The Education-Crime Connection

A Data-Driven Analysis of the Philadelphia Public School System and Crime Rates

Liam Smith Tyler Christianson Mitch Jeter

March 2nd, 2025

Contents

1 Executive Summary

1.1 Goal of the Study

The primary objective of this project is to explore and understand the relationship between the Philadelphia Public Schooling System and crime rates in the city of Philadelphia. The relationship between education and crime has long been a topic of interest for policymakers and researchers, as strong education systems are often associated with lower crime rates. This project aims to analyze the connection between the public school system and crime rates across the city, exploring how factors such as graduation rates influence crime patterns. By leveraging records from 2023 and various statistical and machine learning methods, this study seeks to uncover potential relationships and provide insights that could inform policy decisions geared towards improving both education and public safety in the city of Philadelphia.

Through data-driven analysis, this project will look into specific neighborhoods and evaluate the relationship between crime rates and the quality of public schools in the area. By identifying factors that contribute to this relationship, this study can serve as a potential foundation for discussions on how investments in education can serve as a proactive step towards crime prevention.

1.2 Data description

1.2.1 OpenData Philly - Uncleaned Datasets

1.2.1.1 Crime Incidents Data The Philadelphia Police Department provides and maintains a dataset of crime incidents in the city of Philadelphia going back to 2006. The dataset contains key information about each incident, including the police district it occured in, the date and time an officer was dispatched to the scene, the block that the incident occured on, and the type of crime that was committed. For the purposes of this study, the only year used was 2023.

1.2.1.2 School Department of Philadelphia Data The School District of Philadelphia provides a variety of datasets regarding school performance. This study is based on the School Graduation Rates dataset, which is a longitudinal dataset containing the graduation rate by school broken out by graduation rate type, demographic category, and ninth grade cohort. Students are attributed to the last school that they attended if they did attend multiple schools. For the purposes of this study, the cohort that started high school in 2019 was used.

1.2.1.3 Neighborhood GeoJSON Data The city of Philadelphia provides neighborhood boundaries for over 150 neighborhoods in the city. This dataset will be used for mapping and also for grouping schools and crime incidents by neighborhood.

1.2.2 Team One Data - Cleaned Datasets

1.2.2.1 Crime Incidents by Neighborhood This dataset contains crime incident records from 2023 categorized by neighborhood. It includes 162,032 entries, with attributes such as the date and time of the incident, crime classification, location details, and associated neighborhood. The dataset also includes unique identifiers, district and police service area, and geospatial information. This dataset will be used for analyzing crime trends, spatial distributions, and temporal patterns across neighborhoods.

1.2.2.2 School Data This dataset contains information on school performance metrics for 79 schools. Each entry includes details such as the school year, unique school identifier, school name, and sector (sector, e.g., district or charter). It focuses on four year graduation rates. The dataset also includes numerical values for assessment, as well as geographical coordinates for spatial analysis. This dataset is useful for evaluating school performance trends and comparing graduation rates across different student populations.

2 Exploratory data analysis (EDA)

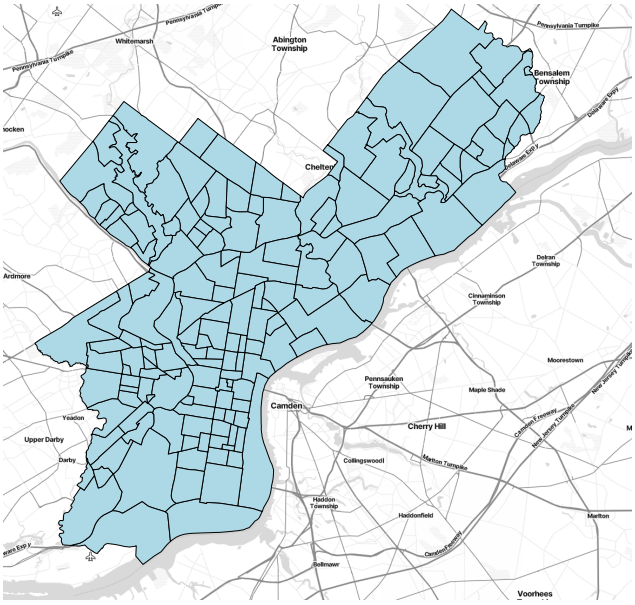
2.1 Exploring Philadelphia

Philadelphia, the largest city in Pennsylvania and the sixth-largest in the United States, is a historic and cultural hub known for its rich colonial heritage and vibrant modern identity. Founded in 1682 by William Penn, the city played a crucial role in the American Revolution and served as the nation's capital before Washington, D.C. Landmarks such as Independence Hall, where the Declaration of Independence and the U.S. Constitution were signed, and the Liberty Bell symbolize Philadelphia's deep-rooted connection to American democracy. Beyond its historical significance, the city boasts world-class museums, including the Philadelphia Museum of Art — home to the famous "Rocky Steps" — as well as a thriving arts, music, and culinary scene that attracts millions of visitors each year.

Today, Philadelphia is a dynamic metropolis that balances tradition with innovation. It is home to leading universities such as the University of Pennsylvania and Temple University, contributing to its reputation as an educational and research powerhouse. The city's economy is diverse, with strong sectors in healthcare, finance, and technology, alongside a passionate sports culture represented by teams like the Eagles, Phillies, Flyers, and 76ers. However, Philadelphia also faces challenges, including socioeconomic disparities, crime, and education system concerns. Despite these issues, the city remains resilient, with ongoing efforts to foster economic growth, community development, and urban revitalization, making it a place of both historical pride and modern ambition.

The city of Philadelphia is home to 159 neighborhoods, the boundaries of which can be seen as shown here :

```
## i © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap contributors.
```



This map will serve as a foundation for this project and the following exploratory data analysis. By undertsanding the city's geographic layout and distribution of neighborhood, we can better examine how our factors differ across neighborhoods.

2.2 Exploring Crime

Philadelphia has long faced challenges with crime, making public safety a key concern for residents and city officials alike. The Philadelphia Police Department (PPD), one of the oldest municipal police forces in the

United States, is responsible for maintaining law and order across the city's diverse neighborhoods. Crime rates in Philadelphia have fluctuated over the years, with violent crime, particularly gun violence, being a persistent issue in certain areas. Factors such as socioeconomic conditions, education, and policing strategies all play a role in shaping crime trends throughout the city. While initiatives like community policing and crime prevention programs aim to reduce criminal activity, disparities in crime rates across neighborhoods highlight the need for a data-driven approach to understanding and addressing public safety concerns.

Crime incidents are categorized two different ways by the city of Philadelphia: Part 1 and Part 2. Part 1 crimes are more serious offenses. Part 1 crimes are as follows: Homicide (Criminal), Homicide (Justifiable), Homicide (Gross Negligence), Rape, Robbery Firearm, Robbery No Firearm, Aggravated Assault Firearm, Aggravated Assault No Firearm, Burglary Non-Residential, Burglary Residential, Theft from Vehicle, Retail Theft, Motor Vehicle Theft, and Recovered Stolen Motor Vehicle. Part 2 crimes are less serious offenses. Part 2 crimes are as follows: Other Assaults, Arson, Forgery and Counterfeiting, Fraud, Embezzlement, Recieving Stolen Property, Vandalism/Criminal Mischief, Weapon Violations, Prostitution and Commericalized Vice, Other Sex Offenses, Narcotic/Drug Law Violations, Gambling Violations, Offenses Against Family and Children, DUI, Liquor Law Violations, Drunkenness, Disorderly Conduct, Vagrancy, and All Other Offenses. The top five most commited crimes by Part are as follows:

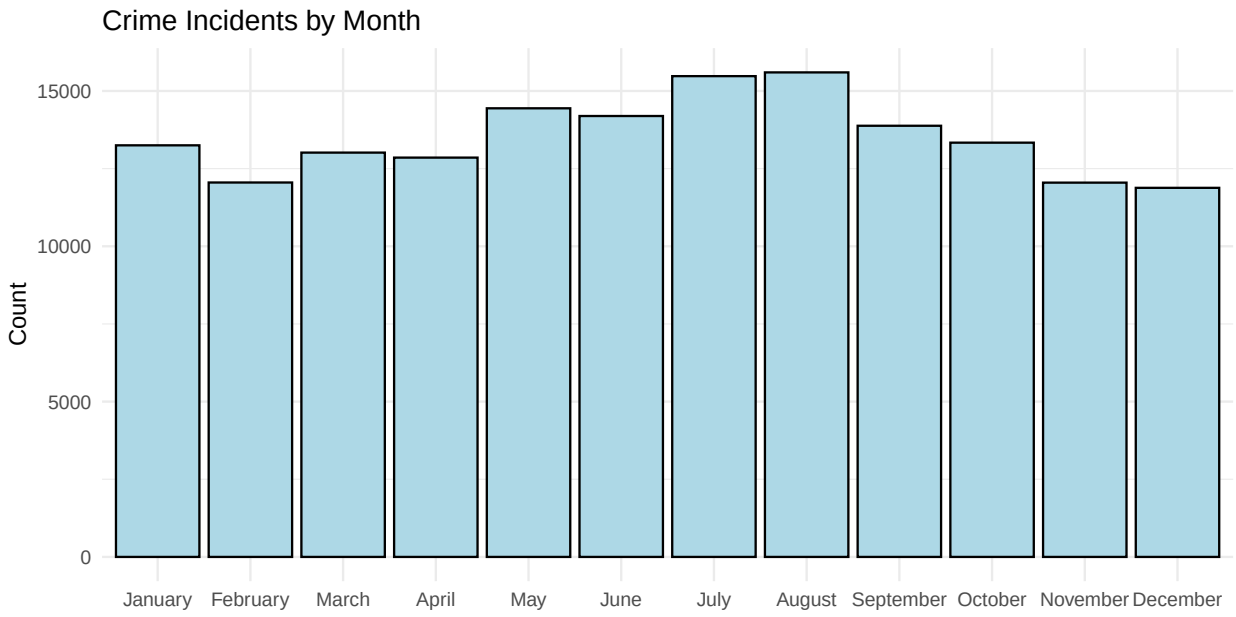
Table 1: Most Committed Part 1 Crimes

Crime	Count
Thefts	35471
Motor Vehicle Theft	21563
Theft from Vehicle	12056
Aggravated Assault No Firearm	4950
Aggravated Assault Firearm	3202

Table 2: Most Committed Part 2 Crimes

Crime	Count
Other Assaults	25109
Vandalism/Criminal Mischief	16760
All Other Offenses	12911
Fraud	7369
Weapon Violations	3121

Crime rates often experience a noticeable increase during the summer months, a trend observed in many cities, including Philadelphia. Several factors contribute to this seasonal spike, including warmer weather, longer daylight hours, and increased social interactions, all of which create more opportunities for criminal activity. Schools being out of session can also lead to higher rates of juvenile crime, as young individuals have more free time and fewer structured activities. Additionally, heat has been linked to increased aggression and heightened tensions, which can escalate conflicts and contribute to violent crimes. Law enforcement agencies often anticipate these seasonal fluctuations and may adjust their patrols and crime prevention strategies accordingly to mitigate the surge in incidents. Philadelphia's number of crimes by month in 2023 is shown here :

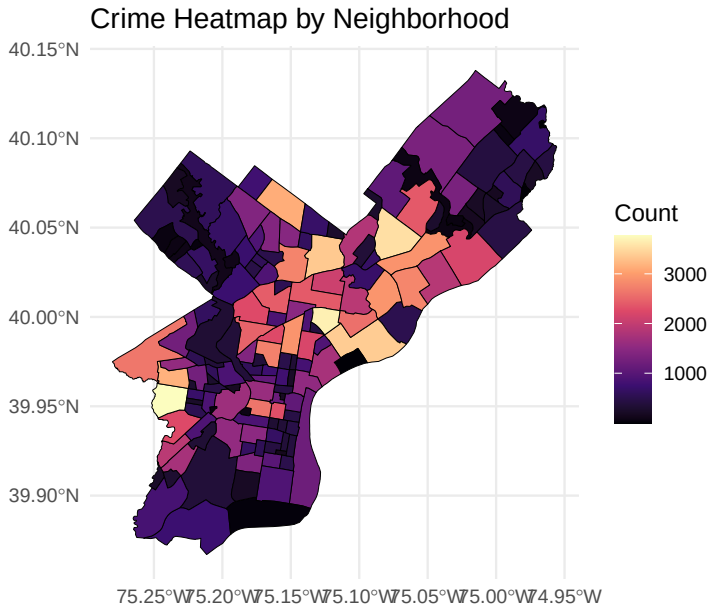


Crime in Philadelphia varies significantly by neighborhood, reflecting differences in socioeconomic conditions, population density, and law enforcement presence. Certain areas, such as Kensington and North Philadelphia, experience higher crime rates, particularly violent crimes, due to factors like poverty, drug activity, and gang presence. In contrast, neighborhoods like Center City and parts of South Philadelphia, while still experiencing crime, tend to have lower rates of violent offenses and more incidents related to property crime or minor offenses. The distribution of crime is also influenced by factors such as proximity to major transit hubs, nightlife areas, and community investment in safety initiatives. Understanding crime patterns at the neighborhood level is crucial for law enforcement, policymakers, and community organizations to develop targeted strategies for crime prevention and public safety improvements. The top ten neighborhoods by total crime incidents are as follows:

Table 3: Top 10 Neighborhoods by Crime Incidents

Neighborhood	Count
Cobbs Creek	3755
Upper Kensington	3648
Oxford Circle	3516
Richmond	3362
Olney	3334
Haddington	3153
West Oak Lane	3114
Frankford	2907
Hartranft	2904
Mayfair	2895

And a crime heatmap can be seen here:



2.3 Exploring Schools

The School District of Philadelphia (SDP) is the largest public school system in Pennsylvania, serving over 200,000 students across a diverse and dynamic urban landscape. With a mix of traditional public schools, charter schools, and specialized programs, the district plays a crucial role in shaping educational opportunities for children in the city. However, SDP has long faced challenges related to funding disparities, aging infrastructure, and achievement gaps among students from different socioeconomic backgrounds. Despite these obstacles, the district continues to implement reforms aimed at improving academic performance, increasing graduation rates, and expanding access to quality education. As education is closely linked to broader social and economic outcomes, understanding the strengths and struggles of Philadelphia's public schools is essential in analyzing larger community trends, including crime and resulting mobility.

Graduation rates are a key metric for assessing school performance and student success. A high school diploma is a critical milestone that opens doors to higher education, career opportunities, and economic stability. Graduation rates can vary significantly across schools and student populations, reflecting differences in academic support, resources, and community factors. Schools with higher graduation rates often have strong leadership, effective teaching practices, and supportive learning environments that help students stay engaged and motivated to complete their education. By analyzing graduation rates by school and demographic group, we can identify patterns, disparities, and areas for improvement within the public school system. Across the city the average four-year graduation rate is 67.3%.

The top ten schools by four-year graduation rate are as follows:

Table 4: Top 10 Schools by Four-Year Graduation Rate

School	Graduation Rate
Girard Academic Music Program	100.0
High School of Engineering and Science	100.0
Parkway Northwest High School	100.0
Lankenau High School	100.0
Arts Academy at Benjamin Rush	100.0
High School for Creative and Performing Arts	99.5
Julia R. Masterman School	99.1

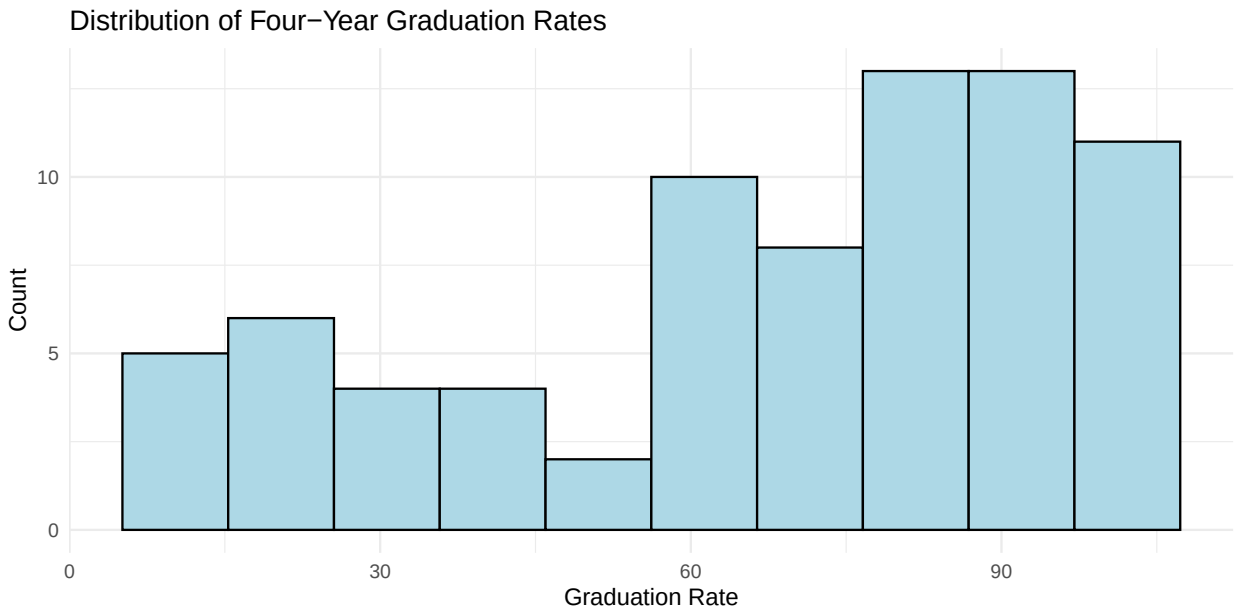
School	Graduation Rate
Science Leadership Academy	98.3
Central High School	98.1
Swenson Arts and Technology High School	98.0

Inversely, the lowest ten schools by four-year graduation rate can be seen here:

Table 5: Bottom 10 Schools by Four-Year Graduation Rate

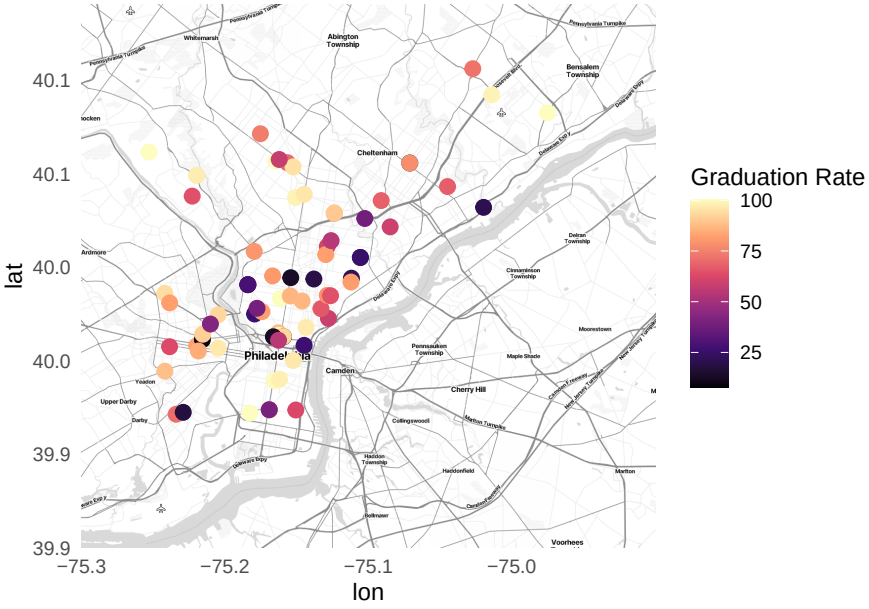
School	Graduation Rate
Ben Franklin High School EOP	8.06
Phila. Juv. Justice Services Ctr.	9.68
Northeast High School EOP	13.25
Gateway to College - Community College of Philadelphia	13.79
YESPhilly	14.00
One Bright Ray - Elmwood Campus	18.18
One Bright Ray - Fairhill Campus	19.61
Pennypack House School	21.05
Liguori Academy-Fortis	22.22
One Bright Ray - Simpson Campus	23.97

The distribution of graduation rates across schools can be visualized in a histogram:



As seen in the histogram, graduation rates among Philadelphia schools vary widely, with some schools achieving high rates of success while others struggle to graduate students on time. Factors such as school leadership, teacher quality, student engagement, and community support all play a role in determining graduation outcomes. Schools with lower graduation rates may face challenges related to student retention, academic performance, and social-emotional support, highlighting the need for targeted interventions and resources to help students succeed. By analyzing graduation rates at the school level, we can identify areas for improvement and develop strategies to increase student achievement and promote educational equity across the city.

The distribution of these schools across the city as well as a heatmap of their four-year graduation rates can be seen here:



It is important to understand how schools perform based on the location that they are in. By understanding the distribution of schools and their graduation rates, we can better understand how schools are performing in different neighborhoods. The top 5 neighborhoods by average graduation rate are as follows:

Table 6: Top 5 Neighborhoods by Average Graduation Rate

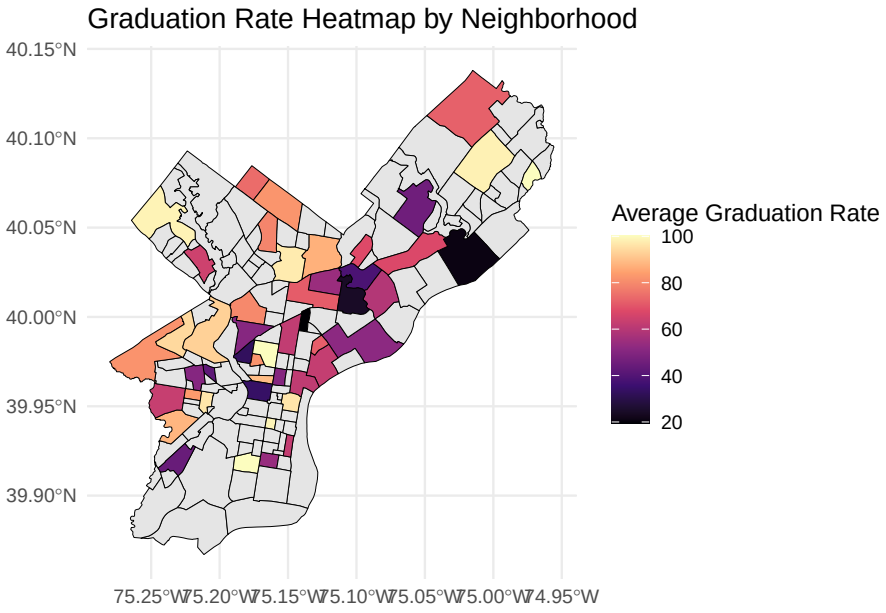
Neighborhood	Average Graduation Rate	Number of Schools
Girard Estates	100.0	1
Millbrook	100.0	1
North Central	100.0	1
Hawthorne	98.5	2
Upper Roxborough	98.3	2

And the bottom 5 neighborhoods by average graduation rate are as follows:

Table 7: Bottom 5 Neighborhoods by Average Graduation Rate

Neighborhood	Average Graduation Rate	Number of Schools
Fairhill	19.6	1
Holmesburg	21.1	1
Juniata Park	24.5	2
Brewerytown	32.9	2
Logan Square	34.2	2

A heatmap of the average graduation rates by neighborhood can be seen here:



Now that there is a good understanding of the city of Philadelphia, the distribution of crime and schools, and the relationship between schools and crime, we can move on to the analysis of the relationship between the two.

3 Regressions

We are going to start by running a simple regression to see if there is a relationship between a school's graduation rate and the total number of crimes within a 1-mile radius of the school.

The results of the simple regression show that there is not a significant relationship between a school's graduation rate and the total number of crimes within a 1-mile radius of the school. The p-value for the crime_total variable is 0.37, which is above the typical threshold of 0.05 for statistical significance. This suggests that the number of crimes near a school does not have a strong impact on the school's graduation rate. However, it is important to note that this is a simple regression model and does not account for other potential factors that could influence the relationship between crime and education.

We then investigated whether violent and non-violent crime rates influence the likelihood of a school achieving a high graduation rate (above 70%). A logistic regression model was applied to predict the likelihood of a school achieving a high graduation rate based on the independent variables of violent crime rate and total crime (which includes both violent and non-violent offenses). The dependent variable, High_Grad, is a binary indicator representing whether a school has a high graduation rate (>70%).

The results of the analysis indicate that none of the predictors were statistically significant ($p > 0.05$), likely due to the limited number of variables included in the model. The Akaike Information Criterion (AIC) for the model was 78.446, providing a measure of model fit relative to other possible models. The model predicts probabilities for High_Grad with a classification threshold of 0.7, meaning that a probability greater than or equal to 0.5 is classified as 1 (high graduation rate), while probabilities below 0.7 are classified as 0. Although an increase in violent crime rate appears to reduce the odds of achieving a high graduation rate, the p-value (0.210) indicates that this effect is not statistically significant. As a result, we cannot confidently conclude that violent crime has a real impact on graduation rates.

On top of that, the confusion matrix provides insight into the performance in predicting whether a school has a high graduation rate. The model correctly identified 18 low-graduation schools (true negatives) and

2 high-graduation schools (true positives). However, it also misclassified 17 high-graduation schools as low-graduation (false negatives) and incorrectly predicted 1 low-graduation school as high-graduation (false positive). This results in an accuracy of 52.63%, which is only slightly better than random guessing. The model demonstrates high specificity (94.74%), meaning it is very effective at correctly identifying low-graduation schools. However, its sensitivity is extremely low (10.53%), indicating that it fails to correctly identify most high-graduation schools. In other words, when a school truly has a high graduation rate, the model only detects it 10.53% of the time, making it unreliable for identifying high-performing schools.

Additionally, the precision (66.67%) suggests that when the model does predict a high graduation rate, it is correct about two-thirds of the time. However, the high number of false negatives significantly impacts its usefulness, as many high-graduation schools are being misclassified as low-graduation. The model is heavily biased towards predicting low-graduation schools correctly

The analysis suggests that while there may be a negative relationship between violent crime and high school graduation rates, the current model does not provide statistically significant evidence to support this claim. The lack of significance could be attributed to a limited number of predictor variables, potential non-linear relationships not captured in the model, or the need for additional socioeconomic factors to enhance predictive accuracy. To improve the model and gain better insights, we recommend incorporating additional predictors, such as school funding, poverty rates, and student-teacher ratios. Furthermore, exploring non-linear regression techniques or machine learning models may help capture more complex relationships, and increasing the sample size could improve statistical power. Further research with expanded datasets and alternative modeling approaches could yield more conclusive findings on the impact of crime rates on educational outcomes.

4 Conclusion

4.1 Future Work

In a perfect world, we have access to as much data as we would like, whenever we would like it. Unfortunately, due to plenty of constraints, things do not work that way. In the future, we would like to have access to more data to better understand the relationship between crime and education in Philadelphia. This could include data on school funding, student-teacher ratios, poverty rates, and other socioeconomic factors that could influence educational outcomes. By expanding the scope of the analysis and incorporating more data sources, we can gain deeper insights into the factors that impact crime rates and graduation rates in Philadelphia, ultimately informing policy decisions and interventions to improve public safety and educational outcomes in the city.

4.2 Final Statement

In conclusion, this project aimed to explore the relationship between crime rates and school performance in Philadelphia. By analyzing crime incidents and school graduation rates, we sought to identify patterns, trends, and potential correlations between these two critical aspects of urban life. Through exploratory data analysis and various types of regression models, we gained insights into the distribution of crime, school performance metrics, and their spatial relationships across neighborhoods in Philadelphia. While our initial regression models did not yield statistically significant results, they provided a foundation for further research and analysis to better understand the complex interplay between crime and education in the city.

Appendix

4.3 Data dictionary

A detailed summary of the variables in each data set follows:

4.3.1 OpenData Philly - Uncleaned Datasets

4.3.1.1 Crime Incidents

Source: https://metadata.phila.gov/#home/datasetdetails/5543868920583086178c4f8e/representationdetails/570e7621c03327dc14f4b68d/?view_287_page=1

Description: Part 1 & Part 2 Crime Incidents from the Police Department's INCT system with generalized UCR codes and addresses rounded to the hundred block. These counts may not coincide exactly with data that is submitted to the Uniformed Crime Reporting (UCR) system.

- the_geom - Unique identifier
- cartodb_id - Unique identifier
- the_geom_webmercator - Unique identifier
- objectid - Unique identifier
- dc_dist - A two character field that names the District boundary
- psa - A single character field that names the Police Service Area Boundary
- dispatch_date_time - The date and time that the officer was dispatched to the scene
- dispatch_date - Dispatch date formatted as a string
- dispatch_time - Dispatch time formatted as a string
- hour_ - The generalized hour of the dispatched time
- dc_key - The unique identifier of the crime that consists of Year + District + Unique ID
- location_block - The location of crime generalized by street block
- ucr_general - The rounded crime code
- text_general_code - The generalized text for the crime code
- point_x - Longitude of crime location
- point_y - Latitude of crime location
- lat - Latitude of crime location
- lng - Longitude of crime location

4.3.1.2 School Department of Philadelphia

Source: https://www.philasd.org/research/#opendata

Description: This longitudinal open data file includes information about the graduation rates for schools broken out by: graduation rate type (four-year, five-year, or six-year), demographic category (EL status, IEP status, Economically Disadvantaged Status, Gender, or Ethnicity), and 9th grade cohort. Students are attributed to the last school at which they actively attended in the respective graduation window, which ends on September 30 each year. Students are classified as EL, as having an IEP, and/or economically disadvantaged if they were designated as such at any point during their high school career.

- cohort - School year that freshman began
- schoolid_ulcs - Unique school identifier
- school_name - School name
- sector - Type of school (Two levels - District and Alternative)
- rate_type - Graduation rate type
- group - Grouping of students
- subgroup - Subgrouping of students
- denom - Number of students in cohort

- num - Number of students that graduated in cohort
- score - Percentage of students that graduated in cohort
- lat - Latitude of school location
- long - Longitude of crime location

4.3.1.3 Neighborhood GeoJSON

Source: https://opendataphilly.org/datasets/philadelphia-neighborhoods/

Description: This dataset includes neighborhood boundaries for 150+ neighborhoods in Philadelphia. The data was gathered from a mix of publicly available maps, including from the City of Philadelphia, the City Archives, the Philadelphia Inquirer, and user feedback.

- NAME - Neighborhood name
- LISTNAME - Neighborhood name
- MAPNAME - Neighborhood name
- Shape_Leng - Length of shape of neighborhood boundary
- Shape_Area - Total area inside shape of neighborhood
- geometry - Polygon coordinates of neighborhood boundary

4.3.2 Team One Data - Cleaned Datasets

4.3.2.1 Crime Incidents by Neighborhood

Description: This clean dataset contains crime incident records from 2023 categorized by neighborhood, including details on crime type, date, location, and police district, making it useful for analyzing spatial and temporal crime patterns.

- the_geom - Unique identifier
- cartodb_id - Unique identifier
- the_geom_webmercator - Unique identifier
- objectid - Unique identifier
- dc_dist - A two character field that names the District boundary
- psa - A single character field that names the Police Service Area Boundary
- dispatch_date_time - The date and time that the officer was dispatched to the scene
- dispatch_date - Dispatch date formatted as a string
- dispatch_time - Dispatch time formatted as a string
- hour_ - The generalized hour of the dispatched time
- dc_key - The unique identifier of the crime that consists of Year + District + Unique ID
- location_block - The location of crime generalized by street block
- ucr_general - The rounded crime code
- text_general_code - The generalized text for the crime code
- point_x - Longitude of crime location
- point_y - Latitude of crime location
- NAME - Neighborhood name
- LISTNAME - Neighborhood name
- MAPNAME - Neighborhood name
- Shape_Leng - Length of shape of neighborhood boundary
- Shape_Area - Total area inside shape of neighborhood
- geometry - Polygon coordinates of neighborhood boundary

4.3.2.2 School Data

Description: This clean dataset contains school performance metrics from 2023, including four year graduation rates by school, sector, and demographic group, along with geographical coordinates for spatial analysis.

- cohort - School year that freshman began
- schoolid_ulcs - Unique school identifier
- school_name - School name
- sector - Type of school (Two levels - District and Alternative)
- rate_type - Graduation rate type
- group - Grouping of students
- subgroup - Subgrouping of students
- denom - Number of students in cohort
- num - Number of students that graduated in cohort
- score - Percentage of students that graduated in cohort
- lat - Latitude of school location
- long - Longitude of crime location

4.3.2.3 School Data with Neighborhood

Description: This clean dataset contains school performance metrics from 2023, including four year graduation rates by school, sector, and demographic group, along with geographical coordinates for spatial analysis and neighborhood information.

- cohort - School year that freshman began
- schoolid_ulcs - Unique school identifier
- school_name - School name
- sector - Type of school (Two levels - District and Alternative)
- rate_type - Graduation rate type
- group - Grouping of students
- subgroup - Subgrouping of students
- denom - Number of students in cohort
- num - Number of students that graduated in cohort
- score - Percentage of students that graduated in cohort
- NAME - Neighborhood name
- LISTNAME - Neighborhood name
- MAPNAME - Neighborhood name
- Shape_Leng - Length of shape of neighborhood boundary
- Shape_Area - Total area inside shape of neighborhood
- geometry - Coordinates of school location

4.4 Data Prep for EDA

The type of crime is very important for grouping incidents. UCR code is grouped as a factor for easier analysis.

```
# Convert crime code to factor
crime <- crime %>%
  mutate(ucr_general = as.factor(ucr_general))
```

Some of the school scores are missing, so any that cannot be converted to numeric will be removed.

```
# Filter schools where 'score' is numeric
schools <- schools %>%
  filter(!is.na(as.numeric(score))) %>%
  mutate(score = as.numeric(score))
```

In order to plot, the school's GPS coordinates have to be split.

```r
# Split out the GPS coordinates
schools <- schools %>%
  mutate(geometry = gsub("c\\(|\\)", "", geometry)) %>%
  separate(geometry, into = c("lat", "long"), sep = ", ", convert = TRUE) %>%
  st_as_sf(coords = c("lat", "long"), crs = 4326)
```

4.5 Data cleaning process

4.5.1 Putting crime incidents into neighborhoods

In order to analyze crime incidents by neighborhood, we needed to spatially join the crime incidents data with the neighborhood boundaries. This was done using the `sf` package in R. The process involved reading in the neighborhood boundaries and crime incidents data, converting the crime incidents to spatial points, and then performing a spatial join to assign each crime incident to a neighborhood. The resulting dataset was then saved as a new CSV file for further analysis.

```r
# Load in data
neighborhood <- st_read('data/philadelphia-neighborhoods.geojson')
crime <- read.csv('data/crime-incidents-2023.csv')

# Filter out missing coordinates in crime
crime <- crime %>% filter(!is.na(lat) & !is.na(lng))

# Convert crime to sf
crime_sf <- st_as_sf(crime, coords = c('lng', 'lat'), crs = 4326)

# Make sure both have same CRS
neighborhood <- st_make_valid(neighborhood)

# Simplify geometries
neighborhood <- st_simplify(neighborhood, dTolerance = 0.001)

# Spatial join to get neighborhood for each crime
crime_w_neighborhoods <- st_join(crime_sf, neighborhood, join = st_intersects)

# Save to a new csv
#write.csv(crime_w_neighborhoods, 'data/crime-by-neighborhood.csv')
```

4.5.2 Combining school coordinates with graduation rates

In order to begin working with the school data, we had to combine the school coordinates with the graduation rates. This was done by reading in the school list and graduation rates data, joining them on the school ID, and then splitting the GPS coordinates into separate latitude and longitude columns. Also, given time constraints on this project, the dataset was reduced to only include four year graduation rates. The resulting dataset was then saved as a new CSV file for further analysis.

```r
# Load in files we need
grad_rates <- read.csv('data/SDP_Graduation_rates_school_S_2024-04-01.csv')
school_list <- read.csv('data/2022-2023 Master SChool List (20230110).csv')

# First, let's clean up the graduation rates
```

```r
# We only want the graduating class of 2023, 4 year grad rates, for all students
grad_rates <- grad_rates %>%
  filter(cohort == '2019-2020',
         rate_type == '4-Year Graduation Rate',
         group == 'All Students')

# Next let's bring in the latitude and longitude from the school list
school_data <- grad_rates %>%
  left_join(school_list %>% select('ULCS.Code', 'GPS.Location'),
            by = c('schoolid_ulcs' = 'ULCS.Code'))

# Split up GPS into lat and long
school_data <- school_data %>%
  separate(GPS.Location, c('lat', 'long'), sep = ', ', convert = TRUE)

# Drop empty coordinate
school_data <- school_data %>%
  filter(!is.na(lat))

# Save as csv
#write.csv(school_data, 'data/school_data.csv', row.names = FALSE)
```

4.6 Putting schools into neighborhoods

```r
# Merge school locations with spatial neighborhood data
# Filter out missing coordinates in schools
schools <- schools %>% filter(!is.na(lat) & !is.na(long))

# Convert crime to sf
schools_sf <- st_as_sf(schools, coords = c('long', 'lat'), crs = 4326)

# Make sure both have same CRS
neighborhoods <- st_make_valid(neighborhoods)

# Simplify geometries
neighborhoods <- st_simplify(neighborhoods, dTolerance = 0.001)

# Spatial join to get neighborhood for each crime
schools_w_neighborhoods <- st_join(schools_sf, neighborhoods, join = st_intersects)

# Save to a new csv
#write.csv(schools_w_neighborhoods, 'data/crime-by-neighborhood.csv')
```

4.7 Aggregate Crime Data By School Location

```r
# Load datasets (update file paths as needed)
crime_data <- read.csv("data/raw/crime-incidents-2023.csv")
school_data <- read.csv("data/clean/school_data.csv")
```

```r
# Remove missing values
crime_data <- na.omit(crime_data)
school_data <- na.omit(school_data)

# Aggregate Crime Data by School Location
# Create a matrix of crime type counts within a 1-mile radius of each school

# Function to count crimes within 1-mile (~0.016 degrees) radius per school
count_crimes_near_school <- function(lat, lon, crime_df) {
  crime_subset <- crime_df %>%
    filter(sqrt((lat - crime_df$lat)^2 + (lon - crime_df$lng)^2) < 0.016)
  return(nrow(crime_subset))
}

school_data$crime_count <- mapply(count_crimes_near_school, school_data$lat, school_data$long, MoreArgs

# Dummy Variable Crime Categories
crime_data$crime_type <- as.factor(crime_data$text_general_code)
crime_matrix <- model.matrix(~ crime_type - 1, data = crime_data)  # Dummy encoding

# Aggregate the one-hot encoded crimes by school proximity
school_crime_features <- school_data %>%
  select(schoolname, lat, long, score) %>%
  mutate(crime_total = school_data$crime_count)

# Add aggregated crime types per school
for (crime_col in colnames(crime_matrix)) {
  school_crime_features[[crime_col]] <- sapply(1:nrow(school_data), function(i) {
    lat <- school_data$lat[i]
    lon <- school_data$long[i]
    sum(crime_matrix[sqrt((crime_data$lat - lat)^2 + (crime_data$lng - lon)^2) < 0.016, crime_col])
  })
}
```

4.8 Create Crime Counts for Logistic Regression

```r
# Load datasets
crime_data <- read.csv("data/raw/crime-incidents-2023.csv")
school_data <- read.csv("data/clean/school_data.csv")

table(crime_data$text_general_code)  # Checking crime categories

# Define binary graduation rate variable (High = 1, Low = 0)
graduation_threshold <- 80  # Example threshold for high graduation rate
school_data$High_Grad <- ifelse(as.numeric(school_data$score) >= graduation_threshold, 1, 0)

# Define crime categories
violent_crimes <- c("Homicide", "Aggravated Assault", "Robbery", "Rape")
crime_data$Violent <- ifelse(crime_data$text_general_code %in% violent_crimes, 1, 0)

# Remove rows with missing latitude or longitude
crime_data <- crime_data %>% drop_na(lat, lng)
```

```r
# Convert to spatial data frames
crime_sf <- st_as_sf(crime_data, coords = c("lng", "lat"), crs = 4326)
school_sf <- st_as_sf(school_data, coords = c("long", "lat"), crs = 4326)

# Buffer schools to create a 1 km radius for crime aggregation
school_buffers <- st_buffer(school_sf, dist = 1000)  # 1000 meters = 1 km

#Spatial join: Count crimes within each school's buffer
crime_counts <- st_join(school_buffers, crime_sf) %>%
  group_by(schoolid_ulcs) %>%
  summarize(
    Violent_Crime_Rate = mean(Violent, na.rm = TRUE),
    Non_Violent_Crime_Rate = 1 - mean(Violent, na.rm = TRUE),
    Total_Crime = n()
  )
```