Data management and visualization Assignment (CA 682)

Topic: Racial and gender Disparities in Police Stops: A Comprehensive Examination in Colorado Before and After Marijuana Legalization

Name : Makarand Thorat

Student number : 23262520

Name : Jai Waghmare

Student number : 23265906

Screencast Recording :
https://drive.google.com/file/d/1DdGuWG25IcJU3oCD5lDyCpvJFrw6m77I/view?usp=drive_link

Racial and gender Disparities in Police Stops: A Comprehensive Examination in Colorado Before and After Marijuana Legalization

Abstract

In Colorado, United States of America law enforcement officers possess the authority to stop a vehicle under various instances. This authority includes violations of traffic rules, such as running over a red light or engaging in any form of activity that violates the state's regulations. Moreover, the officers have the right to stop drivers suspected of engaging in unlawful activities, including the illegal possession of drugs or firearms. With the optimum aim of maintaining the law and protecting the well-being of the community, law enforcement officers are trained to make use of their judgment and adhere to established protocols when stopping a vehicle or detaining an individual.

In Colorado, with regard to police stops there was a biased trend encountered by the black drivers. They were stopped more frequently and a search was conducted when compared to their Hispanic and white counterparts. In 2013 the Colorado state law decided to legalize recreational marijuana across the state territory, a significant drop in the number of searches was seen after this law was sanctioned. But still, according to the data, changes amended in laws pertaining to marijuana influenced search rates, but disparities in traffic stops and searches still persist. This issue raises questions about racial bias faced against black drivers in the state of Colorado.

To draw an inference and conclude the above statement, we visualized the search rate conducted in the state of Colorado and It was evidently visible that despite changing the drug laws the black drivers faced racial discrimination.

Data collection

The dataset used for our assignment was sourced from the open policing website, an insightful open-source website for law enforcement-related datasets. The data comprises more than 3 million rows and weighs 76.6 MB, justifying the volume aspect of big data. Each row in the dataset relates to a specific police stop faced by a driver and explains whether or not a search was conducted, thus capturing a wide array of information. The dataset has numerous columns, each giving distinct information about the stop. Initially, the dataset had 20 columns, some key columns that contributed towards our study and were considered for our visualization were:

- Date
- Subject_race
- Subject_sex
- Search_conducted

The dataset includes the presence of numerical data, strings, Boolean values, and datetime entries that enhance the complexity hence, giving a detailed exploration of policing patterns and race bias.

Data Exploration, Processing, Cleaning and/or Integration

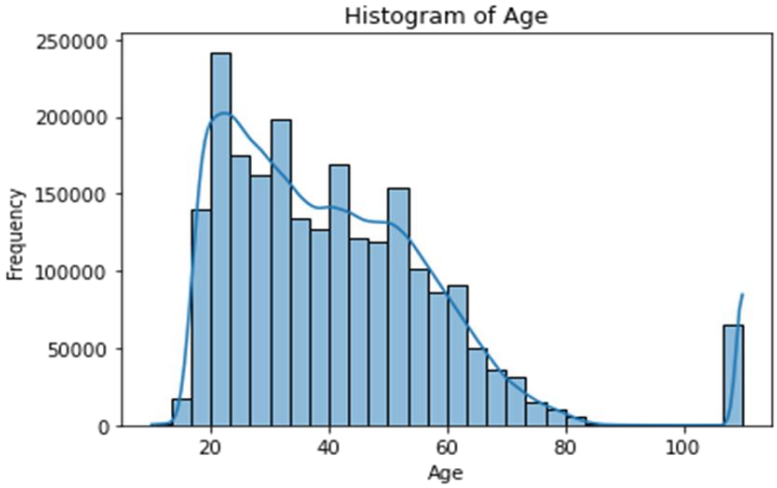
To prepare the dataset for visualization, the dataset was downloaded from the open policing website. This website has data on many states in the United States of America. Our focus was on the state of Colorado, the reason being it was relevant to our project. In 2013, the legalization of recreational marijuana in Colorado resulted in a unique instance for our analysis. The initial dataset comprised 20 columns and over 3 million rows. We carefully examined the data and only kept those columns that aligned with our visualization goals.

"**Date**"(indicating the date of stop), "**subject_race**"(ethical race of the person), "**subject_sex**"(identifies the gender), and "**search_conducted**"(contained a Boolean value indicating whether or not a search was conducted)

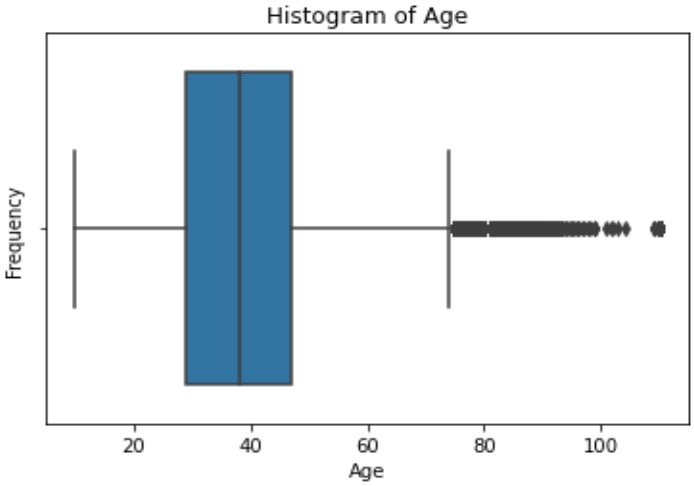
The reason to focus on these particular features helped to fast-track the trends in police stops, specifically examining the racial disparities and search conduct. We observed that black drivers were interrogated and a search was conducted more times than the whites and Hispanics. The thorough exploration, processing, and cleaning of the dataset was essential to deriving significant information and communicating our insights through the presented visualization—A line graph explaining the disparities in police stops and searches conducted among white, black, and Hispanic drivers in Colorado.

The data cleaning process involved several steps to make sure the visualization gave optimum results with the findings being reliable and accurate. Missing values were addressed, correct data types were assigned, outliers were considered, and irrelevant columns were not considered for further evaluation. Feature selection played a crucial role in streamlining the dataset for our study, allowing us to focus on the most pertinent information. The data cleaning was carried out in the following steps :

- The dataset incorporated many duplicate records where multiple rows had the same value and hence these records were dropped.
  subset=["date","location","county_name","subject_age","officer_id_hash"], all these columns contained exact same values
- The dataset included rows where all the values were *"Nan"*, after carefully examining these records are deleted.
- The rows where the values of search_conducted were *"nan"* were also dropped.
- To address missing values in the 'subject_age' column, we opted to replace them with the median instead of the mean. This decision was driven by the desire to avoid potential skewing of results, as indicated by a histogram analysis. Using the median gave a robust solution, reducing the impact of outliers and ensuring a more representative central tendency in the age distribution.



- As per the research conducted it was found that theminimum age for driving a vehicle in Colorado is 15. The records with "subject_age" < 15 were dropped from the database to get desirable output. A boxplot was plotted to confirm to visualize the presence of outliers.



- Our visualization exclusively focused on the black, white, and Hispanic racial categories. This was deliberately done because other categories had significantly lower frequencies, which would have compromised results. Focusing on these 3 major racial groups helped us to draw clearer insights and a more meaningful visualization.
- We conducted an examination of data types and identified the need to convert the 'date' column to a datetime datatype. This conversion helped to extract the year from each entry, enabling us to better analyze and visualize temporal trends in police stops. A new column “Year” was added to the data frame.
- We categorized the 'subject_age' column into three groups based on age magnitude. Individuals aged 60 and above were labeled as 'Senior,' those aged between 30 and 59 as 'Adult,' and those below 30 as 'Young.'

Data Processing

- The dataset was divided into three separate data frames for Black, White, and Hispanic races. This segregation assisted in focused analysis on individual racial groups, allowing us to later combine these data frames for comprehensive visualizations.
- To visualize/analyze the true values in “search_conducted ” we introduced a crucial metric “search_rate”.
- This metric was calculated to represent the percentage of drivers searched among every 500th driver.
- The "search_rate" metric was calculated for each of the three individual data frames representing Black, White, and Hispanic races.
- Subsequently, these data frames were combined and the columns that contributed to the visualizations were :
    - Year
    - Subject_race
    - Search_conducted
    - Subject_sex

Visualization

We have visualized the distribution between the search rate per 500 stops by year and filtered them by race (Black, Hispanic, and White) and also by gender.

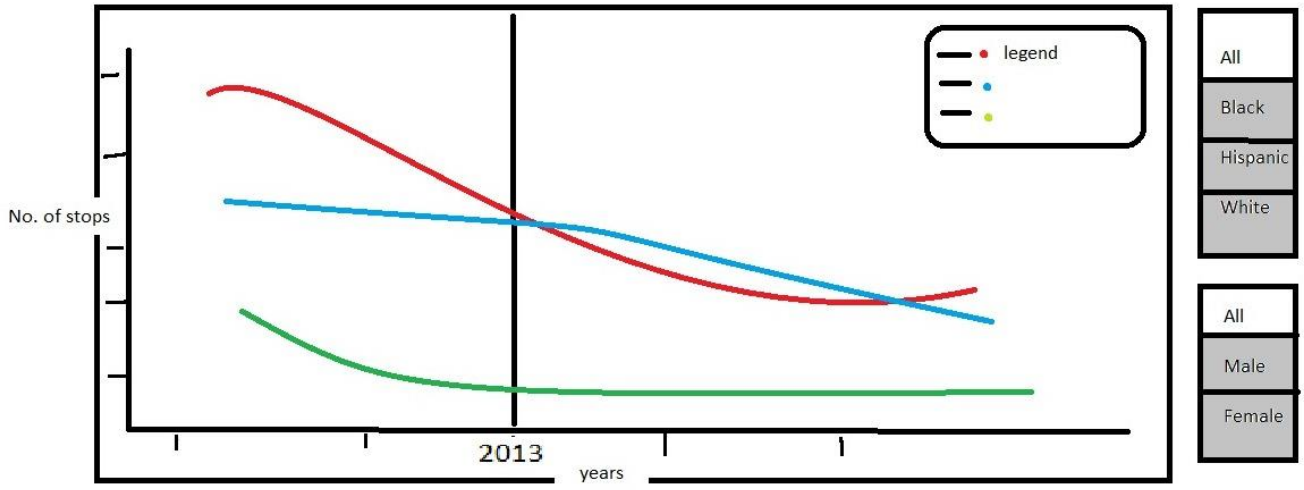
Libraries used for visualization:

1. Seaborn

2. Matplotlib

3. ipywidgets

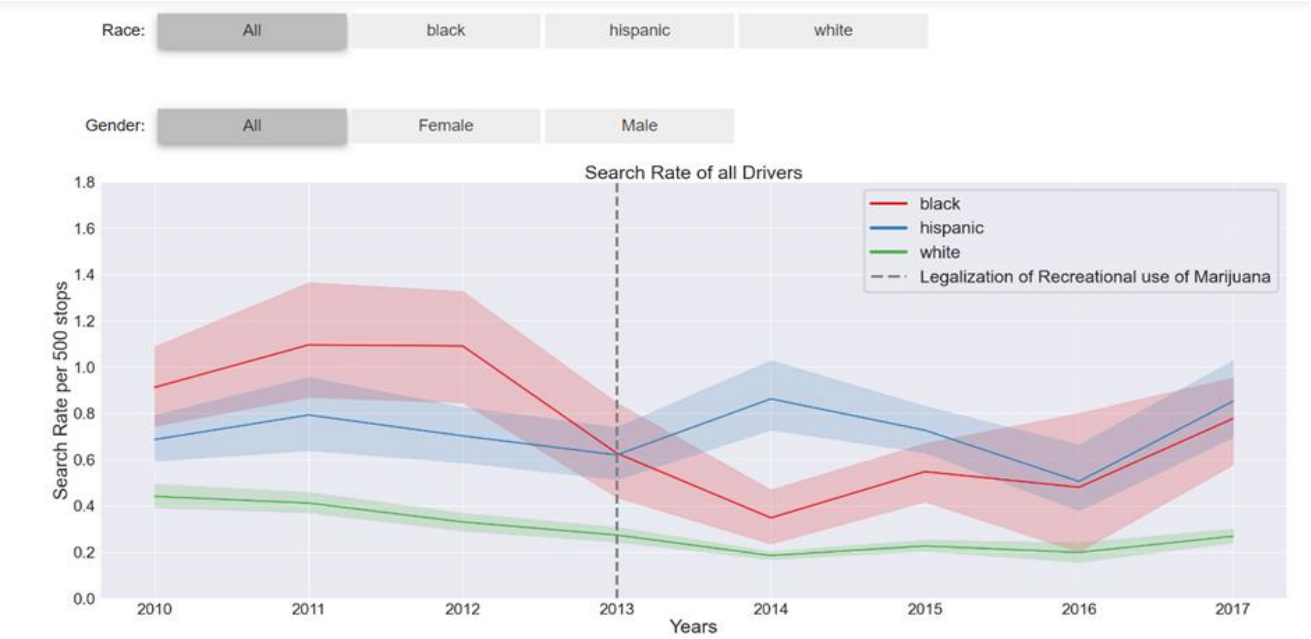
*** The envisioned sketch for our visualization encapsulated the desired appearance and structure we aimed to achieve in the final representation.***



- We first created a basic line plot using the Seaborn library to visualize the entire dataset which had combined search rates of all the drivers of all the races.
- We then set up the line of division for the year 2013 to mark the year when recreational use of marijuana was legalized.
- We then set up the plot details like xticks, yticks, and legend.
- After the basic framework of the plot was completed we introduced the ipywidgets library to make the plots more interactive.
- First, we added only the Race subset where we showed the distribution of data for particular selected data.
- The toggle button seemed to be the appropriate widget to do so and gave a simple yet good look to the overall plot, also tried with a dropdown which is less user-friendly as compared to a toggle button.
- We passed the user-selected value to the function using the interact method of ipywidgets and added if statements to select the data depending on what the value is passed.
- After this we introduced the gender buttons which would make the interaction more complex, We identified the following combinations that could be possible :

  a. All Race / All Gender
  b. Specific race / All Gender
  c. All Race / Specific Gender

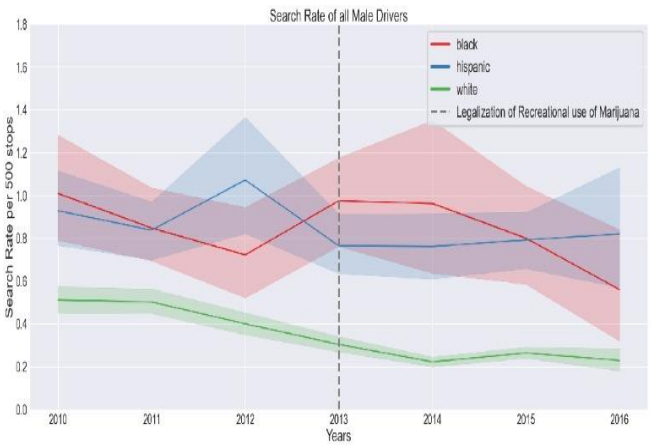
- We used a combination of nested if-else loops to plot the above-mentioned cases and since the data was already divided into separate data frames we just had to change plot titles and the data frame passed in sns.line plot() function
- The last part of the visualization was using the Layouts module of ipywidgets to arrange the layout of buttons and tweak the sizes of elements (lines, legend,..etc) to appropriate sizes.

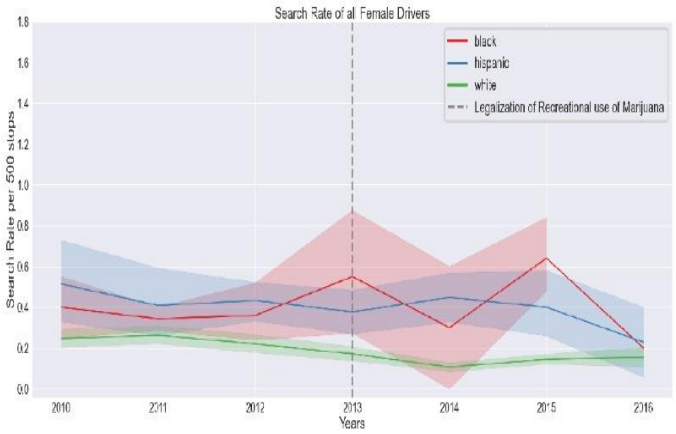


Conclusions

The data exploration, cleaning, and processing steps were time-intensive, with a challenge arising in determining a metric for calculating the "search_rate." The imbalance between True and False values in the "search_conducted" column required careful consideration. Using Python's Jupyter Notebook, our visualization project involved working with a cleaned dataset containing over 2 million rows. The computation of search rates was time-consuming, thus we divided the datasets based on different races and gender categories. Ultimately, we combined the distributed datasets into a unified dataset, forming the foundation for our visualization analysis.

In our visualization, the important question answered was the impact of the legalization of recreational marijuana on search rates in 2013. Notably, an evident decrease in search rates was observed overall following this legislative change. However, the racial disparity still remained, particularly for black drivers. Interestingly, there was no significant change in search rates for Hispanic drivers. In contrast, the number of white drivers subjected to searches displayed an approximately linear decrease over the years.





The diagrams shown above lead to another conclusion, A higher frequency of targeting was observed for males, especially black and Hispanic males, compared to their white counterparts. The disparity in search rates for females was notably less significant, with considerably fewer searches conducted on females compared to males. This finding marks a gender-related disparity, indicating a potential focus on males, particularly those belonging to black and Hispanic racial groups, in law enforcement activities.

References

- Dataset : https://openpolicing.stanford.edu/data/
- Proof of legal driving age in colorado:https://www.codot.gov/safety/colorado-teen-drivers/parent/teen-driving-restrictions.html
- Interactive python plots : https://www.youtube.com/watch?v=jWT-HXv0LUQ&ab_channel=NeuralNine
- Seaborn official website : https://seaborn.pydata.org/generated/seaborn.lineplot.html
- IPywidgets official website : https://ipywidgets.readthedocs.io/en/stable/