Understanding Crime in NYC and Chicago to build Sustainable Cities And Communities

Adithya V Ganesan, Koushik Kumar Reddy Modugu*, Manoj Kumar* & Abhishek Deshmukh**

1. Introduction - Our work is directly connected with SDG 11[1]. More than half of the population now live in towns and cities, and with migration and population growth, that proportion is only going to increase further. The cities are only going to get busier—at the same time it is pertinent to ensure a safe, healthy, and happy place to live. Our work focuses on the analysis of the safety and security of the communities. Estimates suggest that, by 2050 approximately ⅔ of all humanity will be urban. Crime rates are on the rise in urban areas due to various factors such as economic and social. In light of the population growth and rapid urbanization advances all over the world, crime has become one of the most serious social problems. Crime and the fear of crime are serious issues confronting societies and contribute to the decline of the quality of life. *Hence we aim to understand the nature of crimes in two major cities: NYC and Chicago, investigate the potential reasons and suggest what can be done to build safe and human inhabitable communities.*

2. Related Work – In this section we briefly discuss related research efforts and explain why we focus on crime prevention to make our community secure and sustainable.

Crime Prevention v/s Crime Control: Crime Control refers to methods taken in order to reduce crime in the society. But many of the traditional modes of crime control were ineffective at times. It is this notion of crime control that also led to innovation of alternatives such as natural surveillance, natural access control which fall under a methodology called crime prevention. It has been defined in various ways by many different people. Operations to deal with gang offences, securing a correlational facility etc. More often than not, crime prevention refers to preventing criminal offending – before it has been committed. Crime prevention alone holds much promise in preventing crime, but less is known of its effectiveness . In order to go a long way towards building a safer and more sustainable community , a greater balance has to be made between crime prevention and crime control *(Welsh et al, 2012)* , for which we will be addressing effective measures throughout this proposal.

Using Big Data to make Cities Safer : With the emerging field of data analytics and revolution in big data collection, much analysis could be made offering a fresh look at the historical perspectives. Considering that in mind, we look at some of the research work done by Wharton Statistics Department and understand some of the practical principles for community development *(Wood, 2018)*. Working with publicly available data, they investigated the relationship between crime frequency, neighborhood factors and economic conditions to arrive at some conclusions. Two Important findings state that –

- Modest poverty alleviation in the poorest neighborhoods may reduce violent crimes.
- High-vacancy neighborhoods appear to have higher crime, but crime does not happen near vacancies.

These findings gave us a sense of understanding of how to approach our analysis. *Inferring from above, we try to understand the potential cause of crime based on statistical evidence, socio-economic features affecting the neighborhood.* Analyse the impact of hate and non-hate crimes in society and propose measures to mitigate these crimes

3. Data - The following data were used for the project.

NYC Crime Dataset[2]: The NYC arrests dataset comprised all the arrests made since 2006. This transactional dataset includes information relevant to each arrest such as date, offense description, perp's race, perp's sex, perp's age group, location and precinct number.

Chicago Crime Dataset[3]: The Chicago arrests dataset had records since 2001. This transactional dataset includes information relevant to each arrest such as date, offense description, location, whether or not domestic violence and community area.

[1] *http://www.data4sdgs.org/SDG11*

** equal contribution*

[2] https://www.opendatanetwork.com/dataset/data.cityofnewyork.us/8h9b-rp9u

[3] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Unemployment Data[4]**:** The month wise unemployment data for these two cities were used to test the hypothesis if this was an indicator of crime.

Household Income Data[5]**:** The year wise distribution of household income across various buckets were used to test for a potential significant correlation with crime.

Chicago Housing Data[6]**:** The Chicago housing permits dataset had information regarding the fee structure of the house which inturn was summed to represent 'housing value' and the location of the house. This was used to understand the relevance of crime with location.

NYC Housing Data[7]**:** The NYC housing dataset was also used to understand the relevance of a location with crime. This dataset had information such as location, number of units affordable by each income group (segregated as extremely low income, very low income, low income, moderate income, middle income and others) for each building in NYC. Additionally, it had the number of studio units, 1 BR units, 2 BR units etc for each building.

3. Infrastructural Overview -

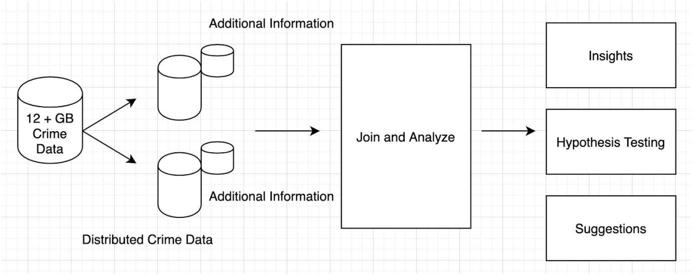


Fig (1): Infrastructure used for the project.

We had 12+ GB of crime data for NYC and Chicago in the hadoop distributed file system *(Shvachko et al, 2010)* and all other data such as unemployment, household income and housing data were broadcasted to all the nodes owing to their small size as shown in *fig(1)*. We used pyspark (*Zaharia et al, 2012*) data frames to load the data set and perform all the necessary operations such as filtering and joining.

For every insight we went ahead and performed hypothesis testing to determine the significance of each new finding. Based on the insights and their significance we provide suggestions that can help decrease the number of crimes.

4. Methods -

NYC crime by type vs Demographics: The demographics of the perps such as sex, race and the age group is recorded for each arrest in the NYC crime dataset. To understand the significance of the correlation between accused perps' demographics for each crime type with the respective crime, the transactional dataset was turned into key-value pairs, with each key representing a crime type and the values having a list of the following column vectors:

I. Monthly count of the crime.
II. Monthly count of each category in perp sex (M/F)
III. Monthly count of each category in perp age group
IV. Monthly count of each category in perp race

Following this processing, each of the column vectors in (II, III, IV) were tested for significant correlation with I and the p-value along with correlation were returned only for the demographics that were significant after applying Bonferroni correction.

Chicago crime vs Unemployment: Both the Crime count and Unemployment Rate of Chicago city from 2012, for every month, are considered for this specific analysis. The correlation of 0.88 was observed. Upon fitting the linear regression, predicting crime count using unemployment rate resulted in a slope of 546.22 with intercept 1945.87 and r-squared value of 0.78. We then performed slope significance hypothesis testing. Considering the null hypothesis that the slope is insignificant, the p-value obtained was 0.0. Null hypothesis was rejected and *thus the unemployment rate is a good predictor of crime rate* in Chicago.

NYC crime by type vs Unemployment: Overall crime count turned out to have a significant correlation with unemployment rate. To understand more about its significance we aggregated arrests data by types of crime and turned it into a dataframe with each column representing its monthly count and broadcasted the unemployment rate column. These monthly count crime vectors were each paired with the unemployment rate to be fed into a regression model. Correlation for each of these crime types with the unemployment rate was calculated and top 10 crimes that are highly correlated were obtained.

[4] https://www.bls.gov/
[5] https://www.datausa.io/
[6] https://data.cityofchicago.org/Buildings/Building-Permits/ydr8-5enu
[7] https://data.cityofnewyork.us/Housing-Development/Housing-New-York-Units-by-Building/hg8x-zxpr/

Unemployment is considered as a feature to predict each of ten crimes individually to validate how significant the observation was.

Chicago crime vs Income: We aggregated the Chicago arrests data based on the year they were committed, to get the total number of crimes per year. We joined this data with the household income data that we had to study the correlation between how total crime changes with different economic buckets. We had 16 different economic buckets where bucket 0 represented the section of people with income < $10,000 and bucket 15 represented people with income $200,000+. We also had the intuition that income tells us a lot about domestic crimes thus, we also did a similar analysis on how domestic crime correlates with different income buckets. The results have been discussed in the results section.

Chicago housing data analysis: To investigate if there are patterns in the cost of a locality with the crimes, we aggregated the cimes data based on the community to get the number of crimes per community area. We then analyzed the total crimes from a community area with the average price of a building in that community. (We believe that average price of a building in a community is a good proxy for the overall quality of the locality). The results have been discussed in the results section.

NYC housing data analysis: A custom metric had to be used to account for missing 'housing values' within a particular area in NYC housing dataset. It is calculated by a weighted average over the number of units of particular type in a community based on their earnings, viz, Extremely Low, Highly Low, Low, Moderate , Middle incomes.

$$\text{Housing Value} = (\sum_{i} (W_i)^* (U_i) \ / \sum_{i} U_i$$

W_i = Weight associated with a particular earning type, determined by the range of earnings in that type.

U_i = Number of units associated with a particular type.

We used a uszipcode[8] search engine to get zipcode by latitude and longitude in both NYC Housing and crime datasets. We aggregated the datasets by zip code and merged them by zip code, while grouping housing values for a zip code we created two separate data frames , one each for aggregation by sum , aggregation by median for analysis. Column vectors for crime count and housing values are created and passed onto a regression model to validate how significant the observations made are.

Modelling crime as a function of Unemployment: Since the crime by type was found to have a significant correlation with unemployment (*results section*), the monthly crime count (by type) for the top ten crimes reported in NYC were modelled as a function of unemployment using an LSTM model *(Hochreiter et al, 1997)*. This LSTM model was provided with two time steps of unemployment data to predict the crime of particular type for the month. *These models were built using pyTorch (Paszke et al, 2019) and all these models were parallely trained on CPUs.*

5. Results -

NYC crime by type vs Demographics: The crime by type column vectors were tested for significance with each demographic column vector and only the ones with a significant correlation were returned along with the p value and correlation value. *Fig(2)* is an example of the top 5 significant correlating demographics for felony assault.

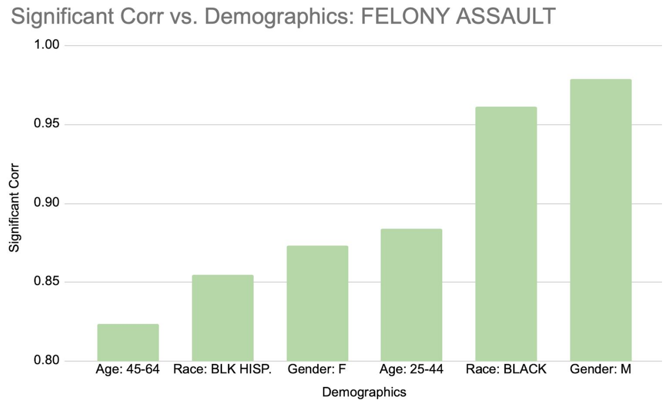


Fig (2): Top 6 significant correlation of demographic features for Felony Assault.

A few more figures, *fig (a-2; a-3; a-4)*, for other crimes can be found in the Appendix. From this analysis we were able to conclude that male, black and white-hispanic were commonly occurring significant demographic features with the top 10 crimes in NYC.

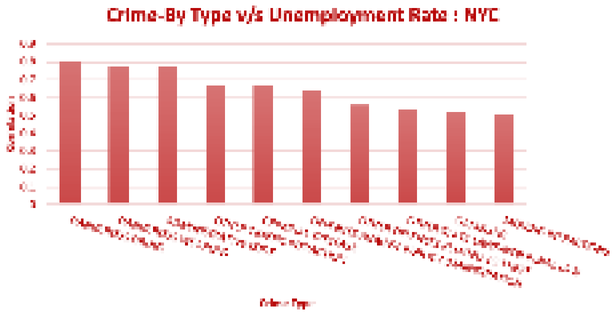
Chicago Community Level Analysis: Considering the data of top 7 crimes by count (Battery charge, Domestic battery, Assault, Retail theft, Criminal land trespass, Unlawful possession of handgun, Possession of heroin

[8] https://pypi.org/project/uszipcode/

(Narcotics)), a very small set of communities out of 77 in total emerged to be the major grounds across different crime types. They are Austin, Near North Side, North Lawndale, Lake View. *Educating these communities, allotting more resources and improving surveillance in these regions should cut down the crime rate in the city significantly.*

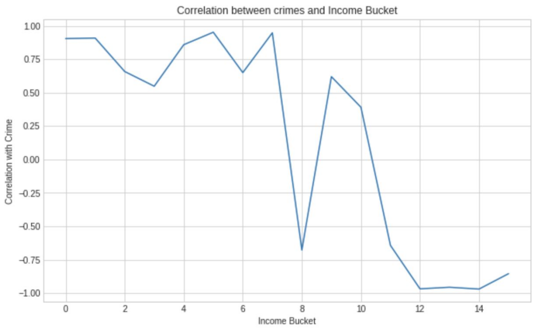
NYC crime by type vs Unemployment: The unemployment rate is used as a feature to test for significance with each of the top 10 crime types based on correlation. *Fig(3)* is the plot for top 10 crimes by type and their corresponding correlation values.

The hypothesis testing results for each of these crime types indicate insignificant slope and p-values , which rejects the null hypothesis and signifies that *the unemployment rate is actually a good predictor for crime rate among these types of crimes for NYC* .



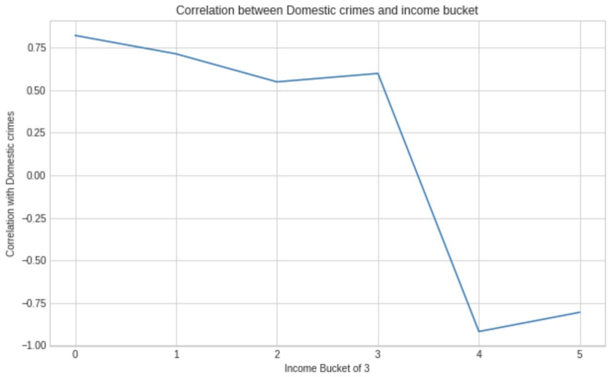
*Fig(3): Correlation values between top 10 crimes and unemployment rat*e

Chicago crime vs Income: We observed that higher income groups are involved with fewer crimes overall.



Fig(4): Correlation values between total crimes and different income buckets.

From *fig(4)*, the correlation decreases and eventually moves into negatives as the income bucket increases. When we studied the low income buckets in particular we saw a correlation of 0.8 with crimes, thus *more and more people accumulating in the low income buckets will result in an increase in the number of crimes*. We also did a hypothesis testing for the same to validate how significant the observation was, we found a p-value of 0.013 and we concluded that *income is actually a good predictor of crimes*.

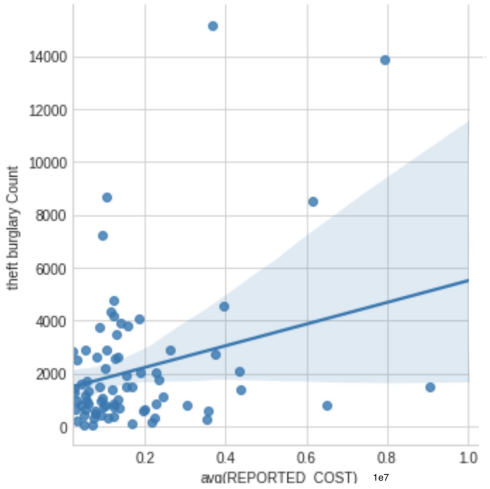


Fig(5): Correlation values between domestic crimes and different income buckets.

We did a similar study between domestic crimes for the low income bucket and observed from *fig(5)* that a similar trend follows where the correlation decreases and eventually moves into negatives as the income bucket increases. For the low income buckets we found a correlation of 0.82. We also did a hypothesis testing for the same to validate how significant the observation was, we found a p-value of 0.045 and hence we concluded that *income is not only a great predictor for crimes in general but also for domestic crimes in particular.*

Chicago housing data analysis: We observed a correlation 0.001 between crimes in a community area and the average price of a building in that community. This means that there is no correlation between community and overall crime.

When we further analyzed just the crimes related to theft and burglary from a community and the average price of a building in that community we saw a correlation of 0.3 (*fig (6)*). On hypothesis testing for the same we got a p-value of 0.008 which confirms that average price of a building in a community is a significant predictor for the total number of theft and burglary. *Since we see that well-off communities are more susceptible to crimes related to theft and burglary we make a suggestion of improved CCTV surveillance in such communities.*



Fig(6): Graph between average price of a building in a community area and the number of thefts and burglary in that area.

NYC housing data analysis: The 'housing value' for an area is used as a feature for a regression model to predict crime count. A positive correlation of 0.386 was found between housing values and crime count when housing value for an area is aggregated by sum , but a negative correlation of -0.1123 was found when aggregated by median. The latter makes more sense since the data can be skewed when sum is considered as an aggregation as a consequence to economic factors in that area, and also similar negative correlation was observed in Chicago housing analysis. The hypothesis testing results for housing value and crime count indicate insignificant slope and p-values, which rejects the null hypothesis and signifies that the *housing value is actually a good predictor for crime rate.*

Similar analysis was made by filtering top 10 crimes - a slightly increased negative correlation of -0.1138 was found between crime and housing values.

Modelling crime as a function of Unemployment: Each of the top 10 crimes in NYC data was modelled as a function of unemployment. The input to the model was two time steps of unemployment data. The results of these experiments in *table(1)* is suggestive that the RMSE values are proportionate to the mean crime over months. A visualization of this can be seen in *fig.(a-1)* in the Appendix.

On experimenting with 3 time steps of unemployment data, it was observed that it took significant time to train and produced very little increase in efficiency. Each LSTM model was parallely run by putting it inside a map function.

CRIME TYPE	RMSE	CRIME RATE PER MONTH
FELONY ASSAULT	35.62	1200.731
OTHER OFFNS. REL. TO THEFT.	39.47	2142.925
CRIM. TRESPASS	21.85	1003.657
DANGEROUS WEAPONS	29.95	1127.120
OTHER STATE LAWS	15.67	1474.870
VEHICLE. & TRFC.	40.75	1480.314
DNGRS. DRUGS	60.83	5880.87
OTHR TRFC. INFRACTION	19.33	787.490
ASSAULT 3 & REL. OFFNS.	53.53	3014.222
PETIT LARCENY	44.16	1490.083

Table (1): RMSE and the Mean crime over months.

6. Conclusion - (1) *Unemployment is a good predictor of major crimes.* Modelling crimes using unemployment with RNN seems to be effective. (2) *The lower income groups seem to have significant cases of domestic crime.* Policy planning should cover mental health and counselling support. (3) *Well off communities are more susceptible to theft and burglary.* CCTV cameras should be installed and other communities should be developed with parks, schools and amicable neighborhoods. Following this would help build safer and peaceful communities.

References -

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, pages 8024–8035, 2019.

Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael JFranklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets:A fault-tolerant abstraction for in-memory cluster computing. In Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, 2012. www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf.

K. Shvachko, H. Kuang, S. Radia, and R. Chansler. The hadoop distributed file system. In Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), MSST '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

Jonathan Wood ,2018. How to Use Big Data to Make Cities Safer, WG.

Brandon C. Welsh , David P. Farrington (2012). Crime Prevention and Public Policy , The Oxford handbook of crime prevention.

Massoomeh Hedayati Marzbali, Aldrin Abdullah, Nordin Abd. Razak, Mohammad Javad Maghsoodi Tilaki (2011). A Review of the Effectiveness of Crime Prevention by Design Approaches towards Sustainable Development, Journal of Sustainable Development.

Appendix

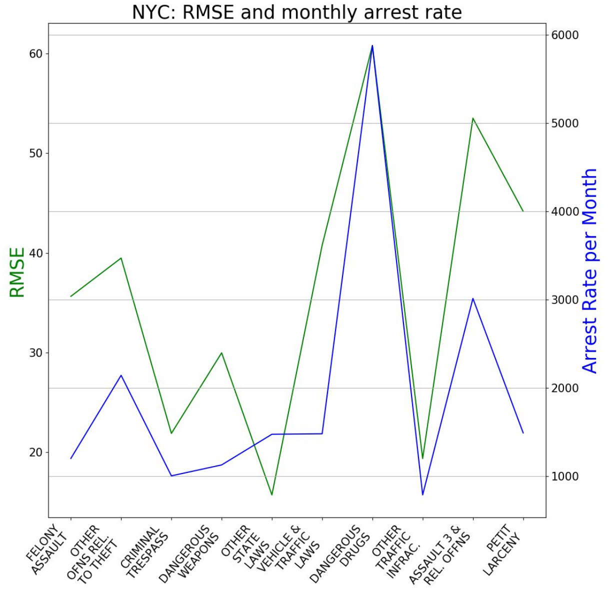
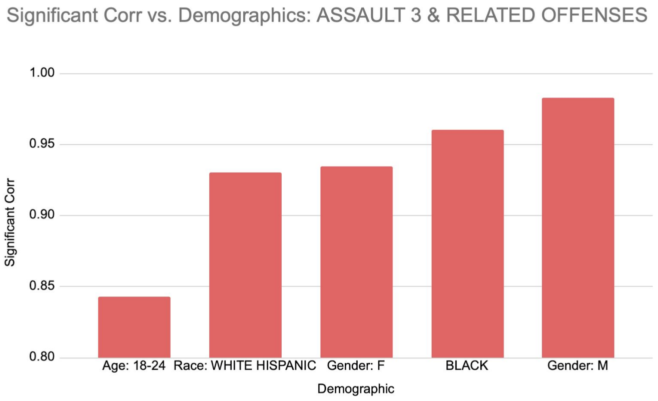
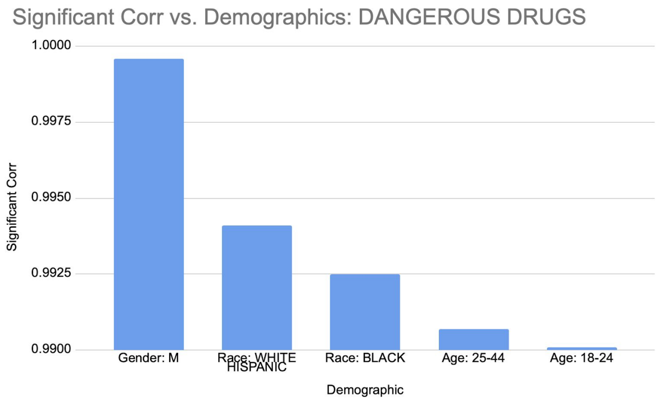


Fig (a-1): RMSE vs Arrest Rate per month. The RMSE values are higher for those crimes that have high mean value.





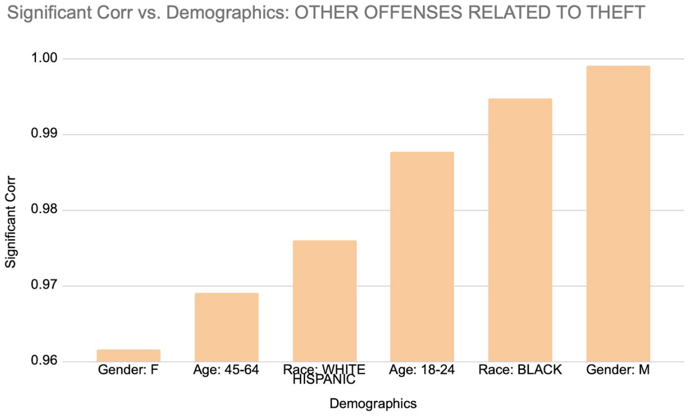


Fig (a-2; a-3; a-4): Top 6 significant correlation of demographic features for Assault 3 & Rel. Ofns., Dngrs. Drugs and Other Ofns. related to theft.