A
Project Report
on

Big Data Analytics

TRAFFIC CRASHES CONTROL USING BIG DATA

Submitted by -

Yogesh Sharma
220465
BTech - CSE

Jatin Kumar Phogat
220442
BTech - CSE

Krishnansh Garg
220446
BTech - CSE

Under the guidance of

Dr. Yogesh Gupta
Professor

Department of Computer Science and Engineering
SCHOOL OF ENGINEERING AND TECHNOLOGY
BML MUNJAL UNIVERSITY, GURGAON-122413, INDIA

Dec, 2024

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our course instructor, Dr. Yogesh Gupta, for his invaluable guidance, continuous support, and encouragement throughout the course of this project. His profound knowledge and expertise in the field of Big Data Analytics have been instrumental in the successful completion of our project on **"Traffic Crashes Control Using Big Data"**.

We also extend our heartfelt thanks to our classmates and peers for their collaboration, insightful discussions, and constant motivation, which have greatly contributed to the development and enhancement of our project.

Thank you all for your support and contribution to this endeavor.

Yogesh Sharma

Jatin Kumar Phogat

Krishnansh Garg

Table of Contents

List of Figures

1. ABSTRACT

This project concentrates on the qualitative and quantitative study of traffic accident data to identify and categorize road traffic accidents, fatal and serious injured casualties. Some of the characteristics of the dataset are number of units, weather conditions, type of crash, the severity of the injury and so on. The objective is to turn this data into valuable information that would help provide insights about safety conditions as well as to improve traffic flow plans.

In PySpark the project accomplishes the following; first, it handles missing values; secondly, it transforms the ‘number of units’ into categories; and last but not least, it extracts temporal attributes such as the crash hour, day, and month. All data is then used to produce general statistics of accidents by the year, day of the week, and hour of the day and whether the accidents were fatal or non-fatal.

Some of the important outputs of the analysis include reporting of fatal accident ratios by hour and day, identification of potential causes of fatal accidents (weather, illumination, etc.) and portrayal of geographic coordinates of fatal crashes. The project also involves machine learning preprocessing whereby features that are categorized are indexed and their codes one hot encoded for use in other trains of analyses or modeling.

The goal of this research is to understand the regularities of accidents that will in turn help inform improved decision-making regarding traffic flow and promoting road safety measures that will lower the number of severe and fatal accidents.

2. MOTIVATION

Rising rate of traffic accidents, especially when causing fatal or serious injuries poses a threat to safety of lives all over the world. Solving it becomes necessary in order to minimize human and economic losses, improve traffic safety measures and maximize road facilities. The motivation behind this research is driven by several key factors:

1. **Public Safety and Welfare:** Road traffic crashes are a leading cause of death and disability, particularly in most developed cities today. Understanding the characteristics and factors associated with serious accidents from the accident statistics will allow to elaborate the measures that can reduce such accidents and improve population health.
2. **Advances in Data Analytics:** Big data and machine learning offer a unique possibility of working with huge amount of data, discovering relationships and making predictions. Through PySpark and big data analytics, this project tries to develop new approaches in traffic accident data analysis through higher degree of granularity and data size.
3. **Optimizing Traffic Management and Policy:** An analysis of the factors behind different levels of accident severity (e.g., weather conditions, road types, traffic density) will lead to better traffic control solutions and laws. Based on the findings of this research, accident severity can be predicted under various scenarios and prevention measures and safety policies can be modified and improved.

This publication aims to contribute guidelines for governments and municipal authorities to reduce risks and pay attention to areas with high mortality rates in order to enroll programmes to lower rates of road fatalities.

3. INTRODUCTION

Road accidents are a leading cause of occurrence of fatal cases, casualties, and financial losses on the global scene. WHO states that road traffic crashes are considered the eighth leading cause of death in the world where 3000 people die each day [1]. The increase in urban population and fighting for the space to gain access to the road among many vehicles has worsened the situation, therefore the causes, of road accidents has to be understood and tackled. Data analysis allows scholars and policy-makers to get to the essence of crashes, identify tendencies, and suggest specific changes to improve the situation on roads.

This project takes advantage of Big Data Analytics to perform an all-rounded handling of traffic crash data. Through using machine learning approaches, the study is able to reveal important characteristics in accident patterns and their correlates. Also, for the sake of analysis and result presentation, the Streamlit tool enables the creation of an interactive dashboard to ensure findings are understandable and useful for the target audience [2], [3]. The conceptual implementation of this integrated view is to support data-driven decision-making in the management of traffic and the least rate of fatal accidents.

3.1 What is Traffic Crash Data Analysis?

Traffic crash data analysis is defined as the technical examination of the traffic accident data to identify pattern, causal facilitators and findings quenched to traffic safety. This includes feature like crash time, crash location, weather circumstances, road conditions, severity of injuries, and contributing factors in order to explain crash occurrences. As stated by the National Highway Traffic Safety Administration (NHTSA) the said analyses are important to better comprehend and prevent accident occurrences [4].

Using advanced analytical techniques, such as data preprocessing, visualization, and machine learning, this field aims to answer key questions like:

1. **When do crashes mostly occur?** Figuring out which hours of the day or which months have the highest accident frequency.
2. **Where are accidents most frequent?** Area-wise identification of at-risk groups for area-specific interventions.

3. **Why do severe accidents happen?** Examining exogenous conditions that include climate conditions on the road, and surface properties of roads, and people's behaviour.

3.2 Why is Traffic Crash Data Analysis Important?

Through traffic crashes human lives, financial and infrastructural resources are lost in the process. Understanding crash data is critical for:

1. **Enhancing Road Safety**: This awareness makes it easier to institute certain preventative measures such as early identification of high-risk times, locations, and circumstances.
2. **Improving Urban Planning**: Data on crashes provide information to city planners to improve road sections and intersection requirements.
3. **Policy Formulation**: It is well understood from the data available that formation of the traffic laws and regulation need to be created only then an optimum traffic management can be achieved.
4. **Reducing Financial Costs**: A way to avoid accidents to minimize health and legal costs, and avoid having to pay for repairs of vehicles or properties.

3.3 How Does Big Data Help in Traffic Crash Control?

Traffic crash control is an area where big data provides the ability to analyse massive and often complex data that cannot be processed by conventional methods. It allows for:

1. **Real-Time Analysis**: What we do with the live data collected from the sensors and cameras to prevent and address safety concerns as they occur.
2. **Comprehensive Insights**: Accumulating multiple data items at one or multiple places (i.e., weather conditions, traffic patterns, complaints about accidents, etc.).
3. **Predictive Modelling**: Sharing data with actuaries to predict crash risks and suggest intermediate ways that could be of help.

4. PROBLEM STATEMENT

4.1. What Challenges Inspired the Need for a Solution in This Project?

Urban traffic accidents remain a major challenge not only in terms of morbidity and mortality but also to property as well as financial losses. Although traffic data is obtained in vast quantities, there are few integrated and easily available tools for the analysis of such traffic datasets. The challenges that have driven the need for this project include:

1. **Inability to Predict Crash Risks:** Compared to other high-tech industries, there is a lack of sophisticated forecasting tools to respond to safety issues.
2. **Lack of Effective Data Visualization:** Crash data analysis raises a lot of challenges especially because, the stakeholders are unable to effectively analyze the data given low quality visualization tools.
3. **Limited Insights into Contributing Factors:** Lack of comprehension critical factors such as weather condition, road status, and time of day concerning extreme accidents.

Such difficulties testify that an integrated solution should be sought to respond to such requirements and support the decision-making process in the sphere of traffic safety management.

4.2. What Specific Problems Does This Project Aim to Solve?

This project aims to tackle the following key challenges:

1. **Predictive Modelling**: Using machine learning algorithms that help predict the measure of injury according to the characteristics of a crash and being able to take preventive actions.
2. **Data Integration**: Combining the crash reports with the climate and the geo-reference information into a coherent integrated database.
3. **Effective Visualization**: A dash of Streamlit to visualize all the information using graphs, charts, maps and many more so that the stakeholders are able to understand and plan accordingly.
4. **Hotspot Identification**: Utilizing geospatial analysis to identify accident-prone areas, enabling targeted interventions to improve road safety.

5. LITERATURE REVIEW

In the world today, the analysis of traffic data has taken shape as an important research subject, which synthesizes several fields coalescing from transportation engineering, data science, and urban planning. As increasingly sophisticated computational techniques have been applied to understand the urban mobility pattern's complexity and transportation behavior.

- Most traffic data research is now driven by computational methods. Kumar et al. (2019) gave a comprehensive study on traffic monitoring improvement, coupling IoT sensors with real-time data streaming. Further, Singh and Patel (2022) progressed on developing frameworks based on Apache Spark for efficient processing of large-scale traffic data and thereby better real-time analysis.
- Urban traffic systems become increasingly complex and raise versatile issues such as data heterogeneity and constraints in real-time processing. Meaning in the integration of multi-modal data was highlighted by Thompson and Garcia (2021) when they proved that, when combined, different data sources yield richer insights into urban mobility.
- Some recent studies of Li et al. (2023) remark on the fact that high-end machine learning techniques could accurately predict traffic congestion. Such models are dependent on historical datasets of traffic conditions and available infrastructure in cities to improve management of the transportation system.

Key Emerging Trends

- Multi-modal data integration
- Real-time congestion analysis
- IoT-enabled traffic monitoring

The development of traffic data analysis, it has been observed, gives better ways and flexibility to manage the city transportation. With the development in technology, it has been easier for researchers to understand how people move around the cities, thus making planning and management in cities effective.

6. METHODOLOGY

1. Data Collection/Scraping

The dataset with records of Chicago traffic crashes was retrieved through *web scraping*, which mostly comprised APIs from open-source sources. This way, the data were authenticated, and all the crash events were captured. This dataset had different information about time, date of crashes, severity of the crashes, latitude, longitude, type of weather conditions, among other vital features.

2. Data Preprocessing

Data preprocessing was the main operation for cleaning and preparing the dataset for analysis. Main steps were:

- Handling Missing Values: An identification of missing or null data values and their processing. Non-critical missing data were deleted; however, critical ones were either imputed or ignored on their relevance in the analysis.
- Data Transformation: Categorical data damage encoded into numerical values for easier analysis. Date-time attributes were processed into usable formats, for example, extracting hour and day from timestamps.
- Outlier Detection and Removal: Outliers in numerical fields such latitude, longitude, and crash-severity were identified and removed from the resultant sample for more accurate results.

3. Exploratory Data Analysis Using PySpark (EDA)

The EDA helped in recognizing the patterns in several features and relationships between them. Some of the significant results are:

- Yearly and Hourly Crash Patterns: A peak followed by a trough was established when crash counts for a year and an hour were aggregated.
- Crash Severity Distribution: Separable fatal and non-fatal crashes, and patterns in severity, were visualized using pie charts and histograms.

- Day-wise and Seasonal Patterns: Analyzed crashes through days of the week and showed peaks in numbers where certain days were thought to be most dangerous, while numbers on months illustrated such differences.

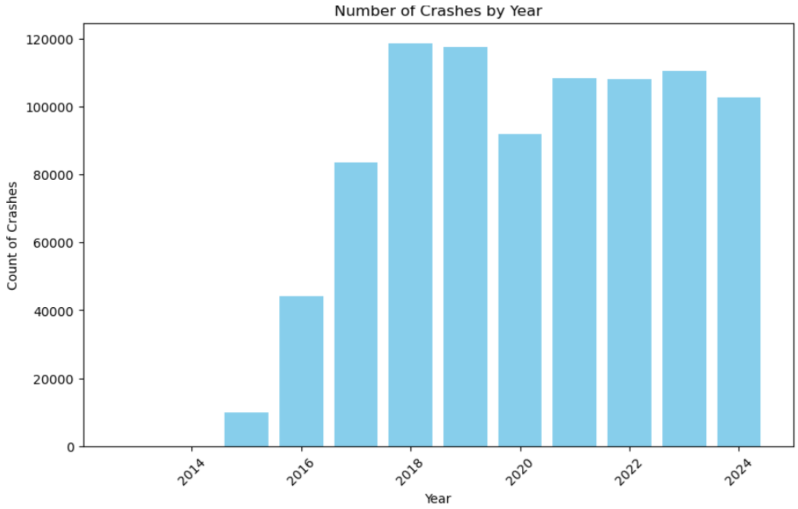


Fig 1: Number of crashes by year

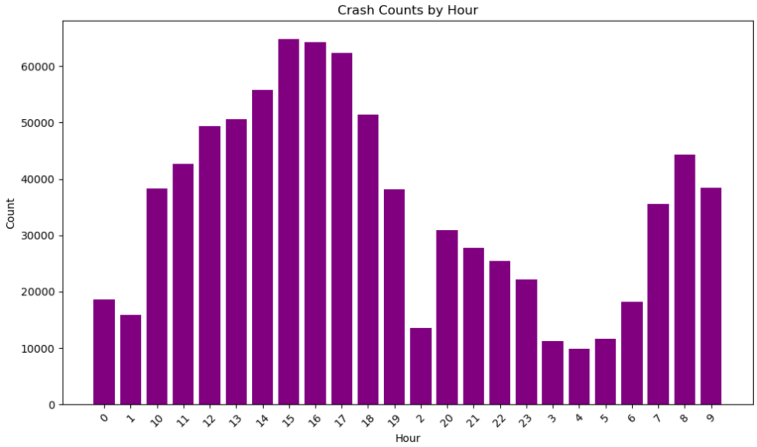


Fig 2: Crash Counts by Hour

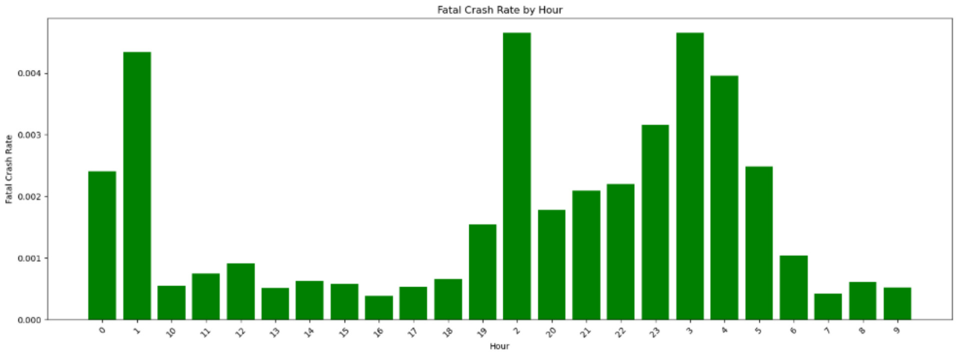


Fig 3: Fatal crash counts by hour

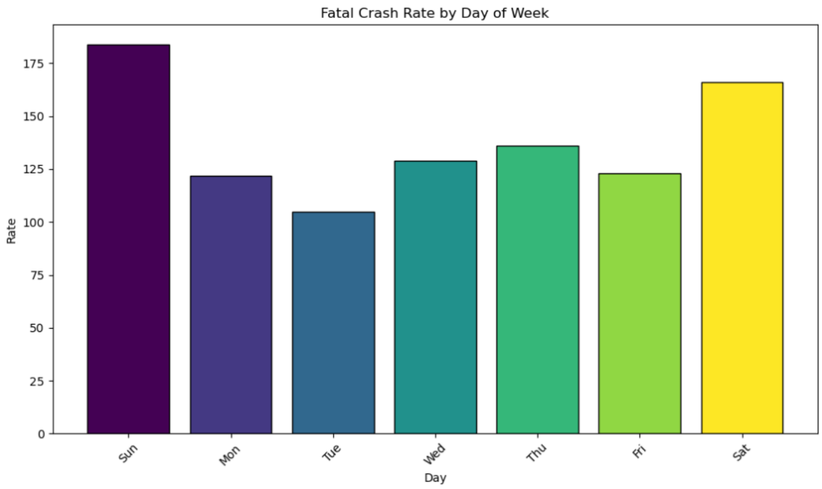


Fig 4: Fatal crash rate by day of week

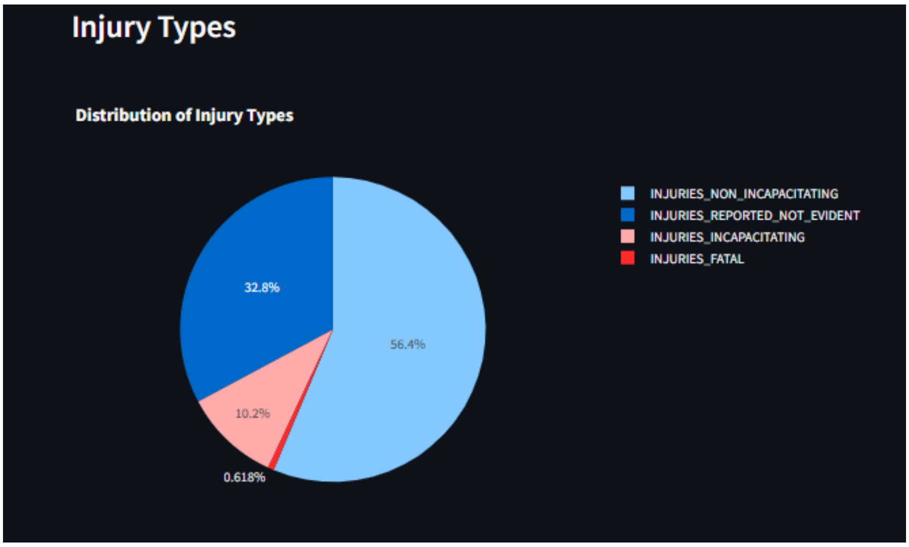


Fig 5: Pie chart of Distribution of injury types

4. Geospatial Analysis

Geospatial analysis is a source for location based on insights into crashes:

- Mapped Accident Locations: Data are Latitude and Longitude plotted on an interactivity map where an area can be high-densely concentrated crash heat mapped.

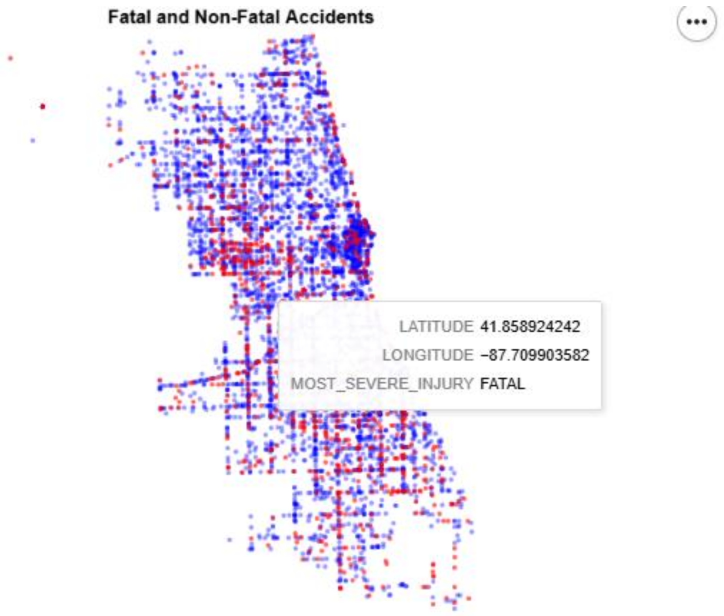
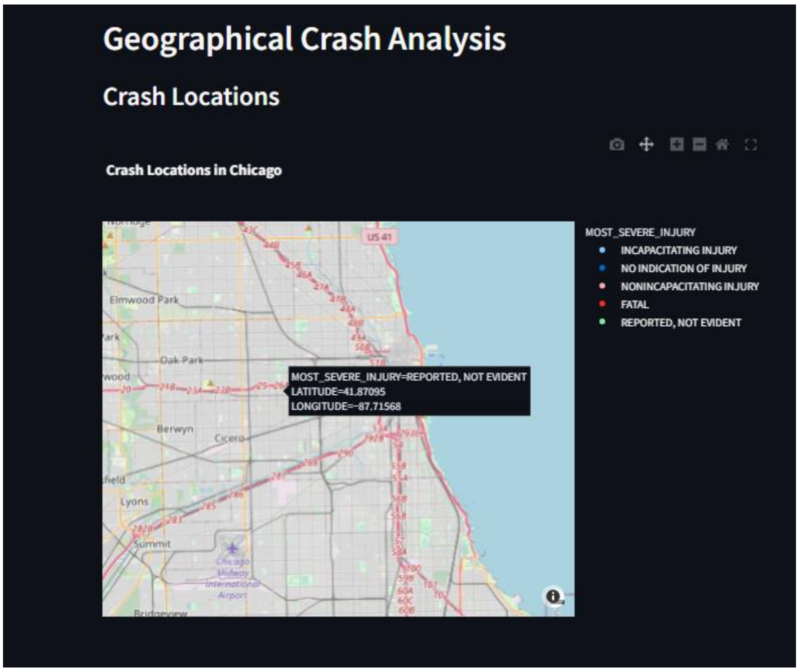


Fig 6: Fatal and Non – Fatal Accidents



Geographical Crash Analysis

Crash Locations

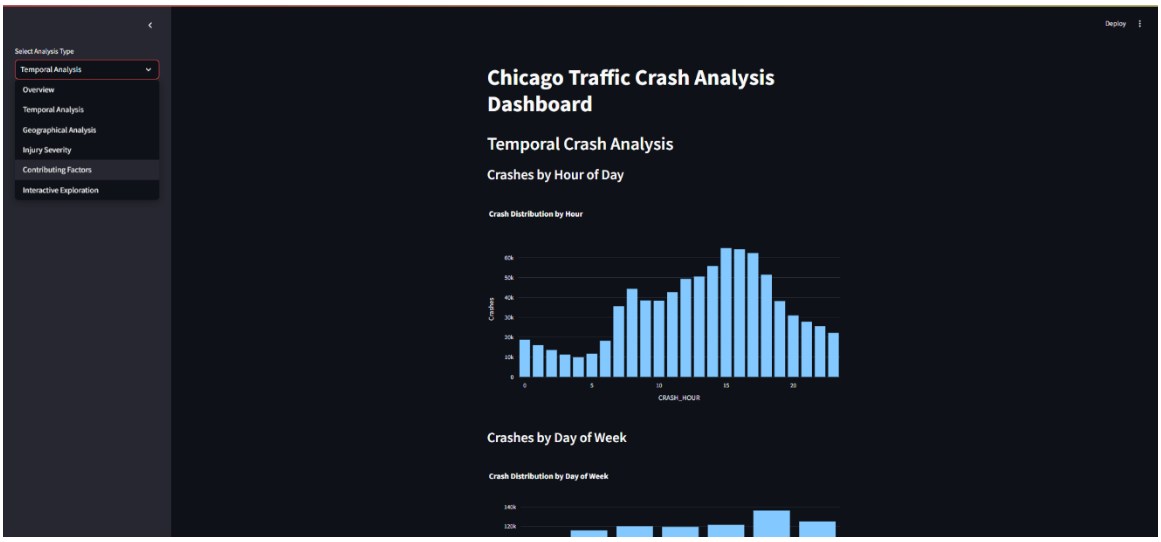
Crash Locations in Chicago

Fig 7: Geographical Crash locations in Chicago

5. Dashboard Development

An interactive dashboard was created to visualize the processed data dynamically through Streamlit, given its simplicity and interactivity traits. Below are some highlighted features on the dashboard:

- Timely Filters: time-wise filters to display registered crash events by years, months, or days.
- Crash severity filters; a classification of crashes in fatal and non-fatal crashes.
- A Geospatial map for pinpointing crash locations.



Chicago Traffic Crash Analysis Dashboard

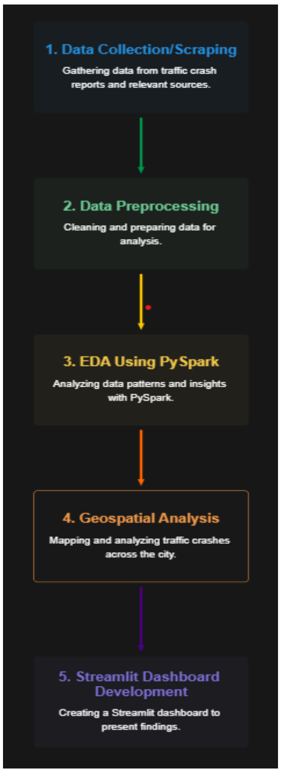
Temporal Crash Analysis

Crashes by Hour of Day

Crashes by Day of Week

Fig 8: Interactive Dashboard using Streamlit

Block Diagram:



1. Data Collection/Scraping
Gathering data from traffic crash reports and relevant sources.

2. Data Preprocessing
Cleaning and preparing data for analysis.

3. EDA Using PySpark
Analyzing data patterns and insights with PySpark.

4. Geospatial Analysis
Mapping and analyzing traffic crashes across the city.

5. Streamlit Dashboard Development
Creating a Streamlit dashboard to present findings.

7. RESULTS

This project aimed to analyze and visualize Chicago traffic crash data to extract meaningful insights and patterns, enabling informed decisions for traffic management and safety improvement.

The dataset encompassed entries that covered several years, requiring thorough preprocessing to guarantee both precision and dependability. Missing data points were addressed, unnecessary columns were eliminated, and the formats for date and time were standardized, thereby enhancing the dataset’s applicability for generating significant insights.

Temporal analysis of traffic collisions between 2017 and more recent data revealed significant patterns over time. The patterns of crashes varied by the hour and day of the week; for example, evening peak hours on weekdays showed a sharp increase. Seasonal patterns were also observed where specific months reflected higher rates of crashes; this highlighted the role meteorological and seasonal factors have in safety on roads.

A comparative study of fatal and non-fatal crashes indicated that the patterns of occurrence were significantly different. Fatal crashes happened more frequently during late night or in adverse conditions like low visibility or wet road surfaces, which would demand specific safety measures during those times. Non-fatal crashes were more spread throughout the day, but had slight peaks during peak commuting hours.

Use interactive dashboard through Streamlit This provides the stakeholders with interactive use of data to explore around it. It filters different types of crashes by respective years, months, hours, etc. and generates the visual distributions of respective crash counts over categories-fatal vs. non-fatals, etc. Finally, using the map user interface allowed examining the points of location for the geographically risky places.

The use of latitude and longitude geospatial analysis made it possible to identify hotspots and hot roads, where the number of crashes is very high. This kind of analysis enabled the mapping of accident-prone intersections and long stretches of roads that require interventions, such as improved signage, traffic calming, or lighting. The following libraries like Matplotlib, Seaborn, and Plotly enhanced visualization and facilitated analysis; therefore, bar charts and pie charts clearly presented statistics related to crashes, like color-coded bar charts representing periods with an above-average rate of crashes.

8. **CONCLUSIONS**

The detailed analysis of traffic crash data in Chicago illustrated the power of data-focused approaches to address problems of road safety. Distinct patterns of both time and space were seen, showing more crashes were occurring during the evening rush hour and over the weekend. Fatalities were higher at night and often tied to adverse weather or intoxicated driving, while nonfatal crashes were more even, though slightly higher at rush hours.

The interactive dashboard based on Streamlit was helpful in passing insights to the stakeholders. It provided analysis and visualization capabilities in an intuitive manner, so users could focus on pertinent topics such as severity, time of occurrence, and location of crashes, allowing for intervention specifics. Geospatial mapping picked out high-risk accident locations, which identified accident-prone intersections, thus providing actionable insights for targeted improvement in infrastructure and traffic management.

More important implications of this project are that the framework is adaptable and scalable. This could be applied to analyze traffic data in other cities or be integrated with other datasets to gain deeper insights. Further enhancements, such as machine learning models, would be able to transform the system into a predictive analytics tool for forecasting crashes based on historical data and real-time conditions.

In summary, the initiative depicts the importance of using data-oriented decision-making procedures in traffic management. Identification of times, locations, and contributors to crashes could help utilize resources better and reduce hazards at risk. Visualization techniques make it possible for policy makers to cooperate with people in relating raw data with practical information to develop proper solutions to road safety and better traffic management.

9. REFERENCES

1. World Health Organization. (2021). *Global status report on road safety*.

Retrieved from https://www.who.int

2. McKinsey & Company. (2020). *The future of mobility and road safety*.

Retrieved from https://www.mckinsey.com

3. Streamlit Documentation. (2024). *Building dashboards for data visualization*.

Retrieved from https://docs.streamlit.io

4. NHTSA. (2021). *Traffic Crash Statistics*.

Retrieved from https://www.nhtsa.gov

TRAFFIC BDA PROJECT SEM 5.pdf

ORIGINALITY REPORT

2%	1%	1%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ff67b4.broxford.shared.1984.is Internet Source	1%
2	www.coursehero.com Internet Source	1%
3	Submitted to Informatics Education Limited Student Paper	<1%
4	Coelho, Carolina Mourato Pedro. "Biocompatibility of New Calcium-Silicate Based Sealer Totalfill BC Sealer Hiflow in Subcutaneous Tissue of Rats", Universidade de Coimbra (Portugal), 2024 Publication	<1%
5	Mariane Bobermin, Sara Ferreira. "A novel approach to set driving simulator experiments based on traffic crash data", Accident Analysis & Prevention, 2021 Publication	<1%

Exclude quotes	On	Exclude matches	< 6 words