Data Visualization A1

Dhruv Kothari
IMT202211
IIIT Bangalore
dhruv.kothari@iiitb.ac.in

Harsh Modani
IMT2022055
IIIT Bangalore
harsh.modani@iiitb.ac.in

Mohammad Owais
IMT2022102
IIIT Bangalore
mohammad.owais@iiitb.ac.in

Dataset

This dataset, created by *The Washington Post*, tracks every individual fatally shot by an on-duty police officer in the U.S. from 2015 to 2024. It was developed after the 2014 Ferguson incident, when it was revealed that FBI statistics significantly underreported these incidents—capturing only about one-third of fatal police shootings by 2021. This database seeks to close that gap by providing detailed information on each case, including the police departments involved, in order to promote greater transparency and accountability. The data fields present in the dataset are:

1) Date: The date on which the shooting has occurred
2) Name: The name of the person shot
3) Gender: The gender of the person shot
4) Armed: If and what the person shot was armed with
5) Race: The race of the person shot
6) City: City in which the shooting has occurred
7) State: 2 letter US state code of the state in which the shooting occured
8) Flee: If and what with the person shot was fleeing with
9) Body Camera: Indicates if the police officer was or not wearing a body camera
10) Signs of Mental Illness: If there were signs of mental illness present in the person shot, as determined by the police officer at the time of shooting
11) Police Departments involved: Every police department involved in this particular case

We also have calculated fields in the data, that include:

1) Number of police departments: The number of police departments involved in the shooting
2) Fleeing: Aggregates 'not' into 'not fleeing', and fleeing by 'car', 'foot', or 'other', to 'fleeing'
3) known_race: Aggregates 'unknown' into 'unknown race' and 'known race'
4) before_2022: Aggregates year 2021 and years before 2021 into 'before_2022' and year 2022 and years after 2022 into 'after_2022'

We have also imported an additional dataset, of which the fields we have used were:

1) City: Cities in the United States
2) State: 2 letter state code of each state
3) Population: The population of each city

Task

Through visual exploratory analysis, we target to gain the following insights and expect the one to reproduce the following tasks:

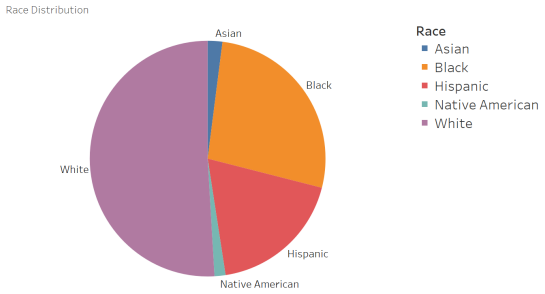
T1: Demographic Analysis
T2: Geopolitical Analysis
T3: Contextual Analysis

Assumption/Data Filtration

Since the data entries were very large, a lot of visualization used won't make much sense. Due to this reason, we applied some sort of data filtration which mostly included the following constraints.

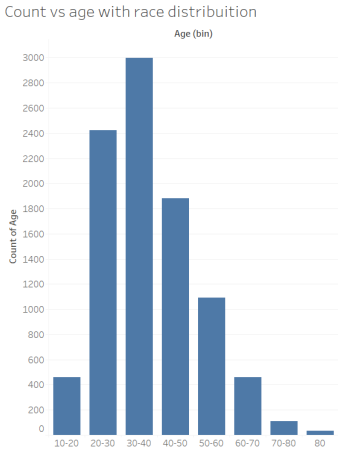
1) To ensure consistency in our analysis of time-related data, we excluded entries from the year 2024, as the data for that year is incomplete.
2) To improve clarity, visualizations exclude outliers from regions with minimal data, focusing instead on areas where the majority of data is concentrated.
3) Null values and entries labeled as 'others' were not included in the visualizations to maintain accuracy and ensure clearer visual representation.

Data Stories



T1.1

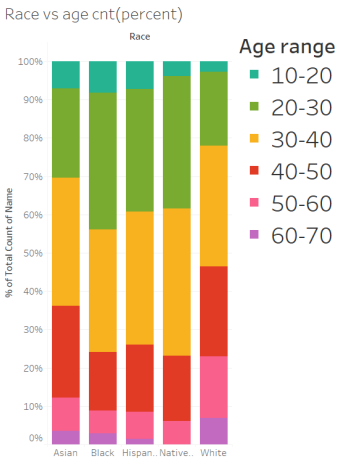
T1 Demographic Analysis:



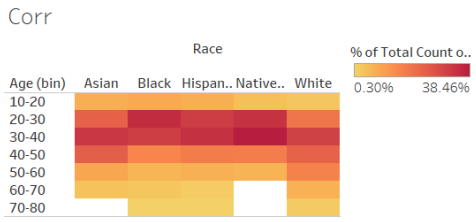
T1.2

The visualizations, Fig T1.1 and Fig T1.2 shows the distribution of age and race in police shooting cases recorded in the dataset.

Hypothesis T1.1: Certain racial groups experience a higher incidence of police shootings at different age groups.
The idea behind this hypothesis is that different racial groups may have varying age distribution in police shooting incidents due to social disparity, like younger individuals from certain racial groups may be more involved in police shooting due to social, economic, or systemic factors, while other groups may see more incidents in older age groups. We can use a stacked bars to analyze the percentage distribution of age groups across different racial groups. This visualization allows us to compare how different age groups are distributed in different races. The hypothesis can be visualized using Fig T1.3 and the correlation can be confirmed using Fig T1.4.



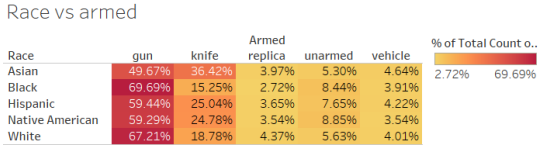
T1.3



T1.4

Based on the analysis, we found that about 42% of police shootings involving Black, Hispanic, and Native American individuals involved people aged 10-30, while only 20% of cases involving White individuals were in this age range. This shows a clear difference in how age groups are distributed across racial groups, with younger people from minority groups being more affected. The visualization reflects this disparity, suggesting that age is an important factor when looking at racial differences in police shootings.
The stacked bars allows for a clear comparison of age group within each racial category. By stacking the bars we can see the proportion of individuals from each race in different ages ranges.

Hypothesis T1.2: There is a Correlation Between Race and Factors Such as Being Armed, Fleeing, or Signs of Mental Illness.
The hypothesis that there is a correlation between race and factors such as being armed, fleeing, or showing signs of mental illness makes sense because these factors can influence the dynamics of police encounters. Different racial groups might experience varying circumstances during such interactions due to social, economic, and systemic factors.
The presence or absence of a weapon can significantly affect how police responds. Racial disparities in weapon possession could be influenced by broader socio-economic conditions or differences in community safety perceptions. By examining the the color intensity, we can analyse certain races being more likely to be armed/unarmed. A higher concentration of one race being unarmed might indicate systemic biases in how people are perceived based on their race.



T1.5

From the visualization Fig T1.5, one key takeaway is that unarmed Black and Native American individuals appear to be more likely to be involved in police

shootings compared to White and Asian individuals. Another observation is that Black and White individuals involved in police shootings are more likely to have a gun on them, whereas this trend is less common among other racial groups, such as Asians and Hispanics. Additionally, about 37% of Asians caught in police shootings were carrying a knife, highlighting a different pattern of armed encounters compared to other races.

Race vs Fleeing

	Race				
Fleeing	Asian	Black	Hispanic	Native A..	White
Fleeing	25.64%	44.20%	40.12%	40.00%	33.92%
Not Fleeing	74.36%	55.80%	59.88%	60.00%	66.08%

% of Total Count o.. 25.64% 74.36%

T1.6

If certain racial groups are more likely to flee, it might reflect a lack of trust or fear of police interactions. From Fig T1.6, we can infer that Black, Hispanic and native Americans are more likely to flee during police encounters compared to Asian individuals who had tried fleeing only 25% of the times as compared to the about 40% for the other races. This could suggest that Asian individuals may experience police interactions differently, potentially due to factors like as cultural attitudes toward authority, differences in socio-economic conditions, or fewer negative prior encounters with law enforcement. Also this could imply that Asian individuals feel less threatened or less threatened or less inclined to escape in such situations, as compared to other racial groups.

Signs of mental illness vs race(percent)

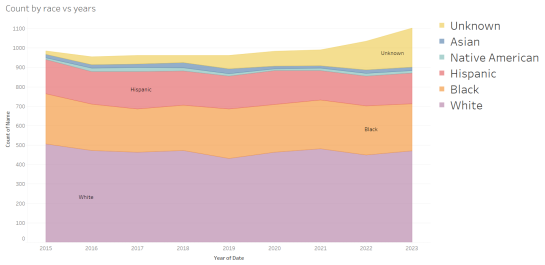
	Race					
Signs Of ..	Asian	Black	Hispanic	Native A..	Unknown	White
False	75.72%	85.27%	84.77%	84.73%	85.07%	74.40%
True	24.28%	14.73%	15.23%	15.27%	14.93%	25.60%

% of Total Count o.. 14.73% 85.27%

T1.7

The presence of mental illness signs across racial groups might reveal disparities in how mental health issues are recognized and addressed in different race groups. From Fig T1.7, we can infer that around 25% of Asian and White individuals involved in police shootings show signs of mental illness at a higher rate compared to 15% for Black and Native American individuals. This suggests that mental health may play a large role in police interactions with Asians and Whites. It could indicate that health issues are more frequently recognized, reported in these groups during police encounters how mental illness is perceived by law enforcement. The lower rates among Black and Native Americans might suggest under reporting or under diagnosis of mental health issues within the communities possibly due to cultural stigma.

Hypothesis T1.3: It is expected that the number of police shootings per year will rise in correlation with population growth. However, the rate of increase for each racial group is likely to vary due to evolving social dynamics and disparities that affect different communities in distinct ways over time.
This hypothesis is valid because population growth generally leads to increase in the number of interactions between law enforcement and public, which can result in a higher number of police shootings. However, the rate of increase in each racial group might differ due to factors like social and economic disparities, policing practices like racial profiling, historical experiences and cultural attitude towards authority.

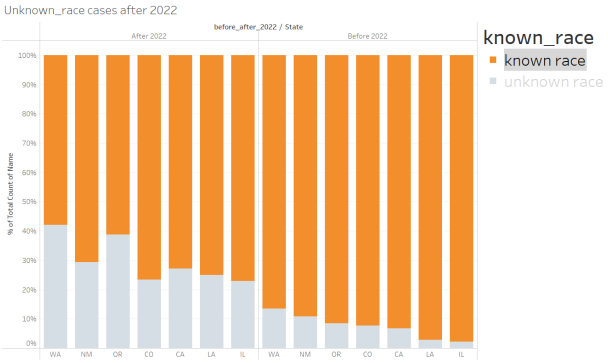


T1.8

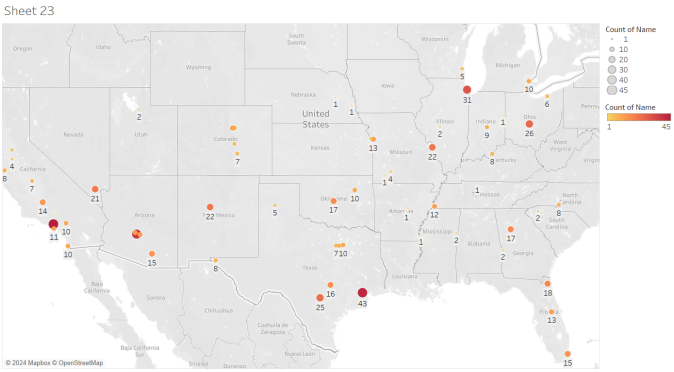
From the visualization Fig T1.8, we can observe a consistent rise in the number of police shooting cases in recent years. However, there has not been a significant increase among specific racial groups. In fact, there has been a slight decline in cases involving racial communities such as Hispanics. Notably, there is a sharp increase in cases categorized under ”Unknown” race. One possible explanation for this could be negligence or inconsistencies in reporting race data by police department in recent years, as the number of cases with unknown race has surged from 70 in 2021 to over 200 in 2023.
To investigate this trend further, it would be useful to analyze the percentage of cases classified as ”Unknown” both before 2021 and after 2022 in major states. This would help assess whether the increase in unknown racial data represents a broader issue in reporting practices.

From the above visualization in Fig T1.9, we can infer that in major states that report large number of cases of police shooting every year like Washington, California and Los Angeles have significant percentage increase in the number of unknown race cases of police shooting after 2022. This clearly highlights the significant negligence within the police system in maintaining accurate records of cases involving police shootings.

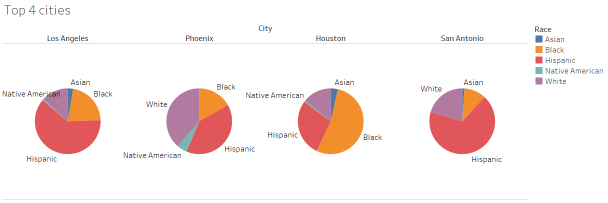
The cities with the highest number of police shooting cases include Los Angeles, Phoenix, Houston, and San Antonio. The racial distribution in these cities is



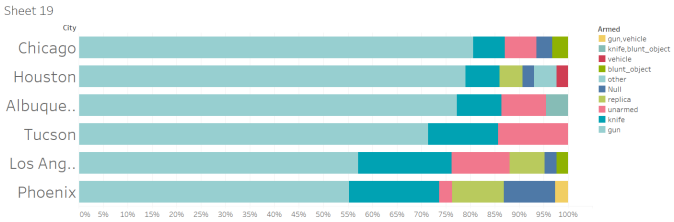
T1.9



T1.11



T1.10

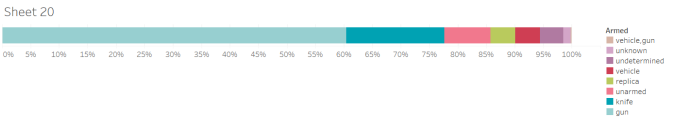


T1.12

visualized above in Fig T1.10 Arizona, California, and Texas, in particular, report elevated numbers of police shootings, likely due to stricter policing in response to their large immigrant populations. Wealthier white individuals are less represented in these records, possibly due to racial profiling and inherent biases in the police force that disproportionately affect non-white communities.

Hypothesis T1.4: How does the involvement of youth in police shootings vary across different cities?
Analyzing the variation in youth involvement in police shootings across different cities helps identify potential factors that contribute to these incidents, such as differences in policing practices, socio-economic conditions, crime rates, or local policies. Understanding these patterns can inform policy recommendations and interventions aimed at reducing the number of shootings and addressing systemic issues affecting youth in specific regions.

The visualization in Fig T1.11 shows a heatmap of police shooting cases involving individuals aged 14 to 28. From this, we can infer that cities like Los Angeles, Houston, Chicago, and Phoenix have a notably high number of youth-involved cases. In Los Angeles, this may be attributed to its large population, while Chicago and Houston are known to struggle with youth gang activity, which could explain the elevated number of police shootings involving young individuals in these areas.



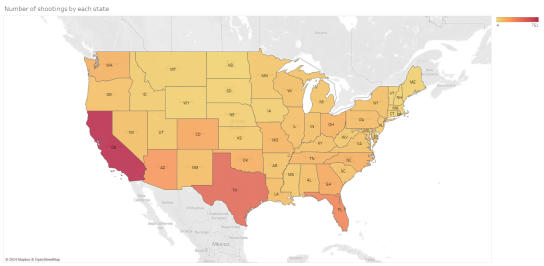
T1.13

Analyzing the youth data from the above visualization, Fig T1.12 and comparing it with the all cities youth data in Fig T1.13, we can infer that cities like Chicago, Houston, and Albuquerque have a high percentage of youth involved in police shootings who were carrying firearms. In Chicago, this is likely tied to youth gang activity, while in Albuquerque and Houston, weaker gun control measures may contribute to the higher involvement of armed youths in these incidents.

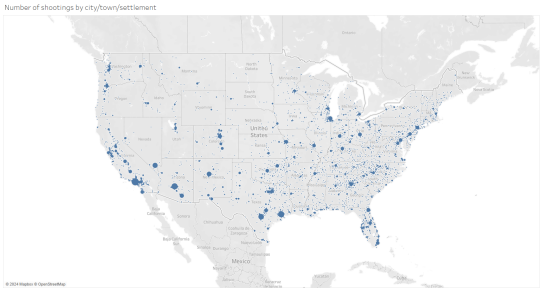
T2 Geopolitical Analysis:

Hypothesis T2.1: The first hypothesis is that areas of police shootings coincide with the areas of higher population, concentrated around cities and mainly in states with higher population as well.
The intuition behind this hypothesis is simple; areas with more population density tend to have both higher policing as well as higher crime rates, as well as incidents in general.
We attempt to verify the hypothesis by visualizing the shootings in map form, both by the state as well as the

hot-spots of the shooting incidents, as shown respectively in figures T2.1 and T2.2. We can also verify the same by plotting bar graphs that showcase the number of cases for each state, and the scatter plot of the number of cases against the fraction of the national population that the state has. These visualizations can be seen in figures T2.3 and T2.4.



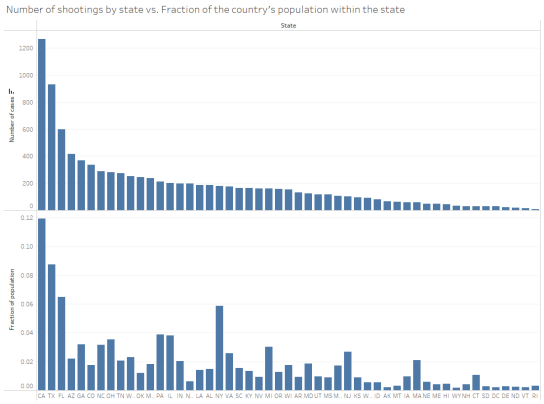
T2.1



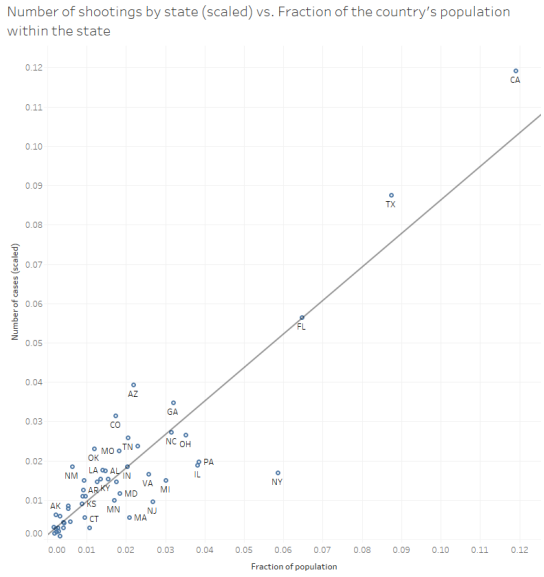
T2.2

As we can see from figure T2.1, the states with more population tend to have more cases, and the density of cases in cities and well-settled areas is seen in figure T2.2 - where, in the Midwest, the low population density results in lower cases, whereas in areas like the Pacific coast and the Atlantic coast, it follows the cities. In the Mississippi and Missouri basins, the population is evenly distributed, and there are not too many large cities - this means that the cases are more evenly distributed.

Figures T2.3 and T2.4 further confirm this hypothesis, where the lengths of the bars in figure T2.3 largely coincide for each state. In figure T2.4, we also see that there is a clear trend line, with a few notable exceptions on both sides. (We can explain the clustering of values near the origin of the scatter plot due to a majority of states in the US having far less population than the likes of California and Texas, as well as almost proportionally fewer cases of shootings.) We will further analyze these states in later hypotheses.



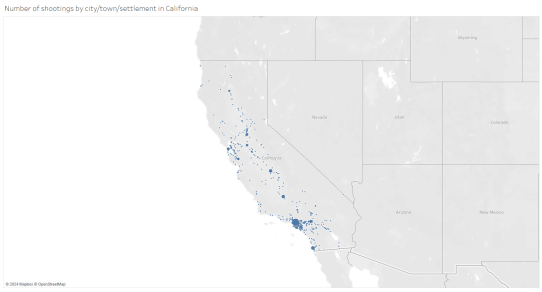
T2.3



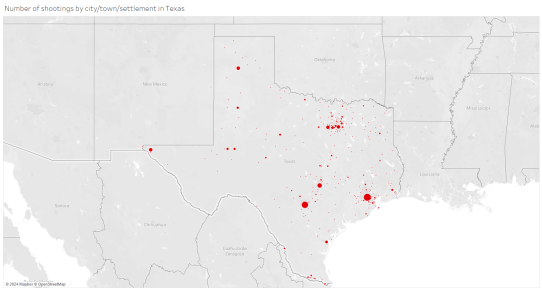
T2.4

Hypothesis T2.2: The second hypothesis is that political alliance (red or blue) plays a role in the demographics of the suspects that encountered police shootings and/or the situation(s) in which cases occur.
Context for Hypothesis T2.2: Red states refer to those states which have predominantly voted for the Republican party since 2000, while Blue states refer to those states which have predominantly voted for the Democratic party since 2000. Texas and California (henceforth abbreviated as TX and CA respectively) are the two states with the largest population, as well as states that have voted Republican and Democratic consistently for the last 5 elections respectively.

To analyze this hypothesis, we will look at visualizations pertinent to CA and TX. Figures T2.5 and T2.6 show the geographical distribution of cases in CA and TX respectively. In California, we see that the cases are

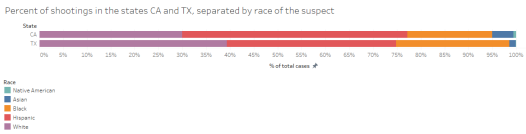


T2.5



T2.6

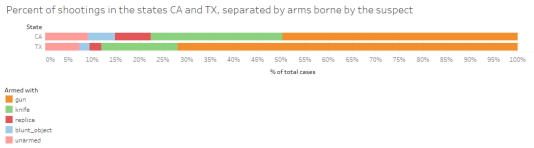
more concentrated towards the southern end of the state (where the Hispanic population is higher), whereas in Texas we see that the cases are mainly concentrated in the four major cities of Houston, Dallas, San Antonio and Austin.



T2.7

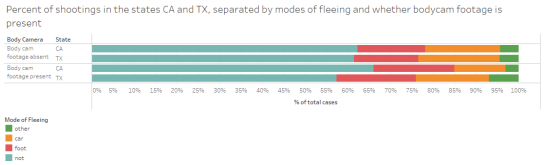
The stacked bar chart T2.7 shows the ratio of shootings in each of the two states segregated by race. There is a noticeably higher proportion of Hispanic civilians shot in CA, whereas the same can be said about Black people in TX. The skew towards Hispanic shootings in CA may be due to the higher proportion of people of Hispanic descent in southern California (where the cases are more rampant), whereas the higher ratio of Black people being shot in TX can be explained by the cases being concentrated in more urban areas. An alternate explanation for the higher Black cases in TX could be the more conservative nature of Texas, and racial profiling done by the police departments in these cities.

The stacked bar chart T2.8 shows the ratio of shootings in each of the two states, segregated by whether the suspect was armed in the encounter. There is not



T2.8

a significant difference in the number of unarmed suspects, but the stark difference in the number of gun-bearing suspects can be attributed to the more lax gun laws and the more commonly-available firearms in Texas.



T2.9

The stacked bar chart T2.9 shows the ratio of shootings in each of the two states, segregated by whether the suspect attempted to flee (and whether there was body cam footage present). Whenever body cam footage is absent, there is not a significant difference in the number of suspects who did not attempt to flee (or the distribution of the means of fleeing used by the same); however, for cases where body cam footage is present, the number of suspects not fleeing is much higher in California. This can be attributed to either false case reports submitted by the police in CA (where they fabricate that the suspect is fleeing) or the lack of ability in police officers to defuse the situation.

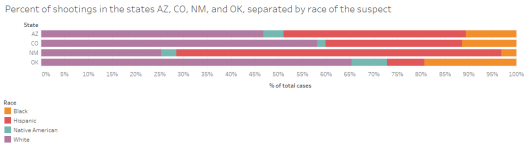
Hypothesis T2.3: The third hypothesis is that there is some geographical relation between the density of cases in each state, at the extreme values (low and high case density).
The 'density of cases', as mentioned before, refers to the number of cases of police shootings in the state, divided by its fraction of the total population.
This hypothesis can easily be inferred by looking at the graphs T2.3 and T2.4, which tell us that the notable outlier states are:

- *higher density*: Arizona, Colorado, New Mexico, Oklahoma
- *lower density*: Connecticut, Massachusetts, New Jersey, New York

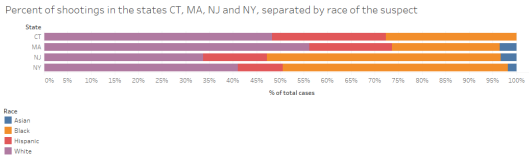
We can see that the states with lower density (CT, MA, NJ, NY) are all states that are located in either New England or the Tri-State Area, both of which are in the north-east of the US and are on the Atlantic Coast. On the other hand, the states (AZ, CO, NM) are in

the south-west of the US near Mexico. The state (OK) is an outlier, which is located in Central US, but still bordering both CO and NM. Hence, there is a clear correlation between the geographical location of the state and its tendency to be an outlier in the trend highlighted by the scatter plot in figure T2.4.

Hypothesis T2.4: The fourth hypothesis is that there are some common characteristics between the outlier states (AZ, CO, NM, OK) and (CT, MA, NJ, NY), on the basis of race and situation of the shooting cases.
We use the stacked bar charts in figures T2.10 through T2.15 constructed in a similar manner to the previous figures T2.7, T2.8, and T2.9 used to analyze California and Texas.

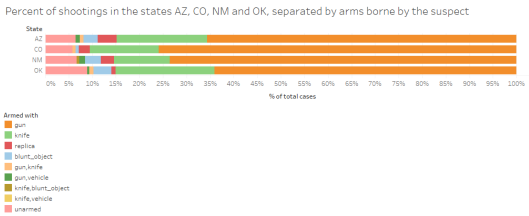


T2.10



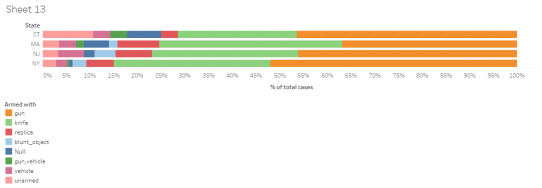
T2.13

In figure T2.10, which contrast the distribution of cases by race, we see no clear trend in the high case density states. These seem to reflect the population distribution by race of the states themselves.
However, in figure T2.13, we see a higher proportion of Black shootings in the states of NY and NJ, which may indicate racial profiling in these two states.



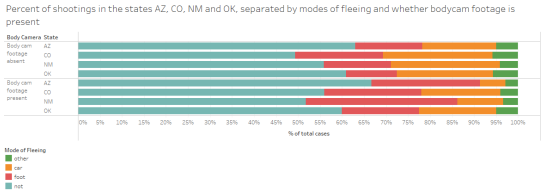
T2.11

In figures T2.11 and T2.14, we see a lot more suspects armed with guns in the high case density states, as well as a lot of unarmed suspects. The higher proportion of gun-bearing suspects in these states can be attributed to the more liberal gun laws in the same states, and the

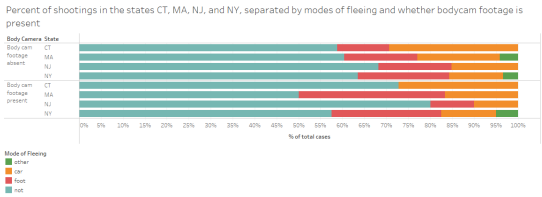


T2.14

slightly higher proportion of unarmed suspects indicates a more belligerent nature of police officers in the same states. The exception to the same is CT, where roughly 10.71% of the suspects shot are unarmed.



T2.12



T2.15

In figures T2.12 and T2.15, we see fewer people who are fleeing in the high case density states than the lower case density states. This can have one (or both) of the following conclusions:

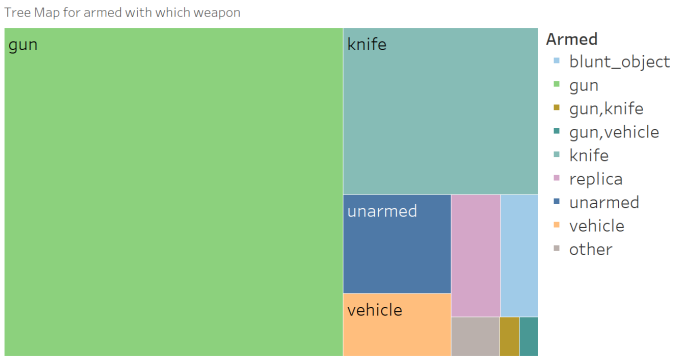
- Police officers in the low case density states are worse at defusing tense encounters, and tend to resolve cases by opening fire;
- Police officers in the low case density states tend to be more cautious about opening fire at suspects who have fled, in order to avoid collateral damage.

This concludes the hypotheses that can be made about the geopolitical aspects of the cases of police shootings in the US.

T3 Contextual Analysis:

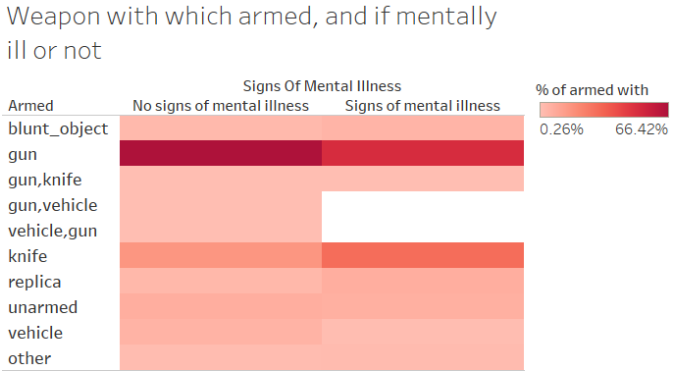
Hypothesis T3.1: The most obvious context for a police shooting is if the suspect was armed, so the more deadly the weapon on the suspect, the likelier they are to be shot at, so, the ratio of suspects with guns should be highest, and those unarmed should be lowest.
This hypothesis is somewhat verified by Fig T3.1,with 63.26% of total suspects being shot having guns, and only 6.06% suspects being unarmed. However, there also

are suspects that are armed with other weapons, present in a lower ratio than those unarmed.



T3.1

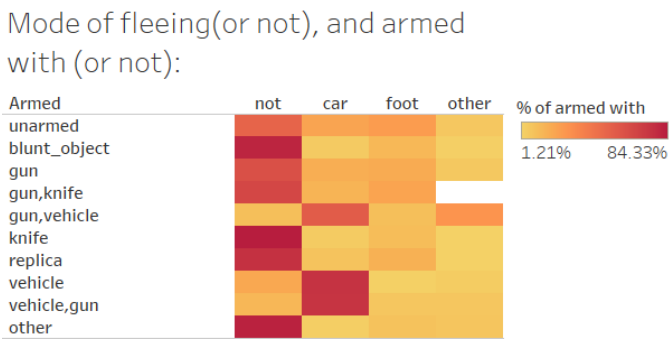
We can further look into this data, as our dataset also gives us information on whether the suspect was deemed mentally ill by the police officer or not, so we look at Fig T3.2, which tells us that the distribution of choice of weapon is similar across both those deemed mentally ill and not, with the only exception being that those deemed mentally ill were only armed with vehicles at 0.87%, whereas vehicle armed and not mentally ill suspects add up to 5.22% of their total.
However there can be no more inference made from this information other than: suspects who were deemed mentally ill have very low tendency to be armed with vehicles.



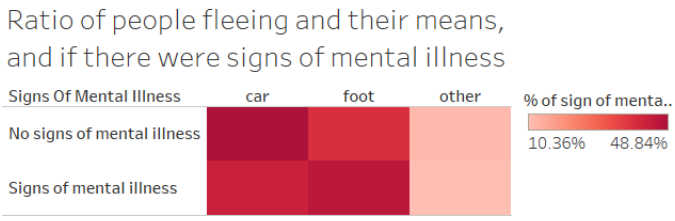
T3.2

Hypothesis T3.2: Since we have data on what suspects were armed with, and whether they attempted to flee, and if they had signs of mental illness, one obvious hypothesis we can make is that those armed with vehicles attempt to flee, and since we see from the inference of *Hypothesis T3.1* that those with mental illness are not armed with vehicles, they are less likely to flee with car, and are likelier to flee on foot or other means.
From Fig T3.3, we can see that the obvious statement of our hypothesis is true, and those armed with vehicles tend to flee in them.

We also see in Fig T3.4, that those that did exhibit signs of mental illness do tend to flee on foot rather than or car. So our *Hypothesis T3.2* is also accurate, as suspects armed with vehicles do tend to flee, and mentally ill suspects tend to flee on foot rather than in cars.

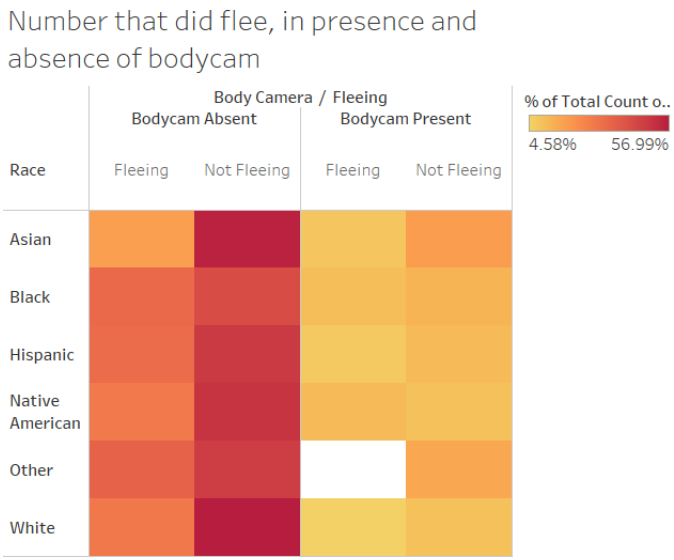


T3.3

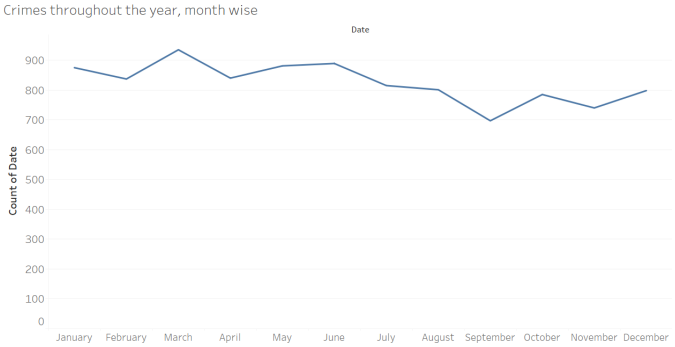


T3.4

Hypothesis T3.3: Since we have data on bodycam footage presence, we can hypothesize that when bodycam footage is present, suspects choose to not flee, as they feel more secure with the police even though they have committed a serious crime, and they would rather flee when the police have no bodycam on them.
We can choose to plot this data segregated by race, so that we can see how comfortable each race is in the context of presence of bodycam on the officer. We can see in Fig T3.5, that our hypothesis is completely wrong and that the presence of bodycam on the officer does in fact not make lesser people to attempt to flee.
The ratio of people fleeing is equal to those not attempting to fleeing under both columns of bodycam present and absent, with the exception of Asians.
We also see that a greater portion of the dataset does not have bodycam present, so we can also analyse bodycam footage throughout the months of the year, to see if we can make an inference on it, and looking at T3.6, we can see that the number of bodycam footages present across the year more or less the same, however, there is a spike in March, and a dip in September, in the number of cases where bodycam footage is absent.
However this can easily be attributed to the observation that there are simply more crimes or less crimes committed in these months of the year, which can be verified by



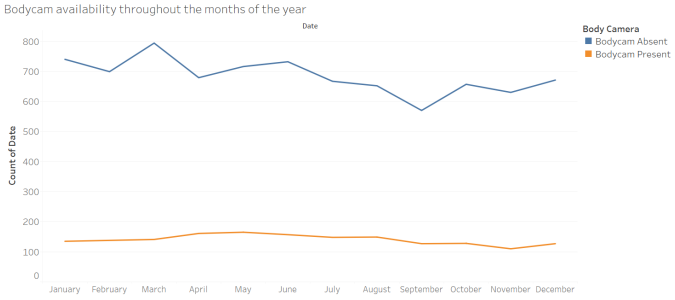
T3.5



T3.7

T3.7. So, we can not make an inference on the absence of bodycam footage.

Our dataset is however limited to suspects that are fatally shot, so the suspects that flee successfully are not counted for in this dataset, so we can not successfully make an inference on successfully fleeing, in presence or absence of . Conclusion: our hypothesis, *Hypothesis T3.3* is wrong.



T3.6

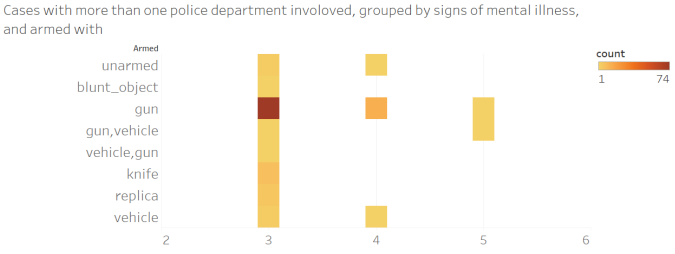
Hypothesis T3.4:We see in our dataset that certain cases have more than one police department involved per case. This is unusual for a typical case of a police shooting, so it must mean that the scale of the crime was greater, or the suspect was deemed highly dangerous, or maybe was a known criminal. We can hypothesize that: The greater the number of police departments involved, the more serious the crime.

We can verify this hypothesis with a heatmap of cases with what a suspect was armed with against the number of police departments involved.

We can see in T3.8 that in most cases which have more than 2 police departments involved, the suspect was armed with a gun.

We do have an outlier in the dataset with there being a case each where 4 police departments were involved for one suspect being armed with a vehicle, and another being unarmed.

Our hypothesis, *Hypothesis T3.4* was correct, as the only cases where more than 2 police departments were involved, were because the suspect was armed with a gun, and for other ways of being armed, there are only upto 2 police departments involved per case.



T3.8

VISUALIZATIONS

Following are the visualizations that are used and described in detail in the section above.

1) Pie Charts
2) Area charts
3) box and whisker plots
4) Stacked bar charts
5) Heat plots
6) Symbol plots
7) Line plots

Also in each of these plots/charts we have employed various marks for making the visualizations more expressive.

MEMBER WISE CONTRIBUTIONS

T1: Dhruv Kothari
T2: Harsh Modani
T3: Mohammad Owais

We independently came up with initial hypotheses for our task, and cross verified with each other for correctness. Additionally, insights derived from one another's tasks were utilized to inform and refine individual analyses. This collaborative approach facilitated a more comprehensive interpretation of the dataset in each task, incorporating perspectives from each task.