The Correlation between Police Salaries and Crime Rate in the City of Boston

Jinghong Chen, Yu Ji, Sijia Peng, Xinghua Peng (Project Leader)

Northeastern University, Boston, MA, USA

Abstract

Federal Bureau of Investigation (FBI) data shows that the number of murders rose nearly 30% from 2019 to 2020 – the largest single-year increase ever recorded in the United States (Lucas, 2021). Because of the surge in criminal activities, redirecting fundings toward the police departments is urgent. Therefore, we will be focusing on crime rate & quantity and police officers' earnings in Boston during the past 10 years. We hypothesized that there is a positive correlation between these two variables. By visualizing crime activities and policemen's payrolls over Boston, we analyze the existing pattern and make future predictions. The pattern will enable college students to better understand the wage standard, percentage growth, and vocational opportunities in the police industry. Police departments can adjust their existing employee distributions to better match the local demand. Government agencies may use the data to switch their subsidies to specific police department sectors and alleviate crime issues or uneven labor distribution problems in the Boston area. However, our project turned out to show a negative correlation between crime quantity and the annual average income of the police department in Boston.

Keywords: *Crime quantity, police department, salary, correlation, funding redirection, Boston*

Introduction

Our objective is to find current payrolls and employment distributions among different sectors in the police department, predict future salary patterns, and get an optimum salary plan for police officers and government agencies. We will present the annual average income in the police department and the total crime cases from 2011-2021 in Boston. To get a specific industry overview, we will sort the data based on overtime records. We aim to find a relationship (either positive or negative) between these two variables by using functions created in python with visualizations generated from packages & libraries. Our working hypothesis is to find a positive correlation between them since the areas with high criminal activity frequencies tend to require higher workloads, more hirings, and thus require higher salary packages. The result can help generate an outlook for job seekers, police officers, and political decision-makers. At the same time, based on the statistical summary of police salary and the number of crime cases, we also project the number of crime cases in the following three years. Beyond the ivory tower, we are trying to provide a broad view of criminal activities in Boston for citizens. Criminal activities are likely to be related to criminals' motivations and government policies, but our project will provide a unique perspective to develop a correlation between crime quantity and the annual average salary of police officers.

Methods

Packages & Libraries that we used during this project:

- CSV: Reads and writes comma-separated values (CSV) files
- Matplotlib: Comprehensive library for creating static/animated/interactive visualizations.
- Pandas/GeoPandas: Powerful data analysis and manipulation tool.
- Seaborn: High-level interface for drawing attractive and informative statistical graphs
- Sklearn: A set of fast tools for machine learning and statistical modeling.
- NumPy: Add support for large, multi-dimensional arrays and matrices.
- WordCloud: Represent text data in which the size of each word indicates its frequency.

I Data Munging

dangerous_streets (data, img)
This function produces a word cloud of the most dangerous streets in Boston. The parameters are crime incident data and the "gun" background picture. Some of the data processing techniques are storing the street word frequency in one dictionary, extracting all streets, and calculating the word frequency statistics with *for* loops. Then we inputted the background image and created and configured word cloud properties.

II Data Filtering

extract_column (data, colidx, coltype = str)
This function returns data from specific columns in the form of a list. The parameters are data, representing all data, and colidx, representing the column.

get_zip (data)
This function aims to get the zip codes in common from each year's data. We created an *if* statement to store the first year's data into the list if there is no data yet and an *else* statement to get the intersection with the previous results. Then, we got the zip codes that are present in all years of data.

III Data Transformation

read_data (filename, header, coltypes)
This function is meant to read the data into a list of dictionaries and return the list of data. During the data cleaning process, we stored and parsed the key-value pairs without the missing values. Then, we removed irrelevant characters (",", "(", ")", "$") and kept the numbers only. For zip codes, we tried to only keep the initial five numbers (e.g., 02124 for 02124-3735), and add zero or zeroes before the numbers if there are less than five digits (e.g., 02124 instead of 2124).

read_all_data (filename, start, end, header, coltypes)
This function is also meant to read the data into a list of dictionaries and return the list of data, but with proper headers to show the years accordingly.

zip_total (data, title)
This function converts all the elements of the year in the list into strings to prevent each value from being omitted. It also helps to retrieve the annual income data from each zip code.

IV Data Visualization

get_color (length)

This function creates a color sequence to show symmetrical gradient colors. The parameter length refers to the number of desired colors. To ensure the colors are symmetrically from the middle, we created the sequence, reverted it, and spliced it with the original one.

bulletgraph (l, aim, title, color)

This function helps to build a bullet graph and return a matplotlib figure. The parameters are l, aim, title, and color. L means the label, measurement, and lists of the target value; value means target value; title is the title of the chart; and color shows the background of the stacked bars. After setting the limit and labels for the chart, we created color values for appended data and added each bar to the graph accordingly. Within the adding process, we got an axis, removed unnecessary labels, designed the background, built a horizontal bar graph, and drew marked lines. Eventually, we produced a well-labeled bullet graph with a proper distance between the figure layer and sub-figures.

annual_forecast (annual)

This function helps to project the police department's gross income and predict the corresponding gross income after three years (2022-2024). We first created a data frame and labeled columns corresponding to the total income after N days. Then we constructed testing and training datasets and features used for model prediction. Through training the data with the linear regression model, we found that the R Square equals 0.80, and plotted the trend comparison of the predicted and actual values. After calculating the predicted and actual values, we drew the comparison figure.

offense_group (data)

This function returns the statistics of the number of crimes with their offense types. We extracted the different crime type data from each year, calculated the word frequency using a *for* loop, converted the statistics into a list, and plotted the bar chart.

crime_hour (data)

This function outputs the time distribution of the crimes in a day. We classified the day into four different categories as a dictionary: "00:00-06:00", "06:00-12:00", "12:00-18:00", and "18:00-24:00". Then we extracted the time of the crime, calculated the category statistics using a *for* loop, and plotted the pie chart.

crime_map (case_data, filename)

This function aims to read crime data and return a crime map in Boston. We read the file of the Boston map, generated Boston's administrative divisions by polyplot under geoplot, and marked the zip codes and the locations for criminal activities. We took the longitudes and latitudes of statistical criminal records into consideration and only kept the coordinates within the mapping area. Then we made a few edits including adding labels and removing axes and edges to make our graph more aesthetically pleasing.

crime_salary (data, start, end, case_data, case_start, case_end)

This function is designed to show the relationship between crime quantity and the average annual salary of police departments. The first part was to test linear regression. We generated an empty list to store crime data and append it from the data files. By selecting the salary data of the same year, and assigning value to salary, year, and crime rate, we created a scatterplot. To show the trend line between two variables, we implemented a regression model with *DataFrame*. The second part of the function is a bubble chart showing the relationship between wage growth rate and the number of crimes according to years. We computed the wage growth rate based on the annual income data and appended the statistics to our empty list *salary_ratio*. Since the growth rate is too low to be observed (close to 0), we expanded it 3600000 times to visualize it.

shooting (case_data)

This function helps to generate the shooting frequency per day of the week within a bar chart. We created an empty dictionary to store the count of gun-shooting cases. We located the shooting data in our data file and assigned the value *y* to indicate a gunshot. With a *for* statement, we added one to the current number if data is already available on that day of the week; otherwise, we added it to a dictionary and set the value to one if data is not available on that day. Then, we converted the statistics into lists, appended them accordingly, and generated the chart.

zip_income_sort(data, zips)

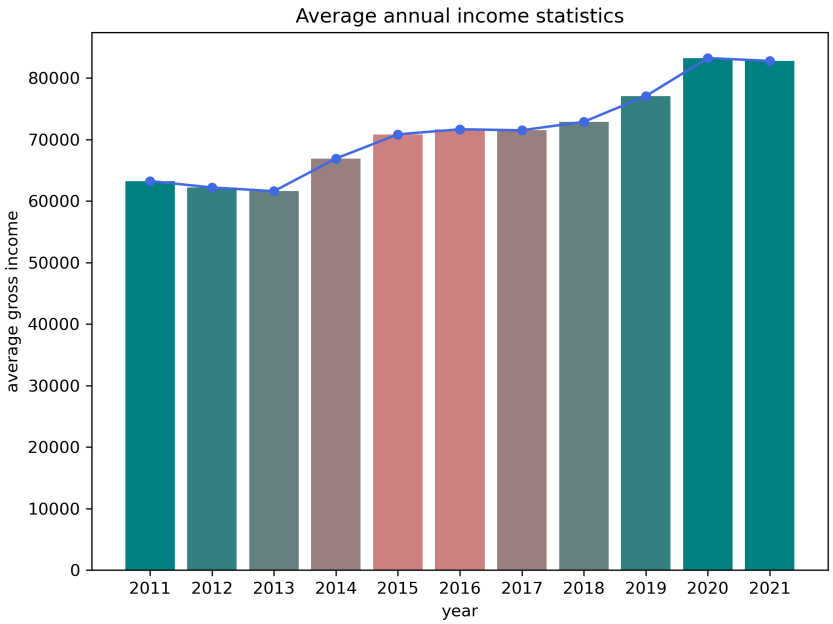
This function intends to get data from five regions with the highest average income and the lowest ones, and then creates one bullet graph for the highest and lowest regions respectively. We generated an empty list and gave a few headers to it: [{zip code: , average income: [], average total income:}]. After getting the annual income from each zip code, we calculated the annual income for each year by using *for* loops. To compute the mean, we divided all appended annual income data by the number of data pieces we have gotten and plotted the bullet graph according to that.

income (data, header, coltypes)

This function aims to get statistics of average annual income and gets a bar chart with a trendline. We created empty data lists with headers, calculated the average value for each column by dividing the gross income by the number of data pieces in that year, and appended corresponding data under different headers. With the total earnings and years, we successfully generated our bar chart and line chart from 2011 to 2021.

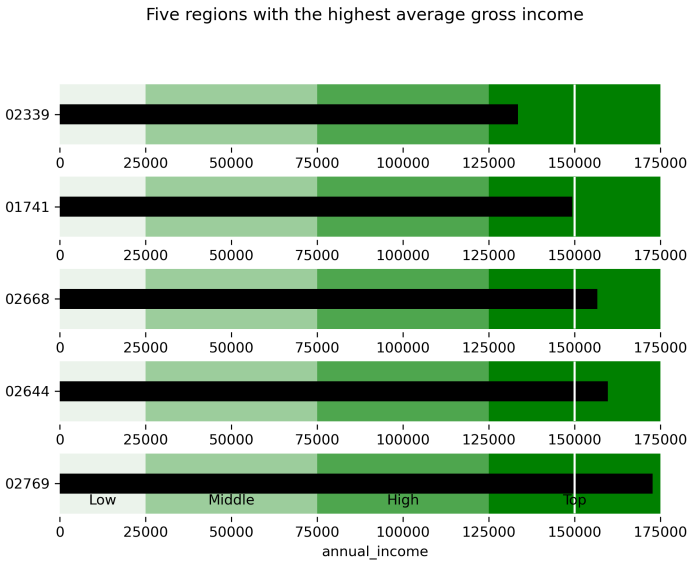
Analysis

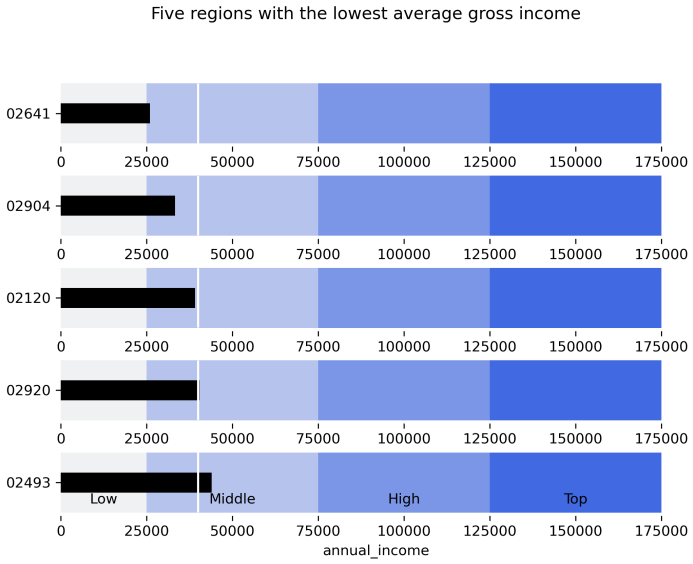
Figure 1 – Average annual income of police departments [2011-2021]



Explanation - The average income of the police department shows an overall upward trend. The overall growth rate is relatively slow and peaks in 2020. The lowest average income is in 2013 and the highest is in 2020.

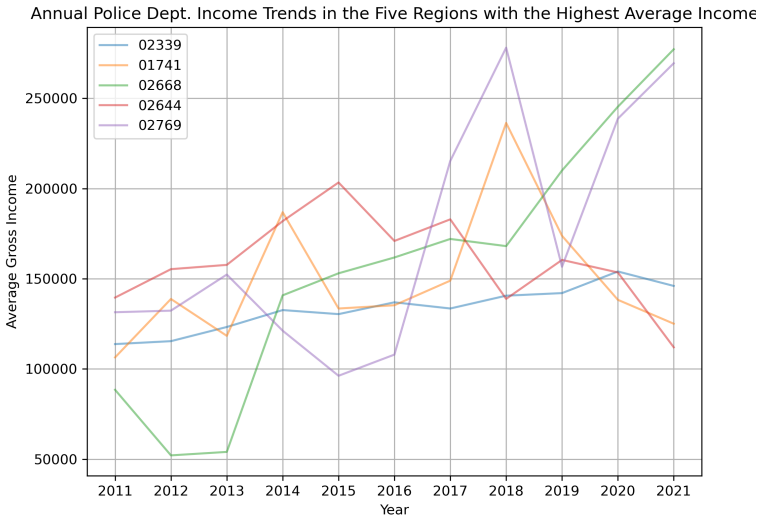
Figures 2 & 3 – Five Regions with the Highest/Lowest Average Gross Income [2011-2021]

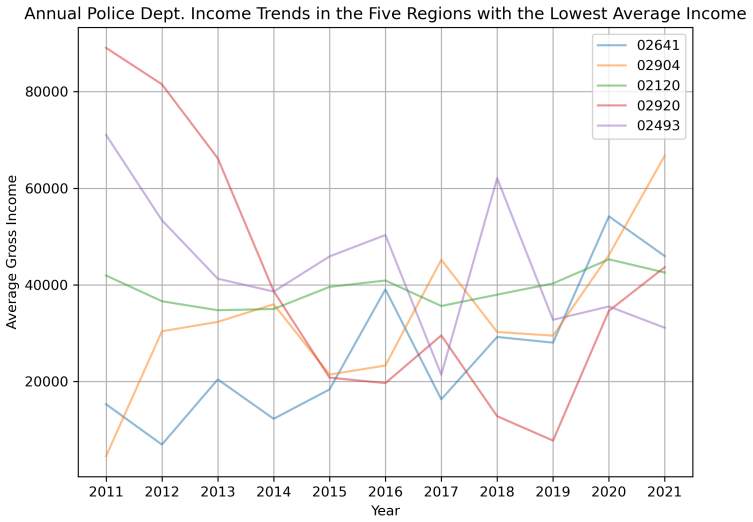




Explanation - The highest average gross income level for all five regions is above $125,000. The highest average gross income level of the three regions exceeds $150,000. The region with the highest number has the zip code of 02769 and the region with the lowest number is 02339. The lowest average gross income level in all five regions is below $50,000. They are in the lower middle-income group. The area with the lowest number is 02641 and is close to the low-income group.

Figure 4 & 5 – Annual Police Department Income Trends in the Highest/Lowest Five Regions in the City of Boston [2011-2021]



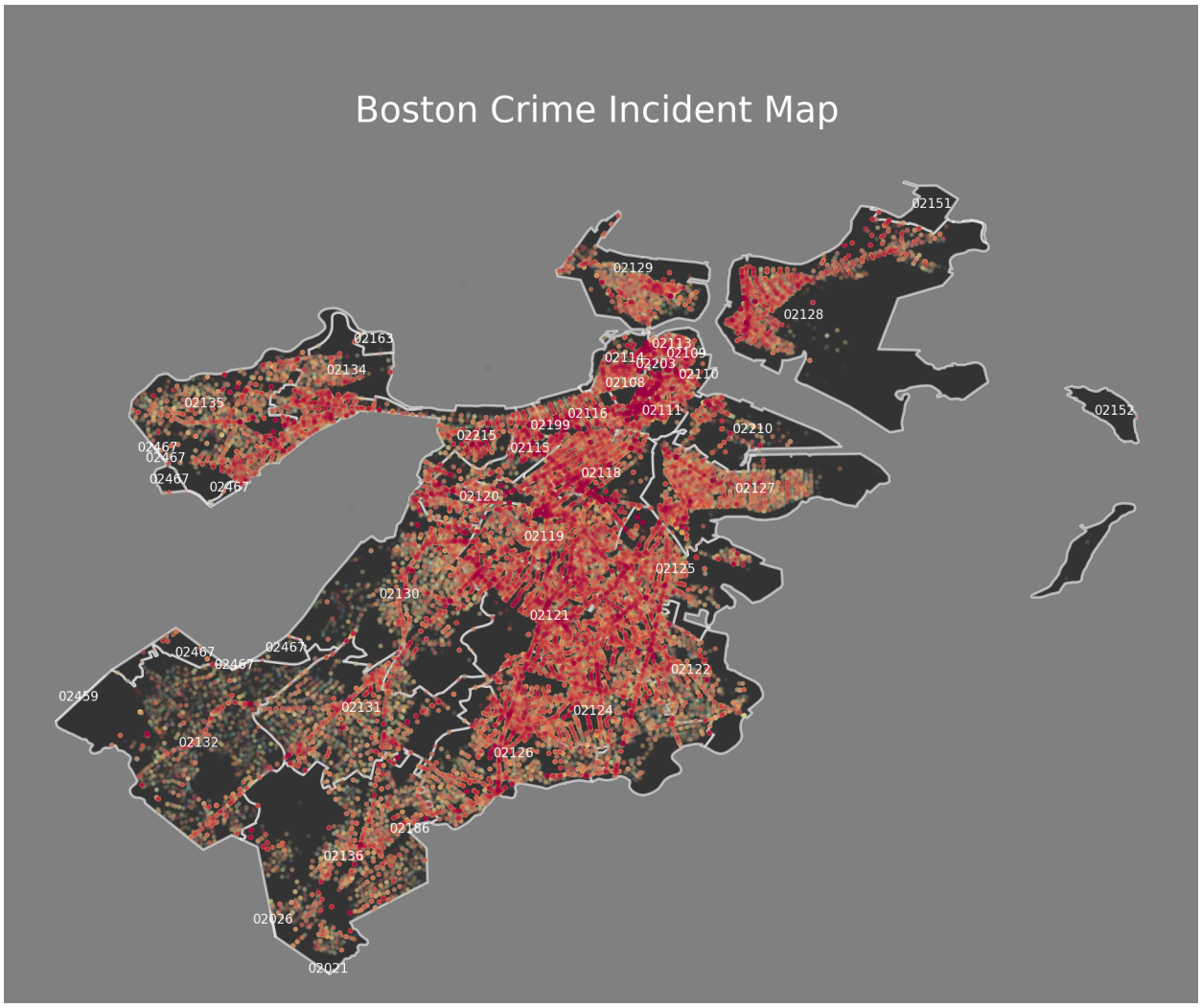


Explanation - The data in these line charts both show a fluctuation pattern of average income levels in selected regions from 2011 to 2021.

Most regions with the highest average income show an upward trend, and 4 out of 5 have higher incomes in 2021 than in 2011. Only a region with the zip code 02644 demonstrates a downward trend in its annual average income from 2011 to 2021. The area with the highest annual average income in 2021 has a zip code of 02668, with the highest overall growth rate at the same time.

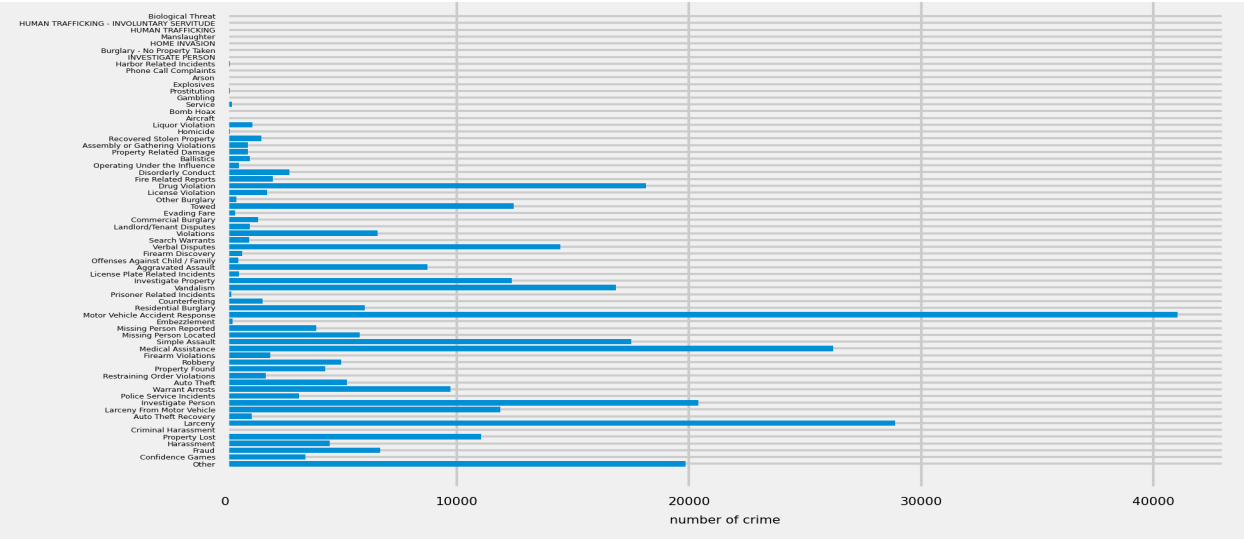
Most regions with the lowest average income show a downward trend. Although the gaps between the five lowest-income regions are wide in 2011, they gradually narrow down in 2021.
Only two regions out of five have a higher annual income in 2021 than in 2011: 02641 and 02904, with 02904 showing the largest increase over the decade.

Figure 6 – Boston Crime Incident Map



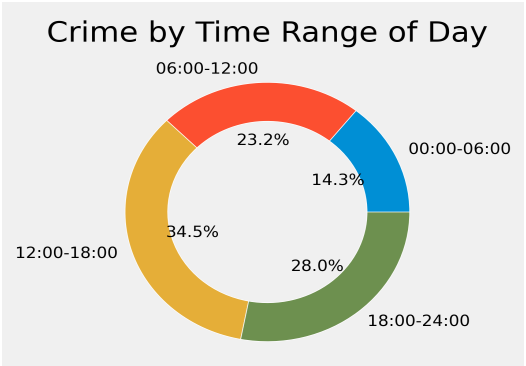
Explanation - We used the data to draw this Boston crime incident map, using a scatter plot to represent each crime. By using the colors, we can see that the darker areas are crime-intensive areas. On the contrary, the lighter areas represent less frequent crimes. By comparison, we can find that the safer areas are 02128 (Boston Logan Airport with high-level security), 02110 (Seaport), 02459 (Oak Hill), and 02026 (Dedham). The more dangerous areas, where crime often occurs, are 02203 (Government Center), 02108 (Beacon Hill), and 02114 (West End).

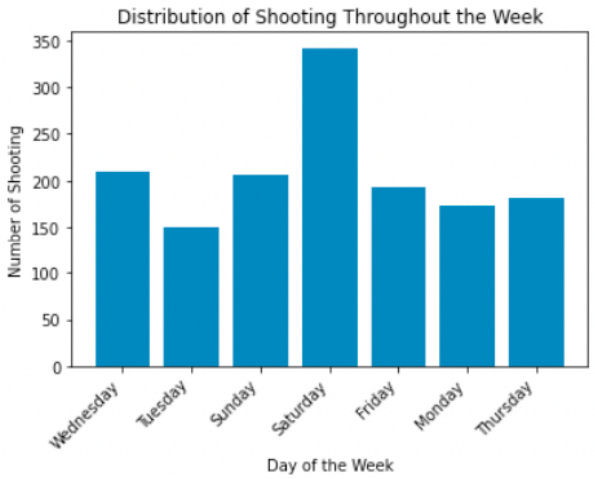
Visualization 7 – Crime Types in Boston

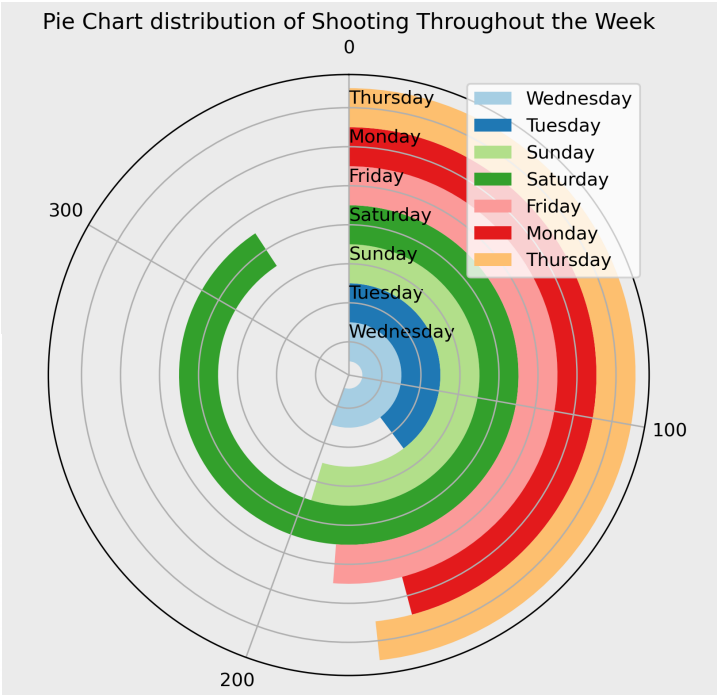


Explanation - We compared the number of each different crime using a bar chart with the x-axis representing the number of crimes and the y-axis representing the type of crime. Based on this bar chart, we were able to infer the most common types of crimes. The most frequent crime is "Motor Vehicle Accident Response". The second and third most numerous are "Medical Assistance" and "Larceny".

Figure 8 – Distribution of Crimes (shooting-related) Quantities Throughout the Week



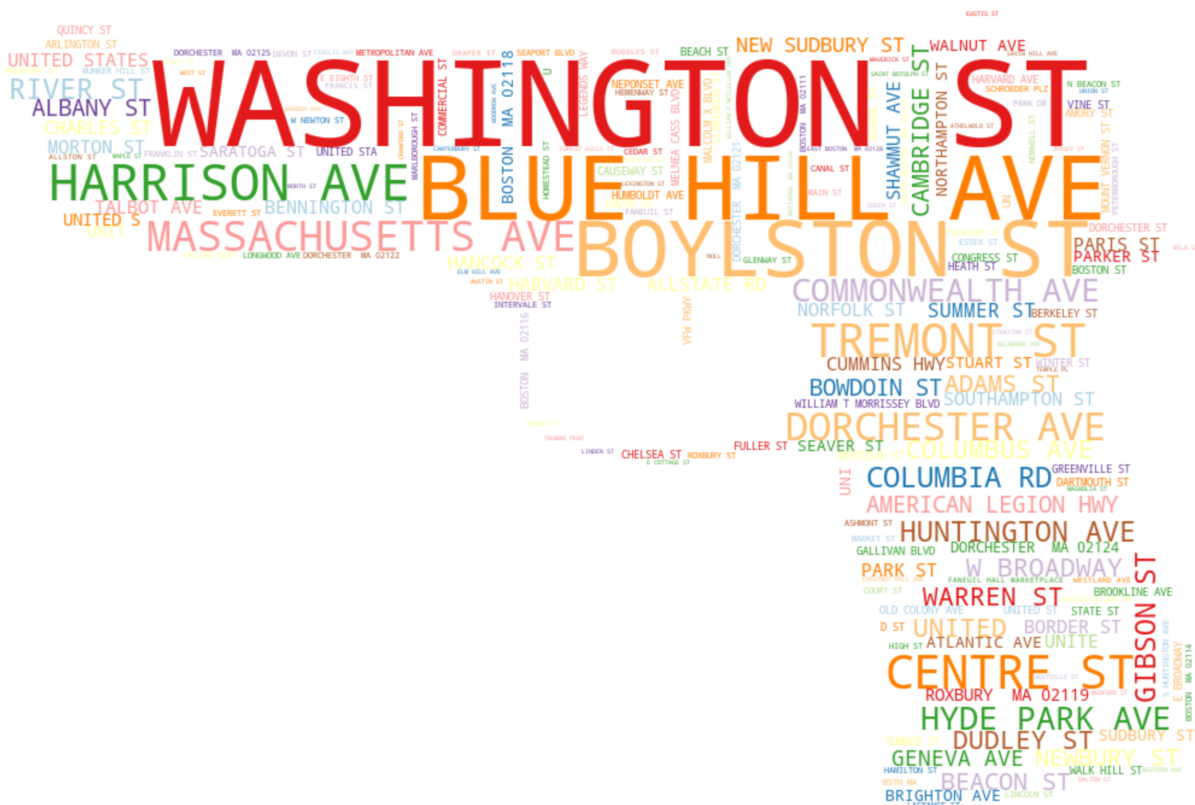




Explanation - According to the graph above, we could see that most of the crimes are committed between the time range of 12:00-18:00, which has the most frequent criminal activities percentage (34.5%), and the time between 00:00-06:00 appears to have the least cases of crimes (14.3%).

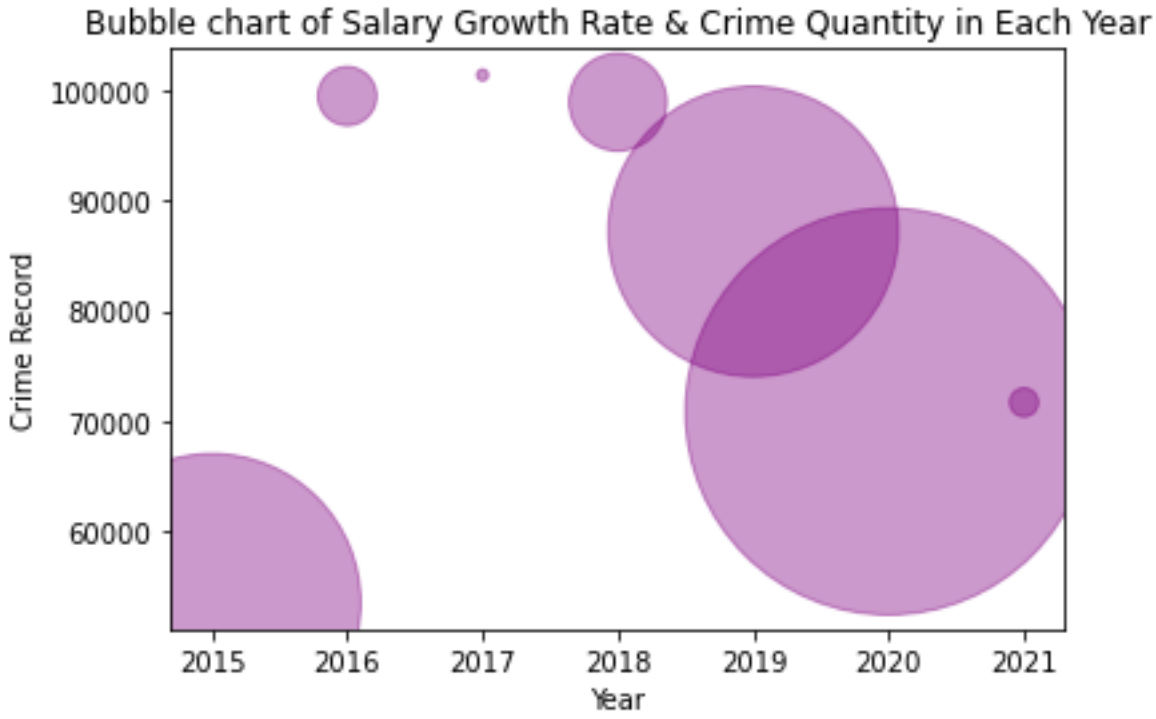
We built two graphs to display the distribution of shooting-related crime throughout the week, one bar chart and one pie chart. Based on these plots, Saturday is the day with the most frequent shooting-related crime while Tuesday with the least. The other days have a similar shooting-related crime frequency, so we suggest redistributing some police force from Tuesdays to Saturdays, while keeping the employment force for other days the same.

Visualization 10 – World cloud of the most dangerous streets in Boston



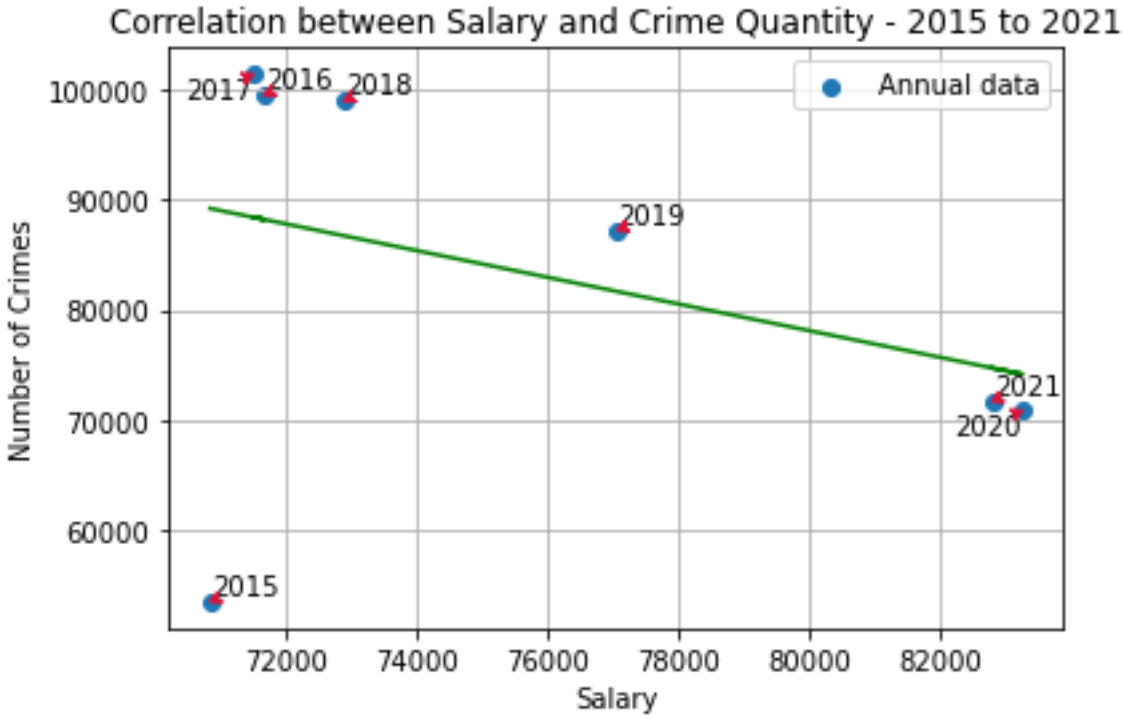
Explanations - This is a word cloud of the most dangerous streets in Boston. Washington St. appears the be the most dangerous street in Boston, however, one limitation is that there are Washington streets in multiple districts. The second dangerous street is shown to be Blue Hill Ave and the third is Boylston St. Based on the data from 2015 to 2021, other streets with a relatively high crime quantity include Harrison Ave, Centre St, Massachusetts Ave, Tremont St, Dorchester Ave, Hyde Park Ave, and Commonwealth Ave.

Figure 11 – Police Average Salary Growth Rate & Crime Quantity [2015-2021]



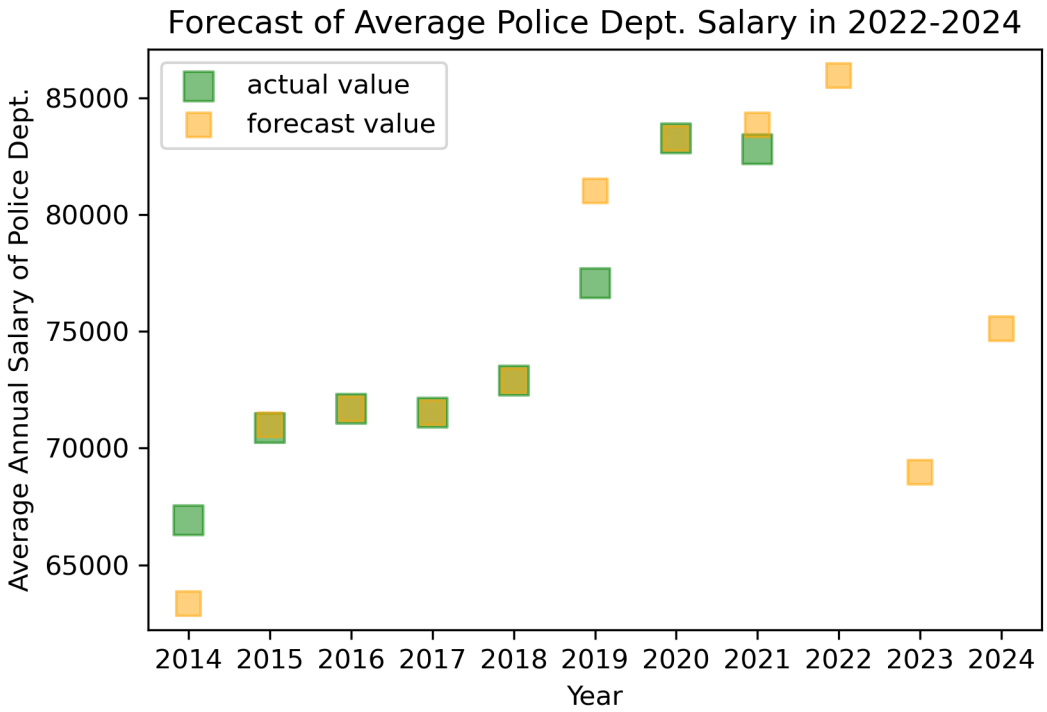
Explanation - We used a bubble chart as an extension of a scatterplot to visualize relationships among three numeric variables – salary growth rate, crime quantity along with time. Each bubble in the chart represents a single data point. There is an overall salary increase and a declining pattern in crime quantity from 2015 to 2021. There were the least cases of crime in 2015, and most in 2017; police officers had owned their highest average annual salary growth rate in 2020 and lowest in 2017. The salary growth rate decreased from 2015 to 2017, then increased from 2017 to 2020, with a sudden decline in 2021. A possible explanation might be the presidential election results in 2017 and 2021, when Donald Trump and Joe Biden's tenures started, respectively. Based on the graph, except for outliers, the salary growth rate is negatively related to the crime quantity. The higher the crime quantity, the lower the salary growth rate.

Visualization 12 – The correlation between salary and crime quantity [2015-2021]



Explanation - We use the data from 2015 to 2021 to illustrate the correlation between salary and crime quantity. There is a negative correlation between these two variables, with an R-square of 0.80, representing approximately 80% of the variability observed in our data is explained by the regression model. As shown by our trendline, the slope is –1.21, suggesting a one-dollar increase in the police annual salary will result in a decrease of 1.21 cases in crime quantity. The y-intercept is 174,820.88, indicating the number of crimes will be 174,820.88 when police officers have zero income (unlikely). There is an overall decreasing crime quantity trend, where crime quantity surged from 2015 to 2016 and decreased annually from 2016 to 2021. However, is an outlier in the year 2015, and a possible explanation would be Hillary Clinton and Donald Trump became the presumptive nominees for the final major state primaries in the 2016 presidential election in the US, and thus the crime quantity surged as a result of the instability in society. The crime rate then gradually lowered as insecurity decreased. Meanwhile, an increasing trend in police officers' annual income is observed, with the highest growth rate in 2019.

Visualization 13 – Forecast of Average Police Department Salary [2022-2024]



Explanation - The R square of our model equals 0.80, which illustrates that 80% of the variability in our actual data is statistically significant and can be explained by our forecast. Therefore, we conclude the prediction as relatively accurate, with five years of forecast data overlapping with the actual values. An overall upward slope is observed from our current data. Based on existing statistics, our forecast model predicts a summit in the average annual salary of the police department will occur in 2022, more than $85,000, with a sudden drop in 2023 and a quick recovery in the following year.

Conclusions

Results

Our statistical analysis reveals the areas with the most (02203) and least (02128) frequent criminal activities; the day with the most frequent criminal activities is Saturday during 12:00-18:00, and the most dangerous street is Washington St. The average annual salary of the police department is negatively correlated with the number of criminal activities but has not been proven causality. There is an overall increasing trend in the average annual income and a growth rate of it in the police department. Meanwhile, a declining pattern is observed in crime quantity from 2011 to 2021. Based on our prediction, the police average income will continue growing but with a sudden fall in 2023. Since the salary is extremely low in 2014, that piece of data may cause a sudden drop in 2023. Sociocultural causes may be the long-lasting effects of covid-19, which contribute to saturation in the police department. With the decrease in salary in 2023, there will be an employment shortage in the police industry and thus the government may increase the budget again to recruit more members in the following year, 2024.

Accomplishments

Our report reveals the most dangerous areas and times in Boston in the past ten years to increase awareness among citizens. The current and future salary patterns in the police department will provide useful insight to college students who pursue a future degree in the police industry. Police officers can increase their labor force in areas with the most frequent crime activities, such as 02203. They may also distribute more police officers from weekdays to Saturday afternoons to prevent and control possible criminal activities. To match the local and police department demand, the government may redirect their funding strategies, such as reallocating more assistant police forces in the areas with the most crime activities to maintain public order in Boston.

Limitations

There are many districts called Washington Street that make it impossible for us to separate the different Washington Streets. So, we can only count them as a whole. This limits the accuracy of our assessment of the danger level of Washington Street.

Algorithmic prediction of crime data would reinforce the prejudice in the existing data. Algorithmic statistics, a method that appears to be scientific, neutral, unbiased, and rational, hide the fact that the crime database they collect is just a summary of the neighborhoods they used to patrol. The neighborhoods that are often patrolled by police are often poor, colored neighborhoods, because some of the police are likely to be systematically racist. The results obtained from such data further reinforce this bias, allowing neighborhoods that have been heavily patrolled in the past to continue being heavily patrolled. The over-policing of these areas leads to huge consequential problems, such as uneven police force allocation and labor shortage. Thus, we need to look at the data critically and be aware of the prejudice, recognize and accept the fact that uncertainty is normal in life, and not allow that unpredictability to be exploited by the authorities or capital corporations. In addition, other determining factors may also contribute to the increase in police pay, such as an urgency to recruit police forces and presidential elections.

For a future direction, we consider taking the actual number of police officers into account to generate a more accurate outlook of criminal activities in the city of Boston.

Author Contributions

We were first individually assigned specific leadership roles according to *The Data Science Hierarchy of Needs*. Each section was taken over by one Lead and two Key Consultants to enhance the quality of our work and ensure that everyone was on the same page regarding each step. Last but not least, we scheduled weekly touch-base meetings to work and collaborate on ad-hoc analysis and tasks like report writing, documentation, and group presentation. Please refer to our team dynamics & details below for more information:

Learn/Optimize – A/B Testing, Experimentation, Simple Machine Learning Algorithms
- **Lead: Xinghua Peng**
- Key Consultants for additional support: Sijia Peng, Jinghong Chen

Aggregate/Label – Analytics, Metrics, Segments, Aggregates, Features, Training Data
- **Lead: Sijia Peng**
- Key Consultants for additional support: Jinghong Chen, Xinghua Peng

Explore/Transform – Cleaning, Anomaly Detection, Prep
- **Lead: Yu Ji**
- Key Consultants for additional support: Xinghua Peng, Sijia Peng

Collect/Move/Store – Data Flow, Infrastructure, Pipelines, Structured/Unstructured Data Storage
- **Lead: Jinghong Chen**
- Key Consultants for additional support: Yu Ji, Xinghua Peng

Signatures: Xinghua Peng Sijia Peng Yu Ji Jinghong Chen

References

1. *Crime incident reports (August 2015 - to date) (source: New system)*. Analyze Boston. (n.d.). Retrieved November 30, 2022, from https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

2. *Employee earnings report*. Analyze Boston. (n.d.). Retrieved November 30, 2022, from https://data.boston.gov/dataset/employee-earnings-report

3. Lucas, R. (2021). FBI Data Shows An Unprecedented Spike In Murders Nationwide In 2020. Retrieved December 3, 2022, from https://www.npr.org/2021/09/27/1040904770/fbi-data-murder-increase-2020

4. *The White House*. (2022). Presidents. Retrieved December 4, 2022, from https://www.whitehouse.gov/about-the-white-house/presidents/

5. Wang, J. (n.d.). *"This Is a Story About Nerds and Cops": PredPol and Algorithmic Policing"*. Retrieved December 2, 2022, from https://www.e-flux.com/journal/87/169043/this-is-a-story-about-nerds-and-cops-predpol-and-algorithmic-policing/

6. *ZIP codes*. Analyze Boston. (n.d.). Retrieved November 30, 2022, from https://data.boston.gov/dataset/zip-codes

7. ZIP Codes - Boston. (n.d.). Retrieved December 1, 2022, from https://www.cityofboston.gov/images_documents/ZipCodes_tcm3-47884.pdf