and a small but substantial number of communities within the two indicating varying levels of suburbia.



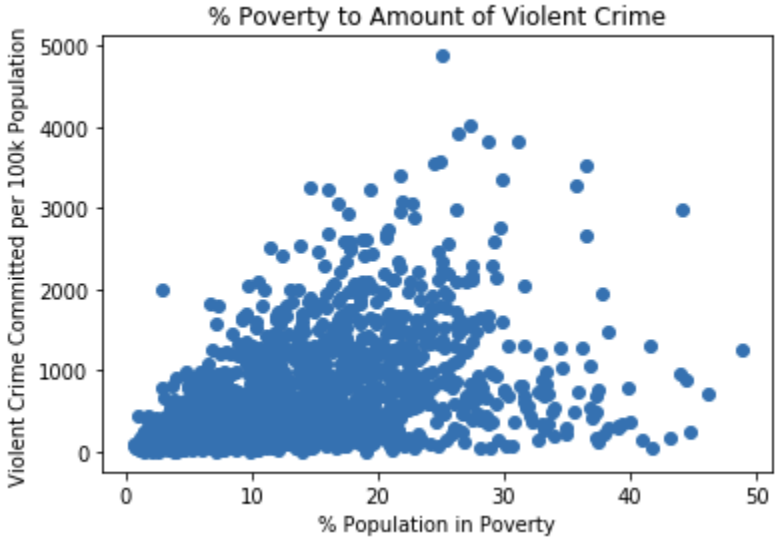
I also plotted urbanism to poverty, but found that it was not as directly correlated as I thought it would be. There was surprisingly a very large amount of percentage of those in poverty that were in completely rural areas. I would have expected very low poverty levels at the rural level since when people think of poverty, they think about dense urban communities, not rural communities.



To truly get the best idea of how, if at all violent crime was affected by these statistics, I graphed them to violent crime and came up with some interesting plots.



The first plot is how urban the dataset is plotted with the amount of violent crime. While this graph on its own does not create any discernable pattern, when compared to the graph above it, urbanism to the poverty %, creates a nearly identical graph. To confirm, I then plotted the poverty percentages of each city the amount of violent crime, and sure enough, there seems to be a relatively strong, somewhat linear correlation between poverty and crime. I am not expecting this graph to clearly show an exact correlation due to poverty not being the sole determining factor of crime rate, yet the ability to show at least some correlation over other plots does indicate that this is a deciding factor.

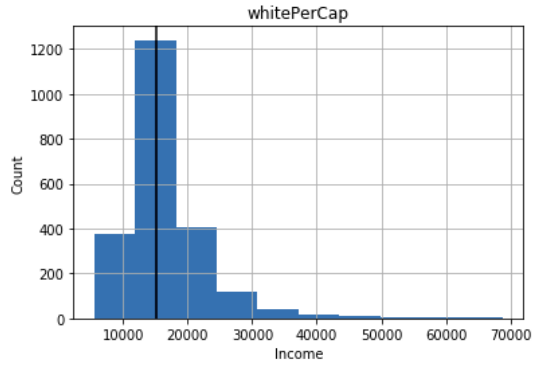


In addition, I plotted two things I found interesting, that weren't directly related to the crime rate, however, I thought was important to note, because I think that it could be a motivating factor when

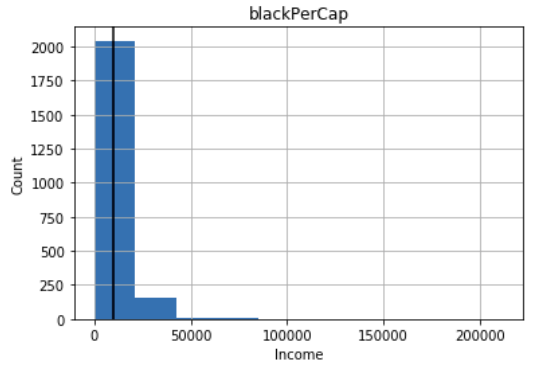
people think that they need to commit crime, and that is the comparison between the unemployment rate and the poverty level. When comparing the two percentages, in a perfect world, only those unemployed would be living in poverty level and we should find that the levels are equal. The fact that so many red points occur above the bulk of the blue points shows a very important point, that people are employed and yet still living in poverty. This can be a contributing factor particularly to non-violent crime such as theft, in order to make sure that their own families have enough to sustain themselves, since just employment is not enough to do so.

While plotting the above, I figured that races would play an important part in how much they were earning and what poverty levels would look like, and ultimately playing a part in how much crime was committed. For example:

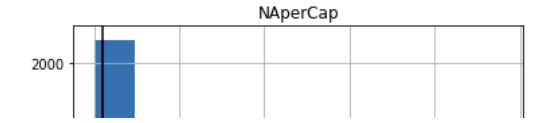
```
The median income for whitePerCap is:15073.0
```




```
The median income for blackPerCap is:9777.0
```




```
The median income for NAperCap is:9895.0
```



While the race histograms do not seem to indicate much relating to crime at the moment, this information came into play later on when finding out individual correlation coefficients to the violent and nonviolent crime rates when I found specific races had stronger positive and/or negative values based on raise. While I refrain from making conclusions of crimes happening due to certain races, the individual race correlation coefficients each play a small role in the final regressions.

B) Statistical Analysis

In addition to the usual exploratory data analysis, I tried to do some statistical analysis as well. In this section I did three things:

1) Trying to figure out whether my two predicted variables would be predicted as related or independent, and seeing if they were similar enough to each other
2) How the dataset compares to the population mean taken from Statista from that year, and
3) The variables' correlation coefficients, the Pearson R's, to my two predicted variables.

The first two T-tests were run on the two variables being predicted to see whether they are related enough to each other to be considered equal. The first test is a two-sided test for the null hypothesis that 2 related or repeated samples have identical average expected values. Although it is general common sense that the violent and nonviolent crime averages would not be the same, it would be later on predicted by the same variables, so statistical confirmation is always useful. When running with the null hypothesis that violent and nonviolent crime is the same, but related, we end up with a large t-statistic and a p-value of zero, leading to a rejection for the null hypothesis. When running the paired t=test as independent variables with equal average, we once again get a p-value of zero, indicating that even when looked at independently the two means are not equal. Although the variables used to predict these values will be the same, the ultimate results are different enough to warrant running them as separate algorithms, since the average differences between the two are statistically significant enough to warrant us to do so.

Then 2 single sample T-tests were run based on means provided by a third-party data aggregator and was converted to violent and nonviolent crime amounts per 100,000 people. The tests are comparing whether the predicted variables are statistically significantly different to those provided by Statista. As the results indicate, it is different enough to reject the null hypothesis that the sample is equal to the mean. However, it is interesting to note that despite the p-values being low enough to reject that the t-statistics are low, meaning that it is not obscenely far from the expected values compared to other possible data. The differences may be contributed to difference in a variety of factors: source data, how nonviolent crime is defined, which I found changed based on source (though I tried to eliminate this as much as possible by trying to match crime to crim in my data set and their definition) as well as my dataset not being the 100% true population representation of the nation despite it having samples from all over the nation. Despite having a diverse sample pool, I think that the data set not being completely representative of the overall nation to be the most likely explanation for these differences. Since the t-values are relatively low, despite it being statistically significantly different enough to not be completely equal to the population mean, I believe that it is close enough to get a general idea of the population.

Finally, the correlation coefficients to each of the predicted variables was calculated. To prepare the dataset, I first had to clean up the dataframe a little bit further than how I had it before. The identifying factors of each variable: the city name, state, community code, and others cannot be reasonably compared, both due to the community code not being able to be matched to others, and the inability to run strings through the correlation function. Therefore, after cleaning up the data further to use only variables with float, which I was able to convert all others to at the beginning of my cleanup

stage, I was able to run correlations and sort from lowest to highest R-value. However, this on its own isn't representative of the most and least correlated, since the least correlated variable would have an R-value of zero. Since I wanted to demonstrate the most influential variables on the final crime rate, I decided to take the first 10 and last 10 variables, excluding where it was correlated with itself, which would have produced a coefficient of 1.0, and compiled them together.

For violent crime the most correlated values were as follows:

	R-values
pctKids2Par	-0.684059
pct2Par	-0.649762
pctWhite	-0.647164
pct12-17w2Par	-0.616963
pctKids-4w2Par	-0.611884
pctWdiv	-0.534563
pctPersOwnOccup	-0.494272
pctHousOwnerOccup	-0.443022
medFamIncome	-0.394124
medIncome	-0.380866
rapesPerPop	0.583146
pctBlack	0.590580
autoTheftPerPop	0.600200
nonViolPerPop	0.627124
murdPerPop	0.635940
burglPerPop	0.676466
pctKidsBornNevrMarr	0.686665
robbbPerPop	0.789590
assaultPerPop	0.852080

And for nonviolent:

	R-values
pctKids2Par	-0.653018
pct2Par	-0.645484
pct12-17w2Par	-0.608651
pctKids-4w2Par	-0.598805
pctPersOwnOccup	-0.495958
pctWdiv	-0.482246
pctWhite	-0.475897
medIncome	-0.462785
pctHousOwnerOccup	-0.457087
medFamIncome	-0.453612
assaultPerPop	0.553505
pctMaleDivorc	0.576979
autoTheftPerPop	0.585534
pctFemDivorc	0.589553
pctAllDivorc	0.596741
robbbPerPop	0.614391
violentPerPop	0.627124
burglPerPop	0.793696
larcPerPop	0.926313

There were some interesting things that popped out to me when I saw these variables that were the most correlated. Looking back to the definition of what constituted violent crime, we have murder, rape, robbery, and assault. For non-violent crime, burglaries, larcenies, auto thefts and arsons were used. It would make clear and total sense that for violent crime the most highly correlated variables should be the crimes that the definition is consistent of. However, among those 4 crimes, we see "pctKidsBornNevrMarr" which as the column title would suggest is percentage of children who were born to parents that were never married. The nearly 0.7 correlation amount shows a pretty strong correlation to those born to parents who were never married and committing violent crime. Where as in the clear opposite direction, to the strong negative correlation, those who were born to 2 parents who were married were less likely to commit or be involved in violent crime. Also, despite them not being part of the definition, plenty of nonviolent crimes made high correlation on the violent crime list, making it seem that in general, for each area, if there was high non-violent crime, there would likely be high violent crime as well. This, in essence, held true in reverse for the nonviolent most correlated variables as well, where some violent crime made for high correlations due to areas being high crime overall regardless of whether it was violent or not. For non-violent crime, the predictive variables that surprised me the most was how much divorce in an area seemed to affect the crime rate. There was not

as high of a correlation in violent crime, but high percentages of divorcees in an area, seemed to be relatively correlated to non-violent crime.

These variable correlations were not factored too much into the regressions that follow. However, knowing these variables is important both in the decision-making process for the client, and could also possibly affect how data is gathered in the future. This will be elaborated more upon in a later section.

6) Regressions Discussion & Results

I decided to take a look at 3 different types of regression here to see which one would best fit the data and create the most accurate predictions. The goal here is to be able to pick the best algorithm to use, along with new updated data, and generate on-demand results of what neighborhoods to focus on. In this project, I compare 3 different types of regression: Linear Regression, Ridge Regression, and Random Forest Regression.

A) Linear Regression

The first regression I attempted was a Linear Regression. It made sense to use this as a starting block for regression. I would make the sweeping generalization that these variables were all linearly correlated to the variable being predicted, although this is most likely not true, it is a good starting point. Since there are over a hundred independent variables, we use a multiple linear regression, where separate small, yet not insignificant slopes and intercept make up the final linear equation. To explain what I did briefly: I removed all non-predictive variables from the data set, then used all of those remaining variables to predict my two dependent variables, violent crime per 100k population and non-violent crime per 100k population. I split the dataset into a 70/30 split of training data and testing data, then fit the model using the training data, and made predictions with the test data. Since crime variables already existed in the form of actual crime rates, these would be used as benchmarks for how accurate our model was. I printed the correlation coefficient overall, which we are trying to maximize as well as the root mean squared error (RMSE), which we are trying to minimize. The best regression would have the highest r^2 and the lowest RMSE. These two things alone could be used to judge how well the regression id, however, I wanted to make it immediately obvious to someone that wasn't into code how accurate the model was, so in addition to r^2 and RMSE, I included the first few entries of the test dataset and the predicted values the model came up with, as well as the city it came from so it becomes immediately obvious how well the prediction worked for each city. In addition, I provided a 10 fold cross validation to see if we could improve the scores from that beyond just the test. I expect this to be slightly higher since I am incorporating the entire dataset rather than just the test values, however, the number of folds averaged out may not overall be higher.

When applying this technique to violent crime, we get an r^2 value of ~0.54 and an RMSE of roughly 392.17. The 10 fold cross validation score was 0.532. I applied this same technique to the nonviolent crime, however, the results were not quite what I expected. I got an r^2 value of 0.47, lower than that of violent crime, and an RMSE of 2139.87. The massively large RMSE value seems to show me that this is not a good way to predict non-violent crime. Other regressions may or may not improve on this. In general, I believe that these algorithms seem to be much better suited to predict violent crime over non-violent crime. However, our cross validation scores did seem to mildly improve the scores.

B) Ridge Regression

After linear regression, I attempted to apply ridge regression. Prior to actually doing the calculations, I predicted that the ridge regression would do better than the linear regression due to how it deals with variables that are correlated to each other. Although in the statistical analysis, I only calculated the correlation to the dependent variables, I assume there was at least sometimes, instance of high correlation between the independent variables and ridge regression accounted for this. In addition, ridge regression having a penalty term allowed some coefficients to be reduced to zero, therefore the algorithm does the work of narrowing down our many variables, and using ridge also prevents overfitting due to collinearity that our linear regression would not be able to adjust for. After following a similar code structure to that used in linear regression I was able to get an r^2 value and an RMSE to compare with the other regressions. Instead of a single cross validation score as used in other regressions, I used a plot of the 10-fold cross validation scores over different alphas and plotted to show a moving average.

For violent crime, I was able to achieve an r^2 of 0.544 and an RMSE of 391. While these are not higher than those by linear regression, they were achieved by using the default alpha of ridge which is 1, whereas the plot later on across multiple alphas shows that the best alpha to maximize cross validation scores is not 1. When replaced and rerun with a number closer to the best alpha value, we see a slight increase in r^2 and RMSE. For nonviolent crime, an r^2 of 0.472 and an RMSE of 2138 was achieved, showing that ridge regression was ever so slightly better at prediction than linear but not by much. When looking across alphas and 10-fold cross validation scores, I saw an average of 0.515 regardless of alpha. This is slightly better the cross-validation scores for linear regression. So, although there seemed to be very little if any difference for violent crime prediction between the two types of regression, there was improvement in nonviolent crime prediction.

C) Random Forest Regression

Last, but certainly not least, I used random forest regression to try to predict. The random forest model is usually better than all other models since it makes predictions by combining decisions from an aggregate of decisions trees where all the base model trees were constructed using a different subset of data. This turned out to be better than linear regression, at least since random forest regressions would capture non-linear interactions between the dependent and independent variables. There is also enough data here to generate good tress, since random forest inherently does not do well with sparse data, but with our given dataset, this probably would have been the best regression to use regardless of outcome. Once again, I followed a similar pattern as before, splitting the data into training and testing, creating the regression, fitting, predicting, then calculating scores.

For violent crime our r^2 was much higher than before, 0.593 and RMSE lower than all other regressions, 370.27. When running our 10-fold cross validation score, we achieved a score of 0.532. With the highest r^2 of the bunch and the lowest RMSE of the regressions I compared, this was the best of them, and what I would use to try to predict violent crime. When run again for nonviolent crime, an r^2 of 0.429 was achieved, with an RMSE of 2222.28. For non-violent crime, random forest was not the best regression. However, it does not seem any of the regressions did a particularly good job of predicting for non-violent crime. However, for violent crime, all algorithms did a better job of predicting overall, with the random forest regression doing the best.

7) Recommendation and Reflection

To choose what the best regression is to recommend to the client, I looked at two criteria for the each of them, the r-squared value that each regression decided to come up with, and the root mean squared error value. The r-square differ slightly for each type of regression because the type of regressor affects how well the data might be fit. In addition, we are looking for the lowest RMSE value, which would be closest to the best fit line. In my situation, the Random Forest Regressor did the best job, and I would recommend using this regression to determine which communities to look at.

In addition, I would recommend that this algorithm only be used to assess violent crime in a certain area, and not non-violent crime. When looking at the RMSEs for non-violent crime, the differences are far too large to be able to say that is a good predictor. In additional work on this project at a later time, I would like to try to find the best regression for non-violent crimes, but I do not feel comfortable saying that any of these 3 are good benchmarks.

Those with STEM backgrounds may be alarmed at the relatively low r-squared values, but these seem reasonable given that human behavior is difficult to predict, and the dataset has an extremely large number of independent variables to consider. To make the model even more accurate, I would attempt to narrow down the independent variables, or eliminate them by the data collection phase, and that would generate a more accurate model.

For just easy visual purposes, I just grabbed the first 5 cities that came up and put them side by side with the predicted values for violent crime in those cities in the order of: linear regression, ridge regression, then random forest. As you can see, the 3rd type of regression I performed, the Random Forest Regression, did the best job of predicting, though far from perfect. This is due to imperfections in the data, where I had to put the median of a variable to fill in missing data in the columns, as well as hundreds of independent variables which all affected the outcome in a small shape or form.

```
         Cities   Actual  Predicted
0       Turlock   847.77     577.00
1     Bridgeton  2241.85    1875.08
2  Harrisonburg   151.78     311.92
3       Hayward  1118.18     925.06
4    Waynesboro   337.96     627.34
         Cities   Actual  Predicted
0       Turlock   847.77     571.65
1     Bridgeton  2241.85    1863.30
2  Harrisonburg   151.78     303.16
3       Hayward  1118.18     916.16
4    Waynesboro   337.96     616.25
         Cities   Actual  Predicted
0       Turlock   847.77     673.53
1     Bridgeton  2241.85    1781.98
2  Harrisonburg   151.78     394.59
3       Hayward  1118.18     859.89
4    Waynesboro   337.96     460.66
```

At this point, we have pinpointed the best regression to use, and possibly even, what data to use, but the question of "What do I do with this?" still remains to be answered. I believe that a company looking to invest, either in improving city infrastructure to reduce crime or increasing law enforcement presence, would benefit from looking at the list of cities where the most violent crime in that year was committed, and use it as a list of where needs the most work first. Of course, size considerations must be taken into place as well. This is in no means meant to be a perfect representation of what crime looks like in certain cities. However, I believe this to be a good starting point and building block towards reducing crime across the nation. Below, I was able to compile the top cities for crime sorted by the highest actual violent crime rate, to the right, a direct comparison to those sorted by the predicted violent crime rate.

	Cities	Population	Actual	Predicted
0	Atlanta	394017	4026.59	2303.77
1	AtlanticCity	37986	3583.48	2619.88
2	Richmond	203056	3239.20	1812.77
3	SanBernardino	164164	3047.66	1419.41
4	BatonRouge	219531	2978.69	1455.40
5	Lynwood	61945	2956.98	1355.16
6	Oakland	372242	2601.60	2010.66
7	Commerce	12135	2584.96	1045.63
8	Hawthorne	71349	2541.38	1606.93
9	Mansfield	50627	2493.17	878.04
10	KansasCity	435146	2460.11	1412.48
11	Grenada	10864	2333.62	1455.05
12	NorthCharleston	70218	2264.09	1767.12
13	Bridgeton	17779	2241.85	1781.98
14	Salisbury	20592	2209.01	1834.24
15	NewHaven	130474	2127.02	1980.26
16	Lawrence	70207	2124.14	1798.63
17	LakeCity	10005	2119.93	1493.09
18	Brunswick	16433	2089.44	1410.75
19	Inglewood	109602	2078.85	1878.21
20	RockHill	41643	2047.92	1196.12

	Cities	Population	Actual	Predicted
1	CollegePark	20457	1874.95	2544.325161
2	AtlanticCity	37986	3583.48	2441.849855
3	Atlanta	394017	4026.59	2199.101243
4	Burton	27617	374.06	2170.767103
5	Inglewood	109602	2078.85	2015.093299
6	Pleasantville	16027	1730.67	1994.936979
7	WarrensvilleHeights	15745	929.91	1936.268545
8	Richmond	203056	1617.17	1877.331541
9	Cambridge	11514	1514.12	1765.474212
10	NewHaven	130474	2127.02	1760.288557
11	Americus	16512	607.80	1753.027037
12	Rochester	70745	1171.84	1740.556593
13	Oakland	372242	2601.60	1728.379966
14	Brunswick	16433	2089.44	1708.134319
15	Bridgeton	17779	2241.85	1699.005901
16	Coatesville	11038	1054.23	1663.377074
17	Berkeley	102724	851.23	1602.324167
18	Albany	29462	1363.32	1601.026643
19	Lynwood	61945	2956.98	1592.867008
20	Hawthorne	71349	2541.38	1577.037532

For the most part, these lists share quite a few cities. The differences between actual and predicted values, may be due to the way I filled in the missing data with the median of the column that it was missing. I chose not to do mean as to skew values towards larger or smaller cities, and thus took a hit in the accuracy level of the algorithm in favor of having more reasonable results overall.

To improve the accuracy of the algorithm, I would first work with the dataset. There was a lot of missing and possibly inaccurate data, and so while I did my best to work with the missing data, there was no way to fill it without compromising the accuracy. Working with the mean would not work as the extremely small or extremely large cities would affect the value. The median, which is what I chose to work with, was not the best option either, as if it tried to fill in a large city with a median value, or a smaller city with the median, the values would not match up, and this is seen clearly when looking at the actual violent crime value next to the predicted. I am not sure what the best solution to this, but my

guess would either somehow be a moving average based on the city size that I had done during visual analysis, or medians based on sectioned out portions of the dataset, once again, most likely done based on city population.

In addition, if we take a look back at our statistical analysis, we started working on determining the most correlated variables to each of the dependent variables. If we are able to pinpoint the variables that are most effective at determining the final dependent variable, the client would then be able to focus on gathering and maintaining the highest quality dataset, and also simultaneously making the most accurate algorithm. I believe that a lot of the error that came as a result of the prediction of violent crime using this dataset, was due to inconsistencies with data, i.e. missing data.

The dataset itself notes that using this data alone to evaluate communities is a bit over simplistic since a lot of data is missing that has a lasting effect on crime rate, such as more visitors in a city than one doesn't have many visitors come in since the crim rate is based on per capita population and visitors would not count towards the resident population but could count towards the amount of crime that occurs. Therefore, to truly have a more thorough look at assessing more at-risk communities, we would need additional corroborating data. It would be impossible to determine all of the variables that would affect the ultimate data, but certainly additional community data would be useful, and entirely beneficial as it would help make the prediction algorithm even stronger.

I understand that the dataset is relatively old at this point, and of course the best-case scenario is to have the most updated data possible, so a massive improvement would to be to incorporate all new data as well. The most current data would be to be able to gather the newest census data at the time, and gather crime data by the year from the FBI, but law enforcement data would be much harder to gather. Testing would show whether or not the law enforcement data is important enough to absolutely necessitate when data gathering. However, my goal with using the existing data was to determine the best algorithm to use. Although rates and data within the communities may change over time, the contributing factors should not, thus all that would need to be changed is running the updated years' data into the working algorithm.

8) Wrap Up/TLDR

1) As one can see for the most part, the list of the top 20 cities is roughly the same between the actual crime rates and the predicted. Therefore, I believe this is the best regression to use to find which cities need the most work.
2) Using the most up to date data with this algorithm would generate the most up to date list of predicted violent crime rate for a certain area, leading to the best decision making by cities themselves, or 3rd party companies to come in and offer opportunities to provide a safer living place for its citizens.
3) None of the regressions I compared did a particularly good job of predicting nonviolent crime, so I would refrain from using any of them to predict, but of the three, ridge regression actually performed the best for nonviolent crime.
4) When going and working on collecting new data, the company/city has the opportunity to look at what is most important when determining the predicted crime rate, and ensure that data is

the most accurate and complete before using the algorithm, to ensure the most accurate output.

5) Given the data that I was, and assuming that while the data itself may change, the causes of people committing crime don't, and thus despite changes in the data, the algorithm should still hold accurate, and Random Forest remains the best regression to use.
6) However, once again, this is in no means meant to serve as an accurate predictor for when and how much crime will happen, only to offer guidelines to those seeking to make cities safer.