Cristian Simms
Nerolu, Meenakshi
December 3, 2024
Intro to Data Science

Traffic and Drug-Related Violations Analyzation

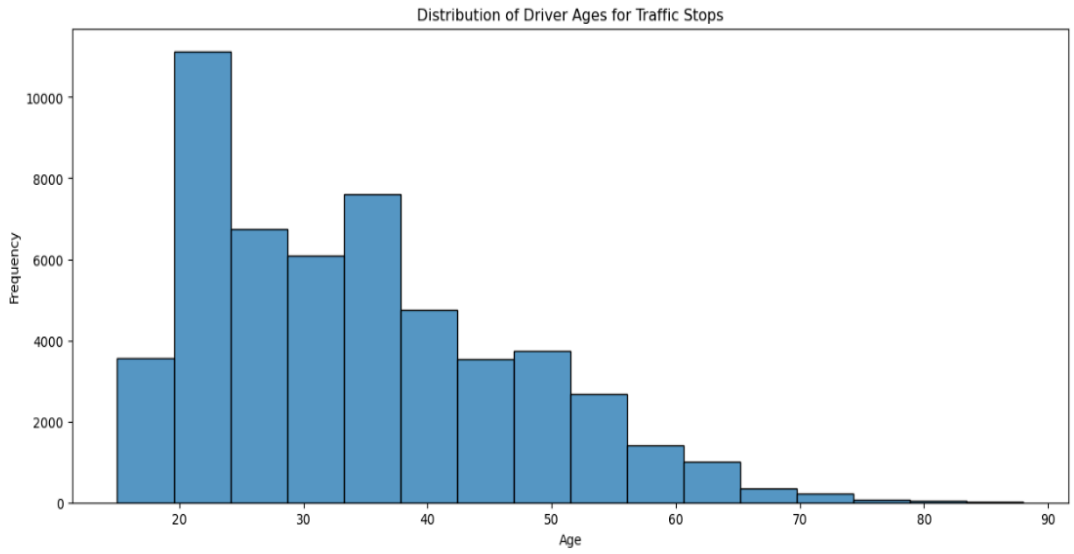
Objective:
The primary objective of this report is to analyze traffic stop data that will lead to conclusions about traffic patterns, trends, and potential disparities that exist in law enforcement tactics. In this report, we will analyze the following; trends regarding the most common violation, reasons for drug-related traffic stops, understanding of demographic disparities, access temporal patterns for traffic stops, as well as insights that can improve law enforcement strategies. This large dataset provides immense flexibility and the freedom to understand a wide range of positive and negative patterns in law enforcement. By understanding this data, we can provide evidence-based solutions and contribute valuable information to not only the enforcement department but also the community.

Introduction:
Law Enforcement plays a major role in upholding public safety and ensuring there is a sense of order throughout. A major department that handles cases leading to public safety is traffic enforcement. They play a major role in maintaining road safety as well as upholding law and order. This report focuses on analyzing the "Traffic and Drugs Related Violations Dataset," which contains over 65,000 records detailing traffic violations and stops. This dataset highlights the diverse categories that make up a daily traffic stop. The key attributes present in this dataset are: *stop_date, stop_time, driver_gender, driver_age, driver_race, violation, search_conducted,* and outcomes such as *is_arrested and drugs_related_stop.*
When analyzing large pools of data related to traffic stops we can better improve the methodologies necessary to ensure public safety. Throughout this process, patterns will be revealed that can shed light on societal issues such as violation trends, geographical and demographic disparities, as well as the presence of criminal activity. This report uses multiple methods through Python to properly analyze this code, such as exploratory data analysis (EDA), data visualization, cleaning missing values, and data wrangling. Throughout this report, we aim to transform the raw data into a concise and readable interpretation which will be used to emphasize the importance of data-driven decision-making for traffic law enforcement.

Method (e.g., Data Wrangling, Handling Missing Values)
This dataset included over 50,000 cases of information that consisted of traffic stops, violations, time of day, and outcomes of traffic stops. The first step taken in the analysis was to read the CSV file into Python (Jupyter Notebook) using the *pandas* library. We accomplished this by using the function *df = pd.read_csv("3_traffic_violaions.csv").* We labeled the dataset variable with the title of *df*. Once we began the process of cleaning the dataset we made a copy of this CSV file and labeled it in *df_new* by using the ode function *df_new = df.copy()*. Before we could begin analyzing and researching patterns, we had to first check if this file was missing any data as we wanted to ensure our findings were as accurate as possible. By using the *isnull* function, we discovered that all but one column had at least one missing piece of data. For the column titled *country_name*, it had almost 53,000 missing values. In this example, we decided to drop

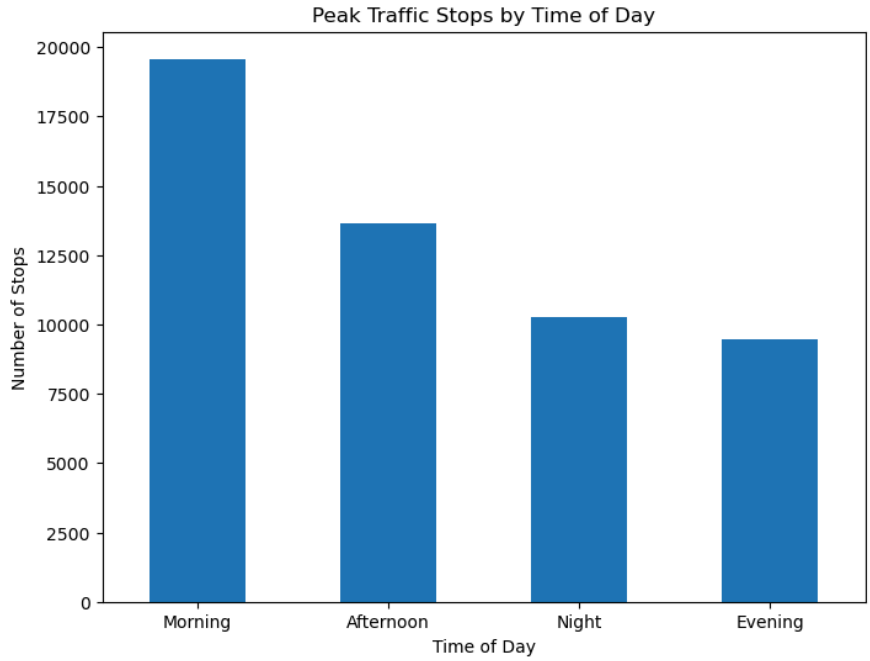
this entire column from the dataset using the input *df_new=df_new.drop(columns=['country_name']).* This was done because with such a large amount absent, there is no value in keeping this column around as our goal in data wrangling is to make this dataset as concise as possible so analyzing it can be as simple as possible. Each column that had missing values was assessed individually and we used specific methods to clean each one based on our goals. For the time-series data and columns with only 1 missing value such as *search_conducted* and *drugs_related_stops*, we used the *fillna* method. This replaced the missing value with the value that occurred either before or after it. For the rest of the columns with missing values, I used the imputation method to clean the data. Numerical columns such as driver_age were imputed with their mean values (df['column'].fillna(df['column'].mean())), ensuring a balanced approach without skewing the data distribution. I used the mode function to fill in the missing values for all the *categorical columns: driver_gender, driver_race, violation_raw, violation, stop_outcome, is_arrested.* Another data wrangling method oftentimes used is the removal of duplicate rows using the *drop_duplicates()* function. I decided to keep all duplicate rows as it could show a potential pattern in the type of arrests later on in the data for analysis purposes. By cleaning this data, we are now able to accurately proceed with data visualizations and statistical analyses that will allow us to uncover valuable insights into the inner workings of all traffic stops in this dataset.

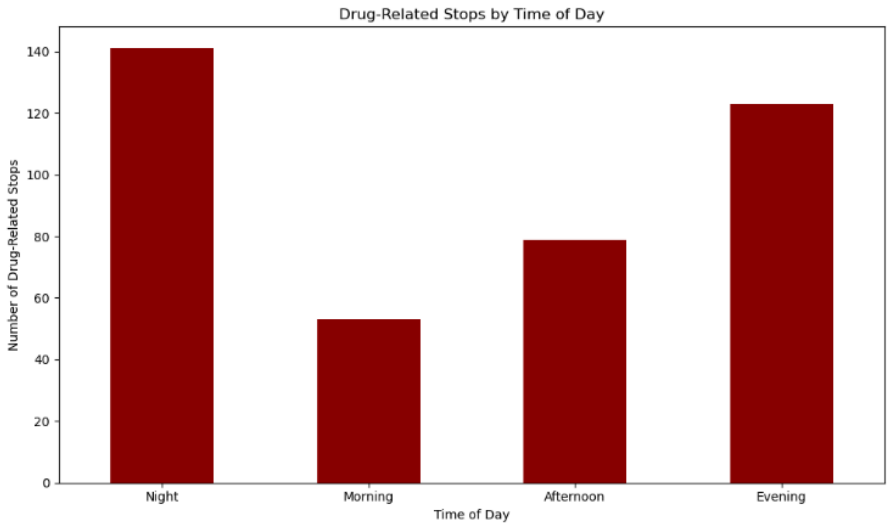
Storytelling:



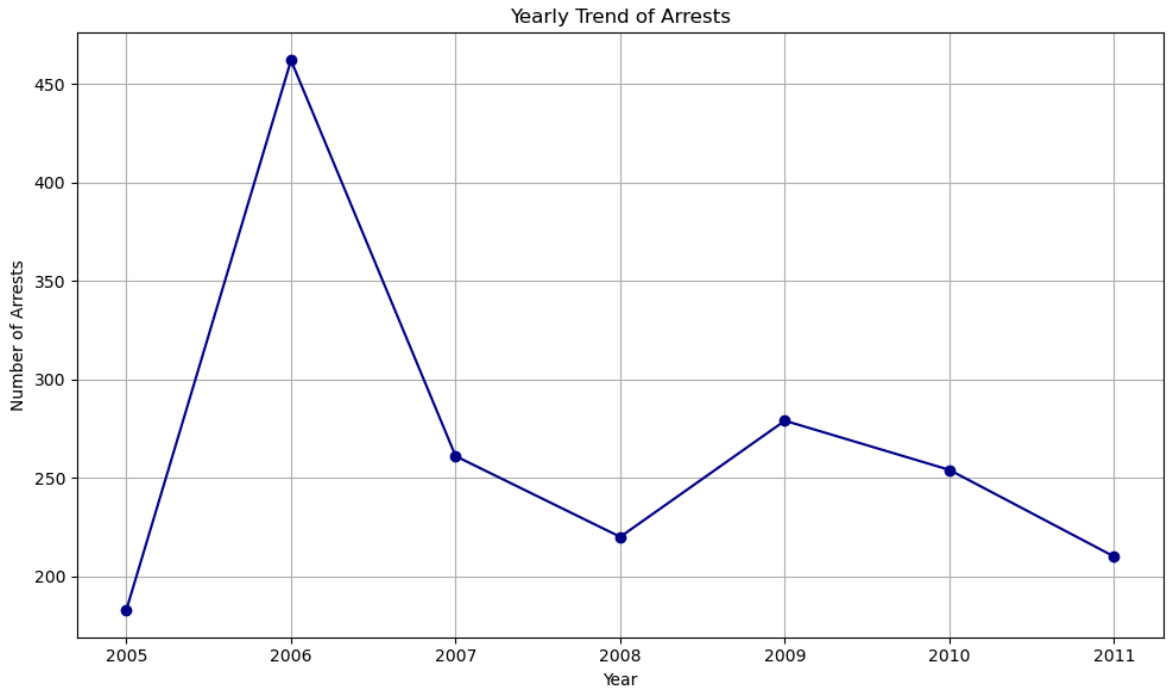
The above histogram, titled “Distribution of Driver Ages for Traffic Stops,” gives us a visualization of the drivers' ages that obtain the most traffic stops. A very valuable insight that can be seen from this plot is the fact the age group with the highest traffic stops spans from 25-40 years old. Also, this histogram is skewed right, which provides valuable insight into the correlation that increased age leads to fewer traffic stops. This distribution highlights the fact that drivers in their prime years either have increased driving activity or are simply reckless drivers. Another inference using real-world knowledge is that many new drivers between 18 and 24 are typically still in school and therefore are not driving as often. This shortens the pool of data for that age group which is a likely reason why they have low counts for traffic stops

.





These two bar charts showcase very valuable information. The blue bar chart showcases, “Peak Traffic Stops by Time of Day,” and the red bar chart showcases, “Drug Related Stops by Time of Day.” For traffic stops in general, which include drug-related stops as well, the time of day with the highest number is the morning (5 AM to 11 AM). This may be attributed to the morning rush hour, when more vehicles are on the road, potentially leading to increased enforcement. The time of day with the lowest number of stops is during the Night which is likely due to reduced road activity as most civilians are asleep. However, a key insight that can be concluded is that despite Night being the lowest overall for traffic stops, it is still the time of day with the highest amount of drug-related stops. This insight can lead to several inferences. The night is a time often associated with increased illegal activity or impaired driving and Law enforcement may target this period due to heightened suspicion of drug-related offenses. This allows us to conclude that the night patrol officers are favoring only stopping a vehicle if they suspect it of drugs.



To analyze the trend of arrests over the years, we deduced that a line plot would visualize this trend in the most readable way. The plot shows variations in arrests from year to year, which could be influenced by changes in enforcement policies, societal factors, or variations in the volume of traffic stops conducted. There is a peak in arrests in the year 2006, reaching over 450 arrests. Between the years of 2005 and 2006, it would be helpful to understand what happened to increase the arrests. This could be due to increased awareness from law enforcement, legislation changes, societal shifts, or the area where this survey was done developed a surge of increased crime. After 2006, there was a decline in over 200 arrests and this trend stayed rather constant with little fluctuation for the next several years. This could likely be due to a decrease in reckless driving due to the drastic number of arrests in 2006.

A stacked bar plot was created to visualize the relationship between the arrest rate and the specific violation that was suspected when a search was conducted. This chart highlights patterns in how specific violations have a higher chance of resulting in an arrest. Arrest rates for registration/plates and equipment are significantly lower when a search is conducted. This can be alluded to by the fact that these violations oftentimes do not warrant an arrest because it is a less severe infractions. Typically arrests will occur after a search has been conducted and these arrests tend to be for more severe offenses such as drug-related crimes.

Conclusion:

The "Traffic and Drugs-Related Violations Dataset" has allowed us to analyze and find conclusions regarding the vast dynamics of traffic stops. By implementing tools such as data visualization, exploratory data analysis, and data wrangling we explored the relationships between driver demographics, time of day, violation categories, searches, and arrests, unveiling trends that will undoubtedly provide meaningful insights for future strategies aimed at increasing public safety. After concluding this report, I have noticed that the incorporation of additional variables could have provided further insights into the patterns of traffic stops. This additional information could be in the form of geographical data and employee tracking which would have allowed for more specific inferences to be made. All in all, this report showcases the extensive data-driven analyses that can be made using tools like Python.

References:

https://www.kaggle.com/datasets/shubamsumbria/traffic-violations-dataset/data
Python Coding Platform: Jupyter Notebook

Acknowledgments:

Shubam Sumbria (creator of "Traffic and Drugs-Related Violations Dataset")
* Special thanks to Professor Nerolu for guidance and feedback during this project