Na’Jae Batts

November 30, 2024

Dr. Nerolu

Final Project: Insights on Traffic and Drug Violations

Objective: The objective of my final project is to present data wrangling, inspecting the data, feature engineering, data analysis, data visualization, and storytelling on traffic stop data. The dataset I chose is called Traffic and Drug Related Violations Dataset and I found it on Kaggle. This dataset has various information about traffic stops and violations along with information of the race, gender, and why the individuals were stopped. The reason I felt like this dataset was important to analyze was because driving is very important but can also be very dangerous. Therefore, it's important to analyze the different violations, and when traffic might be the heaviest. The dataset also shows us how some police conduct their work.

Introduction: The dataset Traffic and Drug Related Violations has a lot of interesting components to it. The Traffic and Data Related Violations Dataset consists of multiple rows of traffic stops and 15 columns. Brief overview of the key columns:

- **stop_date:** Date of violation
- **stop_time:** Time of violation
- **driver_gender:** Gender of violators (Male-M, Female-F)
- **driver_age:** Age of violators
- **driver_race:** Race of violators
- **violation:** Categories of violations: Speeding, Moving Violation (Reckless Driving, Hit and run, Assaulting another driver, pedestrian, improper turns and lane changes, etc),Equipment (Window tint violations, Headlight/taillights out, Loud exhaust, Cracked windshield, etc.), Registration/Plates,Seat Belt, or other (Call for Service, Violation of City/Town Ordinance, Suspicious Person, Motorist Assist/Courtesy, etc.)
- **search_conducted:** Whether search is conducted in True or False form
- **stop_outcome:** Result of violation
- **is_arrested:** Whether a person was arrested in True or False form
- **stop_duration:** Detained time for violators approx(in minutes)
- **drugs_related_stop:** Whether a person was involved in drug crime (True,False)

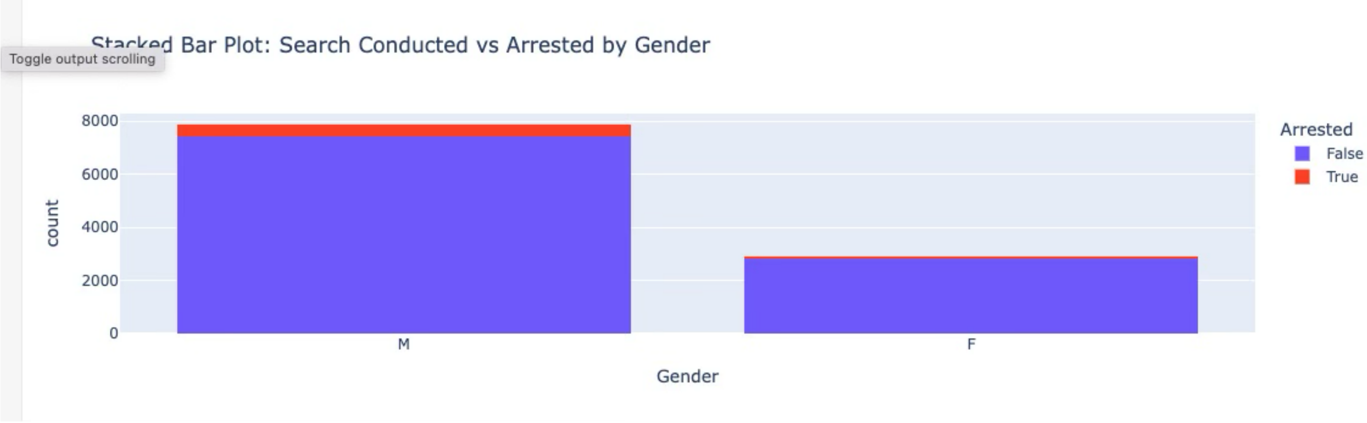
Method:

After loading the dataset into Jupyter notebook using pandas, NumPy, matplotlib, and seaborn it displayed the csv data for me which contained 11292 rows and 15 columns. I then inspected the data by displaying the first 10 rows of the data. After seeing the first 10 rows I wanted to print the column names and the data types so I could know what I'm working with and how to handle it. The data types tell us that our factors (which are our columns) are stored as an object or a float64. An object in our dataset means it contains string data which is regular text. A float64 in our dataset means we have floating point numbers which are decimal points, and 64 is the measure of memory so basically saying it's a very small or very large number.

It was now time to see how much I had to clean up the dataset by looking at duplicate values and missing values. When printing the number of duplicate values, it said that in the dataset there were 76 rows of duplicate values. Those are a lot of duplicate values, and a reason for it could be human error, data collection methods, or inherent characteristics of the system or process generating the data. I knew that after seeing these duplicate rows I had to drop them from the dataset to clean it up and that's what I did. The next step is data cleaning is to see how many missing values I had which ended up being a lot, over 5000 missing values. Before figuring out how I was going to either fill, drop, or manipulate the missing values, I copied the first data frame into a new one. The first column I knew I could drop was the country_name. I knew I could drop this column because the whole column was NULL. The reason for this could be often related to data collection, processing, or structural issues. Next to move forward with data cleaning, I decided to drop rows that had 3 or more empty rows in the dataset. After doing this it removed most of the missing values, and that left me with driver age and search type. For search type I decided to row where 'search type' is NaN, the value of search type is replaced with the string "Search not conducted". The reason why I did this is because most of the rows where there are missing values means that the police didn't conduct a search. The last column left was driver age, and this wasn't that difficult because it only had one missing value. So, I decided to combine stop date and stop time columns to create a new column called stop_datetime. This converts combined string into a datetime object using pd.to_datetime(). Then I added a new column, stop month which extracts the exact month from the stop_datetime column. I also did the same thing by adding stop year, along with extracting the hour from the same stop_datetime column to make a new column named stop hour and this tells us the first number of the hour that the traffic stop happened. Since I converted the columns into a proper date time format, I was then able to calculate the driver's age at the time of the stop. I did this by subtracting the drivers_age_raw column from stop year. This is subtracting the year that the driver was born from the year which they got stopped, which will give us the drivers age of when they got stopped. Doing this will fill the missing value I had in my data for driver age. After this I had one more step which involved handling the outliers. There were two outliers which were 0.0 in the driver age raw which is supposed to the years but that's not possible. Another possible outlier would be 2006 which is under driver age, and it doesn’t make sense for someone to be that age. After this the dataset was clean and was ready to be analyzed.

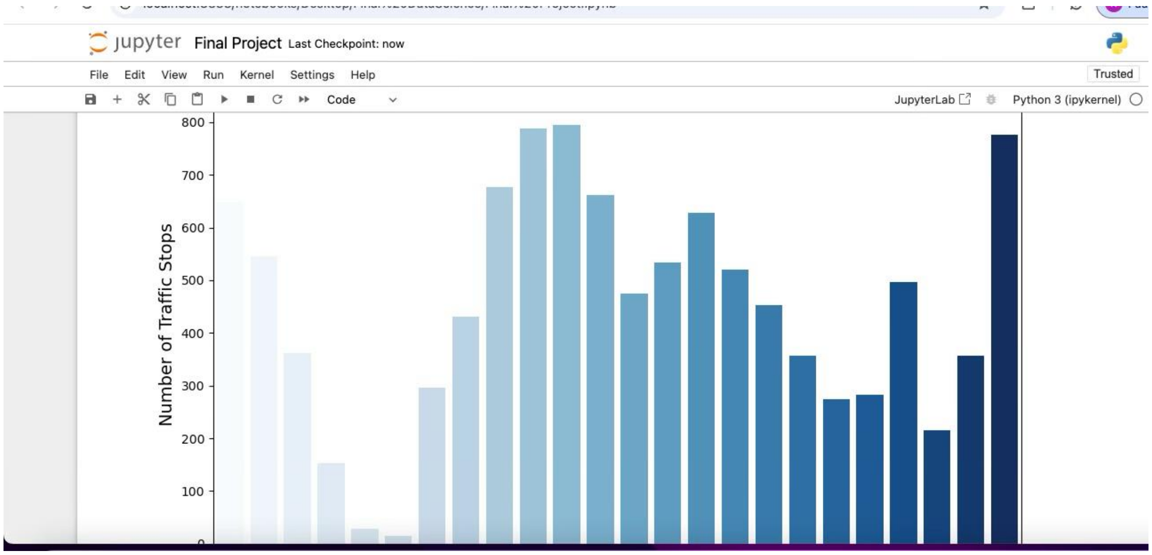
Storytelling:

Stacked Bar Plot showing Search Conducted vs Arrested by Gender



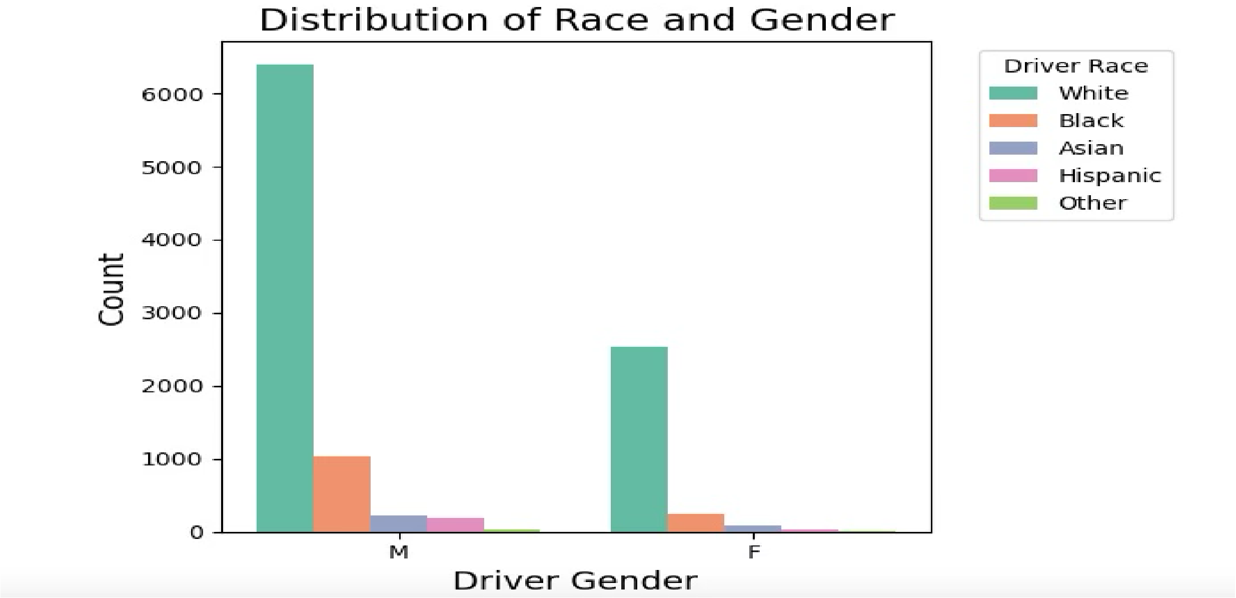
Advanced visualization I used was Stacked Bar Charts to visualize relationships between variables like search conducted and is arrested across gender. This is good for our dataset because we can see the actual count for the small bars like the arrested being true which is only at a count of 83. This visualization definitely helped me understand the dataset better. This visualization shows that mostly males had a search conducted than females. It also shows that in males that had a search conducted only a little got arrested, and the same goes for females.

Hourly Trend of Traffic Stops



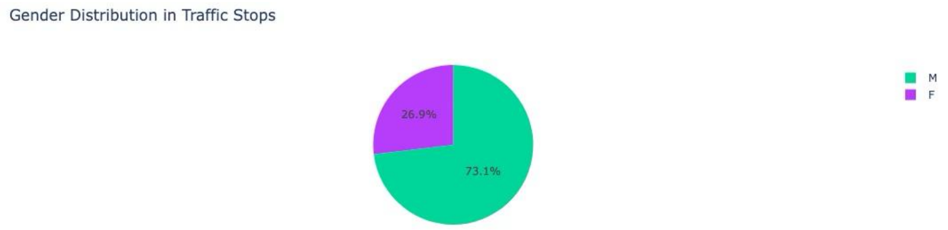
The peak times for traffic stops are in the morning (8-10 am), and at night (11pm), with a little peak around midday around 2pm. During the morning it makes sense because people are making commutes to work and school with lots of traffic. Around 11pm makes sense as well because people usually speed more at night when people aren't on the road. This distribution was interesting to see because further applications to the real world can be applied by these results.

Distribution of Driver Gender and Race



The gender distribution of the drivers is male and female, with male being significantly higher than females. The male count is around 7,000 while the female count is near 3,000. Between both genders, the majority race is white. In males there are more black people than black females for the traffic stop. The other races such as Asian, Hispanic, and other are around the same amount for both genders. A surprising trend was the majority of the race being white in our distribution. I feel like if the dataset was updated to be more recent than the majority race would definitely be different. However, there is a huge imbalance between the number of males vs females in the dataset.

Gender Distribution in Traffic Stops



Using advanced visualizations, I chose to use Plotly to create an interactive pie chart, which allows users to hover over the chart for additional insights such as percentages and counts. As you can see, it's the same percentages as the other pie chart, but for this one you can hover over it and see the actual count. These advanced visualizations help you understand the dataset better. This pie chart does a good job showing how many males are in the dataset and how they are the majority with 73%, while females are 26%.

Conclusion:

In conclusion, this dataset, Traffic and Drug Related violations was an interesting dataset. The quality of the dataset was good however it had many missing values, duplicates, and outliers. After I cleaned the dataset, it was ready to be visualized and analyzed. The dataset was pretty huge so that gave me an advantage to see statistical significance and to generalize findings. Overall, the individuals that had the most traffic stops were white males between the ages of 20- mid 30s. The traffic stops were done the most either in the morning or late at night. Most of these violations were from speeding and didn't end up in being arrested, along with most of these stops not being related to drugs. Some actions that can be put in place from this dataset could be implementing traffic safety especially during the peak times which are morning and late at night such as speeding cameras so that could be limited. Another one is utilizing arrest and search data to determine when it's necessary for police and improve their efficiency in determining when searches are necessary. For instance, if certain violation types (e.g., drugs-related stops) consistently lead to arrests, this data can help in developing predictive policing models to prioritize resources and improve decision-making during stops. The only major thing I would change about this dataset is updating it so it can be recent years. I know that the dataset would be very different in some areas if it was more recent. Overall, this dataset was very interesting to analyze.

References: Kaggle: Traffic and Drugs Related Violations Dataset

Acknowledgments:

I want to thank Dr. Nerolu for introducing my classmates and I to Jupyter Notebook, data science, and teaching us how to visualize datasets. I have learned so much during this semester!