NYC Motor Vehicle Collisions Dashboard Project Report

Team Members:
1. Mohamed Khafagy
2. Habiba Walid
3. Menna Kurdi
4. Aya Moustafa

1. Introduction

This report presents the complete Data Engineering workflow applied to the NYC Motor Vehicle Collisions dataset. It includes data acquisition, exploration, cleaning, integration, and the development of an interactive dashboard using Flask and Plotly. The dataset was sourced from NYC Open Data to analyze traffic safety patterns across New York City.

2. Dataset Description

Two official NYC Open Data datasets were used: the Motor Vehicle Collisions – Crashes dataset and the Motor Vehicle Collisions – Persons dataset. Due to size limitations, a 50,000-row sample from each dataset was used to perform analysis and pipeline testing.

3. Exploratory Data Analysis (EDA)

EDA included evaluating missing values, crash distributions by borough, temporal crash patterns (monthly/annual trends), top contributing factors, and injury-type distributions. Several plots were created including bar charts, line charts, and summary tables.

4. Data Cleaning

Data cleaning addressed missing values, removed duplicates, standardized categorical values, converted improper data types, and detected outliers using the Interquartile Range (IQR) method. Invalid geographic coordinates were also identified and handled appropriately.

5. Data Integration

The crashes and persons datasets were integrated using COLLISION_ID with a left join, ensuring all crash records were retained. Post-merge cleaning involved removing redundant columns and resolving new missing values introduced during integration.

6. Final Dataset (df_site.csv)

A final cleaned dataset was produced to support dashboard filters. It contains details on crash dates, location, vehicle types, contributing factors, injury counts, and engineered temporal features such as crash_year, crash_month, and crash_hour.

7. Dashboard Development

An interactive Flask dashboard was developed featuring 8 dynamic visualizations: crashes per borough, monthly crash trends, contributing factors, heatmap (day × hour), crash density map, injury breakdown pie chart, vehicle type distribution, and a severity chart comparing injuries vs fatalities.

8. Deployment

The dashboard is structured for deployment via Render, using requirements.txt and a Procfile for production configuration. The dashboard is fully compatible with cloud hosting environments.

9. Conclusion

The project demonstrates a complete data engineering pipeline—from raw data to an interactive analytical dashboard. It highlights key insights about NYC traffic collisions and provides a foundation for further predictive or prescriptive analytics.