# Houston Crime Analysis

**Authors**

***Preethu Manjunath, Lohith Ramesh, Raquel Yupanqui***

## 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 socio-economic correlations. By integrating crime data with meticulously web-scraped demographic information, our analysis offers a holistic view of the interplay between crime rates and socio-economic factors across different neighborhoods.

The analysis commenced with the aggregation and preprocessing 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 socio-economic 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 socio-economic 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 socio-economic 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.

# Data Collection/Preparation Information

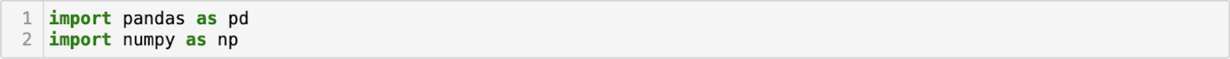
The Houston Crime Analysis project embarked on a multifaceted journey to dissect urban crime through the lens of data-driven insights. Commencing in 2019 and extending through 2024, this endeavor aimed to unearth patterns, trends, and correlations within the vast landscape of crime data, enriched by socio-economic 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.

## Data Collection

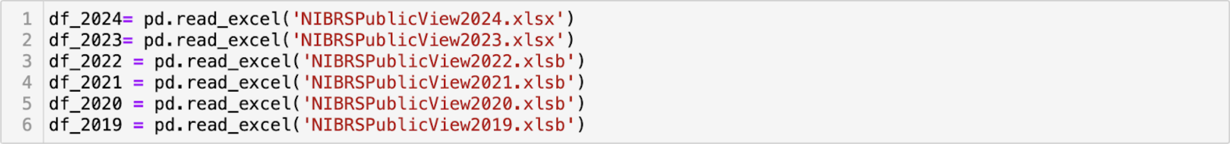
- Source of Data: The primary dataset, encompassing detailed monthly crime reports by street and police beat, was sourced from the City of Houston's official website, www.houstontx.gov. This rich repository of crime incidents provided a comprehensive canvas for the analysis.
- Data Acquisition: The crime 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.

## Initial Loading and Analysis

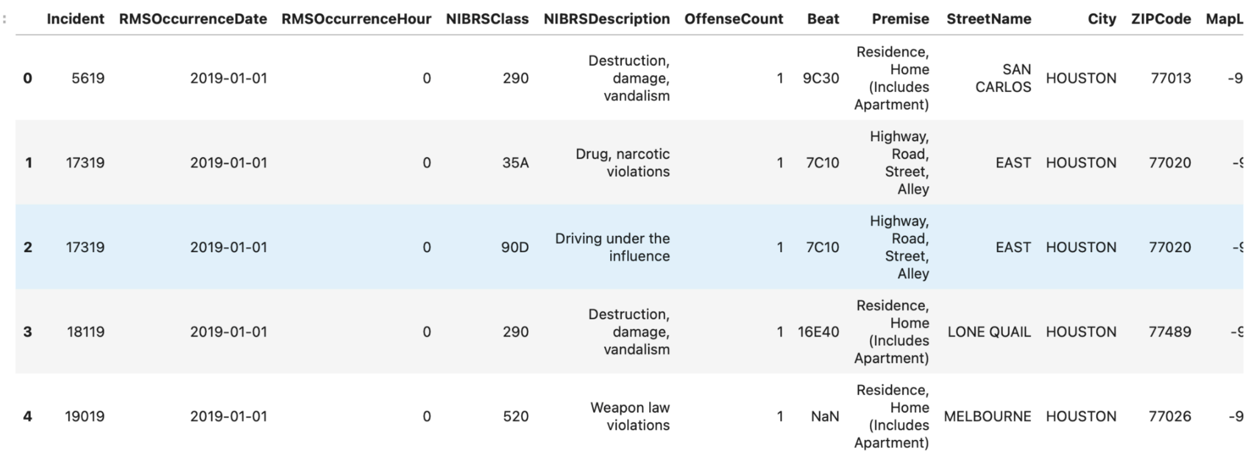
- Utilizing Python libraries Pandas for data manipulation and Matplotlib/Plotly for visualization, the yearly data was loaded into separate Pandas DataFrames. This step was crucial for acclimating to the dataset's nuances, setting the stage for in-depth exploration and analysis.



```
import pandas as pd
import numpy as np
```



```
df_2024= pd.read_excel('NIBRSPublicView2024.xlsx')
df_2023= pd.read_excel('NIBRSPublicView2023.xlsx')
df_2022 = pd.read_excel('NIBRSPublicView2022.xlsb')
df_2021 = pd.read_excel('NIBRSPublicView2021.xlsb')
df_2020 = pd.read_excel('NIBRSPublicView2020.xlsb')
df_2019 = pd.read_excel('NIBRSPublicView2019.xlsb')
```



| : | 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 Apartment) | MELBOURNE | HOUSTON | 77026 | -9 |

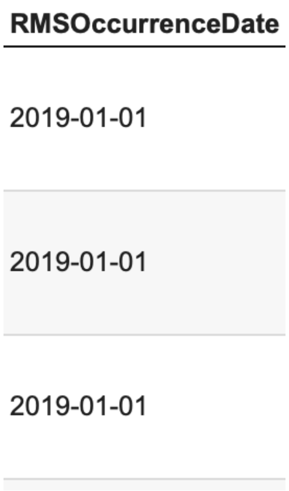
**Preprocessing**
The preprocessing phase was a pivotal endeavor to refine the dataset, encompassing:

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.

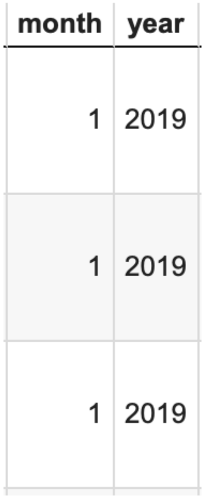
2. Data Type Conversion: Data types were meticulously adjusted for analysis efficiency and accuracy. Zip codes were transitioned from floats to integers, and date strings to Datetime objects, facilitating accurate temporal analyses.

3. Standardizing Values and Removing Duplicates: Efforts ensured data consistency, particularly for textual data, and duplicate checks confirmed the dataset's quality.

4. Feature Engineering: The dataset was enhanced by extracting meaningful features such as month and year from crime incident dates, enabling detailed pattern analysis.



| RMSOccurrenceDate |
|---|
| 2019-01-01 |
| 2019-01-01 |
| 2019-01-01 |



| month | year |
|---|---|
| 1 | 2019 |
| 1 | 2019 |
| 1 | 2019 |

5. Data Reduction: Irrelevant columns, like street type and number, were pruned to focus the analysis, streamlining the dataset for clarity and purpose.

6. 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 preprocessing, feature engineering, and subsequent modeling stages.

7. 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.

8. 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.

9. 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.

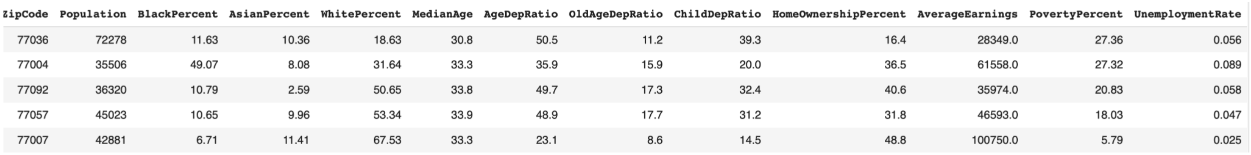
10. Normalization: Normalization is the process of scaling individual data points to ensure consistency across different ranges for comparative and analytical purposes. It is especially important when dealing with features that vary in scale, units, or range, as it can impact the performance of many 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.

11. 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.

### Integration with Demographic Data

- The project expanded its analytical horizon by weaving in web-scraped demographic data from [worldpopulationreview.com](https://worldpopulationreview.com), adding layers of socio-economic 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.



| ZipCode | Population | BlackPercent | AsianPercent | WhitePercent | MedianAge | AgeDepRatio | OldAgeDepRatio | ChildDepRatio | HomeOwnershipPercent | AverageEarnings | PovertyPercent | UnemploymentRate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 77036 | 72278 | 11.63 | 10.36 | 18.63 | 30.8 | 50.5 | 11.2 | 39.3 | 16.4 | 28349.0 | 27.36 | 0.056 |
| 77004 | 35506 | 49.07 | 8.08 | 31.64 | 33.3 | 35.9 | 15.9 | 20.0 | 36.5 | 61558.0 | 27.32 | 0.089 |
| 77092 | 36320 | 10.79 | 2.59 | 50.65 | 33.8 | 49.7 | 17.3 | 32.4 | 40.6 | 35974.0 | 20.83 | 0.058 |
| 77057 | 45023 | 10.65 | 9.96 | 53.34 | 33.9 | 48.9 | 17.7 | 31.2 | 31.8 | 46593.0 | 18.03 | 0.047 |
| 77007 | 42881 | 6.71 | 11.41 | 67.53 | 33.3 | 23.1 | 8.6 | 14.5 | 48.8 | 100750.0 | 5.79 | 0.025 |

### 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 socio-economic insights. This phase required careful data alignment, particularly in matching geographical identifiers, to enable holistic analyses of crime within Houston's broader socio-economic and demographic landscape.
- Post-merge, the dataset underwent additional feature engineering and verification steps to ensure data integrity and analytical readiness.

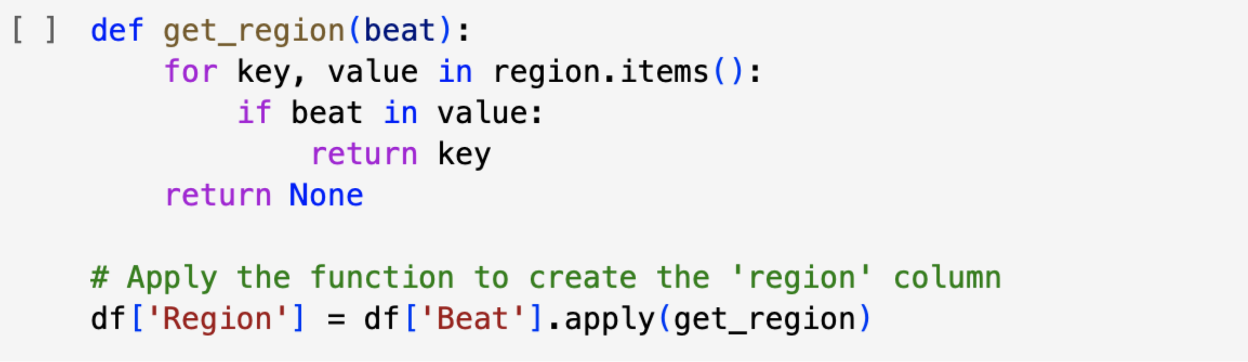
## Cleaning

### 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.

### Handling Missing Values

- Strategic Removals and Imputations: 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.



```
[ ]  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)
```

**Data Type Conversion**

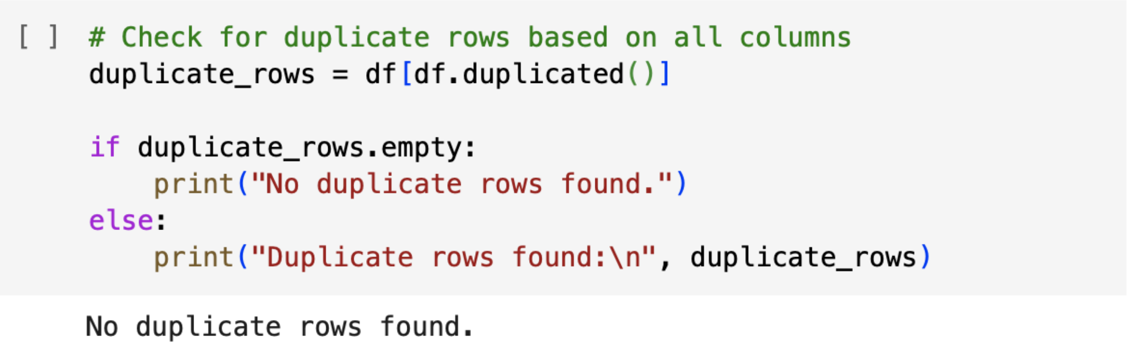
- Enhancing Analytical Readiness: Critical to the analysis was ensuring that each column's data type accurately reflected its content. Zip codes, initially recorded as float values, were converted to integer format to correct their representation and facilitate accurate geographical analysis. Date columns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

**Standardizing Values**

- Uniformity in Textual Data: 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.

**Removing Duplicates**

- Ensuring Data Uniqueness: 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.

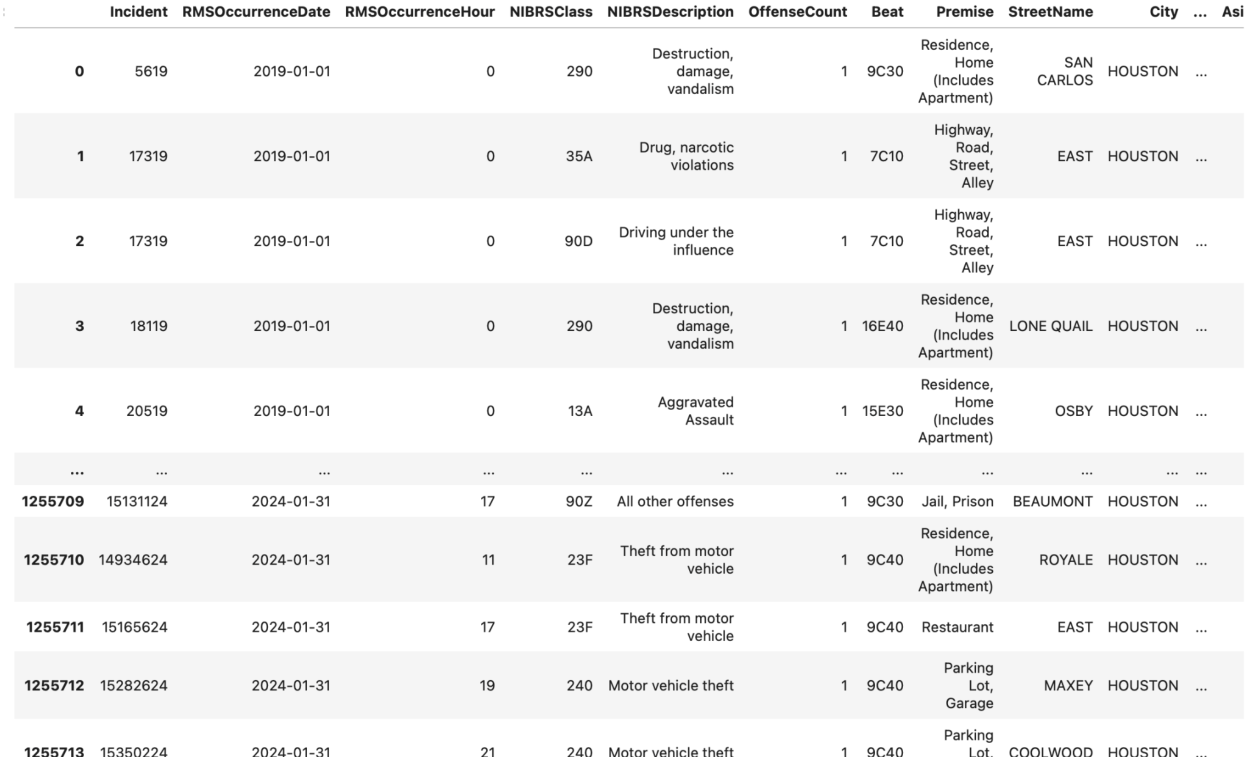


```
[ ]  # Check for duplicate rows based on all columns
     duplicate_rows = df[df.duplicated()]

     if duplicate_rows.empty:
         print("No duplicate rows found.")
     else:
         print("Duplicate rows found:\n", duplicate_rows)

     No duplicate rows found.
```

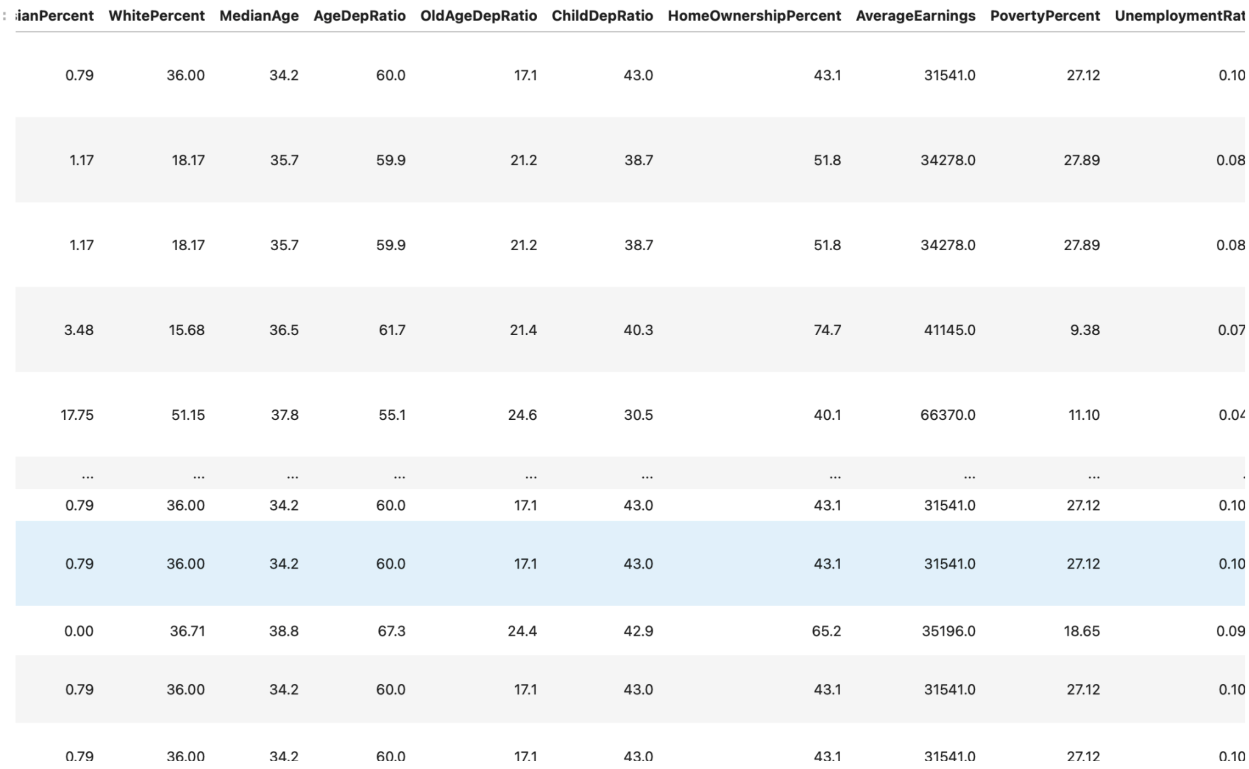
**Feature Engineering**

- Extracting Analytical Value: 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.

### The Final Dataset Snippet



| | 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** | 15131124 | 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 | ... | |



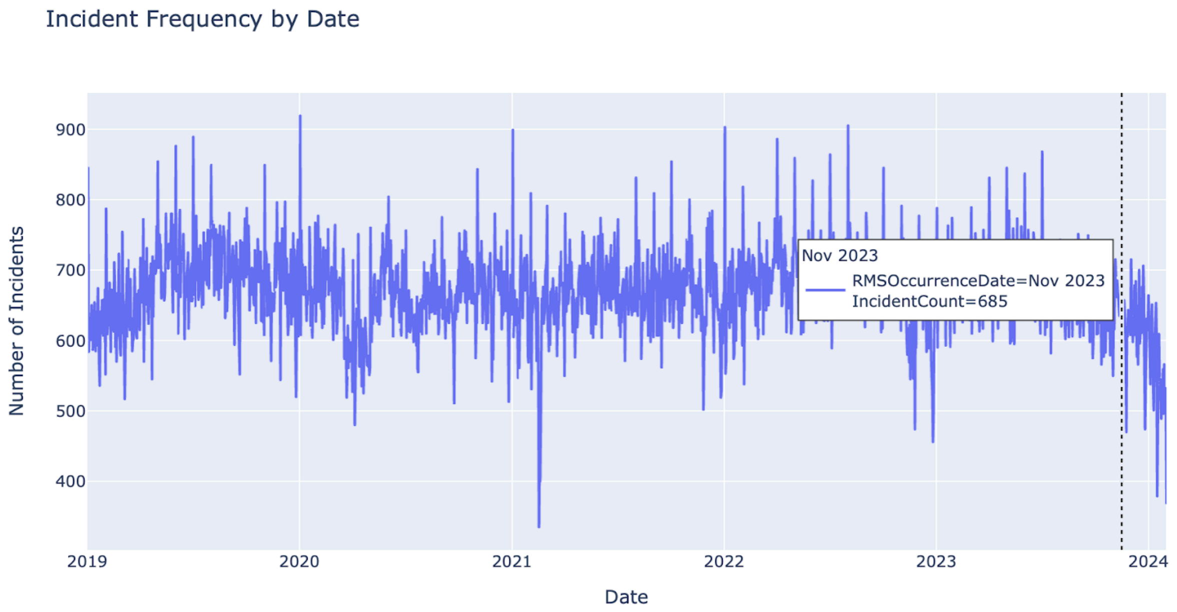
| ianPercent | WhitePercent | MedianAge | AgeDepRatio | OldAgeDepRatio | ChildDepRatio | HomeOwnershipPercent | AverageEarnings | PovertyPercent | UnemploymentRat |
|---|---|---|---|---|---|---|---|---|---|
| 0.79 | 36.00 | 34.2 | 60.0 | 17.1 | 43.0 | 43.1 | 31541.0 | 27.12 | 0.10 |
| 1.17 | 18.17 | 35.7 | 59.9 | 21.2 | 38.7 | 51.8 | 34278.0 | 27.89 | 0.08 |
| 1.17 | 18.17 | 35.7 | 59.9 | 21.2 | 38.7 | 51.8 | 34278.0 | 27.89 | 0.08 |
| 3.48 | 15.68 | 36.5 | 61.7 | 21.4 | 40.3 | 74.7 | 41145.0 | 9.38 | 0.07 |
| 17.75 | 51.15 | 37.8 | 55.1 | 24.6 | 30.5 | 40.1 | 66370.0 | 11.10 | 0.04 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | . |
| 0.79 | 36.00 | 34.2 | 60.0 | 17.1 | 43.0 | 43.1 | 31541.0 | 27.12 | 0.10 |
| 0.79 | 36.00 | 34.2 | 60.0 | 17.1 | 43.0 | 43.1 | 31541.0 | 27.12 | 0.10 |
| 0.00 | 36.71 | 38.8 | 67.3 | 24.4 | 42.9 | 65.2 | 35196.0 | 18.65 | 0.09 |
| 0.79 | 36.00 | 34.2 | 60.0 | 17.1 | 43.0 | 43.1 | 31541.0 | 27.12 | 0.10 |
| 0.79 | 36.00 | 34.2 | 60.0 | 17.1 | 43.0 | 43.1 | 31541.0 | 27.12 | 0.10 |

# 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.

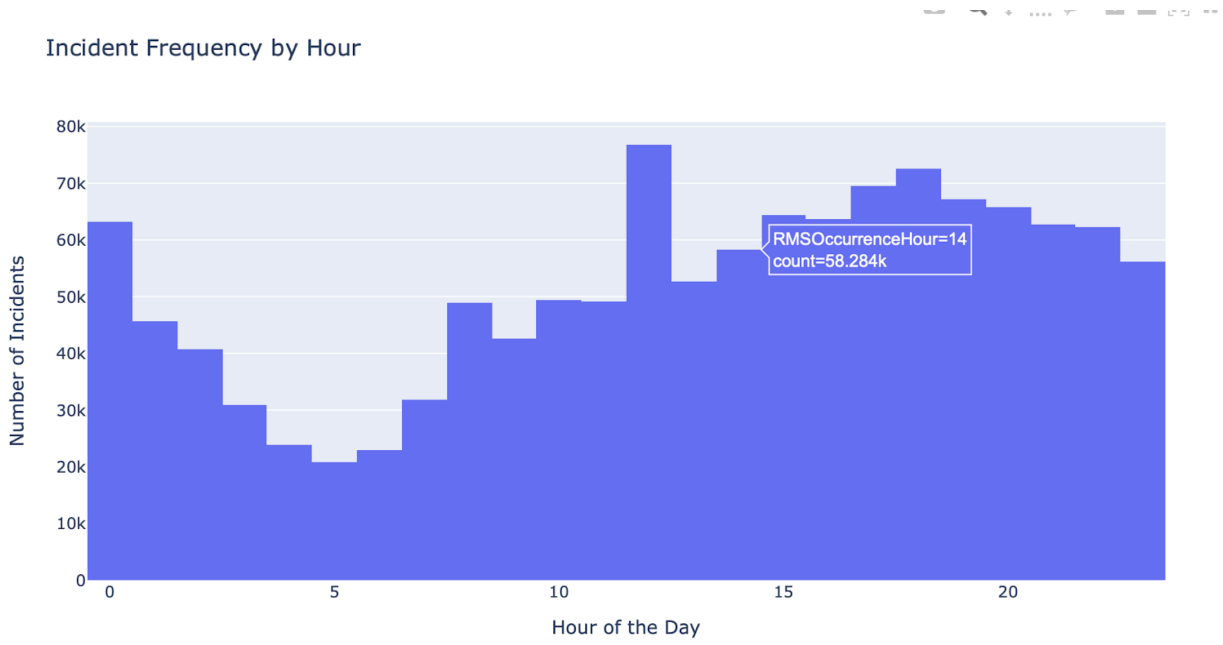
**1. Daily Crime Incidents Over Time**

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.



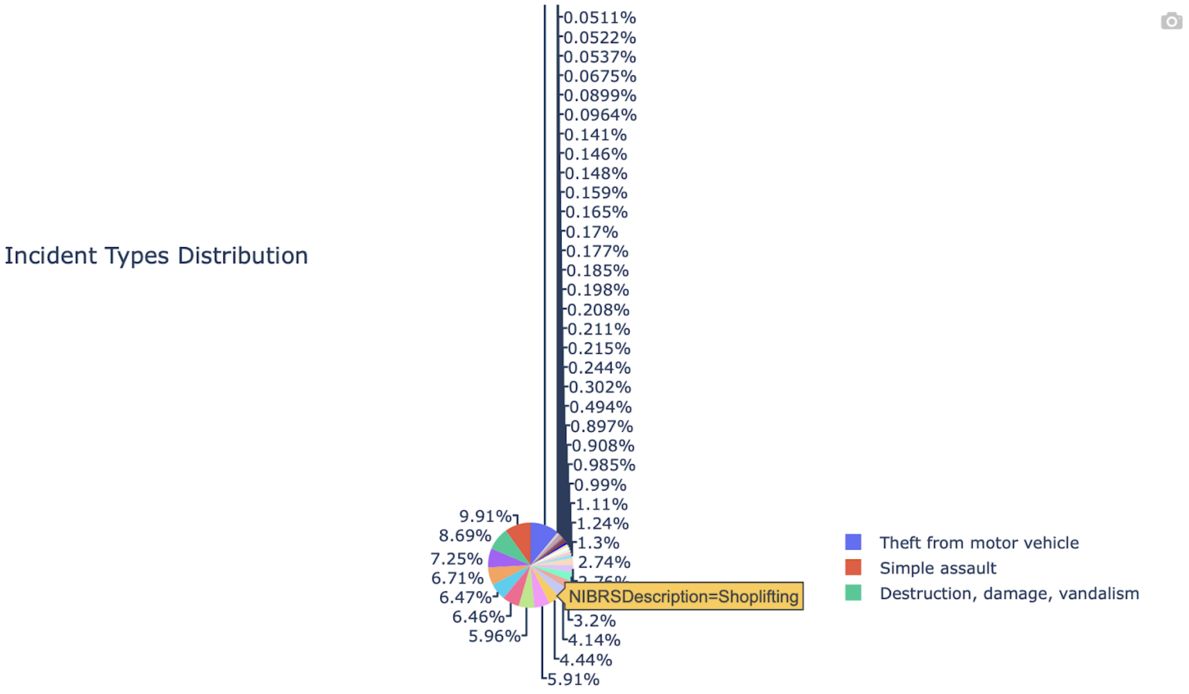
## 2. Geographic Distribution of Crime Incidents

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.



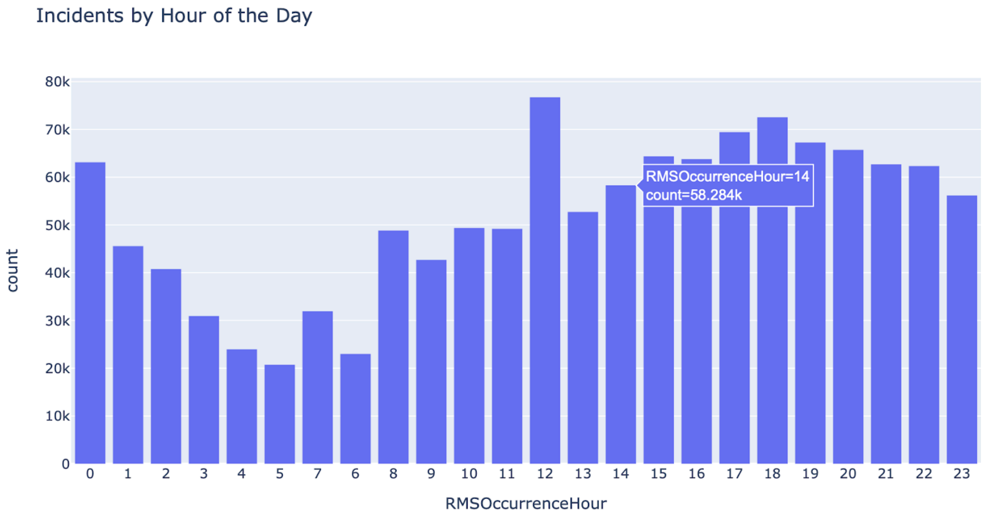
### 3. Distribution of Crime Types

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.

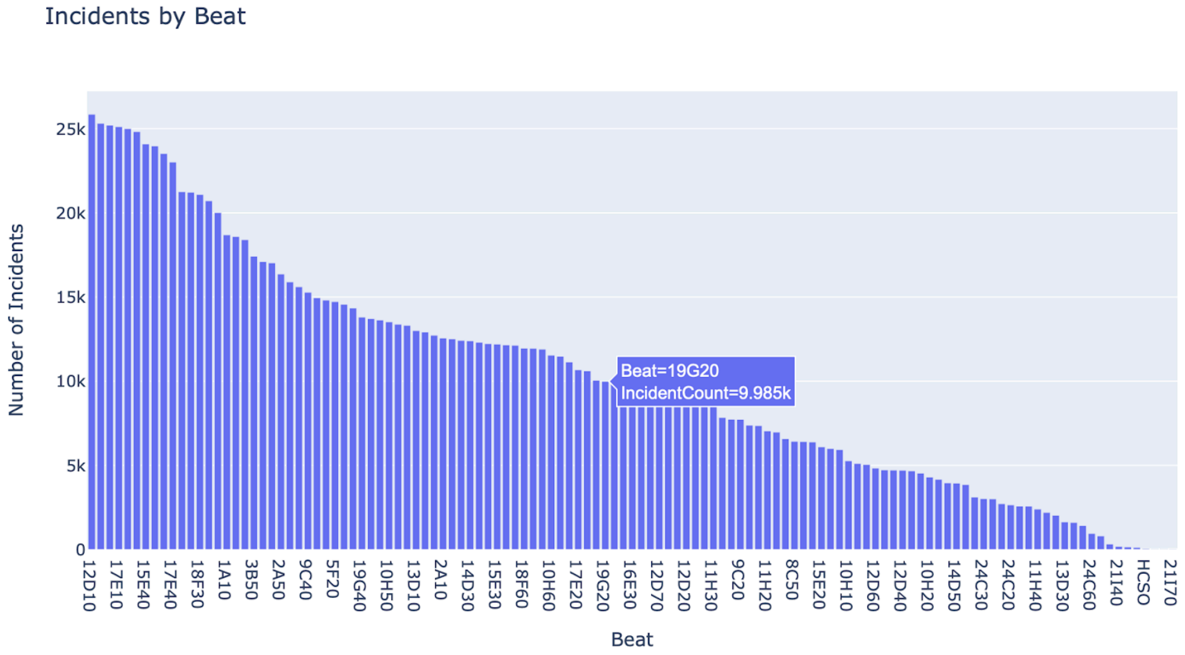


### 4. Crime Incidents by Hour of Day

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.

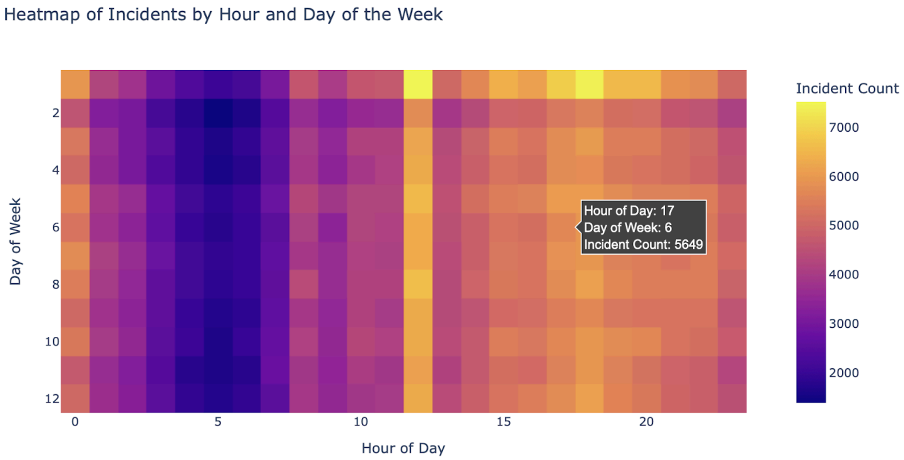


## 5. Incidents by Beat



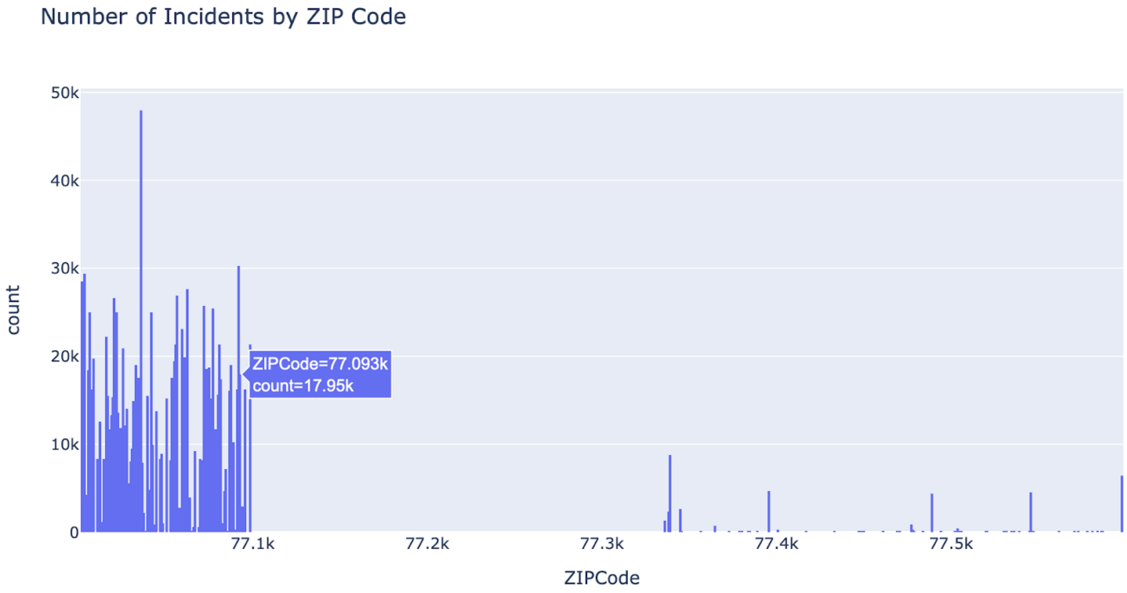
## 6. Heatmap of Incidents by Hour and Day of the Week

A heatmap visualizes crime incidents by day of the week and time of day, providing a dense, color-coded representation of when crimes are most likely to occur. This can reveal patterns, such as higher crime rates on weekends or during night hours.

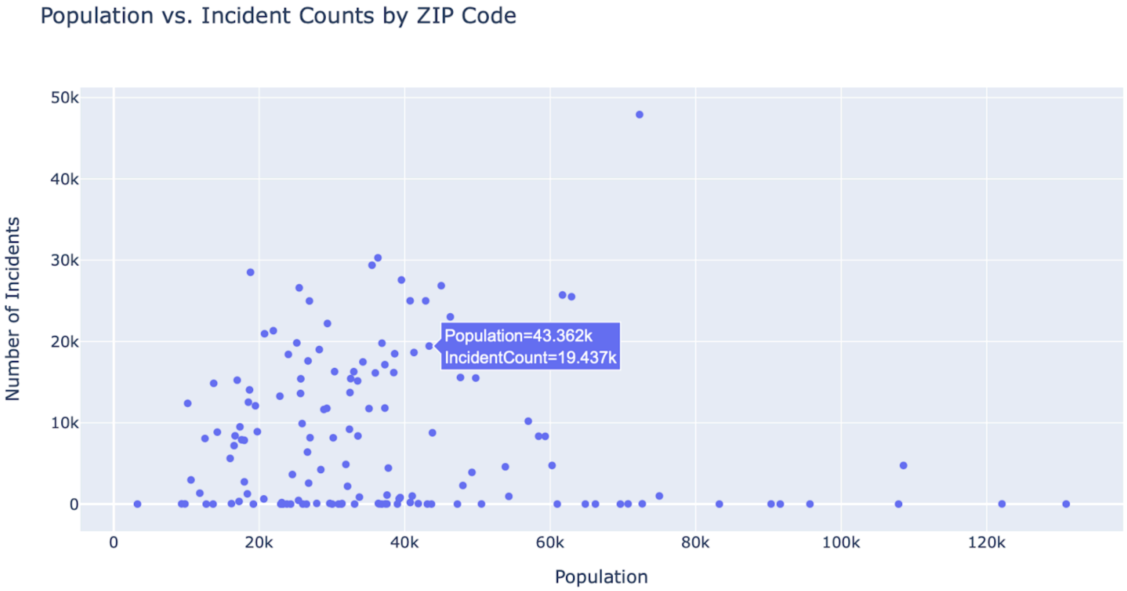


### 7. Number of Incidents by ZIP Code

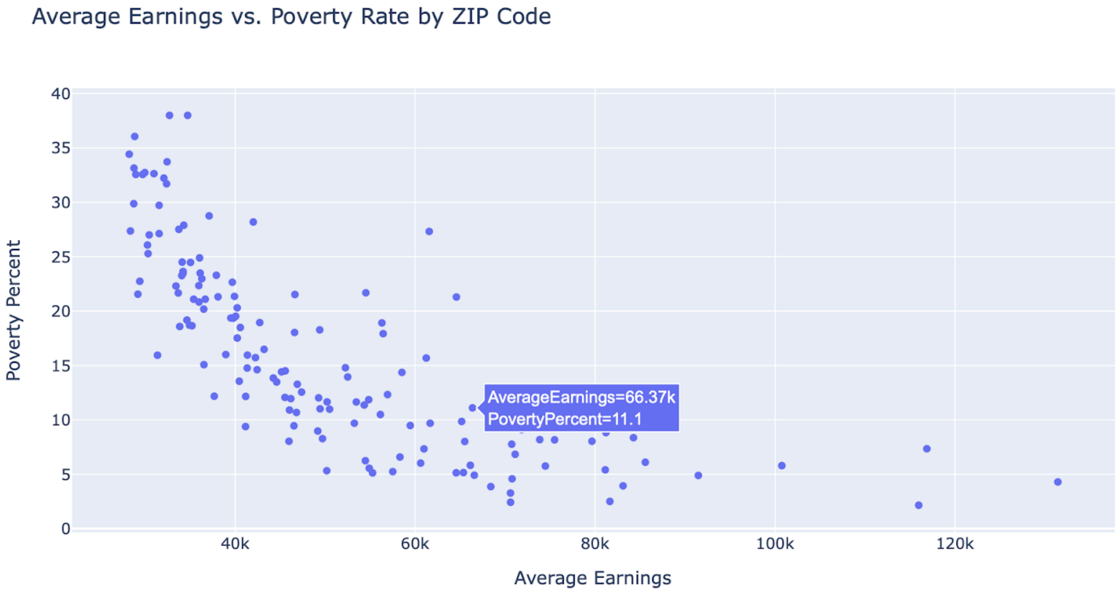
A bar chart or histogram displays the number of crime incidents occurring in different ZIP codes, identifying areas with higher crime rates. This visualization can inform targeted interventions in high-crime areas.



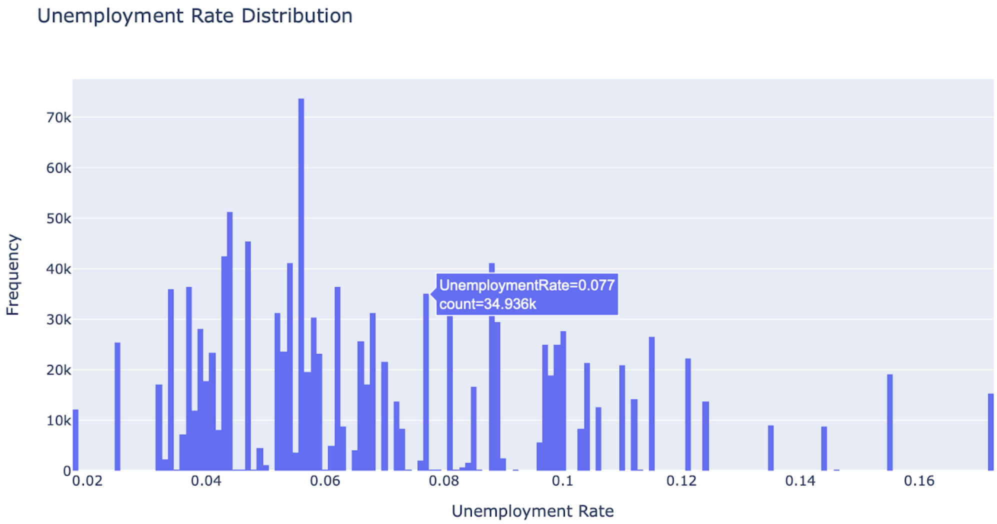
### 8. Population vs. Incident Counts by ZIP Code



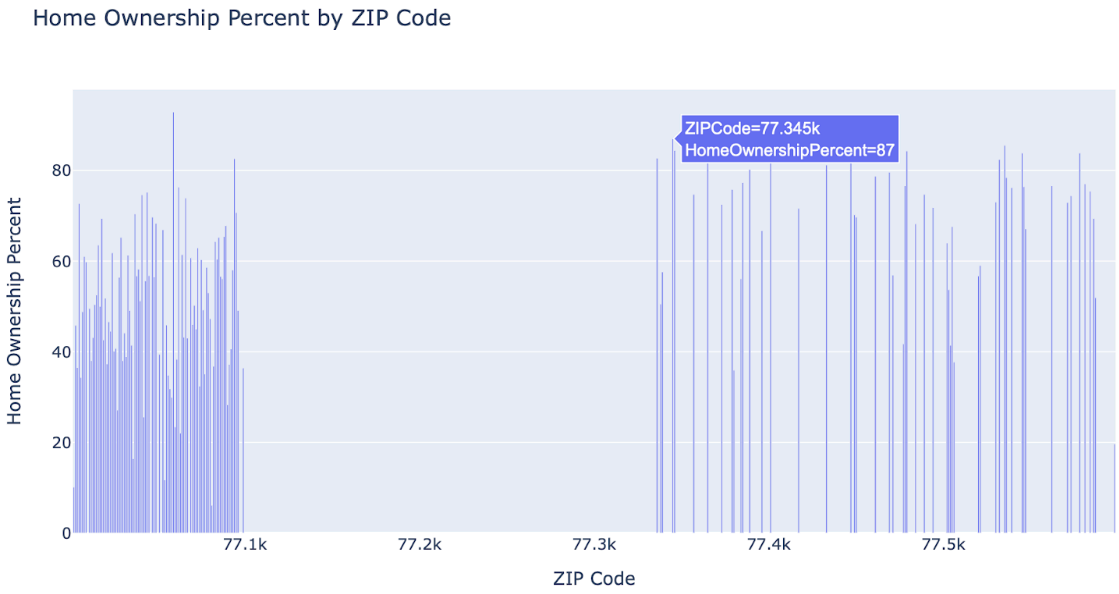
## 9. Average Earnings vs. Poverty Rate by ZIP Code



## 10. Unemployment Rate Distribution



## 11. Home Ownership Percent by ZIP Code



## Conclusion

After a thorough examination of the Houston Crime Data from 2019 to 2024, supplemented with demographic insights, the following overarching conclusions can be drawn:

1. Crime Patterns and Trends: 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.

2. Socio-Economic Factors: The incorporation of demographic data, such as home ownership percentages, unemployment rates, and average earnings, against crime statistics has revealed correlations between socio-economic 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.

3. Resource Allocation: 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.

4. Community Impact and Interventions: By combining crime data with socio-economic 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.

5. Predictive Insights: The report's exploration of incident frequencies over time, by location, and across various socio-economic contexts sets the stage for predictive modeling. Such models could anticipate crime trends and aid in developing proactive public safety strategies.

6. Data-Driven Decision Making: 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.

7. Policy Formulation and Evaluation: 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.

8. Community Engagement: 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.

9. Limitations and Future Work: 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 socio-economic 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.