Data Visualizations on Violent Crime Rate in California (2000-2013)

Santa Clara University

COEN 396B

Advanced Topics in Computer Science & Engineering - Data Visualization

Aishwarya Gupta, Ankit Malasi, Jessica Torres, Priya Jain

Link to Application

I. INTRODUCTION

According to the California Department of Public Health, neighborhood safety is essential for good physical and mental health. Violent crime disproportionately affects communities of color and young adults. Victims, families, and community members who have been involved in or witnessed violent crimes may develop post-traumatic stress disorder. High crime rates also discourage residents from exercising outside because they may not feel safe doing so. When crime rates rise, people are less likely to interact with their neighbors, which can negatively impact their mental health due to social isolation. Aside from the health consequences, high violent crime rates contribute to negative perceptions of neighborhoods. This can hinder economic development and lower real estate values in areas seen as unsafe. Knowledge about crime rates and their impact helps individuals and communities take proactive measures to enhance safety. This can reduce feelings of vulnerability and helplessness, leading to a greater sense of control and reduced anxiety. Raising awareness and educating the public about violent crime in California is essential for ensuring public safety and safeguarding the well-being of communities across the US.

II. MOTIVATION

This project offers data visualization tools for analyzing trends and patterns in violent crime rates across different regions of California. These visualizations are crafted to be accessible and easy to read, making them ideal for curious citizens interested in understanding local safety dynamics. Prospective homebuyers can use this tool to assess the safety of specific neighborhoods before committing to a long-term residence, enhancing their peace of mind. Tourists can identify areas with lower crime rates, ensuring a worry-free visit. Law enforcement agencies can utilize historical crime data to pinpoint emerging hotspots and allocate resources more effectively. By fostering awareness and education about violent crime, this project aims to strengthen community cohesion, support mental health, and reduce the stigma associated with crime victimization. Overall, it strives to provide citizens, businesses, and public

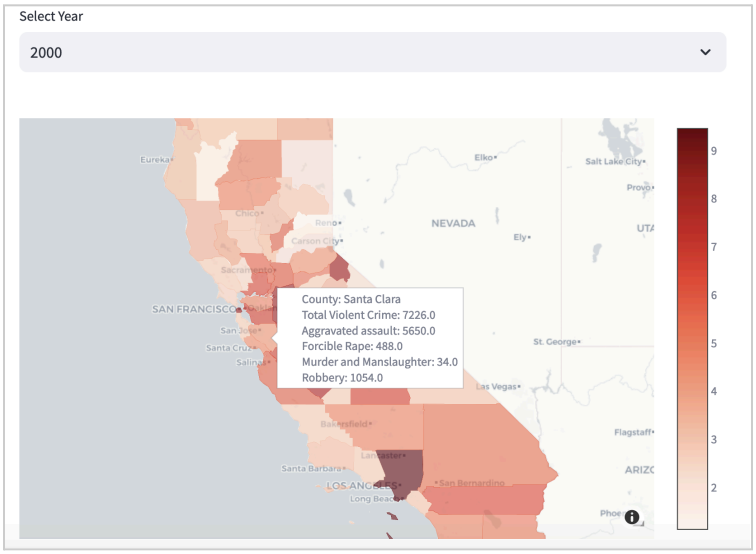

Fig 1. Heatmap of Crime Rate across California Counties per year

institutions with localized insights into violent crime, promoting informed decision-making that can improve public safety and overall well-being in California's communities.

III. VISUALIZATION DESIGN

We implemented three interactive visualizations and two intelligent features to illustrate crime rates across different regions: a heat map, a line graph, a pie chart, k-means geospatial clustering, and an LLM chatbot.

The first visualization, as shown in Fig 1, is a heat map to show geographic variations in violent crime rates in California counties. Heat maps are effective at revealing spatial patterns and trends in location-based data, providing a clear visual representation of how a specific metric varies across a geographic area. In this case, the heat map allows users to quickly identify regions with higher or lower concentrations of violent crime by varying the intensity of the colors. We chose a red sequential color scale because red hues effectively signal warnings or negative connotations, which are consistent with the nature of high crime rates. Furthermore, by using a sequential gradient rather than a categorical gradient, the heat map can represent the entire spectrum of crime rate values, from low to high. To enhance user interaction, the visualization includes a filter that allows users to select a specific year of interest, which dynamically updates the heat map to show the corresponding annual crime rate concentrations per county. Users can explore individual county statistics by hovering over a desired location, which opens a pop-up window with the precise crime rate value and population figures to contextualize the data. This approach combines the strengths of geographic heat maps with interactive elements, allowing users to gain insights into regional crime patterns through an intuitive and exploratory visual interface.

The second visualization uses a line graph to show how the total number of crimes in each California county changed over 13 years from 2000 to 2013. A line graph is an effective tool for displaying trends over time. Users can easily compare crime rate trends by selecting a specific county from the dropdown menu. We chose a line graph for this visualization because it is useful for comparing quantities over time. The line graph feature enables users to track and analyze subtle fluctuations in total crime rates. It allows for the correlation of these changes with legislative developments across various counties, providing a detailed perspective on how policy impacts crime trends. We chose to use red for the line color so that it stands out against the white background. Whether a county experienced a steady decrease, a concerning increase, or alternating periods of higher and lower crime, the line graph effectively illustrates these historical trends. Visualizing crime rates through a line graph can help develop strategies to reduce crime rates and identify the factors that may influence crime rates in different regions. This is represented in Fig 2.

The third visualization is a pie chart that shows the distribution of the total violent crimes committed in each county across different categories. The user can filter through different years and counties. The crime categories

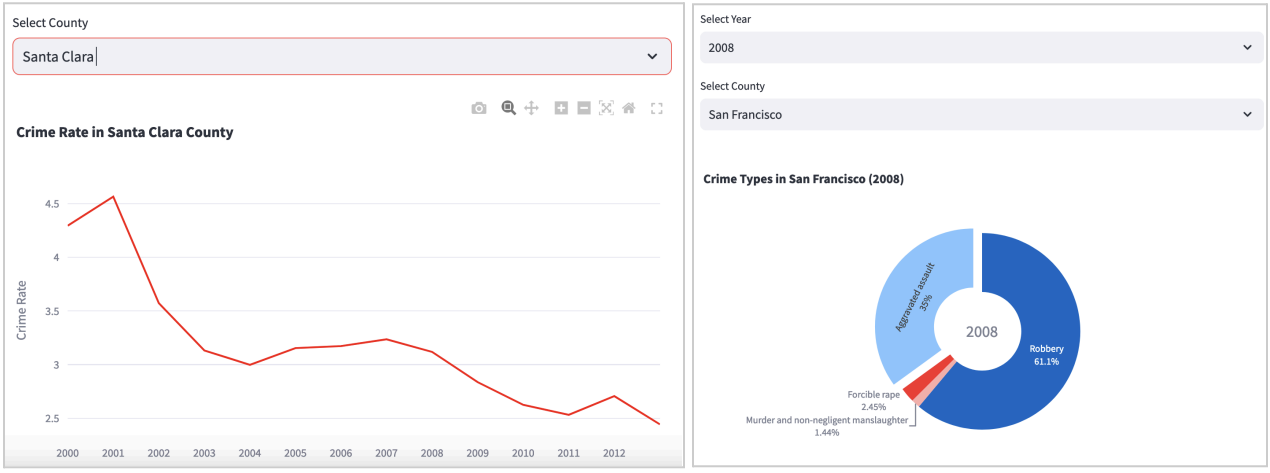

Fig 2. Crime Rate Line Graph over Time and Crime Types Distribution Pie Chart

include robbery, aggravated assault, murder, non-negligent manslaughter, and forcible rape. A pie chart was selected because it shows data as percentages of a whole. In this case, the whole represents the total crimes committed, and the percentages represent the different types of crimes mentioned previously. This visualization provides a clear and straightforward way to pinpoint the most prevalent crimes and track trends in crime types—whether they are on the rise, in decline, or exhibiting cyclical patterns—across each county over the 13-year period.

The fourth visualization is a geospatial analysis with clustering. The user can zoom in on the map and see the different clusters of crimes committed. This type of visualization is effective for identifying and visualizing

clusters of crime rates across different regions in California. The colors are colored using Plotly's qualitative color scheme, which ensures distinct and visually appealing cluster differentiation. Each cluster color corresponds to a specific crime rate category, ranging from low to extremely high. The map shows clusters based on the K-Means clustering algorithm, which categorizes areas into different crime rate levels. K-Means clustering works well with around 5 clusters for spatial data since it mostly captures the major variations in the data without overfitting. Using different combinations of clusters, we found that for our dataset, five clusters provide effective data partitioning. The user can also deselect multiple clusters from our interactive legend to showcase areas ranging from extremely high crime rate areas to low crime rate areas, or any combination of crime rate areas, as shown in Fig 3.

To enhance user interaction and data accessibility, we integrated a chatbot using PandasAI's Bamboo LLM. This chatbot interacts directly with our preprocessed dataset, which is loaded into a DataFrame structure optimized for high-performance querying. The chatbot's primary functionality is to interpret and respond to user queries that

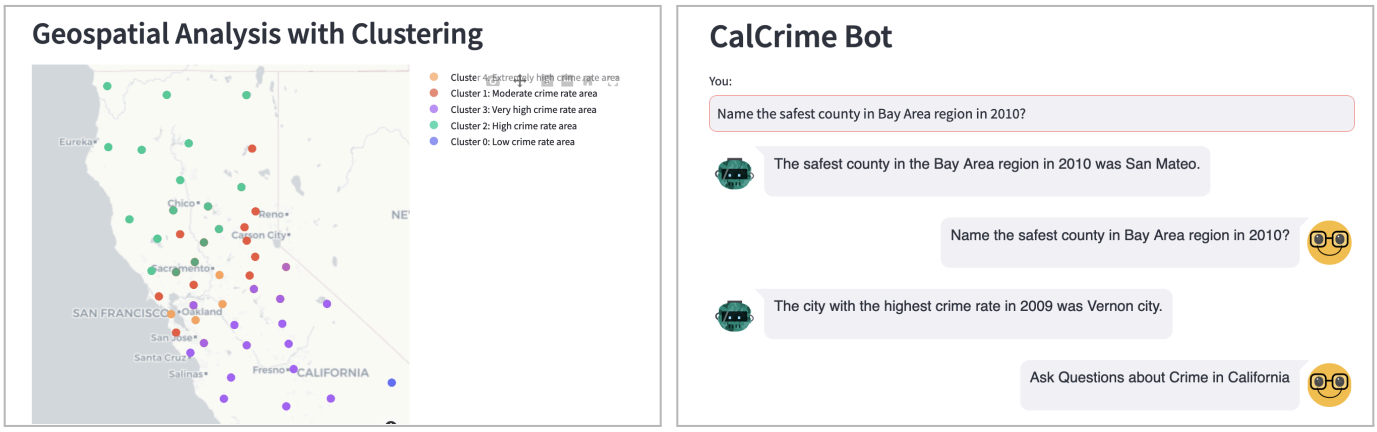

Fig 3. Geospatial Clustering of Crime Density using K-Means and CalCrime Chatbot using Bamboo LLM

require numerical data extraction. For example, if a user asks, "What was the highest crime rate in any county over the last decade?" the chatbot can automatically understand the requirement to search for maximum values within specific features of the dataset—such as 'crime rate' and 'year'. This chatbot dramatically simplifies user interaction with complex datasets by providing immediate answers to data-specific questions without the need for manual data analysis or navigation through complex interfaces. Users can engage with the chatbot through a simple text input interface, making it accessible to individuals with varying levels of technical expertise. This is shown in Fig 3.

IV. METHODOLOGY

Data Collection

The primary source of data for this study was the "Violent Crime Rate" dataset provided by the California Department of Public Health, accessed via, California's Open Data Portal. The dataset includes detailed records of violent crime

rates across various counties from 2000 to 2013, categorized by crime type including robbery, aggravated assault, murder, non-negligent manslaughter, and forcible rape.

Data Preprocessing

The raw data, sourced from the "Violent Crime Rate" dataset provided by the California Department of Public Health, was initially cleaned using Python and its libraries, such as Pandas and NumPy. This cleaning process involved handling missing values, removing outliers, and standardizing data formats to ensure consistency across the dataset. To better utilize this dataset in our visualizations and analysis, we aggregated data entries for each county within the same year to provide a consolidated view of crime statistics per county per year. This aggregation was crucial in managing instances where multiple entries for a single county in one year could skew data interpretation.

Moreover, we calculated the crime rate percentage by utilizing the formula:

$$Crime\ Rate\ = \frac{Crimes\ Reported}{Population\ Reported} * 100$$

This calculation provides a normalized measure of crime, allowing for comparable analysis across counties with varying population sizes.

Additionally, to enhance the dataset's utility for our visualization tools and for consumption by the large language model (LLM), we created contextual summaries of the data. These summaries included key insights and narrative elements around crime trends and demographic impacts, enabling the LLM to generate more informed, contextually accurate responses. This preparation step was instrumental in transforming raw data into a more insightful and narrative-driven resource, facilitating deeper analysis and more engaging visual storytelling in our visualizations.

Visualization Development

The following visualizations were developed using JavaScript and visualization libraries such as D3.js, Plotly, Streamlit, and Pandas AI:

1. **Heat Map:** This visualization shows the spatial distribution of crime rates across California counties. It uses color intensity to represent higher and lower crime areas, providing a quick visual interpretation of geographic crime disparities.
2. **Line Graph:** This tool displays crime trends over time within each county. By selecting a specific county, users can observe how crime rates have evolved, identifying any significant increases, decreases, or stable periods.
3. **Pie Chart:** This visualization breaks down the crime types within each county, illustrating the proportion of each crime type like robbery, assault, and others. It helps in understanding the composition of crimes in different regions.
4. **Geospatial Clustering:** This analysis clusters similar crime rates geographically, allowing users to see how crime rates group together across different parts of the state. This helps in identifying regions with similar safety profiles.
5. **LLM Chatbot:** This was designed to interact with the dataset dynamically. Users can query the

chatbot with specific questions about the data, such as asking for the highest crime rates in a given year or comparisons between counties.

Each visualization was designed with interactivity in mind, allowing users to filter data based on various parameters such as year, county, and crime type, which enhances the user experience and aids in detailed analysis.

Research Questions Addressed

The study aims to answer the following questions:

- How do violent crime rates vary geographically across California?
- What are the trends in violent crime rates over the past decade?
- What types of violent crimes are most prevalent in different counties?
- How do changes in legislation correlate with changes in crime rates?
- How can we address a varying range of user questions on safety and crime, besides the visualizations?

V. EVALUATION PLAN

Accuracy

Data Check: We cross-checked our visualizations, comparing random data points to the visual category of a county in the heatmap and the geospatial graphs.

Error Reporting: On discovering inconsistencies, we tuned the visualization to generate accurate results.

Validation: Each team member thoroughly tested the website, utilizing all interactive and intelligent features to ensure that the visualizations function correctly and display accurate data. This internal testing phase was critical to identify and rectify any potential issues before public deployment.

Usability and Accessibility

The design of the application emphasizes ease of use, ensuring that it is accessible to a broad audience, including those who may not be tech-savvy. Interactive elements, such as dropdown menus for selecting different years and counties, tooltips that provide additional data points, and a responsive chat interface, enable users to navigate the application intuitively. This interactivity not only simplifies the exploration of data but also empowers users by providing them with the tools to find the specific information they need without any technical barriers.

Aesthetic and Design

Aesthetically, the application maintains a clean and professional look, using a consistent color scheme and layout across all visualizations. This uniformity aids in reinforcing user understanding as they switch between different types of data presentations. The use of color, for example, the heatmap uses shades of red to denote severity, aligns with common color associations for warning or danger, enhancing immediate comprehension.

Sustainability and Impact

By making crime data accessible and understandable, the application plays a crucial role in educating the public about safety and crime in their communities or areas they may visit. It enables informed decision-making, whether for

personal safety, real estate investments, or community involvement in crime prevention efforts.

VI. DISCUSSIONS & FUTURE WORK

Discussions

Effectiveness of Visualizations

The heat map, line graph, pie chart, and geospatial clustering visualizations were effective in analyzing violent crime data in California. The heat map displayed geographic patterns and trends, while the line graph depicted how crime rates changed over time in specific counties. The pie chart clearly defined the types of crimes, and geospatial clustering identified areas with similar crime rates. The interactive features improved user engagement and understanding of the data. Users could filter by year, county, and crime type, allowing for a more personalized exploration of crime statistics.

User Feedback

Initial testing indicated high satisfaction with clear and easy-to-use visualizations. Participants appreciated the ability to interact with and analyze crime data dynamically. However, some users suggested improvements such as more detailed hover information and additional filtering options.

Future Work

Future iterations may include more interactive features, such as viewing details on specific crimes and seeing animations of how crime rates evolved.

- Including data on income, education, and employment levels could provide a more comprehensive analysis of the factors influencing crime rates and underlying causes.
- Creating predictive models using machine learning techniques could help predict future crime trends and aid in proactive prevention strategies.
- Including violent crime data from other states may provide comparative insights and improve understanding of national crime patterns.
- Integrating these visualizations into law enforcement systems could help with operational planning, resource allocation, and real-time crime monitoring.
- Conducting more user testing with a variety of groups, such as potential homebuyers, tourists, law enforcement, and policymakers, could improve the visualizations.
- Adding crime data from 2014 to the present would provide a current view of trends and emerging patterns.
- Creating intuitive web and mobile applications would improve accessibility and usability for a broader audience with varying technical backgrounds.
- Improving the LLM Model powering the chatbot would help the users ask more subjective and direct questions and generate custom graphs for data, leading to more user-friendliness, and customizability of the application.

VII. REFERENCES

[1] California Department of Public Health. “Violent Crime Rate.” *California Open Data*, 8 Dec. 2023, data.ca.gov/dataset/violent-crime-rate.

[2] Plotly, “Mapbox County Choropleth,” *plotly.com*, Jul. 03, 2019, https://plotly.com/python/mapbox-county-choroplet/

[3] Pandas-AI, "Pandas AI – Simplify your Data Science Workflow," Pandas-AI. [Online]: https://pandas-ai.com/.

[4] Hugging Face, "pandasai/bamboo-llm," Hugging Face. [Online]. Available: https://huggingface.co/pandasai/bamboo-llm.

[5] OpenAI, "Usage - OpenAI API," OpenAI. [Online]: https://platform.openai.com/usage.

[6] Streamlit, "Chatbot.py," GitHub. [Online]: https://github.com/streamlit/llm-examples/blob/main/Chatbot.py.

[7] USA Facts, "Our Changing Population: California," USAFacts. [Online]: https://usafacts.org/data/topics/people-society/population-and-demographics/our-changing-population/state/california.