Comprehensive Analysis of Traffic Crashes, Speed Camera Violations, and Average Traffic Counts

1st Pawan Kumar
dept. of Computing
National College of Ireland
Dublin, Ireland
x22186115@student.ncirl.ie

2nd Rehan Shariff
dept. of Computing
National College of Ireland
Dublin, Ireland
x22246339@student.ncirl.ie

3rd Yash Bhargava
dept. of Computing
National College of Ireland
Dublin, Ireland
x22220861@student.ncirl.ie

***Abstract*—This study provides a detailed analysis of traffic crashes, speed violations and the average traffic counts in Chicago city USA, depicting a dataset spanning the period from 2014 to 2023. The main goal is to analyse the underlying causes and contributing factors to traffic accidents in the city, thereby notifying effective policies and intervention strategies. Crashes show interesting patterns with greater incident numbers occurring during certain days and times. Findings reveal a meaningful relationship between traffic accidents and certain weather and lighting conditions which provides insights into the underlying causes of such incidents. The speed Camera Violations dataset shows the number of speed violations detected by the cameras while the third dataset i.e., Average Traffic Counts reveals a detailed description of an average number of traffic counts on an average working day and analyses the dataset and derives valuable insights.**

***Index Terms*—Traffic Crashes, Speed Camera Violations, Average Traffic Counts, Chicago, Urban Safety, python analysis, NoSQL, PostgreSQL, and MongoDB**

I. Introduction

The main motive is to find related datasets based on Chicago traffic incidents and store them in different databases such as MongoDB and PostgreSQL and retrieve store data for performing Exploratory Data Analysis(EDA) and Visualization. The first dataset which is a CSV file downloaded from the United States Data Catalog website indicates the number of incidents happening in hours and daily basis and the reason for collisions including the time of the crash, weather and lighting conditions, location of incidents, fatigue, poor road conditions, and failure to obey traffic signals also play roles in accidents etc. other related datasets indicates the speed camera violation which is JSON file which Fetched from MongoDB involves various visualization pair plots and graphs that used to further analysis the third and the last data sets tells the average traffic count in the city the workflow involves retrieving raw, semi-structured data from US Data Catalog website These data, presented in JSON format, are then stored in a NoSQL database known as MongoDB: all these datasets underwent initial pre-processing and exploration in Jupyter Notebook, where they were checked for missing values and cleaned using various techniques. Moreover, it is used for visualization by comparing numerical and categorical values and plotting different graphs and pair plots which show an overview of the accident cause and also tell about at which time the accident happened more and which time the accident happened less. At last, we found the conclusion among these data sets and how it will be used for future work or analysis. In addition, the study also evaluates the distribution of collisions across different climate conditions, representing a clear connection between adverse weather and increased accidents. Moreover, in descriptive statistics, the analysis employs visual tools such as heatmaps and bar charts to describe the trends and patterns in the data. In general, this research contributes to understanding traffic crashes in urban settings, focusing on Chicago. The findings are intended to assist policymakers, urban planners, and public safety officials in developing targeted strategies for reducing traffic accidents and enhancing road safety in the city.

II. Related Work

[1] This related study work was used in New Zealand to Study all the mortalities related to work-related traffic accidents on public roads in New Zealand from 1985 to 1998. The purpose was to systematically identify and characterize these incidents. [2] Various machine-learning models for prediction and better analysis help to reduce such types of crash incidents [2] According to a report published by the National Center for Biotechnology Information (NCBI), traffic accidents cause major health issues and become a problem for health institutions more than 1 million and 35 hundred thousand people die or are disabled in traffic crashes every year. In a 2019 survey, 93 per cent of people affected by these incidents belonged to low – and – middle-income countries and in this survey, car crashes were considered one of the main causes of death. [3] Using Some related research World Health Organization (WHO) reports that in road traffic accidents 65 per cent of deaths are unprotected road users such as pedestrians or cyclists in which rash drivers or red-light violators hit them. [5] Furthermore the work of Brown et al. (2015) delved into the application of ADT counts in real estate development. Their research showed how traffic data can inform decisions related to property development, considering accessibility and traffic flow. This aligns with the present dataset's relevance for real-estate developers aiming to make informed decisions in urban settings.

III. Methodology

In the methodology, we first explain all the software and languages used in this analysis and then we will define all the steps followed in this analysis, we have used the below languages and technologies:

- **Python** – It is an exceptionally good programming language for data science as it has various packages and libraries like Pandas and NumPy and for Machine Learning we have scikit learn as well.
- **Docker**- It is software that allows users and developers to easily create containers for applications like MongoDB, and pgAdmin4, just to create an instance of applications.
- **Jupyter Notebook** – It is a part of Anaconda by which we can easily code for Python to do data analysis which has various features for creating interactive charts and visualizations for better understanding and representation.
- **MongoDB**- It is a database in which we can store our semi-structured data like JSON file data in the form of a collection instead of a table for structured data and we can easily fetch the JSON data from MongoDB using pymongo package in Python to do further analysis on the data.
- **PostgreSQL**- It is a database that we have used in our analysis to store the CSV file in structured format i.e., tables programmatically using pyscopg2 and SQL alchemy packages and then fetch the structured table data from pgadmin4 for further analysis and visualization.

A. DATA COLLECTION

DATASET 1: Speed Camera Violations - This dataset represents the daily volume of speed violations in Chicago, the US, from 2014 to now. The reason for choosing this dataset for analysis was to suggest that to improve the traffic safety of individuals as overspeeding can lead to major accidents and to prevent that we have chosen this dataset as a part of our analysis and extracting various trends and patterns of this dataset to understand the behaviour of the drivers, to improve the safety of traffic. This dataset was downloaded in JSON format from the catalog.data.gov website.

DATASET 2: Traffic Crashes Dataset- This dataset represents the crashes reported to date in the city of Chicago which represents whether there is an injury caused or not, the weather conditions at which crashes happened, the controls are functional or not of a vehicle, and many more. This data was downloaded from catalog.data.gov in CSV format and the reason for choosing this dataset was to optime the flow of traffic, behavioural analysis of driver, updating the traffic control policies, etc. by analysing the dataset and providing insights to provide more insights for improvement in traffic control and policies.

DATASET 3: Average Traffic Counts Dataset- This dataset represents the number of vehicles that were counted on an average weekday. This dataset was downloaded from catalog.data.gov in JSON format and the reason for choosing this dataset was that it is impossible to count each vehicle, sample counts are conducted along bigger roadways to estimate traffic on half-mile or one-mile street segments. ADT counts are used for a variety of planning and operational objectives by city planners, transportation engineers, real-estate developers, marketers, and others.

B. DATA MANAGEMENT

For managing the data, we have used two databases PostgreSQL for storing the Traffic Crashes dataset as it is a CSV file, and MongoDB for managing the Speed Camera Violations dataset which was a JSON file.

- **MongoDB** is a database in which we store semi-structured data in collections. So, we have created a database named Traffic inside which I have created a collection named Speed which we have used to store our dataset (Speed Camera Violations.json) in the collection using the pymongo package from which we have MongoClient to establish a connection with MongoDB and then loaded the data in the Traffic. Speed collection in the MongoDB database and then fetch from that database and do further exploratory data analysis and visualizations.
- **PostgreSQL**- It is a database that was used to store structured data in which we have stored our 2nd dataset i.e., the Traffic Crashes dataset, and then we fetched this dataset from PostgreSQL into our Jupyter Notebook using pyscopg2 and SQL alchemy package in python and then do further analysis and visualizations.
- **Docker**- It is an environment by which we easily create an instance to deploy applications for MongoDB and PostgreSQL.

C. SYSTEM FLOW DIAGRAM

In the below figure, we have explained the system flow diagram of our analysis:

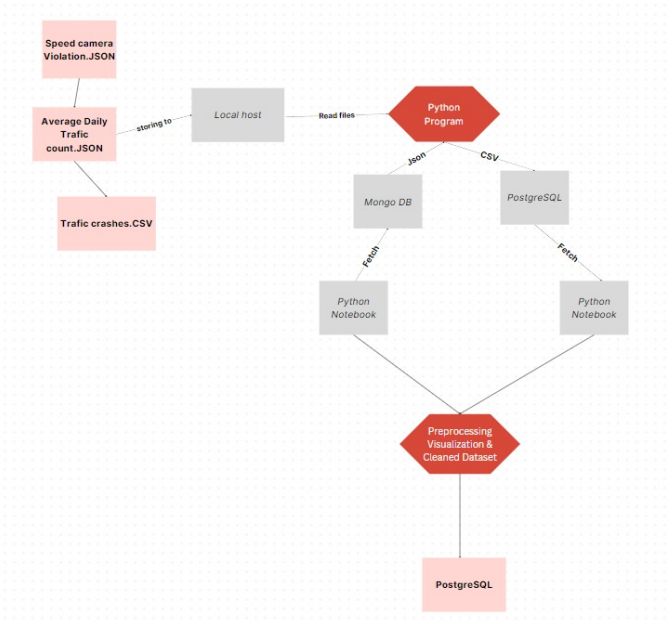

Fig. 1. System Flow Diagram

Firstly, we have stored our first dataset i.e., the Speed Camera

Violations dataset to MongoDB through Python. We also have stored the second dataset i.e., the Traffic Crashes dataset to PostgreSQL using Python. Then we fetched the first dataset did Exploratory Data Analysis and created a visualization to get better insights from the data. Then after the analysis of the first dataset, we put the second dataset into our Jupyter Notebook did exploratory data analysis (EDA) and data pre-processing, and then further transformed it into the cleaned dataset to create interactive and insightful visualization to provide better insights. Then we stored the Speed Camera Violations dataset which was cleaned into PostgreSQL as in structured format.

C. DATA PRE-PROCESSING

In this step of Data Methodology, we have followed the 4 steps:

- **Data Collection-** In this step of data pre-processing, we have extracted the two datasets that are related to the transportation domain i.e., Speed Camera Violations which represents details of violations of speed recorded by the camera was taken from catalog.data.gov website. The other dataset, which is the Traffic Crashes dataset which represents the accidents reported by the Police Department of Chicago was also taken from the same website catalog.data.gov. The third dataset i.e., Average Traffic Counts was taken from the same website which represents the average daily count census of the City of Chicago which gives us the average count of vehicles passing through a certain location on an average weekday.
- **Data Exploration-** In this step of data pre-processing, we have checked the data for any incorrect column names, missing values, outliers, data types of variables, etc. For the first dataset i.e., the Speed Camera Violations dataset we have treated the missing data by filling the continuous with mean and categorical data with the median as it is a good practice in data analysis to always use the mean as all the values were closer to the mean.
- **Data Exploration-**- In this step of data pre-processing, we have checked the data for any incorrect column names, missing values, outliers, data types of variables, etc. For the first dataset i.e., the Speed Camera Violations dataset we have treated the missing data by filling the continuous with mean and categorical data with the median as it is a good practice in data analysis to always use the mean as all the values were closer to the mean.
- **Data Cleaning or Transformation-** In this step of data pre-processing, we have treated outliers with the clip () function in which we have replaced the lower outlier value with p1 or 1st quartile value of that variable and the upper outlier value with the p99 or the 99th quartile value. We have also done some variables data types conversion like date columns and then finally treated the missing values. After doing the data cleaning we created interactive visualizations and finally stored the dataset in PostgreSQL later.
- **Data Visualization-** In this step, we have also done some visualizations like the correlation between variables and creating interactive charts while considering important key performing indicators (KPIs). Further, we have derived interactive charts for all the datasets derived valuable insights from all the datasets and found the relationships between all the key findings by the analysis of the datasets.

IV. RESULTS AND EVALUATION

So, now we have analysed all three datasets individually and created different visualizations based on the key performing indicators (KPIs) and derived valuable insights and key findings of this analysis, concluding them by showing the impact of the relationship between all the three datasets that what the number of traffic vehicles and the speed of those vehicles have on traffic collisions and incidents. So, for the first dataset:

1. **Speed Camera Violations Dataset: -** In this dataset, we have created interactive charts and visualizations using Matplotlib and Seaborn library. Firstly, we stored the JSON file of this dataset in MongoDB and then fetched it into our Jupyter Notebook as a data frame and then performed the exploratory data analysis steps like checking for missing values, and incorrect datatypes of variables, followed by checking for outliers. Further, we have cleaned this dataset with the treatment of outliers in the boxplot as we can see in Figure 2 and handling the missing values, getting the correlation between the variables i.e., shown in below figure 3 as a heatmap. Finally, on the cleaned and transformed dataset, we

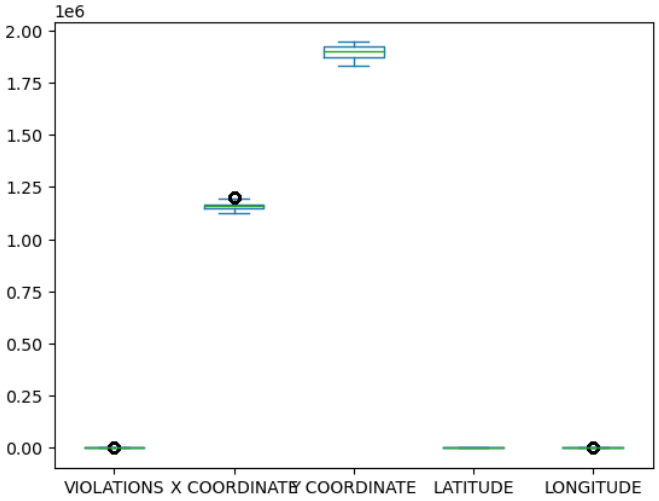

Fig. 2. Treatment of Outliers

have created visualizations like we can see month-wise analysis of the number of violations in the below Figure 3 and see that there are a greater number of violations were recorded in May and December as compared to other months. In Figure 5, we have found the top 10 cameras which recorded the highest number of violations which shows that CHI079 has recorded the highest number of

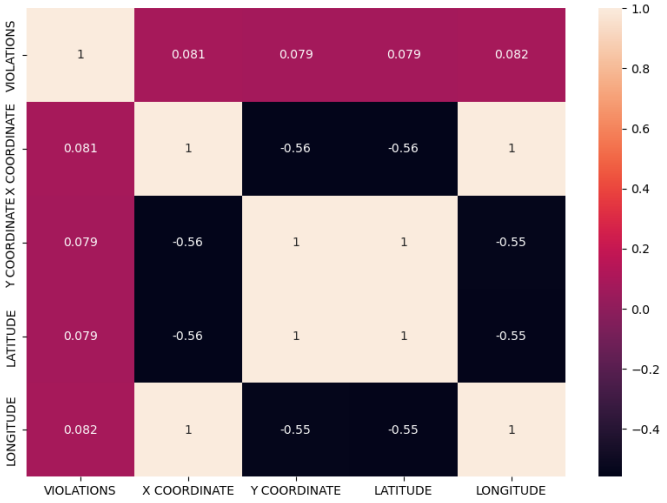

Fig. 3. Correlations for Dataset 1

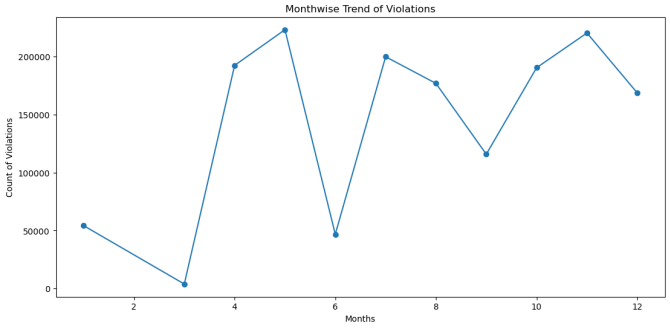

Fig. 4. Monthwise Trends of Violations

violations of speed limits because may be the reason for that is the camera was on highways where most of the vehicles are at high speeds and maybe the drivers are not in a driving state which also causes high number of accidents and crashes In Figure 6, we can see the

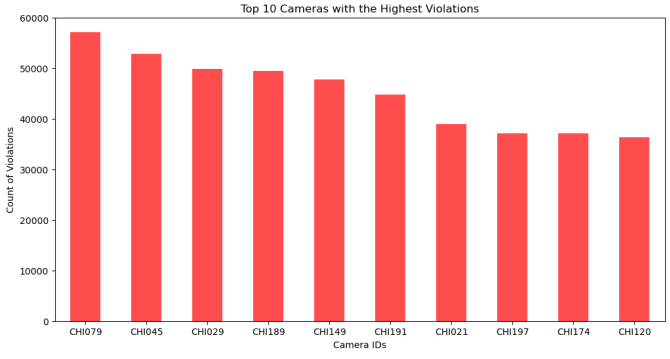

Fig. 5. Bar graph of Top 10 cameras

top 10 locations with the highest number of violations and find out that 2705 W Irving Park has the highest number of speed violations and is among the top 10 locations. Maybe the reason for this was the location was near the highway and the vehicles that were running on the highway were at higher speeds which the camera recorded speed violations while the locations where the vehicles were not recorded with higher speed the location was very crowdy that's why all the vehicles run on low speeds.

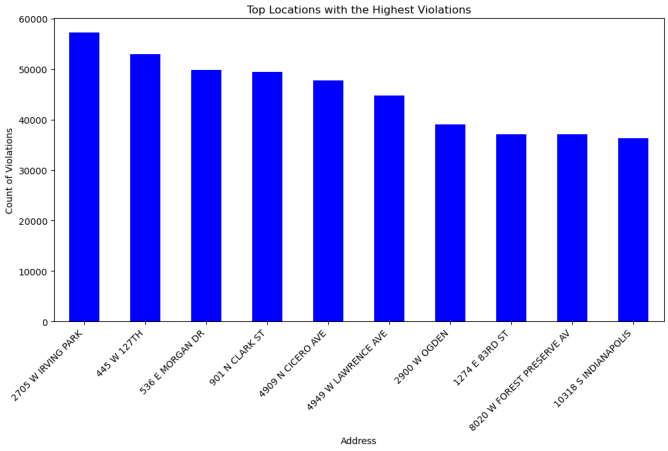

Fig. 6. Bar graph of Top 10 cameras

2. Traffic Crashes Dataset- In the second dataset, provided in a CSV format, we found a substantial volume of unstructured data consisting of 254,702 rows and 16 columns we stored this CSV file in PostgreSQL on my table. The dataset has critical variables such as Crash ID, Crash Hour, Crash Year, Weather, and Lighting Conditions. A Major challenge was the abundance of missing values, Consisting of some columns with a high percentage of missing data. To solve these issues, we implemented a detailed data-cleaning process. After the cleaning phase, Exploratory Data Analysis (EDA) was done to enable clearer visualization and further analysis of the dataset. This procedure was crucial in enhancing the dataset's suitability for knowledgeable and exact analysis. In the cleaning process, there was a high percentage of Missing Column Which we further removed using some method after that we plotted various plots such as shown in Figure 7 which indicates the relation between Crashes Per Hour in the city.

Another plot is a time series which tells the daily number of reported crashes which is shown in Figure 8. In Figure 9 it describes crashes in weather conditions.

Finally, we that found in the evening hours collision is more common compared to morning and afternoon and at the weekend crashes happen on a larger scale and more accidents happen in clear weather conditions.

3. Average Traffic Count-Average Traffic Count- The dataset provides a comprehensive snapshot of traffic activity in the city. The total passing vehicle volume and the breakdown by each direction of traffic offer a nuanced understanding of how vehicular flow varies across different

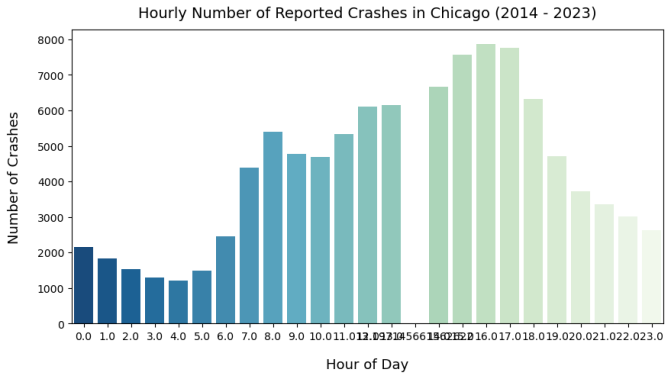

Fig. 7. Hourly Analysis of Crashes

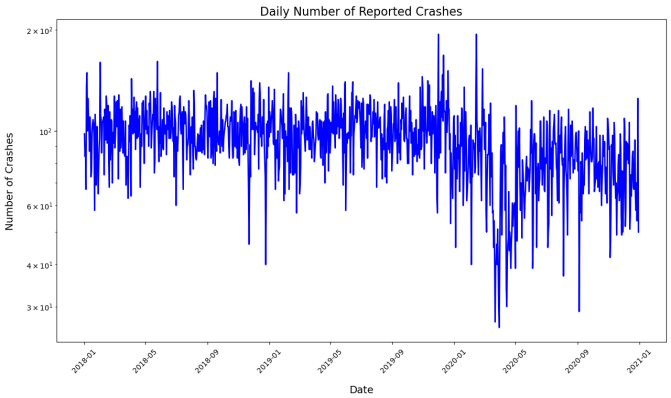

Fig. 8. Daily Number of Reported Crashes

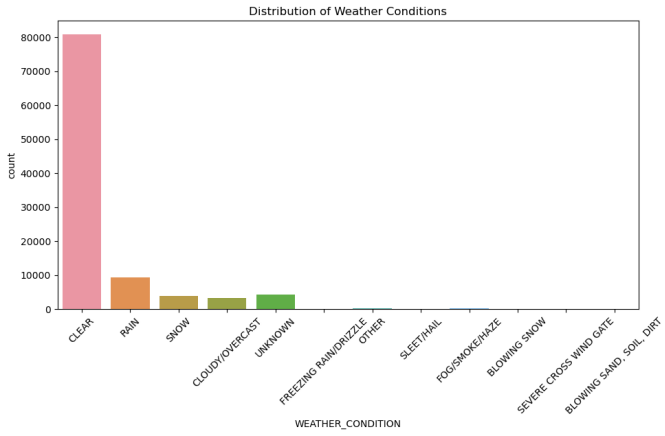

Fig. 9. Distribution of Weather Conditions

strt segments. This information is critical for identifying areas of high traffic concentration, potential connection points, and areas requiring infrastructure improvements. Additionally, the dataset's temporal dimension, captured through the 'Data of Count' column, nabls the identification of any temporal trends or variations in traffic volume. This can be particularly useful for anticipating changes in traffic patterns over the years. In our exploration of the Average Daily Traffic Counts dataset sourced from Data.gov, we conducted a thorough analysis encompassing various dimensions. A particular emphasis was placed on the column "Distribution of Total Passing Vehicle Volume" to understand its nuanced characteristics.

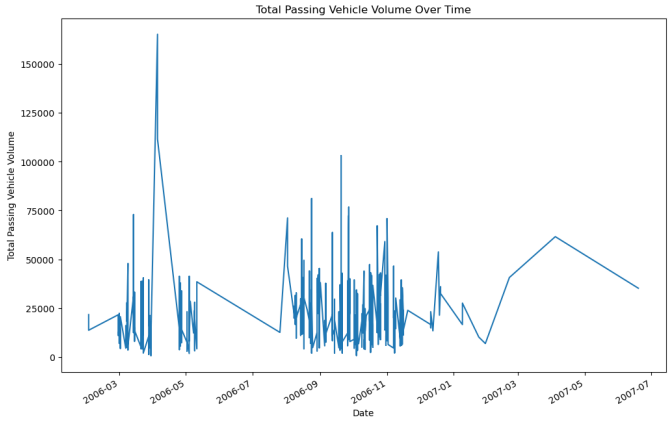

Fig. 10. Total Passing Vehicle by Time

Figure 10 illuminates the temporal trends of total passing vehicle volume over time by data, providing a dynamic visual representation of the dataset's temporal dynamics. Examining the above figure, we noticed a date, 10/04/2006, where traffic counts were usually high. To understand why, we scrutinized external factors such as local events, road closures, or public gatherings that might have influenced traffic patterns on that specific day. A particular date, 09/28/2006, caught our attention, indicating a substantial Total Passing Vehicle Volume of 40,000 on Western Ave. This anomaly prompted a dpr investigation into potential factors influencing this surge, such as local events or road conditions. An observation arises concerning the date 03/30/2006. On this day, the recorded Total Passing Vehicle Volume is notably lower, measuring 700. This discrepancy prompts us to delve into potential factors contributing to this decrease.

Furthermore, in above Figure 11, the heatmap analysis allowed us to identify areas with high correlation. For instance, a high correlation between 'Total Passing Vehicle Volume' and 'Latitude'/'Longitude' may indicate specific geographic locations with consistently high traffic. We have also analysed the total volume count by location using the variables Latitude and Longitude using the scatter plot in Figure 12 where we see that the larger

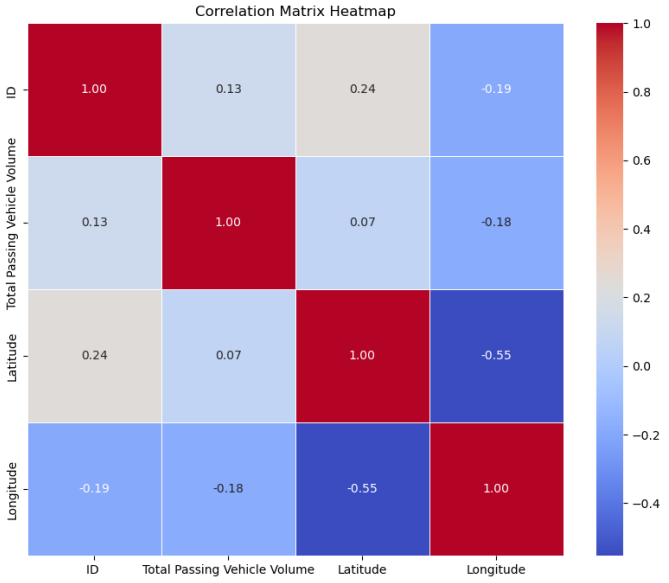

Fig. 11. Correlations of Dataset 2

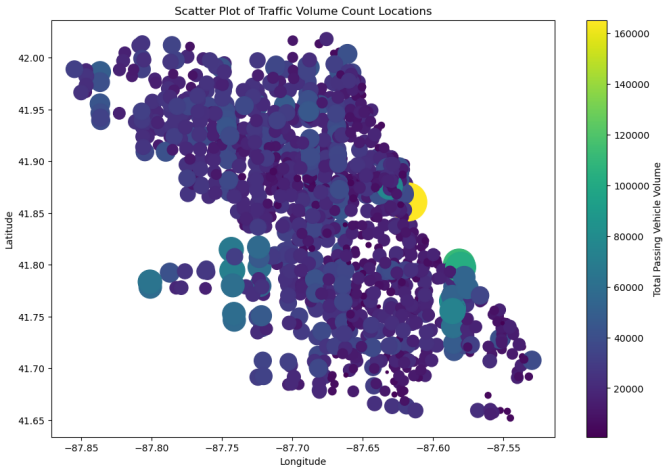

Fig. 12. Total Volume count by Locations

the size of dots more is is the volume of traffic and the shorter the size of points the lesser the volume of traffic. Acknowledging the dataset's origin from Data.gov adds a layer of credibility to our findings, reinforcing the transparency and reliability of the data source. These insights, coupled with the visualizations and correlation analyses, mpowr decision-makers across various domains with a robust foundation for informed planning and strategic initiatives.

V. Conclusion and Future Scope

So, we can say that after the analysis of all three datasets which provided us valuable insights in the domain of transportation and traffic safety on roads. This analysis of speed camera violations, daily traffic volume and road traffic crashes provides us with a comprehensive understating of road safety. One of the key findings of this research analysis was that between all three datasets. The relationship between the locations of the higher number of speed violations often increases in traffic crashes. This shows the importance of control of vehicle speed which prevents road safety hazards. Making such high security in these locations of high-speed violations like increased traffic police, increasing in road safety, making strict policies for road safety which later move towards the reducing number of traffic crashes. Keeping this in mind we can build future initiatives that can made to address these challenges of road safety and more efficient for all individuals. In future, there will be a lot to do in the analysis of the transportation domain such as application of predictive analytics by which we can predict future accidents based on the previous data which further allows us to prevent road crashes. We can also collaborate with the traffic agencies and urban planners to make some strategies that will help to overcome road safety challenges.

VI. Bibliography

References

[1] "Deaths caused by road crashes in New Zealand - Figure.NZ," Figure.NZ. https://figure.nz/chart/KDPiy6QznWSo97LJ

[2] Cappellari P, Weber BS. An analysis of the New York City traffic volume, vehicle collisions, and safety under COVID-19. J Safety Res. 2022 Dec;83:57-65. doi: 10.1016/j.jsr.2022.08.004. Epub 2022 Aug 10. PMID: 36481037; PMCID: PMC9364745.

[3] "Road traffic injuries," Dec. 13, 2023. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[4] Brown, Andrew Colville, Ian Pye, Annie. (2015). Making Sense of Sensemaking in Organization Studies. Organization Studies. 36. 10.1177/0170840614559259.

[5] https://collisionreport.nypdonline.org/