Time Series Cluster and Descriptive Analysis of Crimes in Boston

***Abstract*— This research paper extensively explores crime patterns in Boston from 2015-2018 implementing a data-driven approach consisting of descriptive analysis, time series and cluster analysis. This study focuses on identifying crime hotspots, peak times, and seasonal trends across various districts in Boston. The analytical approach involves visualizations such as heat maps, time series graphs, and clustering. Findings reveal unique crime patterns across districts with some offenses peaking at different times of the day. This approach also uncovers trends and seasonal fluctuations across crime rates. Final results can help authorities make informed decisions on different law enforcement strategies with a common goal of ensuring public safety for all citizens of the city of Boston. The research paper concludes with ideas for further research, emphasizing the importance of hidden crime causes through qualitative research and diverse data sources.**

I. Introduction

In an era where data-driven results are important in informing public policy, analyzing a crime dataset is essential for understanding and reducing future criminal activities. The purpose of this project and research paper is to explore the different patterns and trends that lead to criminality across various districts in Boston using descriptive statistics, clustering, and time series analysis. The motivation behind this extensive data analysis and research paper is to take an urgent course of action to implement the correct law enforcement strategies to create a safer world for everyone to live in.[1] A detailed analysis can identify crime hotspots allowing the police to allocate more officers and resources to effectively decrease the prevention of Boston crimes. Specialized training for state defenders can also be presented, to control all crime types that occur in the city. Devoting appropriate crime prevention programs, and investing in recent technological advances such as cameras, can help prevent crimes daily. Safety and public awareness are consistently lacking in major cities; therefore, a crime analysis can help communities take proactive measures based on criminal understanding in each Boston district, ensuring safety for the general public. In essence, Boston crime analysis plays an extremely important role in enhancing public safety measures and providing insight into the social dynamics that influence a variety of crime activities.

II. Analytical Questions

1) Are there any trends in the most common crime offenses when comparing days of the week with districts?

The question will help us analyze if certain crimes occur more frequently across the different Boston districts. This can help increase law enforcement and police security. We will answer this question by creating pivot tables and plot heatmaps to help us break down the different crimes by day and district.

2) Can we use historical crime data to predict future trends and seasonal crime variations?

This question will help us forecast future crime trends based on past data to prepare public authorities as well as citizens accordingly. We will answer this question by performing time series analysis and exponential smoothing.

3) What underlying factors might explain the clustering of crime types by days of the week?

The purpose of this research question is to observe types of clustering on certain days. Analyzing this question will help us see how different socioeconomic factors impact crime rates. We will use hierarchical clustering to group similar crime types and analyze each crime occurrence per day.

4) Is it possible to identify unique crime pattern clusters among districts by analyzing their crime counts and diversity?

This research question will identify whether certain districts have unique crime features. It can help us identify the specific needs that each district might need. We will answer this question by performing K means clustering to crime counts and crime patterns.

5) What seasonal trends and long-term changes occur in crime rates over time?

This question will identify long-term trends and seasonal variations in crime rates by performing seasonal decomposition. This can help us understand seasonality effects and residual occurrences as well as target the evolving nature of planning with effective strategic planning.

III Data(materials)

The Boston crimes dataset chosen to perform this crime analysis is publicly available on Kaggle. The dataset consists of crime variables such as offense code, which is the type of offense crimes occurred, district, which is the area coordinates of Boston police departments, occurred on date, which is the date and time that the offense occurred, and location, which is the location coordinates.[2]

We also found the exact districts from a referenced Boston Police Department website and changed the police department codes with the relevant districts.*[3]* These extensive variables are suitable for answering our analytical questions since we can perform a detailed to get insight into the frequency, nature, and geography of each criminal activity. The occurred on date column is essential for our time series analysis of how crime rates evolve daily, weekly, and annually. This temporal data enables our crime analysis on different crimes and districts. We have a large amount of historical data from 2015-2018 so we can easily use that data to help answer our time series prediction question where we aim to identify future seasonal variations and take the necessary courses of action to prevent future crimes. The offense code column consists of a variety of categorical criminal activities, which is key to understanding the complexity of criminal behavior and implementing crime prevention strategies. The variety of crime offenses will help us with our K means and hierarchical clustering to cluster the types of crime offenses which can also help to prevent future crimes. Both the occurred on the date and the offense code columns are suitable for our seasonal decomposition question to identify seasonal trends and variations in crimes. Our main data assumptions are that crimes are consistently reported over the years and that the dataset accurately represents Boston crimes.

IV Analysis

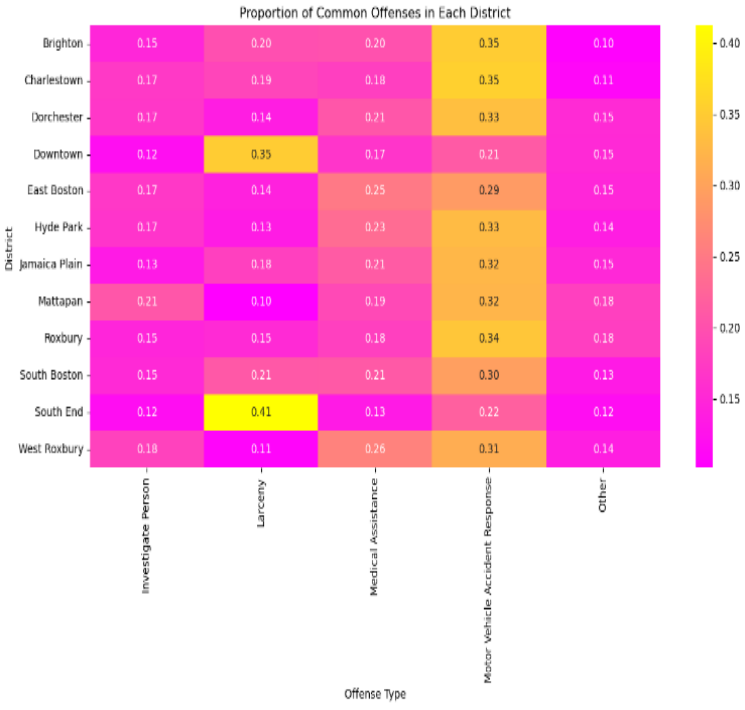
1) Are there any trends in the most common crime offenses when comparing days of the week with districts?

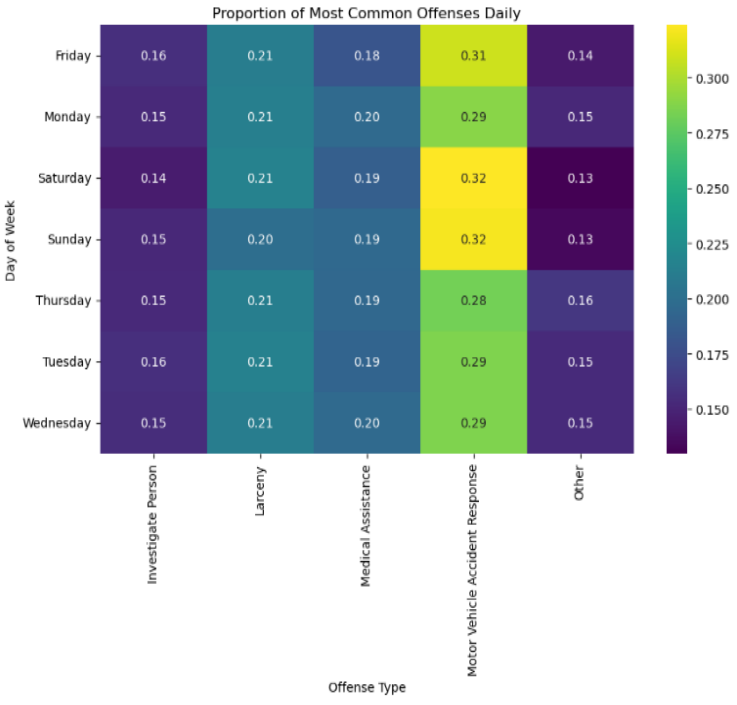
- Data preparation

About the data preparation section for this analytical question, we are working with the offense code group, day of the week, and district columns. We created the district column by changing the Boston Police Department codes that were initially under the district column with the actual district names of each department, which were taken from the Boston Police Department website. The other two variables were already available in our dataset. We used the value counts function to calculate the top 5 most common offenses present in Boston. Furthermore, we decided to organize and prepare our data by creating a pivot table to help us answer our research question. We finally normalized our values and converted them to a 0-1 scale so comparative analysis would be easier.

- Data derivation

In this stage, we derive visual insight by creating a heatmap plot using the heatmap function so that we can visually represent our normalized data from the pivot tables. A heatmap helped us derive valuable insight into crimes across different days and districts, making this initially complex dataset easier to understand. The colorful heatmap is an effective way to transform numerical and categorical data into a visual format.





2) Can we use historical crime data to predict future trends and seasonal crime variations?

- Data preparation

This step involved converting our occurred-on date column into a date and time format using the panda data frame to proceed with our temporal analysis. We then used the resampling method to aggregate the crime data every month.

- Data derivation

In this section, we incorporate trend and monthly seasonality derived from the initial characteristics of the data, which are essential in setting up our exponential smoothing model.

- Construction of models

We construct a predictive model where we use historical data to perform exponential smoothing. We forecast future crime rates for the next 12 months.

- Validation of results

By creating a time series plot we can visualize our historical crime counts as well as the futuristic forecast so we can assess how the model can capture trends and seasonality.

3) What underlying factors might explain the clustering of crime types by days of the week?

- Data preparation

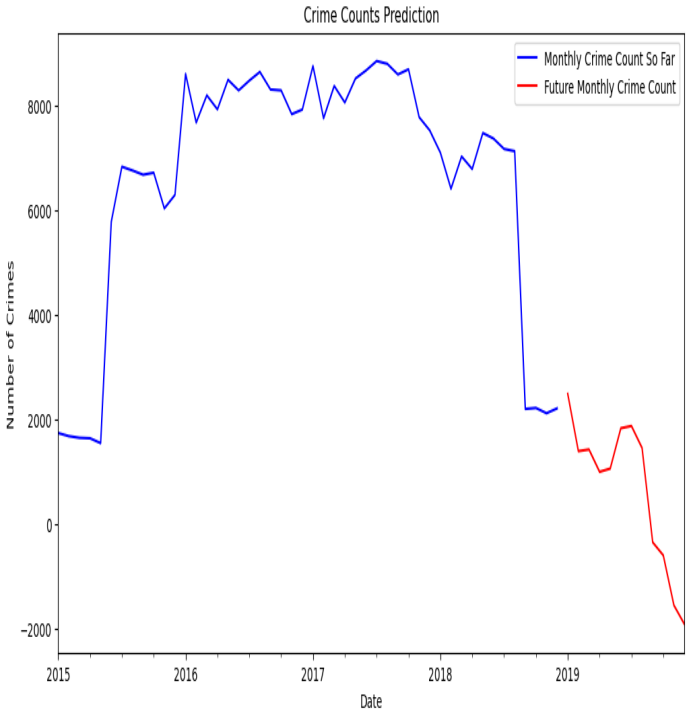
Firstly, we created a pivot table to organize our dataset by offense code and day of the week column. We structure the data in this way so it aligns with our research question of investigating clusters along types of crime by days of the week. Furthermore, we normalized our dataset so we have it in a 0-1 format and facilitate comparisons.

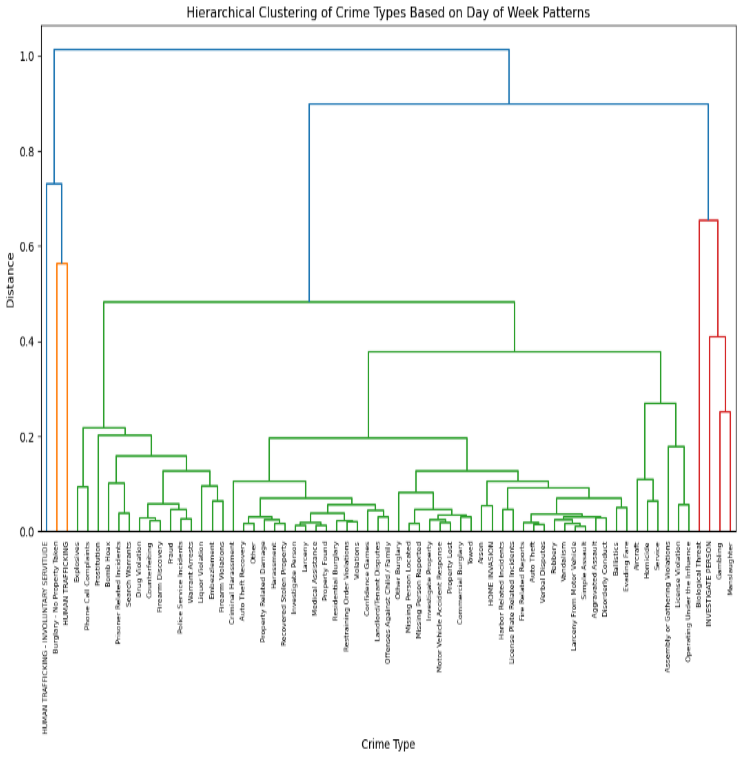
- Construction of models

We perform hierarchical clustering with the use of the linkage command to group crime occurrences across different days of the week to reveal hidden underlying structures and relationships of our dataset.

- Validation of results

We plot a dendrogram to provide a graphical representation of the clustering process, illustrating the different ways that the crime types are being grouped.





4) Is it possible to identify unique crime pattern clusters among districts by analyzing their crime counts and diversity?

- Data preparation

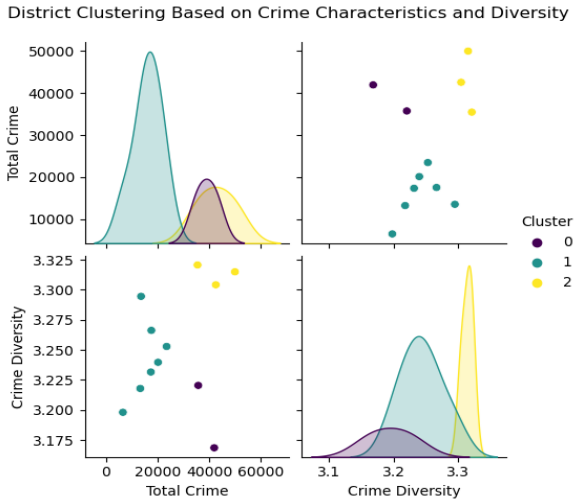
We used the group by function to aggregate the total number of crimes by each district. We then calculated crime diversity using the entropy function to measure the diversity of crime types in each district. Moving on, we created a new dataset by combining the total crime counts with crime diversity. Finally, we again normalize our dataset to have it on a 0-1 scale.

- Construction of models

We implement the K means clustering algorithm to identify clusters among districts based on normalized crime counts and diversity. Finally, we assign each district to one of the 3 clusters that we created.

- Validation of results

We validate our results using a pair plot to visualize and interpret the results of how districts are being grouped into clusters.



5) What seasonal trends and long-term changes occur in crime rates over time?

- Data preparation

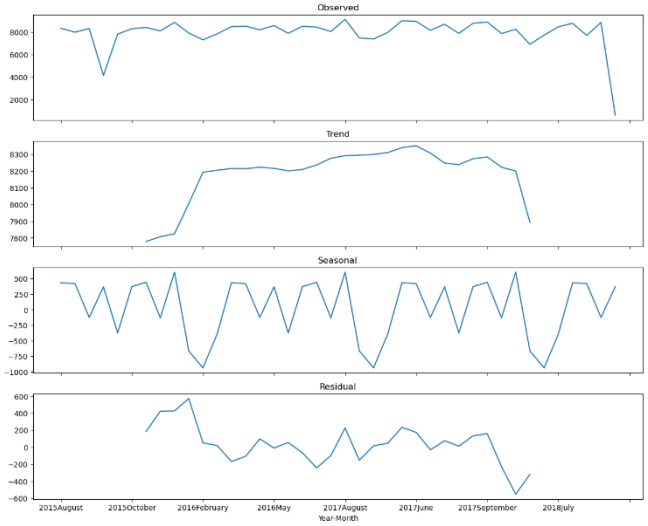
we create a combined time variable which is the year-month variable by concatenating the year and month columns into a single string. After, we group our new variable and count the number of occurrences of each offense so that we can analyze how crime rates evolve every month.

- Data derivation

Using the seasonal decompose function we can break down the monthly offense data into the original time series, as well as the trend, seasonality, and residual components. This helps us isolate and examine individual components of the time series data.

- Validation of results

We create subplots of our time series to visualize the original data which acts as a reference when comparing with the other plots. We also have the trend plot highlighting any long-term changes, the seasonal plot that reveals any seasonal patterns, as well as the residual plot, which includes the random noise of our data.



V) Findings, reflections, and further work

For our first question, we can see from the first heatmap that motor vehicle accident response during the weekend. This helps us identify a trend between traffic-related incidents during the weekend. Medical assistance offenses are also consistently high, with major peaks on weekends. Investigate the person shows an even distribution of crime occurrence indicating that these types of crimes occur frequently. From the second heatmap, we can see that the South End district has a high proportion of larceny offenses indicating that the area needs extra attention and urgent law enforcement. Motor vehicle accidents occur mostly in Downton and Dorchester, indicating that the areas need improvements in road safety as well as better transportation networks. South End and Downtown, have a high medical assistance offense which indicates social problems. Reflecting on our question, we can see that there are some patterns when it comes to certain types of offenses. Also, certain offenses are more common in certain districts which highlights the importance of looking into law enforcement. For our second question, there is an increase in crime rates up to 2018 when there is a significant spike drop. This could be due to effective law enforcement or a decrease in crime rates during that time. Our 12-month prediction shows that there will be a decrease in future crime rates because of the downfall of the red line. Reflecting, the graph helps answer our research question since our model identified a continuous decrease in future crimes providing valuable insight into Boston's future. The dendrogram for our first clustering question shows that the majority of crimes have similar occurrence patterns since they are closely together. Examples include vandalism and graffiti which can both occur when public spaces are more crowded. However, there are also crimes where the cluster distance is very far apart indicating that they need a different approach to limit the number of crimes in Boston. Examples include Homicide and Fraud. Reflecting, indeed some offenses occur more frequently and on the same day, showing the influence of economic and environmental factors. This analysis can provide patrols in vulnerable areas, and also prompt further investigation. For K means clustering there is a visible trend where higher total crime counts also show a higher diversity in crime types. In popular areas, we have more activities going on leading to a variety of crimes. Reflecting, the cluster analysis shows that districts can have distinct crime patterns. This outlines that districts with higher total crime and diversity should have a wider range of protective measures. For our seasonal decomposition graph, the observed graph shows fluctuations between August 2015 and July 2018 and sees an increasing trend until a peak in late 2017. Consistent seasonal patterns and small effects of white noise from the residual trend are also visible. Reflecting, trend, and seasonal components can guide resource allocation and protective measures since we saw a period of rising crimes, followed by a downfall. The data and analytical research conducted during this coursework, provide useful insight into crime trends, highlighting when and where crimes are more likely to occur. We were also able to find specific issues across different districts as well as any anomaly findings that did not follow a general trend. For future work, we can address crime causes by conducting qualitative research such as surveys and evaluations. We can also incorporate additional sources of data to identify crime under various conditions. This approach can lead to a better. understanding of crime dynamics and simultaneously more effective strategies for crime prevention in Boston.

REFERENCES

[1] "About Crime Analysis." Accessed: Dec. 24, 2023. [Online]. Available: https://www.iaca.net/about-crime-analysis

[2] "Crimes in Boston." Accessed: Dec. 24, 2023. [Online]. Available: https://www.kaggle.com/datasets/AnalyzeBoston/crimes-in-boston

[3] "Districts – Boston Police Department." Accessed: Dec. 24, 2023. [Online]. Available: https://police.boston.gov/districts/

Word count

Abstract: 135/150
Introduction: 218/300
Analytical questions: 293/300
Data(materials): 296/300
Analysis: 811/1000
Findings, reflections, and further work: 589/600

Dataset links

https://www.kaggle.com/datasets/AnalyzeBoston/crimes-in-boston
https://police.boston.gov/districts/#:~:text=URL%3A%20https%3A%2F%2Fpolice