GDV

University of Applied Sciences Northwestern Switzerland

Boran Eker

BSc Data Science Student

December 24, 2024

Contents

Introduction

In an era where the digitalization of data grows exponentially, data visualization becomes an essential tool to uncover the underlying narratives within datasets. As Ben Shneiderman aptly stated, "The purpose of visualization is insight, not pictures", reinforcing that visual representations should focus on unveiling insights and actionable information. Visualizing data not only helps simplify complex datasets but also enables a deeper understanding of the context, relationships, and anomalies present within the data.

This report, intended for the general public, delves into the principles of data visualization and demonstrates their application to a dataset related to racism and police shootings in the United States. Through the lens of data, we aim to reveal patterns that tell the unsettling story of racial disparities in fatal police encounters. The data, collected from Kaggle, underwent a thorough cleaning process involving value correction, normalization, and handling of missing information before being prepared for visualization.

With the increase in awareness around racial justice and the incidents involving law enforcement, this dataset provides a timely exploration into how factors like race, mental illness, weapon possession, and other variables contribute to fatal outcomes in police interactions. The following visualizations will explore demographics, regional trends, and the context of these incidents to provide a comprehensive view of the situation in the United States.

The dataset includes information such as the names, ages, genders, and races of the victims, as well as details about the incidents, such as the location, method of killing, weapon used (if any), whether mental illness was a factor, and if the incident was captured on police body cameras. By presenting this data visually, we can uncover critical patterns and insights about the prevalence and nature of police violence across different racial groups.

For the purposes of this report, the data was sourced from Kaggle's public database on police shootings [5], and data preparation was guided by best practices outlined by the European Environment Agency [2], and further cleaned using techniques recommended by experts in data cleansing and visualization [1]. By following these methodologies, we aim to present the data in a clear, accurate, and informative manner, highlighting the importance of effective data preparation and visualization in understanding societal issues.

For further details on the technical implementation, the complete source code of the dashboard is available on my GitHub Repository. This includes the scripts as pdf and the whole code in Python.

LO 1 Visualization Basics and Chart Types

Creating effective data visualizations is essential for conveying insights derived from datasets. Visualizations not only display information but also enhance understanding by tailoring data representation to the audience.[19] Selecting the appropriate type of graph based on the nature of the data and the message to be conveyed is crucial.[15] In this chapter, we explore three key types of visualizations—line charts, maps, and side-by-side bar charts—applied to a dataset of police shootings in the United States. We examine how these charts help us analyze the frequency of shootings by race, trends over time, and the most common weapons involved.

Visualizing Fleeing Types Across Age Groups

When analyzing the relationship between age and police shootings, a **line chart** is an excellent choice for visualizing trends in raw counts.[16] The chart below depicts the number of incidents categorized by fleeing type across different age groups, allowing us to observe how these types vary with age.

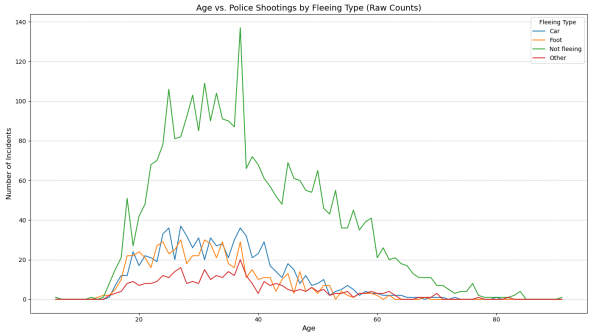

Figure 1.1: Line Chart: Age vs. Police Shootings by Fleeing Type (Raw Counts)

When to Use Line Charts

Line charts are particularly well-suited for visualizing changes or trends over a continuous variable, such as age.[16] In this dataset, age groups are analyzed to see how the frequency of different fleeing types (e.g., car, foot, not fleeing, other) evolves over time. Each line represents a fleeing type, making it easy to compare categories and spot age-related patterns.[19]
For instance, individuals who were "not fleeing" are the most frequent across nearly all age groups, with a noticeable peak in incidents occurring around the 30 to 40-year age range. Other fleeing types, such as "car" and "foot", show much smaller but consistent counts across the age spectrum.

When to Avoid Line Charts

Line charts are not ideal for datasets that do not involve continuous variables, as they rely on the natural progression of data points along an axis (e.g., age).[14] They can also become overwhelming if too many lines are plotted simultaneously, leading to cluttered and difficult-to-interpret visualizations.[8] However, in this case, the manageable number of fleeing types ensures that the chart remains clear and effective.

Visualizing Trends Over Time

To effectively demonstrate the fluctuation in police shootings over the years, we incorporate a visual representation directly into our analysis. The map below shows the distribution of police shootings across different states, offering a geographical perspective on the data.[8] This enhances our understanding of regional trends and anomalies in the frequency of these incidents.

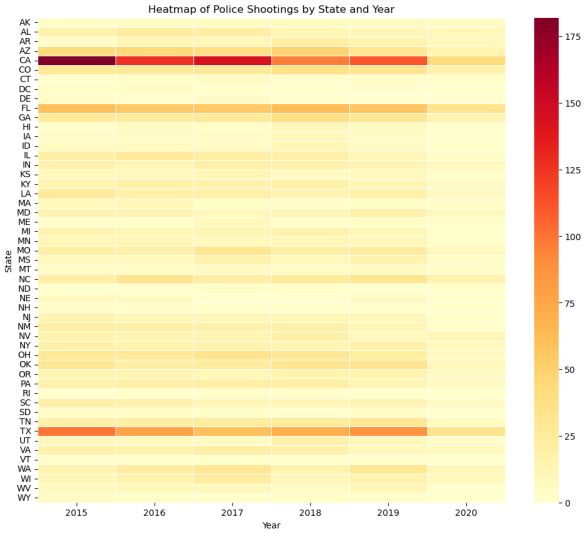

Figure 1.2: Map of Police Shootings by State in the US

When to Use Heatmaps

Heatmaps are particularly effective for visualizing dense data across two dimensions, such as time and geography. They allow for quick identification of hotspots and trends by using color gradients to

represent the intensity of data values.[8] In the context of our analysis, the heatmap highlights variations in police shootings across different states and years, making it easier to detect regional and temporal patterns.

When to Avoid Heatmaps

Although heatmaps are useful for identifying trends and clusters, they may not be the best choice when precise numerical values are required or when the data spans a very large range of values. Additionally, the effectiveness of a heatmap depends heavily on the choice of color scale and resolution. Poorly chosen colors or overly dense data points can obscure patterns and lead to misinterpretation.[14]

Visualizing Race-Based Data in Police Shootings

When analyzing race-based data, it's important to provide a clear visualization of absolute counts to highlight trends and disparities.[15] The chart presented here uses a simple bar chart to show the number of police shootings across different racial groups, offering an accessible overview of the data.

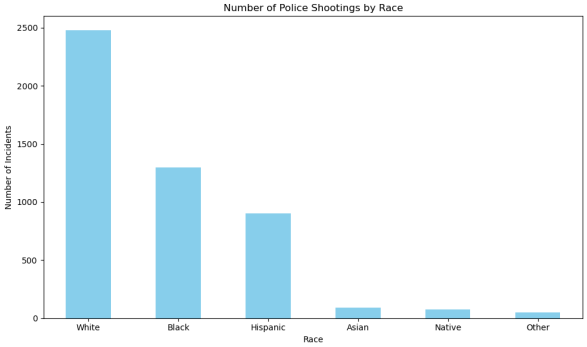

Figure 1.3: Bar Chart: Number of Police Shootings by Race

Advantages of Bar Charts

Bar charts are particularly effective for displaying absolute counts in a way that is both intuitive and visually striking. The bar chart in this analysis provides: - A straightforward visual comparison of police shootings across racial groups, emphasizing the disparities in total incidents. - An approachable format for audiences unfamiliar with complex data visualizations.[15]

Insights from the Chart

The bar chart highlights notable disparities in police shootings among racial groups. For instance, while some groups show a lower absolute number of incidents, others stand out with significantly higher counts, suggesting areas where further investigation and policy action may be warranted. By focusing on absolute numbers, the chart establishes a foundation for understanding broader trends and patterns.

When to Use Bar Charts

Bar charts are best suited for visualizing datasets where absolute counts provide meaningful insights. This type of visualization is particularly effective for: - Presenting an overview of categorical distributions. - Engaging non-technical audiences by simplifying complex datasets into accessible formats.

When to Avoid Bar Charts

While bar charts are excellent for visualizing counts, they may oversimplify analyses that require normalization or proportional representation. For example, accounting for population size differences across racial groups would require complementary visualizations, such as side-by-side bar charts or scatter plots.[14]

Why These Charts Were Chosen

After numerous attempts and careful consideration of different chart options, I ultimately selected these three graphs. Not only do they best align with the characteristics and insights of my dataset, but they also stand out as the most engaging and visually compelling representations.

LO 2 Visual Perception

Basic Heatmap vs. Annotated Heatmap

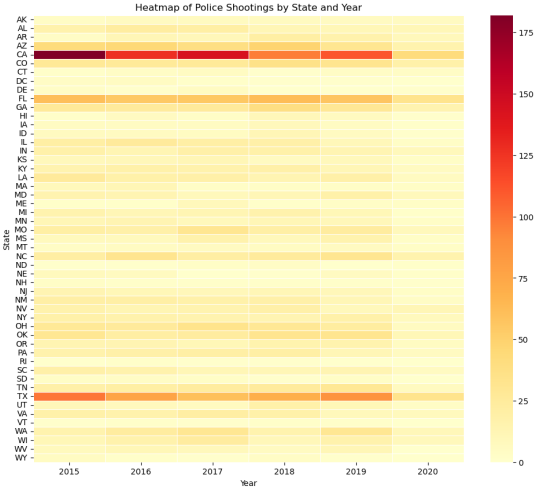

Figure 2.1: Heatmap of police shootings

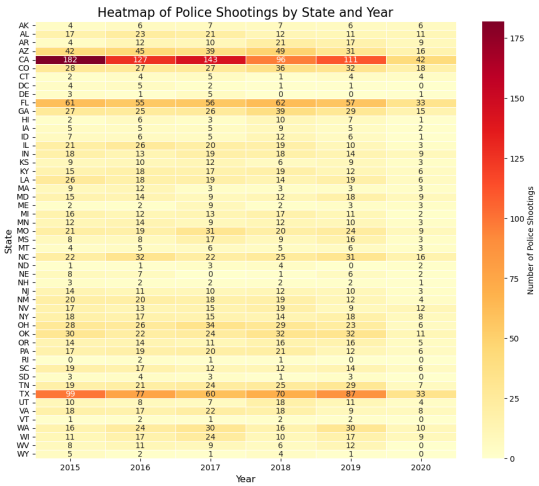

Figure 2.2: Heatmap of police shotings with numbers

The chart on the left is the **Basic Heatmap**, which visualizes the frequency of police shootings by state and year using a gradient color scheme. While it effectively highlights general trends through color intensity, it does not provide specific numerical values for each cell, which can make precise analysis more challenging. The chart is useful for identifying broad patterns but lacks finer detail.[4]

The chart on the right is the **Annotated Heatmap**, which enhances the basic version by adding numerical annotations to each cell, indicating the exact count of incidents. Additionally, the color bar includes a title to clarify the scale, making the chart more informative and user-friendly. This combination of color intensity and precise numbers allows for both quick visual analysis and detailed examination of state-wise and year-wise patterns.[24]

While the basic heatmap provides a clean, high-level overview, the annotated heatmap is more suitable for in-depth analysis due to its clarity and additional detail. Together, these charts demonstrate how small enhancements—like annotations and labeled scales—can significantly improve the interpretability and utility of heatmaps.[9]

Stacked Bar Chart vs. Line Chart

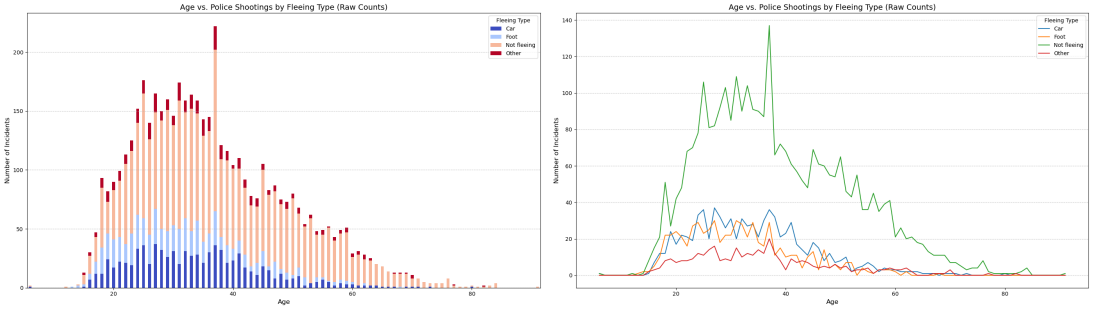

Figure 2.3: Comparison of Stacked Bar Chart and Line Chart

The chart on the left is the **Stacked Bar Chart**, which displays the distribution of police shootings by fleeing type across different age groups. Each bar represents a specific age group, segmented into different fleeing types. While this chart allows for a breakdown of contributions within each group, it becomes challenging to discern trends or comparisons across age groups due to the stacking and overlap of categories.[23]

The chart on the right is the **Line Chart**, which represents the same data but plots the raw counts of fleeing types against age groups using separate lines for each category. This format makes it significantly easier to track trends and patterns over the age continuum, as each fleeing type is clearly represented by a distinct line.[34]

The stacked bar chart is better suited for visualizing the proportions of fleeing types within individual age groups. However, the line chart excels in providing a broader overview of trends and variations across the age spectrum. It is particularly effective for identifying peaks, troughs, and overall patterns, making it the preferred choice for detailed and comparative analyses.[33]

Line Chart vs. Colorblind-Friendly Line Chart

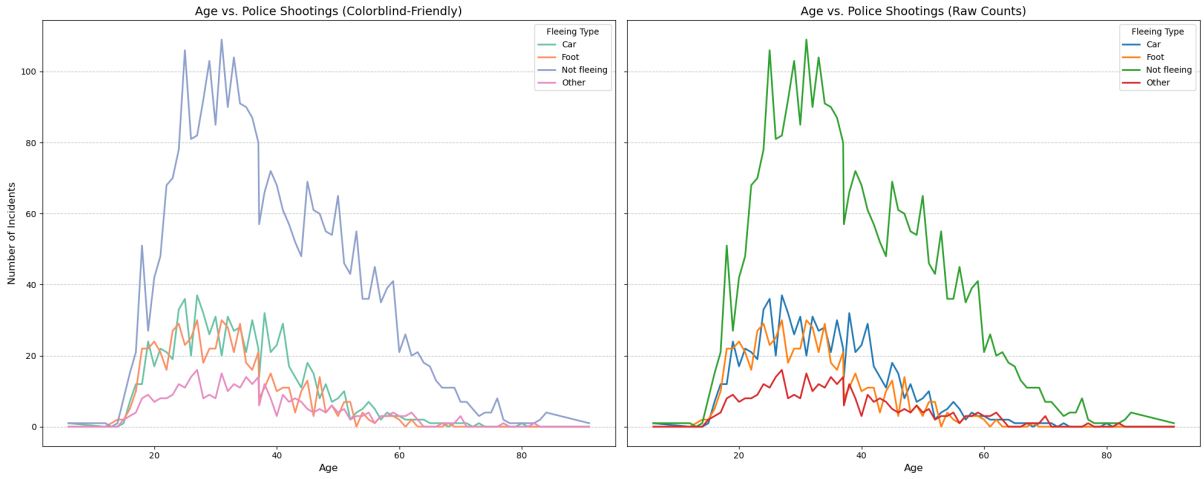

Figure 2.4: Colorblind vs. Raw

The chart on the right presents the **Raw Line Chart**, which depicts the number of police shootings across different ages and fleeing types using default colors. This chart effectively highlights the trends and variations in incidents by age but may pose challenges for individuals with color vision deficiencies due to the lack of an accessible color palette.[11]

The chart on the left showcases the **Colorblind-Friendly Line Chart**, which enhances the original visualization by using a distinct, accessible color palette specifically designed to accommodate colorblind viewers. This improvement ensures that the patterns for different fleeing types remain clear and distinguishable for a wider audience, without compromising the overall interpretability of the data.[3]

While the raw line chart provides a straightforward depiction of trends, the colorblind-friendly version prioritizes accessibility and inclusivity, making the data easier to analyze for individuals with visual impairments. This refinement highlights how small design choices, like color selection, can significantly improve the utility and reach of visualizations.[32]

LO 3 Design Principles vs. Data

In this chapter, we examine the relationship between data preprocessing and design decisions in visualizing racial disparities in police shootings. The goal is to create clear, accessible bar charts that effectively compare both absolute incident counts and population-normalized rates across different racial groups.

Data Preparation

The first step in the visualization process involves careful preparation of the data to ensure it aligns with the goals of the analysis. The dataset provides details of police shootings, including the race of individuals involved. To visualize the data meaningfully, the following steps were taken:

- **Calculating incident frequency:** The data was grouped by racial categories, and the total number of incidents for each group was computed. This serves as the basis for understanding the absolute distribution of shootings.
- **Normalizing for population size:** To account for differences in population size across racial groups, incident counts were normalized using population estimates (per million people). For example, the White population was approximated at 197.5 million, while the Black population was estimated at 42.0 million. By dividing the incident counts by population size, normalized rates were calculated, providing a fairer comparison.[25]
- **Structuring the data:** The data was reorganized to include both absolute incident counts and normalized rates, allowing for a dual analysis that highlights both the raw numbers and adjusted rates.

Selecting and Tailoring Important Features

To ensure the visualizations remain focused and interpretable, the data was filtered and tailored for clarity:

- **Filtering columns:** Non-essential attributes were removed to streamline the analysis. Only race, incident counts, and population-normalized rates were retained to highlight the core message.[29]
- **Population mapping:** Racial categories were matched with corresponding population estimates, ensuring accuracy in the normalization process. Races such as *White*, *Black*, *Hispanic*, *Asian*, *Native*, and *Other* were included.[20]
- **Scaling normalized rates:** To visually compare absolute numbers and normalized rates on the same chart, the normalized values were scaled appropriately. This ensures the two metrics can be juxtaposed without distortion.[27]

Following Design Principles

Bar Chart Design and Visual Hierarchy

The visualization consists of a side-by-side bar chart that presents both absolute incident counts and normalized rates for each racial group. This design choice allows viewers to identify disparities both in raw numbers and population-adjusted metrics.

To enhance clarity, the following design principles were applied:

- **Color choice:** Distinct colors were used for the two metrics: blue for absolute incident counts and orange for normalized rates. This ensures clear differentiation while maintaining accessibility.[22]
- **Bar arrangement:** The bars are positioned side by side for each racial group, providing a direct comparison between the absolute and normalized values.[13]

- **Labels and axes:** The x-axis labels the racial categories, while the y-axis represents the count or scaled rate of incidents. Titles and axis labels were chosen to provide context and improve interpretability.[7]

Scaling and Normalization

The use of normalization is critical to understanding racial disparities in police shootings. While absolute counts provide valuable insights, they can be misleading without considering population size. For instance, a racial group with a smaller population may appear to have fewer incidents in absolute terms, but normalization reveals higher rates per capita.

To address this, incident counts were divided by population size (in millions), producing a rate per million people. This approach highlights disparities that are not immediately visible in raw counts. The normalized rates were then scaled for visualization, ensuring they align visually with the absolute numbers while preserving their relative magnitudes.[17]

Accessibility and Colorblind-Friendly Design

The chosen colors (blue and orange) adhere to accessibility standards, ensuring the chart remains interpretable for viewers with color vision deficiencies. The use of clear labels and a well-structured legend further enhances readability, aligning with best practices in inclusive design.[10]

Conclusion

Through careful data preparation and thoughtful visualization design, this analysis highlights racial disparities in police shootings. By presenting both absolute counts and population-normalized rates, the visualizations offer a balanced and comprehensive perspective. This dual approach ensures that the data is communicated accurately, clearly, and inclusively, enabling a deeper understanding of the underlying patterns.[21]

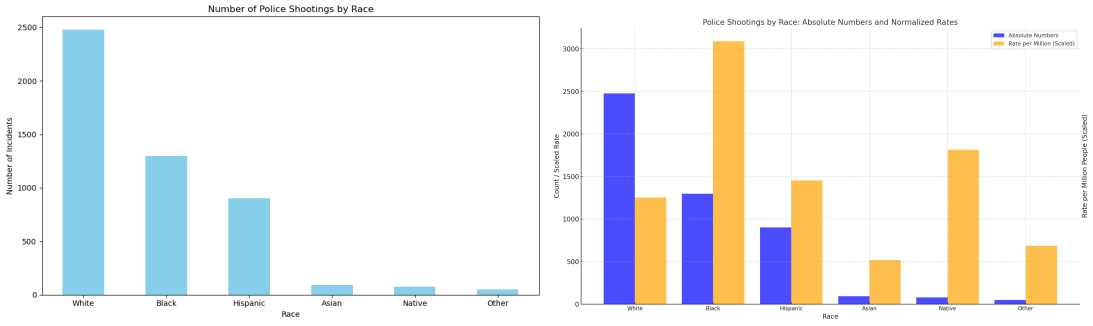

Figure 3.1: Comparison of Stacked Bar Chart and Line Chart

LO 4 The Grammar of Graphics

The Grammar of Graphics is a comprehensive framework for constructing statistical visualizations, systematically mapping data into visual elements. It structures the visualization process into a series of well-defined steps, enabling flexibility and clarity in designing graphics. This approach has become a cornerstone of modern data visualization, supporting the creation of complex, multi-layered visual representations.

Leland Wilkinson introduced this concept in 1999 in his book *The Grammar of Graphics*, which laid the foundation for many popular data visualization tools, such as `ggplot2` for R and `Altair` for Python. These tools utilize the principles of the Grammar of Graphics to provide a structured way to transform data into meaningful visuals. This framework has driven significant advancements in visualizing multidimensional data and large datasets, addressing the challenges of both complexity and scalability in data analysis.

Data

The starting point for any visualization is the dataset, which consists of variables that can either be continuous, such as income or age, or categorical, such as gender or region. Continuous variables represent numerical values within a range, while categorical variables define distinct groups or categories. A well-constructed dataset often includes both types, offering a wealth of opportunities for visualization. Preparing data effectively, as emphasized by the European Environment Agency, is essential to ensure accuracy and reliability in visualizations.[12]

Aesthetics

Aesthetic attributes play a crucial role in visualizations as they determine the visual properties of a graphic that are perceived by the viewer. These attributes include the position of elements on the x- and y-axes, the color used to differentiate groups, the size of points to reflect quantities such as population, the transparency to reduce clutter in dense datasets, and the shape of elements to denote different categories. For example, in a scatter plot, income might be represented on the x-axis and age on the y-axis, while population size could determine the size of the points, and professional categories might be distinguished by color. Accessibility is a vital consideration when choosing colors, as it ensures that the graphic is interpretable by a broader audience, including those with color vision deficiencies. Tools such as the `viridis` palette, available in Python and R, can help create colorblind-friendly visualizations.[30]

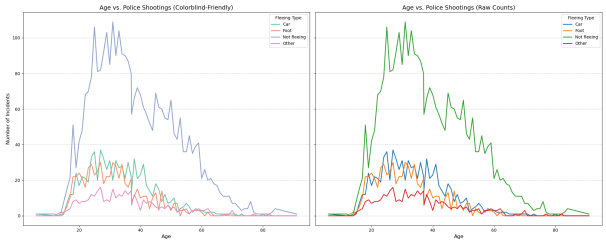

Figure 4.1: Colorblind vs. Raw

Geometric Objects

Geometric objects, often referred to as "geoms," are the elements that visually represent data points in a graphic. These objects include points for scatter plots, lines for time series data, and bars for bar charts. Depending on the dataset and visualization goal, different geoms are used to effectively communicate the desired information. For instance, a line can represent trends over time, such as changes in average income for different age groups, while a point plot might highlight individual data entries in a dataset.[31]

Statistical Transformations

Statistical transformations are integral to the visualization process, as they involve processing raw data to derive summarized insights. These transformations may include calculating averages, medians, or other aggregate statistics, as well as estimating densities or confidence intervals. For example, in a demographic dataset, it could be insightful to compute the average income and age for specific professional groups across several years to observe trends or patterns.[26]

Scales

Scales are an important aspect of any visualization as they define how data values are mapped to visual dimensions. These scales can be linear for evenly distributed variables, logarithmic to handle skewed data, or categorical for discrete groupings. For instance, a logarithmic scale might be applied to income data to better represent a wide range of values while maintaining clarity and interpretability.[28]

Coordinate Systems

The choice of coordinate systems defines the framework for representing data visually. Commonly used systems include the Cartesian coordinate system for x-y plots, the polar coordinate system for circular visualizations such as pie charts, and geographical coordinate systems for mapping spatial data. Selecting the appropriate coordinate system ensures that the visualization effectively conveys the relationships and patterns within the data.[6]

Facets

Faceting is a powerful technique that involves splitting a dataset into subsets and creating multiple plots, one for each subset. This approach is particularly useful for comparing groups or categories within the data. For instance, a faceted visualization could display income distributions for different professional groups or analyze regional variations in demographic data. By using consistent scales and aesthetic attributes across the facets, these plots maintain comparability while providing detailed insights.[18]

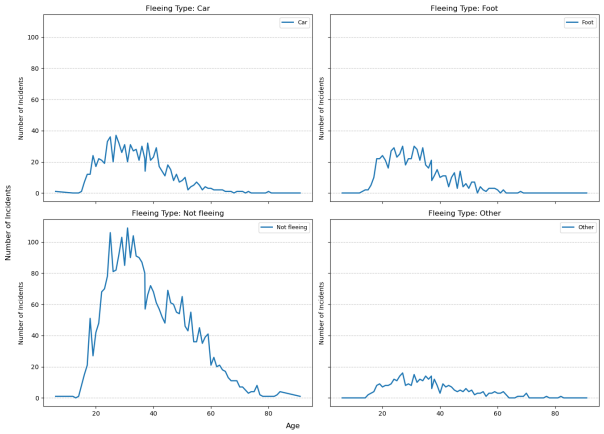

Figure 4.2: facets

In summary, the Grammar of Graphics provides a robust and structured approach to creating data visualizations. By considering aspects such as accessibility, effective scaling, the use of appropriate geometric objects, and the choice of coordinate systems, it is possible to create visuals that are not only clear and accurate but also inclusive and accessible to diverse audiences.

LO 5 Evaluation

Analyzing police shootings is crucial for understanding patterns and trends that may inform public policy, law enforcement training, and societal awareness. Poorly communicated analyses of such data can lead to misconceptions or oversights, while comprehensive and clear evaluations can drive meaningful changes and improved outcomes. This chapter evaluates visualizations derived from the dataset to uncover critical patterns related to race, armed status, mental illness indications, and other contextual variables.

Defining the Analysis Environment

The dataset comprises incidents of police shootings, including details such as the victim's race, armed status, mental health signs, and whether they were fleeing. To ensure a balanced evaluation, two visualizations were designed:

1. A heatmap showing the frequency of shootings by race and armed status.
2. A bar chart illustrating the number of incidents involving individuals with signs of mental illness, segmented by state.

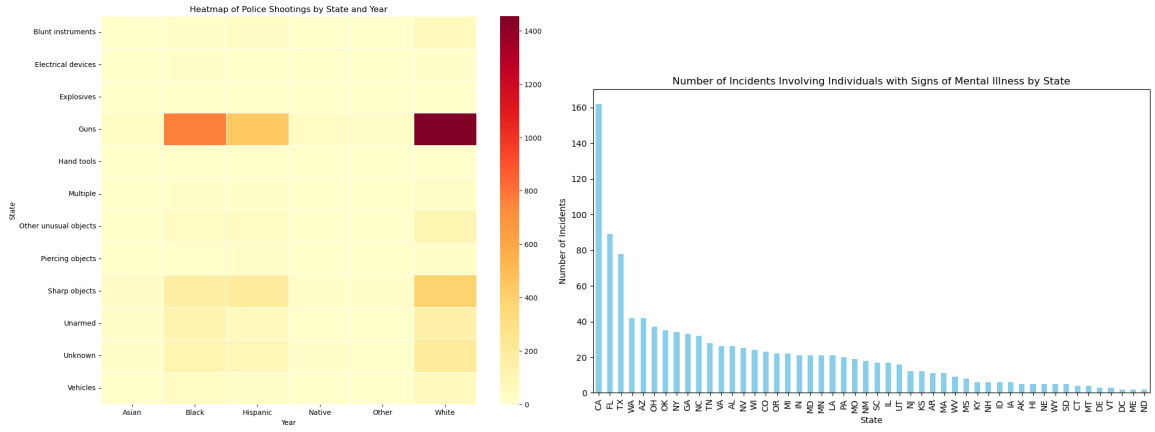

Figure 5.1: Comparison of Stacked Bar Chart and Line Chart

These visualizations were shared with five individuals from diverse backgrounds, including data analysts, social workers, and community advocates, to assess their clarity and effectiveness.

Evaluation Questions

Participants were asked the following questions during a usability test:

#	Question
1	Which race appears to be most frequently involved in police shootings, according to the heatmap?
2	What is the most common armed status of individuals in the dataset?
3	Do you notice any patterns related to mental illness signs across states?
4	Are there elements in the visualizations that make interpreting the data challenging?
5	What insights can you derive about individuals who were fleeing versus those who were not?

Table 5.1: Evaluation Questions for Visualizations

Analyzing the Results

Question 1

The heatmap consistently highlights that individuals identified as White and Black are most frequently involved in police shootings. This trend was clear to all participants, validating the visualization's effectiveness in illustrating racial disparities. However, one participant noted that the categories for "Other" races were not distinctly represented, potentially obscuring nuances in less common cases.

Question 2

Participants observed that the majority of incidents involved individuals identified as armed with a firearm. This observation aligns with the dataset's emphasis on armed encounters, as reflected in the visualization. However, two participants suggested that including a breakdown of "other" weapons could add depth to the analysis.

Question 3

The bar chart effectively highlighted that states like California and Texas have a high number of incidents involving individuals with signs of mental illness. While participants appreciated the clarity of this trend, some noted that the chart could benefit from a normalization approach (e.g., per capita rates) to contextualize state-level differences.

Question 4

Three participants found the heatmap's color scheme somewhat challenging, particularly when differentiating between similar intensity levels. Additionally, overlapping data points in the bar chart were flagged as a minor issue when attempting to identify specific state counts.

Question 5

The visualizations revealed that individuals not fleeing were more likely to be armed, while those fleeing often fell into unarmed or less-lethal categories. Four out of five participants successfully identified this pattern, though one participant expressed difficulty interpreting the fleeing status due to its presentation style.

Summary

The evaluated visualizations effectively conveyed critical patterns, such as racial disparities and the prevalence of armed encounters. However, areas for improvement include better representation of less common categories and normalization for state comparisons. Overall, the visualizations were successful in facilitating insights into police shootings, laying the groundwork for informed discussions and potential policy actions.

Appendix: Visualizations

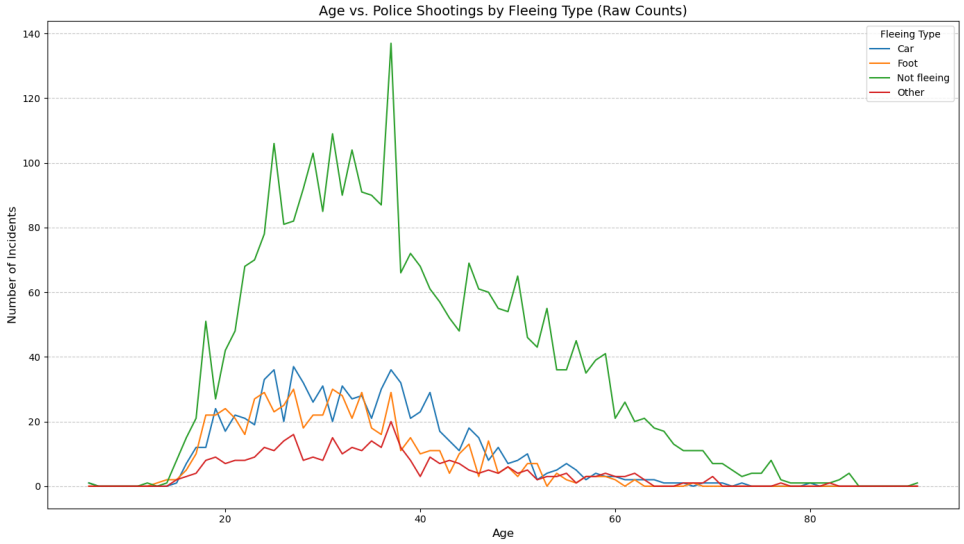

Figure 5.2: line chart about police shootings by fleeing type

Figure 5.3: Normalized Bar Chart: Number of Police Shootings by Race

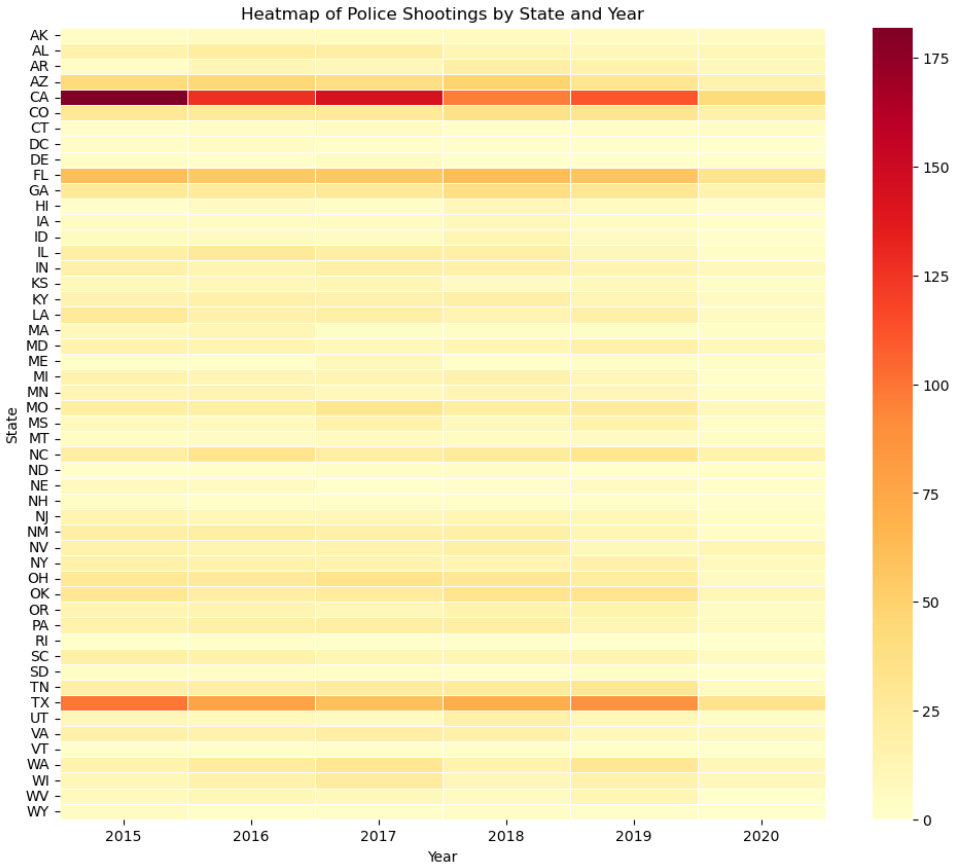

Figure 5.4: State-Year Heatmap

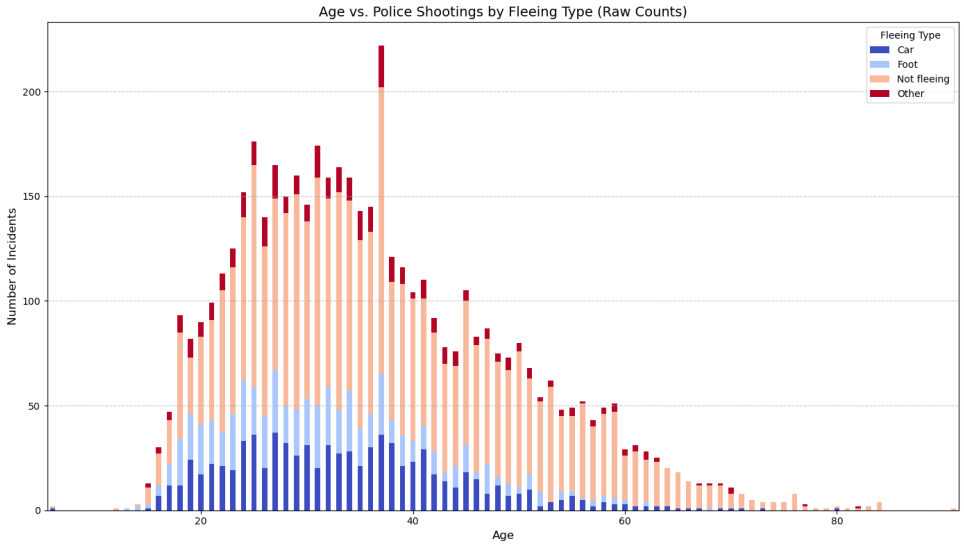

Figure 5.5: Line chart showing trends in fatal police interactions over time.

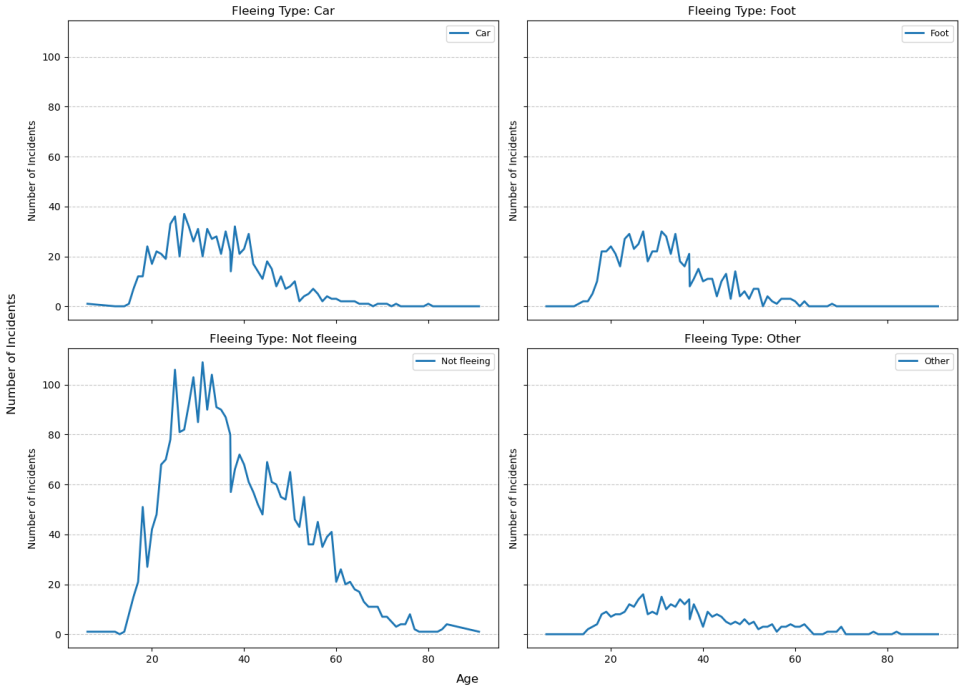

Figure 5.6: Facets of police shootings by fleeing type

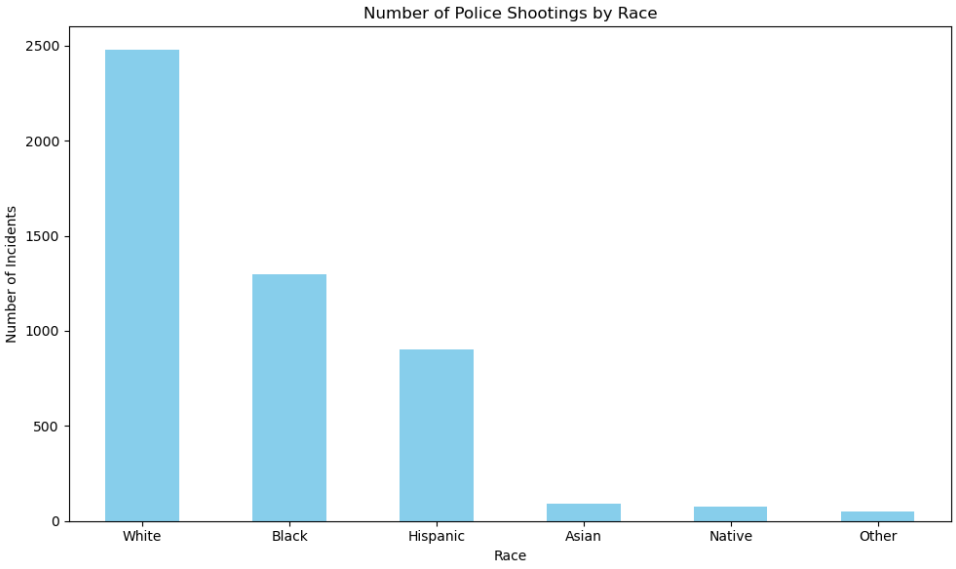

Figure 5.7: Bar Chart: Number of Police Shootings by Race

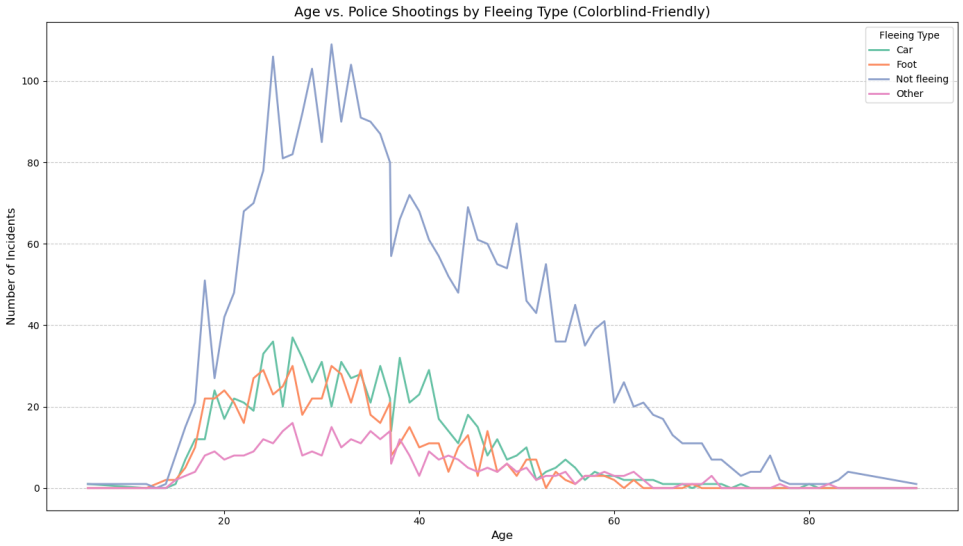

Figure 5.8: Colorblind line chart about police shootings by fleeing type

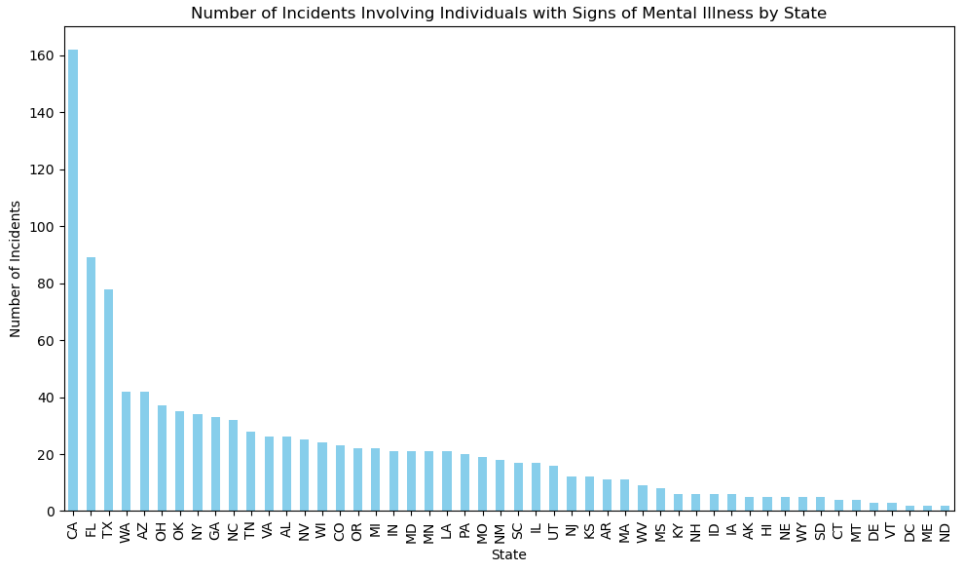

Figure 5.9: Bar chart about number of incidents by state

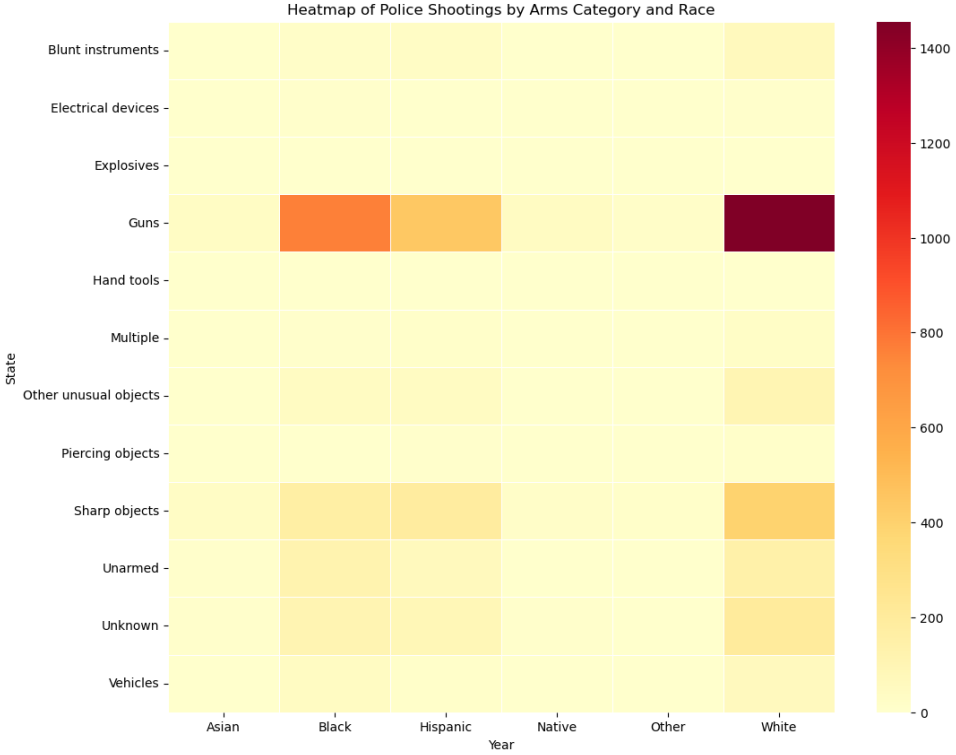

Figure 5.10: Heatmap of Police Shootings by Arms Category and Race

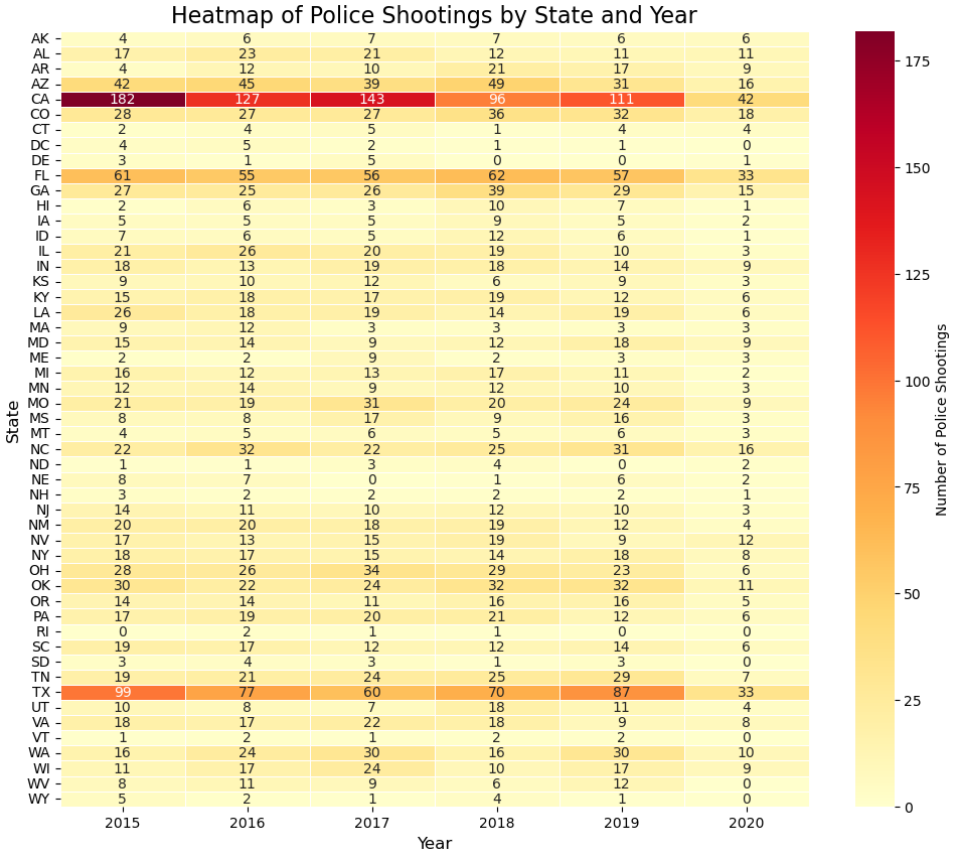

Figure 5.11: Heatmap of police shotings with numbers

Bibliography

[1] European Environment Agency. Chart dos and don'ts: Best practices for data visualization.

[2] European Environment Agency. Preparing data for analysis and visualization.

[3] AppyPie. Color blind-friendly palettes: A guide to enhancing chart accessibility, 2024.

[4] Atlassian. Heatmaps: The complete guide, 2024.

[5] Ahsen W. Chaudhry. Us police shootings.

[6] Claus O. Wilke. Coordinate systems and axes in data visualization, 2024.

[7] Cornell Engineering. Bar charts: Design principles and best practices, 2024.

[8] DataCamp. 11 data visualization techniques you should know.

[9] DataCamp. Seaborn heatmap tutorial: Creating beautiful heatmaps in python, 2024.

[10] Datawrapper. How to choose colors for data visualizations (part 2: Colorblindness), 2024.

[11] EM360 Tech. Accessibility guide: Using colors in data visualization, 2024.

[12] European Environment Agency. Prepare data: Ensuring accuracy and reliability in visualizations, 2024.

[13] FasterCapital. Chart design: The art of clustered bar charts in excel, 2024.

[14] GeeksforGeeks. 5 best practices for effective and good data visualizations.

[15] GeeksforGeeks. Charts and graphs for data visualization.

[16] GeeksforGeeks. Types of data visualization.

[17] GeeksforGeeks. Normalization and scaling: What is the difference?, 2024.

[18] Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen. Facets in ggplot2, 2024.

[19] Tony Hammond. 9 best practices and tips for effective data visualization.

[20] Hands-On DataViz. Normalize: Make comparisons fairer, 2024.

[21] Harvard University IT Accessibility. Data visualization accessibility: Charts and graphs, 2024.

[22] Infogram. Do this, not that: Bar charts, 2024.

[23] InsightSoftware. Comparing data visualizations: Bar vs. stacked, icons vs. shapes, and line vs. area, 2024.

[24] Matplotlib. Annotated heatmap in matplotlib, 2024.

[25] Neilsberg Insights. United states population by race: A comprehensive analysis, 2024.

[26] Number Around Us. Revealing hidden patterns: Statistical transformations in ggplot2, 2024.

[27] Power BI Gate. Power bi filters: Unlocking the full potential of your data visualizations, 2024.

[28] Statistics Easily. What is axis scale: Understanding data visualization, 2024.

[29] Storytelling with Data. Strip away the nonessential: Simplifying your visualizations, 2019.

[30] The R Project for Statistical Computing. Introduction to the viridis color palettes, 2024.

[31] The Tidyverse Team. ggplot2 reference, 2024.

[32] Venngage. How to use color blind friendly palettes in your design, 2024.

[33] Visual Design Network. How to choose the right chart for your data, 2024.

[34] wpDataTables. Line chart vs bar chart: Key differences and use cases, 2024.