

Huma Tejaswi Akkem

Dinesh Dammalapati

Project 1 CS 424

Analysis of traffic crashes in City of Chicago

Background Information

Open chicao data information is taken from chicao data portal service which consists of more than 500 information of accidents occurred in chicao that includes minor and major crashes.

Records include numerous attributes such as road condition, Weather condition, Crashed Date, Posted speed limit, device condition, road number, damage information and more excluding the personal information.

In this project, we study the dataset that contains the accidents occurred in the time period of january 1st of 2021 to january 1st of 2022. The attributes we used are posted speed limit, weather condition, lighting condition, roadway surface condition, damage, number of cars involved, most severe injury, injuries total, crashed hour, crashed day of week and crash month.

Cleaning the dataset and exploratory data analysis

According to Data Portal, the dataset has 42 attributes and we wanted to use only a few attributes in the dataset. We dropped the half of the attributes that contain string unknown, null values and empty values using df.drop. To simplify the dataframe we dropped the column and rows that we are not using. Exploratory data analysis consists of studying the main roles of a data set usually by means of visualization that contains specific interests such as time, space, space + time, distributions, and comparisons.

Answering the questions

1. How many accidents occurred in a month and in which month did the highest number of crashes occurred?

- To study the distribution of crashes per month, we used the pandas Dataset groupby method.
- Groupby method made it easier to simplify the dataset by splitting the objects.
- First we grouped by month and then calculated the highest occurrence in each month.
- **Conclusion** : Based on the dataset we used, more crashes occurred during the month of June. From November to January we see less public because of cold weather.
- Assumed reasons are, June is the month of music festivals and warm weather that attracts the public and also the time period we studied was pandemic time and since lockdown lifted in June public started coming more than other months.

2. On Which day of the week did the maximum number of accidents occur ?

- Same as months, we analyzed the distribution in the month of June according to the day of the week and used a graph as a bar plot.
- Groupby method made it easier to simplify the dataset by splitting the objects.
- **Conclusion** : Based on the dataset we used, more crashes occurred during the weekends as expected because lots of restaurants and bars open at time and also those days are non-working days.

3. On which hour of the day did we have more accidents?

- We plotted the distribution of crashes this time according to time in a day.

- We took Saturday as the highest crash day and analyzed at what time a higher number of accidents occurred .
- **Conclusion**: Based on the dataset we used, most of the crashes are occurring on a Saturday between 3pm and 4pm.

4. Which weather condition leads to more crashes?

- In our study our assumption proved wrong, we assumed most of the accidents occurs in moonlight but according to the analyzed data frame most accidents occur during the daylight and also in dry conditions.
- **Conclusion :** Accidents tend to happen more when it's dry

5. Comparison Graphs

- **Geographical plot**: In this graph we studied which part of the city had the highest number of accidents. Our Final conclusion was based on our data, West, southwest and South have more crashes.
- **Stacked graph**: We also observed the comparisons of which time of the week a higher number of accidents occurred and we concluded between 3pm to 4 pm higher crashes occurred in a week.
- **Stacked graph** : We distributed the most severe injuries into fatal, incapacitating injury, no indication of injury, nonincapacitating injury, and reported not evident. After analyzing the injuries in a week, we concluded that most of them are no indication of injury.
- **Speed limit Bar plot** : We distributed the crashes by the speed limit since the highest speed limit in records is 30 so most crashes occur at 30 speed limit.

- **Vehicles Involved bar plot :** We divided the crashes by how many vehicles involved in the accident and we concluded in an accident mostly two vehicles are involved in a higher amount.
- **Folium :** We analyzed the data in a special form using folium and also it is the best way to see the crash analysis. In this map plot, we can see the weather condition, road condition, number of accidents occurred and also location name.