Keenan Venuti
Final Paper GOV-1430
Professor Sweeney

The Crimes that Divide Us

Abstract

Intro, Background, and Methods:

By almost all measures, crime in America has been steadily decreasing for the last 25 years [1]. Most researchers and civilians agree that we are living in a safer America, in terms of civilian crime. Societally, we are at the pinnacle of conflict resolution and violence avoidance. However, geographically, is crime reducing in every state evenly? Despite the downwards shift in violent and serious criminal offenses, we still view some parts of America as more dangerous than others. Fair or not, I aim to investigate if regionally based claims of differences in crime rates have credibility. It is vital to note the preexisting and overt political, socio-economic, and cultural differences in the U.S.A; most standards show we are living in one of the most divided Americas to date. In two studies, I confirmed that based on crime rates, the states of America show significant divides. In my first study, I used metadata of crime across every state, isolated by crime type, to cluster states via an unsupervised learning algorithm. In a follow up study, I analyzed four specific cities and compared crime dispersion, crime rates per-capita, and the types of crime committed.

Results and Discussion:

In my statewide study, I found significant clusters indicating that the states of America can be divided into two to three separate groups differentiated by crime rate. This compliments a sundry of preexisting research indicating the real and potent divides in America. In my citywide study, I found key differences in crime rates and crime types between and within cities. When cross-comparing with outside research, showing that crime occurs more often in cities, I can show which cities have outlier proportions of crime (more/less crime than their residing state). With these new measurements, researchers can implement further studies to understand the ontology of this divide as we as a nation try to come together.

Introduction

Divide in America:

If America feels more divided than ever, it's because we are. In 2017 we have grown used to vast differences of opinions across the nation. A most recent example is the 2016 presidential election. Polls and experts were confident that they correctly predicted America's decision to elect Hillary Clinton as president but Donald Trump became the next leader of America. It's not necessarily true that polling techniques and statistical analysis has failed us. Instead, the divides in America have become so ingrained in society, we often can't see when a

new one opens. Politically, we are divided as ever. A study performed by Pew Research found that republicans and democrats are more ideologically divided than in the past [2]. Although opinions on how to run the government have always fallen into two main parties with a few tertiary outlets, there seems to be a growing divide between mainstream political parties. This applies to social issues but the study included things like social welfare and personal political opinions. Beyond the cleaved political system, we can see division in other highly important and broad attributes of America. Research found that people don't even realize how huge socio-economic disparity is in America. One study found that people believe the richest 20% of Americans own 59% of wealth and the bottom 40% own 9% [3]. In reality, the top 20% own 84% of wealth whereas the bottom 40% combine to own .3% of total wealth. Clearly these numbers show intense and damning socio-economic inequality, but it may be more pejorative that we don't recognize our differences.

Socially, America is a melting pot comprised and built by people from many different ethnic backgrounds. Despite the prevailing existence of racism, America as a whole is diverse. Still, we often discuss cultural differences amongst our citizens based on family heritage and regional location. When analyzing people from different regions of America, research confirms these cultural differences. Unsurprisingly, large cultural differences have existed long before gross amounts of class imbalance, political divides, and socio-economic inequality rose in the post 2000 era. Many have theorized as to the origin of these cultural rifts. One study claims that vastly different forms of economic structure, dissimilar legislation, and externally imbedded factors account for some of the divides in America [4]. For example, if someone wanted an abortion in a state like Oklahoma, they are mandated to undergo intense counseling to dissuade them from their decision; conversely, in New York, it is a much simpler process with emotional support available. Some states even list abortion as a crime in the datasets I used for research.

Finding the ontology of these national differences is much more difficult than seeing the actual divide. One study insinuated that things like climate can influence culture and personality [5]. However, this type of research is largely correlational and does little to boil out the true reasons of a growing divide. Cultural differences have always existed, but with growing wealth and political division, we need to be careful while planning our country's future leadership and means of communication. I chose to study crime in America as it represents a non-traditional quasi-pulse for a city. It can measure a city or state's economic and emotional health and compliments preexisting research. Understanding alternative metrics showing our country's division will be vital as external divides continue to grow larger. From there, researchers must plan studies to see the causes of this division and analyze how these metrics may influence each other.

Hypotheses:

I performed two studies to understand geographic differences in crime across both states and cities. In my first study, I analyzed the divide in America using criminal report data. Focusing on Northern, Western, and Southern states, I predicted that Northern and Western states will have significantly less violent and theft related crime than Southern States. I based this prediction on cultural psychology research but would like to objectively analyze the states to see if my assumption is fair. In my second study, after getting an idea of how the American map

divides based on criminal data on the state level, I dove into city wide datasets to compare cities in different geographical areas, noting crime dispersion within and between each city. I predicted that all major cities will have a fair amount of crime, regardless of the safety rating of their residing state. However, I also predicted that the kind of crime that occurs will be different city to city, implying that the external factors that uniquely influence each city also influence crime.

Background

Previous Research:

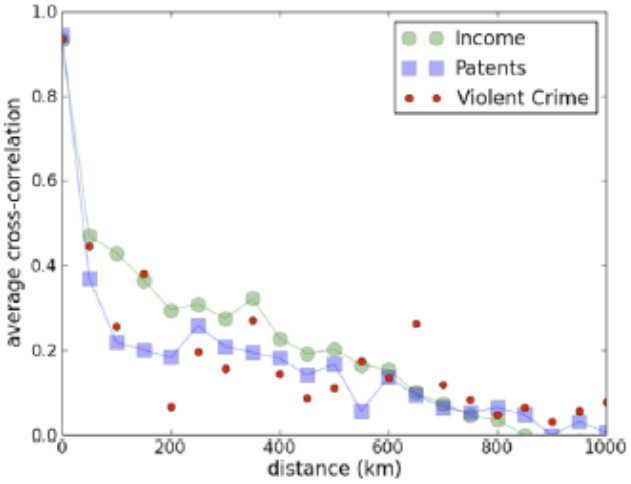
While there are some studies on differences in crime rates across America, I focused my research into possible ontological explanations. I also examined other types of American divides discussed in the abstract. My main interest lies in whether violence is treated differently in different states. Numerous cultural psychology studies have been executed to further understand cultural differences in individuals based on where they grew up. One classic behavioral study, performed by Richard Nisbett and Dov Cohen, found that when bumped physically by a confederate, Southern students showed more aggressive tendencies [6]. For example, after getting bumped, Southern students felt their masculinity was threatened, showed higher levels of cortisol implying they were more stressed or angry, were more primed for aggression measured via higher levels of testosterone, and were more likely to engage in aggressive behavior. However, another study, performed by Dov Cohen and a different research team, found that there are subtle differences in how Northerners and Southerners deal with micro-aggressions [7]. Researchers concluded that Southern politeness norms may promote violence; they found that Southern students show little reaction to an annoying confederate but have an aggressive outburst later on, whereas Northern students consistently show micro-aggressive reactions without an outburst. These studies yield fascinating conclusions, but how can we interpret why we observe this behavior? A similar study found possible reasons for Southern anger in these situations. Namely, researchers found that Southerners overestimate aggressiveness of peers and the encouragement of aggressive actions by peers but are not more likely to encourage aggressive behavior themselves [8]. This would imply that people's actions are a result of a reaction that was original based in a biased prior perception.

Hypothesis Support:

State Level

Despite the vast amounts of cultural data, they can only really imply correlation. Even if a causation exists, it is nearly impossible to understand why. Further, one study found weaknesses in the use of subjective self-reporting measures like the Likert scale, which was used in some of the previously mentioned studies [9]. For this reason, I will use objective analysis based on data harvested from government affiliated websites. The psychology research discussed earlier is not all-telling, especially within the domain of my study, but it implies I will be able to isolate some states as being more violent or crime ridden than other states.

City Level



Regarding my second study, focused on a city level, I predicted crime will be highly focused within all cities. A report created by Plos One found that as cities begin to develop, rises in innovation and wealth are followed by rises in crime [10]. This means positive metrics like innovation increase and decrease at proportional rates to negative metrics like crime. This correlation also shows a decrease in all metrics in communities further away from a city. The adjacent figure (created by Plos One) shows decreases in income, patents, and violent crime the further away from a city you are [10]. Based on this research, I anticipate there will be significant crime in all major cities analyzed as development can't out-grow crime. However, I also predict there will be vast differences in the kinds of crime that occur between cities.

Methods

Study 1

Data Collection:

In my first study, focused on statewide differences, I used metadata to make my analyses. I accessed the datasets from The Uniform Crime Reporting Statistics hosted by the U.S. Department of Justice [11]. The data are comprised of violent crimes and property crimes across all states accumulated from the years 1960 to 2014. Violent crime includes: murder, legacy rape, revised rape, robbery, and aggravated assault. Property crime includes: burglary, larceny-theft, and motor vehicle theft. I was able to accumulate information on crimes per-capita for all 50 states and the District of Colombia.

Data Cleaning:

The UCR has a simple interface to download CSVs with relevant data of your choosing. However, the data came disorganized and requiring cleaning. In excel, I isolated the dataset to contain each crime rate across a 10 year period, 2004-2014. Relevant files are included in the datasets folder.

Data Analysis:

I transferred the datasets into R Studio, an interface to use the statistical analysis language R. I isolated the dataset yearly and looked for trends across states. My data consisted of

rates of each the aforementioned crimes. These data are perfect for an unsupervised learning approach as there is no class or measure; we simply have information on different states and want to classify them using our own categories. States can be classified by interpreting the differences between them based on the criminal data. For example, if two states have a large amount of burglary but a small number of murders, there is a relatively small distance between them. I chose two very different unsupervised clustering algorithms, K-means and Agglomerative clustering, to better understand the data and correctly cluster it.

K-means clustering is a standard and staple clustering algorithm because of its robustness and simplicity. It works to partition N observations into K clusters (the data scientist choses K beforehand). The algorithm begins by choosing K observations at random and randomly places them into individual clusters. From there, it begins analyzing the distance between every non-clustered observation and the preexisting clusters. Each observation is assigned the nearest cluster until all data have been clustered. From there, the algorithm recalculates the mean of each cluster and iterates through the dataset, again assigning each observation to the nearest cluster. It repeats this process iteratively until every data point is in the appropriate cluster.

Hierarchical agglomerative clustering is simple in theory but allows for more user input. It begins by creating a cluster for each observation. It then looks for two clusters with the smallest distance (they are the most similar). It groups those two clusters together, regardless of the size of each cluster. It does this process until all clusters have been grouped into one cluster containing all observations. From there, a data scientist picks how many clusters, K, are appropriate, using specific tools. This methodology is better at picking up minutia detail between observations.

In both of these clustering algorithms, one most chose the number of clusters. This may seem subjective, however, there are two main ways to pick the best K value. I used the GAP statistic and the Silhouette Score. The GAP statistic takes a measure of within-cluster dispersion (how different are the observations within each cluster) and compares it to a null distribution (a clustering where observations are clustered randomly). The Silhouette Score is a very simple calculation done on each observation. It takes each point and sees if it is more similar to the points in its assigned cluster or more similar to points in different clusters. If it is more similar to its own cluster it returns a positive value.

With a full toolbox, I began to analyze the data. To my surprise, after some initial analysis using individual years and all crime rates in the datasets, my algorithms recommend I use one cluster. This means that the data were not significantly separable when using all different types of crimes as a predictive feature set. However, when isolating violent crimes and theft-related crimes, both the GAP statistic and Silhouette Score recommended multiple clusters. From there, I was able to cluster states based on their similarities using a separate violent and theft related feature set. The optimal number of clusters was typically two or three; using fewer clusters produced less error and showed statewide differences. With proper analysis completed on state differences, I focused on individual cities.

Study 2

Data Collection:

City specific data were accessed from the Police Data Initiative [12]. This repository has specialized datasets specific to cities across America. Each entry in each dataset is a categorized arrest. Each dataset represented a different precinct across years usually containing data-ranges from 2000 to 2015. Some precincts divide their arrest data across multiple sets while others use one large file. All datasets are extremely intricate, complicated, and documented/organized very differently. I chose four cities in very different parts of America: Seattle, Washington, Los Angeles, California, New Orleans, Louisiana, and Hartford, Connecticut. Each city is large enough to hold a sizeable population compared to their state. I downloaded each of the datasets that contained 100s of thousands of crimes in each city.

Data Cleaning:

The data spanned across multiple years and held a lot of unnecessary information. It needed intense cleaning and simplification. To help, I chose to only analyze 2014 data. For most precincts, it is the most recent yet complete dataset. Using Python, I began exploratory data analysis to see the best way to analyze crime differences. I quickly discovered a huge problem in how crimes were being reported. Some states call the same crime totally different things while others have multiple ways to classify a single crime. For example, the Los Angeles precinct may use assault with a deadly weapon as a categorization label while the Seattle precinct further divides this kind of crime by having subcategories defining what kind of weapon (assault with a knife, bat, etc.). Every state used a different reporting methodology and different key terms. To account for this, I went through all the different types of crimes and classified them under a standard category. I used broad meta-labels like theft, murder, grand theft auto, etc. to unite each dataset with a common label. For example, robbery denoted all crimes where someone robbed someone else whereas burglary meant breaking and entering/stealing while theft regarded any stealing from public premise, shoplifting, etc. This created around 30 mega-labels; however, in the end, I only used 8 that were most relevant to study 1. With the data cleaned, isolated to 2014, and painstakingly categorized, I began analysis.

Data Analysis:

With the cleaned dataset, I created visualizations of crime dispersion in each city. I found hotspots of crime, often in the most populated areas of the city. I also saw distributions of which crimes were most common in which cities. As a final means of objective analysis, I scaled criminal activity to population size (per-capita conversion) and compared cities based on crimes per civilian. The results as to where/what crimes happen in which cities were surprising.

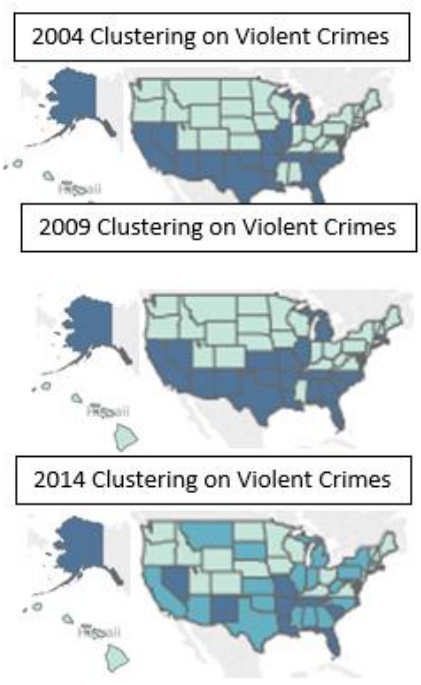
Results

Study 1

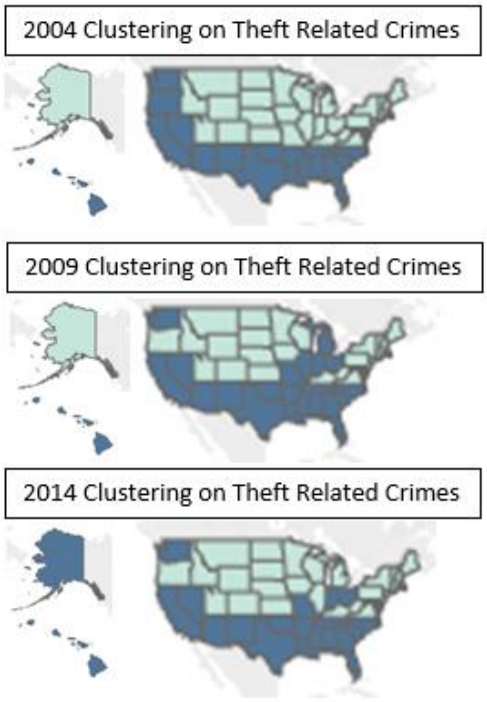
After running a cluster analysis using the 2-3 recommended clusters, I found significant groupings of states across America based on violent and theft related crimes. Some states did have higher rates of crime proportional to their population. Further, these rates change over time with some states becoming more comparatively violent and less violent. To make sure the clustering was accurate I implemented further tests. Sometimes clusters can be chosen but don't properly represent the observations because the data should be singularly clustered. However, I found that these clusters were very accurate using the same Silhouette score as a testing mechanism. Almost all observations (states) have a positive Si score implying they are in the cluster that suits them best.

These are two visualizations of which states fell under which clusters during a 10 year period. The darker a state the more crime occurred there. In the end, the violent crimes clustering was created using K-means whereas the theft related crimes clustering was created using agglomerative clustering. As you can see, Southern states do have more crime per-capita. However, some states I wouldn't have suspected (West Coast) were grouped with the Southern states. Northeast and Midwest states generally cluster together whereas Western and Southern states are more similar.

Clustering of Violent Crimes 2004-2014 (via K-means)



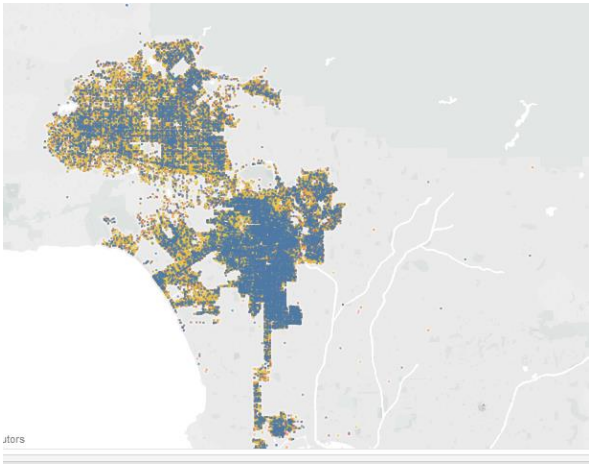
Clustering of Theft Related Crimes 2004-2014 (via Agglomerative clustering)



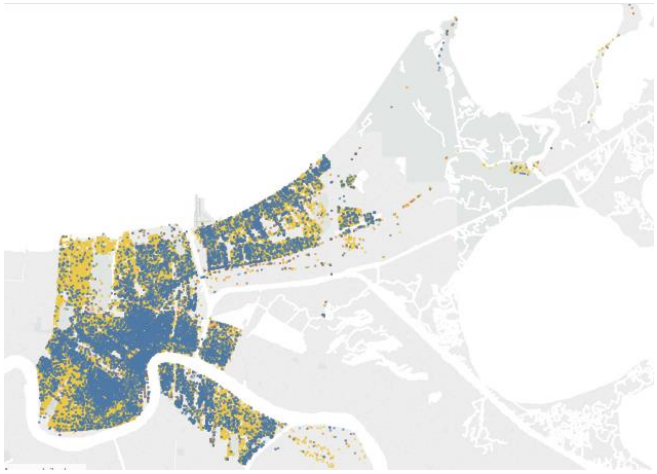
Study 2

Using a visual analysis of 2014 data, I observed large differences in both the amount of crime and type of crime across cities. I created the following maps to better understand crime dispersion in each studied city. However, these visuals don't account for city population, each dot simply represents a crime that occurred there. Cities with smaller populations will naturally have less dots. Looking through the data, it does seem like some types of crime (burglary and assault) are more common in cities like Los Angeles and New Orleans than Hartford.

Los Angeles, California

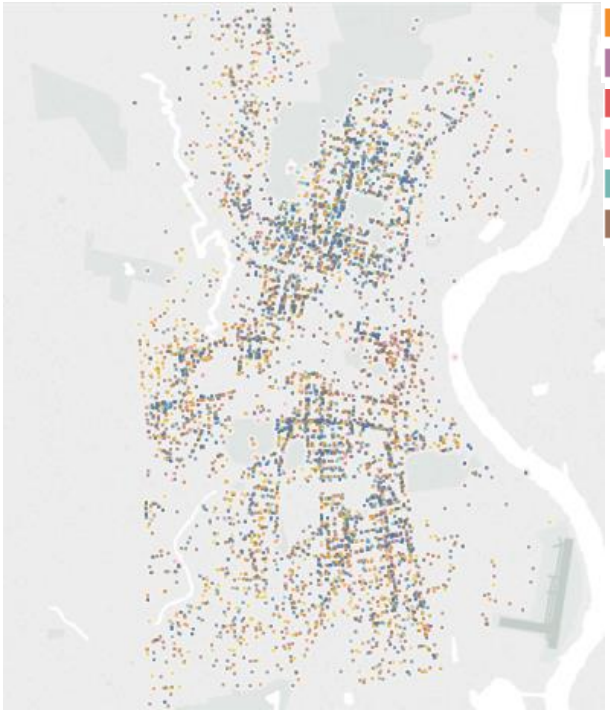


New Orleans, Louisiana

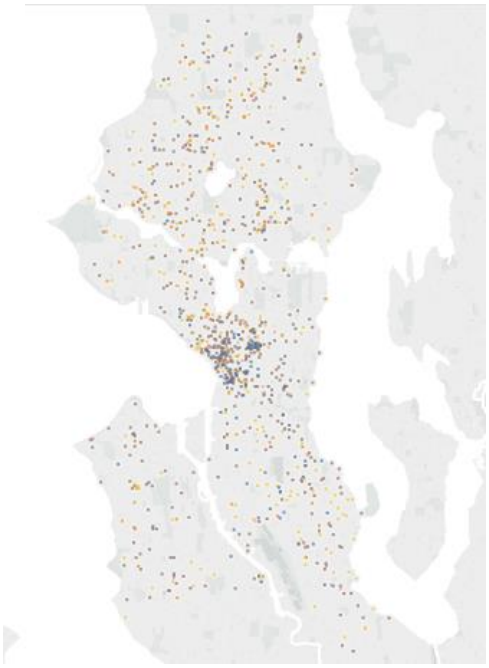


Legend (by crime type)

Hartford, Connecticut



Seattle, Washington



As one can see, cities like Los Angeles and New Orleans do have a ton of crime compared to Seattle and Hartford. In fact, Seattle was ranked on of the safest cities in America [13]. Yet, with this kind of data, one must consider the population of each city before reading too much into the results. Sure a city may have a lot of crime but it could also have a huge population. Below is a table of crimes per thousand citizens categorized appropriately. A 0 indicates missing data or a number too small to be calculated significantly.

Table of Crime Per-Capita

	Assault	Burglary	Grand Theft	Murder	Rape	Robbery	Shooting	Theft	Population
Hartford	16.719458	7.898641	6.110421	0.168397	0	4.081633	4.707109	38.579047	124,705
Nola	69.996357	150.413718	14.938072	0.400708	0.850853	4.589925	7.790383	46.398834	384,320
LA	10.22143	7.238992	3.766607	0.0682	0.756172	2.020107	0.158819	13.716467	3,929,000
Seattle	0.327677	0.644879	0.315707	0.0299	0	0.0823	0	1.584518	668,342

For example, despite the map showing 20 times as many assaults in Los Angeles as Hartford, Hartford actually had more assaults per citizen. This is due to the enormous Los Angeles population compared to Hartford's much smaller population.

Discussion

Conclusions:

After weighing the results, my analysis and visualizations allow for a better understanding of the data. One of the most important things to consider is population dispersion. Research regarding rates of innovation and wealth associated with crime seem to be true [10]. Cities tend to have higher rates of crime. Fascinatingly, even states clustered into the lower crime index can contain cities with comparatively (to other cities) higher amounts of crime; the inverse is also possible, but much less likely. This implies that a state with less crime may have a dispersed population with a lot of people living in rural areas (like Connecticut). There is less crime outside the city and thus less crime per person in the overall state.

A state may seem safe overall, but one must consider each major city as an important mark as to how safe the state is as a whole. What we envision as the safest places to live may not actually be the safest. Anomalies in the data show we may undervalue the importance of crime per-capita rather than overall crime. For example, despite the conception that Los Angeles is a dangerous city, when accounting for population these data show there is less crime than one would guess. However, some cities, like Seattle, are still unequivocally safer than other cities.

Hypothesis Findings:

Regarding state research across the country, my hypothesis was proven. America can be divided based on crime per-capita data across certain states. Further, we see irrefutable geographical trends linking certain states together. However, I was incorrect about where the exact divide would lie. It appears that Western states are more similar to Southern states and

Midwest states group with Northeastern states based on both violent and theft related crime indexes. This may be due to larger and more concentrated populations in Western and Southern cities within each states with less population dispersion to rural areas. It can also be because of cultural differences across regions in America. Finding the true cause of these differences will be a long process involving a combination of the metrics we use to analyze, other dividing factors, and searching for lesser known causes.

Within city to city research, my hypothesis was partly confirmed. I predicted that cities will all have large amounts of general crime, but specific differences as to type. There were indeed cities like Seattle that had much less crime. Further, cities did show differences in crime rates; to my surprise, some of the "safer" cities (Hartford) had more crime per-capita than the more "dangerous" cities (Los Angeles). Again, I would hypothesize that the cities with more crime residing in states clustered to have less crime, contain a more dispersed statewide population.

Limitations:

It is hard to check the statewide UCR dataset for completeness as it is based on metadata. There are few resources with this kind of information, making fact-checking difficult. Within the citywide Police Records data, I noted several weaknesses. Firstly, there are huge limitations within the datasets I was able to find. It seemed like there are missing or misplaced data regarding individual crimes. For example, both Seattle and Hartford reported 0 rapes in 2014, but a quick Google search shows this is completely false. Latitude and Longitude was even reported incorrectly in some observations. Similarly to the UCR dataset, it is hard to corroborate the accuracy of each precincts arrest data, as there is no outside repository with equal information. Cross-referencing against a different dataset would allow me to pinpoint any weakness in the data.

Secondly, after completing these studies, it is clear that everyone harvests and reports criminal activity differently. Each precinct must have their own methodology for documenting every crime and encoding it into their dataset. This creates huge data analysis limitations, as the lack of a common system forces one to analyze each dataset individually. I believe that each precinct uses different codes, categories, and even software when creating these datasets.

Future Research:

Future research would greatly benefit from a more complete and confirmed dataset. That way, a data analyst can be confident their conclusions stem from accurate criminal data rather than an incomplete dataset. Researchers would also benefit from using a standardized dataset with similar categorization and features across cities. It would allow for a more rigorous analysis without time spent cleaning and checking. This kind of research also benefits greatly from using as much data as possible. The city data were helpful, especially for visualizations, but it would be even more effective to analyze every major city in America, instead of a pre-selected four. County specific datasets would also allow a merger between these two studies as one could observe geographical trends across state lines. It is important to note that my research is partially

a proposition for a new methodology of analysis and any results must be taken with a grain of salt due to dataset limitations.

Recommended Policy Change:

After performing my research, I can attest that there are two clear and necessary changes to crime reporting. Firstly, there should be an automated crime report application. A simple software that allows any police division with a completely standardized process would help data purity immensely. Each crime is very different, but if we can find a way to simplify and automate, it will help the police, civilians, and even the government. It is very difficult to analyze crimes when everyone records data differently or incorrectly. With a new simplified software, we must find a way to make it an industry standard. Every precinct across America should have the ability to accurately and confidently report crimes with this application. Again this would help all parties involved and allow the government to easily keep a large scale repository of all crime across cities, counties, and states for in-depth analysis.

References:

[1] Ford, M. (2016). What Caused the Great Crime Decline in the U.S.? The Atlantic. Retrieved 5/5/17. https://www.theatlantic.com/politics/archive/2016/04/what-caused-the-crime-decline/477408/

[2] (2014). Political Polarization in the American Public. Pew Research. Retrieved 5/5/17. http://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/

[3] Fritz, N. (2015). Economic Inequality: It's Far Worse Than You Think. Scientific American. Retrieved 5/5/17. https://www.scientificamerican.com/article/economic-inequality-it-s-far-worse-than-you-think/

[4] Milligan, S. (2014). The Divided States of America. U.S. News. Retrieved 5/5/17. https://www.usnews.com/news/articles/2014/04/10/the-divided-states-of-america

[5] Triandis, H. C., & Suh, E. M. (2002). Cultural influences on personality. Annual review of psychology, 53(1), 133-160.

[6] Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the Southern culture of honor: An" experimental ethnography." Journal of personality and social psychology, 70(5), 945.

[7] Cohen, D., Vandello, J., Puente, S., & Rantilla, A. (1999). "When You Call Me That, Smile!" How Norms for Politeness, Interaction Styles, and Aggression Work Together in Southern Culture. Social Psychology Quarterly, 257-275.

[8] Vandello, J. A., Cohen, D., & Ransom, S. (2008). US Southern and Northern differences in perceptions of norms about aggression mechanisms for the perpetuation of a culture of honor. Journal of cross-cultural psychology, 39(2), 162-177.

[9] Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. Journal of personality and social psychology, 82(6), 903.

[10] Bettencourt, L. M., Lobo, J., Strumsky, D., & West, G. B. (2010). Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. PloS one, 5(11), e13541.

[11] Crime State by State. UCR. Accessesd 4/26/17. https://www.ucrdatatool.gov/Search/Crime/State/StatebyState.cfm

[12] Police Records Data. URL Defense. Accessed 4/26/17. https://www.policedatainitiative.org/datasets/

[13] Greenburg, Z. O. (2009). Full List: America's Safest Cities. Forbes. Retrieved 5/5/17. https://www.forbes.com/2009/10/26/safest-cities-ten-lifestyle-real-estate-metros-msa_chart.html