DSC 465 Data Visualization
Final Project Report

Police Killings

Group-Data Vizards

Ankita Mishra
Jonathan Lindahl
Kavana Manvi Krishnamurthy
Sanchal Sunil Dhurve

Table of Contents

Introduction

The 2014 killing of Michael Brown in Ferguson, Missouri, sparked a movement for police accountability, leading to the creation of Black Lives Matter. In response, since January 1, 2015, The Washington Post has maintained a detailed database of every fatal police shooting in the U.S., covering critical aspects such as race, age, gender, and whether the individual was armed or experiencing a mental-health crisis. This comprehensive effort addresses the previous lack of reliable data on police killings.

This project aims to analyze how police killings disproportionately affect certain racial groups. By integrating The Washington Post's data with U.S. Census information on poverty rates, high school graduation rates, median household incomes, and racial demographics, we seek to uncover patterns and correlations. This analysis will inform policy changes and improve police practices, striving for justice and equity in law enforcement.

Most of our datasets were sourced from Kaggle, which provides detailed data on fatal encounters involving police officers in the United States. We meticulously merged these datasets based on the State column, ensuring consistency and comparability across various data points. To further enrich our analysis, we calculated the share of each race within the population and the number of killings relative to the total population in each state. This approach allowed us to identify and quantify the disproportionate impact of police killings on different racial groups, shedding light on systemic disparities and informing our broader analysis and recommendations. By integrating this data with additional socioeconomic factors, we aim to provide a nuanced and comprehensive understanding of the issues at hand.

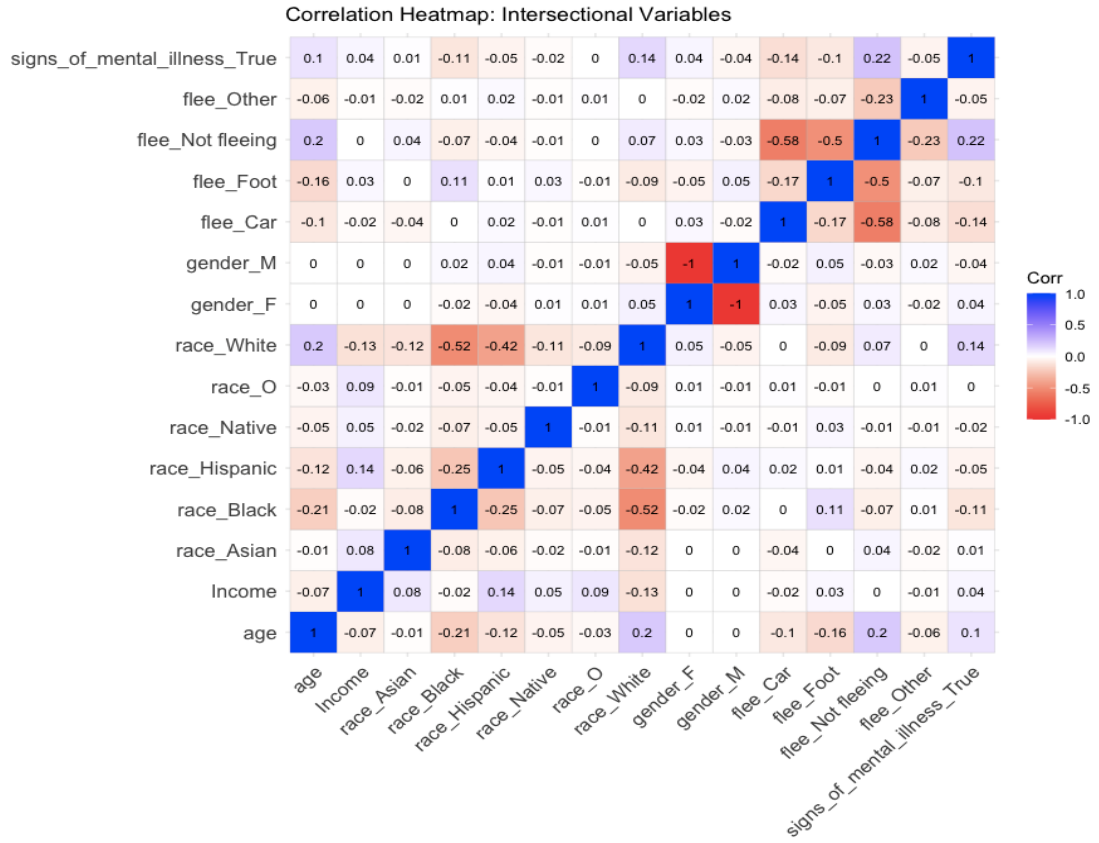
Overview of Variables

In our analysis, the key variables included armed status, age, gender, race, state, city, threat level, flee status, and several demographic and socioeconomic factors such as total population, racial demographics (Black, White, Hispanic, Native, Asian), income, and income per capita. These variables were crucial in understanding the context and impact of police killings across different communities.

Variable	Description	Data Type
Armed	Indicates whether the individual was armed and, if so, the type of weapon	Categorical
Age	The age of the individual at the time of the incident.	Numerical
Gender	The gender of the individual, represented as "M" for male or "F" for female.	Categorical
Race	The racial identity of the individual, e.g., "White", "Black", "Hispanic", "Asian", "Native"	Categorical
City	The city where the incident took place	Geographical
State	The state where the incident took place	Geographical
TotalPop	The total population of the area (likely the city or county) where the incident occurred.	Numerical
White	Percentage of population identifying as White	Numerical
Black	Percentage of population identifying as Black	Numerical
Hispanic	Percentage of population identifying as Hispanic	Numerical
Asian	Percentage of population identifying as Asian-Pacific Islanders	Numerical
Native	Percentage of population identifying as Native Americans	Numerical
Threat_level	The perceived threat level associated with the individual, e.g., "attack", "other."	Categorical
Income	The median household income of the area	Numerical
Income_per_cap	The per capita income of the area	Numerical

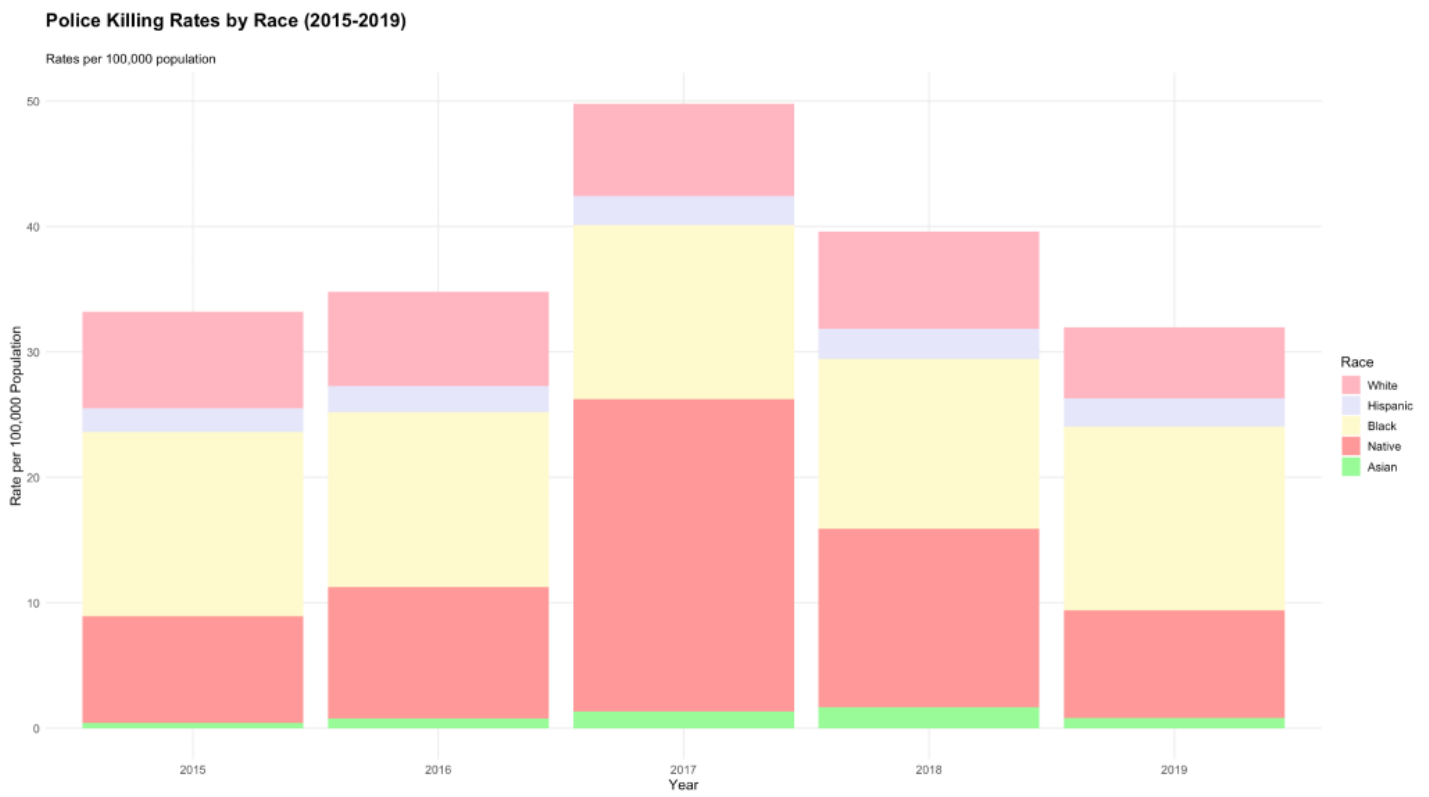
Additionally, our dataset contains a wealth of other variables that we did not use in this analysis but are available for future exploration. These include socioeconomic and demographic variables, income and employment variables, transportation and commute variables, and employment and economic variables. These additional data points can provide further insights and opportunities for in-depth analysis in subsequent studies.

Exploratory Analysis



A correlation heatmap was created in R using a dataset on police killings, including both numerical and categorical variables like race, gender, age, income, fleeing status, and mental illness. Categorical variables were converted to dummy variables, while numerical variables were retained. Income was divided into percentiles for better segmentation. The correlation matrix was computed from the transformed data, with missing values replaced by zeros. The heatmap, visualized with a diverging color scheme, highlighted the relationships between variables. Blue represents perfect positive correlation of 1. While red represents perfect negative correlation of -1.

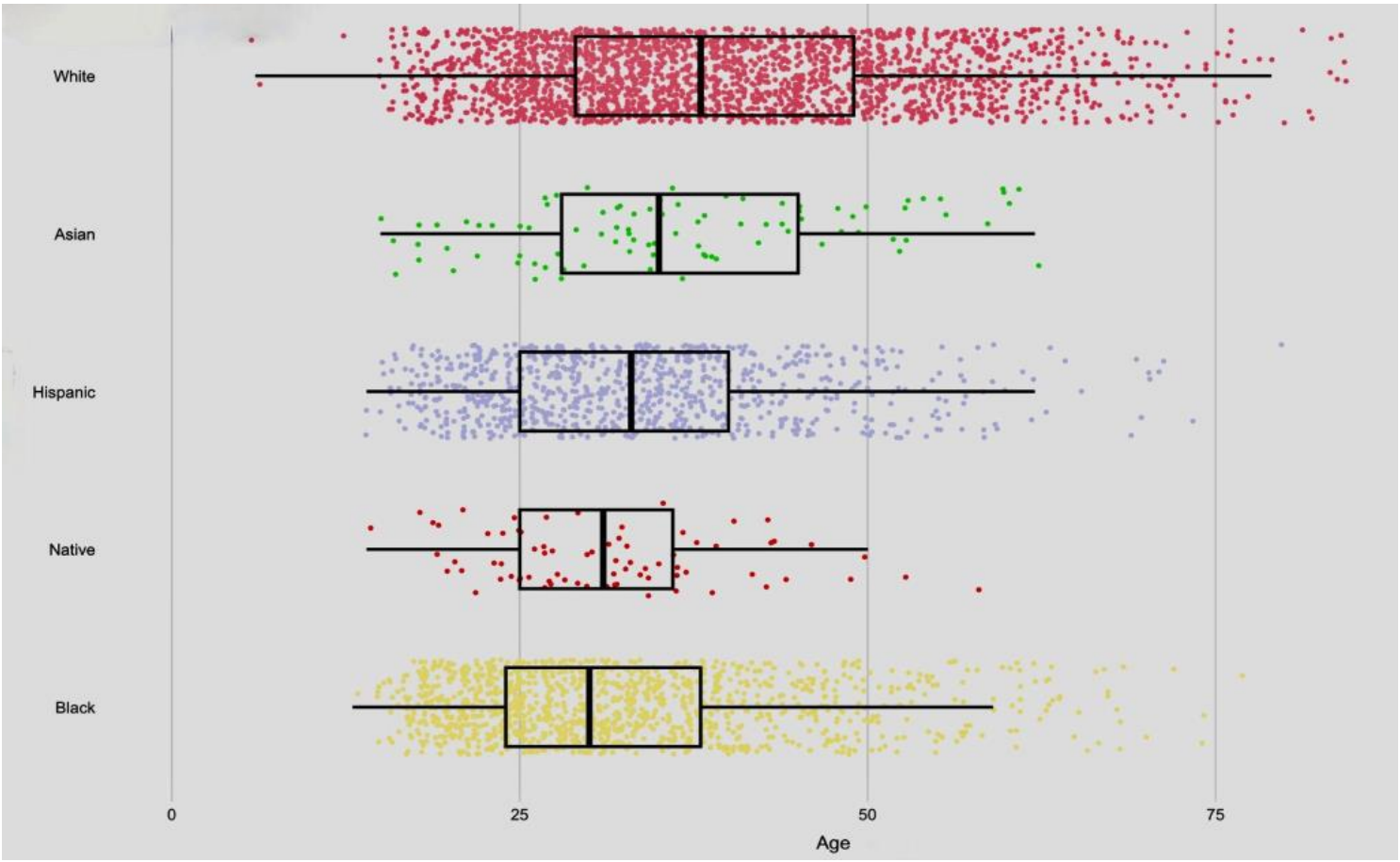
Key observations include a negative correlation between age and being Black (-0.21), suggesting younger Black victims, and a positive correlation between age and being White (0.20), indicating older White victims. Income showed a slight positive correlation with being Hispanic (0.14) and a negative correlation with being White (-0.13). Gender also displayed weak correlations with fleeing behavior, with females more likely to flee in a car (0.03) and less likely to flee on foot (-0.05). Overall, the heatmap revealed complex intersections of race, gender, income, and fleeing status in police killings, though correlations should not be mistaken for causation.



For our second exploratory visualization, we analyzed police killing rates from 2015 to 2019 using a stacked bar chart. Initially, we included data from 2020, but upon reviewing the dataset, we realized the 2020 data was incomplete due to a cutoff in data collection. To avoid misrepresenting the trends, we decided to exclude 2020 from the analysis.

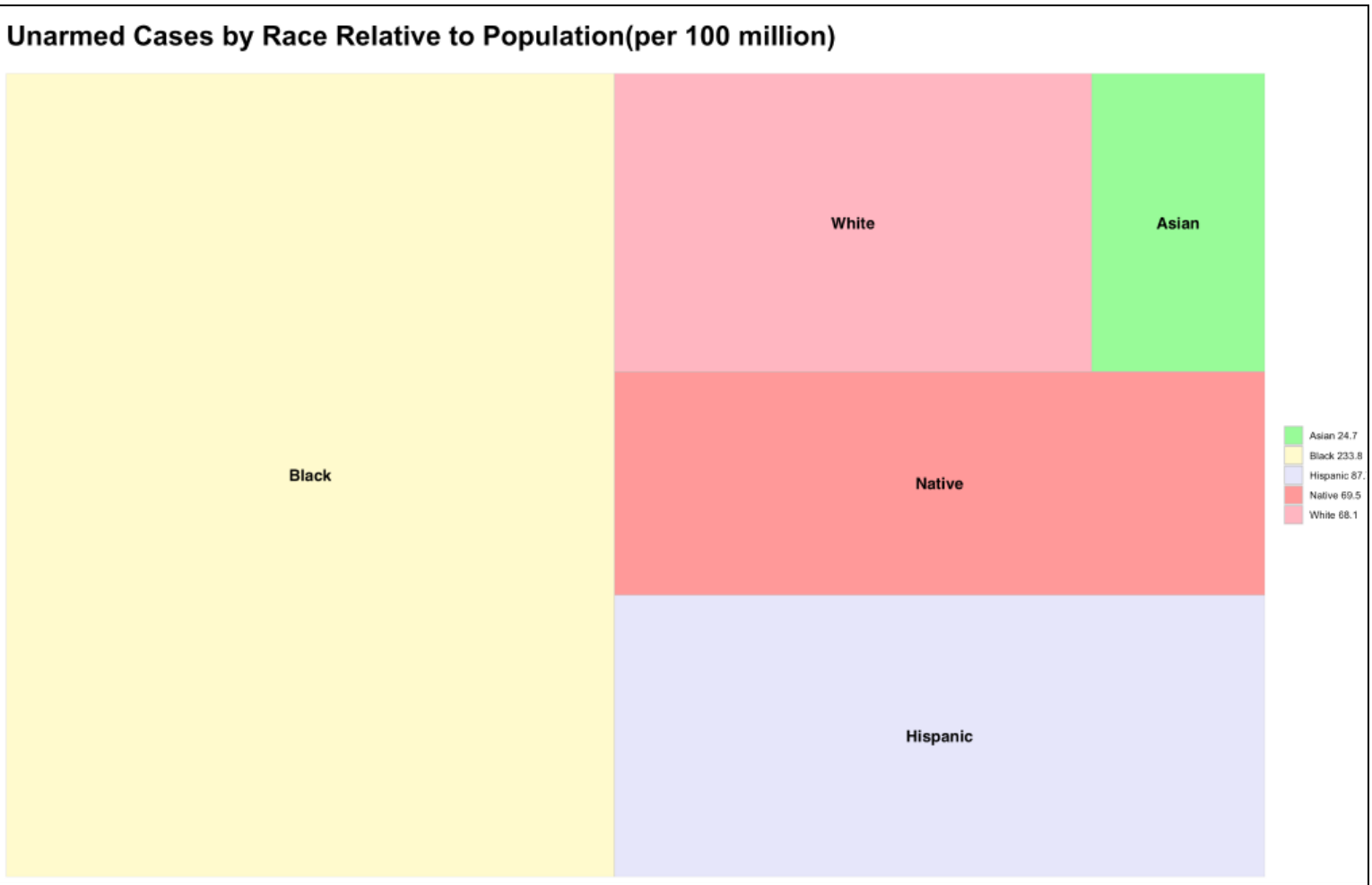
We began by visualizing the total number of police killings. However, we quickly observed that the large white population in the U.S. dominated the total cases, which overshadowed the rates for other racial groups. This made it difficult to compare the impact across different populations. To address this, we shifted to calculating the rate per 100,000 people, allowing for a more balanced comparison and providing a clearer picture of the police killings across racial groups over time. This approach helped ensure that the visualization was more meaningful and highlighted the disparities more effectively.

Age distribution of deaths across racial groups



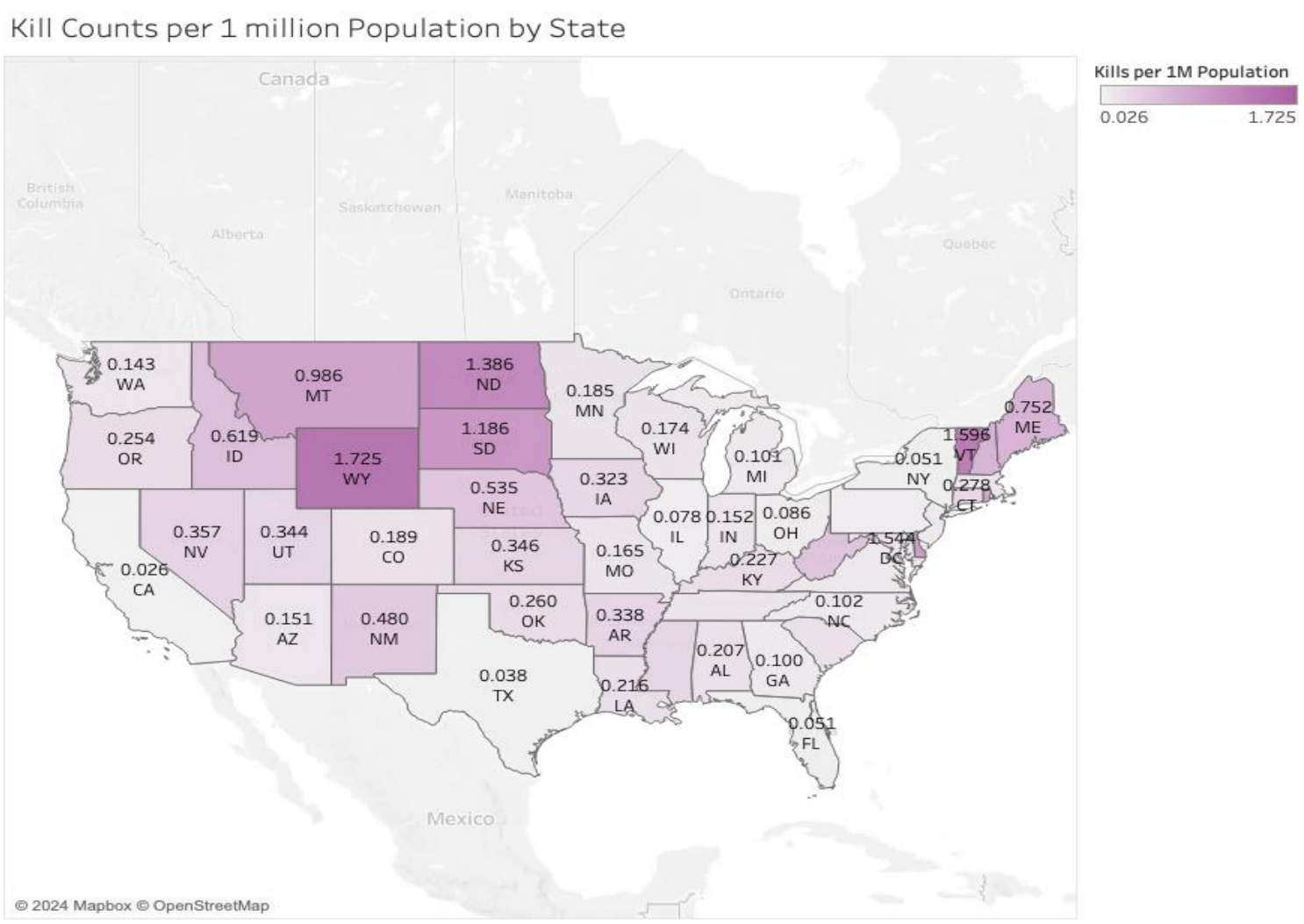
The chart illustrates the age distribution of deaths across racial groups using boxplots and scatterplots. The boxplots provide a summary of the median age and typical age range, while the scatterplots show individual cases for a more detailed view. The White group has the widest age distribution, with deaths reaching nearly 100 years. In contrast, the Black, Asian, Hispanic, and Native groups display more compact distributions, with the Black group having the youngest median age. This suggests that deaths in the Black group tend to occur at younger ages.

The presence of a younger median age in the Black group suggested that additional variables such as the presence of a weapon and flee status could play a role in these patterns.



The treemap visualization examines cases of deadly force used against unarmed individuals across different racial groups, normalized per 100 million population. This data point is particularly significant as it addresses one of the most controversial aspects of police use of force - situations involving unarmed individuals where the necessity of deadly force is highly debatable. The treemap's proportional visualization effectively communicates these disparities through the stark differences in block sizes, providing compelling data for discussions about police reform, accountability measures, and the critical need for enhanced de-escalation training and protocols.

The visualization reveals stark racial disparities, with Black individuals facing the highest rate at 235.8 cases per 100 million - more than three times the rate for white individuals (68.1) and nearly ten times the rate for Asian individuals (24.7). Native Americans experience 65.5 cases per 100 million, while Hispanic individuals face 37.7 cases per 100 million. These disparities raise profound questions about police training, implicit bias, and systemic inequities in law enforcement practices, particularly given that these cases specifically involve unarmed individuals who, by definition, did not pose the same level of immediate lethal threat that might justify deadly force.



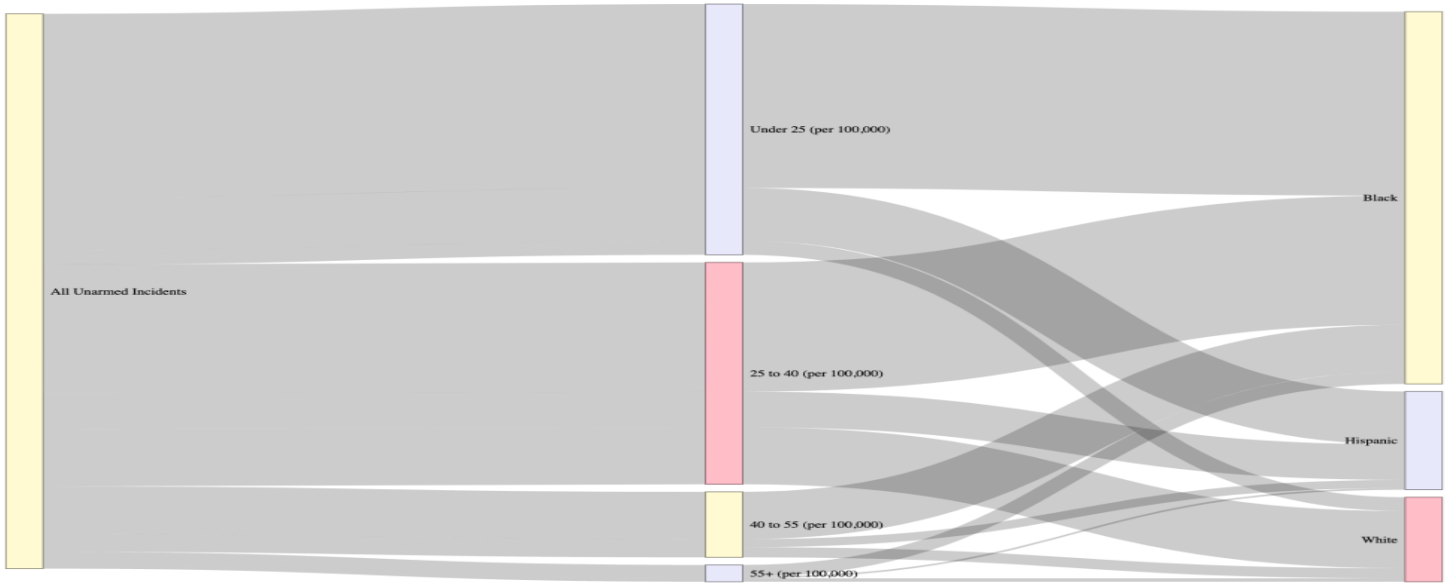
Map based on Longitude (generated) and Latitude (generated). Color shows Kill rate per pop. The marks are labeled by Kill rate per pop and State. Details are shown for State.

This choropleth map visualizes kill rates per million population across the United States using varying color intensities to represent higher rates. Created in Tableau, the visualization combines geographic data (latitude and longitude) with calculated kill rates derived from raw count data and population figures. Killing count was a calculated field created using COUNT([Names]). This field was used to create the kill count per 1 million Total population coulmn [Killing Count] / SUM([Total Pop])*1000000. The new column was added to color under marks and a sequential color scheme was chosen to represent the killing rates.

The resulting map reveals unexpected patterns that challenge conventional assumptions about violence and population density. Notably, rural states like Wyoming and Vermont show surprisingly high kill rates, in contrast to more urbanized states like California, Texas, and New York. This suggests that factors beyond population density - such as socioeconomic conditions, mental health resource accessibility, and gun regulations - may significantly influence regional kill rates. These findings indicate the need for deeper analysis to identify high-risk geographic areas and develop targeted regulatory approaches to promote public safety and equity.

Final Visualizations

Age and Race Collision

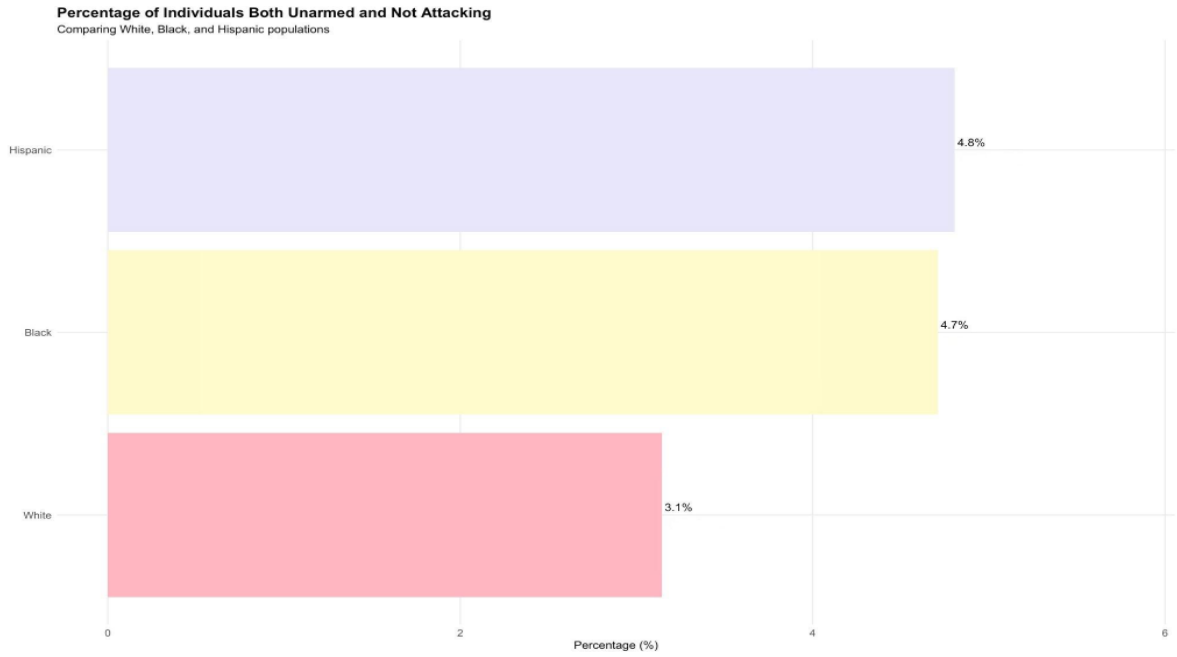


Our story begins with a Sankey diagram used to illustrate the relationship between unarmed police incidents, age demographics, and racial groups across America. The age group categories are: Under 25, 25-40, 40-55, and 55+. The width of each flow represents the volume or proportion of cases, providing an intuitive visual representation of the magnitude of cases. The visual focuses on the three largest racial groups in America: White, Black, and Hispanic represented by pastel colors- light pink, light yellow and lilac.

During the exploratory phase, a box plot was initially used to examine the relationship between unarmed police incidents, age, and racial demographics. However, while box plots effectively show statistical distributions and outliers, they fell short in clearly communicating the complex flow and interconnections between these critical variables. The Sankey format allows viewers to trace paths from incident source through age groups to racial categories, revealing patterns that might be obscured in traditional statistical plots.

Within the broader analysis of police killings across America, this visualization serves as a crucial piece in understanding the intersectionality of age and race in police encounters. It reveals how unarmed incidents disproportionately affect different racial groups across age brackets, with particularly striking patterns among younger Black individuals. This adds a vital layer of nuance to the overall story by showing that racial disparities in police use of force aren't uniform across age groups, suggesting that certain demographic combinations face heightened risk. The Sankey diagram's ability to show these relationships simultaneously, rather than in separate distributions as a box plot would, strengthens the narrative by making these patterns immediately visible and comprehensible to viewers who might not have statistical training.

Killings of Unarmed Individuals Across Races

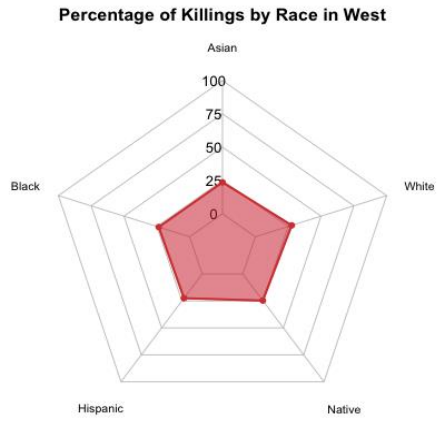


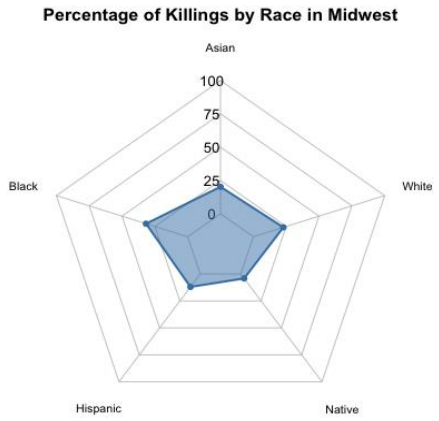
The second part of our analysis focuses on examining bias in the killing of unarmed individuals across different racial groups. We used a Horizontal Bar Graph to explore cases when individuals were both unarmed and not attacking. This is a significant variable, because it is highly debated whether deadly force is necessary for those who are unarmed, especially when they are not attacking. It shows the percentage of unarmed individuals not attacking. It uses the same pastel color scheme to represent different races.

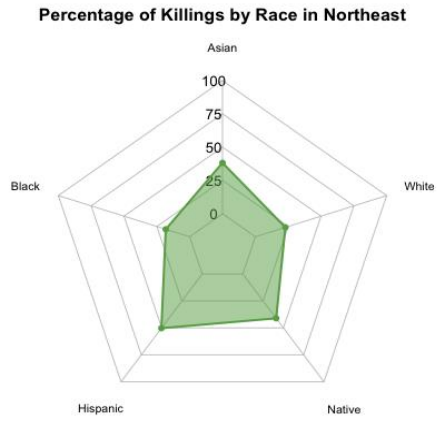
During our drafting process, we noticed that certain racial groups were disproportionately targeted despite being unarmed, leading us to question whether they were killed due to perceived threats to law enforcement. However, after adding the *threat_level* variable, we discovered that many of these incidents involved low or no perceived threat, challenging the assumption that these deaths were mainly due to threats. This additional data provided a clearer understanding of the circumstances behind these cases.

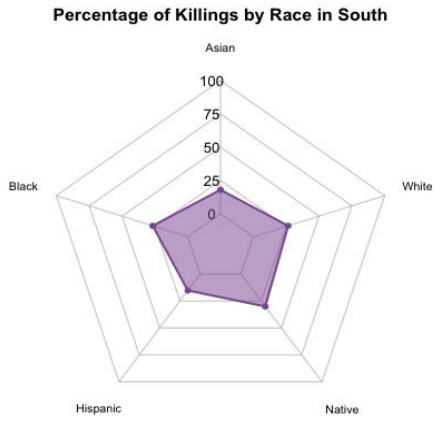
The visualization starkly reveals concerning racial disparities: 4.8% of Hispanic victims and 4.7% of Black victims were both unarmed and not attacking when killed, compared to 3.1% of White victims. These statistics translate to Hispanic and Black individuals being approximately 1.5 times more likely than White individuals to be killed in situations where they posed no apparent threat to law enforcement. This disparity becomes even more significant when considering that these cases represent situations where standard law enforcement protocols would typically call for de-escalation rather than deadly force.

Geography of Risk







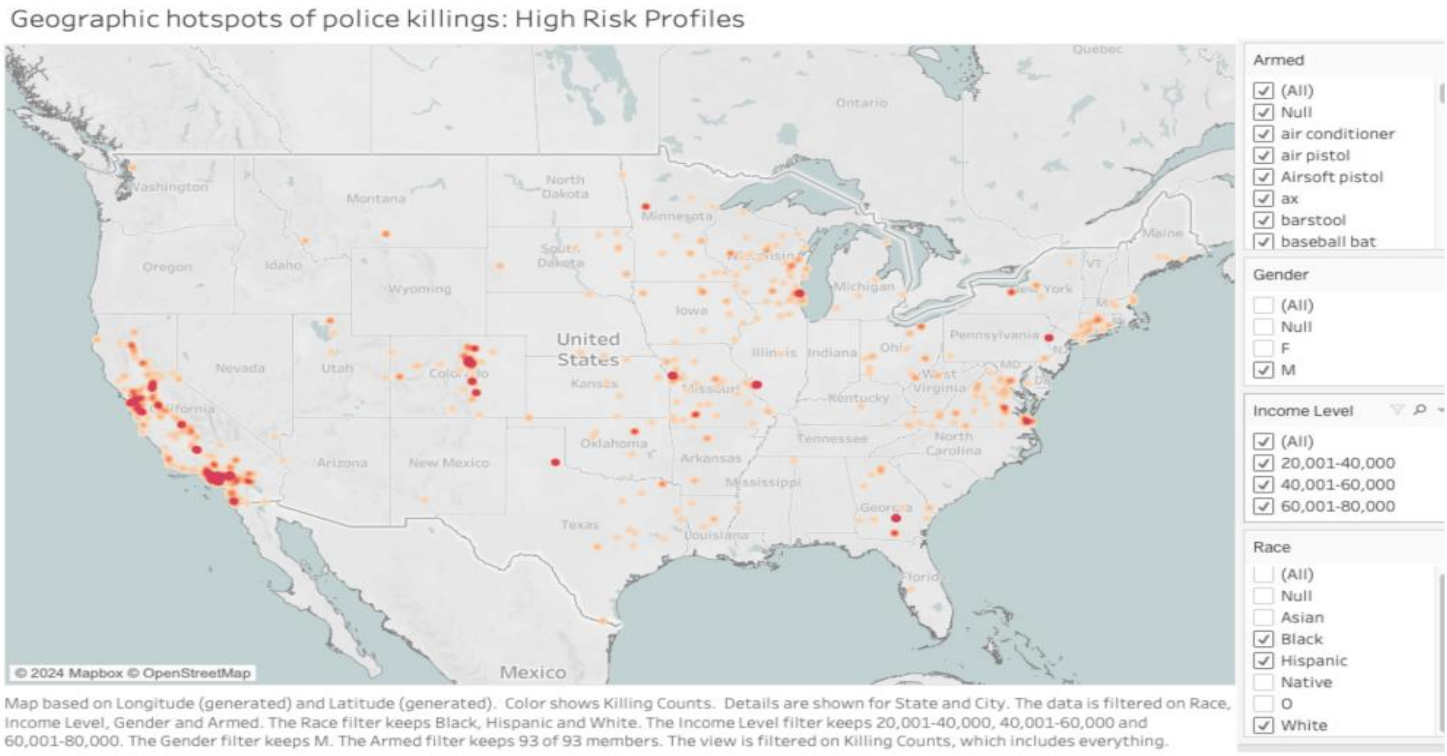


We utilized a star plot to display the percentage of killings by race in four distinct regions: West, Midwest, Northeast, and South. This visualization helps us identify regional disparities and highlights areas where racial groups face elevated risks. The star plot reveals how these percentages vary by region, with notable differences in the distribution of killings, illustrating how geographic location can influence racial vulnerabilities.

First the state abbreviations were mapped to 4 regions. The dataset was prepared by counting the occurrences of each racial group within these regions, normalizing these counts to percentages, and then plotting the results on separate radar charts using library(fmsb) in R. These visualizations allow for easy comparison of racial disparities in police killings within and across regions. Different colors represent different regions and title is added to every sub plot.

Initially, bar charts were considered to represent the data, but they lacked the ability to highlight regional disparities effectively. During the drafting phase, we explored different chart types like violin plots to show the distribution by region. We found that radar charts better emphasized the relative proportions within each region. A single radar chart with 4 regions was first created. But the regions were overlapping even though alpha values were reduced. It was hard to make out the patterns as Northeast was overlapping. So, four plots were faceted, shown side by side which removed clutter and increased clarity.

Intersection of Race, Gender, Armed Status and Income



During the drafting process, the choropleth visualization evolved as we refined how we displayed and interacted with the data to uncover meaningful patterns. Initially, the map simply showed the raw frequency of police killings by location, but it lacked a clear context for understanding the underlying factors contributing to the concentration of these incidents. We even tried to use a mosaic plot to show relationships between categorical variables. But we could not include geography. So in order to incorporate multiple variables, such as location, *armed status*, *gender*, *race*, and *income level*, we created a Choropleth with Interactive feature. Thus, we were able to reveal patterns based on demographic and geographic intersections. We adjusted the color intensity to highlight areas with higher frequencies of killings, making it easier to visually identify geographic hotspots. The addition of interactivity allowed users to filter by these variables, enabling us to explore how different profiles—such as specific racial groups, income levels, or armed statuses—correlated with higher rates of police killings in particular regions. Income was binned into 3 categories as shown in the filter. Kill count was a calculated field created using COUNT([Names]). Tooltips were added to provide detailed information about each data point, giving users insight into the demographic makeup of each location.

This refinement process allowed the visualization to move from a simple depiction of incidents to a powerful tool for understanding the intersectionality of race, income, armed status, and geography in police killings. The story behind this analysis centers on identifying high-risk areas and profiles where compounded vulnerabilities, such as certain racial groups or lower-income individuals, correlate with an increased likelihood of police killings. By incorporating these refinements, the visualization not only revealed patterns but also provided a deeper, more actionable insight into the factors contributing to police violence. The interactivity and detailed breakdown of these factors enhance the

story, allowing users to explore and identify key correlations and make informed calls to action based on the patterns that emerged.

Analysis and Discussion

Age and Race Collision

Initially, a box plot was used to explore the relationship between unarmed police incidents, age demographics, and racial groups in America, but it struggled to convey the complex flow between the variables. The Sankey diagram, however, allowed for a more intuitive understanding of how age and race intersect in unarmed police encounters, highlighting that younger Black individuals-under 25 and 25 to 45 years of age are disproportionately affected. Younger white demographic is the second most affected. This visualization adds depth to the broader analysis of racial disparities in police use of force, demonstrating that these disparities are not uniform across age groups and that certain demographic combinations face a heightened risk.

Killings of Unarmed Individuals Across Races

The analysis revealed significant racial disparities in cases where individuals were unarmed and not attacking. Hispanic and Black victims were about 1.5 times more likely than White victims to be killed in such situations, despite posing no apparent threat to law enforcement. After incorporating the *threat_level* variable, it became evident that many of these incidents involved little to no perceived threat, which contradicted the initial assumption that deadly force was used in response to imminent danger. This underscores significant racial disparities in the use of deadly force, especially in situations where de-escalation should have been prioritized.

Geography of Risk

These radar charts reveal distinct regional patterns. For instance, the Northeast shows a higher proportion of Hispanic victims, while the Midwest has a notably high percentage of Black victims whereas an equal proportion is represented in the West. This visualization fits into the broader analysis by providing a geographic lens on racial disparities in police killings, demonstrating how these disparities vary regionally. These insights support a more nuanced understanding of the intersection of race and geography, emphasizing the need for region-specific policy interventions and reform strategies.

Intersection of Race, Gender, and Income

The provided map offers a compelling visual representation of geographic hotspots for police killings across the United States. By filtering the data to include Black, White, and Hispanic individuals across all armed statuses and income levels, the map highlights areas with a higher concentration of such incidents. The map reveals a notable concentration of police killings along the West Coast, particularly in California. This region appears to have a higher density of incidents compared to other parts of the country. So do some parts of the East coast and the Midwest region, including states like Illinois, also show a significant number of police killings, suggesting a potential hotspot in this area.
Further analysis could involve building distinct profiles based on various factors, such as armed status, mental illness, and fleeing behavior. If time permits, additional filters for these variables could be explored to create more refined profiles and to better understand the compounded risks. By examining these profiles, strategies could be developed to address systemic issues, reduce bias, and promote reforms aimed at ensuring equality and justice in policing practices.

Conclusion

Our visualizations provide multiple perspectives on police killings. The Correlation Plot identifies relationships between variables, informing other visualizations. The Stacked Bar Chart shows trends in police killings by race over time, while the Boxplot highlights age distribution across racial groups. The Treemap displays unarmed cases by race relative to population, and the Choropleth Maps illustrate the geographic distribution of killings.

The final visuals further refine the analysis: the Sankey Plot shows the flow between age, race, and unarmed incidents, and the Bar Graph focuses on killings of unarmed, non-attacking individuals by race. The Star Plot compares geographic risk across regions and races, and the Interactive Choropleth Map allows dynamic exploration by location, armed status, gender, race, and income.

Our visualizations reveal significant racial disparities in police killings, with Black and Hispanic communities disproportionately affected. The data suggests that socioeconomic factors and geographic location also play critical roles in these disparities. These insights can guide policy changes and targeted interventions to address systemic issues in law enforcement.

With more time, we would develop interactive dashboards and incorporate more socioeconomic data, qualitative data, and predictive modeling for a richer analysis.

Appendix-A

Individual Reports

Ankita Mishra

In this project, I embarked on a comprehensive data gathering mission, sourcing datasets from both Kaggle and the U.S. government census website. This involved meticulously merging the datasets to create a unified and comprehensive database. I created new columns to enhance our ability to visualize the data effectively and ensure it captured the relevant insights we were aiming to analyze. Cleaning missing values was a critical step in this process, ensuring the integrity and accuracy of our data analysis.

Beyond data preparation, I played a significant role in shaping the narrative of our project. I helped craft a storyline that guided our selection of visuals, ensuring that each visualization was purposeful and aligned with our overarching goal of understanding the racial disparities in police killings. One of my key contributions was developing a visual that examined the number of killings by race in relation to the proportion of each race within the total U.S. population. This visual was crucial in highlighting the disproportionate impact on certain racial groups. I also built the initial mosaic plot, which illustrated the relationship between age, race, gender, and police killings. This plot provided a multi-dimensional view of the data, revealing intricate patterns and intersections that might otherwise have gone unnoticed. This was improved by Kavana to include income level instead of age to look at a different aspect of our dataset. Each visual was carefully analyzed to understand its implications, particularly in terms of systemic racism within law enforcement.

This project has been an enlightening journey into the power and importance of data visualization. I learned that visualizations are not just about presenting data; they are about telling a story and revealing hidden insights that raw numbers alone cannot convey. Effective visualizations can highlight disparities, uncover patterns, and drive home the impact of systemic issues like racial disparities in police killings. Additionally, the process of creating these visuals reinforced the importance of data integrity, the need for a clear narrative, and the power of diverse visualization techniques to complement and enhance each other's insights.

Jonathan Lindahl

For our group project, we divided the work up 4 ways equally, and my role was to analyze unarmed cases, and which variables are significant in this group of people who were killed by police.

As an online student I was lucky I didn't have to present and so I wanted to contribute as much as I could in other ways. I helped with creating the powerpoint and also a number of different visualizations.

The visualizations that I worked on were the Sankey Diagram that shows how Unarmed cases connect to race and age. I also created the Box Plot with the jittered data points.

Then for the third one I created the Horizontal Bar plot that compares White, Black, and Hispanic cases where the individual was both unarmed and not attacking.

For the exploratory visuals I created the treemap of unarmed cases by race by population, and also the police killings by race bar graph.

The Sankey Diagram took me the longest. At first I was trying to figure out which variables connect logically, and which would tell the most compelling story. After trial and error, I finally arrived at all unarmed cases and how they connect with race and age.

All the visuals I created were to clearly show the differences between races when it came to unarmed cases. This variable was important to me because there are many societal questions that come with unarmed victims who are killed by police.

I learned a great deal from creating these visualizations. I used R for each one, and got the chance to explore new libraries I never used before.

I also learned a lot from the presentations. From what I gathered, the audience should know the message your visual is portraying within a few seconds, or else it's no good. At times it seems people want to try including as many variables as possible in a complex graph, but what I found worked best was focusing on just a few variables that matter the most with a simple graph that people truly understand.

Some of the key takeaways from the class and throughout this quarter were the role of color. I know now that pastels are softer on the eye, where neon is tough to look at for a while. Also that green is perceived the most by us compared to other colors.

I also learned about divergent and sequential color schemes, and when it's best to use them. I never knew about these before attending this class and I find them very useful.

Another thing I learned is that there is no one size fits all way to visual data. There are many different graphs, colors, and designs that can be used to portray the same message effectively.

This class felt like a combination of psychology, design, and data science all combined into one. I work as an email marketing strategist and I have to create reports weekly. Thanks to this class I now understand Tableau and R much better, and I can use what I learned in this class directly with what I do at work.

I now find myself analyzing visuals whenever I come across them, whether that's on tv, driving down the street or just using my computer. I start thinking about why they chose certain colors, and used certain graphs. This class really sparked my curiosity and I am excited to keep improving the visualizations I create in the future.

Kavana Manvi Krishnamurthy

Role and Contribution:
Exploratory Analysis:

- **Analyzing Kill Rates per Million Population in the United States, Choropleth:** The choropleth map highlights kill rates per million population across the U.S., revealing high kill rates in unexpected states like Wyoming and Vermont.
- **Intersectional Variables, Correlation Heatmap:** It shows how race, age, gender, flee, mental illness, and income intersect in police killings.
- **Milestone 3:** I created a heatmap that emphasized the intersection of race, gender, and socioeconomic status in relation to police violence, highlighting the compounded vulnerabilities faced by marginalized communities.

Final Visualizations:

- **Geography of Risk, Radar chart:** The radar chart displays the percentage of killings by race in each region (Midwest, Northeast, South, and West) by first counting occurrences, normalizing them to percentages, and creating four separate charts for each region.
- **Intersection, Interactive Choropleth:** The choropleth map highlights police killings across high-risk groups (e.g., Black, Hispanic, White males) and lower income levels, with color intensity showing the frequency of killings.

Reflections and Learnings:

Working on this project taught me how to choose visualizations that effectively tell a complex story. I started by exploring geographic variables and important variables like gender,race, income etc. I created a heat map to understand the correlation and created a choropleth to identify high-risk profiles. This project reinforced the importance

of aligning data exploration and visualization techniques with the narrative we want to tell, particularly when handling both numeric and categorical variables.

Throughout the project, I gained valuable technical and non-technical skills. Theoretically speaking, I learnt a lot about- Graph types, guidelines, clutter, distortion, chart junk, color, transformation, etc. Technically, I learned various types of visualizations in R, experimenting with different graphs, and became proficient in using Tableau, especially for creating interactive dashboards. On the communication front, I realized that it's not just about the graphs and code but about the data's message and the audience—this became my mantra. I learned to choose the right graph for the right mix of numerical and categorical data to effectively tell a story. Managerially, I embraced the concept of "divide and conquer," improving my ability to communicate ideas with teammates and collaborate better. It was a fun learning experience, and I gained a lot from it.

Sanchal Sunil Dhurve

In this project, I played a significant role in both the exploratory and milestone analyses, contributing to the overall understanding and narrative of our findings. My work began with researching datasets and the team finalizing the one which could stand out in the storytelling. We started with exploratory analysis and I was responsible for creating a stacked bar chart that highlighted the top 10 states with the highest police killings per 100,000 population by race. This visualization provided critical insights, showing that California and Texas had the highest numbers of police killings, with White and Hispanic populations bearing the largest impacts, closely followed by Black populations, who consistently experienced disproportionately high rates. Although this chart was not included in the final report, it played an essential role in our presentation and helped shape our initial understanding of regional and racial disparities. Similarly, in Milestone 3, I developed another stacked bar chart examining the relationship between poverty rate and race in police killings, revealing how socioeconomic factors exacerbate vulnerabilities, particularly for Hispanic and Black populations. This chart, while pivotal during analysis, was also not included in the final report but contributed significantly to discussions that refined our focus on key disparities.

Even though none of the visualizations I created were directly featured in the final report, they were instrumental in informing the broader analysis and narrative of the project. My contributions helped guide the team’s understanding of critical patterns and laid the groundwork for final visualizations that emphasized racial, socioeconomic, and geographic disparities in police killings. Beyond creating visualizations, I played an important role in shaping the project’s narrative, ensuring that our analysis and findings were cohesive and aligned. I also worked closely with the team to structure the presentation, focusing on communicating insights effectively to a wider audience.

Reflecting on this project, I have learned that data visualization is not just about creating visually appealing charts but about uncovering insights and telling a meaningful story.

The process taught me the importance of crafting a clear narrative and selecting visuals in Tableau and R that align with the project's objectives. Even when certain visualizations are not included in the final deliverable, their role in shaping the understanding and direction of the analysis is invaluable. Collaborating with my team further highlighted the power of collective effort and diverse perspectives in refining and enhancing the quality of work. Overall, this experience underscored the potential of data visualization to illuminate systemic issues, foster awareness, and drive impactful discussions.

Appendix - B

Formative Exploratory Data Analysis

Story	Exploratory Visualizations	Final Visualizations
● Age and Race Collisions	Box Plot: Age Distribution of Deaths across Races	Sankey: Age and Top 3 races
● Armed vs Unarmed across ages	Unarmed Treemap: Biased towards certain Races	Bar Graph: Killings of Unarmed Individuals across Ages
● Geography of Risk	Choropleth Graph: Kill Counts per 1M by state	Radar Plot: Percentage of killings by Region
● Intersection of Race, Gender and Income	Correlation Heatmap: Gender, Age, Race, Income, Mental Health, Flee	Interactive Choropleth: Density- Kill count Filter- Gender, Race, Income, Armed

Appendix-C

R-codes for Exploratory and Final Visualizations

Exploratory Visualizations Codes:

1. **Stacked Bar Chart: Police Killing Rates by Race (2015-2019)**

```
library(ggplot2)

# Create the stacked bar chart

 ggplot(pk, aes(x = Year, y = Rate, fill = Race)) + geom_bar(stat = "identity") + labs(title = "Police Killing Rates by Race (2015-2019)", x = "Year", y = "Rate per 100,000 Population") + scale_fill_manual(values = c("White" = "lightblue", "Black" = "lightcoral", "Hispanic" = "lightgreen", "Native" = "lightgrey", "Asian" = "lightpink")) + theme_minimal()
```

2. **Box Plot: Age Distribution of Deaths by Race**

```
library(ggplot2)

 # Create the jittery bar plot

ggplot(pk, aes(x = Race, y = Age, color = Race)) + geom_boxplot() + geom_jitter(alpha = 0.5, size = 1.5) + labs(title = "Distribution of Age by Race", x = "Race", y = "Age") + theme_minimal()
```

3. **Unarmed Treemap : Biased towards certain races**

```
pk$Rate <- (pk$KillCount / pk$Population) * 100000
# Melt the data for heatmap
pk_melted <- melt(pk, id.vars = "Race", measure.vars = "Rate")

# Create the heatmap
ggplot(pk_melted, aes(x = Race, y = variable, fill = value)) + geom_tile() +
scale_fill_gradient(low = "lightblue", high = "darkblue", name = "Killing Rate") + labs(title =
```

```
"Unarmed Cases by Race Relative to Population (per 100 million)", x = "Race", y = "") +
theme_minimal()
```

4. Correlation Heatmap

```
#pk is the data frame with police killings data
pk<- read.csv("~/Desktop/DV/Final Project/cleaned_data.csv", stringsAsFactors=TRUE)

data <- pk[, c("race", "gender", "age",
               "Income","flee","signs_of_mental_illness")]
data_dummy <- fastDummies::dummy_cols(data,
                                      select_columns = c("race", "gender",
"flee","signs_of_mental_illness"),
                                      remove_first_dummy = TRUE,
                                      remove_selected_columns = TRUE)
corr_matrix <- round(cor(data_dummy), 2)


# Replace NA, NaN, and Inf values with 0
corr_matrix[is.na(corr_matrix)] <- 0
corr_matrix[is.nan(corr_matrix)] <- 0
corr_matrix[is.infinite(corr_matrix)] <- 0

# Plot the heatmap with exact values
ggcorrplot(corr_matrix,
           lab = TRUE,
           lab_size = 3,
           method = "square",
           colors = c("red", "white", "blue"),
           title = "Correlation Heatmap: Intersectional Variables",
           ggtheme = theme_minimal())
```

Final Visualization Codes:

1. Sankey: Age and Race Collision

R Code Sankey

```
library(tidyverse)
library(networkD3)

population_adjusted <- tree_data %>%
  mutate(age_group = case_when(
    age < 25 ~ "Under 25",
    age >= 25 & age < 40 ~ "25 to 40",
    age >= 40 & age < 55 ~ "40 to 55",
    age >= 55 ~ "55+"
  )) %>%
  group_by(age_group, race) %>%
  reframe(
    total_cases = n(),
    unarmed_cases = sum(armed_status == "Unarmed"),
    population = first(case_when(
      race == "White" ~ White,
      race == "Black" ~ Black,
      race == "Hispanic" ~ Hispanic
    )),
    total_rate = (total_cases / population) * 100000,
    unarmed_rate = (unarmed_cases / population) * 100000
  ) %>%
  filter(!is.na(population)) %>%
  arrange(desc(unarmed_rate))

age_groups_ordered <- c("Under 25", "25 to 40", "40 to 55", "55+")

nodes <- data.frame(
  name = c(
    "All Unarmed Incidents",
    paste(age_groups_ordered, "(per 100,000)"),
    unique(population_adjusted$race)
  )
)
```

```
population_adjusted <- population_adjusted %>%
  mutate(age_group = factor(age_group, levels = age_groups_ordered)) %>%
  arrange(age_group, race)

links <- rbind(
  data.frame(
    source = match("All Unarmed Incidents", nodes$name) - 1,
    target = match(paste(population_adjusted$age_group, "(per 100,000)"), nodes$name) -
1,
    value = population_adjusted$unarmed_rate
  ),
  data.frame(
    source = match(paste(population_adjusted$age_group, "(per 100,000)"), nodes$name) -
1,
    target = match(population_adjusted$race, nodes$name) - 1,
    value = population_adjusted$unarmed_rate
  )
)

sankeyNetwork(Links = links,
              Nodes = nodes,
              Source = "source",
              Target = "target",
              Value = "value",
              NodeID = "name",
              fontSize = 12,
              nodeWidth = 30,
              height = 800,
              width = 1200,
              sinksRight = TRUE,
              colourScale = JS(sprintf(
                'd3.scaleOrdinal()
                  .domain(["Black", "Hispanic", "White"])
                  .range(["#FFFACD", "#E6E6FA", "#FFB6C1"])'  # Light yellow, Light purple, Light
pink
              ))) %>%
  htmlwidgets::onRender(
    "function(el) {
      d3.select(el)
        .select('svg')
        .append('text')
        .attr('x', width/2)
        .attr('y', 30)
        .attr('text-anchor', 'middle')
        .style('font-size', '16px')
        .style('font-weight', 'bold')
```

```
      .text('Unarmed Police Killings by Age Group and Race');
  }"
)
```

2. Armed vs. Unarmed Across Races

```
# Filter data for "White", "Black", and "Hispanic" races
filtered_data <- pk %>% filter(race %in% c("White", "Black", "Hispanic"))

 # Calculate the percentage of unarmed and not fleeing individuals for each race
unarmed_not_fleeing_percentages <- filtered_data %>% group_by(race) %>% summarize(
total_kills = n(), unarmed_not_fleeing = sum(armed == "unarmed" & threat_level == "other")
) %>% mutate(percentage = (unarmed_not_fleeing / total_kills) * 100)

# Create the sorted horizontal bar plot
ggplot(unarmed_not_fleeing_percentages, aes(x = reorder(race, percentage), y =
percentage, fill = race)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), hjust = -0.3) +
  labs(title = "Percentage of Individuals Both Unarmed and Not Attacking\nComparing White,
Black, and Hispanic Populations",
       x = "Race",
       y = "Percentage (%)") +
  scale_fill_manual(values = c("White" = "lightpink", "Black" = "lightyellow", "Hispanic" =
"lavender")) +
  theme_minimal() +
  coord_flip()
```

3. Geography of risk:Radar Plot

```
pk<- read.csv("~/Desktop/DV/Final Project/cleaned_data.csv", stringsAsFactors=TRUE)
head(pk)

# Load necessary library
```

```
library(dplyr)
library(ggplot2)

# Create a mapping of state abbreviations to regions
state_to_region <- c(
  "CT" = "Northeast", "ME" = "Northeast", "MA" = "Northeast",
  "NH" = "Northeast", "NJ" = "Northeast", "NY" = "Northeast",
  "PA" = "Northeast", "RI" = "Northeast", "VT" = "Northeast",

  "IL" = "Midwest", "IN" = "Midwest", "IA" = "Midwest", "KS" = "Midwest",
  "MI" = "Midwest", "MN" = "Midwest", "MO" = "Midwest", "NE" = "Midwest",
  "ND" = "Midwest", "OH" = "Midwest", "SD" = "Midwest", "WI" = "Midwest",

  "AL" = "South", "AR" = "South", "DE" = "South", "FL" = "South",
  "GA" = "South", "KY" = "South", "LA" = "South", "MD" = "South",
  "MS" = "South", "NC" = "South", "OK" = "South", "SC" = "South",
  "TN" = "South", "TX" = "South", "VA" = "South", "WV" = "South",

  "AK" = "West", "AZ" = "West", "CA" = "West", "CO" = "West",
  "HI" = "West", "ID" = "West", "MT" = "West", "NV" = "West",
  "NM" = "West", "OR" = "West", "UT" = "West", "WA" = "West", "WY" = "West"
)

# Add the 'region' column to the 'pk' dataframe
pk <- pk %>%
  mutate(region = state_to_region[pk$state])

#remove null race
pk<-pk %>%
    filter(!is.na(race) & race != "O" & race!="")
# Drop unused levels from the 'race' factor
pk$race <- droplevels(pk$race)
unique(pk$race)

#remove null region
pk<-pk %>%
    filter(!is.na(region))

unique(pk$region)
# View the updated dataframe
```

```
head(pk)


library(fmsb)
library(dplyr)
library(tidyr)
library(RColorBrewer)

# Prepare the summary dataset: Count occurrences of races within each region
region_race_count <- pk %>%
    group_by(region, race) %>%
    summarise(count = n(), .groups = 'drop') %>%
    pivot_wider(names_from = race, values_from = count, values_fill = list(count = 0))

# Normalize the data (scale it to percentages for proportional comparison)
region_race_count_normalized <- region_race_count %>%
    mutate(across(-region, ~ . / sum(.) * 100))

# Prepare the data for the radar charts
region_race_count_normalized <- as.data.frame(region_race_count_normalized)
row.names(region_race_count_normalized) <- region_race_count_normalized$region
region_race_count_normalized <- region_race_count_normalized[, -1]

# Set max-min values for the radar charts (scale from 0 to 100%)
max_min <- data.frame(matrix(c(rep(100, length(unique(pk$race))), rep(0,
length(unique(pk$race)))),
                             ncol = length(unique(pk$race)), byrow = TRUE))
colnames(max_min) <- colnames(region_race_count_normalized)

# Define colors for regions using a new palette
region_colors <- brewer.pal(4, "Set1")
names(region_colors) <- unique(pk$region)

# Set up the plot area for 4 plots
par(mfrow = c(2, 2), mar = c(1, 1, 2, 1))

# Create a radar chart for each region
for (region in unique(pk$region)) {
    data_for_radar <- rbind(max_min, region_race_count_normalized[region, , drop = FALSE])
```

```
  radarchart(data_for_radar,
             axistype = 1,
             pcol = region_colors[region],
             pfcol = scales::alpha(region_colors[region], 0.5),
             plwd = 2,
             cglcol = "grey",
             cglty = 1,
             axislabcol = "black",
             caxislabels = seq(0, 100, 25),
             cglwd = 0.8,
             vlcex = 0.8,
             title = paste("Percentage of Killings by Race in", region))
}
```