Data Visualization A3

Dhruv Kothari
IMT2022114
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/after_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
4) Crime rates: Crime rates over time
5) Socio-economic data: Socio-economic data like poverty per capita income and higher education rates of different states overtime.
6) unemployment rate: unemployment rate in different years.
7) demographic dataset: This dataset provides population estimates from 2020 to 2023 for various demographic groups in states, categorized by sex, age, and race.

TASK

We had presented our hypothesis and analysis on different aspects of the dataset in A1 report. This report is an extension to that report that was submitted for assignment-1, including visual workflow and a feedback mechanism to make more accurate and conclusive inferences. Overall it's again divided into 3 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.

Visual Analytics Workflow

The workflow for A3 continues directly from the analyses we have made for A1, and we make further explorations from A1.
We first analyze the data using statistic and machine learning methods, and implement a visual analytic workflow that relies on creating new visualizations from the knowledge gained from the inferences of the previous visualizations we have made.

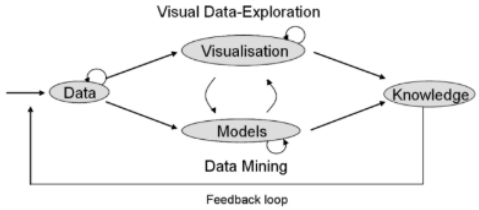


Fig. 1: Kiem et al. Visual Analytics Workflow, Image tesy:

Visualizations

Following are the visualizations that are used and de 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

In each of these plots/charts we have also employed marks to make 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.

Appendix

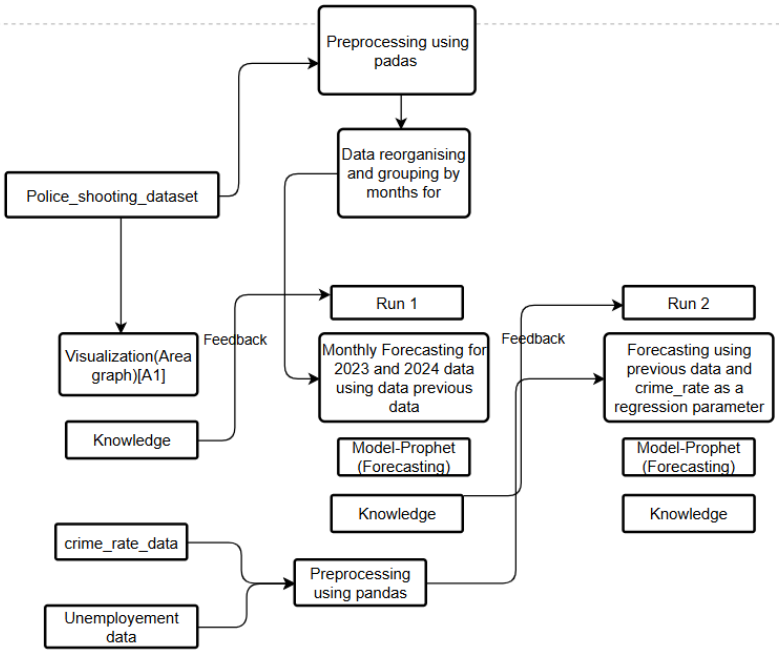
For the sake of completeness, the report for Assignment A1 has been added as an appendix at the end of the document.

Data Stories

T1 Demographic Analysis:

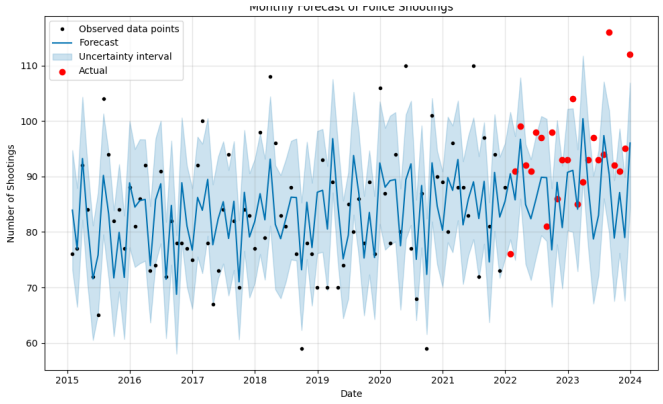
1) Visual analytic workflow T1.1

According to the hypothesis stated in A1, in appendix, under T1, there is a variation in police shootings over time, accompanied by mild fluctuations across different racial groups. These variations are attributed to evolving social dynamics and disparities that uniquely impact various communities over time. The inference in A1 was drawn using an area chart that visualizes these variations across different racial groups. To better analyze the factors, we design a workflow as follows, as depicted by Fig T1.1



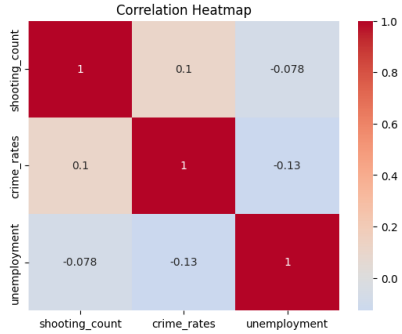
T1.1 A sketch of the Visual Analytics Workflow diagram for time-forecasting

First, we preprocess the data by grouping it by months. Based on the analysis and visualizations from Assignment A1, we plot a forecast of police shootings for future data, specifically for the years 2022–2023, using historical data as the training set. Additionally, we superimpose a scatter plot of the actual future data available to compare it with our forecast and analyze whether there is any growth or decline in the trend. On the initial run, the forecast is depicted by Fig T1.2
For the overall data, as shown above, the forecast, based solely on previous statistics, predicts only a small rise in the number of police shooting cases. However, when comparing this forecast with the actual data, we observe that some data points fall outside the confidence interval boundaries of the forecast.
In the second run, we explored additional dependent variables to improve the prediction of future police shooting data. Specifically, we incorporated two datasets: crime rate over



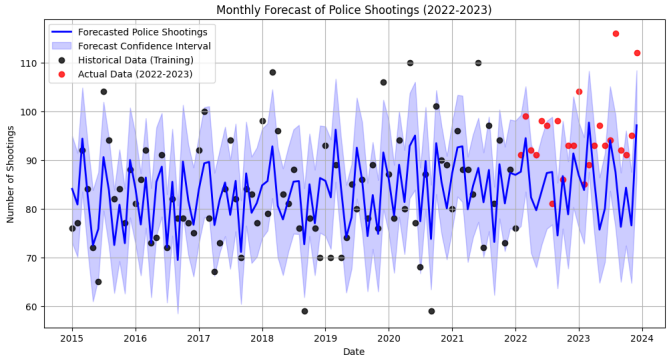
T1.2 forecast model trained on previous data

time and unemployment rate over time, while also examining the correlation between these parameters. Although crime rate inherently accounts for unemployment rate to some extent, we further analyzed their relationships by plotting a correlation heatmap, as shown in Fig T1.3.



T1.3

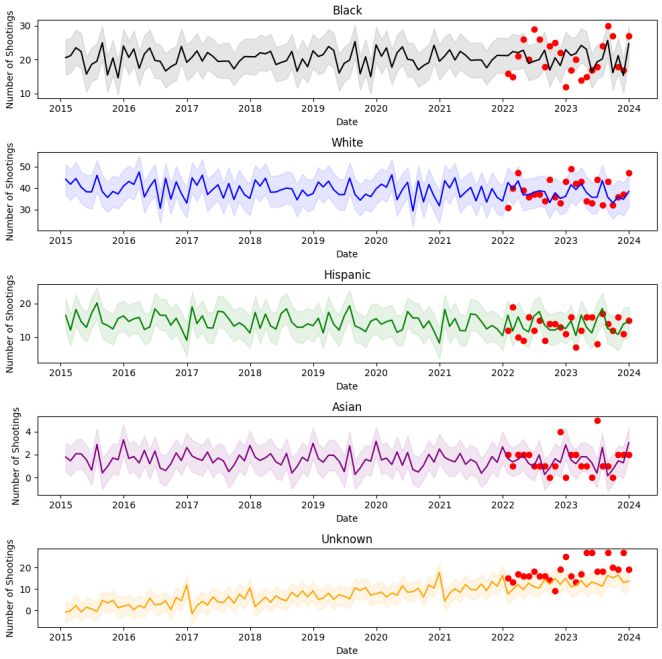
We integrated the crime rate and unemployment rate into the time forecasting model to enhance its accuracy and improve the prediction of future trends. The results of this updated model are illustrated in Fig T1.4.



T1.4 Forecasting model trained on previous data along with crime and unemployability rate, and predicting future on just crime and unemployability rate.

There is minimal change in the plot, except that the forecasted line and the confidence interval now better accommodate the actual data points. Additionally, the confidence interval becomes narrower, which indicates improved precision. This aligns with our hypothesis that police shootings have shown a slight increase over the years.

Also trying out forecasting model for different racial groups based on similar parameters to analyse the changes in the police shooting with time, we obtained the visualizations shown by Fig T1.5



T1.5 Forecast model on different races

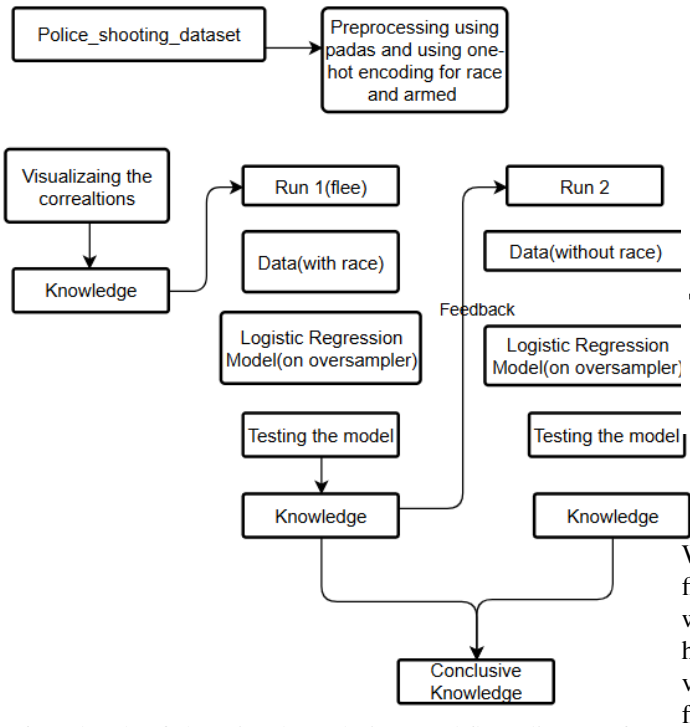
The inference one can make could be is that the police shooting hasn't changed much over time for most races, although they have seen a small dip in the year 2020-21, could be beacause of COVID lockdown. But according to actual data had increased by a bit during some months in the recent years.

An interesting inference stated in Assignment A1 can be confirmed: the number of unknown cases has risen significantly over the years. Notably, in recent years, the actual data shows a sharp increase, exceeding the forecasted values. This trend could be attributed to various factors, such as incomplete datasets or potential negligence by law enforcement in properly recording these cases.

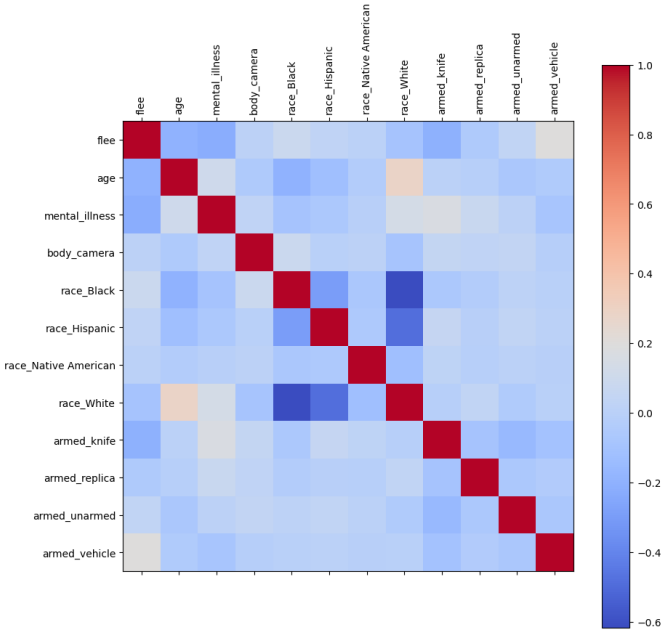
2) Visual analytic workflow T1.2

According to the hypothesis states under T1 in A1 assignment, there is a correlation between race and factors such as being armed, fleeing, or exhibiting signs of mental illness. Our analysis revealed a relationship between gun ownership and race. Among individuals shot by police, The Whites and Blacks show approximately 65% gun possession,

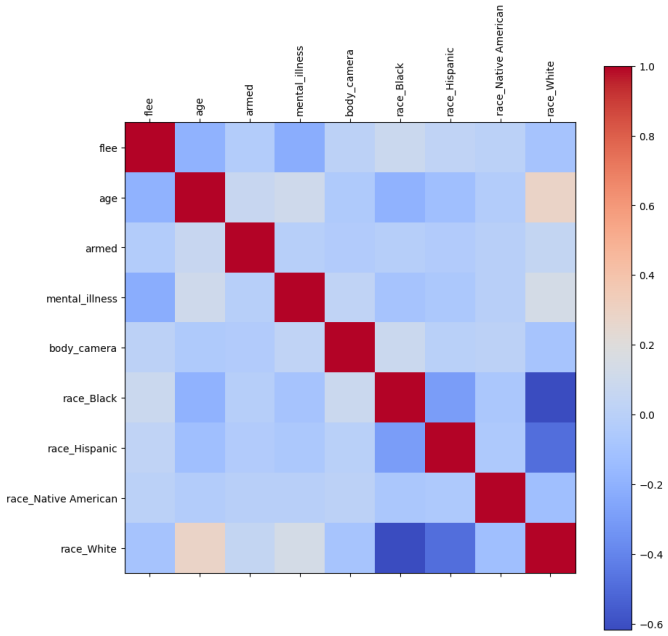
whereas Hispanics and Native Americans have lower rates of gun possession at the time of the incident. To further analyze the armed status and fleeing behavior, we propose a visual workflow to better understand the hypothesis. The workflow is illustrated in Fig T1.6.



T1.6 A sketch of the Visual Analytics Workflow diagram for Analysing race influence infactors like flee or armed



T1.7 Correlational Heatmap(Analyse race influence on flee)



T1.8 Correlational Heatmap(Analyse race influence on flee)

We began by one-hot encoding the races and the 'armed' field, followed by plotting correlation heatmaps to analyze whether race influences fleeing behavior or being armed. The heatmaps are presented in Fig T1.7 and Fig T1.8. From these visualizations, it is evident that the influence of race on both fleeing behavior and being armed is close to zero. To validate this observation, we follow the workflow diagram, training a logistic regression model both with and without race as a feature, and compare their scores to test and validate the hypothesis.

```
Accuracy: 0.68

Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.62      0.66      1317
           1       0.66      0.75      0.70      1316

    accuracy                           0.68      2633
   macro avg       0.69      0.68      0.68      2633
weighted avg       0.69      0.68      0.68      2633


ROC-AUC Score: 0.73
```

T1.9 Score of Logistic regression model, for flee trained on data with race

We trained logistic regression models to predict whether an individual flees or not, using oversampled random data to mitigate skewness in the training set. One model included race and armed features (both one-hot encoded), while the other excluded race but retained one-hot encoded armed data. The model with race achieved an accuracy of 0.68, while the model without race achieved an accuracy of 0.69. This negligible difference in accuracy indicates that the correlation

```
Accuracy: 0.69

Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.63      0.67      1317
           1       0.67      0.74      0.70      1316

    accuracy                           0.69      2633
   macro avg       0.69      0.69      0.68      2633
weighted avg       0.69      0.69      0.68      2633


ROC-AUC Score: 0.73
```

T1.10 Score of Logistic regression model, for flee, trained on data without race

between race and the likelihood of fleeing is minimal, as the model performances are nearly identical.

```
Accuracy: 0.55

Classification Report:
              precision    recall  f1-score   support

           0       0.55      0.57      0.56      1855
           1       0.55      0.54      0.55      1854

    accuracy                           0.55      3709
   macro avg       0.55      0.55      0.55      3709
weighted avg       0.55      0.55      0.55      3709


ROC-AUC Score: 0.58
```

T1.11 Score of Logistic regression model for armed, trained on data without race

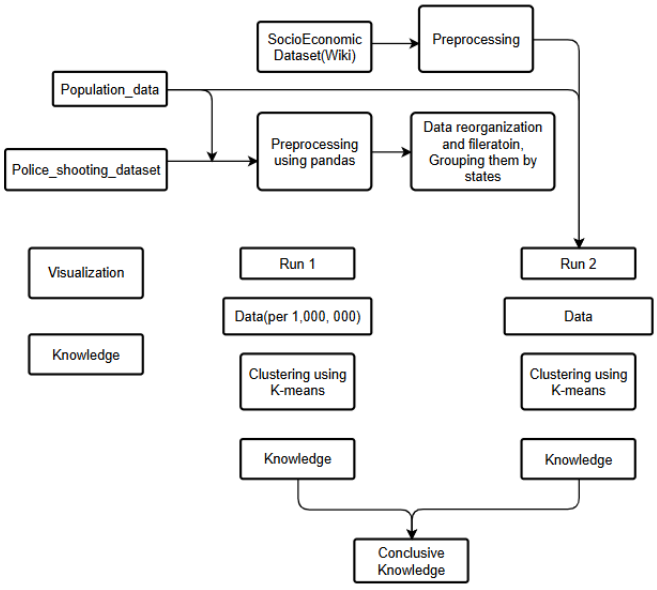
```
Accuracy: 0.56

Classification Report:
              precision    recall  f1-score   support

           0       0.56      0.56      0.56      1855
           1       0.56      0.56      0.56      1854

    accuracy                           0.56      3709
   macro avg       0.56      0.56      0.56      3709
weighted avg       0.56      0.56      0.56      3709


ROC-AUC Score: 0.59
```

T1.12 Score of Logistic regression model for armed, trained on data without race

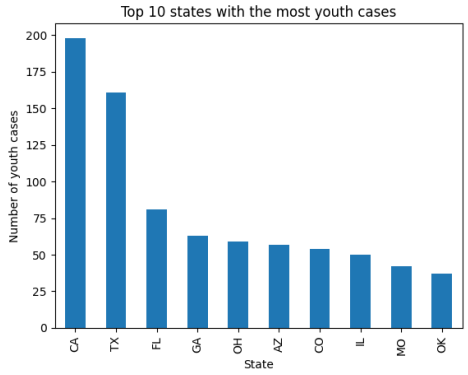
Similarly, we repeated the process to predict whether an individual was armed, treating 'armed' as a binary variable. Logistic regression models were built following the same approach, with one model including race (one-hot encoded) and the other excluding it. The results showed that the accuracy and ROC scores for both models were very similar. This further supports the conclusion that race has a negligible relationship with the likelihood of being armed or not.

3) Visual analytic workflow T1.3

According to the hypothesis states under T1 in A1 assignment, some cities have higher number of police shootings of youth than the others due to varied difference in socio-economic status of the states and also the political condition of the state. Analysis from A1 dataset revealed that states like California and and Texas have very high number of police shooting cases. A possible reason for this is both these states being densely populated. Also Texas has high hispanic population, which could be a potential indicator of difference in social conditions than other states. We are gonna use the workflow diagram shown in to do our analysis on the data.



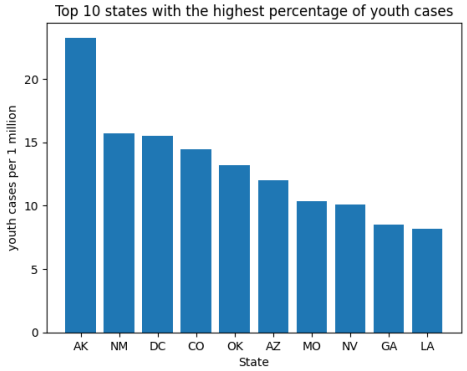
T1.13 A sketch of the Visual Analytics Workflow diagram for youth police shooting



T1.14 States with highest number of police shooting cases.

Fig T1.14 highlights the states with the highest number of police shooting cases involving youths (25 years old). California and Texas together account for 25% of the total cases, primarily due to their large populations, significant social disparities, and, in the case of Texas, permissive gun laws. An alternative analysis could focus on the number of cases normalized by population size to provide a more proportional perspective.

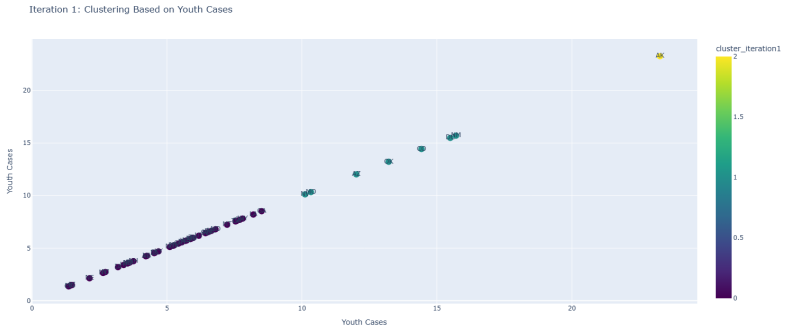
Fig T1.15 shows the states with the highest number of police shooting case per 1 million population. This is a more



T1.15 States with highest number of police shooting cases per 1 million population.

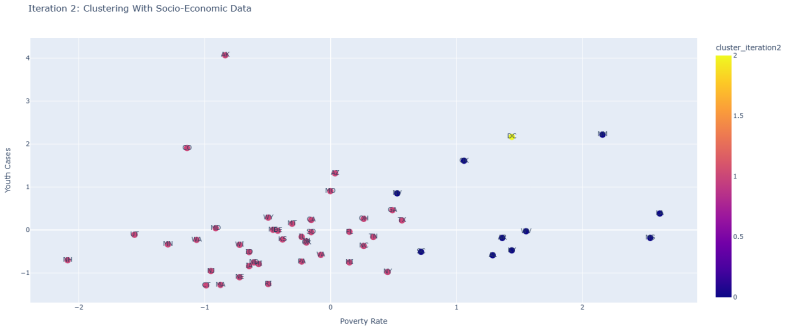
normalized measure for doing analysis. This could majorly be due to small population, so having just a few cases could increase the ration significantly. Also District of Columbia has frequent police activities, could contribute to higher police interaction and hence higher police shooting cases. Also states like New Mexico, Colorado and Alaska also have highrates of oppoid addiction and alcohol abuse. These states also have high poverty rate.

We'll run a clustering model to split to cluster based on police shooting per millions and another model with police shooting per million and other socio-economic factors like per capita income, poverty rate, without health insurance rates. We'll form clusters and check for similarities in the clusters.



T1.16 Clustering only on the number of cases

By comparing Fig T1.16 and Fig T1.17, we observe that states such as New Mexico, the District of Columbia, Oklahoma, and Alaska consistently belong to the same cluster in both plots. This consistency indicates a strong relationship between police shootings and socio-economic factors. These states likely share similar underlying characteristics, such as higher social disparities, economic challenges, or differences in law enforcement practices, which contribute to the clustering observed. Thus, we can conclude that socio-economic factors play a significant role in influencing the prevalence of police shootings.



T1.17 Clustering on Socio-economic factors

T2 Geopolitical Analysis: We will establish a visual analytics workflow for the geopolitical aspects of shooting cases and, through the use of the feedback loops and external datasets used alongside the original data, draw insights about the same.

Visual analytic workflow T2.1

- *External Datasets Used*
  The following external datasets were used in the workflow:
  - A dataset of the populations of various US states, and the fraction of the national population resident in the state.
  - A dataset containing a list of gun laws, along with the laws upheld by each state.
  - A dataset containing the political alignment of the states of the US.
- *Analysis of the number of cases in each state*
  One of the most basic geographical analyses we can run on this dataset is counting the number of police shootings recorded per state. This can give us a broad idea of the distribution of cases among the states of the US and analyze areas of higher cases, if any such emerge from the visualization.

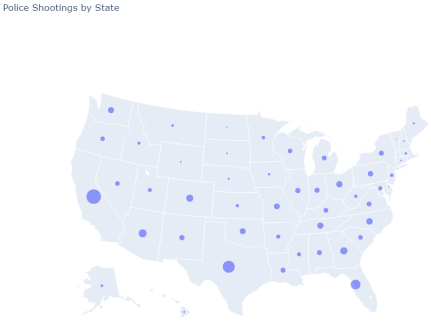


Figure T2.1

Figure T2.1 shows the scatterplot created by the number of cases in each state. From this scatterplot, it seems to be

the case that the number of shooting cases in each state is roughly proportional to the population of the state itself. We can run some analysis on the population of each state to confirm this observation.

- *Analysis of the number of shootings versus the population of the state*
  For this part of the workflow, we input new data in terms of the state-wise population dataset. The dataset we use in this section is a dataset of state population data, where each entry has `2020_census` and `percent_of_total` attributes that respectively correspond to the population of the US state as estimated by the 2020 census and the percentage of the total US population resident in the state.
  A basic method to analyze whether there is any correlation between state population (or rather, the percentage of the total population, which is equivalent since the two values are proportional) is to plot bar graphs of both next to each other and compare the outcomes.
  The bar graphs are plotted as shown in Figure T2.2. As we can see from the bar graphs, the states are fairly similar in the fraction of national population and the number of police shootings recorded. However, the two are not entirely consistent; to study this in more detail, we will perform clustering on these states.
- *Clustering states by fraction of population and police shootings in the state*
  Here, we attempt to cluster states into 4 categories based on the number of police shootings in the state and the fraction of the national population resident in the state.
  We plot the states by the cluster assigned to them in a scatterplot, with the X-axis as the number of police shootings in the state and the Y-axis as the fraction of the national population resident in the state. Figure T2.3 visualizes the same scatterplot.
  We note that the states in the yellow cluster with overall low police shootings and population fraction concentrate in one area and form a meaningless cluster. However, since the dimensionality of the dataset is low, we can see that the equivalence diagonal is a principal component for almost all points here. We can thus perform another K-Means clustering with the equivalence diagonal as one principal element and the unit vector normal to the diagonal as the other.
  Figure T2.4 visualizes the result of the K-Means clustering algorithm with the chosen principal components. This time we get a much more desirable result; the two clusters, one containing California and Texas and the other containing New York are both used for filtering outliers. The rest of the states fall into two clusters, which can be separated entirely by Y = 1.3. These clusters correspond to low and high **case density**, where case density is the number of cases in the state divided by the population fraction of the state.
- *Correlation of case density of the state with gun laws in the state*
  Most shooting cases occur when the victims are armed with firearms or guns. Thus, it would be useful to look into states and see the number of police shootings per capita and the laxness of gun laws in the state and try to find a correlation between them.
  The expected outcome is that since criminals armed with firearms are more present in states with more gun laws, there should be a positive correlation between the laxness of the laws in the state and the number of shooting cases per capita.
  The dataset we use here is a dataset that stores a list of gun strictness laws, and each entry is whether a state has any provision for the same, as of 2019. There are 134 such laws, and the final total is the count of the number of laws that the respective state upholds. This is the **strictness score**; the **laxness score** can be calculated simply as `134 - strictness_score`. We will try to cluster states by their gun laxness and their case density. Figure T2.5 shows the clustering output as well as the respective scatterplot. We notice from the graph here that states with higher gun laxness may or may not have a high case density, but states with lower gun laxness definitely have a lower case density. This indicates that having stricter gun laws is usually a successful preventive measure for police shootings.
  Moreover, in our A1 hypothesis, the points we had chosen for low case density (CT, MA, NJ, NY) are all present in `cluster = 1` (states with low case density as well as low gun laxness); and the point we had chosen for high case density (AZ, CO, NM, OK) are all present in `cluster = 2` (states with high case density as well as high gun laxness).
  We extend the same hypothesis to the other states in the clusters `[1, 2]` and perform further analysis on those states.
  Figure T2.6 shows the states in the clusters formed by states of low case density, low gun laxness and high case density, high gun laxness respectively. We will attempt to analyze the respective states' political alignment in order to see whether it corresponds with any trends in the cluster or not.
- *Analysis of states with low case density and low laxness against states with high case density and high laxness*
  Political alignment refers to the state's overall tendency to vote **Democratic** (blue) or **Republican** (red) in the presidential elections.
  We have used the political alignment dataset that stores the **alignment** of each state in the form `2 - (no. of times the state has voted blue in the last 4 elections)`. For example, if a state voted for the Republican party in 3 of the last 4 elections, their alignment will be `2 - 3 = -1`.
  Figure T2.7 shows the political alignment of each state in the US. Crimson states are those states that have voted Republican in all of the last four presidential elections; light red are those states that have voted Republican

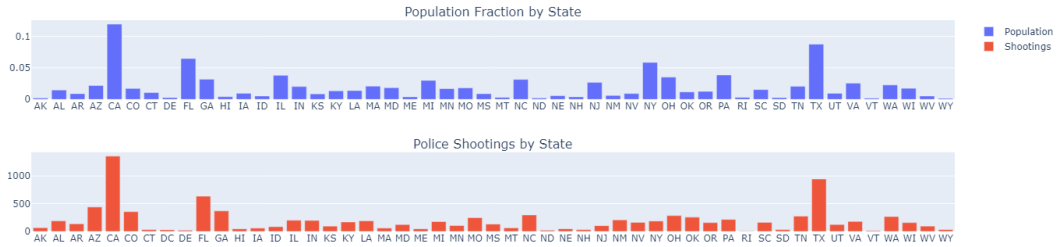


Figure T2.2

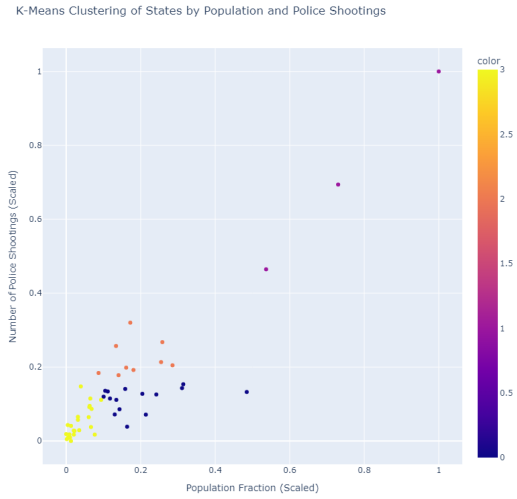


Figure T2.3

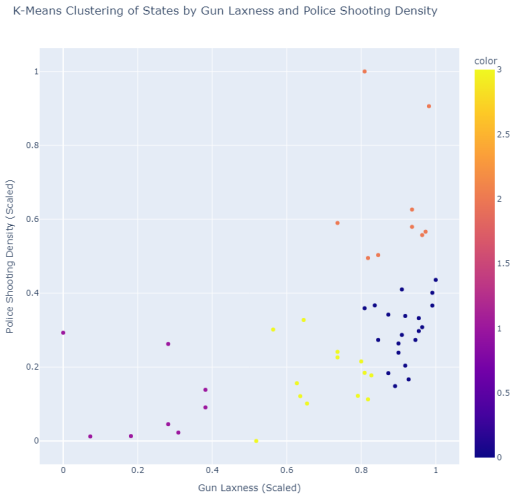


Figure T2.5

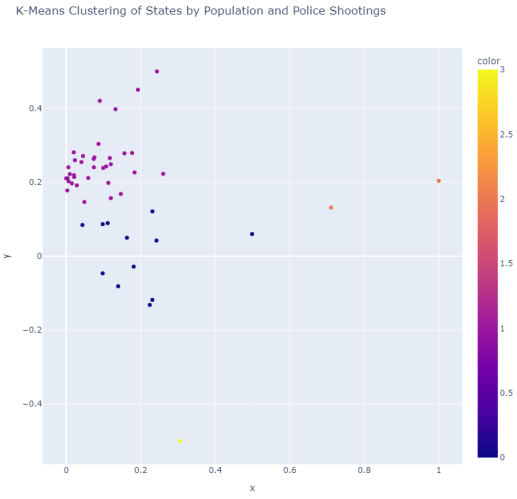


Figure T2.4

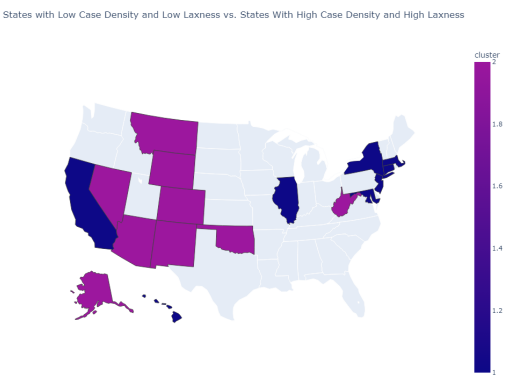


Figure T2.6

in three out of the last four presidential elections; light blue are those states that have voted Democratic in three out of the last four presidential elections, and blue are those states that have voted Democratic in all of the last four presidential elections. The white states, also known as swing states, are the states that have voted for both Republican and Democratic in two out of the last four

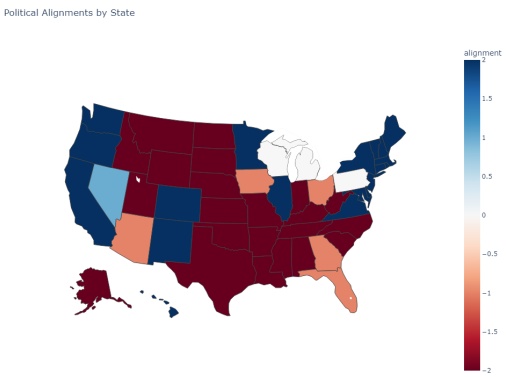


Figure T2.7

elections each.
To set up a comparison between states of clusters 1 and 2, we will identify the number of states that voted red and voted blue inside each cluster.

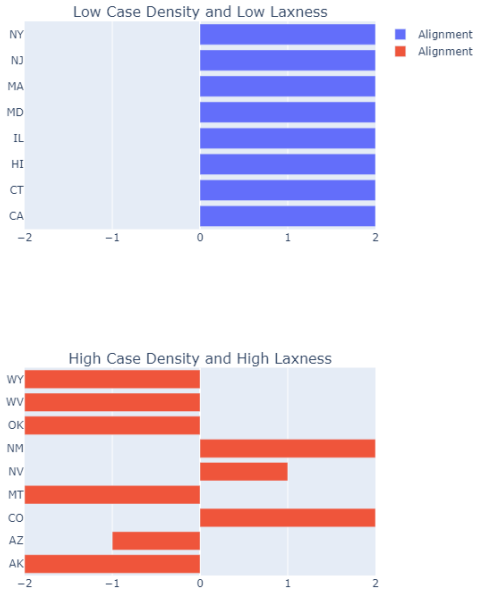


Figure T2.8

Figure T2.8 shows the double bar chart that shows the alignment of each state within the two clusters. While states that have a high case density and high laxness tend to be politically aligned towards the Republican party, there is no clear pattern as to their alignment. On the other hand, states that have a low case density and low gun laxness are politically aligned towards the Democratic party.

- *Conclusion*

In the manner outlined above, we have constructed a visual analytics workflow which starts with analyzing the case count in each state, and moves on to utilize state population fraction, state gun law laxness, state case density and state political alignment to make more thorough insights. The workflow diagram can be seen in Figure T2.9.

T3 Contextual Analysis:

Visual analytic workflow T3.1

We start our analysis by first analysing the most relevant statistic that we can find in a police shooting, which is if the person was armed or not, and we do this by plotting a treemap of all weapons with which the people were armed with.Fig:T3_1. Then, we go on to analyze the same statistic, but over the years(excluding 2024, as the data is incomplete), by plotting a line graph Fig:T3_2. By looking at this graph we see that the statistics for each weapon remains relatively similar, and the increase in values of each weapon is likely to reflect increase of population in the United States. We then go on to plot an even finer graph, where the data is aggregated at each month, and not each year, and see in Fig:T3_3, that the line graphs now see random, with there being some kind of variation with each year that appears to follow no fixed pattern. To understand these variations, we assume that the rate of shootings is standard if not uniform throughout each year, and we find the total shootings just by month, as in Fig:T3_4. Now instead of assuming that each year, the deviation from average is not too high when compared to the average across each month, we plot the variation in a composite visualization, that depicts the line graph showing average, and a scatter of points at each time interval, around the line graphs to indicate data points around each average, as shown in Fig:T3_5. Now that we have noticed that the deviation from the average is not too high for each month across a year, for each weapon, we can now try to run an ML model that can predict/forecast the amount of crimes that occur, weapon wise.
However, we cannot verify this claim as we do not yet have the data for future crimes. So instead, we split our current dataset into test and train, with our test data containing data uptil 2019, and then use timestamps from 2020 to 2023 to predict monthly crime aggregate statistics, and we see the result in Fig:T3_6. The actual statistics are shown in solid line, whereas the predicted output is shown in dotted lines.
We see that the predicted output is extremeley close to the actual values. However, there is too much clutter to make an actual inference, so we only see the output vs predicted for the top 3 categories for armed, which are gun, knife and unarmed, as shown in Fig:T3_7.
From these visualizations we have inferred that there is indeed a general trend to the shootings that happen, and that we can run ML algorithms to predict what an approximate number of shootings will be and what the context of arming of the person being shot will be.
We now show the visual analytic flowchart for the current

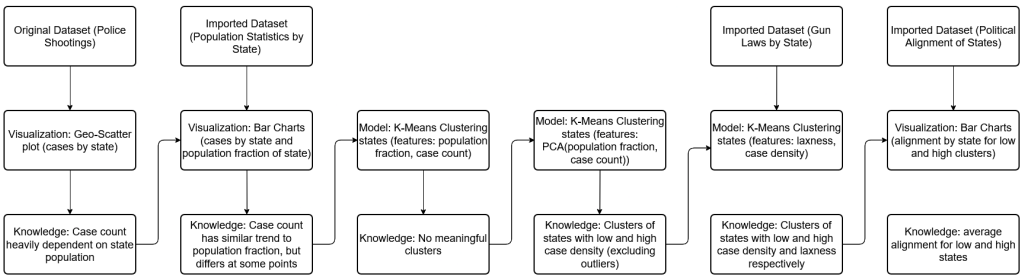


Figure T2.9: Workflow diagram for the geopolitical task

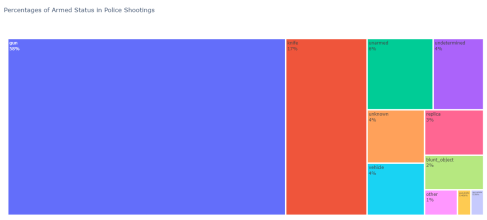


Fig:T3_1: Treemap of weapon armed with

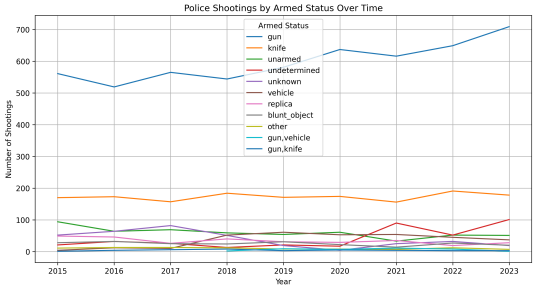


Fig:T3_2: Line graph of weapons armed with over the years

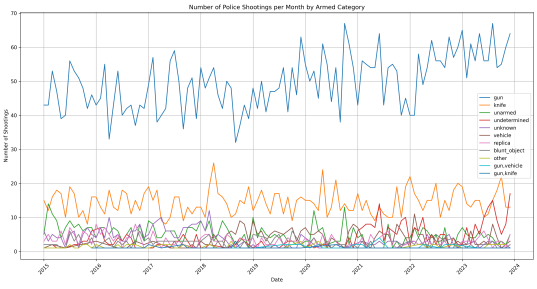


Fig:T3_3: Line graph of weapons armed with over the years over the months

analysis in Fig:Workflow_1.

Visual analytic workflow T3.2

We reiterate the inferences from our experimentation in A1, and we try to run a mathematical/statistical model to verify our findings, i.e. our visual inferences.

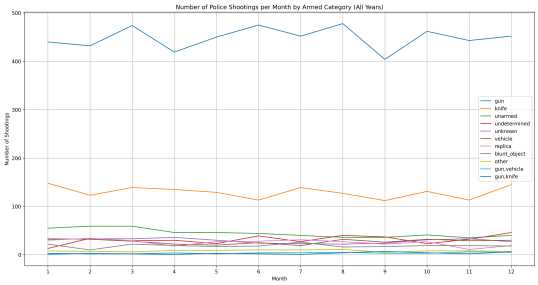


Fig:T3_4: Line graph of weapons armed with over each individual month

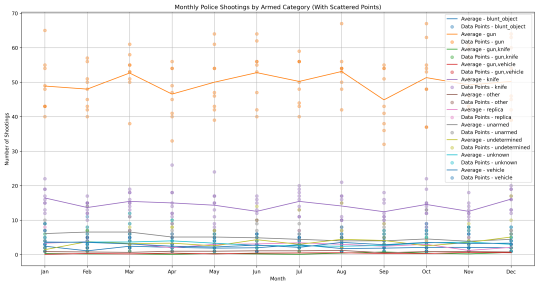


Fig:T3_5: Line graph of weapons armed with over each individual month, with integrated composite visualization of deviation of actual data points from the mean

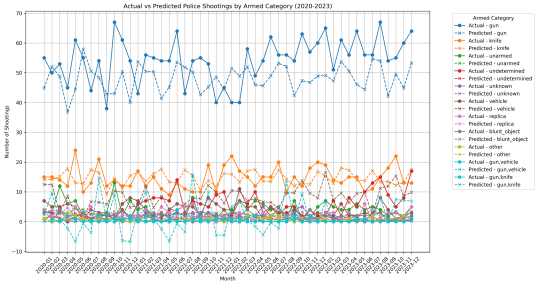


Fig:T3_6: Predicted vs actual line graphs of shootings with weapon with which armed

We inferred that there was no corelation between choice of weapon and presence of signs of mental illness (Hypothesis

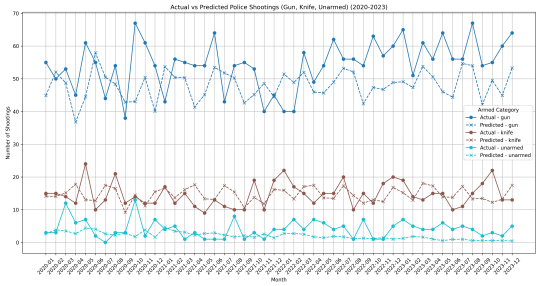


Fig:T3_7: Predicted vs actual line graphs of shootings with weapon with which armed, for gun, knife and unarmed

T3.1 from A1), people who were armed with vehicle attempted to flee (Hypothesis T3.2 from A1), and there is no corelation between the fact if bodycam is present on the officer or not, and if the suspect tries to flee.

Now we try to run a χ^2 distribution model and see the results of the statistics that are output by the model, and see if our inferences had any statistical backing, and we see the figures: Fig:T3_8_1, Fig:T3_8_2, Fig:T3_8_3 and Fig:T3_8_4, which are the contingency plots for {armed and flee}, {signs of mental illness and flee}, {armed and signs of mental illness}, and {bodycam and flee}.

We have run χ^2 model and see that for each of these graphs, the p values for the distributions are $0.0, 1.0005414136127088e-96, 8.022268880024748e-69$ and $7.108074707231791e{-}07$. This leads us to the conclusion that we can refute the null hypothesis($\mathscr{H}_0$) for correlation in each of the 4 pairs of columns chosen, and it is indeed possible to draw a conclusion from these.

We now show the visual analytic flowchart for the current analysis in Fig:Workflow_2.

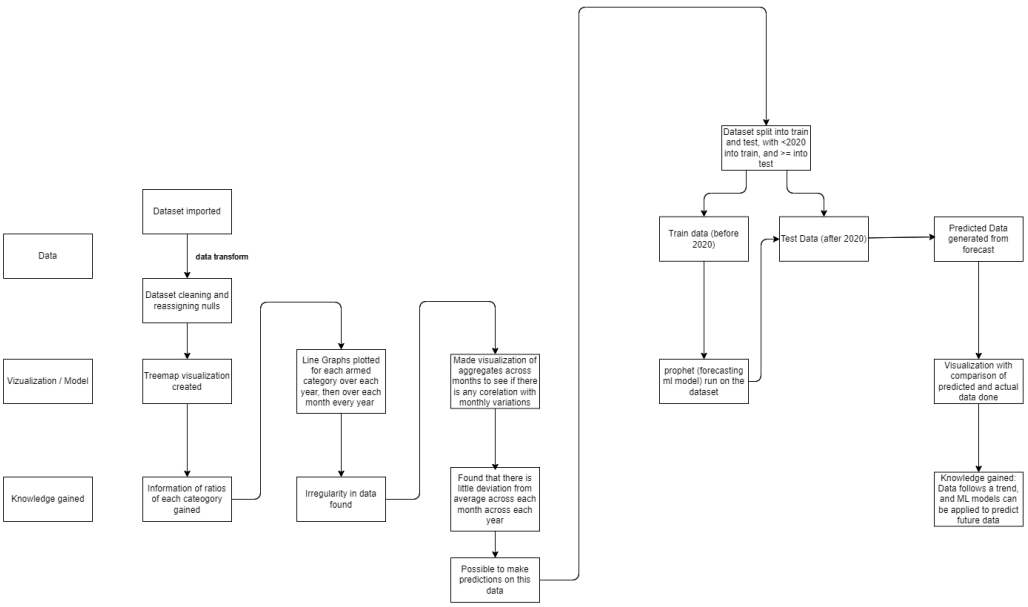


Fig:Workflow_1: Visual Analytic Workflow T3.1

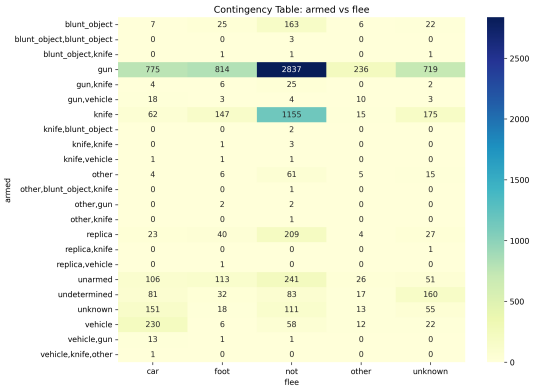


Fig:T3_8_1: Contingency plot for armed and flee

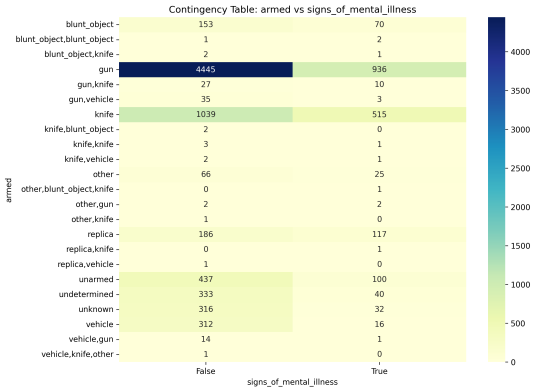


Fig:T3_8_3: Contingency plot for armed and signs of mental illness

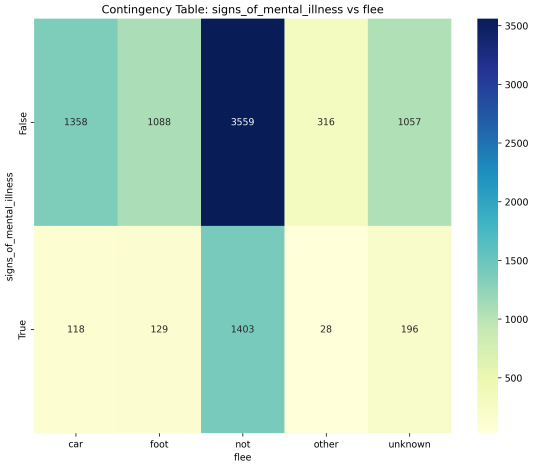


Fig:T3_8_2: Contingency plot for signs of mental illness and flee

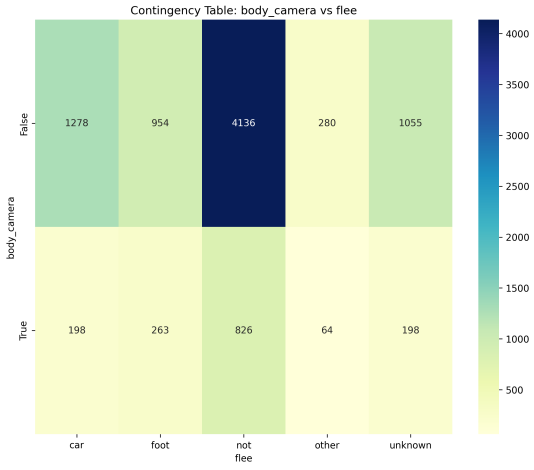


Fig:T3_8_4: Contingency plot for bodycam and flee

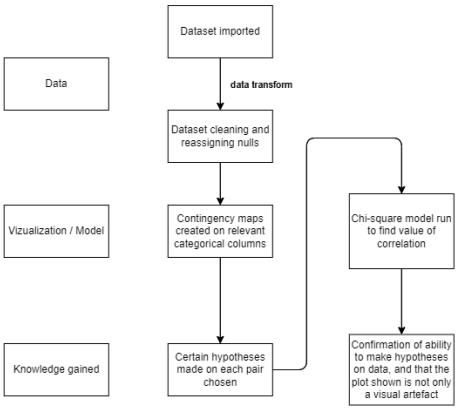


Fig:Workflow_2: Visual Analytic Workflow T3.2

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/after_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:

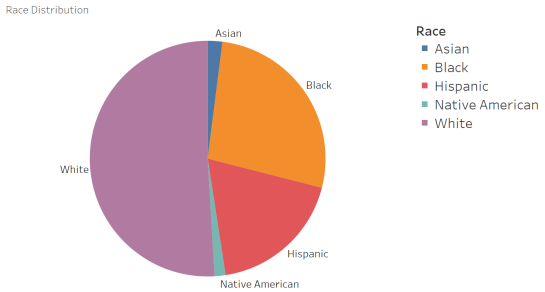
T1: Demographic Analysis
T2: Geopolitical Analysis
T3: Contextual Analysis

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.

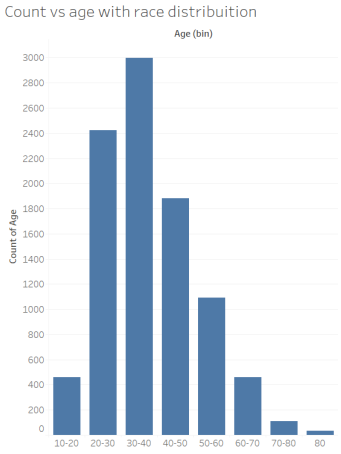
1) To ensure consistency in our analysis of time-related data, we excluded entries from the year 2024, as the data for that year is incomplete.
2) To improve clarity, visualizations exclude outliers from regions with minimal data, focusing instead on areas where the majority of data is concentrated.
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

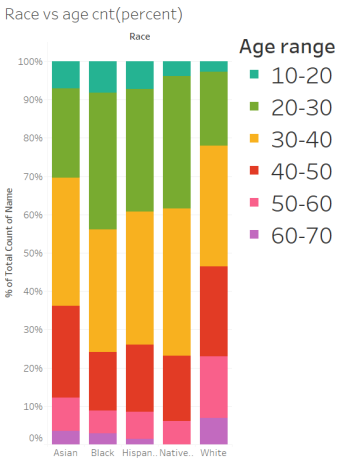
T1 Demographic Analysis:



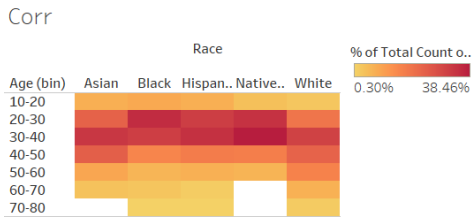
T1.2

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.



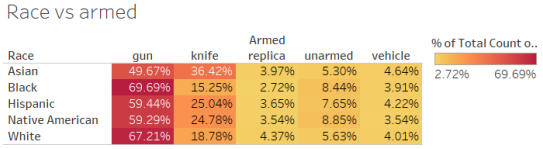
T1.3



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)

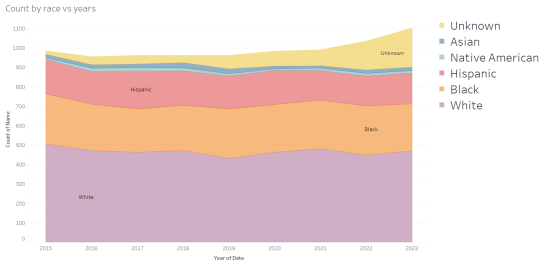
	Race					
Signs Of ..	Asian	Black	Hispanic	Native A..	Unknown	White
False	75.72%	85.27%	84.77%	84.73%	85.07%	74.40%
True	24.28%	14.73%	15.23%	15.27%	14.93%	25.60%

% 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.

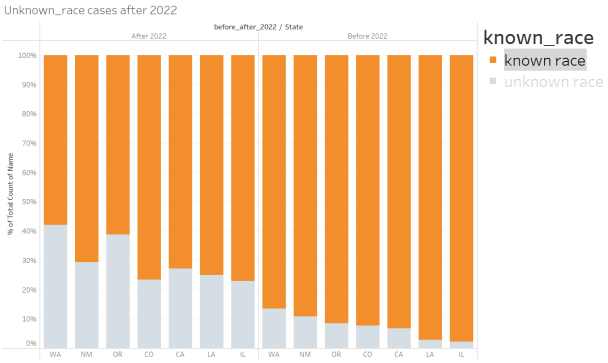


T1.8

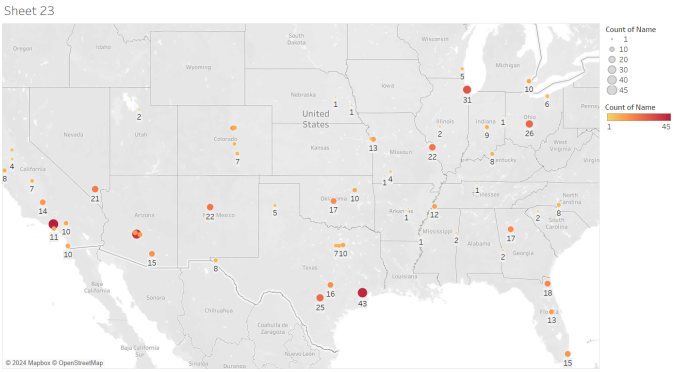
From the visualization Fig T1.8, we can observe a consistent rise in the number of police shooting cases in recent years. However, there has not been a significant increase among specific racial groups. In fact, there has been a slight decline in cases involving racial communities such as Hispanics. 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.

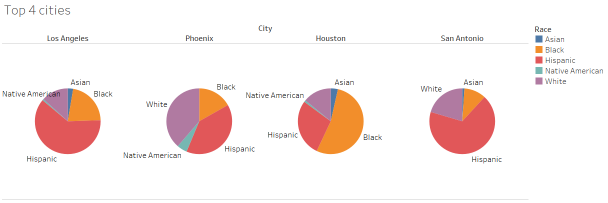
The cities with the highest number of police shooting cases include Los Angeles, Phoenix, Houston, and San Antonio. The racial distribution in these cities is



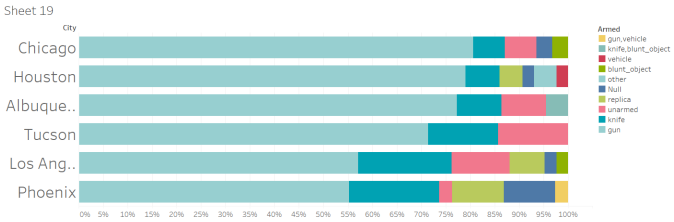
T1.9



T1.11



T1.10

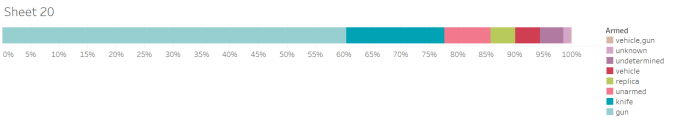


T1.12

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.



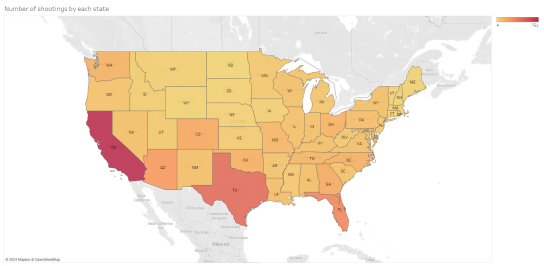
T1.13

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.

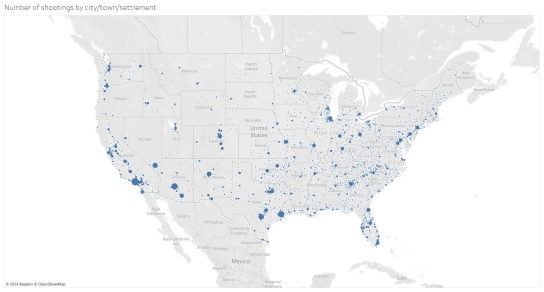
T2 Geopolitical Analysis:

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.



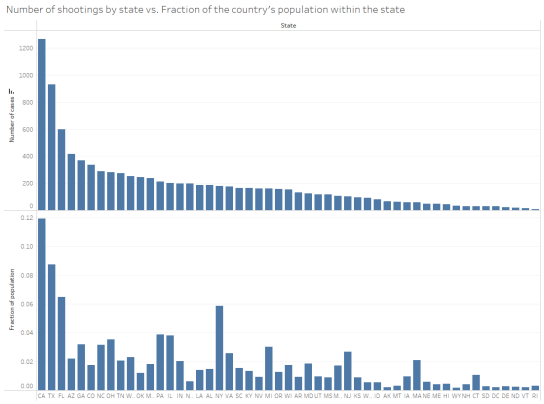
T2.1



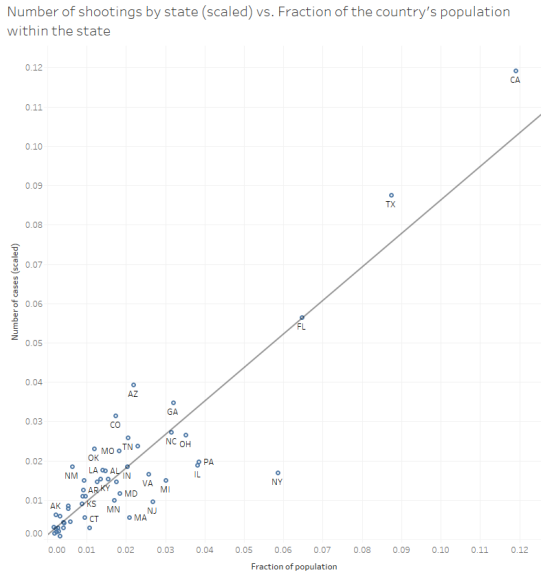
T2.2

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.



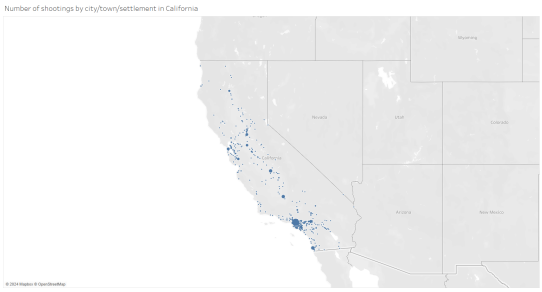
T2.3



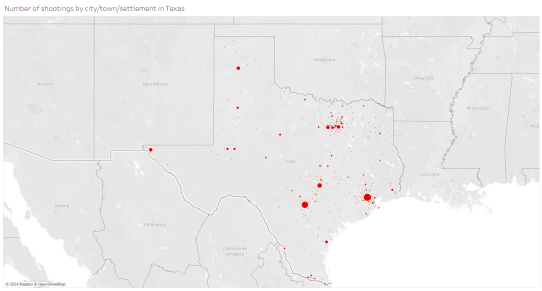
T2.4

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

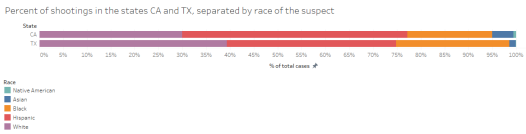


T2.5



T2.6

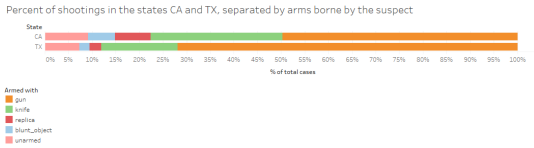
more concentrated towards the southern end of the state (where the Hispanic population is higher), whereas in Texas we see that the cases are mainly concentrated in the four major cities of Houston, Dallas, San Antonio and Austin.



T2.7

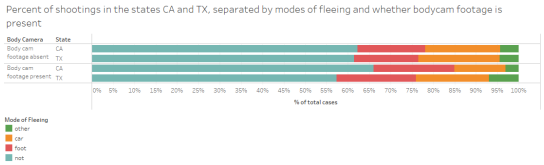
The stacked bar chart T2.7 shows the ratio of shootings in each of the two states segregated by race. There is a noticeably higher proportion of Hispanic civilians shot in CA, whereas the same can be said about Black people in TX. The skew towards Hispanic shootings in CA may be due to the higher proportion 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.

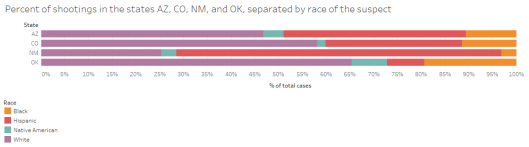
Hypothesis T2.3: The third hypothesis is that there is some geographical relation between the density of cases in each state, at the extreme values (low and high case density).
The 'density of cases', as mentioned before, refers to the number of cases of police shootings in the state, divided by its 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

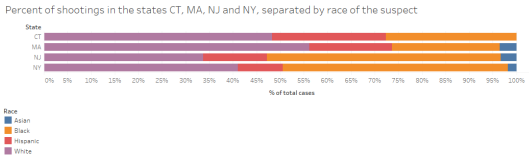
We can see that the states with lower density (CT, MA, NJ, NY) are all states that are located in either New England or the Tri-State Area, both of which are in the north-east of the US and are on the Atlantic Coast. On the other hand, the states (AZ, CO, NM) are in

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.

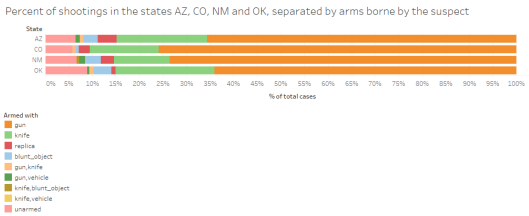


T2.10



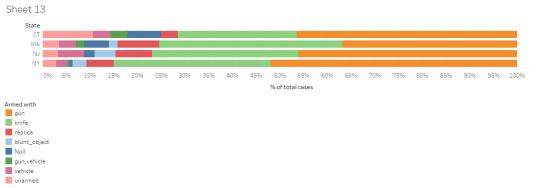
T2.13

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.



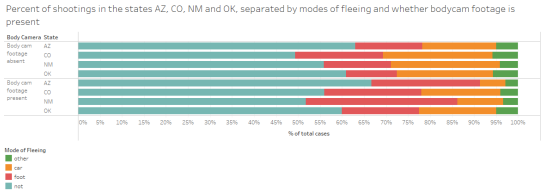
T2.11

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

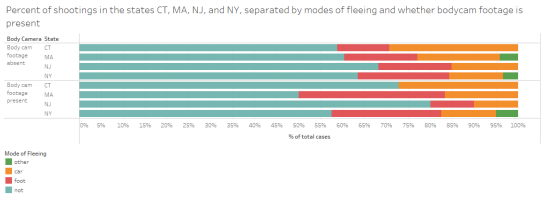


T2.14

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:

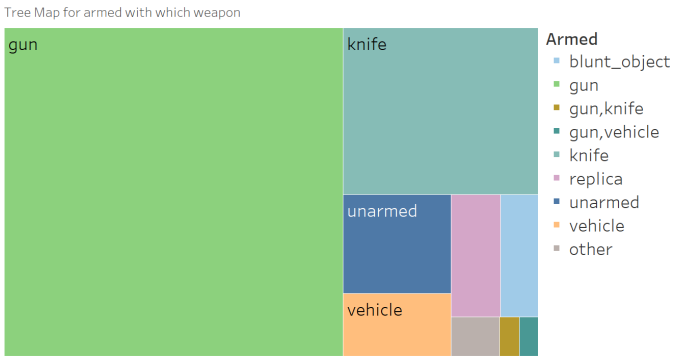
- Police officers in the low case density states are worse at defusing tense encounters, and tend to resolve cases by opening fire;
- Police officers in the low case density states tend to be more cautious about opening fire at suspects who have fled, in order to avoid collateral damage.

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

T3 Contextual Analysis:

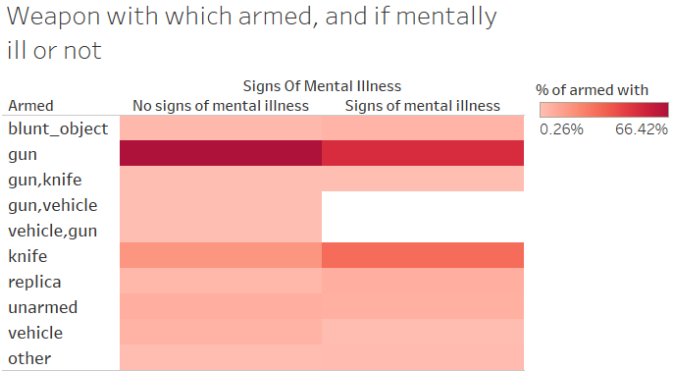
Hypothesis T3.1: The most obvious context for a police shooting is if the suspect was armed, so the more deadly the weapon on the suspect, the likelier they are to be shot at, so, the ratio of suspects with guns should be highest, and those unarmed should be lowest.
This hypothesis is somewhat verified 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

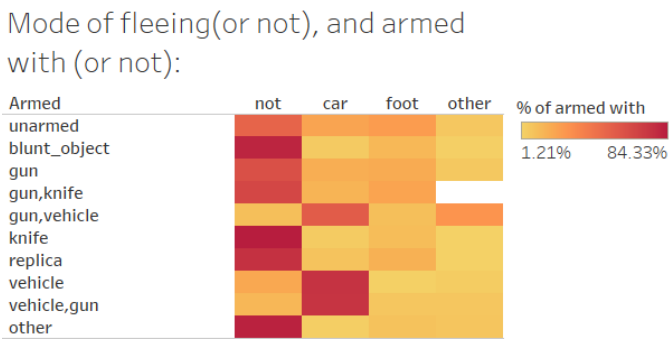
We can further look into this data, as our dataset also gives us information on whether the suspect was deemed mentally ill by the police officer or not, so we look at Fig T3.2, which tells us that the distribution of choice of weapon is similar across both those deemed mentally ill and not, with the 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.



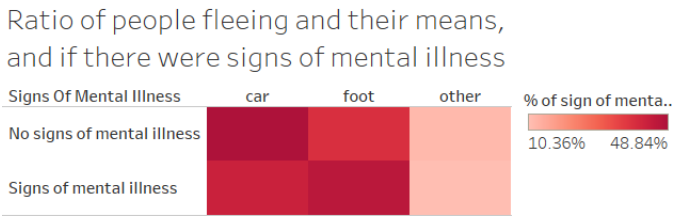
T3.2

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.

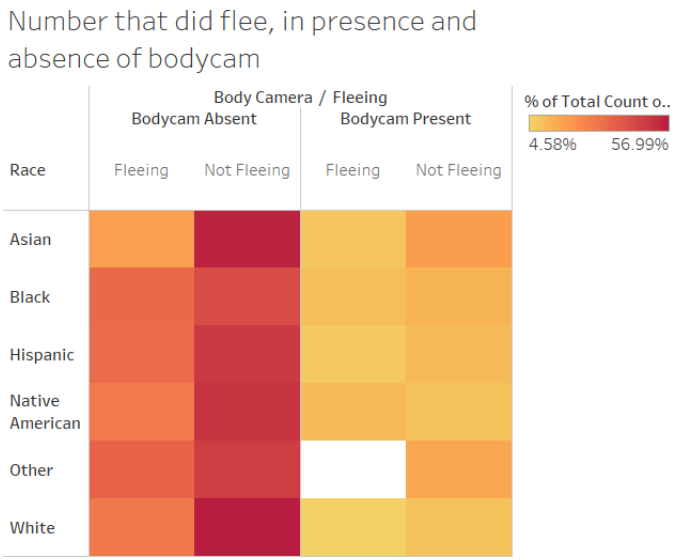


T3.3

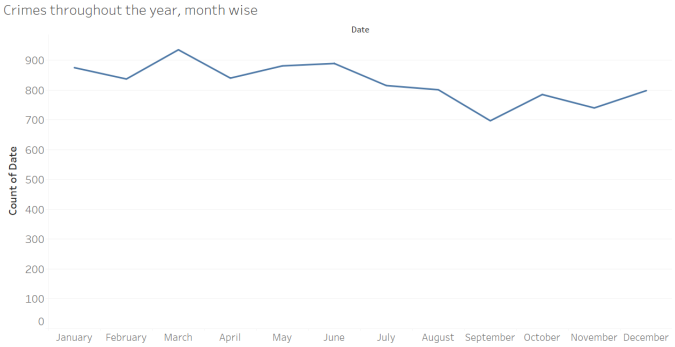


T3.4

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



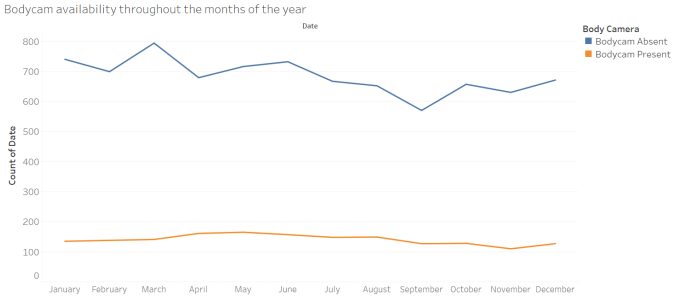
T3.5



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

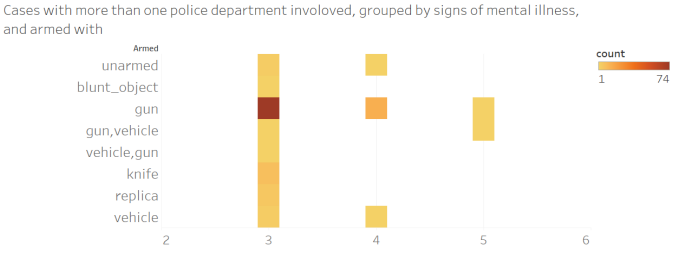
Hypothesis T3.4:We see in our dataset that certain cases have more than one police department involved per case. This is unusual for a typical case of a police shooting, so it must mean that the scale of the crime was greater, or the suspect was deemed highly dangerous, or maybe was a known criminal.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