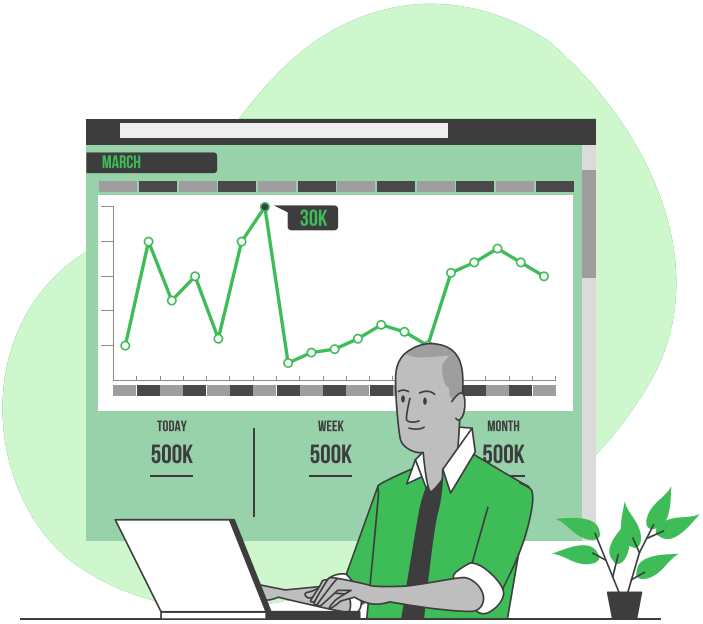
Crash Data Analysis

Subtitle: An In-depth Analysis of Traffic Crash Data
Date: 22.05.2024
Presenter: Louis

Overview



As a traffic stop analyst for the US Traffic Stop Agency, I would be conducting an in-depth analysis of car crash data. Here is an Overview of the data set:

- **Number of records**: 84,278 rows and 44 columns
- **Key columns**: Road Condition, Collision Type, Weather, Light, Road Grade, Agency Name, Crash Date/Time
- **Date range**: 2015-2022

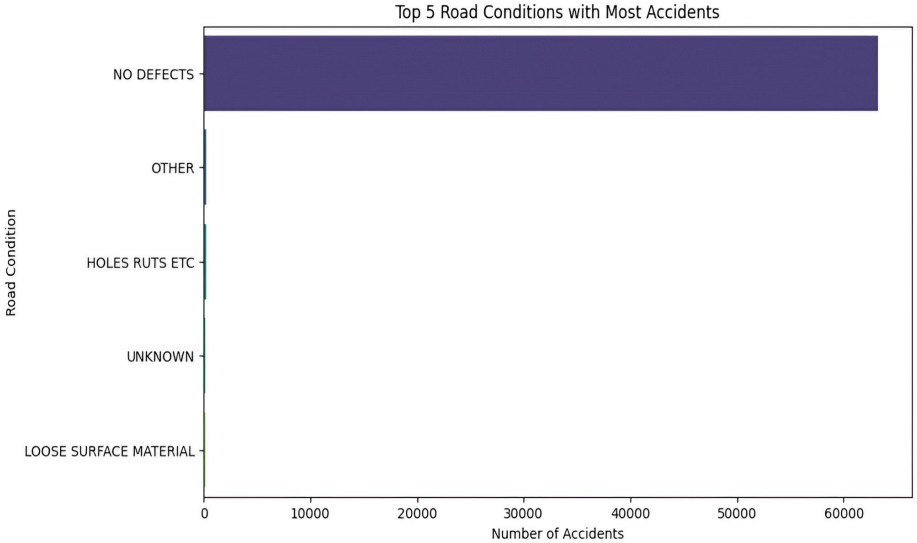
The objective was data exploration to make sense of the data set using sound statistical and visualization methods.

OBJECTIVES

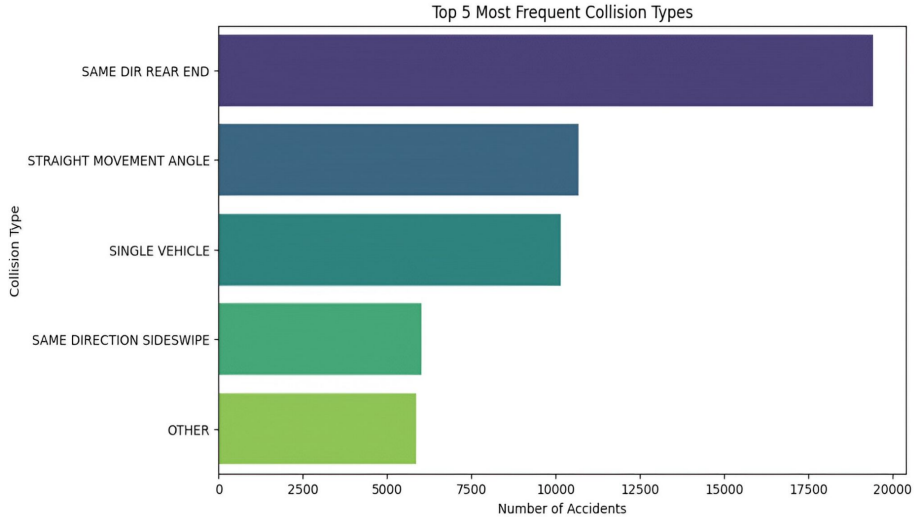
After the necessary cleaning and data wrangling, we were left with a data set of rows and columns being (64295, 47) respectively. We set out to answer the following question:

- What are top 5 “Road Conditions” with most accidents?
- What are the top 5 most frequent “'Collision Type”?
- What are the top 5 most frequent “Weather” conditions?
- What are the top 5 most frequent “'Light” conditions?
- What are the most frequent “Road Condition” and “Road grade”?
- How the number of crashed changes over day time?
- Which day of the week has more crashes?
- Get the evolution of the number of crashes per hour.
- Get the evolution of crashes per year.
- Which agency has reported more crashes?
- We could expect that the mean number of crashes over the weekends is bigger than during labour days. Make a two sample t-test to evaluate this hypothesis

Frequency Analysis

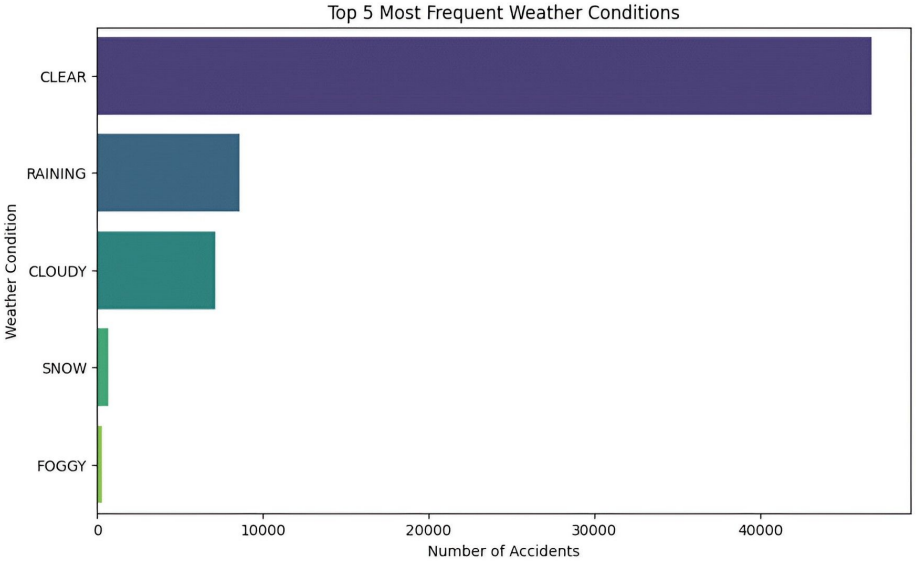


No defect roads have the most accidents: First world data might explain this.

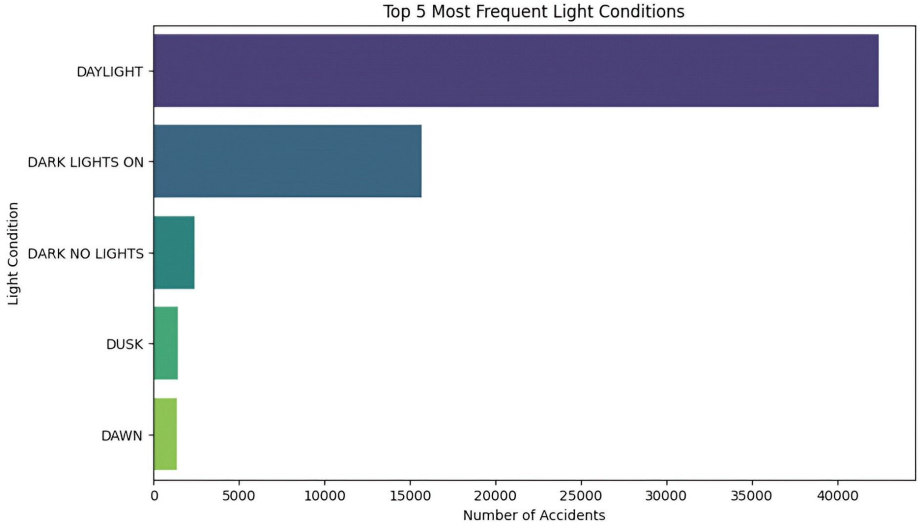


Cars moving in the same direction getting hit from the rear end is the most frequent type of accident.

Frequency Analysis

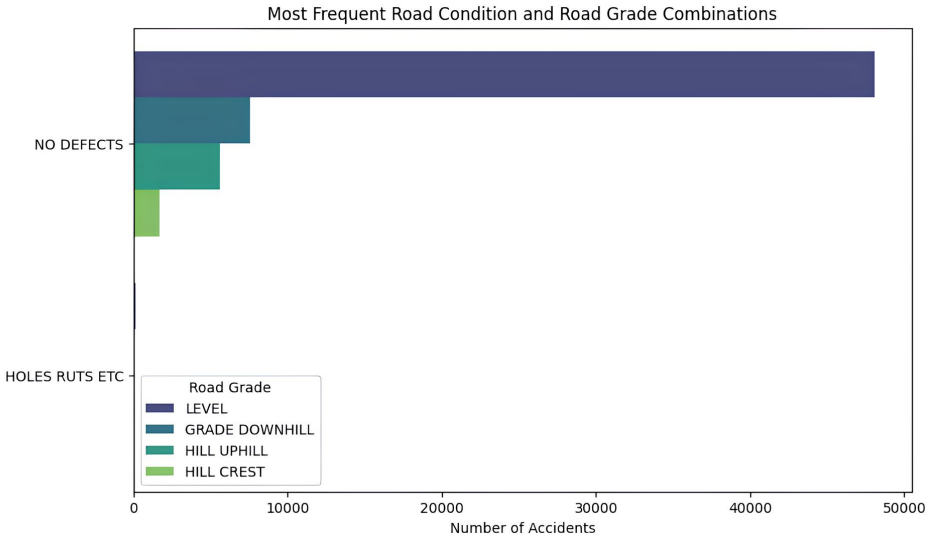


Clear weather conditions had the highest number of accident.

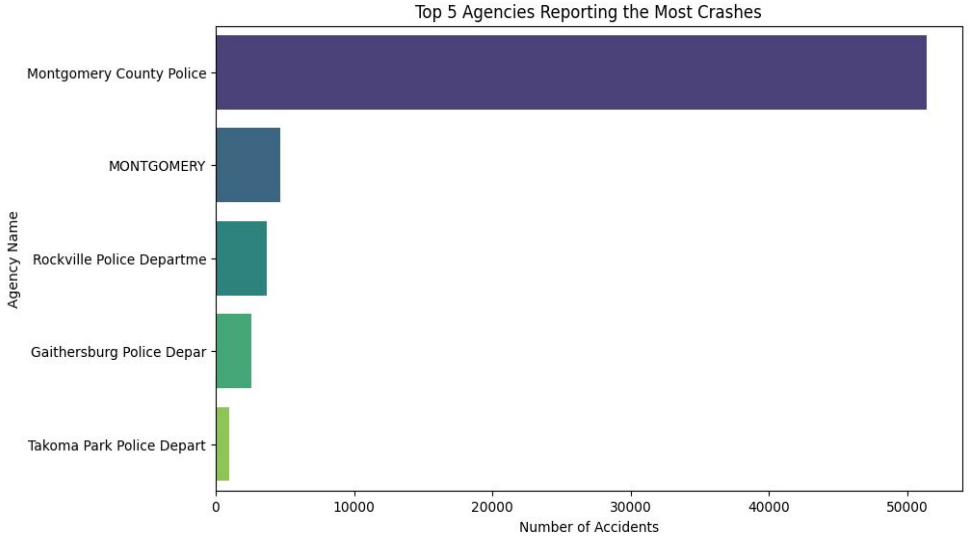


Daylight lighting conditions had the highest number of accidents.

Frequency Analysis

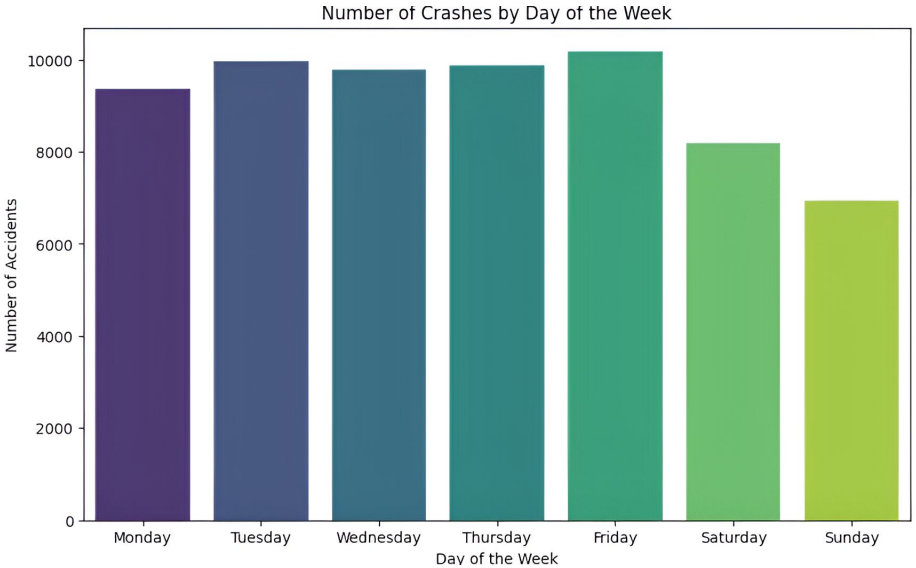


No Defects was the most frequent Road Condition as was seen and level was the most frequent Road Grade.

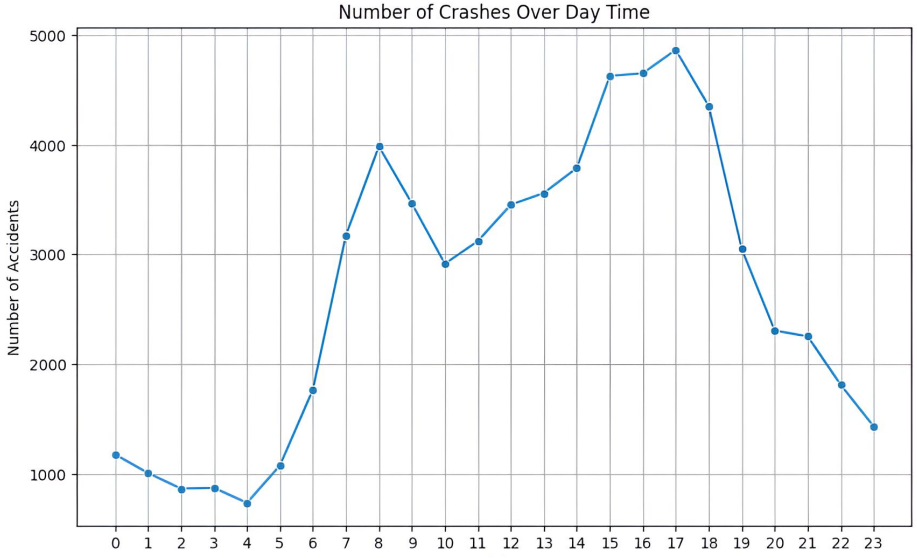


Montgomery County Police Station reported most of these crashes.

Time Series Analysis

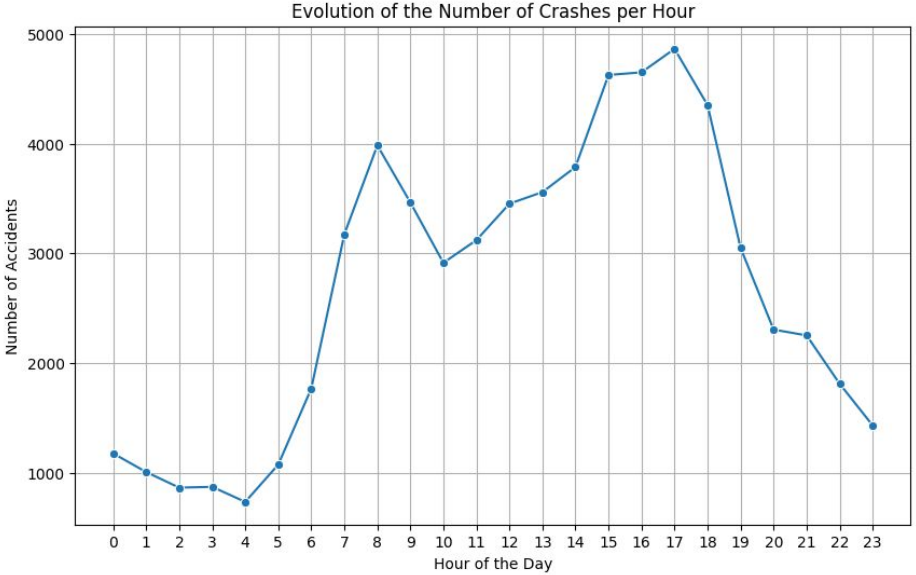


This is a spread by weekly of the number of accidents in a bar chart. Fine spread but generally more over week days.

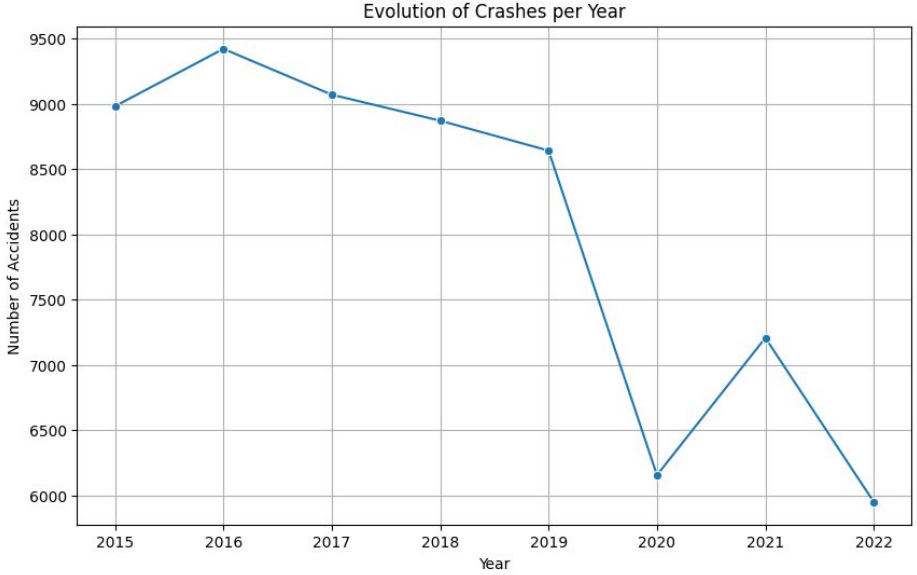


Peaks at morning busy hours and late afternoon hours

Time Series Analysis



Evolution of crashes by hourly shows there are spikes in peak activity hours as we would expect.



Evolution of crashes over the year generally has trended down shy of a spike in 2021.

Statistical Analysis

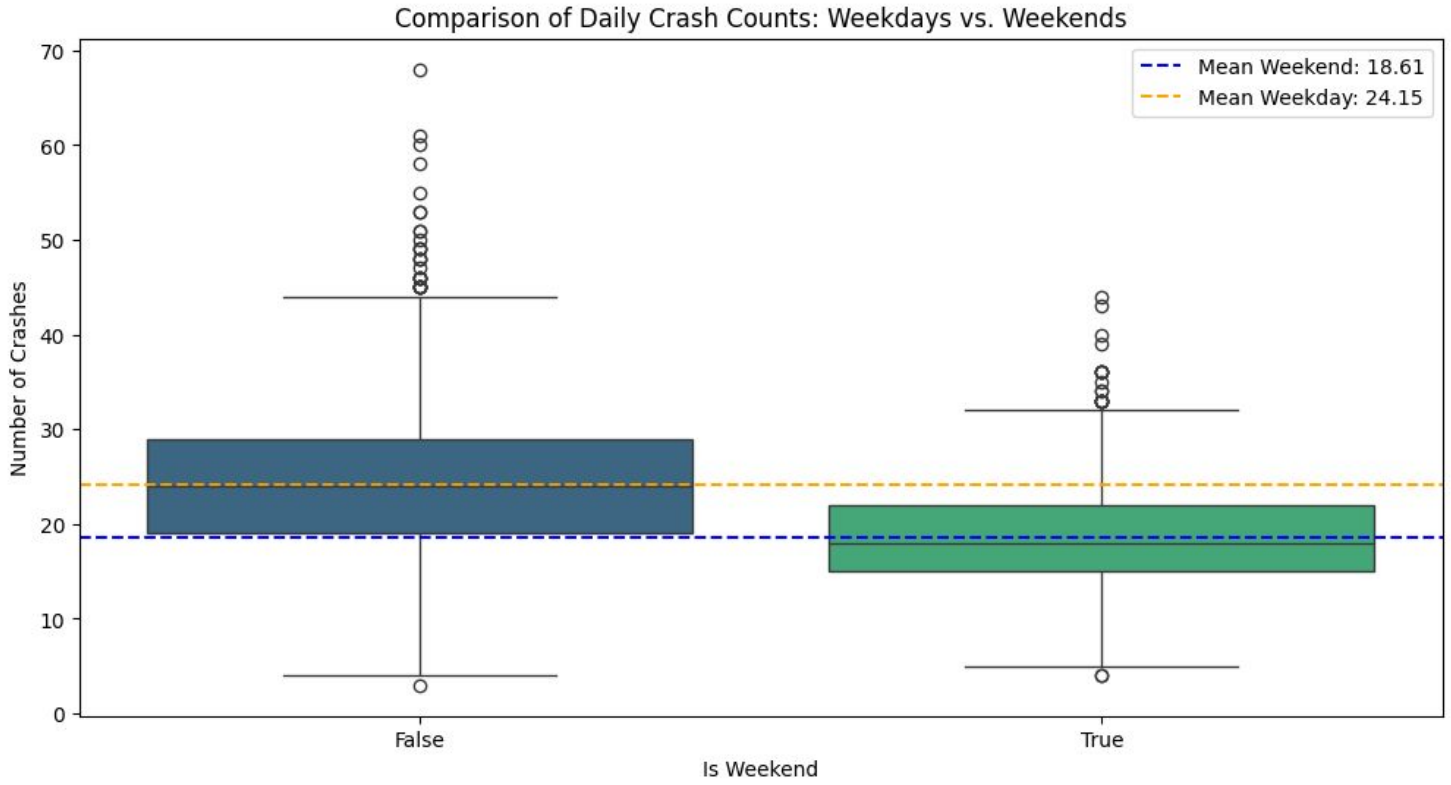
A t-test was used and we got the following:

(-17.9140755420159, 4.367038190000841e-68)

p-value < 0.05: The result is statistically significant, and we reject the null hypothesis. There is a statistically significant difference in the mean number of crashes between weekends and weekdays. Since the t-statistic is negative, it suggests that the mean number of crashes on weekends is lower than on weekdays. These results indicate that the mean number of crashes is significantly different between weekends and weekdays, with weekdays having a higher mean number of crashes

Let us try and visualise this on the next page:

Statistical Analysis (Visualized)



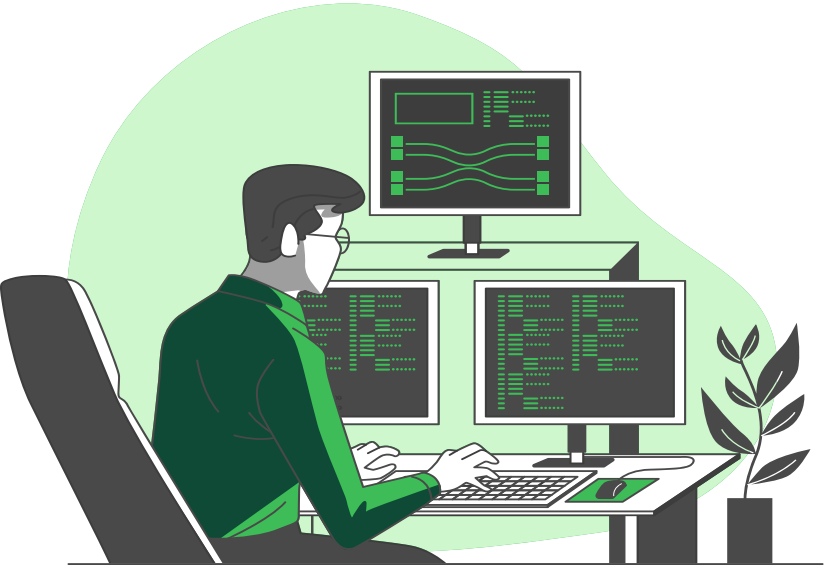
Statistical Analysis (Explained)

Box Plot: The box plot shows the distribution of daily crash counts for weekends and weekdays.

Box: Represents the interquartile range (IQR), containing the middle 50% of the data. Whiskers: Extend to the minimum and maximum values within 1.5 times the IQR. Dots: Represent outliers beyond the whiskers. Mean Overlay: The dashed lines show the mean number of crashes for weekends and weekdays.

Blue Line: Mean number of crashes on weekends. Orange Line: Mean number of crashes on weekdays. This visualization helps to see not only the difference in the means but also the spread and variability of the daily crash counts for both groups.

Summary



In summary, a 64295 by 47 cleaned data set of crash data in the USA between 2015 and 2022 was analysed and here are some key findings:

- **Frequency Analysis**: Crashes happen in more normal road, lighting and weather conditions than otherwise. Also, most if that data came from Montgomery County Police Station.
- **Time Series Analysis**: Accidents happen largely during peak activity times in the morning and late afternoons.
- **Statistical Analysis**: More accidents occur during the weekdays than the weekend.

Thanks!