DUI Incident Rates in Boston from 2019, 2021-2023

Asma Syed

2023-12-17

# About The Dataset

Is it true that there are more drunk driving incidents around the holidays and the weekends? Doing a quick search will yield a startlingly large amount of alarming statistics about DUI related crime.  
According to Driver’s ed, “25% of adults admit that they drink more during the holiday season,” and according to Injury Facts, “Saturday is the most dangerous day of the week to drive.” While these statistics are a culmination of data collected throughout the years and across the United States, an inquiry may arise about the DUI statistics within a particular city. This is where my research question comes into play, specifically about the number of DUI incidents in the city of Boston. I wanted to investigate the correlation between drunk driving incidents around the holidays, weekends, and the summer months within the city of Boston. I would then use the data to determine how well my model would be for predicting DUI’s in the next year. My data set comes from the Boston Police Department Database: <https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system> For this project, I decided to focus on the years 2019, 2021, 2022, and 2023.

# Data Management

Before doing any analysis, I needed to complete a few steps of data management.

After loading the data onto R studio, I had 4 seperate data sets with 17 variables each. The 2019 data set had a total of 87,184 observations, the 2021 data set had a total pf 71,721 observations, the 2022 data set had a total of 73,852 observations, and, lastly, the 2023 data set had a total of 74,040 observations. Note that the 2023 is incomplete as the full year has not yet elapsed, the last crime incident that was entered occured on December 12, 2023. Next, I wished to combine all 4 data sets into one. I accomplished this by utilizing cbind:

The combined data set, crimedata, consists of 306,797 observations in total. Since my focus was exclusively on DUI incidents, I needed to filter out all other crimes from the data set. I created a new data set, *crimeDUI* which consists of reported DUI crimes only and which has the following variables: + OFFENSE\_CODE - All of the crimes in the data set have a particular code + OFFENSE\_DESCRIPTION - Description of the Crime + OCCURED\_ON\_DATE - When the Crime Occured (time and date) + YEAR + MONTH + DAY\_OF\_WEEK

My new dataset has a total of 373 observations or a total of 373 reported DUI incidents within the four years.

Next, I created a column, called **dates** within the dataset that substrings the character values under the variable **MONTH** and translates them into the format I wish to work with. In this case, I wanted the yy-mm-dd format:

Then, I needed columns that distinguish holidays, weekends and summer months from ‘regular days’, ‘work days’, and ‘colder months’. I considered the following holidays: + New Years + Super Bowl Sunday + President’s Day + St. Patrick’s Day + Memorial Day + Juneteenth + Fourth of July + Labor Day + Halloween + Thanksgiving + Christmas Eve + Christmas + New Years’ Eve For the summer months, I considered May until (and including) September To do that, I used the mutate function and the case\_when function:

The last 6 observations of the data look like this:

tail(crimeDUI)

## OFFENSE\_CODE OFFENSE\_DESCRIPTION  
## 368 2101 OPERATING UNDER THE INFLUENCE (OUI) ALCOHOL  
## 369 2101 OPERATING UNDER THE INFLUENCE (OUI) ALCOHOL  
## 370 2101 OPERATING UNDER THE INFLUENCE (OUI) ALCOHOL  
## 371 2101 OPERATING UNDER THE INFLUENCE (OUI) ALCOHOL  
## 372 2101 OPERATING UNDER THE INFLUENCE (OUI) ALCOHOL  
## 373 2101 OPERATING UNDER THE INFLUENCE (OUI) ALCOHOL  
## OCCURRED\_ON\_DATE YEAR MONTH DAY\_OF\_WEEK dates holiday weekend  
## 368 2023-11-28 23:59:00+00 2023 11 Tuesday 2023-11-28 0 0  
## 369 2023-11-29 01:33:00+00 2023 11 Wednesday 2023-11-29 0 0  
## 370 2023-12-01 00:23:00+00 2023 12 Friday 2023-12-01 0 0  
## 371 2023-12-09 14:39:00+00 2023 12 Saturday 2023-12-09 0 1  
## 372 2023-12-10 00:04:00+00 2023 12 Sunday 2023-12-10 0 1  
## 373 2023-12-12 20:49:00+00 2023 12 Tuesday 2023-12-12 0 0  
## summer  
## 368 0  
## 369 0  
## 370 0  
## 371 0  
## 372 0  
## 373 0

# Statistical Analysis and Model Building

For my analysis, since the data set has only one value for the response, a linear model wouldn’t have been the ideal method to analyze data. Instead, I performed some summary statistics. i first wanted to see how many incidents there were each year:

crimeDUI %>%   
 group\_by(YEAR) %>%   
 summarize(count= n())

## # A tibble: 4 × 2  
## YEAR count  
## <int> <int>  
## 1 2019 124  
## 2 2021 79  
## 3 2022 81  
## 4 2023 89

Then I wanted to see how many incidents occurred during the holidays each year and how many occurred on regular days

crimeDUI %>%   
 group\_by(YEAR, holiday) %>%   
 summarize(count=n())

## # A tibble: 8 × 3  
## # Groups: YEAR [4]  
## YEAR holiday count  
## <int> <chr> <int>  
## 1 2019 0 122  
## 2 2019 1 2  
## 3 2021 0 75  
## 4 2021 1 4  
## 5 2022 0 78  
## 6 2022 1 3  
## 7 2023 0 87  
## 8 2023 1 2

prop.table(table(crimeDUI$holiday))

##   
## 0 1   
## 0.97050938 0.02949062

It seems like more accidents occurred outside of the holidays than during the holidays. Only about 2.6% of DUI incidents occured during the holidays

Next, I grouped the data by year and weekends, and year and summer months:

crimeDUI %>%   
 group\_by(YEAR, weekend) %>%   
 summarize(count=n())

## # A tibble: 8 × 3  
## # Groups: YEAR [4]  
## YEAR weekend count  
## <int> <chr> <int>  
## 1 2019 0 70  
## 2 2019 1 54  
## 3 2021 0 47  
## 4 2021 1 32  
## 5 2022 0 54  
## 6 2022 1 27  
## 7 2023 0 52  
## 8 2023 1 37

prop.table(table(crimeDUI$weekend))

##   
## 0 1   
## 0.5978552 0.4021448

crimeDUI %>%   
 group\_by(YEAR, summer) %>%   
 summarize(count=n())

## # A tibble: 8 × 3  
## # Groups: YEAR [4]  
## YEAR summer count  
## <int> <chr> <int>  
## 1 2019 0 72  
## 2 2019 1 52  
## 3 2021 0 49  
## 4 2021 1 30  
## 5 2022 0 53  
## 6 2022 1 28  
## 7 2023 0 51  
## 8 2023 1 38

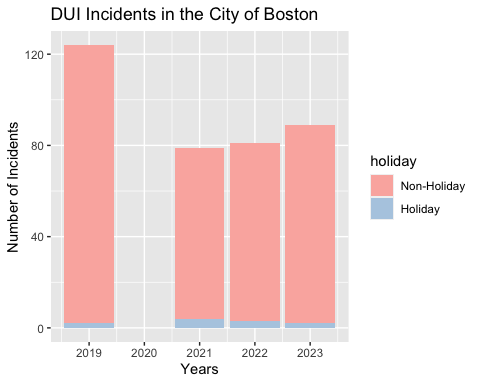
prop.table(table(crimeDUI$summer))

##   
## 0 1   
## 0.6032172 0.3967828

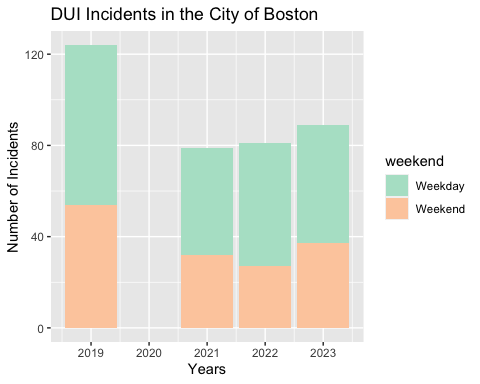
For weekends, more accidents occurred during the weekdays- about 40% of DUI incidents occurred during the weekend and 60% during the weekdays.  
For the summer months, about 40% of DUI incidents occurred between May and September while 60% occurred during the colder months.

Below are bar plots of each of the variables of interest by year:

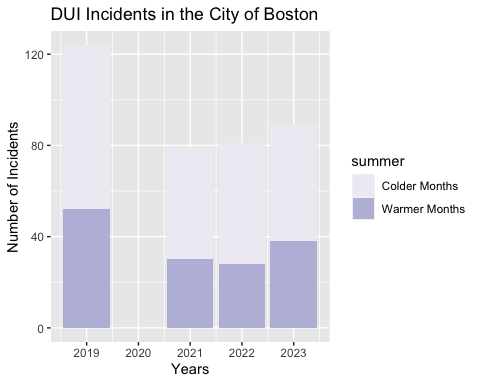
ggplot(crimeDUI, aes(x=YEAR, fill=holiday)) + geom\_bar() + ggtitle("DUI Incidents in the City of Boston") + xlab("Years") + ylab("Number of Incidents") + scale\_fill\_brewer(palette="Pastel1", labels = c("Non-Holiday", "Holiday"))



ggplot(crimeDUI, aes(x=YEAR, fill=weekend)) + geom\_bar() + ggtitle("DUI Incidents in the City of Boston") + xlab("Years") + ylab("Number of Incidents") + scale\_fill\_brewer(palette="Pastel2", labels = c("Weekday", "Weekend"))



ggplot(crimeDUI, aes(x=YEAR, fill=summer)) + geom\_bar() + ggtitle("DUI Incidents in the City of Boston") + xlab("Years") + ylab("Number of Incidents") + scale\_fill\_brewer(palette="Purples", labels = c("Colder Months", "Warmer Months"))



# Conclusions

Although my initial assumption of there being a larger proportion of DUI accidents in Boston during the holidays, weekends, and the summer months was incorrect, I can assume that because I only had data for four years, I cannot make that larger assumption that a large number of DUI accidents in Boston occur around these times of the year. Based on my analysis, I can see why there are a larger number of accidents in the winter months and this is due to the holiday season, even though not many accidents occurred on holidays themselves. Perhaps different conclusions can be reached if I were to analyze data from 10 or 15 years and that I could have gotten a different picture if I also factored in what times of the day do DUI accidents most frequently occur.