UIC HONORS COLLEGE

Digging Deeper Into Chicago Gun Crime:

Analyzing the Socioecomic Contributors to Chicago Gun Violence

Mersim Rizmani, University of Illinois at Chicago

Abstract

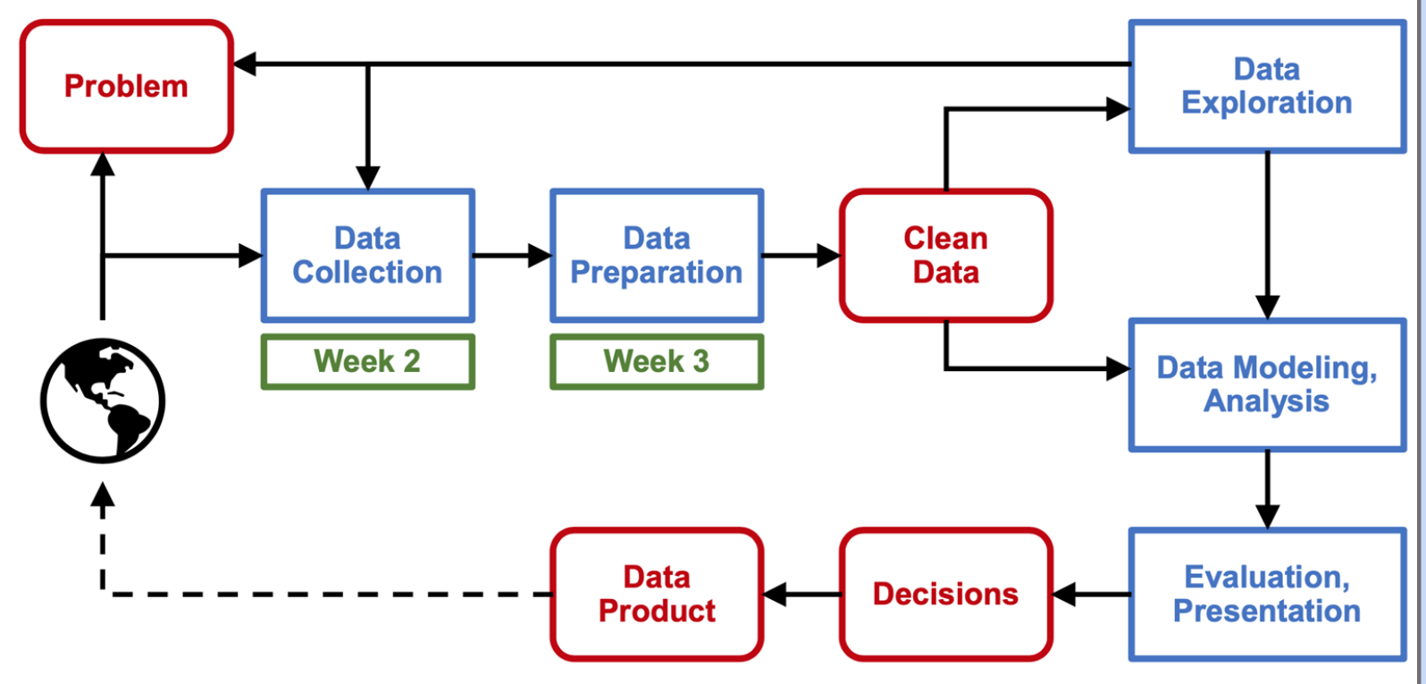
The crime rate in the city of Chicago, especially violent crime, is higher than the United States average. Over the years, the crime rates have drastically increased, with 2021 being the worst year since 1996. This research dives into different socioeconomic and public health indicators from 2005-2011, 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.

Methods

The Data Science Pipeline



Adapted from: Cathy O'Neil and Rachel Schutt, *Doing Data Science* (2013)

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 statistics 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. The data was split into training and testing sets using 5-fold cross validation.

Evaluate Results: Computed accuracy of predictions and classifications and visualized predictions using heatmaps.

Data Collection

All of the data collected for this research came from the **City of Chicago Open Data Portal.** Four datasets from the Chicago Data Portal were analyzed: "**Crimes 2001-pres**", "**Shootings**" (a subset of the Crimes dataset), "**Boundaries - Community Areas (current),**" and "**Public Health Statistics- Selected public health indicators by Chicago community area.**" Public Health Statistics contain a selection of 27 indicators of public health significance by the 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. Note that these indicators range from the years 2005-2011, so for the purposes of this research, only those years were analyzed.

CHICAGO DATA PORTAL

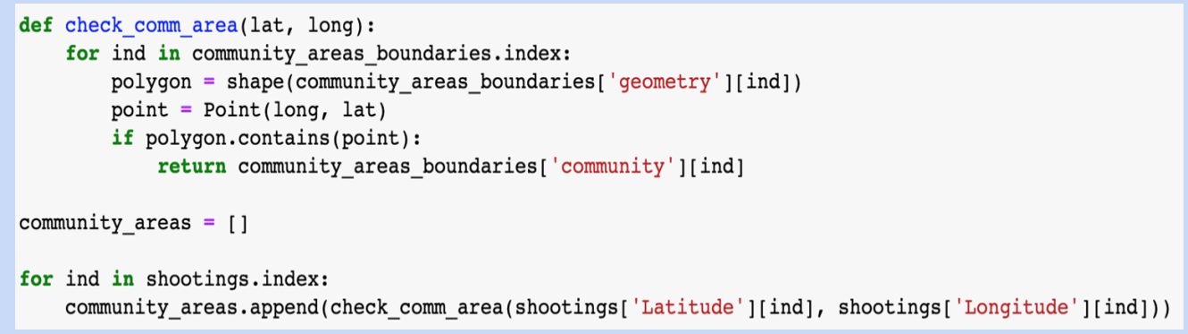
Tables of Socioeconomic Indicators:

Category	Measure	Units
Economic	Per Capita Income	2011 inflation-adjusted dollars
	Unemployment	Percent of persons in labor force aged 16 years and older
	Below Poverty Level	Percent of households
	Crowded Housing	Percent of occupied housing units
	No High School Diploma	Percent of persons aged 25 years and older
	Dependency	Percent of persons aged less than 16 or more than 64 years
Natality	Birth Rate	Per 1,000 persons
	Teen Birth Rate	Per 1,000 females aged 15-19
	Preterm Births	Percent of live births
	Infant Mortality Rate	Per 1,000 live births
	Prenatal care beginning in first trimester	Percent of females delivering a live birth
Health	Cancer (all sites) [Deaths]	Per 100,000 persons (age adjusted)
	Diabetes-related [Deaths]	Per 100,000 persons (age adjusted)
	Stroke (cerebrovascular disease) [Deaths]	Per 100,000 persons (age adjusted)
	Childhood lead poisoning	Per 100
	Childhood blood lead level screening	Per 1,000 children aged 0-6 years

Data Preparation

Several tasks need to be completed to clean and prepare the data for analysis. First, the Shootings dataset was filtered to only include incidents from the **years 2005-2011** to match the socioeconomic data. Second, 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**. Once that was done, the Shootings dataset was grouped by the community area, all columns except the community area and the total shootings were removed, and the dataset was **merged with the socioeconomic data**. 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 **Total Shootings = 400** was a sufficient split point for defining low risk and high risk community areas. A **classification** was added to the dataset using this split point.

GeoData Mapping

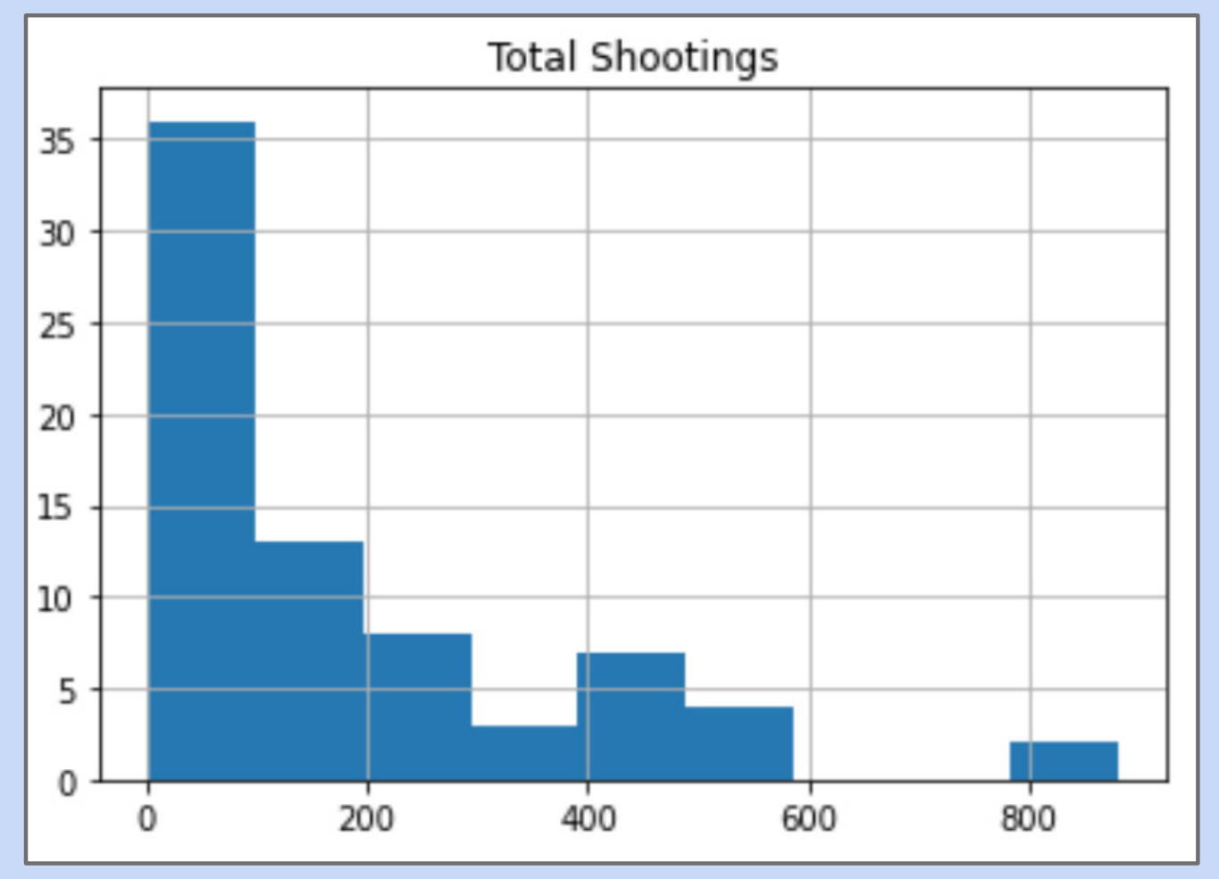


```
def check_comm_area(lat, long):
    for ind in community_areas_boundaries.index:
        polygon = shape(community_areas_boundaries['geometry'][ind])
        point = Point(long, lat)
        if polygon.contains(point):
            return community_areas_boundaries['community'][ind]

community_areas = []

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

Community Areas (#) vs Total Shootings



Data Exploration

Prior to beginning exploratory data analysis, the dataset was **partitioned into low risk and high risk tables** based on the figure defined in data preparation. 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**.

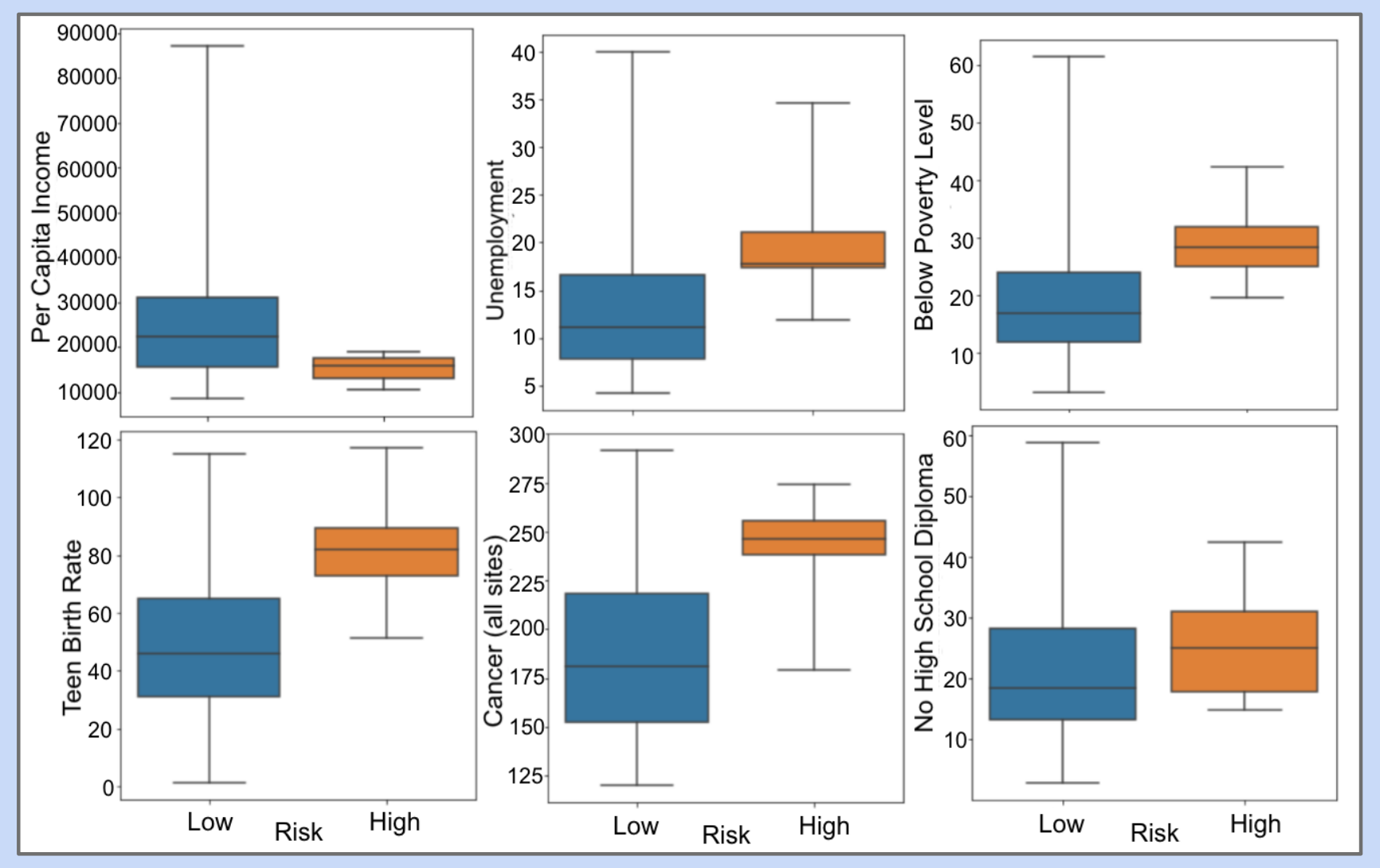
Two-sample hypothesis testing was conducted using an unpaired ***t***-Test for population means. The **null hypothesis in all the tests was that the means were equal**, and the alternative hypothesis was that they were not equal. The table on the right shows the means for each socioeconomic statistic 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.

Hypothesis Testing Results

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

*a ***p***-value is the probability of obtaining a result that is as extreme or more extreme than the observed result if the null hypothesis is true

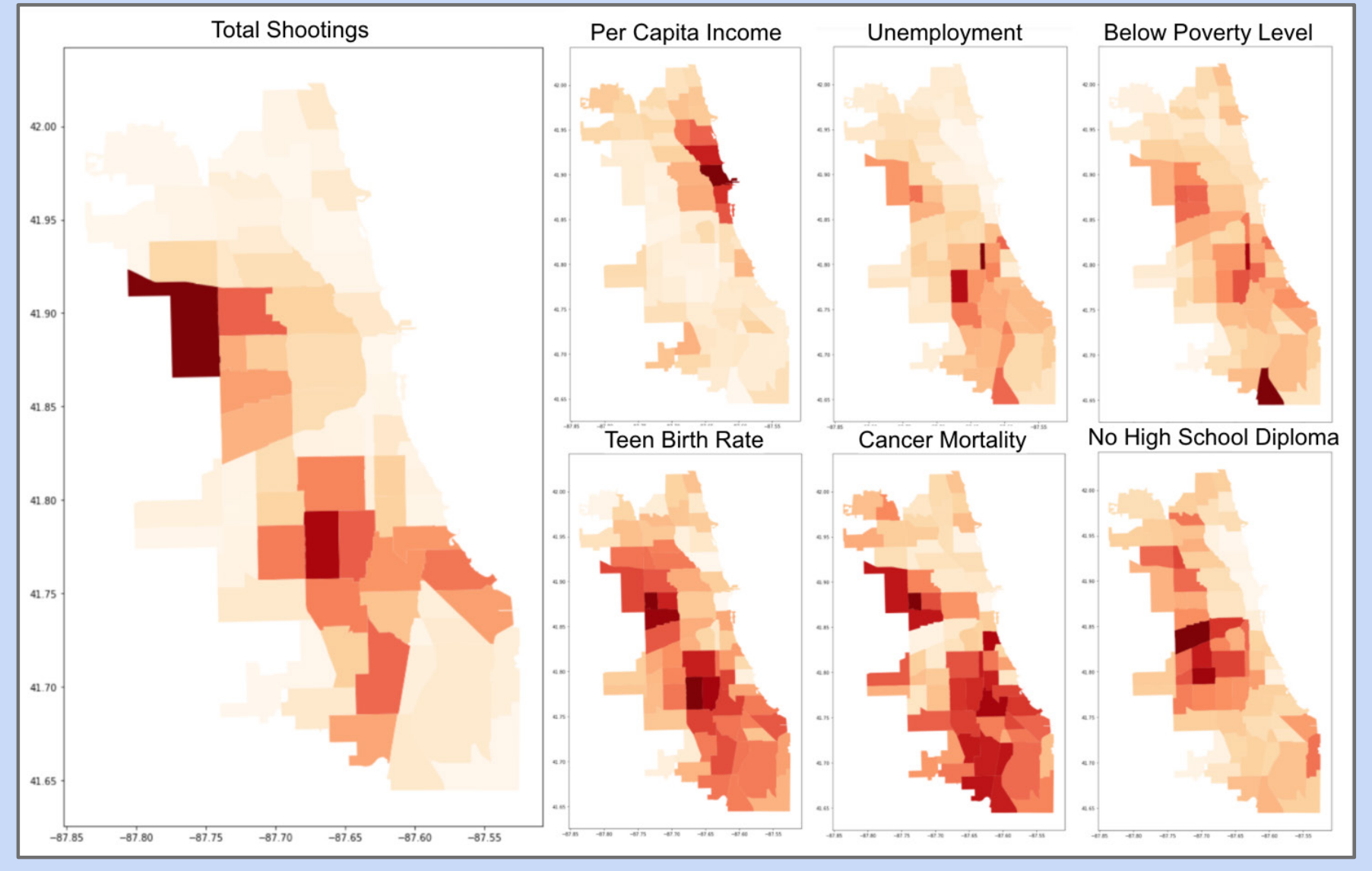
Low Risk vs High Risk Comparison



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 figure on the left shows 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, while high risk community areas have higher unemployment, more poverty, less education, and more deaths from cancer.

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 figure on the right features heatmaps for each of the six variables from the boxplots above, as well as the Total Shootings heatmap highlighted on the left hand side of the figure. 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.

Community Area Comparisons



Data Modeling

Predictors Groupings

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}

Linear Regression Results

	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

Building a Classification Model

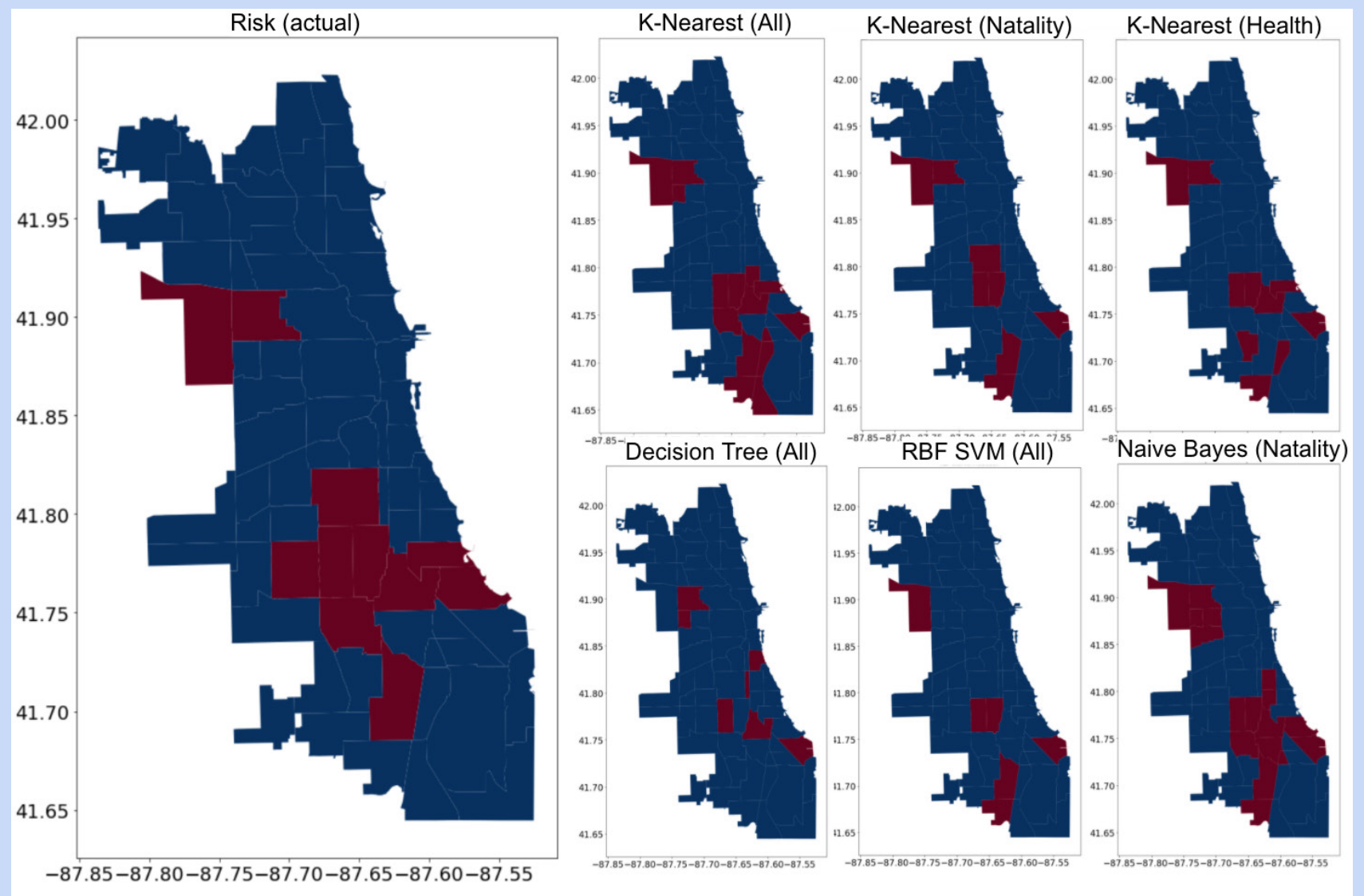
- **58 total classification models** were built using several different classification techniques, two different splitting methods, and four different groupings of variables.
- The classification models did well classifying the high risk and low risk community areas collectively, averaging **83.4% accuracy** 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**. The models were rebuilt with oversampled data, and they performed slightly better with more balanced data points.
- The figures on the right show the best performing models, and the maps show the class predictions from those models.

Building a Regression Model

- **17 total regression models** were built, that is simple linear regression, multiple linear regression, and LASSO regression models.
- The 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.

Best Performing Classification Models

Classifier Type (params)	Predictors	Average Accuracy	Average Precision (Low / High)	Average Recall (Low / High)	Average F1 Score (Low / High)
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



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.

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

Acknowledgements

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