Crime in Philadelphia

Final Report

08/12/2022

CompassRed

Intern Mission Project

Akshay Jain, Luke Halko, Lindi Mukurazita, Sunita Barik

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Overview

With a crime rate of 39 per one thousand residents, Philadelphia has one of the highest crime rates in America compared to all communities of all sizes - from the smallest towns to the very largest cities.

Problem Statement

According to a NeighborhoodScout analytics report one's chance of becoming a victim of either violent or property crime here is one in 25. Within Pennsylvania, more than 95% of the communities have a lower crime rate than Philadelphia.

Proposed Solution

As per the analysis in our project we can see that poverty, population and homelessness are highly related to crime. Therefore, the government should act upon the factors to decrease the crime rate in Philadelphia.

Report Summary

The project includes detailed analysis of the crime in Philadelphia and its factors. It shows a high correlation between the factors population, poverty, homelessness, and slight correlation between weather but no correlation with temperature.

INTRODUCTION

Philadelphia consistently ranks above the national average in terms of crime, especially violent offenses. It has the highest violent crime rate of the ten American cities with a population greater than 1 million residents as well as the highest poverty rate among these cities. It has been included in real estate analytics company NeighborhoodScout's "Top 100 Most Dangerous Cities in America" list every year since it has been compiled. Much of the crime is concentrated in the North, West, and Southwest sections of the city.

DATA

Crime in Philadelphia is an umbrella project which includes multiple datasets. The main dataset is about crime in Philadelphia acquired from the Philadelphia Police department. The data used to find the factors of crime are poverty, population, homelessness, temperature, weather.

Dataset Description

Data acquisition is the process of sampling signals that measure real-world physical conditions and converting the resulting samples into digital numeric values that can be manipulated by a computer.

Crime Dataset

The crime dataset contains crime records recorded by the Philadelphia Police Department from 2016-2021.

Source: OpenDataPhilly.org

Crime Attributes and Data Types

COL NAME	DATA TYPE	DESCRIPTION
zip code	Ordinal	zip code associated with lat/lon coordinates
psa	Nominal	police service areas
datetime	Ordinal	date & time of police dispatch (YYYY-MM-DD hh:mm:ss)
date	Ordinal	date of police dispatch (YYYY-MM-DD)
time	Ordinal	time of police dispatch (hh:mm:ss)
hour	Numerical	current hour when crime was committed (integer)
Weekday	Nominal	day of the week
location	Nominal	street name and block number of crime location (e.g. ''5400 BLOCK ARCH ST")
crime_type	Nominal	categorical text description of the crime
crime_category	Nominal	generalized/reduced categories for crime type
lat	Numerical	x coordinate for location of crime
lng	Numerical	y coordinate for location of crime

Dropped columns: objectid, dc_dist, psa, dc_key, point_x, point_y

Sample: Crime Data

dispatch_date_time	location_block	text_general_code	lat	lng
2016-12-21 21:29:00	6600 BLOCK ESSINGTON AVE	Robbery No Firearm	39.91443023	-75.22059229
2016-12-27 1:47:00	6600 BLOCK ESSINGTON AVE	Robbery Firearm	39.91443023	-75.22059229

Weather Dataset

The weather dataset contains data over five years of hourly weather data for the City of Philadelphia from 2016-2022.

Source: timeanddate.com

Weather Attributes and Data Types

COL NAME	DATA TYPE	DESCRIPTION
date	Ordinal	date of weather recording (M-D-YYYY)
time	Ordinal	hour/minute of weather recording (e.g. "1:54 AM")
temp_f	Numerical	current temp at time of recording in Fahrenheit (e.g. "41 °F")
temp_c	Numerical	current temperature in degrees Celsius
weather	Nominal	text description of weather (categorical)
wind_mph	Numerical	current wind speed in miles per hour
humidity	Numerical	current humidity percentage
barometer_in_hg	Numerical	current air pressure recording in Inches of Mercury (or "Hg)
visibility_miles	Numerical	visibility estimate in miles (needs more info)

Sample: Weather Data

date	time	temp	weather	wind	humidity	barometer	visibility
1-1-2016	12:54 am\nFri, Jan 1	42 °F	Mostly cloudy.	14 mph	55%	30.11 "Hg	10 mi
1-1-2016	1:54 am	41 °F	Mostly cloudy.	12 mph	55%	30.13 "Hg	10 mi

Population Dataset

This dataset includes the population data from the Census 2016-2020. The dataset has columns United States, Philadelphia, and each Philadelphia zip code. Each row has different demographic categories, including age, race, and education.

Source: 2020 Census

Population Attributes and Data Types

COL NAME	DATA TYPE	DESCRIPTION
zip	Ordinal	a column for each zip code in Philadelphia
population	Numerical	Total population in zip code
year	Ordinal	Year when census was recorded

Sample: Population Data

zip	population	year
19102	4,583	2016
19103	22,564	2016

Poverty Data

This dataset contains the poverty data from the Census 2016-2020. The original poverty dataset contained a single row for the population of each variable. We extracted this into its own dataset, see more in Data Transformation

Source: 2020 Census

Poverty Attributes and Data Types

COL NAME	DATA TYPE	DESCRIPTION
label	Nominal	demographic grouping (age group, race, gender, etc)
United States	Numerical	metrics for United States as a whole
Philadelphia	Numerical	metrics for just philadelphia
zip codes	Ordinal	a column for each zip code in philadelphia

Sample: Poverty Data

Label	United States	Philadelphia	19102	19103	19104	…
Under 18 years	**20.30%**	**36.00%**	**0.00%**	**3.80%**	**48.70%**	…
18-34 years	**18.10%**	**26.40%**	**18.40%**	**14.30%**	**53.90%**	…

Homelessness Dataset

This dataset contains the data of homeless shelters across the United States from the year 2016 to 2021.

Source: The United States Department of Housing and Urban Development
https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/

Homelessness Attributes and Data Types

COL NAME	DATA TYPE	DESCRIPTION
CoC State	Nominal	Continuum of Care State
CoC/ID	Ordinal	Continuum of Care ID
HudNum	Nominal	Housing and Urban Development Number
Status	Nominal	Status of information provided ;in this case is "Submitted"
Year	Nominal	Year of collection
Organization ID	Ordinal	ID associated with the Organization Name
Organization Name	Nominal	Organization with the most direct responsibility of administering and carrying out an activity
HMIS Org ID	Ordinal	Homelessness Management Information System Organization ID
Project ID	Ordinal	ID associated with the Project Name
Project Name	Nominal	the type of housing provided to the homeless
HMIS Project ID	Ordinal	Homelessness Management Information System Project ID
HIC Date	Ordinal	Housing Inventory Count Date
Project Type	Nominal	Type of project under Continuum of Care: ES-Emergency Shelter;TH-Transitional Housing;RRH-Rapid Re-Housing;PSH-Permanent Supportive Housing;OPH-Other Permanent Housing;SH-Safe Haven
Geocode	Ordinal	Housing and Urban Development GeoCode
HMIS Participating	Nominal	Homelessness Management Information System Participating
Inventory Type	Nominal	Type of Inventory:- C- Child Support; U-Unemployment benefits
Target Population	Nominal	in this instance the homeless population
Zip Code	Ordinal	Zip Code of the homelessness observations

Sample: Homelessness Data

RowNum	CoC State	HudNum	Organization Name	Total Beds	zip
274038	**AR**	**AR-512**	**Sanctuary Inc.**	**17**	**85018**
244299	**OR**	**OR-502**	**Community Works**	**12**	**97211**

Data Acquisition

Crime Dataset
Weather Dataset Data scraped using the Selenium Python package, code here .
Poverty Data Manually applied the filter for all the zip codes in Philadelphia County and downloaded the csv files for Poverty for the years 2016-2020.
Population Dataset Population was extracted from the poverty data. It had columns for population for each zip code.
Homelessness Dataset Homeless data was extracted from The United States Department of Housing and Urban Development by manually applying filters and downloaded the csv files for homeless for years 2016-2021.

DATA TRANSFORMATION

Time and Date Conversion

Extracted month, year, name of weekday and hour using datetime library in python.

Outlier Removal

Removed some of the outliers for latitude and longitude after the descriptive analysis of the crime dataset.

Crime Type Reclassification

Reduced crime type from 35 categories to 9, and then to a binary choice of violent or non-violent to design a binary classifier.

Weather Type Reclassification

Reduced weather type from 113 unique descriptions to 7 categories.

Crime Rates

Generated a new table representing crime rates per 1000 people based on zip code, population, and the crime category

year	zip	Robbery	Theft	Assaults	…	total	population
2016	**19142**	**19.18703809**	**28.1409892**	**30.91245026**	…	**137.3294486**	**28144**
2016	**19123**	**15.26507598**	**70.82445163**	**30.32386715**	…	**215.2925806**	**14543**

Latitude Longitude to Zip Code Conversion

The crime analysis is based on zip codes, so to convert the longitude and latitude present in the crime data to zip code we made use of googlemapsAPI library in python.

DATA ANALYSIS AND VISUALIZATION

Descriptive Statistics

Table 2: Descriptive Statistics for numerical Variables

	temperature	Total Beds	Under_poverty	population
mean	60.49	239.22	0.28	41121.93
std	17.70	207.19	0.11	17333.46
minimum	6.000000	0	0.05	4564

Table 3: Mode for Categorical Variables

	crime_datetime	crime_type	crime_location	crime_weekday	crime_category	crime_zipcode
count	745611	745611	745611	745611	745611	745611
unique	638061	33	54334	7	9	46
top	2016-01-21 15:30:00+00:00	Other Offenses	5200 BLOCK FRANKFORD AVE	Tuesday	Theft	19134
frequency	17	116338	2771	114085	191801	53021

Data Visualizations

Analysis of crime and the factors associated with it.

Analysis 1 Crime incidence in Philadelphia

An investigation of the crime incidence in Philadelphia was done by computing a metric of crimes per 1000 persons in a Zip Code by average population per year.

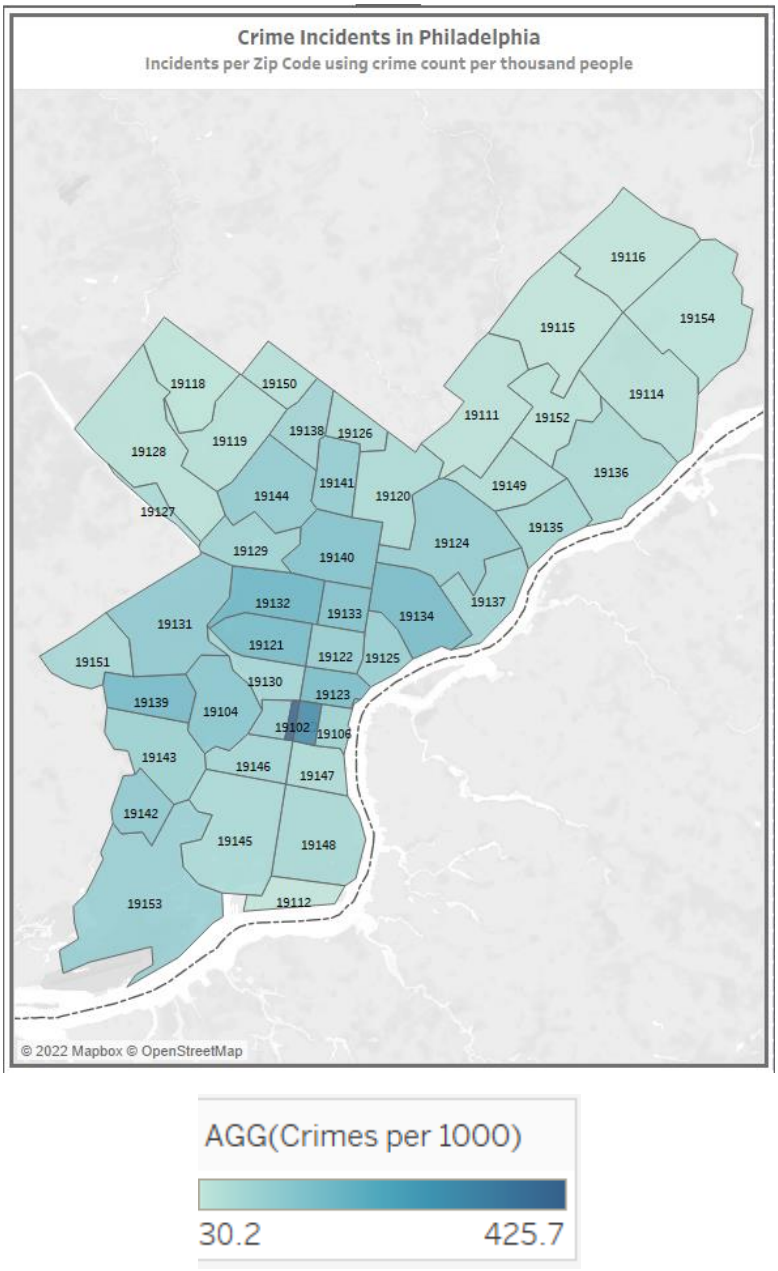

Figure 1: Crime Incidents In Philadelphia

Analysis 2 Yearly Trend

The total number of crimes per year were plotted and it was observed that the highest crime count was experienced in 2016 ,and the lowest in 2020, probably due to the Covid-19 pandemic. There was a slight percentage decrease in crime from 2016 to 2018.An increase of 4.37% was seen from 2018 to 2019 and a sharp decrease by 12.77% from 2019 to 2020 ,likely because of the Covid-19 pandemic.

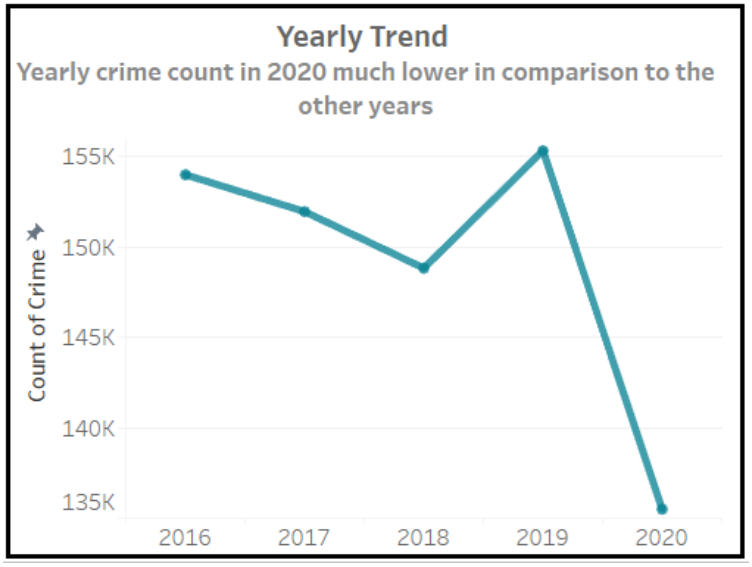

Figure 2: Yearly Trend of Crime in Philadelphia

Analysis 3 Monthly Trend

The crime count per 1000 persons was plotted for the months of the year. Generally, a peak of crime count was observed in the warmer months of each year in the dataset.

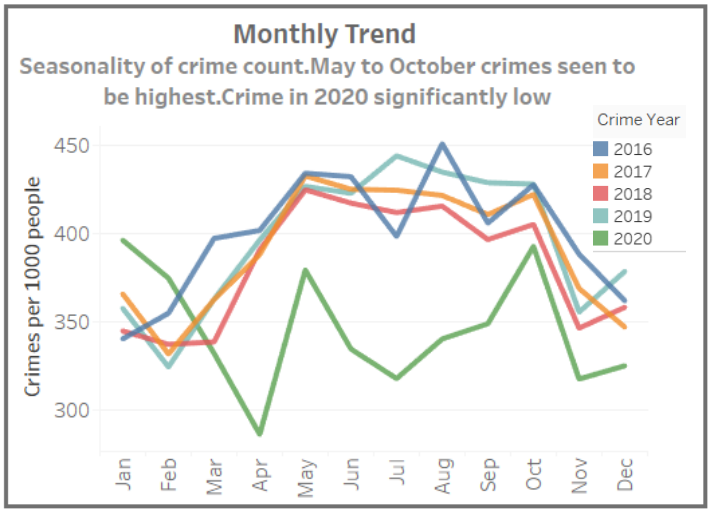

Figure 3: Monthly Trend of Crime in Philadelphia

Analysis 4 Hourly Trend

During the 5 years, an increase in crime count occurred in between 12.00pm to 6.00pm, and the lowest crime count between midnight and 6.00am.

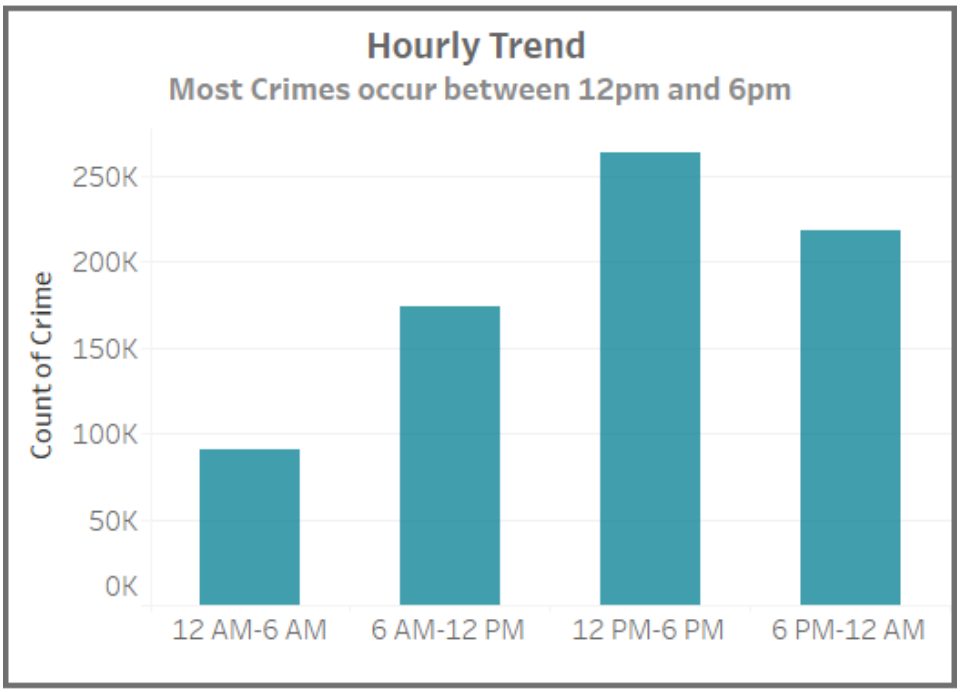

Figure 4: Hourly Trend of Crime in Philadelphia

Analysis of factors with Crime Incidence in Philadelphia

There are four factors which were analyzed and compared with crime incidence to check which were strongly correlated with incidence of crime.

The four factors are :

- Homelessness
- Poverty
- Temperature
- Weather Category

Analysis 6 Crime and Homelessness

Our dataset had information about shelter provided to homeless people for 35 Zip Codes. The number of homeless people per Zip Code were compared to the number of crimes. There is a positive correlation, that is, where there is more homelessness, there is more crime.

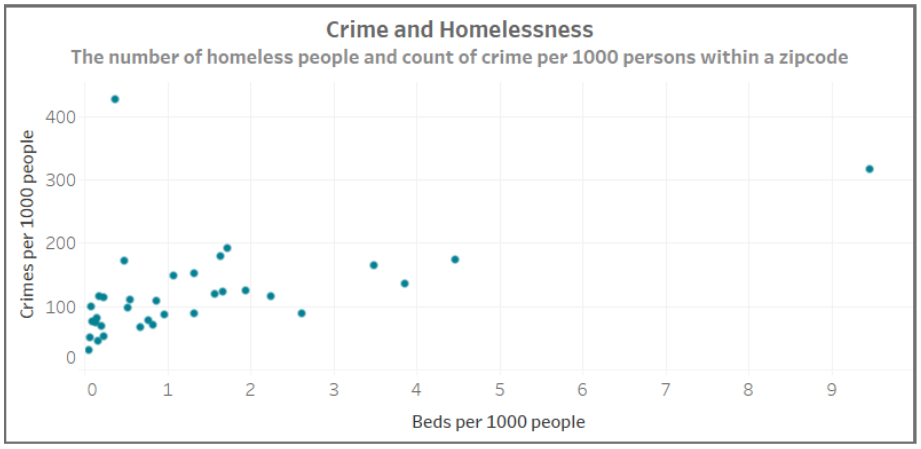

Figure 6 Correlation between Homelessness and Crime

Analysis 7 Crime and Poverty

We classified the areas which had less than 20 percent of the population living under the poverty line as areas with a lower percentage of poverty. The areas which had 20 percent or more of the population living under the poverty line are classified as areas with a higher concentration of poverty.

Since poverty and homelessness have a relationship, we see a similar trend, in that impoverished areas exhibit a higher rate of crime. Therefore, there seems to be a positive correlation between poverty and incidence of crime.

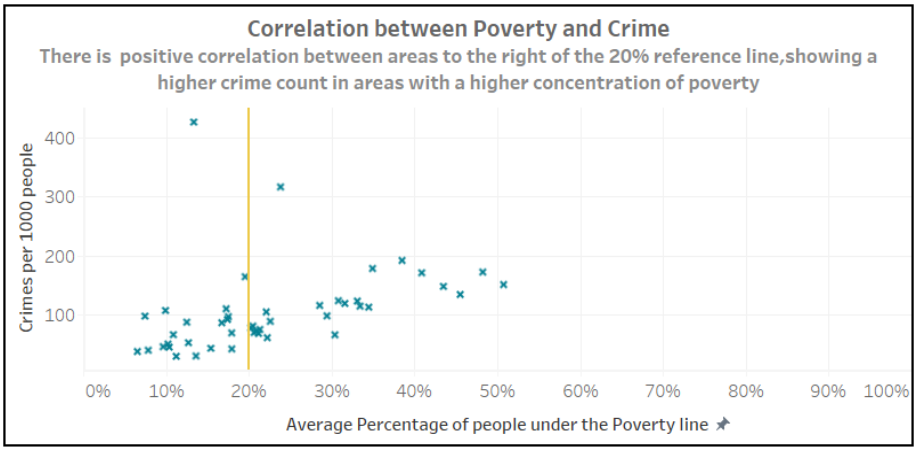

Figure 7 Correlation between Poverty and Crime

Analysis 8 Crime and Temperature

The average temperature does not have a conclusive effect on crime. Within the same city, there will not be many variations of temperature. Generally, there was more crime in warmer weather. However, in terms of average weather, the highest number of crimes in Zip Code 19102 and the lowest number of crimes was in Zip Code 19115.

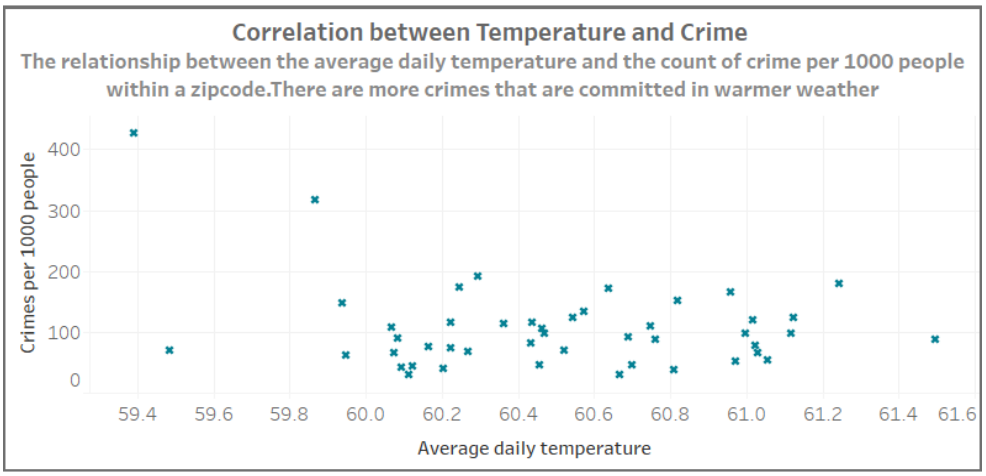

Figure 8 Correlation between Average Daily Temperature and Crime

Analysis 9 Crime and Weather

Most of the days in our dataset were either sunny or cloudy. This introduced a bias because a higher number of crimes were associated with either a cloudy day or a sunny day. A further analysis of crimes per day for each category of weather showed that an equal number of crimes occurred on days which were either sunny or cloudy as compared to the other weather categories, which are foggy, light rain, rainy, snowy and thunderstorms

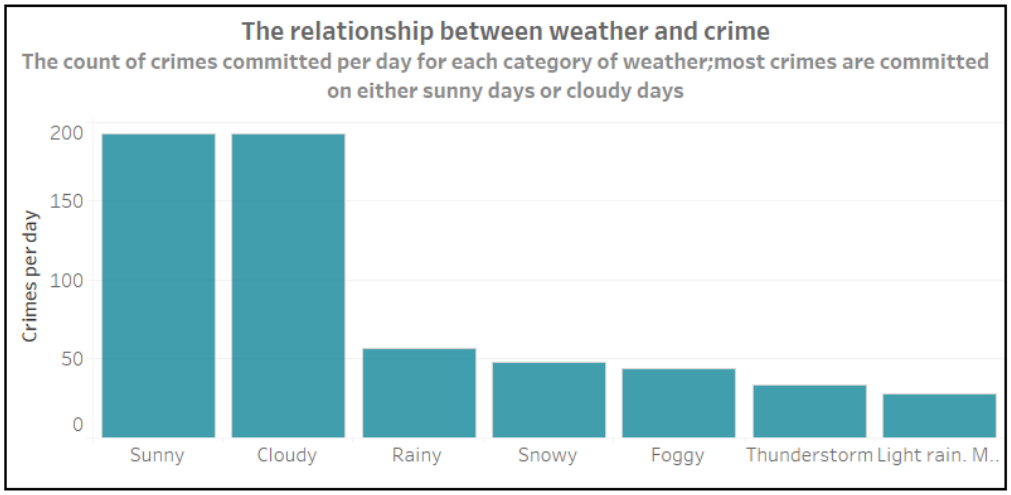

Figure 9 The relation between Weather Category and Crime Incidence

MODEL FITTING

After cleaning the data and deciding the features, we can go ahead and generate one, final dataset to use for modeling and analysis

Our model will attempt to predict the type of crime, first from a choice of 9 categories, and then from a binary choice of violent or non-violent crime

Columns from several other categories were merged with our final crime dataset to be used for modeling

COL NAME	DESCRIPTION
crime_datetime	date & time of police dispatch (YYYY-MM-DD hh:mm:ss)
crime_month	date of police dispatch (YYYY-MM-DD)
crime_year	time of police dispatch (hh:mm:ss)
crime_hour	current hour when crime was committed (integer)
crime_day	day of the month when the crime was committed
crime_location	street name and block number of crime location (e.g. ''5400 BLOCK ARCH ST")
crime_weekday	day of the week
crime_latitude	y coordinate of the crime's location
crime_longitude	x coordinate of the crime's location
crime_category	generalized/reduced categories for crime type
crime_zipcode	zip/postal code associated with latitude and longitude
population	number of residents in the given zip/postal code
under_poverty	percentage of residents under the poverty line
crime_type	categorical text description of the crime
weather	text description of the current weather
temp_f	current temperature in degrees fahrenheit
total_beds	number of beds at homeless shelters in the given zip code
crime_class	crime_type reduced to simply violent or non-violent crime
crime_class_violent	crime_class, expressed as 0 or 1

Analysis Methods

Missing Data

There were three columns which had missing data.

1. Weather category : Replaced NA with the most frequent weather category.
2. Temperature : Used forward fill to impute the missing values in the data.

Encoding

Made use of the get dummies function in python to create encoded variables. get dummies convert categorical data into dummy or indicator variables. This gives a similar result as one hot encoding.

Splitting Data

First step was to split the data into train and test. Our target variable is “crime_class_violent” which has binary values of 0 and 1. We split our data into 80:20 ratio into X_train, y_train, X_test, y_test.

Imbalanced Classification

The target variable in our dataset is not balanced. Count of label '1’: 179586 counts of label '0' is 416678. To avoid the class imbalance, we are going to use the SMOTE library from python which will generate synthetic samples which will not add any new information to the data, but it will solve the imbalance problem. Hence, our model will not be biased towards the majority class anymore.

Supervised Method: Decision Tree

We trained our model on a basic decision tree first, by fitting the X_train and y_train variables. The hyperparameters were set to default. The results that we got from this model are mentioned below.

Table 6: Decision Tree Performance

	Training	Testing
Accuracy	0.92	0.61
Precision	0.97	0.35
Recall	0.77	0.30
F1	0.86	0.32

Supervised Method: Random Forest

The model was hypertunes and n_esimators were set to 120 and this model was also trained on X_train and y_train. This is the data which we got after handling the imbalance.

The results about this model that we received are mentioned below

Table 7 Machines Model Performance

	Training	Testing
Accuracy	0.72	0.64
Precision	0.55	0.37
Recall	0.44	0.27
F1	0.49	0.31

Supervised Method: XGBoost

We have chosen XGBoost as our final model to train our data on. Hyperparameter tuning was done for XGBoost and we set the learning rate to 0.005 and n_estimators to 150. The results we got for test and train are mentioned below

Table 8 Machines Model Performance

	Training	Testing
Accuracy	0.73	0.64
Precision	0.57	0.39
Recall	0.43	0.27
F1	0.49	0.32

We are choosing XGBoost as our final algorithm for this model since it is giving similar accuracy for testing and training that means our model is not over fit. We also have a recall of 39% which means our model is 39% accurate to predict the violent crimes that can happen in Philadelphia County based on the zip code, hour, day of the week and month of the year.

Analysis Results

Table 9: Performance Comparison

	Decision Tree		Random Forest		XGBoost	
	train	test	train	test	train	test
Accuracy	0.92	0.61	0.72	0.64	0.73	0.64
Precision	0.96	0.34	0.55	0.37	0.57	0.39
Recall	0.77	0.31	0.44	0.27	0.43	0.27
F1	0.86	0.33	0.49	0.31	0.49	0.32

Table 10: Variables of Importance

Variable	F-score
temperature	69584
population	62145
under poverty	55044
crime_month_7	5672
crime_month_5	5588
crime_month_10	5283
crime_month_8	5172
crime_month_3	5061

SUMMARY AND RECOMMENDATION

Summary

Overall Statistics:

1. 30% of the total crimes recorded are violent crimes.
2. Assaults and Homicides cases were higher for Zip 19132 and 19133.
3. The average count of crimes is 13 per thousand residents in a zip code in Philadelphia.
4. Center City, 19102 had a crime rate of 4.26 thefts per day compared to 8.5 thefts per day for 19107.
5. After Feb 2020, there was a general dip in crime particularly in April.

Zip Code Specific Statistics:

1. Zip 19102: 1558 crimes committed vs Zip 19107: 3103 crimes committed.
2. Zip 19102: Theft 115.73 per thousands, Assault 51.57 per thousands and Homicide 0.41 per thousands.
3. Zip 19107: Theft 102.45 per thousands, Assault 39.47 per thousands.
4. Zip 19132: Theft 38.26 per thousands, Assault 57.10 per thousands and Homicide 1.21 per thousands.
5. Zip 19133: Theft 28.40 per thousands, Assault 35.16 per thousands and Homicide 0.94 per thousands.

Model:

1. Among Decision Tree, Random Forest and XG Boost; XGBoost model was the best fit with an accuracy of 64%, precision of 39% and a recall of 27%.
2. The XG Boost model can correctly predict 39% of occurrences of violent crime based on the Zip code, the hour of the day, the day of the week and the month of the year.

Recommendations

WHAT COMPASSRED CAN DO ?

As a part of Corporate Social Responsibility, companies can start corporate programs to help the homeless or people under poverty which can positively impact the company's reputation and help the society.

Challenges

Project Management:

1. Delayed Project Setup
2. Delayed Sprint deadlines

Data Acquisition:

3. Opioid Addiction Data
4. Substance Abuse Data

Data Quality Concerns:

5. Duplicate values
6. Many poorly documented columns in crime dataset
7. Too many distinct crime categories
8. Incorrect data entry
9. No supplementary data

REFERENCES

Crime Maps & Stats | Philadelphia Police Department

https://pypi.org/project/googlemaps/

http://philadelphiaofficeofhomelessservices.org/

https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/