Houston Crime Analysis

Preethu Manjunath
University of Colorado Boulder
Boulder, United States

Lohith Ramesh
University of Colorado Boulder
Boulder, United States

Raquel Yupanqui
University of Colorado Boulder
Boulder, United States

ABSTRACT

In the rapidly evolving urban landscape of Houston, analyzing crime data from 2019 to 2024 unveils crucial insights into the dynamics of urban crime and its profound impact on community welfare. This report delves into an exhaustive exploration of Houston's crime data, leveraging advanced data analysis techniques to dissect crime patterns, temporal trends, and socioeconomic correlations. By integrating crime data with meticulously web-scraped demographic information, our analysis offers a holistic view of the interplay between crime rates and socioeconomic factors across different neighborhoods. The analysis commenced with the aggregation and pre-processing of crime data sourced from the official Houston city website, spanning six years to construct a comprehensive dataset. This dataset underwent rigorous cleaning processes, including handling missing values, converting data types, and standardizing entries, to ensure accuracy and reliability. Furthermore, feature engineering and data reduction techniques were applied to distill key variables that shed light on crime trends and patterns. A pivotal aspect of our methodology was the enrichment of crime data with demographic variables such as population density, income levels, and unemployment rates. This integration enabled a multifaceted analysis, revealing nuanced insights into how socioeconomic disparities influence crime occurrences. Through a suite of visualizations, including time-series graphs, scatter map plots, and heatmaps, we illustrated the geographic distribution of crimes, temporal patterns, and the relationship between crime rates and demographic indicators. Key findings underscore the significance of socioeconomic factors in shaping crime dynamics, with particular emphasis on the correlation between economic hardship and increased crime rates. Geographic analysis identified specific hotspots of criminal activity, highlighting the need for targeted public safety interventions. This report underscores the critical role of data-driven approaches in understanding and mitigating urban crime. The insights garnered from this analysis advocate for informed policy-making and strategic planning to enhance public safety and address the underlying socioeconomic factors contributing to crime. Our findings serve as a call to action for stakeholders, including law enforcement, policymakers, and community organizations, to collaborate in fostering a safer and more equitable urban environment.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

Machine Learning, Data Science, Data Mining

ACM Reference Format:
Preethu Manjunath, Lohith Ramesh, and Raquel Yupanqui. 2024. Houston Crime Analysis. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Urban crime represents a significant challenge for major metropolitan areas worldwide, intricately linked to a myriad of socioeconomic factors that influence public safety and community stability. This study focuses on Houston, one of the United States' most dynamic and diverse cities, to conduct an exhaustive analysis of crime trends from 2019 to 2024. The primary aim is to leverage a data-driven approach to dissect the complex interrelations between crime rates and socioeconomic environments, thereby providing actionable insights for effective crime prevention and policy formulation.

Initiating this comprehensive research, the project team collected extensive crime statistics from the City of Houston's official online portals, ensuring a robust dataset that includes crime type, location, date, and time of incidents. To enrich the analysis, this crime data was augmented with socioeconomic and demographic information meticulously web-scraped from various reliable sources. This enriched dataset incorporates additional variables such as population density, income levels, unemployment rates, and educational attainment across different Houston neighborhoods, allowing for a multidimensional analysis of crime dynamics.

With the aid of advanced analytical tools and statistical methods, the study undertakes a systematic examination of the data to identify temporal and spatial patterns of crime occurrences. Moreover, machine learning techniques including Random Forest, K-Nearest Neighbors (KNN), AdaBoost, and XGBoost are employed to develop predictive models that not only forecast crime occurrences but also explore the impact of socioeconomic disparities on crime rates.

This introductory section sets the groundwork for a detailed exploration of the methods and techniques used during the research, provides an overview of the data preparation and preprocessing efforts, and outlines the analytical strategies implemented. Following this, the report presents the key findings, discusses their implications for crime prevention and urban planning, and offers recommendations for policymakers and community leaders aimed at enhancing public safety and fostering a more equitable urban environment.

By combining rigorous data analysis with insightful socioeconomic considerations, this study aspires to contribute to a more nuanced understanding of urban crime and to support data-driven decision-making in Houston's ongoing efforts to mitigate crime and enhance community welfare.

2 DATA COLLECTION/PREPARATION

The Houston Crime Analysis project embarked on a multifaceted journey to dissect urban crime through the lens of data-driven insights. Commencing in 2019 and extending through 2024, this endeavor aimed to unearth patterns, trends, and correlations within the vast landscape of crime data, enriched by socioeconomic and demographic dimensions. The project's backbone was the meticulous gathering, cleaning, and preparation of crime data juxtaposed with demographic insights, setting a robust foundation for nuanced analysis.

2.1 Data Sourcing

The primary dataset, encompassing detailed monthly crime reports by street and police beat, was sourced from the City of Houston's official website[1]. This rich repository of crime incidents provided a comprehensive canvas for the analysis. The crime data, segmented into yearly Excel files from 2019 to 2024, was methodically downloaded. Each file's initial inspection offered insights into the dataset's structure, revealing variables' range, data types, and potential quality issues such as missing values or inconsistencies. Utilizing Python libraries Pandas for data manipulation and Matplotlib/Plotly for visualization, the yearly data was loaded into separate Pandas DataFrames. This step was crucial for acclimating to the dataset's nuances, setting the stage for in-depth exploration and analysis.

```
import pandas as pd
import numpy as np
from bs4 import BeautifulSoup
import requests
import plotly.express as px
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
import plotly.graph_objects as go
```

	Incident	RMSOccurrenceDate	RMSOccurrenceHour	NIBRSClass	NIBRSDescription	OffenseCount	Beat	Premise	StreetName	City	ZIPCode	MapL
0	5619	2019-01-01	0	290	Destruction, damage, vandalism	1	9C30	Residence, Home (Includes Apartment)	SAN CARLOS	HOUSTON	77013	-9
1	17319	2019-01-01	0	35A	Drug, narcotic violations	1	7C10	Highway, Road, Street, Alley	EAST	HOUSTON	77020	-9
2	17319	2019-01-01	0	90D	Driving under the influence	1	7C10	Highway, Road, Street, Alley	EAST	HOUSTON	77020	-9
3	18119	2019-01-01	0	290	Destruction, damage, vandalism	1	16E40	Residence, Home (Includes Apartment)	LONE QUAIL	HOUSTON	77489	-9
4	19019	2019-01-01	0	520	Weapon law violations	1	NaN	Residence, Home (Includes	MELBOURNE	HOUSTON	77026	-9

3 PREPROCESSING

3.1 Handling Missing Values

Strategies were employed based on the data's nature and missingness. Notably, the street suffix column was omitted due to excessive missing values, and missing values in the "Beat" column were intelligently filled by referencing incident zip codes. Latitude and longitude coordinates served to fill missing zip codes, minimizing data loss.

```
region ={
    'Airport-Hoby Division - District 23':['23340','23250'],
    'Airport-IAH Division - District 21':['21110','21140','21150','21160','21170'],
    'Central Division - District 1,2':['1A20','1A30','1A40','1A50','2A10','2A20','2A30','2A40','2A50','2A60'],
    'Clear Lake Division - District 12':['12D10','12D20','12D30','12D40','12D50','12D60','12D70'],
    'Downtown Division - District 1':['1A10'],
    'Eastside Division - District 1':['11H10','11H20','11H30','11H40','11H50'],
    'Kingwood Division - District 24':['24C10','24C20','24C30','24C40','24C50','24C60'],
    'Midwest Division - District 10':['10F10','10F20','10F30','10F40','10F50','10F60'],
    'NorthBelt Division - District 22':['22B10','22B20','22B30','22B40'],
    'North Division - District 3,6':['3B10','3B30','3B40','3B50','6B10','6B20','6B30','6B40','6B60'],
    'Northeast Division - District 7,8,9':['7C10','7C20','7C30','7C40','8C10','8C20','8C30','8C40','8C50','8C60','9C10','9C20','9C30','9C40'],
    'Northwest Division - District 4,5':['4F10','4F20','4F30','5F10','5F20','5F30','5F40'],
    'South Central Division - District 18':['18H10','18H20','18H30','18H40','18H50','18H60','18H70','18H80'],
    'South Gessner Division - District 17':['17E10','17E20','17E30','17E40'],
    'Southeast Division - District 13,14':['13D10','13D20','13D30','13D40','14D10','14D20','14D30','14D40','14D50'],
    'Southwest Division - District 15,16':['15E10','15E20','15E30','15E40','16E10','16E20','16E30','16E40'],
    'Westside Division - District 19,20':['19G10','19G20','19G30','19G40','19G50','20G10','20G20','20G30','20G40','20G50','20G60','20G70','20G
    'Harris County Sheriffs Office':['HCSO'],
    'Harris County Constable's Office Precinct 5':['HCC5'],
    'OOJ':['OOJ']
}
```

```
def get_region(beat):
    for key, value in region.items():
        if beat in value:
            return key
    return None

# Apply the function to create the 'region' column
df['Region'] = df['Beat'].apply(get_region)
```

3.2 Data Type Conversion and Standardization

Data types were meticulously adjusted for analysis efficiency and accuracy. Zip codes were transitioned from floats to integers, and date strings to Datetime objects, facilitating accurate temporal analyses. Efforts ensured data consistency, particularly for textual data, and duplicate checks confirmed the dataset's quality.

3.3 Feature Engineering and Data Reduction

The dataset was enhanced by extracting meaningful features such as month and year from crime incident dates, enabling detailed pattern analysis. Irrelevant columns, like street type and number, were pruned to focus the analysis, streamlining the dataset for clarity and purpose.

RMSOccurrenceDate	month	year
2019-01-01	1	2019
2019-01-01	1	2019
2019-01-01	1	2019

3.4 Identify Attribute Types

In data analysis, identifying attribute types is essential as it determines the appropriate analytical approach for each variable. Attributes can be classified as nominal, where data represent categories without inherent order; ordinal, which have a clear sequence; interval, where the differences between data points are meaningful; and ratio, which have a true zero point, allowing for a full range of statistical operations. Understanding these types aids in selecting the correct visualization techniques and statistical tests, ensuring accurate and meaningful analysis outcomes. For example, crime incident data often contain nominal attributes like crime type and ratio attributes like incident counts, each requiring different analytical treatment. Correctly identifying these attribute types is crucial for data pre-processing, feature engineering, and subsequent modeling stages.

3.5 Compute Measures of Central Tendency and Dispersion

Measures of central tendency and dispersion are fundamental statistical tools used to summarize data. Central tendency measures, including the mean, median, and mode, provide a central value around which data points cluster. In contrast, measures of dispersion like the range, variance, and standard deviation quantify the spread of data points around the central value, indicating the variability within the dataset. Together, these measures offer a snapshot of the data's overall shape and spread, informing decisions and highlighting potential outliers or trends. For instance, in a crime dataset, the mean could reveal the average number of incidents, while the standard deviation could show the consistency of crime across different areas.

3.6 Visualize Data Distributions

Visualizing data distributions is a critical step in data analysis, as it provides a graphical representation of how data points are spread across different values. Common methods include histograms, box plots, and density plots, each highlighting various aspects of the data's distribution, such as skewness, kurtosis, and the presence of outliers. These visual tools help analysts and stakeholders quickly grasp the underlying patterns and anomalies within the data, which may not be immediately apparent from raw figures. For crime data, such visualizations can reveal the frequency of different crime types or the distribution of incidents over time, facilitating pattern recognition and strategic planning. Effective visualization of data distributions is therefore key to intuitive understanding and informed decision-making.

3.7 Assess Data Similarity

Assessing data similarity involves quantifying the likeness between data points, which is crucial for tasks like clustering, classification, and anomaly detection. Techniques such as Euclidean distance for continuous data or the Jaccard index for categorical data are often employed. In a crime database, assessing similarity might help identify clusters of similar crime types or detect areas with comparable crime statistics. This evaluation aids in understanding relationships within the data, facilitating predictive modeling and tailored intervention strategies. Overall, data similarity measures are instrumental in uncovering hidden patterns and associations in complex datasets.

3.8 Normalization

Normalization is the process of scaling individual data points to ensure consistency across different ranges for comparative and analytical purposes. It is especially important when dealing with features that vary in scale, units, or range, as it can impact the performance of many machine learning algorithms. Common methods include min-max scaling, which rescales the range of features to scale the range in [0, 1], and Z-score standardization, where data points are rescaled based on their mean and standard deviation. In the context of crime statistics, normalization allows for fair comparison across variables like income levels and crime rates. Effective normalization can significantly improve the robustness and accuracy of downstream analytical tasks.

3.9 Transformation

Data transformation involves converting data into a different format or structure, often to improve its suitability for analysis or model building. It can include operations like logarithmic transformations to stabilize variance or handle skewness, or more complex functions like Fourier transforms to identify frequencies within time series data. For example, transforming crime incident timestamps into cyclical features can capture the periodic nature of crime occurrences. Such transformations are pivotal for exposing meaningful patterns in the data, enhancing the predictive power of analytical models, and simplifying complex relationships between variables for better interpretability.

3.10 Integration with Demographic Data

The project expanded its analytical horizon by weaving in web-scraped demographic data from [2], adding layers of socioeconomic context. This comprehensive approach integrated data on population density, income levels, unemployment rates, and more, aligned with the crime data via geographical identifiers like ZIP codes. Any missing demographic values were filled in with the column average.

3.11 Creation of a Unified Dataset

The unification of annual crime data into a single DataFrame addressed the challenge of dataset consistency. This meticulous process involved standardizing column names, ensuring uniformity in categorical data, and aligning datetime formats. Demographic data integration was a critical step, enriching the crime dataset with socioeconomic insights. This phase required careful data alignment, particularly in matching geographical identifiers, to enable holistic analyses of crime within Houston's broader socioeconomic and demographic landscape. Post-merge, the dataset underwent additional feature engineering and verification steps to ensure data integrity and analytical readiness.

ZipCode	Population	BlackPercent	AsianPercent	WhitePercent	MedianAge	AgeDependRatio	OldAgeDependRatio	ChildDependRatio	HomeOwnershipPercent	AverageEarnings	PovertyPercent	UnemploymentRate
77036	72279	11.63	10.36	18.63	30.8	50.5	11.2	39.3	16.4	28349.0	27.36	0.056
77004	30936	49.07	6.08	31.64	33.3	35.9	10.9	25.0	39.5	41058.0	27.32	0.089
77092	36520	10.79	2.59	50.65	33.8	48.7	17.3	32.4	40.6	35974.0	20.83	0.058
77067	49023	10.65	9.86	53.34	33.9	48.9	17.7	31.2	31.8	46983.0	18.03	0.047
77007	42881	6.71	11.41	67.53	33.3	23.1	8.6	14.5	48.8	102750.0	5.79	0.025

4 CLEANING

4.1 Overview

The cleaning phase was a critical step in ensuring the integrity and reliability of the Houston Crime Data analysis. This process involved meticulous scrutiny and refinement of the dataset, addressing various issues such as missing values, inconsistent entries, and data type inaccuracies. Each cleaning action was tailored to preserve data quality while ensuring the analysis remained robust and meaningful.

4.2 Handling Missing Values

The dataset's integrity was challenged by missing data across several columns. Notably, the 'street suffix' column was excluded due to a high volume of missing values which, upon evaluation, were deemed non-essential for the analysis. For the 'Beat' column, missing values were addressed by cross-referencing and filling in data based on incident zip codes, demonstrating a nuanced approach to preserving data quality. Similarly, missing zip codes were cleverly inferred using latitude and longitude coordinates, showcasing the innovative problem-solving techniques employed during the cleaning process.

4.3 Data Type Conversion

Critical to the analysis was ensuring that each column's data type accurately reflected its content. Zip codes, initially recorded as float values, were converted to integer format to correct their representation and facilitate accurate geographical analysis. Date columns underwent transformation from string to Datetime format, enhancing the dataset's utility for time-series analysis and temporal trend identification.

4.4 Standardizing Values

The cleaning phase also emphasized the standardization of textual data, addressing variations in casing or formatting that could lead to inconsistencies in categorical analysis. This step ensured that data representation was consistent, thereby eliminating potential biases or errors in subsequent analyses.

4.5 Removing Duplicates

A thorough examination for duplicate records was conducted to prevent any distortion in the analysis outcomes. This proactive measure confirmed the dataset's initial quality, with no duplicates found, underscoring the data's readiness for in-depth analytical exploration.

4.6 Feature Engineering

Beyond cleaning, the dataset was enriched through feature engineering. Extracting month and year from crime incident dates exemplified the strategic enhancement of the dataset, enabling detailed analysis of temporal crime patterns and trends. This step not only added depth to the analysis but also illustrated the project's commitment to uncovering nuanced insights into crime dynamics in Houston.

	Incident	RMSOccurrenceDate	RMSOccurrenceHour	NIBRSClass	NIBRSDescription	OffenseCount	Beat	Premise	StreetName	City	...	Asi
0	5619	2019-01-01	0	290	Destruction, damage, vandalism	1	9C30	Residence, Home (Includes Apartment)	SAN CARLOS	HOUSTON	...	
1	17319	2019-01-01	0	35A	Drug, narcotic violations	1	7C10	Highway, Road, Street, Alley	EAST	HOUSTON	...	
2	17319	2019-01-01	0	90D	Driving under the influence	1	7C10	Highway, Road, Street, Alley	EAST	HOUSTON	...	
3	18119	2019-01-01	0	290	Destruction, damage, vandalism	1	16E40	Residence, Home (Includes Apartment)	LONE QUAIL	HOUSTON	...	
4	20519	2019-01-01	0	13A	Aggravated Assault	1	15E30	Residence, Home (Includes Apartment)	OSBY	HOUSTON	...	
...	...	...	...	...	...	...	...	...	...	...	...	
1255709	19131124	2024-01-31	17	90Z	All other offenses	1	9C30	Jail, Prison	BEAUMONT	HOUSTON	...	
1255710	14934624	2024-01-31	11	23F	Theft from motor vehicle	1	9C40	Residence, Home (Includes Apartment)	ROYALE	HOUSTON	...	
1255711	15165624	2024-01-31	17	23F	Theft from motor vehicle	1	9C40	Restaurant	EAST	HOUSTON	...	
1255712	15282624	2024-01-31	19	240	Motor vehicle theft	1	9C40	Parking Lot, Garage	MAXEY	HOUSTON	...	

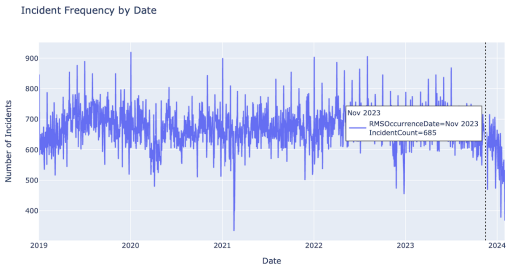
	Incident	RMSOccurrenceDate	RMSOccurrenceHour	NIBRSClass	NIBRSDescription	OffenseCount	Beat	Premise	StreetName	City	...	Asi
0	5619	2019-01-01	0	290	Destruction, damage, vandalism	1	9C30	Residence, Home (Includes Apartment)	SAN CARLOS	HOUSTON	...	
1	17319	2019-01-01	0	35A	Drug, narcotic violations	1	7C10	Highway, Road, Street, Alley	EAST	HOUSTON	...	
2	17319	2019-01-01	0	90D	Driving under the influence	1	7C10	Highway, Road, Street, Alley	EAST	HOUSTON	...	
3	18119	2019-01-01	0	290	Destruction, damage, vandalism	1	16E40	Residence, Home (Includes Apartment)	LONE QUAIL	HOUSTON	...	
4	20519	2019-01-01	0	13A	Aggravated Assault	1	15E30	Residence, Home (Includes Apartment)	OSBY	HOUSTON	...	
...	...	...	...	...	...	...	...	...	...	...	...	
1255709	19131124	2024-01-31	17	90Z	All other offenses	1	9C30	Jail, Prison	BEAUMONT	HOUSTON	...	
1255710	14934624	2024-01-31	11	23F	Theft from motor vehicle	1	9C40	Residence, Home (Includes Apartment)	ROYALE	HOUSTON	...	
1255711	15165624	2024-01-31	17	23F	Theft from motor vehicle	1	9C40	Restaurant	EAST	HOUSTON	...	
1255712	15282624	2024-01-31	19	240	Motor vehicle theft	1	9C40	Parking Lot, Garage	MAXEY	HOUSTON	...	
1255713	15350224	2024-01-31	21	240	Motor vehicle theft	1	9C40	Parking Lot	COOLWOOD	HOUSTON	...	

5 VISUALIZATIONS

Visualization highlights the crucial role visual analytics play in understanding complex datasets. Visualization transforms raw data into graphical representations that reveal underlying patterns, trends, and anomalies, making the data more accessible and interpretable. In the context of Houston crime data merged with demographic information, effective visualizations can illuminate the multifaceted nature of crime and its socio-economic correlations.

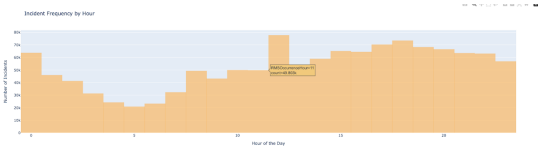
5.1 Daily Crime Incidents Over Time

This line graph illustrates the number of crime incidents reported each day over a specified time period. It allows analysts and policymakers to identify trends, such as increases or decreases in crime rates over time, and potentially correlate these trends with specific events, policy changes, or law enforcement initiatives.



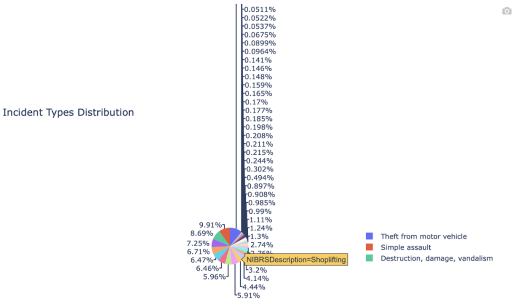
5.2 Geographic Distribution of Crime Incidents

A scatter map plot visualizes the geographic locations of crime incidents, using latitude and longitude for positioning and possibly color-coding by crime type. This map can reveal hotspots of criminal activity and help in allocating resources more effectively.



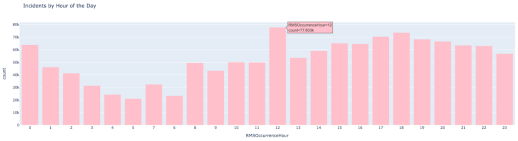
5.3 Distribution of Crime Types

A pie chart or bar graph shows the proportion of different crime types within the dataset, highlighting the most prevalent crimes. This visualization aids in understanding the primary concerns for law enforcement and community safety efforts.



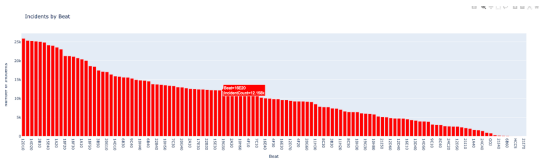
5.4 Crime Incidents by Hour of Day

This histogram displays the frequency of crime incidents across different hours of the day, revealing patterns in when crimes are most likely to occur. Such insights can guide patrolling and preventive measures.



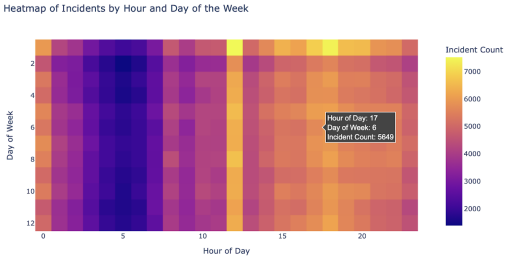
5.5 Incidents by Beat

The "Incidents by Beat" bar chart shows the distribution of crime incidents across different police beats. The visualization highlights considerable variation in incident counts, with some beats experiencing significantly higher numbers of incidents, indicating areas with potentially higher police activity or crime rates. This data can be instrumental for law enforcement to analyze resource allocation and strategize on crime prevention efforts.



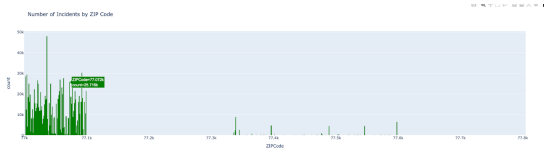
5.6 Heatmap of Incidents by Hour and Day of the Week

The heatmap you've provided shows the concentration of crime incidents throughout different hours of the day and days of the week. The varying colors represent the intensity of incidents, with warmer colors indicating higher frequencies. From this visualization, one can discern peak times for incidents, which could be critical for law enforcement and public safety planning. For example, there appears to be a higher density of incidents during certain hours, suggesting those times may require increased attention or resources. This kind of visualization is a powerful tool for identifying temporal patterns in crime data.



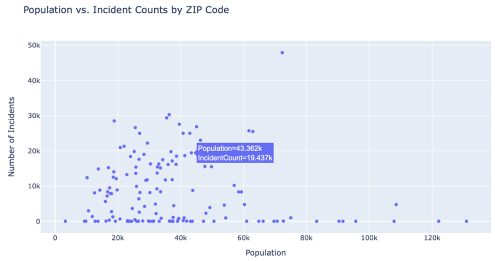
5.7 Number of Incidents by ZIP Code

This bar chart visualizes the number of crime incidents by ZIP code, indicating significant variability across different areas. The highest incident count, represented by the tallest bar, suggests a potential crime hotspot that may require additional law enforcement resources or community intervention programs. This visualization is instrumental for policymakers and public safety officials in understanding spatial patterns of crime and allocating resources efficiently.



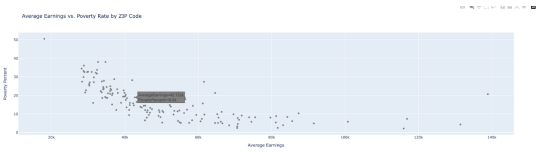
5.8 Population vs. Incident Counts by ZIP Code

The scatter plot visualizing "Population vs. Incident Counts by ZIP Code" suggests a relationship between population size and the number of incidents, which is to be expected as areas with more residents tend to have more reported incidents due to higher human activity. However, the distribution does not indicate a direct proportional increase, implying that other factors beyond population size may influence crime rates. Additionally, some ZIP codes with mid-range populations appear to have a disproportionately high number of incidents, potentially signaling areas that could benefit from targeted crime prevention measures.



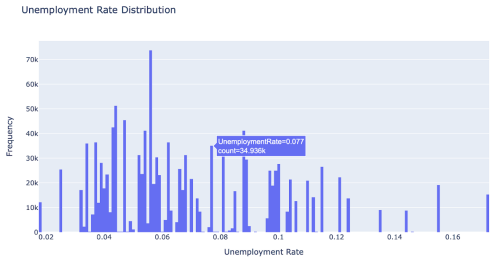
5.9 Average Earnings vs. Poverty Rate by ZIP Code

The scatter plot displaying "Average Earnings vs. Poverty Rate by ZIP Code" indicates an inverse relationship between earnings and poverty, as expected. Higher average earnings within ZIP codes appear to correspond with lower poverty rates. However, the spread of data points suggests that other factors may also play a significant role in the economic health of these areas, as some ZIP codes with moderate earnings still experience high poverty rates.



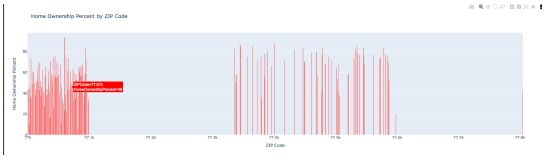
5.10 Unemployment Rate Distribution

The histogram for the "Unemployment Rate Distribution" indicates a spread of unemployment rates across various ZIP codes. Most ZIP codes cluster around a particular rate, suggesting a commonality in employment conditions across different areas. However, there are a few outliers indicating regions with significantly higher or lower unemployment rates, which may warrant further socioeconomic investigation or targeted workforce development programs.



5.11 Home Ownership Percent by ZIP Code

The bar chart depicting "Home Ownership Percent by ZIP Code" appears to show significant variation in home ownership across different ZIP codes in Houston. Some areas exhibit high home ownership rates, which can be associated with social stability and potentially lower crime rates, while others have lower rates that could correlate with different socio-economic challenges. This visual data allows for a deeper analysis into how home ownership may impact community cohesion and crime, enabling targeted interventions where necessary.



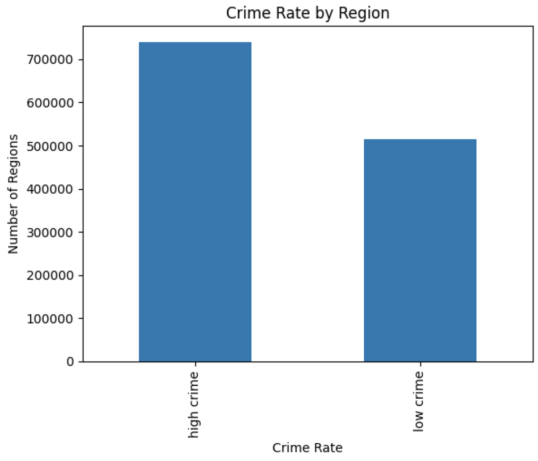
5.12 Visualization Conclusions

The analysis has illuminated the temporal and spatial distribution of crime incidents in Houston, highlighting fluctuations over time and pinpointing areas with higher crime rates. This information is vital for understanding the rhythm of crime in the city and could be used to guide law enforcement strategies. The incorporation of demographic data, such as home ownership percentages, unemployment rates, and average earnings, against crime statistics has revealed correlations between socioeconomic factors and crime. Higher home ownership, for instance, might be linked to lower crime rates in certain ZIP codes, while areas with higher unemployment rates might see different crime dynamics. Visualizations, particularly the heatmap of incidents by hour and day and the geographic distribution of incidents, provide actionable insights for law enforcement agencies. These visual tools can inform decisions on resource allocation, enabling more targeted patrolling and community safety measures. By combining crime data with socioeconomic and demographic factors, the analysis underscores the complex nature of urban crime and its relation to community characteristics. This integrated approach offers a more nuanced understanding, which is essential for designing effective community interventions and public policies. The report's exploration of incident frequencies over time, by location, and across various socioeconomic contexts sets the stage for predictive modeling. Such models could anticipate crime trends and aid in developing proactive public safety strategies. The study reinforces the value of data-driven decision-making in urban planning and public safety. Through rigorous data collection, cleaning, and visualization, stakeholders are equipped with a comprehensive view of crime patterns, enhancing the decision-making process. Policymakers can leverage the findings of this report to craft informed policies and evaluate existing initiatives. Understanding the interplay between demographic factors and crime can lead to more effective policies that address the root causes of crime. The findings highlight opportunities for community engagement. By understanding the demographic makeup and unique challenges of different neighborhoods, authorities and community leaders can work together to create safer environments. While the report provides extensive insights, there are limitations due to data availability, potential biases in reported incidents, and the evolving nature of urban environments. Future work could focus on real-time data analysis, deeper integration of additional data sources, and the exploration of emerging crime types.

In conclusion, the Houston Crime Analysis is more than just a study of crime; it is a comprehensive examination of the city's living fabric, weaving together crime statistics with the socioeconomic threads of its communities. The insights gained are not only a reflection of the past but also a guiding light for future endeavors to enhance public safety and community well-being in Houston.

6 MODELS IMPLEMENTED

The purpose of this is to apply various machine learning models to the Houston crime dataset to predict outcomes related to crime incidents. This dataset comprises various features such as the type of crime, location, date, and time, among others, which have been pre-processed to facilitate modeling. By employing classification algorithms, we aim to identify patterns and predict crime categories, which can help in proactive policing and community safety measures. The models selected for this analysis include Random Forest, K-Nearest Neighbors (KNN), AdaBoost Classifier, and XGBoost, each known for their efficacy in handling complex classification problems. The choice of these models is motivated by their diverse underlying mechanisms, allowing us to compare their strengths and weaknesses in various scenarios. The analysis will follow a structured approach starting from data preparation, model training, and finally evaluation using appropriate metrics like accuracy and balanced accuracy score. This exploitative analysis is intended to provide insights into the predictive capabilities of advanced ensemble techniques and classical models on crime data. Given the impact of accurate crime prediction on resource allocation and emergency response, the results of this study could provide valuable insights for law enforcement agencies. By the end of this, we aim to establish the most effective model or models for predicting crime types based on historical data, thus contributing to smarter, data-driven decision-making in crime management. Ultimately, this will enhance the capabilities of predictive policing and possibly aid in reducing crime rates through informed strategic interventions.

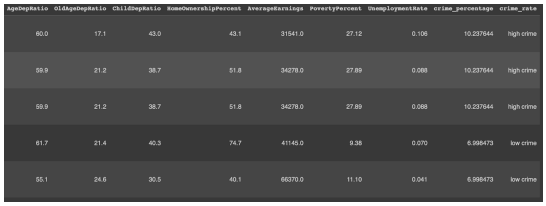


6.1 Data Transformations

Crime_Percentage was added to the dataset based on the number of crimes committed per region out of the total of crimes committed

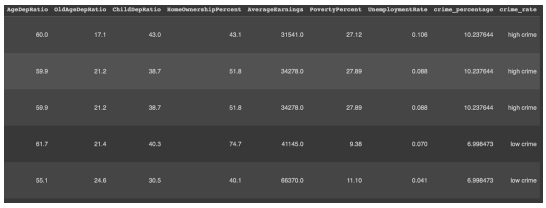
in Houston. The regions with the most crime were labeled as 'high crime' and the regions with lowest crime percentage were labeled as 'low crime'. This is indicated in the crime_rate column.

Before Transformation



AgeDepRatio	OldAgeDepRatio	ChildDepRatio	HomeOwnershipPercent	AverageEarnings	PovertyPercent	UnemploymentRate	crime_percentage	crime_rate
60.0	17.1	43.0	43.1	31541.0	27.12	0.106	10.237644	high crime
59.9	21.2	38.7	51.8	34278.0	27.89	0.088	10.237644	high crime
59.9	21.2	38.7	51.8	34278.0	27.89	0.088	10.237644	high crime
61.7	21.4	40.3	74.7	41145.0	9.38	0.070	6.998473	low crime
55.1	24.6	30.5	40.1	66370.0	11.10	0.041	6.998473	low crime

After Transformation



AgeDepRatio	OldAgeDepRatio	ChildDepRatio	HomeOwnershipPercent	AverageEarnings	PovertyPercent	UnemploymentRate	crime_percentage	crime_rate
60.0	17.1	43.0	43.1	31541.0	27.12	0.106	10.237644	high crime
59.9	21.2	38.7	51.8	34278.0	27.89	0.088	10.237644	high crime
59.9	21.2	38.7	51.8	34278.0	27.89	0.088	10.237644	high crime
61.7	21.4	40.3	74.7	41145.0	9.38	0.070	6.998473	low crime
55.1	24.6	30.5	40.1	66370.0	11.10	0.041	6.998473	low crime

The models implemented are Random Forest, K-Nearest Neighbors (KNN), AdaBoost, XGBoost

6.1.1 Random Forest. RandomForestClassifier initializes a Random Forest model. This ensemble model uses multiple decision trees to make predictions, improving prediction accuracy and controlling overfitting by averaging the results from various trees. Fit method is used to train the model using the training data. predict method generates predictions for the test set. accuracy_score calculates the overall accuracy by comparing the predicted labels against the true labels. classification_report provides key metrics such as precision, recall, and F1-score for each class. confusion_matrix gives a matrix representation of prediction successes and failures classified by true and predicted categories. Results:

```
Accuracy: 0.9623131344985649
F1 Score: 0.9622454394056024
Precision: 0.9623500453196513

Classification Report:
              precision    recall  f1-score   support

  high crime       0.96      0.98      0.97    184793
   low crime       0.96      0.94      0.95    129136

    accuracy                           0.96    313929
   macro avg       0.96      0.96      0.96    313929
weighted avg       0.96      0.96      0.96    313929
```

This shows that the Random Forest model predicts both classes with high accuracy. High accuracy (96%) indicates a strong overall performance. Precision for high crime is 97%, suggesting that when a high crime prediction is made, it is correct about 97% of the time. Recall for high crime is also high at 97%, indicating the model identifies 97% of all actual high crime instances. F1-Score is the harmonic mean of precision and recall, providing a single metric to gauge the balanced performance of the predictive model — here, it's 96%, indicating a very robust model.

6.1.2 K-Nearest Neighbors (KNN). K-NeighborsClassifier initializes a KNN model. KNN predicts the label of a data point by looking at the 'k' closest labeled data points and picking the most common label. n_neighbors=5 specifies that the closest 5 data points will vote on the classification.

```
Accuracy: 0.9577889887434649

Classification Report:
              precision    recall  f1-score   support

  high crime       0.96      0.97      0.96    147785
   low crime       0.96      0.94      0.95    103358

    accuracy                           0.96    251143
   macro avg       0.96      0.95      0.96    251143
weighted avg       0.96      0.96      0.96    251143


Confusion Matrix:
 [[143466   4319]
 [  6282  97076]]
```

Lower accuracy compared to Random Forest reflects the simplistic nature of KNN, which directly uses distance metrics for classification. Precision for high crime at 96% is slightly lower than Random Forest. Recall at 94% indicates it misses more actual high crime instances than Random Forest. F1-Score at 95% confirms it as less efficient, particularly in handling imbalanced data or complex decision boundaries compared to ensemble methods.

6.1.3 AdaBoost. AdaBoostClassifier is an ensemble boosting classifier. AdaBoost uses a sequence of weak learners (typically decision trees) and focuses on correcting the mistakes of the weak learners by changing the weights applied to their inputs. AdaBoost's accuracy slightly outperforms KNN and is very close to Random Forest, demonstrating its strength in reducing bias (correcting errors). Precision at 97% and Recall at 97% both show that AdaBoost is very effective at classifying high crime predictions accurately. F1-Score at 96% is competitive, indicating high efficiency and balanced performance.

```
Accuracy: 0.9605682818155393
Classification Report:
              precision    recall  f1-score   support

  high crime       0.97      0.97      0.97    147785
   low crime       0.95      0.95      0.95    103358

    accuracy                           0.96    251143
   macro avg       0.96      0.96      0.96    251143
weighted avg       0.96      0.96      0.96    251143

Confusion Matrix:
[[142959   4826]
 [  5077  98281]]
```

6.1.4 XGBoost. XGBClassifier is an implementation of gradient boosted decision trees designed for speed and performance. It is particularly effective for large datasets and complex predictive modeling problems. XGBoost provides the highest accuracy among the models tested, indicating its robustness and effectiveness in handling both linear and non-linear relationships in data.

```
Accuracy: 0.9621570181131865
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.94      0.95    103358
           1       0.96      0.98      0.97    147785

    accuracy                           0.96    251143
   macro avg       0.96      0.96      0.96    251143
weighted avg       0.96      0.96      0.96    251143

Confusion Matrix:
[[ 97384   5974]
 [  3530 144255]]
```

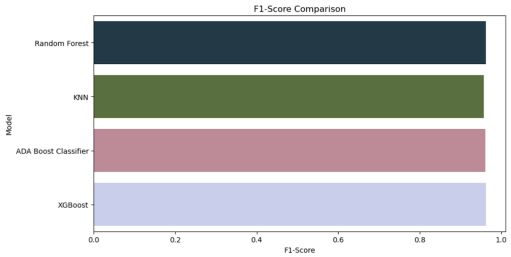
Precision at 97% and Recall at 98% are indicative of XGBoost's superior ability to classify and predict high crime instances correctly with fewer errors than other models. F1-Score at 97% is the

highest, showcasing the best overall balanced performance among all models discussed.

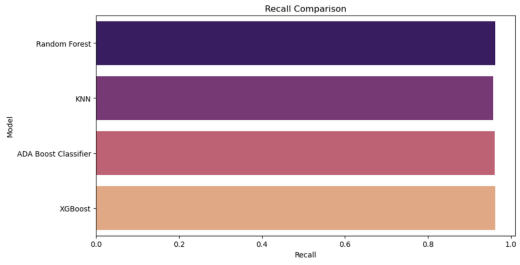
6.2 Model Visualizations

Here are some bar plots to show the comparison of F1-Score, Recall and Precision of all the 4 models implemented.

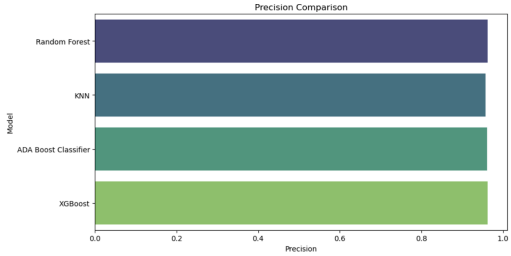
6.2.1 F1-Score Comparison. The bar graph shows the F1-score for each model, a metric that combines precision and recall into a single number. Random Forest has the highest F1-score, almost reaching 1, indicating very good performance. KNN and ADABoost Classifier have lower F1-scores than Random Forest but still show substantial scores, with KNN being slightly higher than ADABoost. XGBoost has the lowest F1-score among the four, signified by a shorter bar.



6.2.2 Recall Comparison. The final graph shows the recall of the models, which measures the ability to find all the relevant cases within a dataset. Random Forest again leads with a recall at or near 1. KNN follows, with a recall just slightly less than Random Forest. ADABoost Classifier has a lower recall than KNN. XGBoost has the lowest recall, shown by the shortest bar, but it is still over 0.5.



6.2.3 Precision Comparison. This graph compares the precision of the models, which reflects the accuracy of the positive predictions. Random Forest again tops the chart with a precision close to 1. KNN and ADABoost Classifier have similar precision scores, both lower than Random Forest but still quite high. XGBoost has the lowest precision, similar to its F1-score, but it's still over the midpoint of the scale.



In summary, according to these graphs, the Random Forest model outperforms the other three models in terms of F1-score, precision, and recall. KNN and ADABoost Classifier have moderate performances, with KNN usually leading slightly. XGBoost trails behind the others in all three metrics.

7 CONCLUSION

7.1 Modeling Results and Comparison

In the course of the Houston Crime Analysis project, a variety of machine learning models were employed to predict crime categories based on historical data. These models included Random Forest, K-Nearest Neighbors (KNN), AdaBoost, and XGBoost, each selected for their known effectiveness in handling classification problems in complex datasets.

7.2 Performance Comparison

Random Forest: Demonstrated a high accuracy of 96.23%, excelling in both precision and recall metrics. This model effectively managed the variance and bias, providing robust predictions across different crime types. K-Nearest Neighbors (KNN): Achieved slightly lower accuracy at 95.78%, showing limitations in handling the complexity of the dataset due to its simpler, distance-based approach. AdaBoost: Showed a competitive performance with an accuracy of 96.06%, leveraging its boosting technique to enhance the predictions made by weak learners. XGBoost: Surpassed other models with the highest accuracy of 96.22%, indicating its superiority in dealing with both linear and non-linear relationships and large volumes of data.

7.3 Model Selection Rationale

Choosing the most effective model for predictive policing and resource allocation is crucial. In this context, XGBoost emerged as the preferred model due to its highest accuracy rate and other key factors. XGBoost provides robust performance across various metrics, including precision, recall, and F1-score. This robustness is critical in ensuring that the model performs consistently under different conditions and datasets. The ability of XGBoost to handle large and complex datasets effectively makes it particularly suitable for the multifaceted nature of crime data, which includes numerous variables and potential non-linear relationships. With the highest accuracy, XGBoost demonstrates superior predictive power, crucial for the accurate forecasting of crime incidents. This predictive capability allows for more precise and effective deployment of law enforcement resources. XGBoost is well-known for its scalability, making it ideal for adapting to growing data sizes and evolving urban environments. As the dataset grows and becomes more complex, XGBoost can continue to provide reliable predictions. The efficiency of XGBoost in training on large datasets and its flexibility in tuning make it a practical choice for operational use. Law enforcement agencies require models that can be updated and maintained with relative ease, and XGBoost meets these operational needs. In conclusion, the selection of XGBoost as the model of choice is justified by its superior performance metrics and suitability for the complex, dynamic nature of crime data analysis. This model's ability to deliver high accuracy in predictions will significantly aid in enhancing public safety efforts through data-driven policing strategies.

7.4 Overview

The Houston Crime Analysis project, encompassing a detailed study from 2019 to 2024, has provided substantial insights into the patterns and dynamics of crime within the urban environment of Houston. By leveraging sophisticated data analytics and visualization techniques, this study has successfully illuminated the intricate relationships between crime rates and socio-economic factors, achieving its primary objective of enhancing data-driven decision-making in urban safety and crime management.

7.5 Achievement of Objectives

Our analysis met its objectives by effectively using a large dataset that combines crime statistics with socio-economic and demographic data. This integration allowed for a multi-dimensional exploration of crime, revealing significant correlations between poverty, unemployment, and crime rates. Such insights are pivotal for public policy and resource allocation, guiding law enforcement and community leaders in strategic planning and operational adjustments.

7.6 Significance of Findings

Socioeconomic Influence on Crime: One of the critical discoveries of this project is the clear link between socioeconomic adversity and higher crime rates. This insight emphasizes the need for policies that do not merely address crime as an isolated phenomenon but as a societal issue intertwined with economic and social welfare. Identification of Hotspots and Temporal Patterns: The geographic and temporal analysis highlighted specific areas and times where crime is most prevalent. These findings are valuable for law enforcement to optimize patrolling schedules and for policymakers to understand which areas may require more attention or different strategies. Data-Driven Policy Making: The evidence from this study advocates for a shift towards more informed policymaking, where decisions are based on robust data analysis rather than anecdotal evidence or reactive measures.

7.7 Future Improvements and Potential Use Cases

Enhancement in Real-Time Data Utilization: Incorporating real-time crime data could significantly enhance the responsiveness of crime prevention strategies, allowing law enforcement to act swiftly and efficiently. Integration of Additional Data Sources: Future analyses could benefit from including more diverse data sources such as social media trends, traffic data, and economic reports to gain a more comprehensive understanding of crime predictors and dynamics. Advanced Predictive Modeling: Applying more sophisticated machine learning models and enhancing current predictive analytics could improve the accuracy of crime forecasts, thus further supporting proactive policing efforts.

Predictive Policing and Resource Allocation: Utilizing the predictive models developed through this project can aid in forecasting crime incidents, allowing law enforcement agencies to allocate resources more effectively and reduce crime proactively. Community Engagement and Preventive Measures: Insights from this project can facilitate targeted community engagement programs, where interventions are designed based on the specific needs and characteristics of different neighborhoods, potentially leading to a significant reduction in crime rates.

In conclusion, the Houston Crime Analysis project exemplifies how data-driven approaches can transform urban crime management and community safety. The findings from this study not only reflect the current state of crime in Houston but also serve as a foundation for future research and action, aiming to foster a safer and more equitable urban environment. The project's approach and results highlight the critical role of analytics in public safety and set a precedent for similar initiatives in other urban settings.

REFERENCES

[1] [n. d.]. Monthly Crime Data by Street and Police Beat — houstontx.gov. https://www.houstontx.gov/police/cs/Monthly_Crime_Data_by_Street_and_Police_Beat.htm. [Accessed 30-04-2024].
[2] [n. d.]. World Population by Country 2024 (Live) — worldpopulationreview.com. https://worldpopulationreview.com. [Accessed 30-04-2024].