Investigating the Causal Relationships between Public Health and Violence/Crime in Chicago

Morgan Kelley
Department of Data Science
UC San Diego
mokelley@ucsd.edu

Mika Philip
Department of Data Science
UC San Diego
mmphilip@ucsd.edu

Abstract

For this data science project, we aimed to analyze the relationship between different attributes related to public health and violence/crime in Chicago by leveraging causal inference methods, such as PC Algorithm, ICA, HSIC, Fisher's Exact Test, and GES. Drawing our data from the city of Chicago, we explored a different method of data analysis that did not utilize traditional regression methods. With some assumptions, we found some interesting structures that provided insight worth further exploring.

1 Introduction

Chicago is a major U.S. city that has grappled with public health and crime issues for a while. Its crime rate is frequently over the national average, and it has been rated the 10th most dangerous city to live in by Property Club [10]. The complex and intertwined issues of health and safety significantly affect the overall well-being of Chicago's residents. It is essential to comprehend the causal connections among these factors to devise impactful interventions and policies that can enhance public health outcomes and alleviate violence and crime rates.

This research endeavor focuses on examining the causal links between diverse factors influencing public health and crime in Chicago. By exploring these relationships, we want to identify potential interventions and strategies that can foster improved mental health and contribute to the reduction of violence and crime rates in the city.

2 Methodology

In this section, we will discuss how the data was collected and the causal methods used to analyze the data. All analysis was done in the Python programming language that read in CSV files of data downloaded from the City of Chicago website.

2.1 Data Collection

We begin by collecting relevant data sets provided by the City of Chicago. We obtained three data sets that contain information on different aspects related to public health and violence/crime in Chicago. These data sets include information such as public health indicators, crime reports, and police sentiment scores. The data sets share a common attribute of community areas, allowing us to combine them for a comprehensive analysis. The three data sets that we use are Police Sentiment Scores [1], Public Health Indicators (PBHLT) [2], and Chicago Crime (CRIME) [3].

Preprint. Under review.

2.2 Causal Discovery

In our analysis, we incorporated the ”causal-learn” Python package to facilitate the implementation of our causal discovery. Alongside the PC algorithm, we employed independent component analysis (ICA), Hill Climb Search, the Hilbert-Schmidt Independence Criterion (HSIC), and Fisher’s exact test.

The PC Algorithm, created by Peter Spirtes and Clark Glymour, allowed us to find a directed acyclic graph (DAG) for the data and finds independence conditional up to n variables. The PC algorithm requires the four assumptions of acyclicity, Markov property, faithfulness, and sufficiency [4]. ICA allowed us to extract distinct components from complex and noisy datato help disentangle different underlying sources of variation. Hill climb search was utilized to create a locally optimized Directed Acyclic Graph (DAG) that represented the relationships among attributes in the data. The HSIC test was employed to assess the level of independence between certain pairs of attributes in the dataset.[5]

Since the PBHLT dataset has 77 rows, a smaller dataset that may not hold well against the chi-squared test, we also employed the Fisher’s Exact Test to check for a significant relationship between two attributes based on the nominal proportions relative to each other. The output of this statistical method is a p-value that says whether or not to reject the null hypothesis (or, the proportion of one variable does not differ depending on the other). This can be used for independence testing, as rejecting the null hypothesis means that there is some dependency between the two attributes. [6]

Another algorithm that was used was the GES (Greedy Equivalence Search) algorithm, a score-based algorithm that tries to find a graph with the optimal BDeu (Bayesian Dirichlet equivalent uniform) score (for discrete data) and BIC (Bayesian Information Criterion) score (for numerical data). In order to use this algorithm, there are two assumptions that must hold between the attributes: 1) they form a DAG, and 2), there are no hidden latent confounders. Of course, those are very big assumptions to make considering there can be more real-world factors. Nevertheless, there can still be meaningful results that come from this method.

Using causal inference-related packages in Python, we were able to implement these algorithms to approximate the underlying causal structure of the data.

3 Experiments

In preparation for analysis, we had a data preprocessing phase that involved cleaning the data and addressing missing values. Additionally, we conduct exploratory data analysis to delve into the distributions and correlations within the datasets, allowing us to gain valuable insights before proceeding with further analyses. All figures can be found in the Appendix section.

3.1 Data Preprocessing

First, we provide a brief overview of the three datasets we use for causal analysis.

The Police Sentiment Scores dataset has 75 columns and around 7,000 rows. The columns consist of:

{ Org level, City, Area, District, Sector, Safety- 17 columns grouped by race, age, sex, education, income, Trust-17 columns grouped by race, age, sex, education, income, TrustListen- 17 columns grouped by race, age, sex, education, income, and TrustRespect- 17 columns grouped by race, age, sex, education, income}

The CRIMES data set has crime records from 2001 to 2018 that took place in Chicago. CRIMES has 22 columns and around 7 million rows. The schema of the 22 columns includes:

{’ID’, ’Case Number’, ’Date’, ’Block’, ’IUCR’, ’Primary Type’, ’Description’, ’Location Description’, ’Arrest’, ’Domestic’, ’Beat’, ’District’, ’Ward’, ’Community Area’, ’FBI Code’, ’X Coordinate’, ’Y Coordinate’, ’Year’, ’Updated On’, ’Latitude’, ’Longitude’, ’Location’}.

For our purposes, we mainly care about the Community Area, Primary Type, and Location Description. Primary Type contains a string of the type of crime that was committed, (e.g.: ”BATTERY’, ’THEFT’, ’NARCOTICS’), and Location Description contains a string of the main location the crime was committed, (e.g.: ’RESIDENCE’, ’CTA BUS’, ’SIDEWALK’).

The PBHLT dataset contains aggregated data of public health statistics, grouped by community area from 2005 to 2011. There are 77 rows, one for each of the 77 community areas in Chicago, and there are 29 columns with the following schema:

{Community Area and name, Birth Rate, General Fertility Rate, Low Birth Weight, Prenatal Care Beginning in First Trimester, Diabetes-related, Firearm-related, Teen Birth Rate, Assault (Homicide), Breast cancer in females, Cancer (All Sites), Colorectal Cancer, Diabetes-related, Infant Mortality Rate (per 1000 live births), Lung Cancer, Prostate Cancer in Males, Stroke (Cerebrovascular Disease), Childhood Lead Poisoning (per 1000), Childhood Blood Lead Level Screening (per 100), Gonorrhea in Males and Females, Tuberculosis, Below Poverty Level, Crowded Housing, Dependency, No High School Diploma, Unemployment, Per Capita Income }.

This data set also provides the national averages for each of the attributes, which we later make use of for data preprocessing for the HSIC test.

In this step to prepare to use the GES algorithm, we referred to the documentation provided by the city of Chicago and one-hot encoded each column. 1 would indicate that that community was at or above the U.S. baseline and 0 for below. For example, if the national baseline for lung cancer was 50.6 persons per 100,000 people and the Edison Park community reported having 45.2 per 100,000, then that cell would be one-hot encoded to 0. The only issue we ran into was with the "Childhood blood lead level screening" attribute having no baseline associated with it. Unfortunately, that attribute became extraneous with this pre-processing step.

During the data exploration phase, our main objective was to identify notable patterns and characteristics within the datasets, particularly in relation to different features and districts/community areas. Our findings revealed that District 11 exhibited the highest crime rate, while District 20 had the lowest (Figure 1).

The data preprocessing step ensures the quality and consistency of the data and prepares it for further analysis. It is essential to address any issues in the data that could affect the accuracy and reliability of the causal discovery and causal effects estimation processes. Data exploration allowed us to understand our data at a coarse level and connect conceptual dots as we proceed into causal discovery experiments.

We then create a correlation matrix of all of the values in the the PBHLT dataset to determine where we should initially focus our search efforts and to see where there are high (or low) correlations between different attributes (Figure 2).

4 Results of CRIMES

In this section, we present the results obtained from our experiments on investigating the causal relationships among violence/crime in Chicago. We discuss the inferred causal graphs, estimated causal effects, and key findings from the integrated data analysis.

4.1 ICA on CRIMES

First, we performed independent component analysis on crimes. Please note that ICA can only be applied in certain times. It requires the relations to be non gaussian, linear, and continuous. After analysis of the data, we found that no two columns had a correlation coefficient larger than 0.5, so we decided that ICA is most likely not appropriate for this data set. With that being said, we did still perform ICA to see if we could yield any meaningful results. To perform ICA, we first convert date to date/time format and then drop all columns that occurred before 2013. This leaves us with the Crimes dataset with 2.6 million rows. Then we once hot encode the "Primary Type" column and the "Location Description" column, which yields 256 columns. Then we created an occurence matrix to find the most common paired occurrences of crimes and locations. We find the highest pairings are: (THEFT, STREET, 147856), (BATTERY, APARTMENT, 128941), (CRIMINAL DAMAGE, STREET, 102442), (MOTOR VEHICLE THEFT, STREET, 94624), (BATTERY, RESIDENCE, 94580), where the tuples represent (crime, location, count). The occurrence matrix has form of (Figure 3).

We then ran ICA on this occurrence matrix to tease out the independent components. We chose to use 5 as the number of components, but note that there may be more significant results if using a larger number of components. We show the individual weights in (Figure 4).

We then find the largest 5 values for each component (Figure 5).

Mapping these back to the occurrence matrix shows that ASSAULT has a large impact on component 1, 4, 5, BATTERY has a large impact on component 1, BURGLARY has a large impact on component 1, and so forth. We then compute the mixing matrix to reconstruct the data. We find the reconstructed data weights as (Figure 6).

The reconstructed data obtained through the dot product of the estimated independent components and their transpose represents the data reconstructed from the extracted independent components using the mixing matrix. It can be interpreted as an approximation or reconstruction of the original data based on the independent components. Again, we believe that although the results of ICA are interesting, there must be more research into how to properly complete ICA with this dataset due to the lack of linearity.

4.2 PC Algorithm on CRIMES

In order to accurately run the PC Algorithm, we first tested the columns of Crimes to determine which columns were gaussian. We found that there were no columns that were gaussian distributed in this dataset. We only want to run the PC Algorithm on the numerical values of Community Area, the most common crimes, and the most common locations. Thus, we drop (["Ward","FBI Code", "ID", "Case Number", "Date", "Block", "IUCR", "Primary Type", "Description", "Location Description", "Arrest", "Domestic", "Beat", "X Coordinate", "Y Coordinate", "Updated On", "Latitude","Longitude","Location"] before running the algorithm. We then took a random sample of 10,000 and ran the PC Algorithm using the chi-squared test during pc.estimate() and a significance level of 0.0075. The results can be seen in (Figure 7). Note that each tuple (A,B) represents a directed edge from A to B, A → B.

4.3 Hill Climb Search on CRIMES

We then ran the Hill Climb Search on the crimes data set using the pgmpy HillClimbSearch package.We found very similar results to the PC Algorithm results. The only other connections we found were that BATTERY → DECEPTIVE PRACTICE, BATTERY → NARCOTICS, and SIDEWALK → NARCOTICS.

5 Results of PBHLT

5.1 PC Algorithm on PBHLT

When calculating which columns have a gaussian distribution in the PBHLT data set, we find that [General Fertility Rate, Prenatal Care Beginning in First Trimester, Preterm Births, Teen Birth Rate, Breast cancer in females, Diabetes-related, Lung Cancer, Prostate Cancer in Males] have a gaussian distribution. Thus, when we run PC Algorithm, we must exclude these columns. We then typecast all values to numeric and run PC Algorithm with chi-squared test and significance level of 0.00725 from the pgmpy package in Python. We find that with two sets of two conditional variables, "Childhood Lead Poisoning" → "Assault(Homicide)" and "Preterm Births" → "Infant Mortality Rate". Running the Hill Climb Search did not yield any connections.

5.2 Joining PBHLT and CRIMES

In this section, we aimed to see if we could find any new insights by joining CRIMES and PBHLT on community area. In order to do this, more preprocessing was necessary. Because PBHLT is aggregated data from 2005 to 2011, we ensured that we only took samples from CRIMES that were in the same time period. We then one hot encoded CRIMES and found the overall sum of each column. The City of Chicago Census of 2010 [7] reports the population of each community area, so we chose to use that as the population size for each area and then divided each value by this population to get a percentage of each crime/location, grouped by community area. Then we could join PBHLT and CRIMES on community area and run the PC algorithm again. The results are shown in (Figure 8).

5.3 HSIC Test - Preterm Births to Infant Mortality Rate

We wanted to check if “Infant Mortality Rate” (denoted as Y) is dependent on “Preterm Births” (denoted as X). Using the Hyppo Python package, we generated the HSIC score and got a score of 0.6 for X to Y and a score of 0.5 for Y to X. Both directions had a p-value ¡ 0.05, which means both rejected the null hypothesis. However, this means the algorithm found that they are dependent on each other. This could be because they actually point to each other directly or indirectly or that there is a latent confounder. When fitting these attributes into the Gaussian process regression (using the Sklearn Python package), both directions (Figure 9 and Figure 10) gave similar results where the actual and predicted seemed to somewhat overlay each other and not give an indication of causal direction.

5.4 HSIC Test - Childhood Lead Poisoning to Assault (Homicide)

HSIC was performed to test the relationship between “Childhood Lead Poisoning” (denoted as X) and “Assault (Homicide)” (denoted as Y). HSIC found that X to Y gave a score of 0 with a p-value of 0.06. Though the p-value is high, or that this fails to reject the null hypothesis depending on the acceptance threshold, this, nevertheless, means it found that Y is dependent on X. HSIC on Y to X gave a score of 0.3 with a p-value ¡ 0.05. So, there seems to be some indication that “Childhood Lead Poisoning” may have some effect on “Assault (Homicide)”. This is particularly fascinating, as research has shown that correlation. For example, there was a study that showed a higher lead level in blood and aggression scores (using the Buss-Perry Aggression Questionnaire) in shooters as compared to archers at a shooting range.[8] Gaussian process regression also agreed with that $X \rightarrow Y$ (Figure 11 and Figure 12).

5.5 Fisher’s Exact Test - Preterm Births to Infant Mortality Rate

Using the one-hot encoded data from the pre-processing stage, we counted the ones and zeroes for two attributes to create a two-by-two matrix. One column represented “Preterm Births”, and the other represented “Infant Mortality Rate”. The first row represented all the times one occurred, or when the community was above the national baseline, and the second row had the counts of all the zeroes, or when the community was below the national baseline. To check if there was dependence between the two columns, we used the fisher_exact tool in the Scipy Python package. We found a p-value of 0.004, indicating that there is a significant difference between the proportion of “Preterm Births” depending on “Infant Mortality Rate”. Thus, we can reject the null hypothesis and say that the proportion of “Preterm Births” is affected by or affects “Infant Mortality Rate”.

5.6 Fisher’s Exact Test - Childhood Lead Poisoning to Assault (Homicide)

Similar to the previous section, we did the same process except for the attributes ”Childhood Lead Poisoning” and ”Assault (Homicide)”. Similary, we got a low p-value of approximately 0.01. Thus, we can reject the null hypothesis and say that the proportion of “Childhood Lead Poisoning” is affected by or affects “Assault (Homicide)”. These results are very similar to those from HSIC.

5.7 GES Algorithm on One-Hot Encoded Data

Using the one-hot encoded data as mentioned in the preprocessing section, it was plugged into the GES algorithm with the BDeu score (as it was a discrete input), and it produced this causal graph (see Figure 12). Some attributes that were not included final graph were “Prenatal Care Beginning in First Trimester” and “Breast Cancer in Women”. Both did not have much variety in terms of being below or above average, so that could have affected the output. As aforementioned, “Childhood Blood Lead Level Screening” had no matching national baseline, so that factor was not considered.

There are some interesting outputs produced from this method. One, it showed that the gonorrhea in females affects gonorrhea in males, which seems reasonable considering how STDs spread. Second, it showed that all the types of cancer this study covered were related to each other but the direction was not clear. That result is also understandable, as it is not clear from the data alone which would cause what, especially if the attribute “Cancer (All Sites)” seems to be the collection of all types of cancer. Thirdly, income affected people not having a high school diploma, which indicated crowded housing, which thus affected the tuberculosis rate. This is a fascinating find, as tuberculosis is more likely to spread between people who spend time with each other every day. Crowded housing is an extremely likely a living condition that would cause that. [9]

This graph also had some interesting callbacks to sections 6.1, 6.3, and 6.4. One is that this DAG indicated that a community being above or below the national baseline for ”Infant Mortality Rate” determines if its ”Preterm Birth Rate” is above or below the national baseline. But, this may just be because of the assumption and the fact that the PC algorithm is a constraint-based method and not a score-based one. And when examining the relationship between ”Childhood Lead Poisoning” and ”Assault (Homicide)”, GES showed they may have shared a common ancestor with preterm births. Regardless of the assumptions, this method produced some intriguing findings with some explainable backing.

5.8 GES Algorithm on Original Data

We hoped to run GES on the original data with no modifications using BIC scoring. However, with 20+ attributes in the dataset, the computer ran out of computing space. We were hoping to see if there would be any relations pertaining to ”Childhood Blood Lead Level Screening” (which the previous section mentioned how and why it did not show up on the final graph). Regardless, performing one-hot encoding significantly reduced computation time and still gave interesting results.

6 Conclusion

Our research focuses on the intersection of public health and public safety. We showed strong causal evidence of public health issues causing crimes, most notably the causal relationship between childhood lead poisoning and adult homicide and the causal relationship between preterm births and infant mortality. We relied on many different tests to find the weights of causal relations. We ran the PC algorithm, Hill Climb Search, Greedy Equivalence Search, Hilbert-Schmidt Independence Criterion, Fisher’s Exact Test, and Independent Component Analysis. It is important to note that these tests require specific data criteria that are mostly satisfied by the sentiment scores, CRIMES, and PBHLT data, but there is still room for improvement. We hoped to run GES on the original data with no modifications using BIC scoring. However, with 20+ attributes in the data, the computer ran out of computing space. Some further efforts would be to run GES with more computing power to see if there are significant results with this different scoring method.

Our results show the importance of public health in order to ensure public safety. More importantly, causal inference can be a tool that is leveraged to figure out where to investigate sociological or public health phenomena. In the future, we hope to find more datasets that can be further joined with the datasets of our paper to find additional revelations of public health in Chicago.

7 Acknowledgements

We would like to thank Professor Dr. Biwei Huang for giving us this opportunity and for her guidance on which methodologies to use for analysis. We would also like to thank Dylan Stockard for generating the graphs in Figure 1. And finally, we would like to thank Fitz and Sasha for all their support.

8 References

[1] Levy, Jonathan (2023). *Police Sentiment Scores* [Data Set]. https://data.cityofchicago.org/Public-Safety/Police-Sentiment-Scores/28me-84fj

[2] Cocadmin (2022). *Public Health Statistics - Selected public health indicators by Chicago community area - Historical* [Data Set]. https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu

[3] Chicago, City of (2018). *Chicago Crime* [Data Set]. https://www.kaggle.com/datasets/chicago/chicago-crime

[4] Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search (Vol. 81). MIT press.

[5] Gretton, A., Bousquet, O., Smola, A., Schölkopf, B. (2005). *Measuring Statistical Dependence with Hilbert-Schmidt Norms*. In: Jain, S., Simon, H.U., Tomita, E. (eds) Algorithmic Learning Theory. ALT 2005. Lecture Notes in Computer Science(), vol 3734. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11564089_7

[6] McDonald, J. H. (2009). *Handbook of Biological Statistics* (2nd ed., pp. 70-76). Sparky House Publishing. http://www.biostathandbook.com/HandbookBioStatSecond.pdf

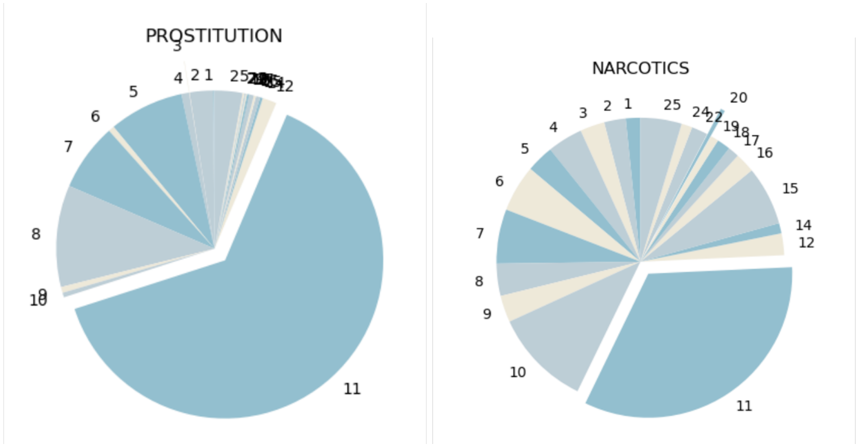
[7] (n.d.). *Community Area 2000 and 2010 Census Population Comparisons*. City of Chicago. https://www.chicago.gov/content/dam/city/depts/zlup/Zoning_Main_Page/Publications/Census_2010 _Community_Area_Profiles/Census_2010_and_2000_CA_Populations.pdf

[8] Naicker, N., de Jager, P., Naidoo, S., & Mathee, A. (2018). *Is There a Relationship between Lead Exposure and Aggressive Behavior in Shooters?*. International journal of environmental research and public health, 15(7), 1427. https://doi.org/10.3390/ijerph15071427

[9] (n.d.). *How TB Spreads*. Centers for Disease Control and Prevention. https://www.cdc.gov/tb/topic/basics/howtbspreads.htm

[10] "Most Dangerous Cities in America." PropertyClub, propertyclub.nyc/article/most-dangerous-cities-in-the-us. Accessed 30 June 2023.

9 Appendix



(a) Prostitution Arrests by District (b) Narcotics Arrests by District

Figure 1: Data Exploration shows up District 11 is high in crime.

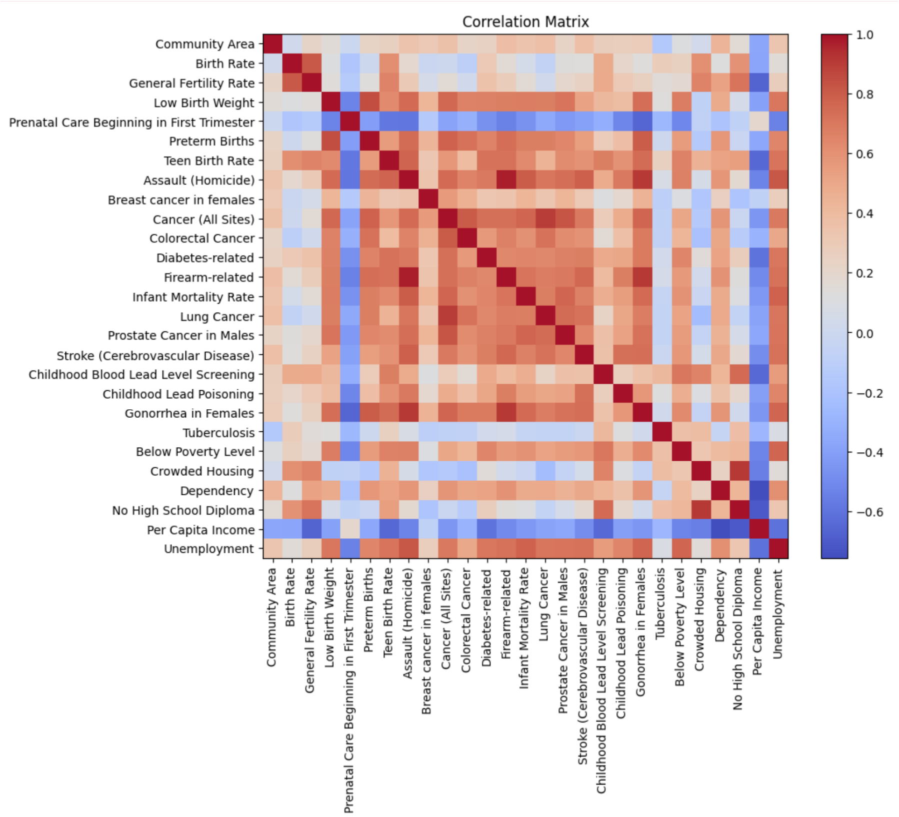
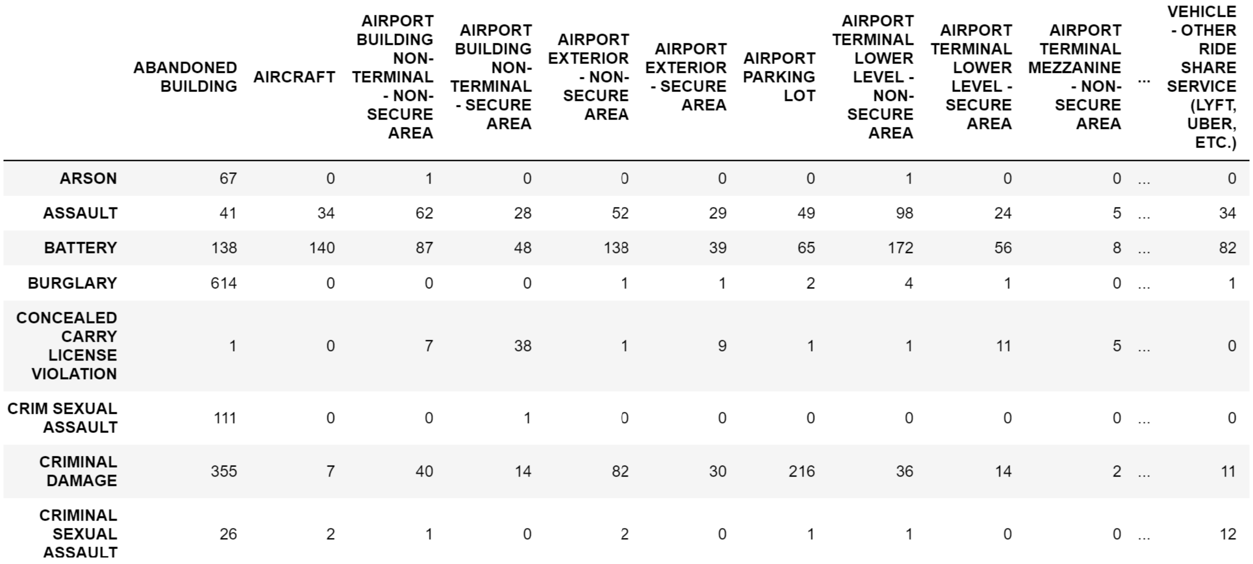
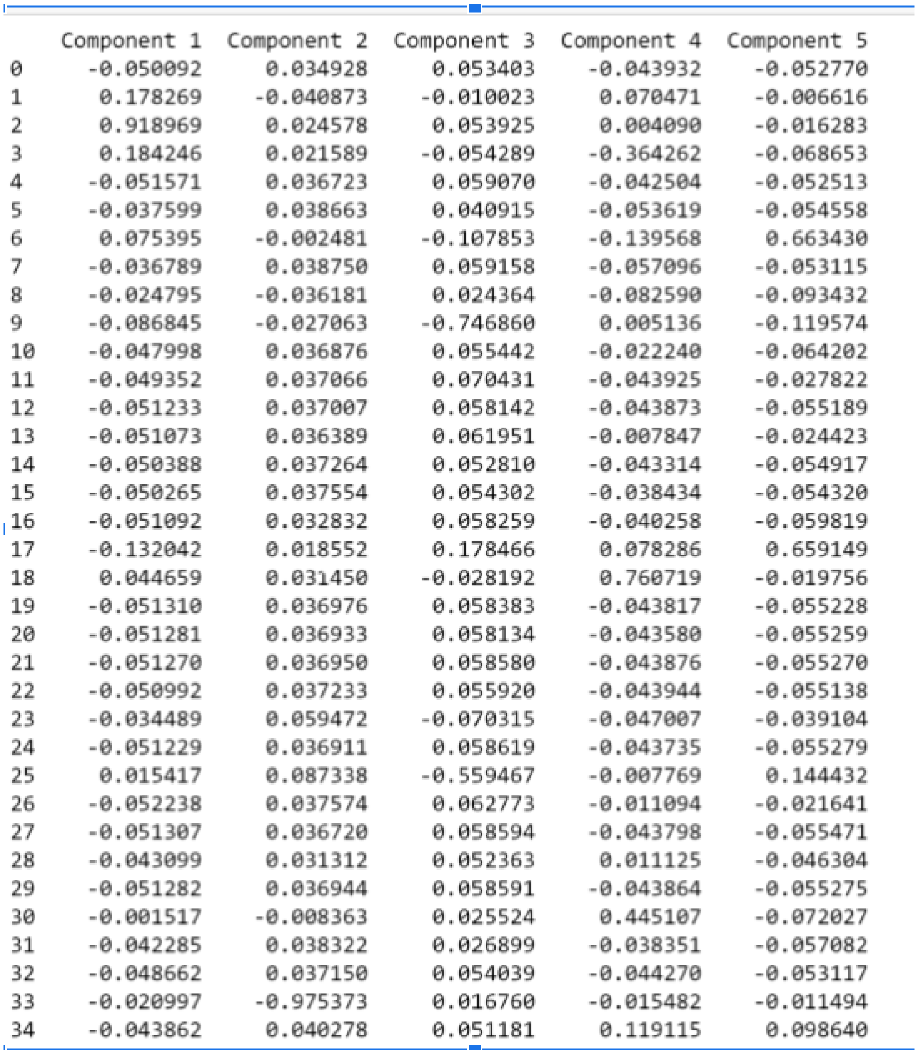


Figure 2: Correlation matrix of PBHLT



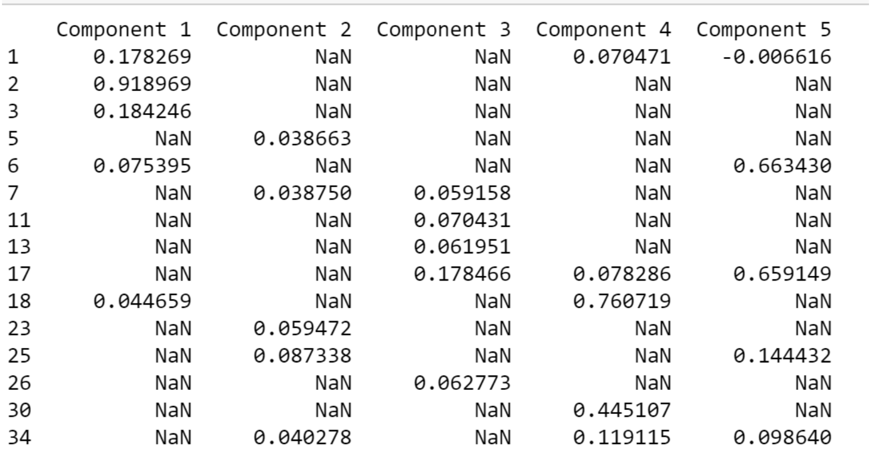
	ABANDONED BUILDING	AIRCRAFT	AIRPORT BUILDING NON-TERMINAL - NON-SECURE AREA	AIRPORT BUILDING NON-TERMINAL - SECURE AREA	AIRPORT EXTERIOR - NON-SECURE AREA	AIRPORT EXTERIOR - SECURE AREA	AIRPORT PARKING LOT	AIRPORT TERMINAL LOWER LEVEL - NON-SECURE AREA	AIRPORT TERMINAL LOWER LEVEL - SECURE AREA	AIRPORT TERMINAL MEZZANINE - NON-SECURE AREA	...	VEHICLE - OTHER RIDE SHARE SERVICE (LYFT, UBER, ETC.)
ARSON	67	0	1	0	0	0	0	1	0	0	...	0
ASSAULT	41	34	62	28	52	29	49	98	24	5	...	34
BATTERY	138	140	87	48	138	39	65	172	56	8	...	82
BURGLARY	614	0	0	0	1	1	2	4	1	0	...	1
CONCEALED CARRY LICENSE VIOLATION	1	0	7	38	1	9	1	1	11	5	...	0
CRIM SEXUAL ASSAULT	111	0	0	1	0	0	0	0	0	0	...	0
CRIMINAL DAMAGE	355	7	40	14	82	30	216	36	14	2	...	11
CRIMINAL SEXUAL ASSAULT	26	2	1	0	2	0	1	1	0	0	...	12

Figure 3: Occurence Matrix of CRIMES



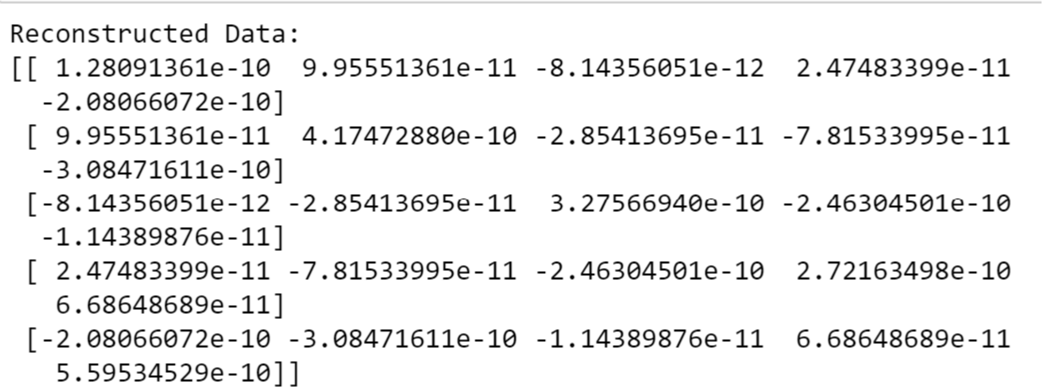
```
    Component 1  Component 2  Component 3  Component 4  Component 5
0     -0.050092     0.034928     0.053403    -0.043932    -0.052770
1      0.178269    -0.040873    -0.010023     0.070471    -0.006616
2      0.918969     0.024578     0.053925     0.004090    -0.016283
3      0.184246     0.021589    -0.054289    -0.364262    -0.068653
4     -0.051571     0.036723     0.059070    -0.042504    -0.052513
5     -0.037599     0.038663     0.040915    -0.053619    -0.054558
6      0.075395    -0.002481    -0.107853    -0.139568     0.663430
7     -0.036789     0.038750     0.059158    -0.057096    -0.053115
8     -0.024795    -0.036181     0.024364    -0.082590    -0.093432
9     -0.086845    -0.027063    -0.746860     0.005136    -0.119574
10    -0.047998     0.036876     0.055442    -0.022240    -0.064202
11    -0.049352     0.037066     0.070431    -0.043925    -0.027822
12    -0.051233     0.037007     0.058142    -0.043873    -0.055189
13    -0.051073     0.036389     0.061951    -0.007847    -0.024423
14    -0.050388     0.037264     0.052810    -0.043314    -0.054917
15    -0.050265     0.037554     0.054302    -0.038434    -0.054320
16    -0.051092     0.032832     0.058259    -0.040258    -0.059819
17    -0.132042     0.018552     0.178466     0.078286     0.659149
18     0.044659     0.031450    -0.028192     0.760719    -0.019756
19    -0.051310     0.036976     0.058383    -0.043817    -0.055228
20    -0.051281     0.036933     0.058134    -0.043580    -0.055259
21    -0.051270     0.036950     0.058580    -0.043876    -0.055270
22    -0.050992     0.037233     0.055920    -0.043944    -0.055138
23    -0.034489     0.059472    -0.070315    -0.047007    -0.039104
24    -0.051229     0.036911     0.058619    -0.043735    -0.055279
25     0.015417     0.087338    -0.559467    -0.007769     0.144432
26    -0.052238     0.037574     0.062773    -0.011094    -0.021641
27    -0.051307     0.036720     0.058594    -0.043798    -0.055471
28    -0.043099     0.031312     0.052363     0.011125    -0.046304
29    -0.051282     0.036944     0.058591    -0.043864    -0.055275
30    -0.001517    -0.008363     0.025524     0.445107    -0.072027
31    -0.042285     0.038322     0.026899    -0.038351    -0.057082
32    -0.048662     0.037150     0.054039    -0.044270    -0.053117
33    -0.020997    -0.975373     0.016760    -0.015482    -0.011494
34    -0.043862     0.040278     0.051181     0.119115     0.098640
```

Figure 4: Individual Components of CRIMES



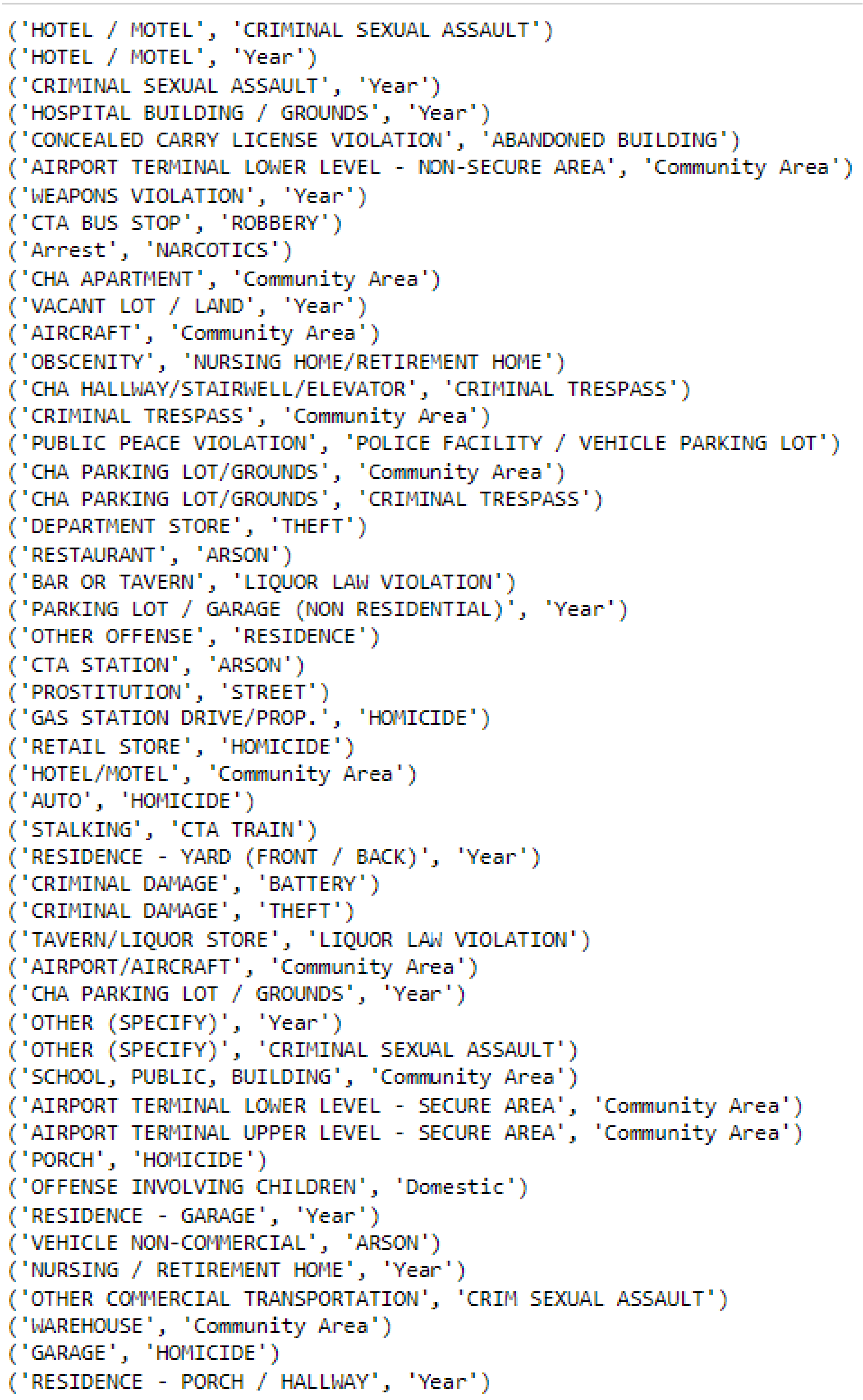
```
    Component 1  Component 2  Component 3  Component 4  Component 5
1      0.178269          NaN          NaN     0.070471    -0.006616
2      0.918969          NaN          NaN          NaN          NaN
3      0.184246          NaN          NaN          NaN          NaN
5           NaN     0.038663          NaN          NaN          NaN
6      0.075395          NaN          NaN          NaN     0.663430
7           NaN     0.038750     0.059158          NaN          NaN
11          NaN          NaN     0.070431          NaN          NaN
13          NaN          NaN     0.061951          NaN          NaN
17          NaN          NaN     0.178466     0.078286     0.659149
18     0.044659          NaN          NaN     0.760719          NaN
23          NaN     0.059472          NaN          NaN          NaN
25          NaN     0.087338          NaN          NaN     0.144432
26          NaN          NaN     0.062773          NaN          NaN
30          NaN          NaN          NaN     0.445107          NaN
34          NaN     0.040278          NaN     0.119115     0.098640
```

Figure 5: Largest 5 values for each component CRIMES



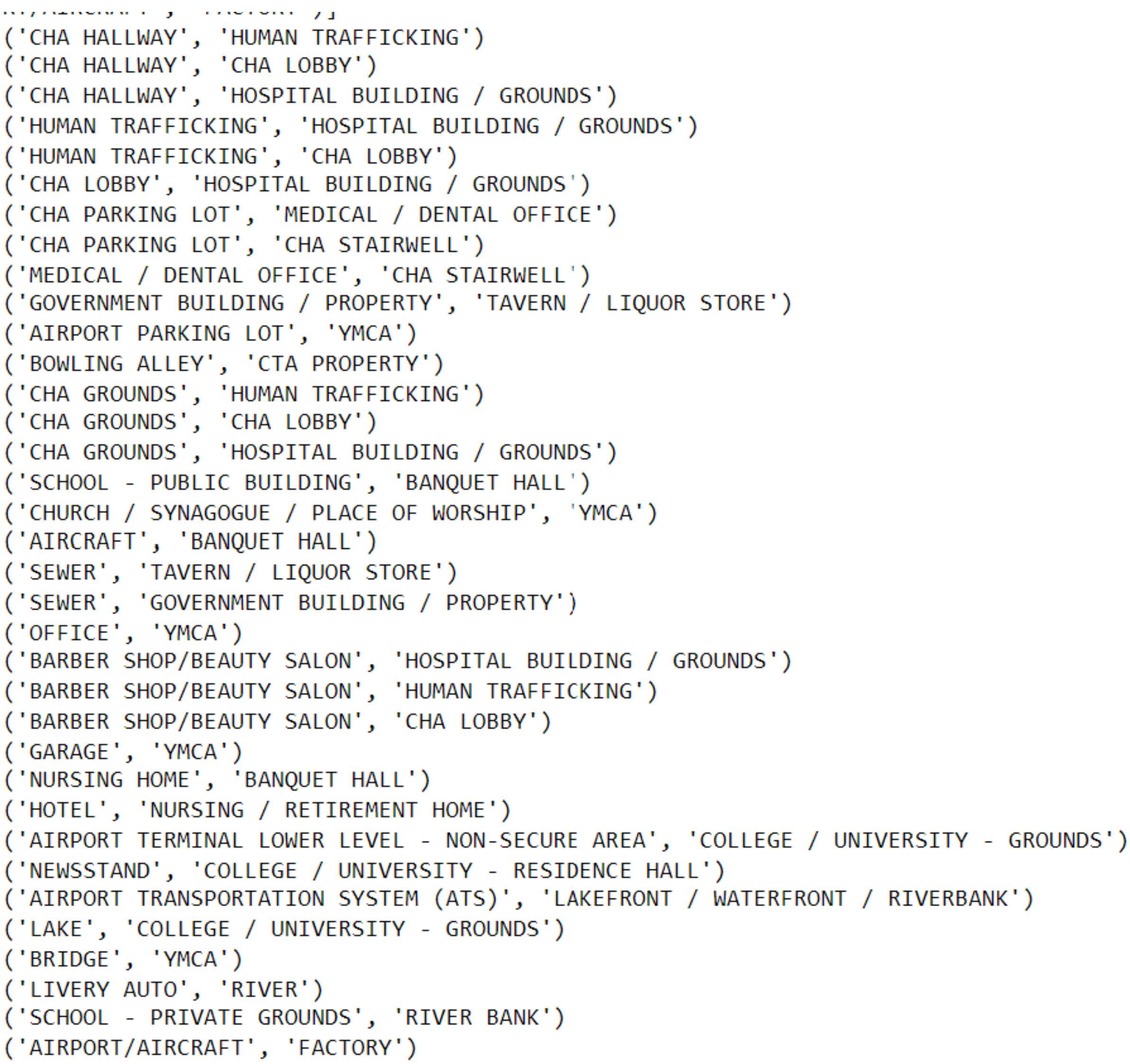
```
Reconstructed Data:
[[ 1.28091361e-10  9.95551361e-11 -8.14356051e-12  2.47483399e-11
  -2.08066072e-10]
 [ 9.95551361e-11  4.17472880e-10 -2.85413695e-11 -7.81533995e-11
  -3.08471611e-10]
 [-8.14356051e-12 -2.85413695e-11  3.27566940e-10 -2.46304501e-10
  -1.14389876e-11]
 [ 2.47483399e-11 -7.81533995e-11 -2.46304501e-10  2.72163498e-10
   6.68648689e-11]
 [-2.08066072e-10 -3.08471611e-10 -1.14389876e-11  6.68648689e-11
   5.59534529e-10]]
```

Figure 6: Reconstructed data weights for ICA on CRIMES



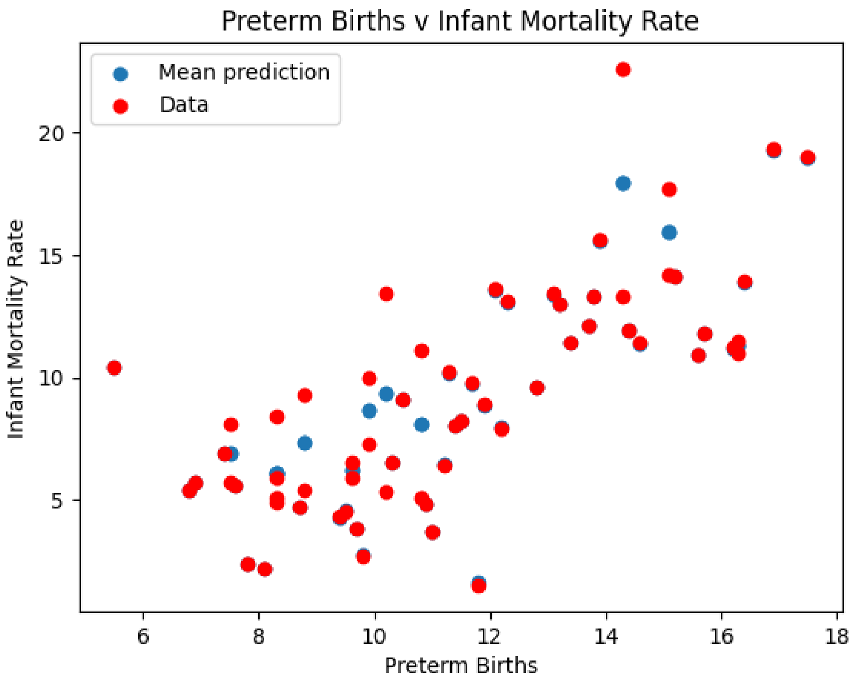
```
('HOTEL / MOTEL', 'CRIMINAL SEXUAL ASSAULT')
('HOTEL / MOTEL', 'Year')
('CRIMINAL SEXUAL ASSAULT', 'Year')
('HOSPITAL BUILDING / GROUNDS', 'Year')
('CONCEALED CARRY LICENSE VIOLATION', 'ABANDONED BUILDING')
('AIRPORT TERMINAL LOWER LEVEL - NON-SECURE AREA', 'Community Area')
('WEAPONS VIOLATION', 'Year')
('CTA BUS STOP', 'ROBBERY')
('Arrest', 'NARCOTICS')
('CHA APARTMENT', 'Community Area')
('VACANT LOT / LAND', 'Year')
('AIRCRAFT', 'Community Area')
('OBSCENITY', 'NURSING HOME/RETIREMENT HOME')
('CHA HALLWAY/STAIRWELL/ELEVATOR', 'CRIMINAL TRESPASS')
('CRIMINAL TRESPASS', 'Community Area')
('PUBLIC PEACE VIOLATION', 'POLICE FACILITY / VEHICLE PARKING LOT')
('CHA PARKING LOT/GROUNDS', 'Community Area')
('CHA PARKING LOT/GROUNDS', 'CRIMINAL TRESPASS')
('DEPARTMENT STORE', 'THEFT')
('RESTAURANT', 'ARSON')
('BAR OR TAVERN', 'LIQUOR LAW VIOLATION')
('PARKING LOT / GARAGE (NON RESIDENTIAL)', 'Year')
('OTHER OFFENSE', 'RESIDENCE')
('CTA STATION', 'ARSON')
('PROSTITUTION', 'STREET')
('GAS STATION DRIVE/PROP.', 'HOMICIDE')
('RETAIL STORE', 'HOMICIDE')
('HOTEL/MOTEL', 'Community Area')
('AUTO', 'HOMICIDE')
('STALKING', 'CTA TRAIN')
('RESIDENCE - YARD (FRONT / BACK)', 'Year')
('CRIMINAL DAMAGE', 'BATTERY')
('CRIMINAL DAMAGE', 'THEFT')
('TAVERN/LIQUOR STORE', 'LIQUOR LAW VIOLATION')
('AIRPORT/AIRCRAFT', 'Community Area')
('CHA PARKING LOT / GROUNDS', 'Year')
('OTHER (SPECIFY)', 'Year')
('OTHER (SPECIFY)', 'CRIMINAL SEXUAL ASSAULT')
('SCHOOL, PUBLIC, BUILDING', 'Community Area')
('AIRPORT TERMINAL LOWER LEVEL - SECURE AREA', 'Community Area')
('AIRPORT TERMINAL UPPER LEVEL - SECURE AREA', 'Community Area')
('PORCH', 'HOMICIDE')
('OFFENSE INVOLVING CHILDREN', 'Domestic')
('RESIDENCE - GARAGE', 'Year')
('VEHICLE NON-COMMERCIAL', 'ARSON')
('NURSING / RETIREMENT HOME', 'Year')
('OTHER COMMERCIAL TRANSPORTATION', 'CRIM SEXUAL ASSAULT')
('WAREHOUSE', 'Community Area')
('GARAGE', 'HOMICIDE')
('RESIDENCE - PORCH / HALLWAY', 'Year')
```

Figure 7: PC Algorithm on CRIMES



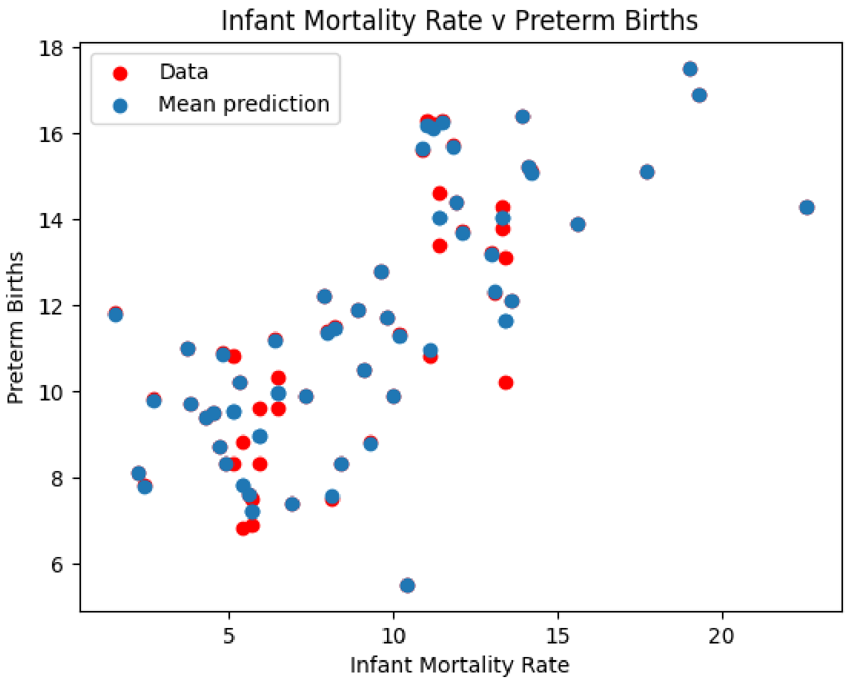
```
RT/AIRCRAFT', 'FACTORY')]
('CHA HALLWAY', 'HUMAN TRAFFICKING')
('CHA HALLWAY', 'CHA LOBBY')
('CHA HALLWAY', 'HOSPITAL BUILDING / GROUNDS')
('HUMAN TRAFFICKING', 'HOSPITAL BUILDING / GROUNDS')
('HUMAN TRAFFICKING', 'CHA LOBBY')
('CHA LOBBY', 'HOSPITAL BUILDING / GROUNDS')
('CHA PARKING LOT', 'MEDICAL / DENTAL OFFICE')
('CHA PARKING LOT', 'CHA STAIRWELL')
('MEDICAL / DENTAL OFFICE', 'CHA STAIRWELL')
('GOVERNMENT BUILDING / PROPERTY', 'TAVERN / LIQUOR STORE')
('AIRPORT PARKING LOT', 'YMCA')
('BOWLING ALLEY', 'CTA PROPERTY')
('CHA GROUNDS', 'HUMAN TRAFFICKING')
('CHA GROUNDS', 'CHA LOBBY')
('CHA GROUNDS', 'HOSPITAL BUILDING / GROUNDS')
('SCHOOL - PUBLIC BUILDING', 'BANQUET HALL')
('CHURCH / SYNAGOGUE / PLACE OF WORSHIP', 'YMCA')
('AIRCRAFT', 'BANQUET HALL')
('SEWER', 'TAVERN / LIQUOR STORE')
('SEWER', 'GOVERNMENT BUILDING / PROPERTY')
('OFFICE', 'YMCA')
('BARBER SHOP/BEAUTY SALON', 'HOSPITAL BUILDING / GROUNDS')
('BARBER SHOP/BEAUTY SALON', 'HUMAN TRAFFICKING')
('BARBER SHOP/BEAUTY SALON', 'CHA LOBBY')
('GARAGE', 'YMCA')
('NURSING HOME', 'BANQUET HALL')
('HOTEL', 'NURSING / RETIREMENT HOME')
('AIRPORT TERMINAL LOWER LEVEL - NON-SECURE AREA', 'COLLEGE / UNIVERSITY - GROUNDS')
('NEWSSTAND', 'COLLEGE / UNIVERSITY - RESIDENCE HALL')
('AIRPORT TRANSPORTATION SYSTEM (ATS)', 'LAKEFRONT / WATERFRONT / RIVERBANK')
('LAKE', 'COLLEGE / UNIVERSITY - GROUNDS')
('BRIDGE', 'YMCA')
('LIVERY AUTO', 'RIVER')
('SCHOOL - PRIVATE GROUNDS', 'RIVER BANK')
('AIRPORT/AIRCRAFT', 'FACTORY')
```

Figure 8: PC Algorithm on CRIMES and PBHLT



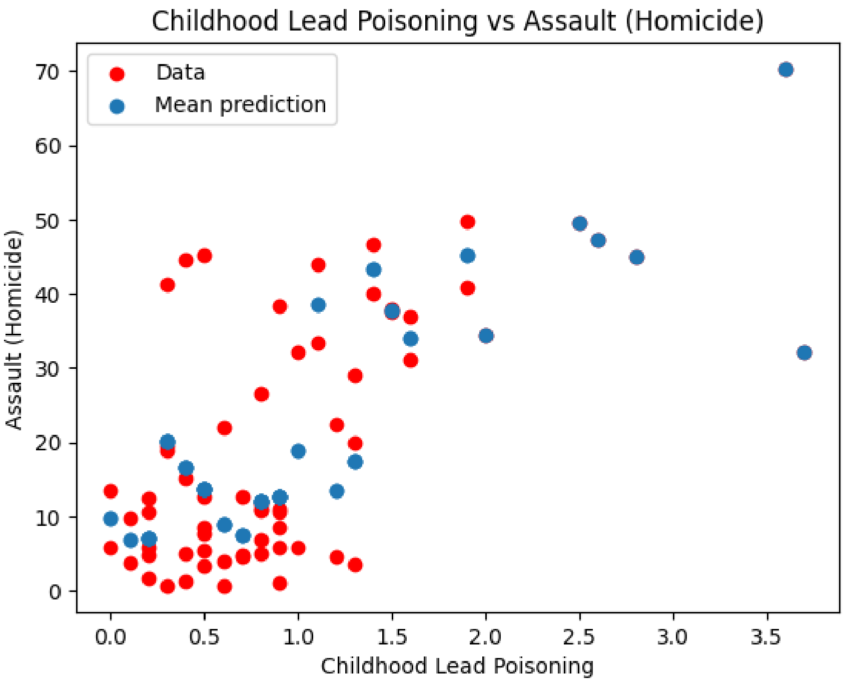
[H]

Figure 9: Preterm Births to Infant Mortality Rate



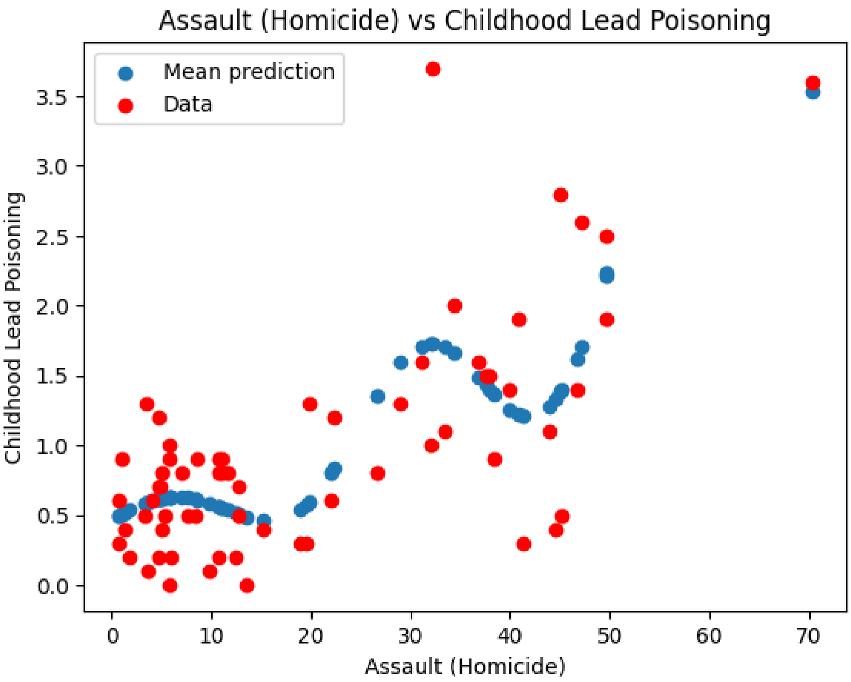
[H]

Figure 10: Infant Mortality Rate to Preterm Births



[H]

Figure 11: Childhood Lead Poisoning to Assault (Homicide)



[H]

Figure 12: Assault (Homicide) to Childhood Lead Poisoning

[H]

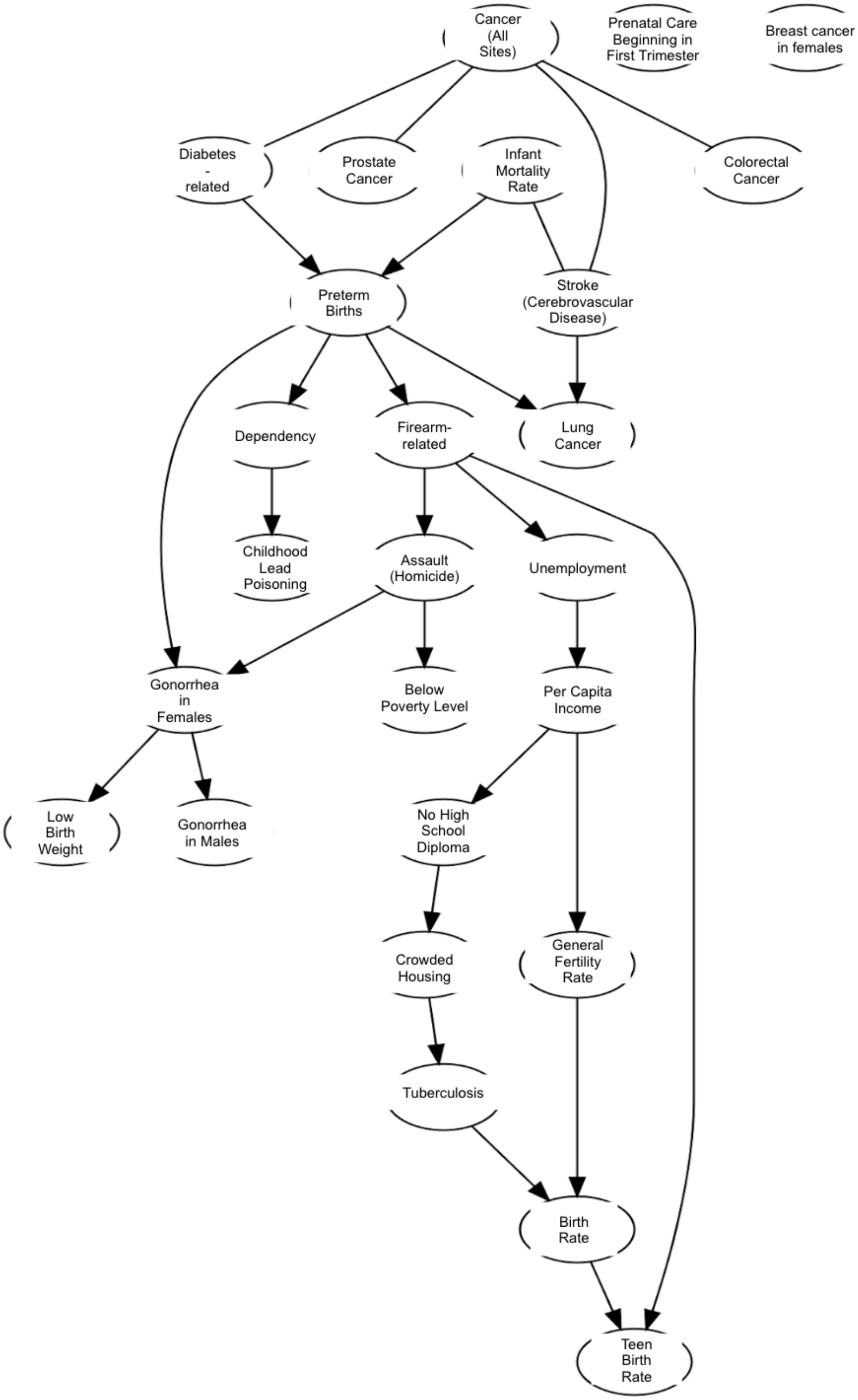


Figure 13: GES Algorithm on PBHLT.