CS 424 Visualization & Visual Analytics

**Exploratory data analysis**

Red Light Camera Violations

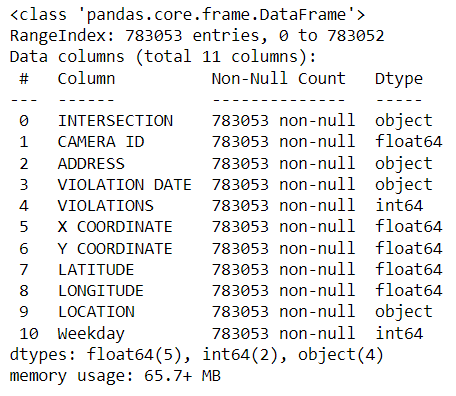
horizontal line

# Introduction

The dataset that our team chose was Red Light Camera Violations from the City of Chicago Data Portal. This dataset reflects the daily volume of violations created by the City of Chicago Red Light Program for each camera. The data reflects violations that occurred from 2015 until present day. The objective of this project is to perform data exploration using Pandas, GeoPandas and the Jupyter notebook to analyze the dataset.

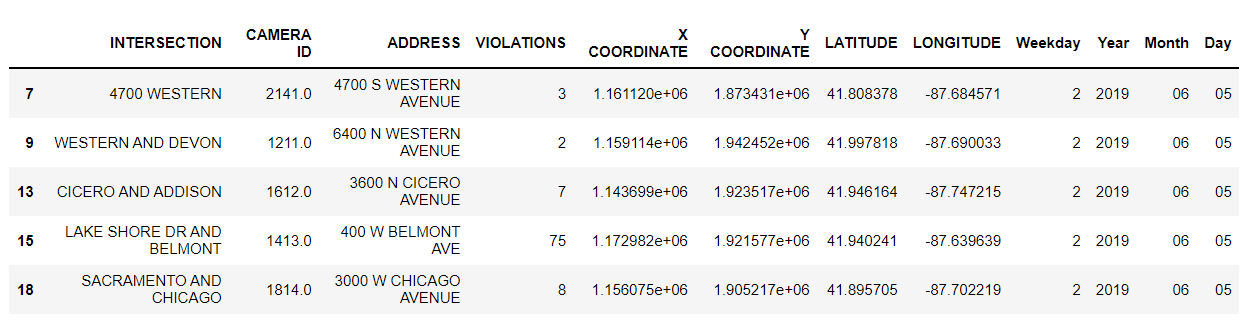
## Dataset

As mentioned the dataset we chose is the Red Light Camera Violations. This dataset originally contained 788K rows and 10 columns. For our analysis we selected data from years 2019 to 2021. The dataset was pretty clean as there were no null values nor NaN values. The dataset contained 4 categorical columns, 5 numerical float columns, and 1 numerical int column. The information of the columns was related to the position of the camera (as latitude and longitude) and the date of the violation.



There needed to be some changes in the structure of the dataset to perform our analysis, and those were the following:

We filtered data from year 2019 to 2021and the Violation Date column was split into Day, Month, Year and Day of week and then dropped in order to have different aggregation levels for our visualizations. The Location column was also dropped, as we had the Latitude and Longitude of the camera sensors, representing the same kind of information. Some violations had no coordinates, so we decided to remove the corresponding lines, reducing the entries of our database. We created a new column with the number of total violations for each camera id and a geometry column representing the positions of the cameras according to longitude and latitude values.



## Objectives

The main objectives of the project, based on the years from 2019 to 2021, were the following:

* Understand the locations of the cameras around the city and the intersections with the greatest number of violations
* Find out if there are any yearly trends for the number of violations.
* Find out if there are any monthly trends for the number of violations.
* Find out if there are any seasonal trends for the number of violations.
* Find out if there are any trends in the days of the week for the number of violations.

## Analysisc

Chart, scatter chart

Description automatically generatedAt first we wanted to see where all the cameras are present over the city of Chicago. As we can see the cameras are pretty spread out and it's denser in the city and some of the northern suburbs compared to the southern part of the city.

*Chart, scatter chart

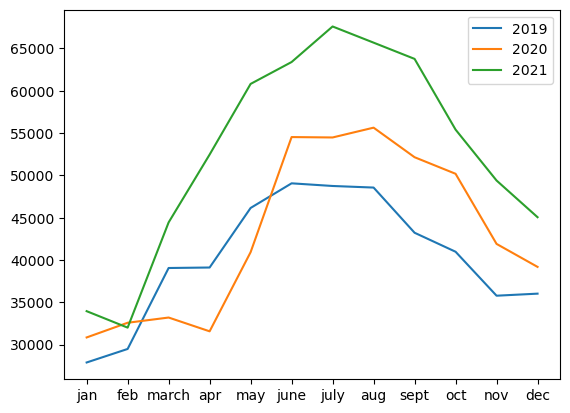
Description automatically generated*

Chart

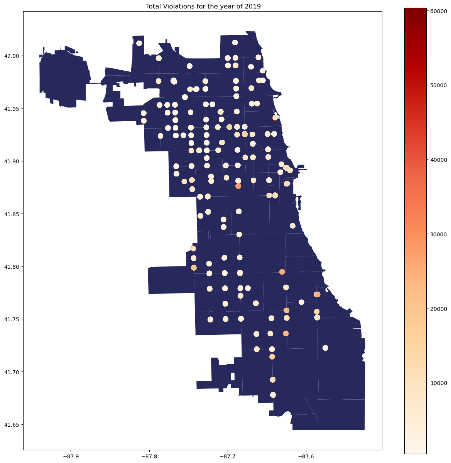
Description automatically generatedWe plotted the map for the number of total violations for the years 2019-2021 with default color, but the color combination used in the plot above, the color gradient of the legend and the clusters make it hard for some people to understand, as most people are not that sensitive to the change from violet to blue, so we decided to change the color palette for the plots moving forward.

*Analysis on Annual and Monthly Trend:*  
We created 3 new dataframes for the years 2019, 2020 and 2021 respectively, so that we can calculate the number of violations captured by each camera sensor in each year.

Here’s the results that we were able to observe through these plots when we plotted the number of violations detected by the camera for each month of the years.



As we can see from the line plot above, there is an annual increase in the number of violations going from 2019 to 2021. We can also observe that the months of June, August and July were the highest for the years of 2019, 2020, and 2021 respectively. Lowest number of violations were seen in the months of January and February across the 3 years. Also, there is an anomaly, where the number of violations didn’t increase significantly in the month of March and it decreased going to April in the year of 2020, this can be due to the COVID restrictions during that period of time.

 Chart

Description automatically generated Chart

Description automatically generated Comparing each of the years on the map of Chicago it is possible to notice that the number of violations in the north area doesn’t change as much as the south one, where we can see an increase. Considering the single sensors, some of them had a high increase of violations, as we can see in the graphs thanks the change of the colors.

*Analysis on Seasonal Trend:*  
We had to create a new column in the dataset where we added the seasons in which the violation was captured by the camera. The seasons were distributed according to the correspondent months. After grouping them into seasons, we applied summation on the number of violations, and this is the result that we have obtained.

Chart, bar chart

Description automatically generated

As we can see Summer has the highest number of red light violations, our hypothesis is that people during summer are more reckless, and there are more people driving during this period, might be because they are on holidays. Spring and Fall have similar numbers of red light violations, this may be due to the fact that the road conditions are quite similar in these two seasons and there are fewer young drivers out as most of them will have school. Winter has the lowest amount of red light violations, this may be due to the road conditions in Chicago, not great during the winter season, with heavy snow days and hence fewer people on the road.

*Analysis on the weekly trend:*Here, to represent the flow of violations during the week, we grouped the data frame on day of the week column and added the violations on each. It gives us the spread of the violations throughout the week which we observe in the following plot.

Chart, bar chart

Description automatically generated

We rescaled the plot to underline the difference between the days.

Chart, bar chart

Description automatically generated

We can see that the number of red light violations increases during the weekend, and stays relatively the same on weekdays. Saturday is the day with most red light violations. Our hypothesis is that many people go out to parties and other social gatherings on the weekends, where they might have been under the influence while getting back to their places.

## Conclusion

From the findings we were able to accomplish the objectives, using Pandas and GeoPandas, that we had set out in the beginning as we were able to see an increasing trend in the number of red light violations annually. We saw the locations of the sensors around the city and the intersections with the highest number of violations. We were also able to see the highest number of red light violations during the months of summer, and the lowest during the winter months. We were also able to see that weekends had an higher number of violations compared to that of the weekdays.

## 