Analysis of 2018 Crime Data in the United States

Group 14
Cai Xinlu 2230034002
Li Weilin 2230034027
LYU Shirui 2230026117
Zhan Hanhui 2230034060
Wang Shumin 2130034036

1 Introduction

1.1 Background

Crime seriously affects the security and stability of American society.

Incidents like "zero dollar purchase" robberies and shootings have become more frequent in the United States in recent years. By analyzing crime data, patterns of crime distribution, causes, and trends can be revealed. Our project focuses on exploring the relationship between factors such as income, race, age, and family size and crime.

1.2 Dataset introduction

Dataset contains 146 columns and 2216 rows, which describe crime-related data and factors in the United States for 2018.

For example

- **Crime-related metrics:**

  "murders", "murdPerPop", "rapes", "rapesPerPop", "robberies", "robbbPerPop", "assaults", "assaultPerPop", "burglaries", "burglPerPop", "larcenies", "larcPerPop", "autoTheft", "autoTheftPerPop", "arsons", "arsonsPerPop", "ViolentCrimesPerPop", "nonViolPerPop"

- **Population:**

  "population", "householdsize", "racepctblack", "racePctWhite", "racePctAsian", "racePctHisp", "agePct12t21", "agePct12t29", "agePct16t24", "agePct65up", "numbUrban", "pctUrban"

- **Income:**

  "medIncome", "pctWWage", "pctWFarmSelf", "pctWInvInc", "pctWSocSec", "pctWPubAsst", "pctWRetire", "medFamInc", "perCapInc", "whitePerCap", "blackPerCap", "indianPerCap", "AsianPerCap", "OtherPerCap", "HispPerCap", "NumUnderPov", "PctPopUnderPov"

1.3 Data Processing

Missing Values

The first is to deal with missing values.

We first print out the columns in the dataset with missing values.

state	state	0
countyCode	countyCode	1221
communityCode	communityCode	1224
population	population	0
householdsize	householdsize	0
racepctblack	racepctblack	2
racePctWhite	racePctWhite	0
racePctAsian	racePctAsian	0
racePctHisp	racePctHisp	0
agePct12t21	agePct12t21	0
agePct12t29	agePct12t29	0
agePct16t24	agePct16t24	0
agePct65up	agePct65up	0
numbUrban	numbUrban	598
pctUrban	pctUrban	598
medIncome	medIncome	0
pctWWage	pctWWage	0
pctWFarmSelf	pctWFarmSelf	24
pctWInvInc	pctWInvInc	0

Figure 1: The missing value section is displayed (Partial screenshot)

For the missing values, we used following steps:

- First, Columns with >50% missing values, removed.
- Second, Remaining missing values filled (numerical: median).

In addition, we have removed columns that are not meaningful for data analysis.

- Removed constant-value columns: countryCode, conmunityCode

Finally, we write the data after missing value processing and deleting the redundant columns into a new dataset file: processed_crimedata

1.4 Initial Settings

```
title: "Analysis of 2018 Crime Data in the United States"
output:
  flexdashboard::flex_dashboard:
    orientation: rows
    vertical_layout: scroll
    source_code: embed
    fig_width: 8
    fig_height: 7
runtime: shiny
```

Figure 2: The setting of output

1.5 Graph: Heatmap

We did a heat map to look at the relationships of all the variables in the processed dataset, and after processing, we found correlations between the factors we wanted to study and crime.

Because the dataset has more than a hundred columns and cannot be centrally represented on a single heat map, we divided the data into three groups, each with crime-related variables, to see the relationship between the crime variable and the other variables.

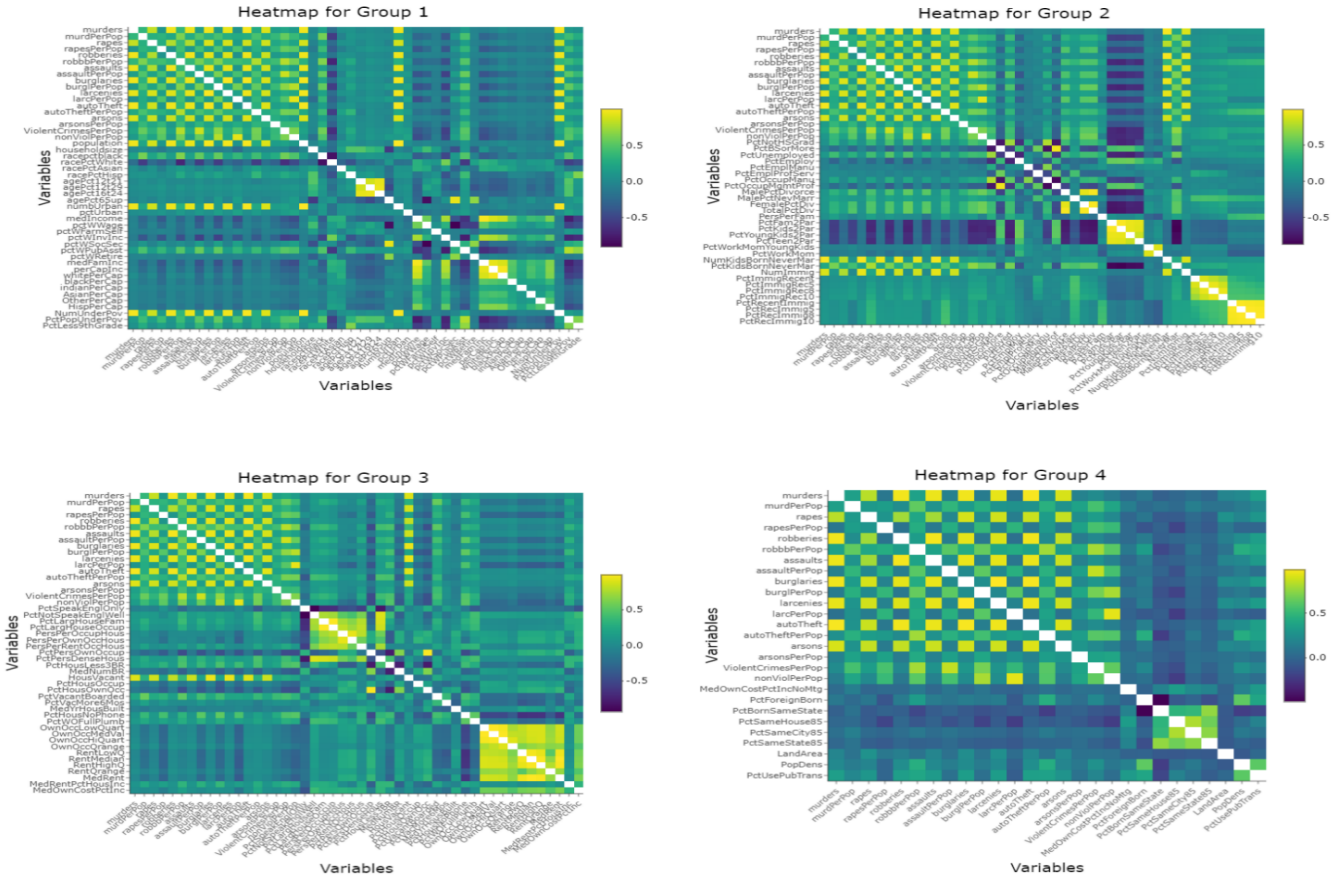

Figure 3: Heat maps of four groups

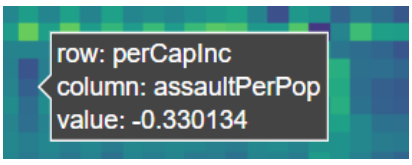

Figure 4: The corelation between PerCapIncome and assaultPerPop

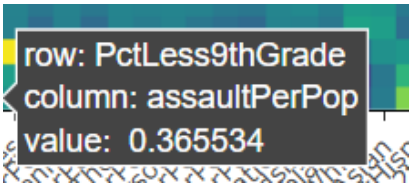

Figure 5: The corelation between PctLess9thGrade and assaultPerPop

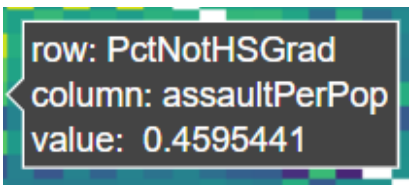

Figure 6: The corelation between PctNotHSGrad and assaultPerPop

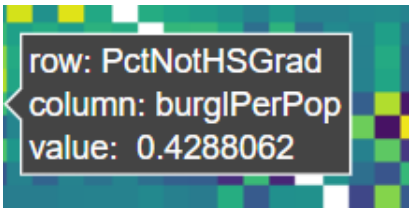

Figure 7: The corelation between PctNotHSGrad and burglPerPop

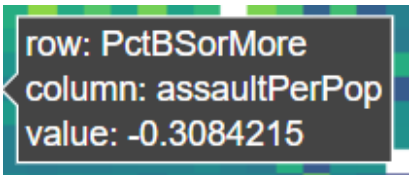

Figure 8: The corelation between PctSorMore and assaultPerPop

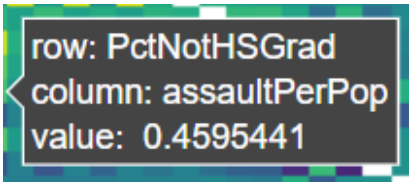

Figure 9: The corelation between population and assaultPerPop

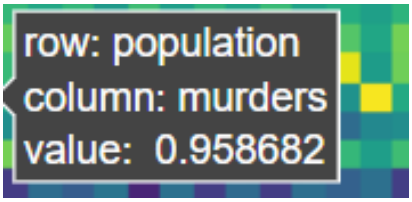

Figure 10: The corelation between population and murders

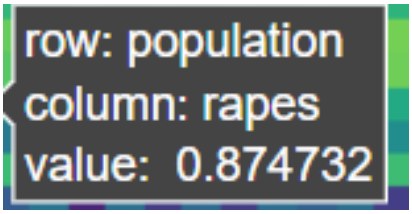

Figure 11: The corelation between population and rapes

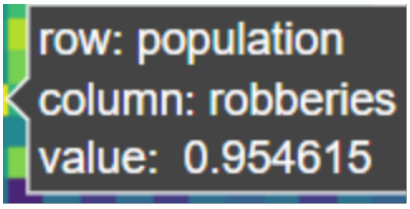

Figure 12: The corelation between population and robberies

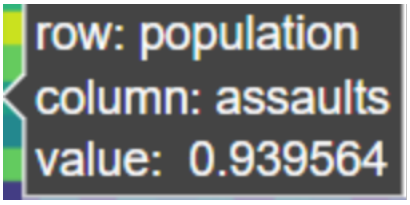

Figure 13: The corelation between population and assaults.

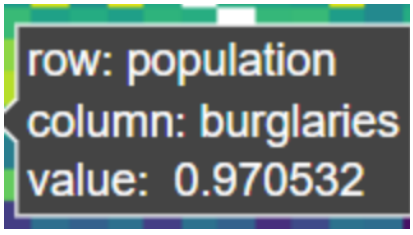

Figure 14: The corelation between population and burglaries.

2 Task Abstraction

- **Task1**: A heat map analyzed variable relationships in the processed dataset. Correlations between studied factors and crime were found. As the dataset had over a hundred columns, it was split into 4 groups of crime-related variables to assess their relationships.
- **Task 2**: The objective of this task is to visualize key demographic and socio-economic variables across the United States, including crime rates by type, racial composition, educational attainment, total population, per capita income, land area, and so forth. This analysis aims to uncover how these social and demographic factors correlate with the geographical location of each state. By doing so, we seek to gain a comprehensive understanding of the spatial dis-

tribution of crime, population composition, and other socio-economic factors, while exploring potential relationships between these variables.

- **Task 3**: In this task, we aim to first visualize the distribution patterns of various crime types across the U.S. states. Additionally, we will examine the relationship between state-level crime rates and socioeconomic factors, such as per capita income and educational attainment.
- **Task 4**: In this task, we aim to first visualize the distribution of various types of crime across states in the United States. The relationship between state-by-state crime rates and socioeconomic factors, such as age distribution, was then examined.
- **Task 5**: In this task, we aim to study the relationship between race and crimes of different states. Crimes are divided into violent crimes and nonviolent crimes. We study the relationship between these two factors and race separately. Besides, relationship between population and race is also taken into considered.

3 Task 1

3.1 Visualization Design

- **Reason for Choosing Heatmaps**

  In this data analysis task, the main goal is to determine the correlation between the variables. The heat map was chosen as the visualization method because it uses color gradients to represent the correlation coefficients (from the correlation matrix) in an intuitive way. Compared to scatter plots or raw tables, heat maps are particularly beneficial for high-dimensional data because they can represent relationships between multiple variables simultaneously, making it easier to observe overall trends and clustering patterns.

- **Mapping the Analytical Task to the Visualization Design**

  - **Analytical Task**: Understand the relationships between all variables, identify highly correlated or uncorrelated variables, and facilitate further modeling and data cleaning.
  - **Design Mapping**:

    Input: Variables were divided into 4 groups (Group 1 to Group 4), with each group used to compute a correlation matrix.

    Output: Corresponding heatmaps were generated to visualize correlations between variables using color depth, where darker shades indicate stronger correlations (positive or negative).

    Design Details:

    * **Diagonal Set to NA**: Diagonal elements (self-correlation) were excluded to avoid unnecessary distractions.
    * **Row and Column Ordering**: Clustering (Rows and Columns) was disabled to preserve the original variable order for easier interpretation.
    * **Distinct Titles for Each Group**: Each heatmap is labeled (e.g., 'Heatmap for Group 1') to clearly distinguish between variable groups and improve readability.

3.2 The Corresponding Theories/Principles

- **Tufte's Rules**

  According to Edward Tufte's principles of data visualization, effective visualizations maximize the data-to-ink ratio while minimizing unnecessary elements. The heatmap design adheres to the following principles:

  - **Maximized Data Density**: Heatmaps compress large amounts of numerical data into a single grid-based visualization using color gradients, providing high data density.
  - **Minimized Non-Data Elements**: Diagonal self-correlations were set to NA to remove redundant information and avoid distractions.
  - **Clear and Concise Presentation**: The grid structure and color coding directly convey the strength and direction of correlations without adding unnecessary complexity.

- **Not to Lie with Data Visualization**

  A critical principle of data visualization is to truthfully represent data without misleading the audience.

  - **Accurate Correlation Representation**: The heatmaps are derived directly from the correlation matrices, ensuring that all values accurately reflect the relationships between variables.
  - **Consistent Color Encoding**: The use of color gradients consistently maps the strength of correlations without exaggerating or distorting the data.
  - **Removal of Misleading Information**: Self-correlations were excluded from the heatmap to avoid any misinterpretation of diagonal elements showing perfect correlation.

- **Chart-Junk Debate**

  Tufte criticized unnecessary visual decorations, referring to them as "chart junk," which distracts viewers and reduces the efficiency of information delivery. This heatmap design strictly follows a minimalistic approach:

  - **No Redundant Decorations**: The heatmaps use a clean color grid without 3D effects, unnecessary embellishments, or distracting backgrounds.
  - **Simple Labels**: Only essential labels, such as variable names and axis markers, are included to maintain focus on the data.
  - **Efficient Information Delivery**: The color gradient effectively conveys correlation strengths, ensuring clarity, accuracy, and simplicity in the visualization.

3.3 Task 1 Conclusion

Heat map visualization effectively shows the correlation between multiple variables and meets the requirements of the analysis task. The design follows Tufte's principles of clarity and minimalism, respecting the principle that data visualization is not misleading and avoiding chart junk by presenting clean and efficient graphics. This visualization not only simplifies the identification of relationships between variables, but also provides valuable insights for subsequent modeling and feature selection.

4 Task 2

4.1 Visualization Design

- **Graph 1: Map**

  We initially plan to develop an interactive map that allows users to view detailed data for the United States based on selected crime types and social-economic indicators. This interactive map will enable users to zoom, pan, and explore specific regions. At the top of the map, four selection menus and a search box will be provided. Users can input the name of a state into the search box, which will cause the map to highlight the corresponding state based on the user's search query.

  For the first selection menu, "Select Crime Type," the map's color will change as users select different crime types. A darker red color will indicate a higher prevalence of the chosen crime type in that area, while lighter colors will signify a lower incidence of the crime. The remaining three selection menus allow users to choose between different age groups, racial demographics, state-level information (the total land area, population density, per capita income, and total population), and the number of violent and non-violent crimes in each state.

  When the user hovers over a state, detailed data will be displayed according to the selected criteria, including various crime rates, racial composition, and other socio-economic indicators mentioned above. It is worth noting that, for the sake of simplicity and ease of understanding, each selection menu is designed to allow only one choice at a time.

- **Graph 2: Crime Type Distribution Pie Chart**

  The first interactive pie chart displays the distribution of various crime types within the given dataset. Different colors are used to distinguish the crime categories: "Murder," "Rape," "Robbery," "Assault," "Burglary," "Larceny," "Auto Theft," and "Arson." The data is grouped by state, and the number of incidents for each crime type is averaged across the states. These values are then normalized and converted into percentages, allowing users to clearly observe the proportion of each crime type in the overall dataset. Hovering the mouse over different sections of the pie chart reveals the name of the crime type, the specific number of incidents, and its percentage share. Users can click on the corresponding legend to toggle the inclusion or exclusion of each crime type from the chart.

- **Graph 3: Racial Composition Pie Chart**

  The second interactive pie chart provides a snapshot of the racial composition within the dataset, displaying the percentage of the total population represented by different racial groups: Black, White, Asian, and Hispanic. Each group is marked with a distinct color in the pie chart. Hovering the mouse over a section of the chart shows the name of the racial group, its percentage of the total population (in %, without listing units in the chart to avoid confusion), and its proportion relative to the four groups under study. Users can click on the corresponding legend to toggle the inclusion or exclusion of each racial group from the chart.

- **Graph 4: Education Level Distribution Pie Chart**

  The third interactive pie chart visualizes the average percentage of the population with different educational backgrounds among those aged 25 and older: "Individuals with less than 9 years of education," "Individuals who have not completed high school," and "Individuals with a Bachelor's degree or higher." Each education level is marked with a different color in the pie chart. Hovering over a section of the chart reveals the name of the education level, its average percentage of the total population (in %, without listing units in the chart to avoid confusion), and its

proportion within the three categories under study. Users can click on the corresponding legend to toggle the inclusion or exclusion of each educational level category from the chart.

- **Graph 5: Crime Rate Pie Chart (Violent vs Non-Violent Crimes)**

  The fourth interactive pie chart illustrates the proportion of violent and non-violent crimes in the dataset, calculated as the average crime rate per 100,000 people for "Violent Crimes per Population" and "Non-Violent Crimes per Population." These two crime categories are represented as percentages derived from the respective averages in the dataset, highlighted with different colors in the pie chart. Hovering over a section of the chart reveals the name of the crime type (Violent or Non-Violent), the average number of incidents, and its percentage share in the total crime dataset. Users can click on the corresponding legend to toggle the inclusion or exclusion of each crime type from the chart.

4.2 The corresponding theories/principles

- **Graph 1: Map**
  - **Tufte's Rules for Data Visualization**

    Use Graphics: In this visualization, a map is employed to effectively display spatial patterns, enabling users to examine key socio-economic and crime data across different states. This approach allows for an intuitive understanding of the geographic trends in data distribution.

    Use Labels: Each state is labeled on the map, ensuring that users can easily identify states. This enhances clarity and usability by providing both visual and textual cues.

    Avoid Chartjunk: Following Tufte's rule, the map avoids unnecessary decoration or extraneous elements (like 3D effects or excessive gridlines) that could distract users from the core information.

    Utilize Micro/Macro: This map enables both micro (individual states) and macro (entire U.S.) analyses. Users can zoom into specific states to see more detailed information, or they can view the entire U.S. to understand broader patterns.
  - **Not to Lie with Data Visualization**

    Cumulative Graphs: There is no use of cumulative graphs, as each crime type or socio-economic variable is shown independently for clear, individual analysis. By focusing on one metric at a time, the map ensures that users can interpret each dataset without confusion or aggregation issues.

    Ignoring Conventions: The map adheres to conventional cartographic and data visualization best practices, such as using a common state outline map for spatial analysis. This makes the visualization accessible and understandable to users familiar with typical geographic data presentations.

    Inconsistent Scales: The map avoids inconsistent scales by ensuring that color gradients are used uniformly across all states. Each metric (crime rate, population, etc.) uses its own scale, but the map's color scheme maintains consistency to avoid misleading visual comparisons.
  - **Chart-junk Debate**

    Avoiding Excessive Decoration: The map follows the principle of minimizing "chartjunk" —extraneous visuals that do not add useful information. There are no unnecessary visual elements, 3D effects, or complex decorations. The map's simplicity aids in clarity, ensuring that the focus remains on the data itself.

Minimizing Redundancy: The map uses color gradation to convey important data points, eliminating the need for redundant or cluttered visual cues. For example, rather than adding multiple symbols or markers on the map, the states' shading directly communicates the data, making the visualization more efficient and easier to understand.

- **Graph 2-5: Pie Chart**
  - **Tufte's Rules**

    Use Graphics: Pie charts are chosen because they are effective in showing proportions of a whole. The use of pie charts helps in easily communicating the percentage distribution of each category in a clear and intuitive manner.

    Let the Data Speak:The pie charts directly present the proportions of various categories (e.g., crime types, racial distribution, education levels, crime rates), allowing the data to be interpreted naturally. The use of percentages ensures clarity and precision in the visualization.

    Use Labels: Each slice in the pie chart is labeled with both the category name and its corresponding percentage. This ensures that viewers can immediately identify both the specific crime type, race group, or education level and its relative proportion.

    Avoid Chartjunk: Our pie charts design avoids unnecessary decoration or extraneous elements that do not contribute to the understanding of the data. The layout is clean, and there are no extraneous visual elements (such as 3D effects or gradients) that distract from the data's message.

    Utilize Micro/Macro: The pie charts effectively balance micro (individual categories) and macro (overall proportions) views. The individual segments allow for detailed inspection, while the overall pie chart presents a holistic view of each data distribution.
  - **Not to Lie with Data Visualization**

    Cumulative Graphs:

    The charts are not cumulative; each pie chart is independent and represents a distinct set of data. This avoids misleading interpretations that could arise from stacking values in a single chart.

    Ignoring Conventions:

    The use of pie charts is in line with standard conventions for representing categorical data as proportions. The chart titles are appropriately descriptive, and the percentages are clearly labeled.

    Inconsistent Scales:

    The pie charts use a consistent scale to represent percentages. Each chart is scaled to 100%, ensuring that comparisons between categories within each chart are meaningful and consistent.

    Size & Volume Encoding:

    The pie chart format avoids using size and volume encoding (such as 3D pie charts), which can distort perception. The values are encoded as percentages, which are easy to interpret visually.
  - **Chart-Junk Debate**

    Minimalist Design: The design prioritizes clarity over decoration. The pie charts focus on presenting accurate data, avoiding unnecessary visual elements (e.g., 3D effects, excessive gridlines, or excessive color).

    Effective Use of Color: The colors used for each category are distinct but not overly saturated, ensuring that they are easily distinguishable while not overwhelming the viewer.

4.3 Visualization Results

- **Graph 1: Map**

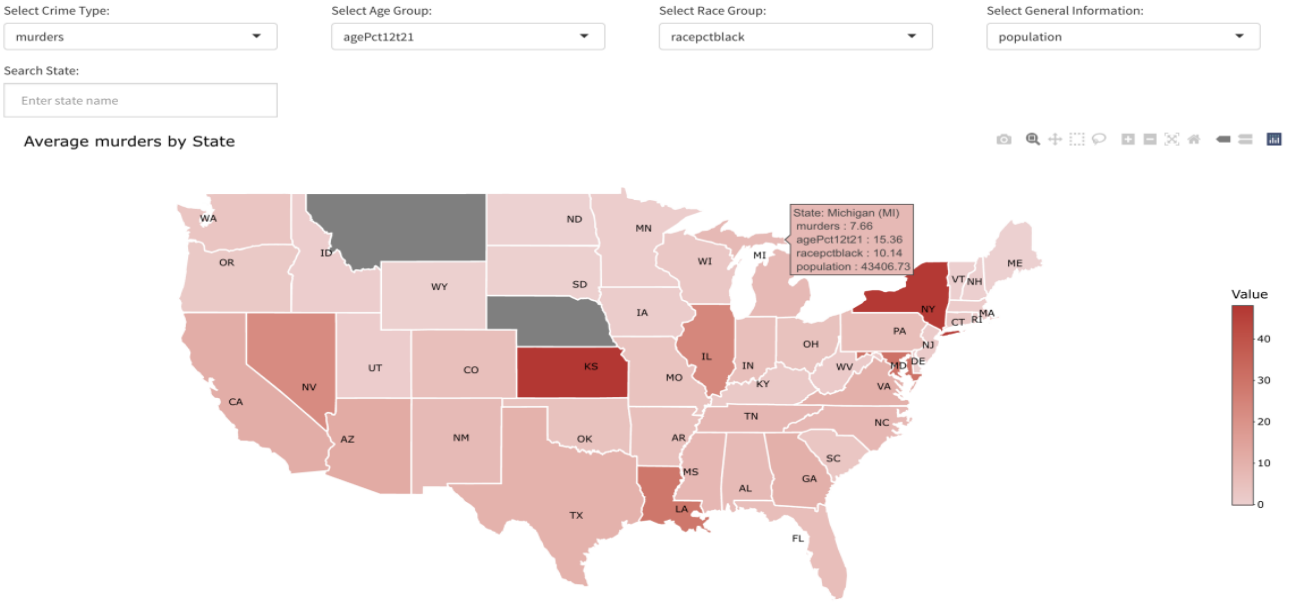

Figure 15

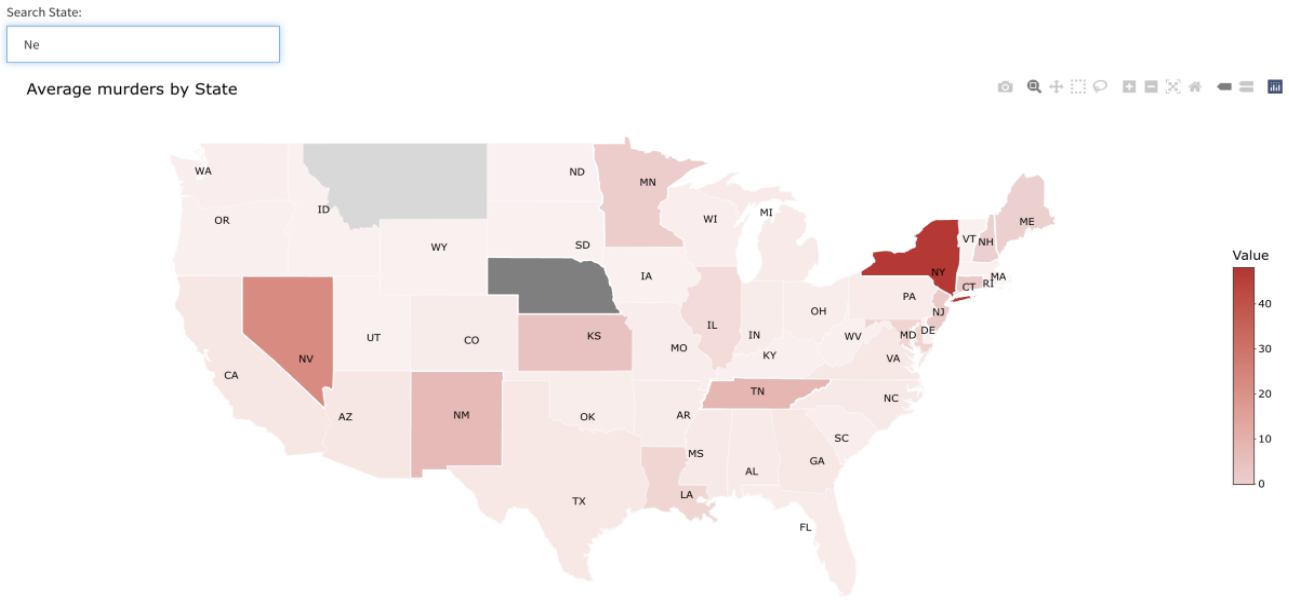

Figure 16

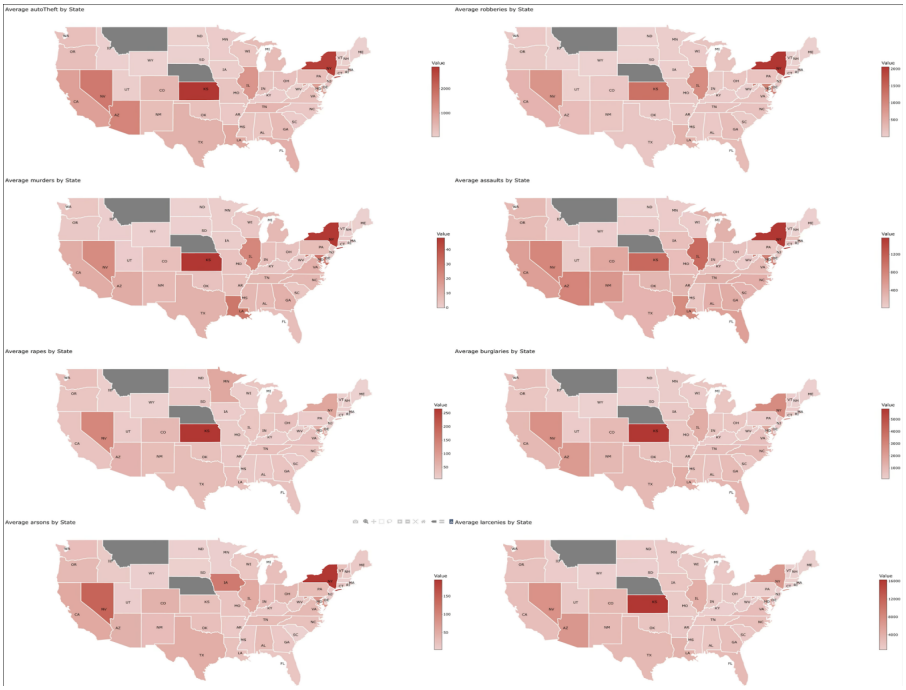

Figure 17

From the map, we can initially observe that various crime types are relatively common in the states of Nevada, Kansas, and New York.

- **Graph 2: Crime Type Distribution Pie Chart**

  Note that due to the floating-point precision problem, we encountered an issue with the first pie chart. Despite multiple adjustments to the distribution percentages, the total still doesn't add up to exactly 100%. This is something that we'll need to address in future updates of the analysis.

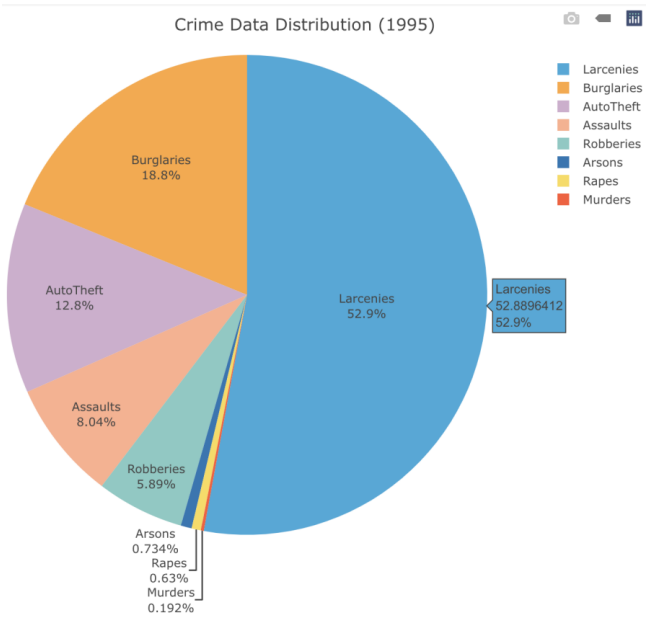

Figure 18: Crime Data Distribution

- **Graph 3: Racial Composition Pie Chart**

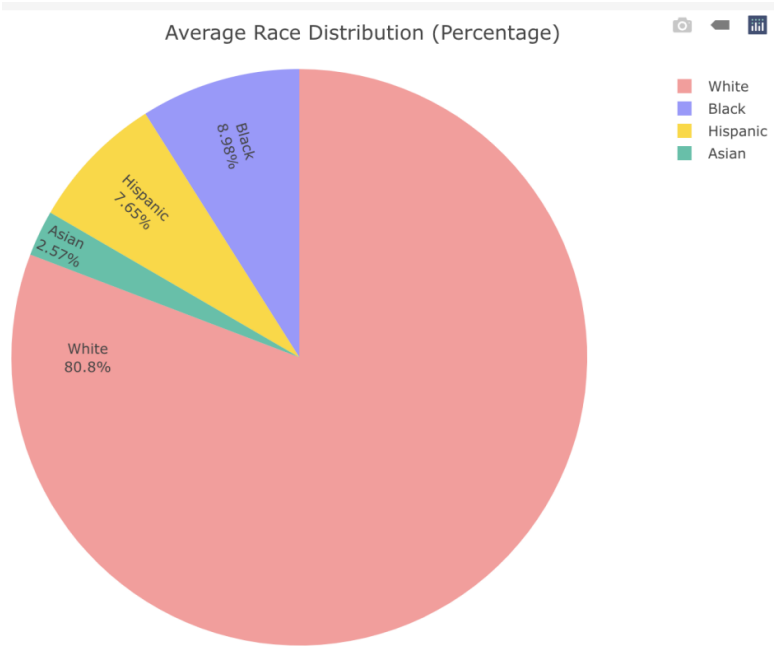

Figure 19: Average Race Distribution

- **Graph 4: Education Level Distribution Pie Chart**

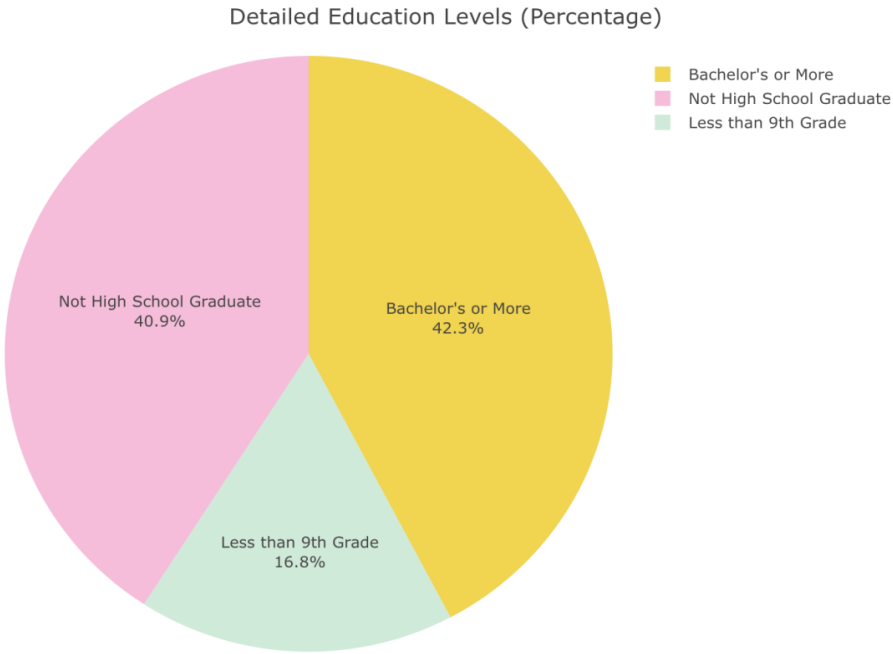

Figure 20: Detailed Education Levels

- **Graph 5: Crime Rate Pie Chart (Violent vs Non-Violent Crimes)**

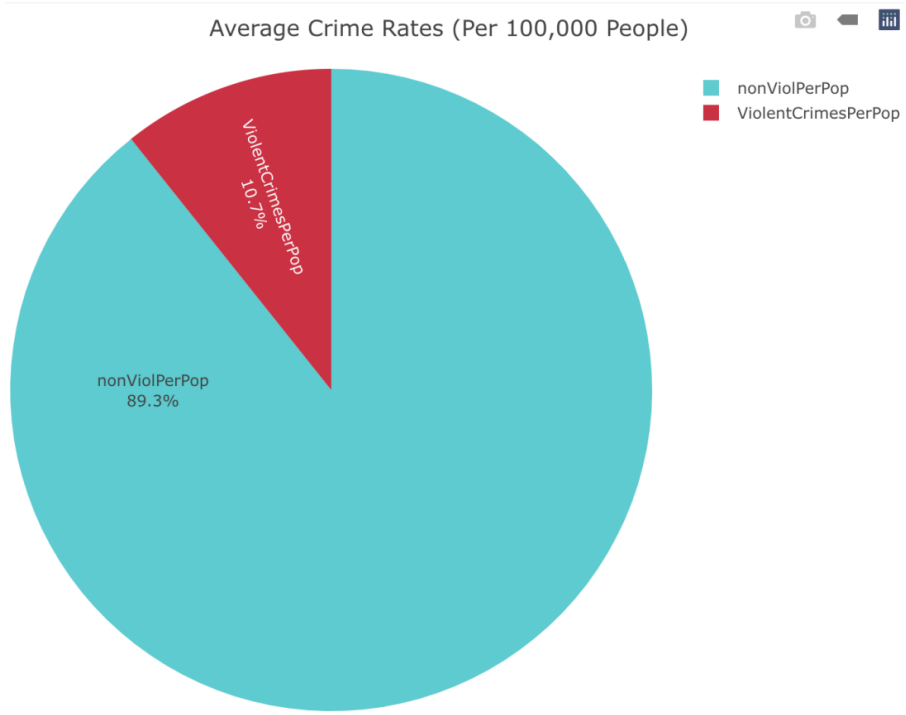

Figure 21: Average Crime Rates

From these pie charts, we can observe several key trends: non-violent crimes are more common overall, with Larcenies being the most frequently occurring crime. In terms of racial distribution, white individuals make up the largest proportion. As for educational attainment, the percentages of people with lower and higher levels of education are roughly equal, and together, they represent the majority.

5 Task 3

5.1 Visualization Design

- **Graph 1**

  The first chart is designed to analyze the total number of various types of crimes in each state in the United States, as well as their distribution. A bar chart is used for this analysis. The height of the bars provides a direct representation of the rankings. Additionally, interactive features are incorporated, allowing users to hover over the bars to clearly view the total number of crimes in each state. A gradient color scheme is applied to the bars, enhancing the visual clarity of crime totals, which enables viewers to quickly distinguish between states with higher and lower crime counts. The bar chart ranks the states in descending order based on total crime numbers, effectively displaying the rankings.

- **Graph 2**

  The second chart analyzes the proportion of different crime types across states. A stacked bar chart is used, which allows multiple data series to be displayed within the same bar. The height of each segment represents the value of different subcategories. This is ideal for showing the proportion of each crime type in different states. Text labels are placed above each stacked

section to display the percentage of each part, providing clear information about the specific data for each state. Additionally, a dropdown menu is added, enabling users to select and display the top N states by total crime numbers (Top 5, Top 10, or Top 15). By selecting states with different rankings, users can compare the commonalities and differences in crime types among high-crime states.

- **Graph 3**

  The third chart analyzes the relationship between crime rates and per capita income, as well as between crime rates and educational attainment. We chose to use a scatter plot with a linear regression line. By adding the regression line, we can clearly observe the trend of changes between the two variables and determine whether they are negatively or positively correlated. A dropdown menu is included, allowing users to select different states, enabling us to examine whether the trends in the relationships between these variables are consistent across states or if there are differences.

5.2 The corresponding theories/principles

- **Graph 1**
  - **Tufte's rules**: The bar chart clearly displays the differences in data size through the use of a color gradient (scale_fill_gradient), which enhances data density while avoiding redundant information.
  - **Not to lie with data visualization**: The bar chart does not over-scale or compress the data during processing. In terms of the color gradient, states with lower crime numbers are represented in light blue, while states with higher crime numbers are shown in purple. This ensures that viewers can accurately understand the true distribution of crime across states.
  - **Chart-junk debate**: The bar chart uses only the gradient color to convey data, without the addition of complex decorative elements.
- **Graph 2**
  - **Tufte's rules**: The stacked bar chart uses color to distinguish between different crime types, providing an efficient visual representation of the data. The position stacking conveys the proportion of each crime type in relation to the total crime count, eliminating the need for additional explanation. Additionally, the stacked bar chart displays the proportions of all crime types at once, increasing the data presentation density.
  - **Not to lie with data visualization**: Our stacked bar chart accurately labels the percentage of each crime type using proportion labels (geom_text), ensuring data transparency.
  - **Chart-junk debate**: Our chart uses a simple color palette (scale_fill_brewer(palette = "Set3")) to clearly illustrate the proportion of each crime type. Additionally, the dropdown menu provides a clean design, allowing users to quickly select and focus on states with higher crime rates that they wish to explore.
- **Graph 3**
  - **Tufte's rules**: The scatter plot effectively represents a large number of data points, directly displaying the relationship between two variables. By adding a regression line (geom_smooth), it helps viewers quickly observe the trend between the variables, revealing potential relationships within the data.

– **Not to lie with data visualization**: The data points in the chart have not been distorted or misleading, and the regression line accurately reflects the distribution of the data.
– **Chart-junk debate**: The scatter plot uses a clean design, distinguishing whether individuals have higher education by color and shape, making it intuitive and easy to understand without unnecessary visual elements. Additionally, the inclusion of the regression line enhances the analytical value of the chart. Finally, the dropdown menu allows users to quickly select and explore the states they are interested in.

5.3 Visualization results

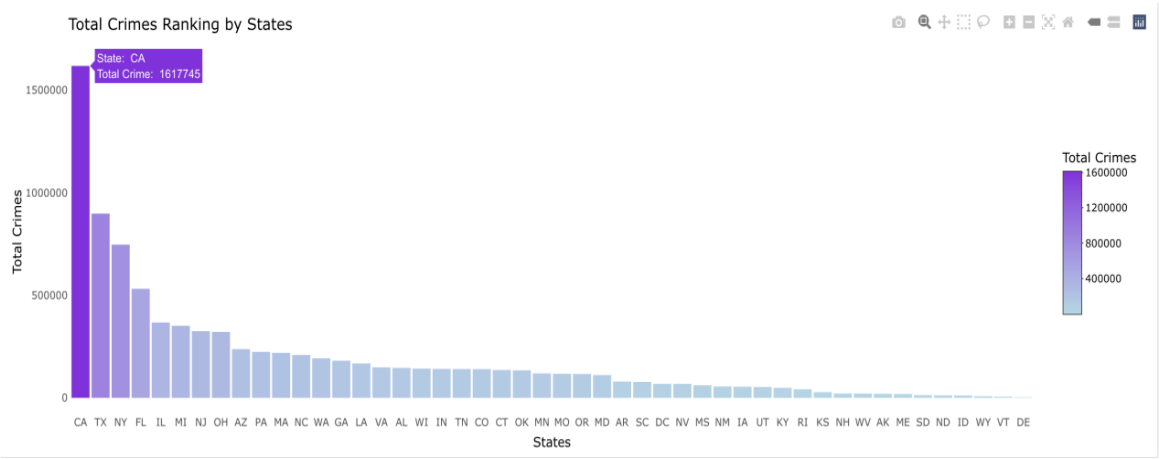

Figure 22: Total crime ranking by states

From a macro perspective, we learn that there is an uneven distribution of the total number of crimes across the United States.

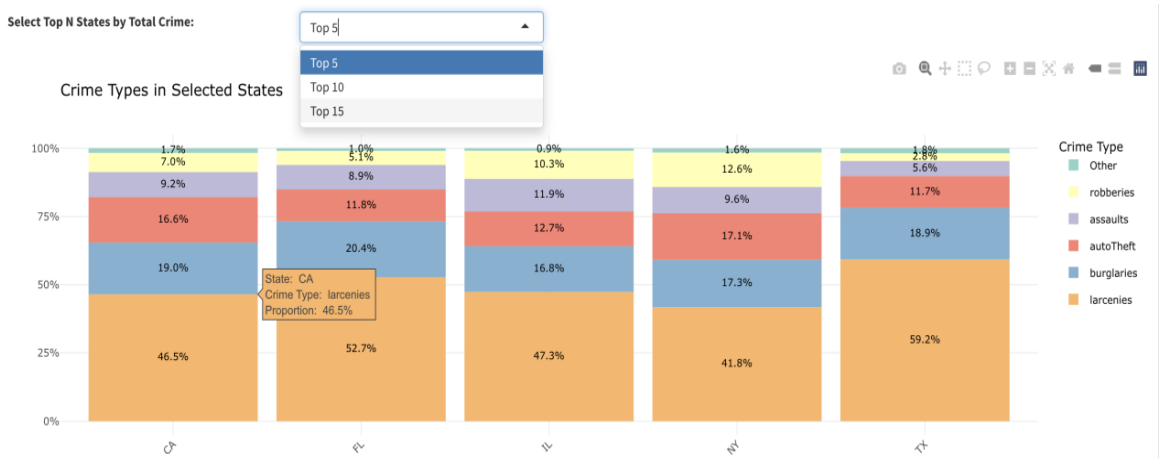

Figure 23: Crime types in selected states

We found a common trend: Theft-related crimes (larcenies, burglaries and autoTheft) make up a significant portion in all of these states.

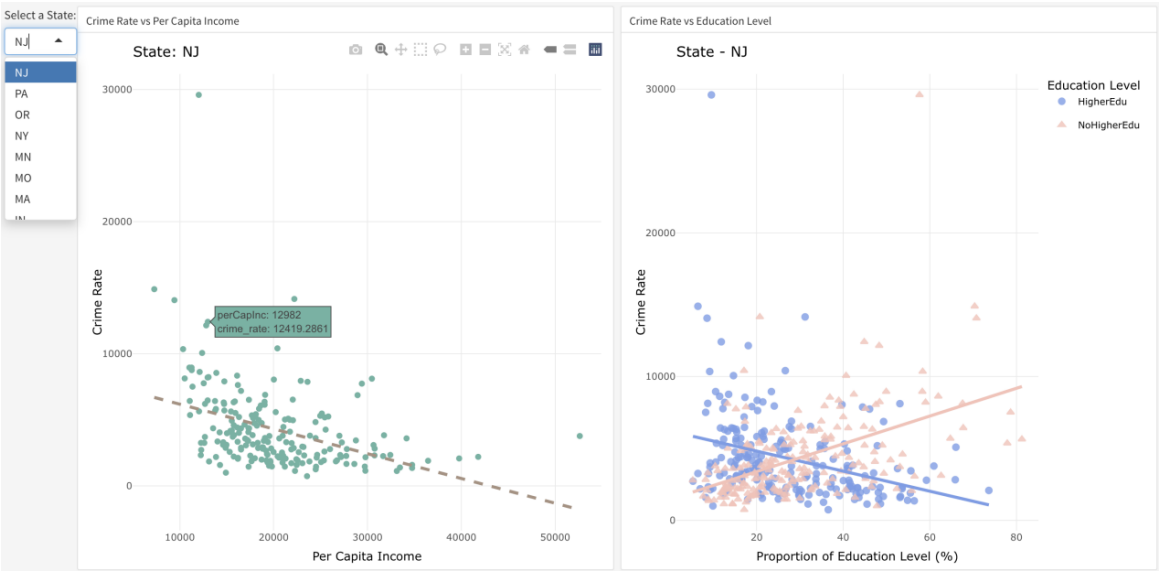

Figure 24: Crime Rates v.s. per capita/education levels

We found a consistent pattern: In a state, per capita income is negatively correlated with crime rates, meaning the higher the income, the lower the crime rate. Additionally, the higher the proportion of people with higher education in a state, the lower the crime rate. Conversely, states with a higher proportion of people without higher education tend to have higher crime rates.

6 Task 4

6.1 Visualization Design

- **Graph 1**

  The first chart is used to analyze the number of violent crimes in the United States in relation to the age distribution. Scatter plots are used for this analysis. The distribution of the scatter plot directly represents the corresponding number of crimes, while the regression curve is plotted to show the general trend. In addition, interactive features are combined, allowing users to hover over the scatter point to clearly see the age range and number of crimes corresponding to the crime incident. Different colors are used in different age ranges to enhance the degree of differentiation of crime incidents in different age ranges, so that the audience can quickly distinguish crime situations in different age ranges.

- **Graph 2** The second chart, similar to the first, analyzes the number of nonviolent crimes in the United States by age range. A scatter plot is also used for this analysis. The distribution of the scatter plot directly represents the corresponding number of crimes, while the regression curve is plotted to show the basic trend. Interactive features are also incorporated, allowing users to hover over the scatter point to clearly see the age range and number of crimes corresponding to the crime incident. Different colors are used in different age ranges to enhance the degree of differentiation of crime incidents in different age ranges, so that the audience can quickly distinguish crime situations in different age ranges.

6.2 The corresponding theories/principles

- **Graph1-2**
  - **Tuft's rule**: Scatter plots clearly show the distribution of data by using different colors, which enhances data density while avoiding redundant information.
  - **Don't lie about data visualization**: Scatter plots don't overscale or compress data during processing. In terms of color, different colors are used for different age distributions. This ensures that viewers can accurately understand the distribution of true crime across age groups.
  - **Chart Junk debate**: Scatter plots use only four different colors to convey data without adding complex decorative elements.

6.3 Visualization results

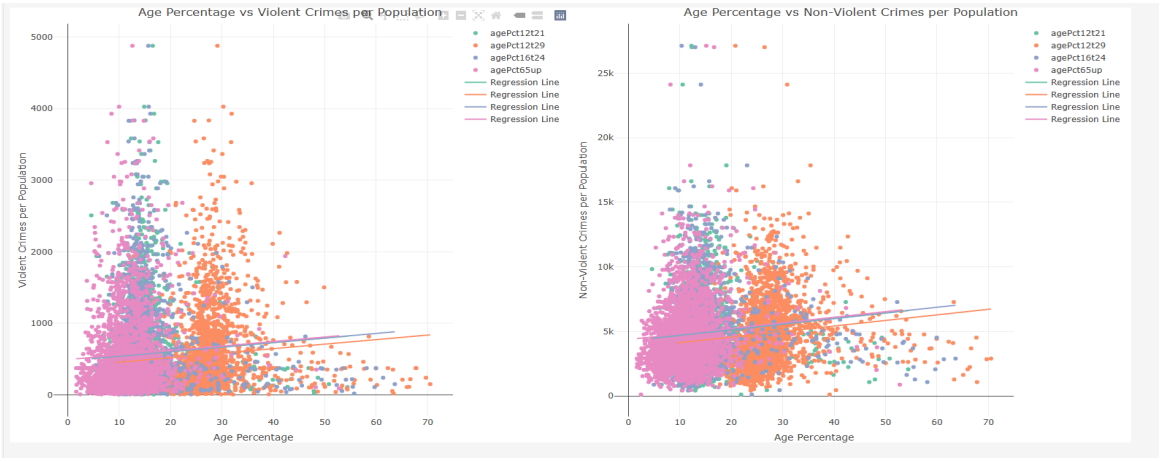

Figure 25: Age percentage v.s. Violent crimes/Non-violent crimes

So we see a very slow positive correlation between age and the number of crimes in a state, which means that the higher the age, the higher the crime rate, but it is possible that there is no strong correlation between age and crime rate.

7 Task 5

7.1 Visualization Design

- **Graph 1** The first chart analyzes the violent crimes in different states in the US about races. Scatter plots can better display the racial distribution of different communities in different states. This analysis used scatter plots. Replacing the race of each community in different states with the one with the highest proportion of race. When the mouse passes through the scatter point, the corresponding community name, state and crime count can be seen, with the largest proportion being the race. The four colors represent different races. After click the race label, the number of crimes committed by a single race can be displayed separately.
- **Graph 2** The second chart analyzes the nonviolent crimes in different states in the US about races. Scatter plots can better display the racial distribution of different communities in different

states. This analysis used scatter plots. Replacing the race of each community in different states with the one with the highest proportion of race. When the mouse passes through the scatter point, the corresponding community name, state and crime count can be seen, with the largest proportion being the race. The four colors represent different races. After click the race label, the number of crimes committed by a single race can be displayed separately.

- **Graph 3** The third chart analyzes the population of states in different communities with different color of races. A bar chart can more intuitively display the population of different states and races. When the mouse passes through the bars, the corresponding community name, state, race and population can be seen. Different colors of bars can be shown in each bar. It is convenient to see the population distribution in each state.

7.2 The corresponding theories/principles

- **Graph1**
  - **Tuft's rule**: The scatter plot uses graphic to let the data speak, scatter can better summarize a large amount of data into the same graph, a picture in place of thousand words.
  - **Don't lie about data visualization**: The y-axis shows entire scale. Not use cumulative graphs, but the scatter points to show the data.
  - **Chart Junk debate**: Scatter plot itself is a very intuitive way to show the relationship between two variables. Different colors or labels distinguish different races, making it easier to observe and discover trends.
- **Graph 2**
  - **Tuft's rule**: The scatter plot uses graphic to let the data speak, scatter can better summarize a large amount of data into the same graph, a picture in place of thousand words.
  - **Don't lie about data visualization**: The y-axis shows entire scale. Not use cumulative graphs, but the scatter points to show the data.
  - **Chart Junk debate**: Scatter plot itself is a very intuitive way to show the relationship between two variables. Different colors or labels distinguish different races, making it easier to observe and discover trends.
- **Graph 3**
  - **Don't lie about data visualization**: Longer bars indicate larger numbers, the bar chart shows size of populations in each state. Avoid size encoding by bars.
  - **Chart Junk debate**: The bar chart is intuitive, making it easy to compare the population of different states. Using colors to distinguish races can also show the proportion of racial composition.

7.3 Visualization results

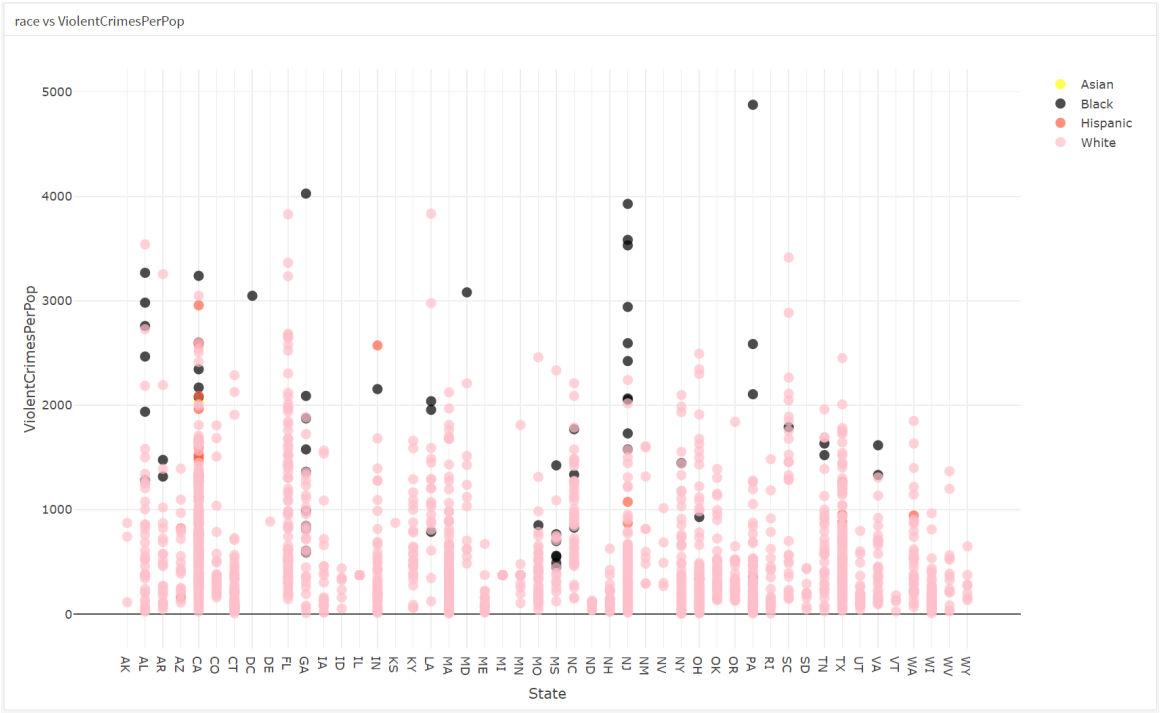

Figure 26: Race v.s. Violent Crime Rate

The community with the highest number of black people always has more violent crimes, followed by white people. However, community with the highest number of white people has totally more violent crimes.

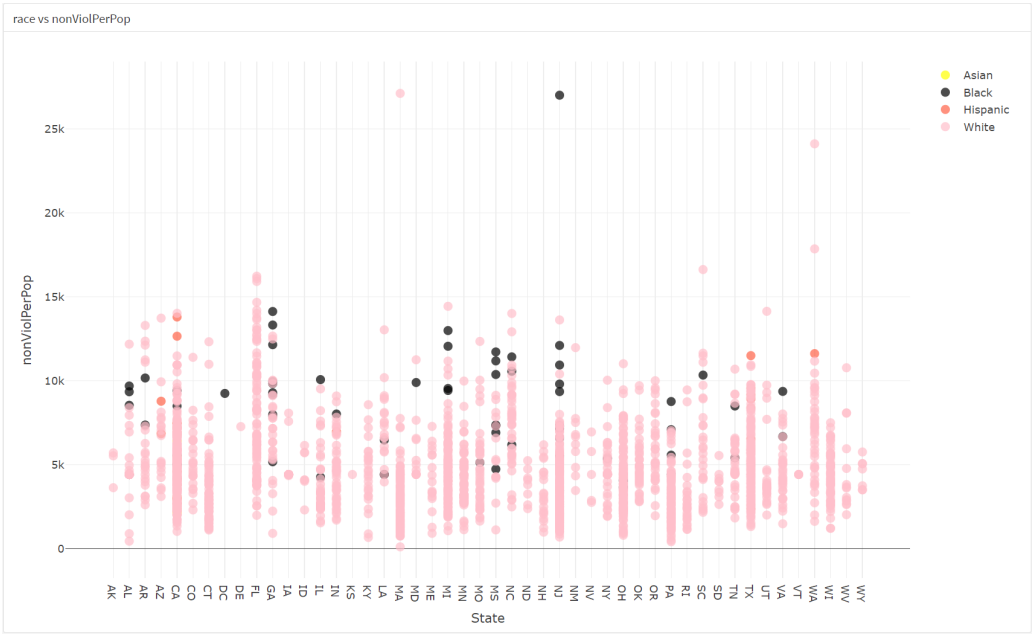

Figure 27: Race v.s. Non-violent Crime Rate

The community with the highest number of black people always has more violent crimes, followed by white people. However, community with the highest number of white people has totally more nonviolent crimes. Nonviolent crimes are large more than violent crimes no mater the race.

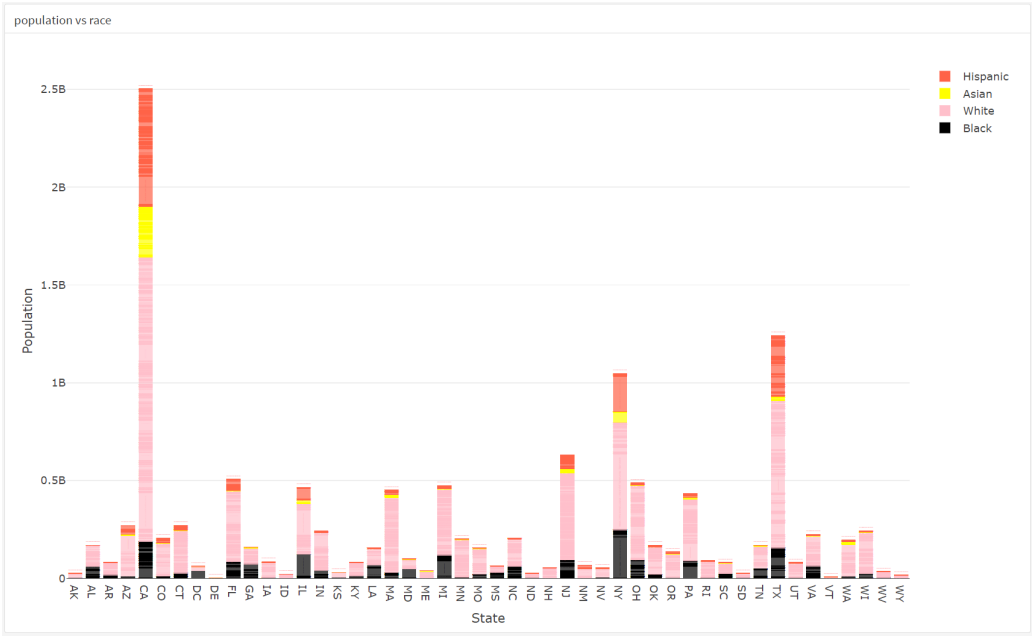

Figure 28: Population v.s. Race

California has largest population ,Hispanic people and white people. Texas has second largest population, followed by New York and New Jersey.

8 Conclusion

These analyses reveal a complex relationship between crime rates and economic factors, racial composition and age, suggesting that crime is not just a single social problem, but is influenced by the interaction of multiple factors. An in-depth analysis of the association between a number of social factors and crime rates in the United States reveals that areas with higher per capita incomes typically have lower crime rates, while areas with higher levels of poverty and lower levels of education tend to face higher crime pressures. This phenomenon may be closely related to the scarcity of resources, higher unemployment and the existence of social inequalities.

In addition, racial composition plays a significant role in crime rates. There are differences in the extent to which different racial groups are involved in different types of crime, and certain racial groups may be exposed to a higher risk of crime in particular neighborhoods.

These analyses provide a more complete understanding of the spatial distribution of crime rates across U.S. states and the multiple factors that influence this distribution. This understanding provides policymakers with a more specific social context to help them design more targeted and effective interventions to reduce crime rates.

9 Contribution

Name_studentID	Description of responsibility	Rating chosen from
Cai Xinlu 2230034002	Task 1 + Data Preprocessing	100
Li Weilin 2230034027	Task 2	100
LYU Shirui 2230026117	Task 3	100
Zhan Hanhui 2230034060	Task 4	100
Wang Shumin 2130034036	Task 5	100

10 Declaration

Since shiny was used in our code, it was not possible to generate html files, so two versions of our code were committed. One is using shiny code and the other is a simplified version of the previous code with a slight adjustment without using shiny features.