University of Illinois at Chicago

Digging Deeper Into Chicago Gun Crime:

Analyzing the Socioeconomic Contributors to Chicago Gun Violence

Mersim Rizmani

Honors College Capstone Project

Professor Gonzalo Bello Lander

3 December 2021

Table of Contents

I. Introduction and Problem Selection

Abstract

The crime rate in the city of Chicago, especially violent crime, is higher than the United States average. Over the years, the crime rates have drastically increased, with 2021 being the worst year since 1996. This research dives into different socioeconomic and public health indicators from 2005-2011, including unemployment, per capita income, housing, and poverty. Furthermore, this research analyzes whether or not these indicators have a correlation with the violent crime rates in different community areas across Chicago. Hypothesis tests were conducted and plots were created to highlight the key differences between low risk and high risk areas. Additionally, linear regression and classification models were built in an attempt to predict the total shootings in a particular community area and to classify these areas as low risk or high risk based on these socioeconomic variables. Data exploration found that there was a statistically significant difference between low risk and high risk community areas across many different indicators. Several of the classification models were able to accurately classify low risk and high risk community areas using the socioeconomic variables. Based on these models, it was suggested that there exists a relationship between the total shootings in each community area and these socioeconomic indicators.

Motivation

As a student in the city of Chicago, this topic hits close to home. A recent article from NBC states that, “according to newly-released crime statistics for the month of July, murders in the city were nearly the same as the number reported last year, but shootings increased by 15% and the number of people shot in the city rose by nearly 10% year-over-year.” Gun violence in the city of Chicago has been on a significant rise, and each year it seems to get worse.

Data Science Solution

Formulate Problem: Using different socioeconomic variables from the community areas in Chicago, predict the total shootings for each community area, and classify each community area as high risk or low risk according to those same variables.

Collect Data: Collected crime data from the city of Chicago, specifically crimes that involved shootings, as well as socioeconomic information by community area.

Prepare Data: Filtered the shootings dataset to only include data from 2005-2011, mapped each shooting incident to a community area, merged the shootings dataset with the socioeconomic data, removed inconsistent or missing values, and defined high risk and low risk community areas.

Explore Data: Computed average socioeconomic statistics for low risk and high risk community areas, determined whether the differences are statistically significant using hypothesis testing, visualized the differences with boxplots, and plotted heatmaps for each statistic by community area.

Build Models: Built regression models to predict total shootings using socioeconomic variables, and classification models to classify low risk and high risk community areas.

Evaluate Results: Computed accuracy of predictions and classifications, and visualized predictions using heatmaps. Also computed several other evaluation metrics for classification models such as precision, recall, and F1 score.

II. Data Collection

All of the data used in this research was collected from the City of Chicago Open Data Portal.

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2
This dataset reflects reported incidents of crime that occurred in the City of Chicago from 2001 to present, minus the most recent seven days.

https://data.cityofchicago.org/Public-Safety/Shootings/vqmv-zqjm
This dataset is a subset of the one above. It contains all of the crimes in which a firearm was involved, i.e. a shooting.

https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu
This dataset contains a selection of 27 indicators of public health significance by Chicago community area, with the most updated information available. The indicators are rates, percents, or other measures related to natality, mortality, infectious disease, lead poisoning, and economic status.

https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6
This dataset contains geodata for the boundaries and areas of all 77 community areas in Chicago.

Below are the public health / socioeconomic features used in the exploratory data analysis and data modeling:

Category	Measure	Meaning/Units	Years
Economic	Per Capita Income	Per capita income in 2011 inflation-adjusted dollars	2007-2011
	Unemployment	Percent of persons in labor force aged 16 years and older unemployed	2007-2011
	Below Poverty Level	Percent of households below poverty level	2007-2011
	Crowded Housing	Percent of occupied housing units considered crowded	2007-2011
	No High School Diploma	Percent of persons aged 25 years and older with no high school diploma	2007-2011
	Dependency	Percent of persons aged less than 16 or more than 64 years	2007-2011
Natality	Birth Rate	Births Per 1,000 persons	2009
	Teen Birth Rate	Births Per 1,000 females aged 15-19	2009
	Preterm Births	Percent of live births that are preterm	2009
	Infant Mortality Rate	Infant mortality Per 1,000 live births	2005-2009
	Prenatal care beginning in first trimester	Percent of females delivering a live birth that received prenatal care beginning in first trimester	2009
Mortality / Lead	Cancer (all sites) [Deaths]	Deaths from Cancer Per 100,000 persons (age adjusted)	2005-2009
	Diabetes-related [Deaths]	Diabetes-related Deaths Per 100,000 persons (age adjusted)	2005-2009
	Stroke (cerebrovascular disease) [Deaths]	Stroke Deaths Per 100,000 persons (age adjusted)	2005-2009
	Childhood lead poisoning	Childhood lead poisoning incidents Per 100	2011
	Childhood blood lead level screening	Number of children that received a blood lead level screening per 1,000 children aged 0-6 years	2011

III. Data Preparation

Preparing the Shootings Dataset

Several tasks needed to be completed to clean and prepare the data for analysis. First, the Shootings dataset needed to be filtered to only include incidents from the years 2005-2011 to match the socioeconomic data. It's worth noting that not all of the variables from the public health indicators dataset spanned across the full range of 2005-2011. There were several instances where a particular variable only spanned across a subset of that range, like per capita income, for example, which ranged from 2007-2011. This was one of the challenges that was come across in this research, however, there was a difficulty in finding good datasets to depict several different socioeconomic variables. So with the desire to use this dataset, this research was conducted under the assumption that there wouldn't be a major statistical difference if the entire 2005-2011 range was used for total shootings.

The next major challenge in preparing the Shootings dataset was to add the community area names to each incident. This was needed to be able to merge this dataset with the others. The Shootings dataset was missing the community area names, so the geodata in the Boundaries dataset was used in conjunction with the longitude and latitude data from the Shootings dataset to map each incident to a community area. First, the community areas boundaries were extracted from the geospatial data (boundaries.geojson file) and put into a new dataframe. Next, an algorithm was devised to map each incident in the Shootings dataset using the longitude and latitude of each incident, to a community area based on geometric location. The algorithm is shown below:

```
# Map coordinates of shootings to community areas

def check_comm_area(lat, long):
    for ind in community_areas_boundaries.index:
        polygon = shape(community_areas_boundaries['geometry'][ind])
        point = Point(long, lat)
        if polygon.contains(point):
            return community_areas_boundaries['community'][ind]

community_areas = []

for ind in shootings.index:
    community_areas.append(check_comm_area(shootings['Latitude'][ind], shootings['Longitude'][ind]))
```

The community_areas array was then appended to the Shootings dataset. Once that was done, the Shootings dataset was grouped by the community area, with the grouping function size(), to get a dataset that contained the community area names and the total shootings for each community area. All columns except the community area name and the total shootings were dropped as they were irrelevant for this research.

The Shootings dataset was then merged with the Boundaries dataset to create a new dataset called community_areas_shootings. The purpose of this was to be able to plot heatmaps of the different variables in the dataset. Then finally, the community_areas_shootings dataset was merged with the public health indicators to create the merged dataset.

Preparing the Merged Dataset

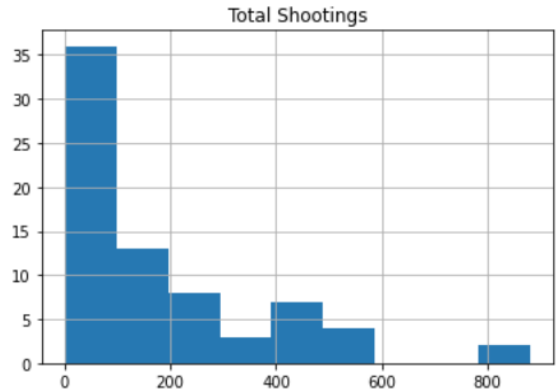
The first task in the merged data set was to drop irrelevant or possibly biased columns. The Assaults (homicide) and Firearm-related variables were dropped from the dataset as they would drastically alter any predictions involving Total Shootings. Next, missing or NaN values were identified for the entire dataset.

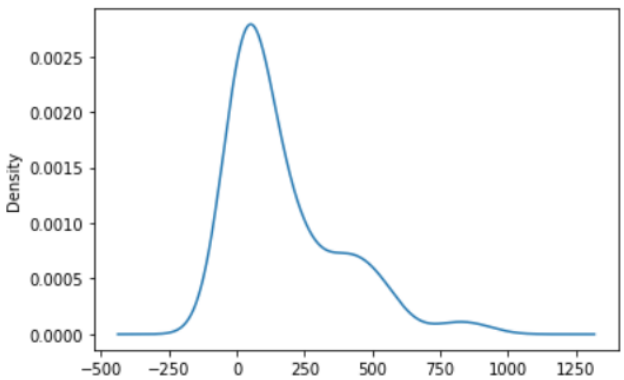
```
Count total NaN at each column in a DataFrame :

 community                                   0
area                                         0
shape_area                                   0
perimeter                                    0
area_num_1                                   0
area_numbe                                   0
comarea_id                                   0
comarea                                      0
shape_len                                    0
geometry                                     0
Total Shootings                              0
Community Area                               0
Birth Rate                                   0
General Fertility Rate                       0
Low Birth Weight                             0
Prenatal Care Beginning in First Trimester   0
Preterm Births                               0
Teen Birth Rate                              0
Breast cancer in females                     0
Cancer (All Sites)                           0
Colorectal Cancer                            0
Diabetes-related                             0
Infant Mortality Rate                        0
Lung Cancer                                  0
Prostate Cancer in Males                     0
Stroke (Cerebrovascular Disease)             0
Childhood Blood Lead Level Screening         1
Childhood Lead Poisoning                     1
Gonorrhea in Females                         8
Gonorrhea in Males                           0
Tuberculosis                                 0
Below Poverty Level                          0
Crowded Housing                              0
Dependency                                   0
No High School Diploma                       0
Per Capita Income                            0
Unemployment                                 0
Risk                                         0
Risk (int)                                   0
dtype: int64
```

```
community                                  object
area                                       object
shape_area                                 object
perimeter                                  object
area_num_1                                 object
area_numbe                                 object
comarea_id                                 object
comarea                                    object
shape_len                                  object
geometry                                 geometry
Total Shootings                             int64
Community Area                              int64
Birth Rate                                float64
General Fertility Rate                    float64
Low Birth Weight                          float64
Prenatal Care Beginning in First Trimester float64
Preterm Births                            float64
Teen Birth Rate                           float64
Breast cancer in females                  float64
Cancer (All Sites)                        float64
Colorectal Cancer                         float64
Diabetes-related                          float64
Infant Mortality Rate                     float64
Lung Cancer                               float64
Prostate Cancer in Males                  float64
Stroke (Cerebrovascular Disease)          float64
Childhood Blood Lead Level Screening      float64
Childhood Lead Poisoning                  float64
Gonorrhea in Females                      float64
Gonorrhea in Males                         object
Tuberculosis                              float64
Below Poverty Level                       float64
Crowded Housing                           float64
Dependency                                float64
No High School Diploma                    float64
Per Capita Income                           int64
Unemployment                              float64
Risk                                       object
Risk (int)                                  int64
Predicted                                   int64
dtype: object
```

The two NaN values found for Childhood Blood Lead Level Screening and Childhood Lead Poisoning were replaced with the mean values for each variable respectively. The Gonorrhea in Females variable was removed from analysis due to it missing 8 values which account for over 10% of the dataset. Because of this, it only made sense to remove Gonorrhea in Males from analysis as well.

A histogram was generated to visualize the number of community areas that fall within a specific range for total shootings, and it was determined that 400 total shootings was an adequate split point for defining low risk and high risk community areas. A classification of “low” or “high” risk was added to the dataset using this split point. This classification will be used later on in building the classification models.





IV. Data Exploration

Hypothesis Testing

Prior to beginning exploratory data analysis, the dataset was partitioned into low risk and high risk tables based on the split point defined in data preparation (low risk < 400 total shootings, and high risk >= 400 total shootings). Using these partitioned datasets, several hypothesis tests were conducted to determine if the difference between low risk and high risk community areas for each socioeconomic indicator was statistically significant.

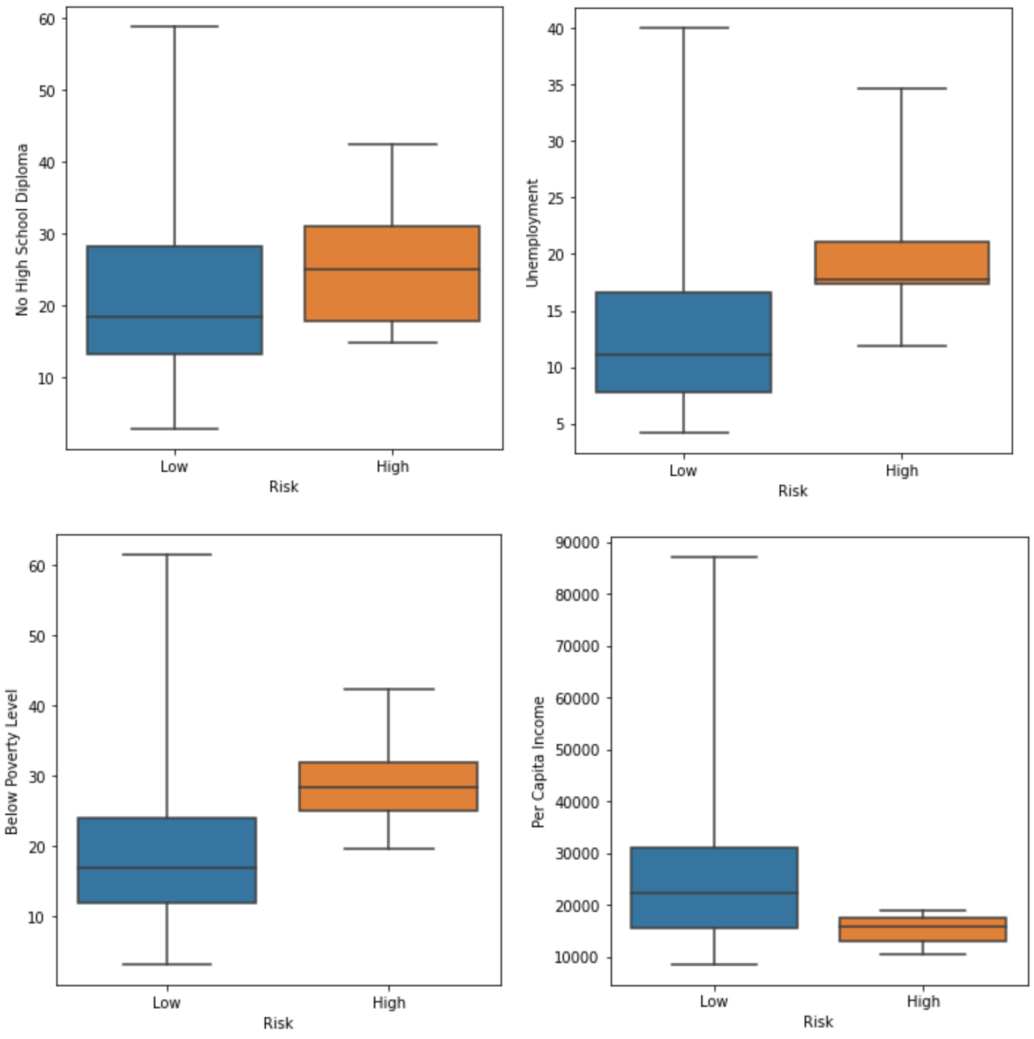
Two-sample hypothesis testing was conducted using an unpaired ***t***-test for population means. The null hypothesis in all the tests was that the means were equal, and the alternative hypothesis was that they were not equal. The table below shows the means for each socioeconomic statistic for both low risk and high risk community areas, the p-values for each test with a significance level of 0.05, and the test conclusions. Below are the results from all the hypothesis tests conducted for each variable:

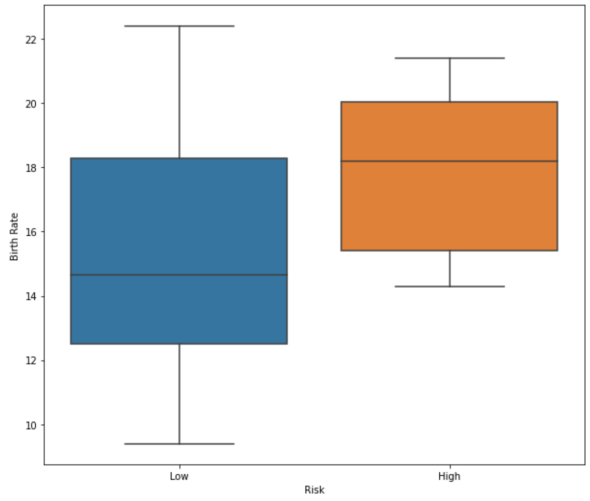
Variable	Low Risk Mean	High Risk Mean	Difference	p-value	Test Conclusion
Per Capita Income	26377.73	15208.64	11169.09	2.52e-06	Reject H_0
Unemployment	12.60	19.50	6.90	0.0045	Reject H_0
No High School Diploma	21.38	25.74	4.36	0.187	Fail to Reject H_0
Below Poverty Level	19.56	28.75	9.19	0.000679	Reject H_0
Crowded Housing	4.92	5.76	0.839	0.460	Fail to Reject H_0
Dependency	35.09	40.43	5.34	2.24e-05	Reject H_0
Cancer (all sites) Deaths	188.56	240.95	52.39	1.92e-05	Reject H_0
Diabetes-related Deaths	69.49	91.24	21.74	2.71e-05	Reject H_0
Stroke Deaths	44.91	59.01	14.10	5.63e-06	Reject H_0
Childhood Lead Poisoning	0.75	1.7	0.95	0.00034	Reject H_0
Childhood Blood Lead Level Screening	385.18	454.80	69.62	0.004	Reject H_0
Birth Rate	15.44	17.95	2.51	0.01	Reject H_0
Teen Birth Rate	46.12	82.52	36.39	2.36e-05	Reject H_0
Infant Mortality Rate	8.29	11.92	3.63	0.0004	Reject H_0
Preterm Births	11.02	13.4	2.38	0.0006	Reject H_0
Prenatal Care	77.59	71.79	5.79	0.0004	Reject H_0

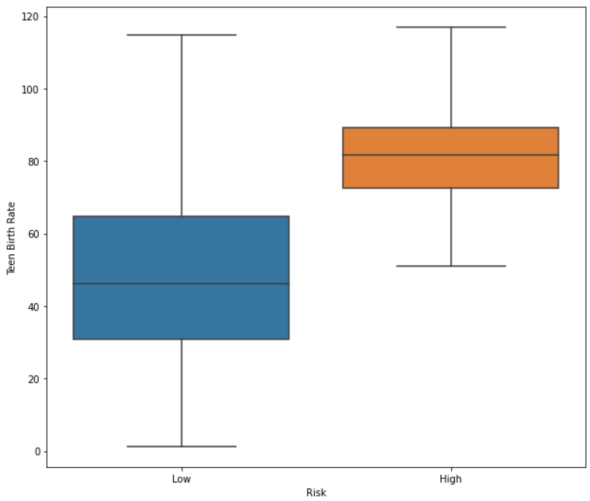
The tests highlighted in green above indicate the variables where the difference between high risk and low risk community areas was determined to be statistically significant. There was sufficient evidence to indicate that the differences in Per Capita Income, Unemployment, Below Poverty Level, Dependency, Cancer (all sites) Deaths, Diabetes-related Deaths, Stroke Deaths, Childhood Lead Poisoning, Childhood Blood Lead Level Screening, Birth Rate, Teen Birth Rate, Infant Mortality Rate, Preterm Births, and Prenatal Care, were statistically significant. There were only two instances where there was not sufficient evidence to assert that the differences were statistically significant between low risk and high risk communities, and that was for No High School Diploma, and Crowded Housing.

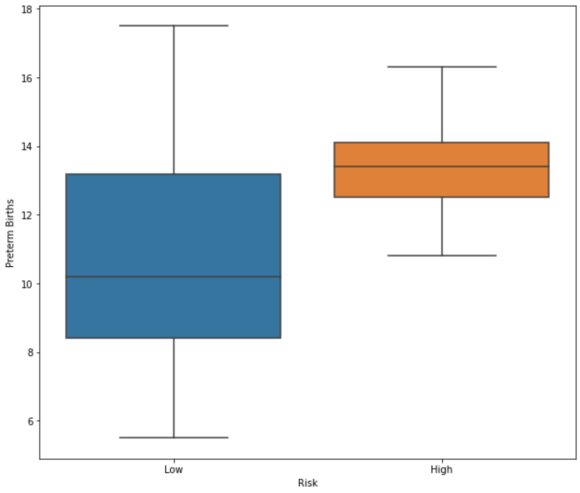
Comparing Low Risk and High Risk Communities with Box Plots

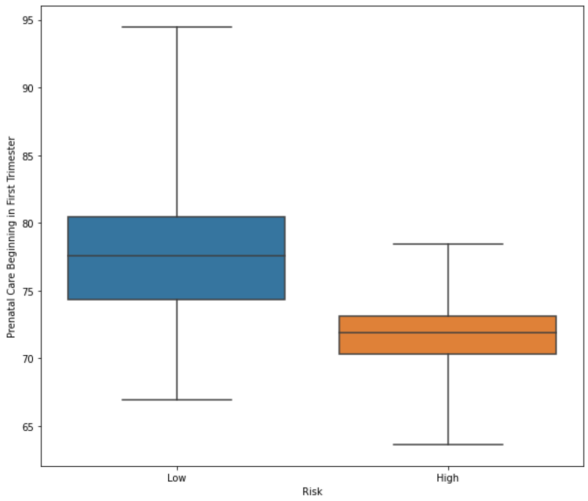
To better visualize these statistical differences between the low risk and high risk community areas, box plots were generated. Box plots depict the distribution of data values using boxes and whiskers, with marks for the first quartile, median, and third quartile, as well as the maximum and minimum. The figures below show box plots depicting the differences in distribution between low risk and high risk community areas for income, unemployment, poverty and more. Low risk community areas generally have higher income and more prenatal care, while high risk community areas have higher unemployment, more poverty, less education, and more deaths from cancer, diabetes and strokes.

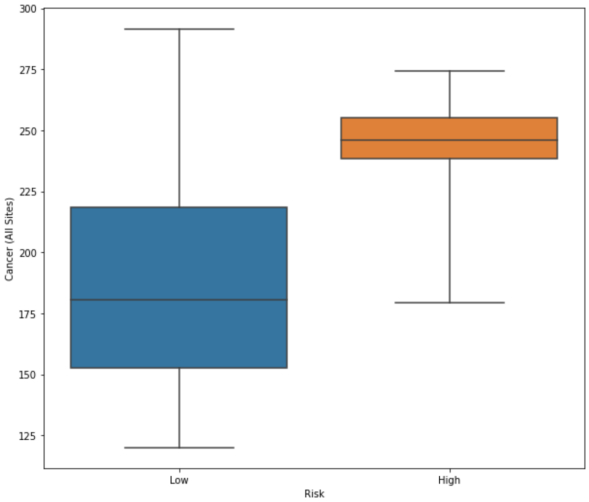


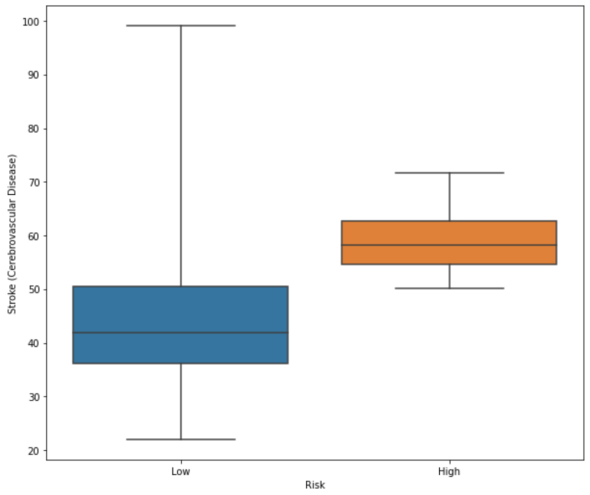

Birth Rate
22
20
18
16
14
12
10
Low
High
Risk

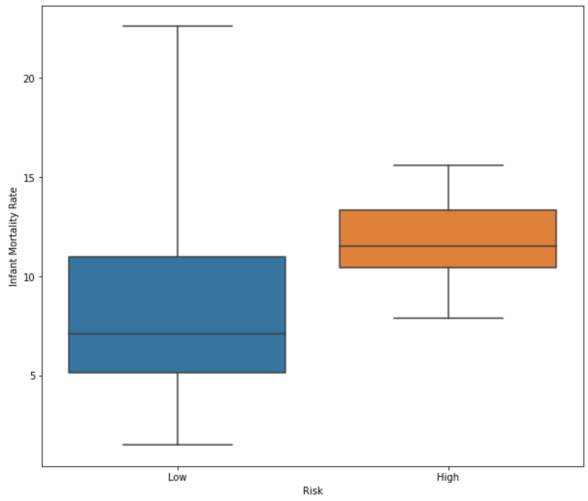

Teen Birth Rate
120
100
80
60
40
20
0
Low
High
Risk

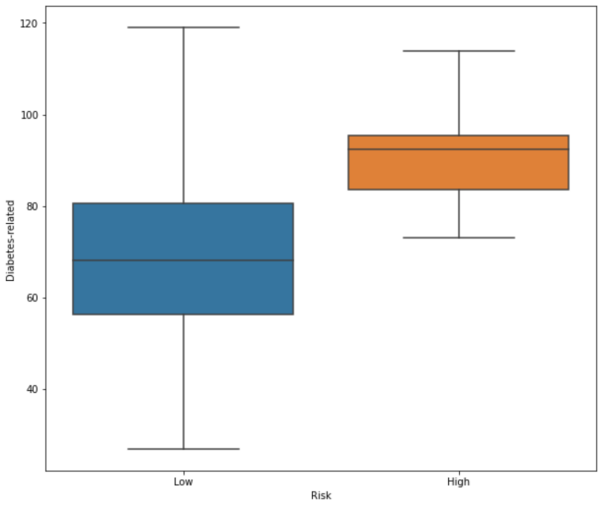

Preterm Births
18
16
14
12
10
8
6
Low
High
Risk


Prenatal Care Beginning in First Trimester
95
90
85
80
75
70
65
Low
High
Risk


Cancer (All Sites)
300
275
250
225
200
175
150
125
Low
High
Risk

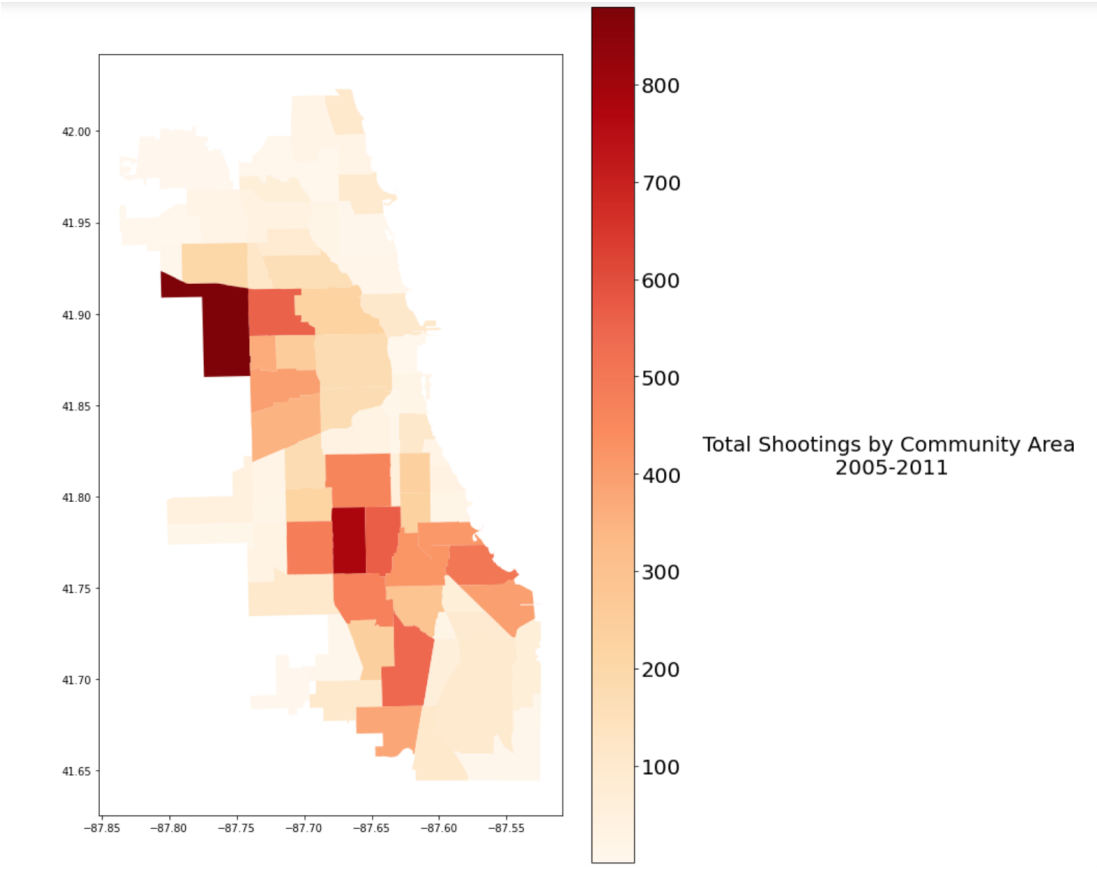

Stroke (Cerebrovascular Disease)
100
90
80
70
60
50
40
30
20
Low
High
Risk

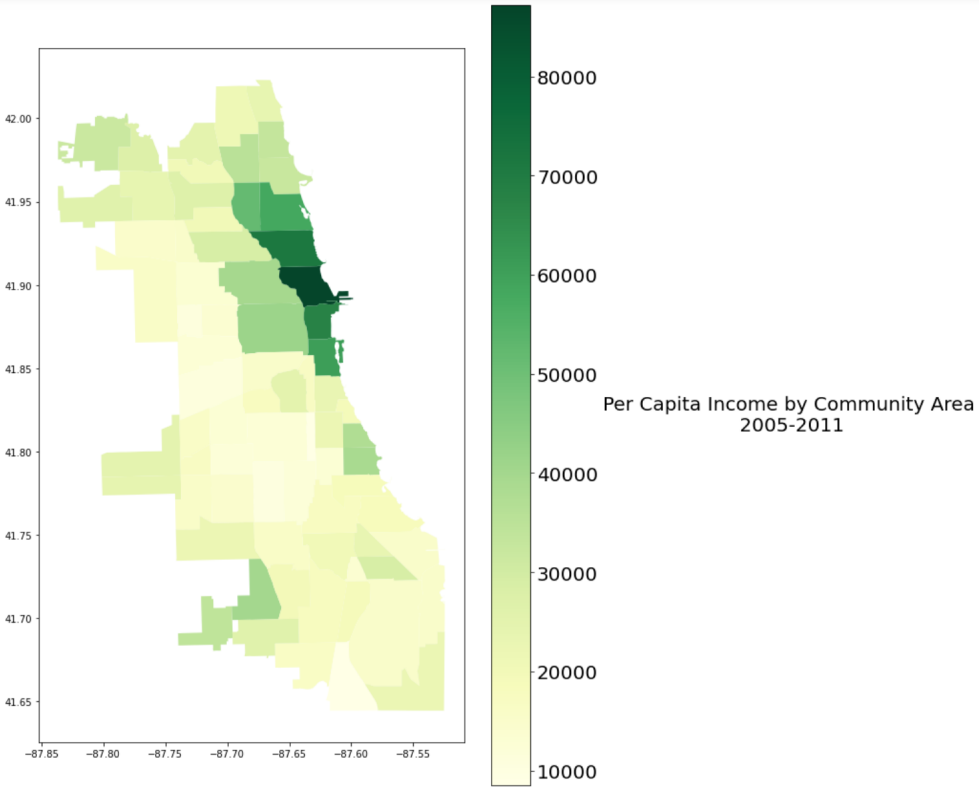


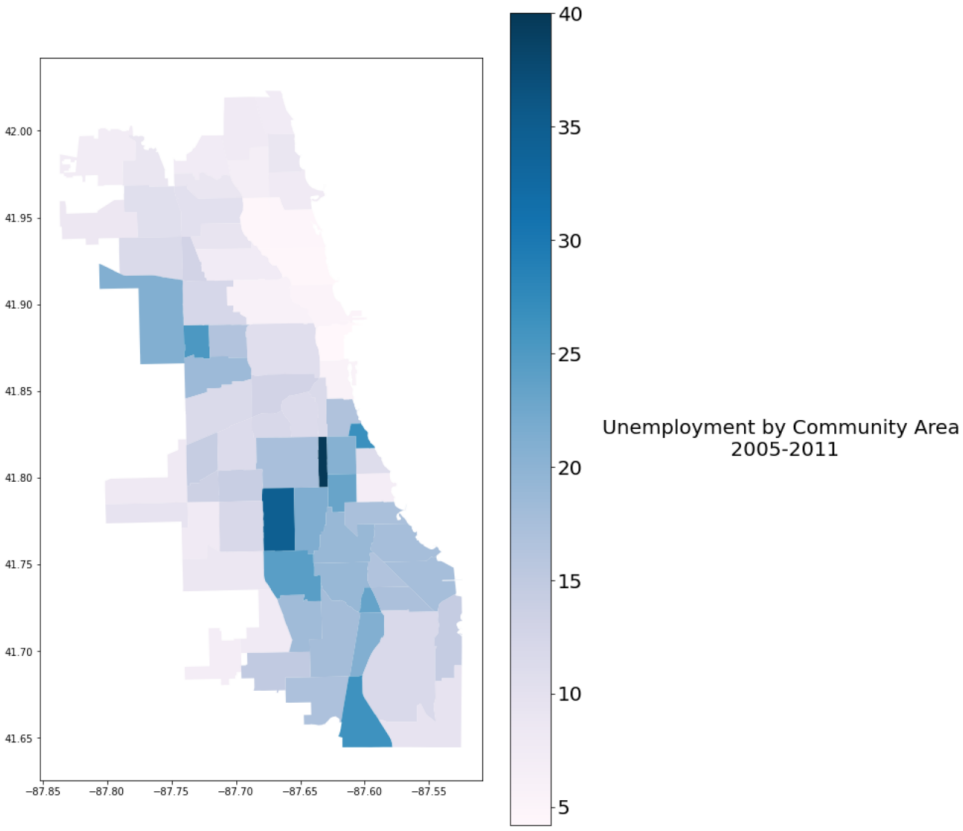


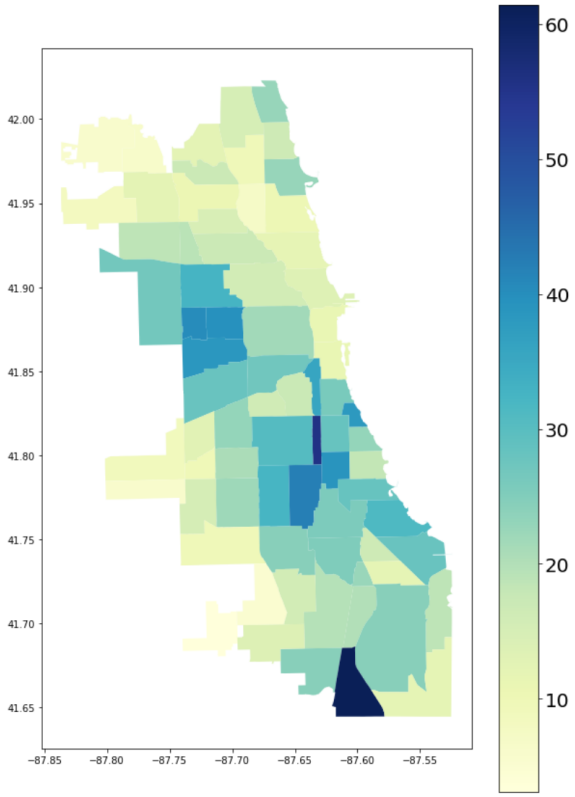
Visualizing Community Areas Differences with Heatmaps

In addition to the distribution comparison between low risk and high risk community areas for the socioeconomic variables, heatmaps were plotted to show the magnitude of each variable for each community area. The figures below feature heatmaps for all of the variables analyzed from the boxplots above, and some additional ones as well. The Total Shootings heatmap is shown first at the top. As shown from the heatmaps, there is a noticeable relationship between Total Shootings and Per Capita Income, Unemployment, Below Poverty Level, Teen Birth Rate, and more.

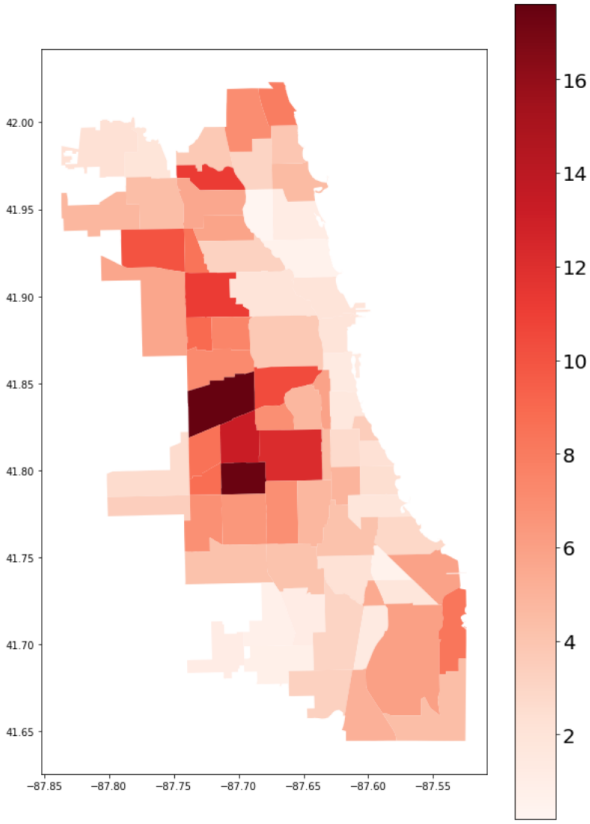



42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
−87.85
−87.80
−87.75
−87.70
−87.65
−87.60
−87.55
80000
70000
60000
50000
40000
30000
20000
10000
Per Capita Income by Community Area
2005-2011

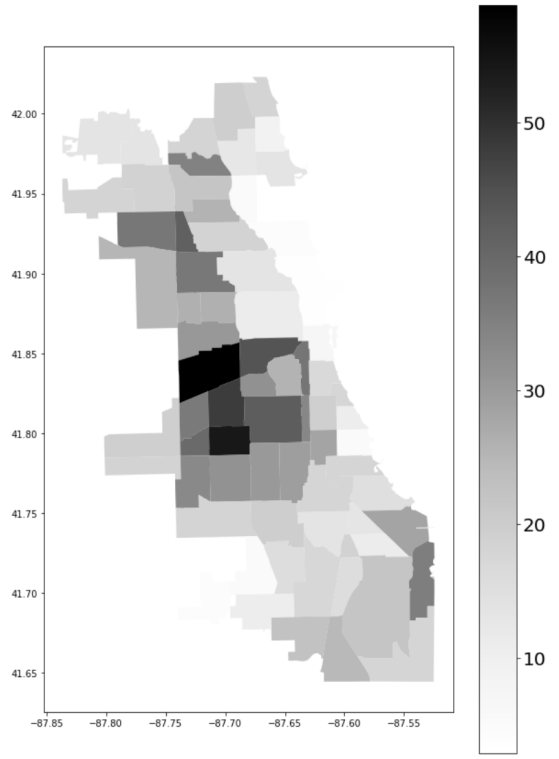

42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
−87.85
−87.80
−87.75
−87.70
−87.65
−87.60
−87.55
40
35
30
25
20
15
10
5
Unemployment by Community Area
2005-2011



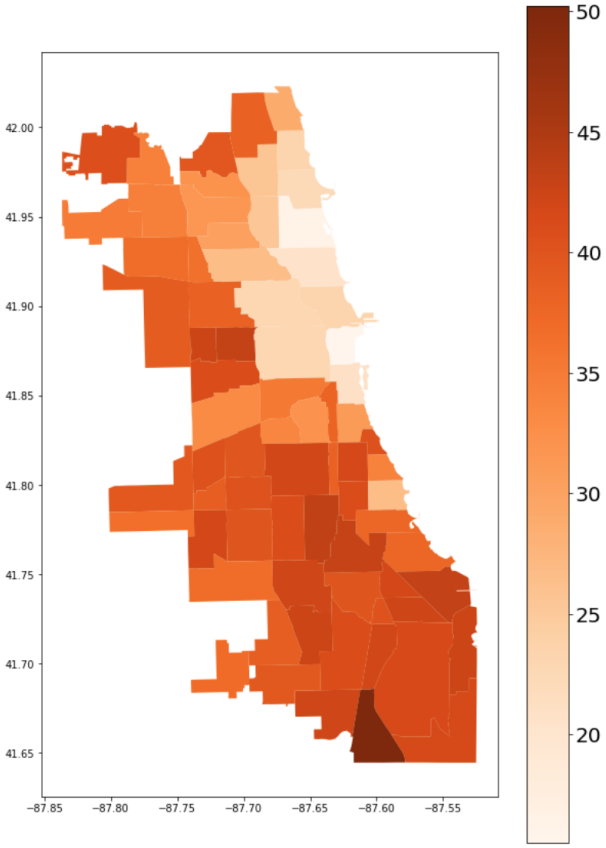
Households Below Poverty Level (%)
by Community Area 2005-2011



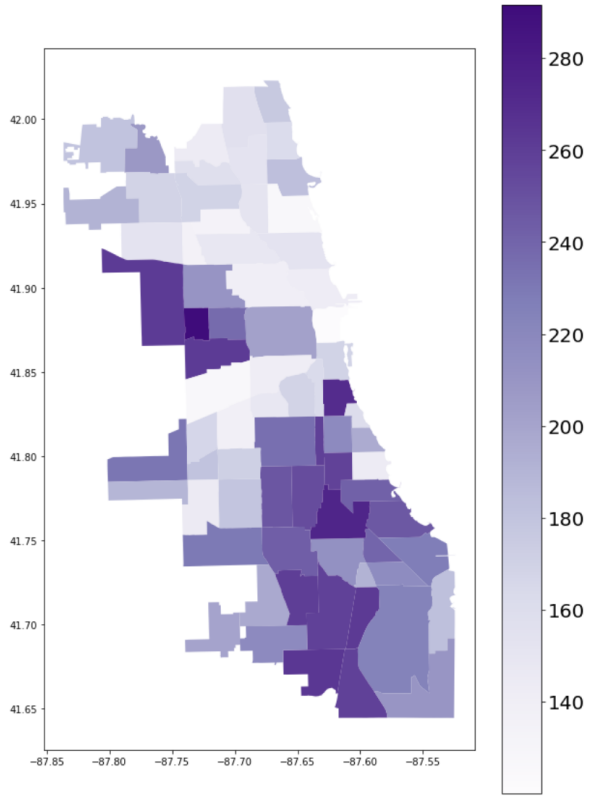
Crowded Housing (%)
by Community Area 2005-2011



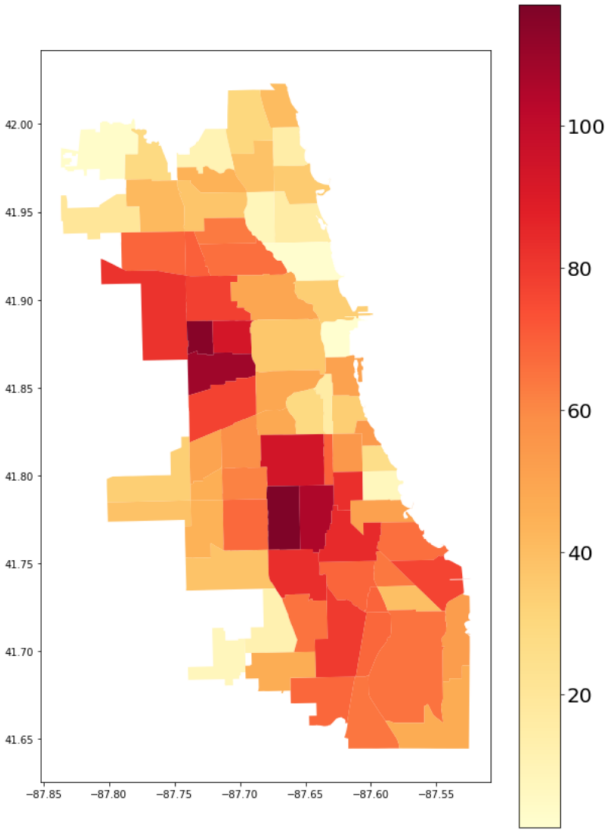
No High School Diploma (%)
by Community Area 2005-2011



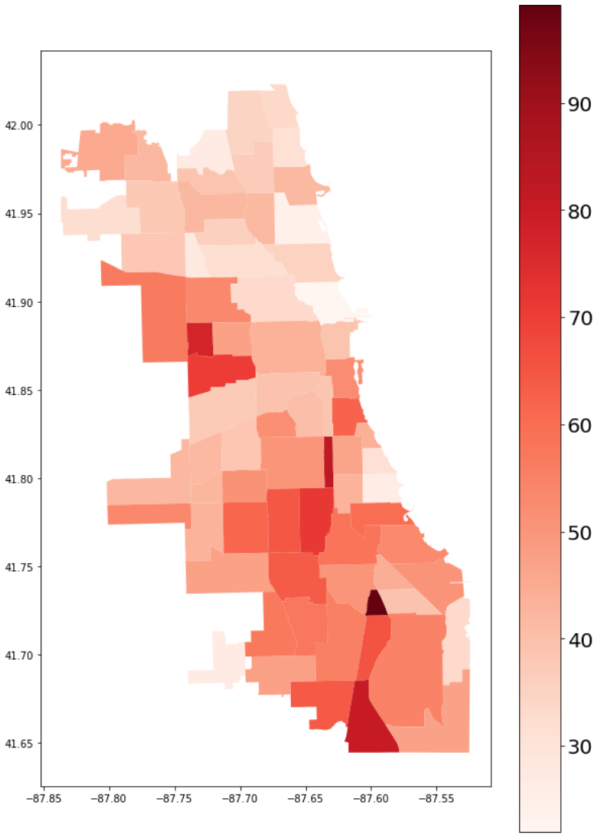
Dependents (%)
by Community Area 2005-2011



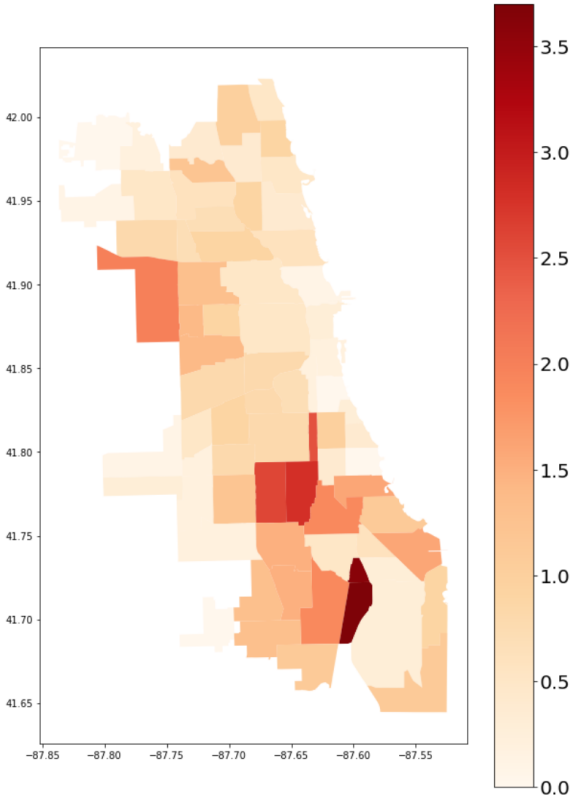
Cancer (per 100,000 persons)
by Community Area 2005-2011



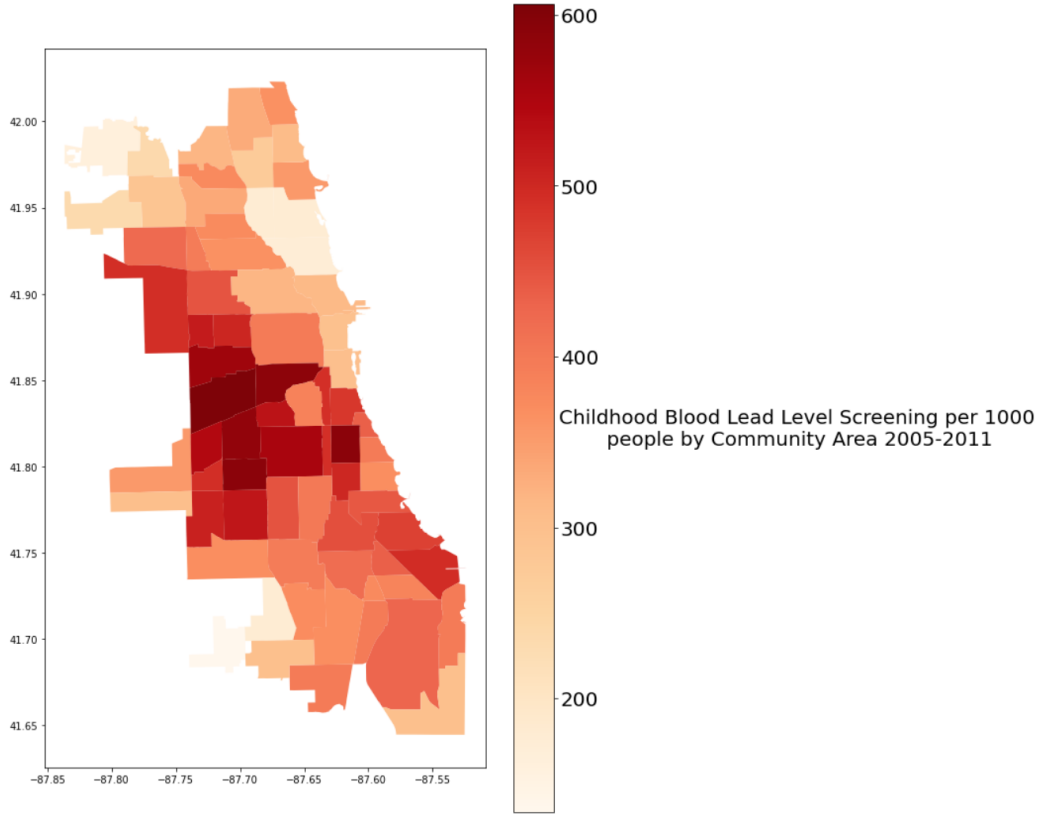
Teen Birth Rate
by Community Area 2005-2011

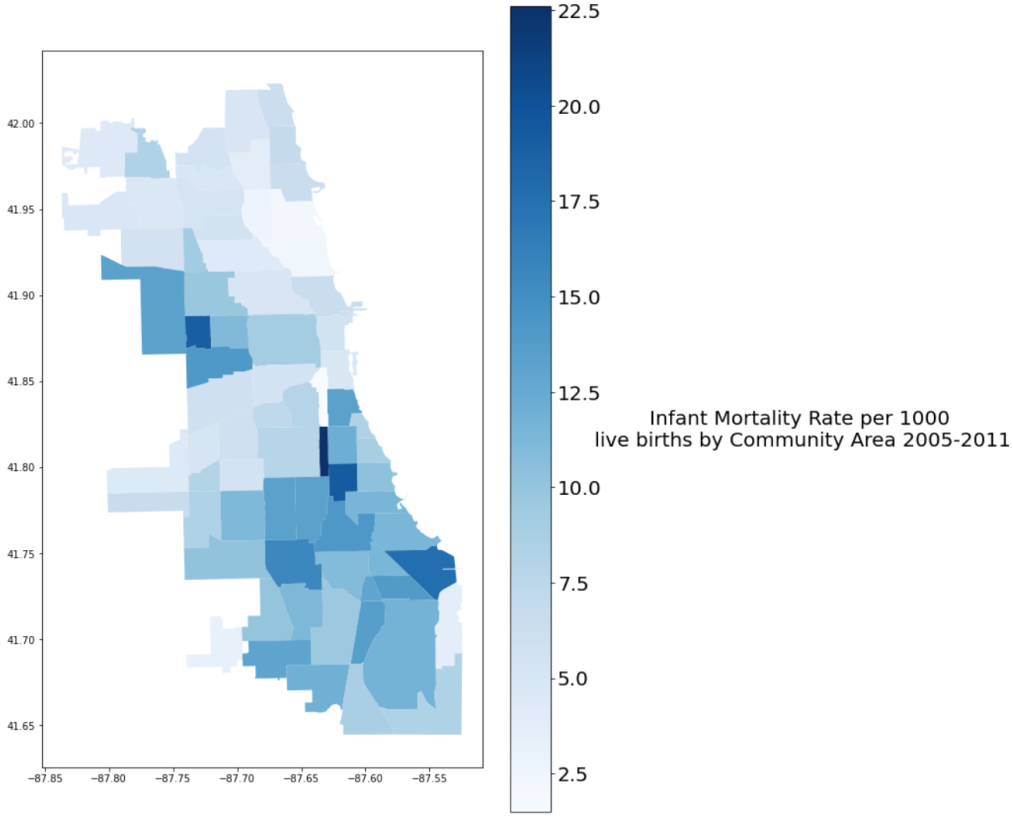


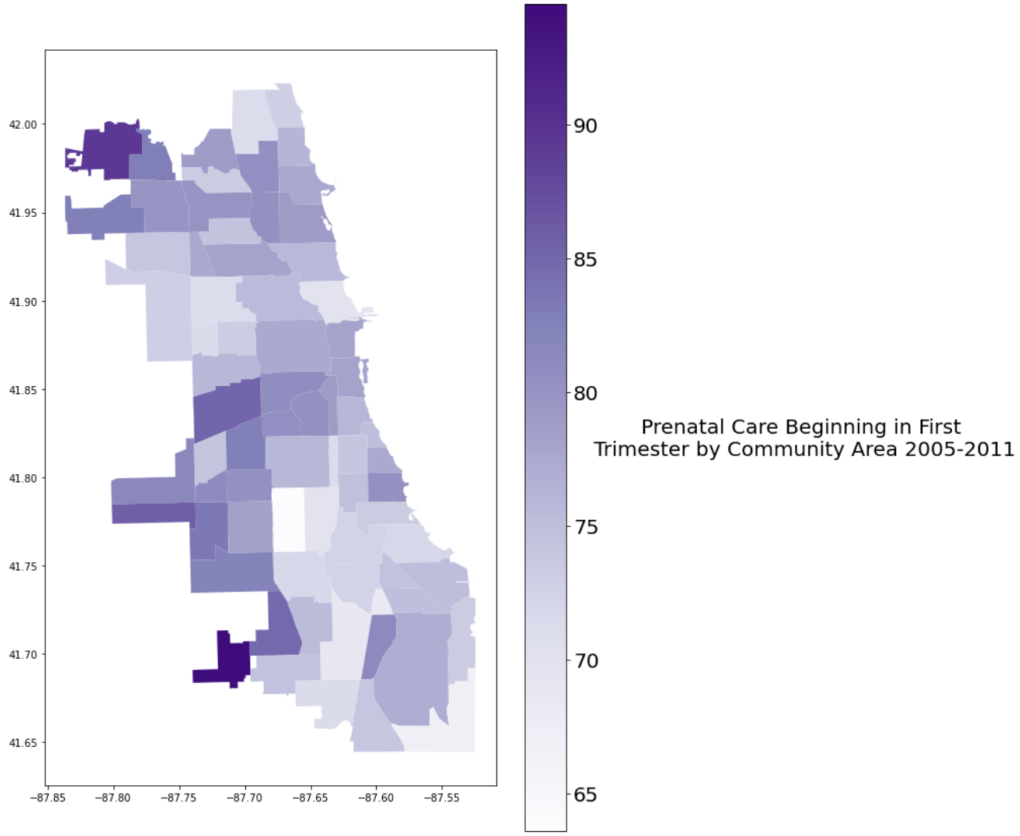
Stroke Deaths per 100,000 people by Community Area 2005-2011



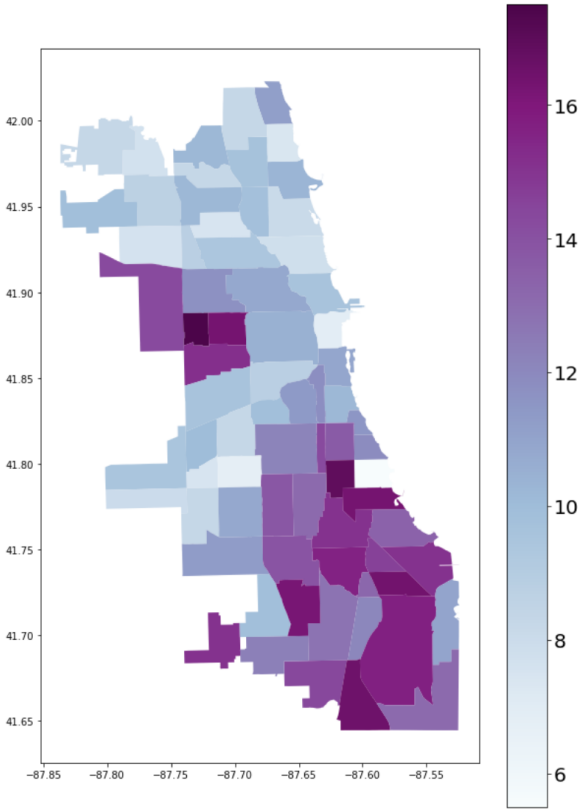
Childhood Lead Poisoning per 100 people by Community Area 2005-2011


42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
−87.85
−87.80
−87.75
−87.70
−87.65
−87.60
−87.55
600
500
400
300
200
Childhood Blood Lead Level Screening per 1000
people by Community Area 2005-2011

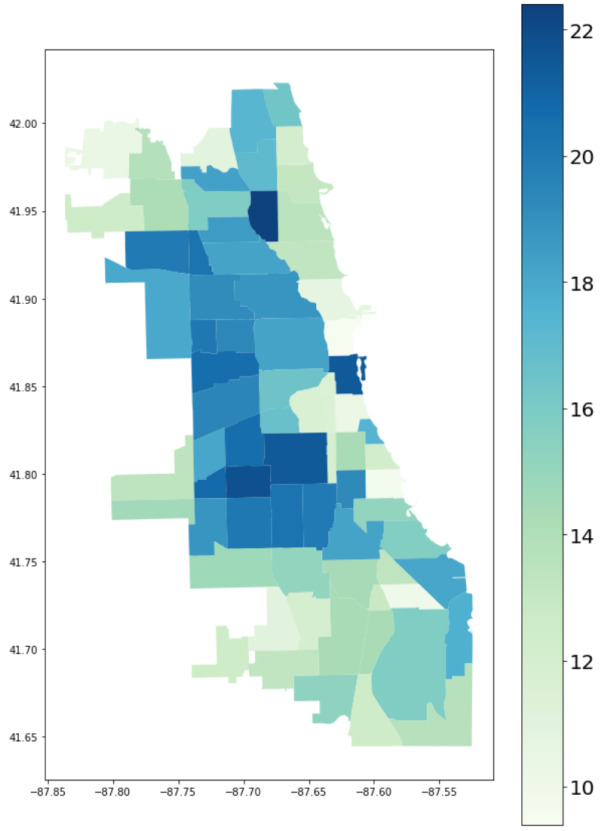

42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
−87.85
−87.80
−87.75
−87.70
−87.65
−87.60
−87.55
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
Infant Mortality Rate per 1000
live births by Community Area 2005-2011



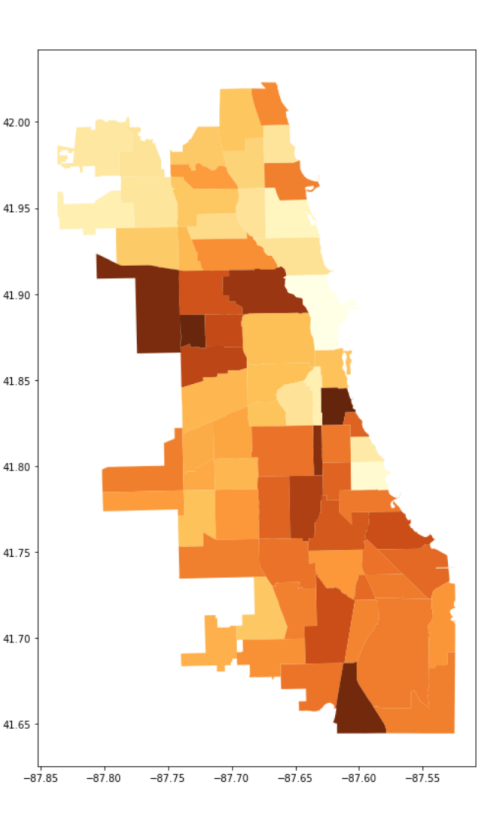
Prenatal Care Beginning in First Trimester by Community Area 2005-2011



Preterm Births by Community Area 2005-2011



Birth Rate by Community Area
2005-2011



Diabetes-related Deaths per 100,000
people by Community Area 2005-2011

V. Data Modeling

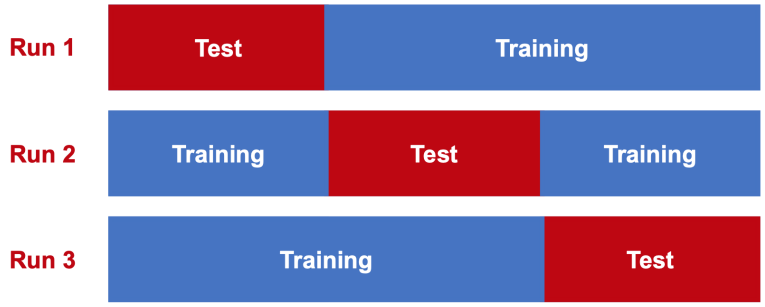
Variable Groups and Modeling Techniques

In order to create classification and regression models that would yield the most meaningful and accurate results, in terms of correlation, precision, and other evaluation criteria, six different classification techniques and three different regression methods were used in combination with four different variable groups. One variable group was the entire predictor set, so all variables, and the other three groups are shown in the table below. Regression models were built using simple linear regression, multiple linear regression, and LASSO regression. Classification models were built using decision tree classifiers, K-nearest neighbors classifiers, linear, polynomial, and radial basis function support vector machines, and a naive Bayes classifier.

Group	Predictors
Economic Predictors	{Per capita income, unemployment, below poverty level, no high school diploma, dependency, crowded housing}
Natality Predictors	{Birth rate, preterm births, teen birth rate, prenatal care beginning in first trimester, infant mortality rate}
Health Predictors	{Cancer (all sites), diabetes-related deaths, stroke-related deaths, childhood lead poisoning, childhood blood lead level screening}

Partitioning the Data into Training Sets and Test Sets

Two different methods were used to split the data into training sets and testing sets. The training sets were used to build the models, and the testing sets were used to evaluate them. Since the amount of data collected is relatively small, it was determined that it would be best to use K-Fold Cross Validation to split the data. The two different methods that were used were K-Fold Cross-Validation, and a variation of K-Fold Cross Validation called Stratified K-Fold Cross Validation. The K-Fold Cross Validation method consists of randomly partitioning the dataset into K equal-sized sets (or folds) and building the model K times. During the *i*th run, the *i*th partition is used as the test set and the rest of the partitions are used as the training set. For example, for 3-Fold Cross Validation, the training and test sets would be split as such:



Stratified K-Fold Cross Validation is the same as just K-Fold Cross Validation, but in Stratified K-Fold Cross Validation, it does stratified sampling instead of random sampling. What this means is that test sets will be generated such that all contain the same distribution of classes, or as close as possible. The advantage of using K-Fold Cross Validation is that all the data is used for both training and testing.

For the models built in this project, K was selected to be 5, thus the data was partitioned into training and testing sets with 5-Fold Cross Validation. This means that the training set will contain 80% of the data, and the test set will contain 20% of the data. To compute the evaluation metrics for the regression models and classification models such as correlation coefficient, accuracy, precision, recall, and F1 score, the average of each of those metrics was taken across all 5 folds.

Building a Regression Model

Linear regression models a linear relationship between a response variable and predictor variables. 17 total regression models were built, that is simple linear regression, multiple linear regression, and LASSO regression models. Simple linear regression modeled the linear relationship between Total Shootings and only one predictor. Multiple linear regression modeled the linear relationship between Total Shootings and multiple predictors. LASSO regression is a regularization technique that adds an additional constraint to the regression model that limits its complexity. In LASSO regression, some coefficients may become exactly 0 and the corresponding predictor variables can be dropped from the model. The correlation coefficient measures the linear relationship between two variables, and is a value between -1 and 1. The closer the coefficient is to -1 or 1, the stronger the relationship.

The results found that simple linear regression did not perform well in terms of being able to predict the total shootings for a given community area. The best relationship found in the simple regression models was with Teen Birth Rate, with a correlation coefficient of 0.749. The multiple and LASSO regression models performed slightly better. There was an average correlation coefficient of 0.634 between all predictors and Total Shootings. Natality predictors did slightly better with an average correlation coefficient of 0.740. The values highlighted below reflect the best performing regression models.

Results of Simple Linear Regression Models

Predictor	Average Correlation Coefficient Across All Folds	Average R^2 Across All Folds
Per Capita Income	0.424	0.220
Unemployment	0.468	0.273
Below Poverty Level	0.507	0.303
Dependency	0.304	0.129
Cancer Deaths	0.509	0.266
Diabetes-related Deaths	0.596	0.368
Birth Rate	0.396	0.188
Teen Birth Rate	0.749	0.575
Infant Mortality Rate	0.480	0.247

Results of Multiple Linear Regression and Lasso Regression Models

	Multiple Linear Regression			LASSO Regression	
Predictors	Average Correlation Coefficient Across All Folds	Average R^2 Across All Folds	Average Adjusted R^2 Across All Folds	Average Correlation Coefficient Across All Folds	Average R^2 Across All Folds
All Predictors	0.617	0.412	0.309	0.634	0.427
Economic Predictors	0.418	0.234	0.102	0.428	0.240
Natality Predictors	0.738	0.555	0.478	0.740	0.557
Health Predictors	0.602	0.386	0.280	0.604	0.389

The R^2 values in these results show the assessment of the fit of the models for simple linear regression, and the proportion of variance of the response variable explained by the model for multiple linear regression. The value of R^2 is between 0 and 1, and the higher the value of R^2 , the better the fit of the model. In multiple linear regression, the adjusted R^2 is a modified version of R^2 that has been adjusted for the number of predictor variables in the model.

Building a Classification Model

The results of the linear regression models weren't expected to be satisfying, which is why in addition to those, classification models were also built. Because there is a minimal amount of data to work with, the quality of the classification models were expected to be far better. Classification is the task of assigning class labels to observations in a dataset. A classification task involves two steps: Induction, which is applying a learning algorithm to labeled observations in a training dataset to build a classification model (or classifier), and deduction, which is applying the classifier to unlabeled observations in a test dataset to predict their class labels. 58 total classification models were built using several different classification techniques, the two different splitting methods, and four different groupings of variables mentioned previously. The classification techniques used were Decision Trees, K-Nearest Neighbors, Linear, Polynomial, and Radial Basis Support Vector Machines, and Naive Bayes.

Overview of Classification Techniques

Decision Trees: A decision tree uses a tree structure to represent a number of possible decision paths and an outcome for each path.

K-Nearest Neighbors: A K-nearest neighbors classifier assigns class labels to observations based on the class labels of the K "most similar" observations (K nearest neighbors).

Support Vector Machines: A support vector machine (SVM) is a classifier that finds the separating hyperplane with the maximum margin.

Naive Bayes: A Naïve Bayes classifier is a probabilistic model that uses Bayes' Theorem for classification.

Average Evaluation Metrics of All Classification Models Using KFold Cross Validation

Classifier Type (params)	Predictors	Average Accuracy	Average Error	Average Precision	Average Recall	Average F1 Score
Decision Tree	All	0.849	0.151	0.900	0.929	0.908
Decision Tree	Economic	0.726	0.274	0.832	0.866	0.836
Decision Tree	Natality	0.834	0.166	0.878	0.943	0.900
Decision Tree	Health	0.823	0.177	0.888	0.914	0.892
K-Nearest **(k=5)**	All	0.805	0.195	0.882	0.912	0.884
K-Nearest **(k=8)**	All	0.806	0.194	0.857	0.943	0.886
K-Nearest **(k=8)**	Economic	0.861	0.139	0.861	1.000	0.917
K-Nearest **(k=3)**	Natality	0.832	0.168	0.899	0.927	0.899
K-Nearest **(k=4)**	Health	0.833	0.167	0.883	0.941	0.899
K-Nearest **(k=5)**	Health	0.807	0.193	0.898	0.898	0.883
Linear SVM	All	0.737	0.263	0.860	0.847	0.839
Linear SVM	Economic	0.847	0.153	0.847	1.000	0.909
Linear SVM	Natality	0.818	0.182	0.869	0.943	0.892
Linear SVM	Health	0.846	0.154	0.870	0.971	0.908
Poly SVM	All	0.832	0.168	0.869	0.956	0.900
Poly SVM	Economic	0.806	0.194	0.842	0.954	0.886
Poly SVM	Natality	0.846	0.154	0.870	0.971	0.908
Poly SVM	Health	0.846	0.154	0.869	0.969	0.907
RBF SVM	All	0.833	0.167	0.859	0.973	0.902
RBF SVM	Economic	0.847	0.153	0.847	1.000	0.909
RBF SVM	Natality	0.819	0.181	0.858	0.958	0.894
RBF SVM	Health	0.820	0.180	0.871	0.941	0.892
Naive Bayes	All	0.807	0.193	0.891	0.896	0.882
Naive Bayes	Economic	0.807	0.193	0.891	0.887	0.878
Naive Bayes	Natality	0.793	0.207	0.873	0.898	0.875
Naive Bayes	Health	0.779	0.221	0.879	0.883	0.866

Standard Deviation of Evaluation Metrics of All Classification Models (KFold)

Classifier Type (params)	Predictors	Std. Dev. Accuracy	Std. Dev. Error	Std. Dev. Precision	Std. Dev. Recall	Std. Dev. F1 Score
Decision Tree	All	0.053	0.053	0.091	0.073	0.040
Decision Tree	Economic	0.117	0.117	0.146	0.089	0.078
Decision Tree	Natality	0.130	0.130	0.143	0.071	0.086
Decision Tree	Health	0.095	0.095	0.123	0.081	0.064
K-Nearest **(k=5)**	All	0.126	0.126	0.159	0.054	0.082
K-Nearest **(k=8)**	All	0.117	0.117	0.149	0.051	0.079
K-Nearest **(k=8)**	Economic	0.157	0.157	0.157	0.000	0.100
K-Nearest **(k=3)**	Natality	0.141	0.141	0.166	0.065	0.091
K-Nearest **(k=4)**	Health	0.141	0.141	0.160	0.057	0.092
K-Nearest **(k=5)**	Health	0.131	0.131	0.166	0.069	0.085
Linear SVM	All	0.139	0.139	0.164	0.095	0.097
Linear SVM	Economic	0.146	0.146	0.146	0.000	0.095
Linear SVM	Natality	0.135	0.135	0.152	0.071	0.080
Linear SVM	Health	0.140	0.140	0.152	0.035	0.092
Poly SVM	All	0.134	0.134	0.152	0.036	0.088
Poly SVM	Economic	0.138	0.138	0.145	0.038	0.089
Poly SVM	Natality	0.140	0.140	0.152	0.035	0.092
Poly SVM	Health	0.146	0.146	0.152	0.038	0.094
RBF SVM	All	0.134	0.134	0.150	0.053	0.089
RBF SVM	Economic	0.146	0.146	0.146	0.000	0.095
RBF SVM	Natality	0.127	0.127	0.149	0.055	0.085
RBF SVM	Health	0.133	0.133	0.158	0.057	0.088
Naive Bayes	All	0.086	0.086	0.142	0.059	0.062
Naive Bayes	Economic	0.167	0.167	0.179	0.058	0.113
Naive Bayes	Natality	0.049	0.049	0.112	0.085	0.042
Naive Bayes	Health	0.074	0.074	0.142	0.086	0.055

Average Evaluation Metrics of All Classification Models Using Stratified KFold

Classifier Type (params)	Predictors	Average Accuracy	Average Precision (Low / High Risk)	Average Recall (Low / High Risk)	Average F1 Score (Low / High Risk)
Decision Tree	All	0.837	0.881 / 0.300	0.936 / 0.233	0.906 / 0.248
Decision Tree	Economic	0.729	0.841 / 0.067	0.842 / 0.100	0.838 / 0.080
Decision Tree	Natality	0.864	0.909 / 0.567	0.936 / 0.467	0.921 / 0.480
Decision Tree	Health	0.837	0.905 / 0.400	0.904 / 0.433	0.904 / 0.414
K-Nearest **(k=3)**	All	0.850	0.948 / 0.537	0.872 / 0.700	0.907 / 0.597
K-Nearest **(k=5)**	All	0.865	0.947 / 0.607	0.891 / 0.700	0.916 / 0.637
K-Nearest **(k=5)**	Economic	0.821	0.870 / 0.133	0.936 / 0.200	0.899 / 0.160
K-Nearest **(k=3)**	Natality	0.907	0.926 / 0.800	0.969 / 0.567	0.946 / 0.633
K-Nearest **(k=4)**	Health	0.877	0.920 / 0.750	0.937 / 0.533	0.927 / 0.593
K-Nearest **(k=5)**	Health	0.850	0.933 / 0.517	0.888 / 0.633	0.909 / 0.560
Linear SVM	All	0.850	0.907 / 0.400	0.919 / 0.433	0.912 / 0.413
Linear SVM	Economic	0.850	0.850 / 0.000	1.000 / 0.000	0.918 / 0.000
Linear SVM	Natality	0.864	0.872 / 0.400	0.985 / 0.167	0.925 / 0.233
Linear SVM	Health	0.809	0.855 / 0.067	0.937 / 0.100	0.892 / 0.080
Poly SVM	All	0.878	0.885 / 0.600	0.985 / 0.267	0.932 / 0.367
Poly SVM	Economic	0.850	0.850 / 0.000	1.000 / 0.000	0.918 / 0.000
Poly SVM	Natality	0.878	0.885 / 0.600	0.985 / 0.267	0.932 / 0.367
Poly SVM	Health	0.850	0.870 / 0.400	0.967 / 0.200	0.916 / 0.267
RBF SVM	All	0.878	0.896 / 0.800	0.969 / 0.367	0.930 / 0.500
RBF SVM	Economic	0.850	0.850 / 0.000	1.000 / 0.000	0.918 / 0.000
RBF SVM	Natality	0.878	0.896 / 0.800	0.969 / 0.367	0.930 / 0.500
RBF SVM	Health	0.850	0.895 / 0.550	0.937 / 0.400	0.912 / 0.433
Naive Bayes	All	0.808	0.933 / 0.350	0.840 / 0.600	0.883 / 0.438
Naive Bayes	Economic	0.754	0.897 / 0.260	0.809 / 0.400	0.848 / 0.307
Naive Bayes	Natality	0.837	0.948 / 0.497	0.858 / 0.700	0.899 / 0.566
Naive Bayes	Health	0.836	0.936 / 0.403	0.872 / 0.600	0.900 / 0.477

The Class Imbalance Problem and Informed Oversampling

The class imbalance problem occurs when there are significantly more observations from one class than from the other in a classification task. When the data is imbalanced, classifiers tend to be biased towards the majority class and to ignore the minority one, which is usually the class of interest. In this case, there was a significant difference between the amount of low risk community areas and the amount of high risk community areas. To try and improve the performance of the models, an informed oversampling method was used called Synthetic Minority Oversampling Technique (SMOTE). Informed oversampling methods add synthetically generated observations to improve the performance of the classifier. SMOTE generates new observations by interpolating two minority class observations. Below are the results prior to oversampling, and then the result after oversampling. The data was oversampled such that there was a 0.4 ratio of the minority class (high risk) over the majority class (low risk).

Results Before Oversampling

Classifier Type (params)	Predictors	Average Accuracy	Average Precision (Low / High Risk)	Average Recall (Low / High Risk)	Average F1 Score (Low / High Risk)
Decision Tree	All	0.837	0.881 / 0.300	0.936 / 0.233	0.906 / 0.248
K-Nearest **(k=5)**	All	0.865	0.947 / 0.607	0.891 / 0.700	0.916 / 0.637
K-Nearest **(k=3)**	Natality	0.907	0.926 / 0.800	0.969 / 0.567	0.946 / 0.633
K-Nearest **(k=4)**	Health	0.877	0.920 / 0.750	0.937 / 0.533	0.927 / 0.593
RBF SVM	All	0.878	0.896 / 0.800	0.969 / 0.367	0.930 / 0.500
Naive Bayes	Natality	0.837	0.948 / 0.497	0.858 / 0.700	0.899 / 0.566

Results After Oversampling

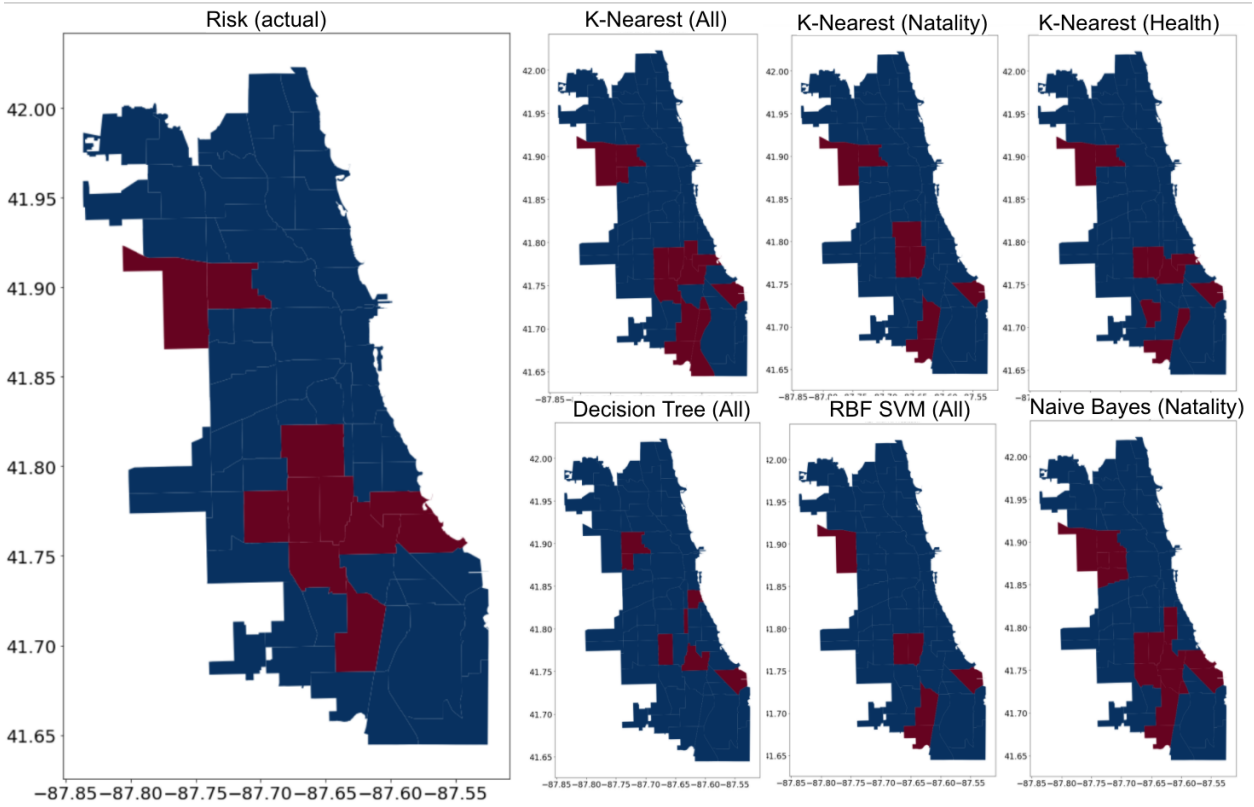
Classifier Type (params)	Predictors	High / Low Risk Ratio	Average Accuracy	Average Precision (Low / High Risk)	Average Recall (Low / High Risk)	Average F1 Score (Low / High Risk)
Decision Tree	All	0.400	0.891	0.922 / 0.733	0.953 / 0.533	0.937 / 0.600
K-Nearest **(k=5)**	All	0.400	0.838	1.000 / 0.552	0.808 / 1.000	0.889 / 0.687
K-Nearest **(k=3)**	Natality	0.400	0.905	0.969 / 0.683	0.919 / 0.800	0.943 / 0.705
K-Nearest **(k=4)**	Health	0.400	0.805	0.950 / 0.503	0.872 / 0.700	0.908 / 0.577
RBF SVM	All	0.400	0.906	0.954 / 0.683	0.937 / 0.700	0.945 / 0.665
Naive Bayes	Natality	0.400	0.837	0.967 / 0.463	0.841 / 0.800	0.898 / 0.579

There was a slightly performance improvement when additional observations were added to the data. This seems to suggest that the performances of the models in terms of predicting the high risk community areas would be much more accurate if there was a larger amount of observations to work with.

Classification Model Conclusions

The classification models did well classifying the high risk and low risk community areas collectively, averaging 83.4% accuracy across all models. The models did very well in classifying the low risk community areas but struggled with the high risk community areas. This can be explained by the class imbalance problem mentioned above. The models were rebuilt with oversampled data, and they performed slightly better with more balanced data points. The figures below show the best performing models, and the maps show the class predictions from those models. The map on the left hand side shows the actual class labels, high risk or low risk, and the rest show the performance of each of the six classification models from the table above.

Classification Model Heatmap Results



VI. Conclusions

Based on the results from the linear regression models, the socioeconomic indicators are not good predictors for estimating the total shootings for a particular community area. However, based on the classification models, the socioeconomic / public health indicators are good at estimating whether or not a community area can be classified as low risk (< 400 Total Shootings) or high risk (>= 400 Total Shootings). In both the regression models and classification models, it was found that natality variables did much better than other categories in predicting total shootings and classifying community areas. With these conclusions from the data models, in conjunction with the comparisons detailed in the exploratory data analysis, it is suggested that there exists a relationship between most of these socioeconomic / public health variables and the crime rate in the city of Chicago.

VII. References

City of Chicago: Data Portal, https://data.cityofchicago.org/.

“Introduction to GeoPandas.” GeoPandas, geopandas.org/en/stable/getting_started/introduction.html.

“SMOTE.” Imbalanced Learn, imbalanced-learn.org/

“sklearn.model_selection.KFold.” scikit learn, scikit-learn.org/

“Check if a point falls within a multipolygon with Python.” GIS Stack Exchange, gis.stackexchange.com

Bello, Gonzalo “What is Data Science?” *UIC Blackboard*, 25 August 2020. Slideshow.

Bello, Gonzalo “Exploratory Data Analysis” *UIC Blackboard*, 17 September 2020. Slideshow.

Bello, Gonzalo “Introduction to Data Modeling” *UIC Blackboard*, 22 September 2020. Slideshow.

Bello, Gonzalo “Regression (I): Simple Linear Regression” *UIC Blackboard*, 22 September 2020. Slideshow.

Bello, Gonzalo “Regression (II): Multiple Linear Regression” *UIC Blackboard*, 29 September 2020. Slideshow.

Bello, Gonzalo “Classification (I): Basic Concepts & Decision Trees” *UIC Blackboard*, 01 October 2020. Slideshow.

Bello, Gonzalo “Classification (II): More Classification Techniques” *UIC Blackboard*, 06 October 2020. Slideshow.

Bello, Gonzalo “Classification (III): Ensemble Methods and the Class Imbalance Problem” *UIC Blackboard*, 13 October 2020. Slideshow.

VIII. Acknowledgements

This work was supervised by Dr. Gonzalo Bello, of the University of Illinois at Chicago Computer Science Department.