Data Science Challenge

Data Innovation Lab, AXA Deutschland

Summary Prepared By: Timileyin David Oyedeji

1 Main Objective

Use publicly available CitiBike trip and NYPD traffic accidents datasets to **generate value-added insights**, particularly to obtain **potential partnerships with an insurance company.**

2 CitiBike Data Analysis

The summarized analysis of the CitiBike 2023 data is given here, highlighting the data cleaning, preprocessing, as well as exploratory data analysis.

Data Cleaning, Preprocessing
The following steps were performed before the data was analyzed:

1. The initial 35 million-row dataset underwent cleaning, where missing values, invalid geographic coordinates, and outliers in ride duration, distance, and speed were removed. The refined dataset comprised 34.79 million rows (≈0.9% loss).
2. Ride duration, speed, and distance columns were engineered.
3. Time-based features, including day, hour, month, and season were extracted using pandas .dt accessor.

Exploratory Data Analysis
Several analyses were carried out based on the cleaned and processed data including

1. Analysis of user types: members vs. casuals (trip share, distance, duration)
2. Breakdown of bike types (classic vs. electric) and preference by user type
3. Trip volumes by day of week, hour of the day, month and season
4. Investigation of ride duration and distance patterns
5. Geospatial analysis

3 NYPD Data Analysis

The summarized analysis of the NYPD accidents data (up to 18th June 2025) including data cleaning, data prepocessing (extracting bike related accidents), feature engineering and exploratory data analysis.

Data Cleaning, Preprocessing
The following steps were performed before the data was analyzed:

1. The initial 2.18 million-row dataset was cleaned (missing values, invalid geographic coordinates removed), yielding 1.94 million rows.
2. Bike-related accidents were filtered by identifying cyclist injuries/fatalities or vehicle codes for bicycles, bikes, scooters, and e-bikes.
3. Time-based features, including day, hour, month, and season were obtained using pandas .dt accessor.

Exploratory Data Analysis
Several analyses were carried out based on the cleaned and processed data:

1. Accident volumes by day of week, hour of the day, month and season
2. Accident severity (injuries and fatalities)
3. Borough-level accident distribution and top contributing factors

4 Combined CitiBike and NYPD Data Analysis

The combined cleaned 2023 citibike data and NYPD bike accidents data were analysed for identifying high-risk area detection, time-based risk pattern analysis, and route risk scoring analysis.

High-Risk Station Detection:

1. Spatial matching to rank stations by accident proximity
2. Mapping of top 20 risky stations using Folium

Time-Based Risk Patterns:

1. Comparison of CitiBike usage volume vs. accident frequency
2. Identification of overlapping peak-risk hours, days, and months

Route Risk Scoring:

1. Scoring of top routes based on proximity to past accidents
2. Ranking of routes to inform future safety measures

5 Machine Learning Based Analysis

Task: create a binary risk label (risky vs. safe) using accident proximity.

1. Build Logistic Regression and Random Forest models using features (ride hour, ride day of the week, ride duration, ride distance, borough, membership, type of bike) and target (whether a trip is ”risky”)
2. Train model, and evaluate performance of model using recall, precison and f1 score