Jericho J. Layos

Project Overview

The notebook conducts an Exploratory Data Analysis (EDA) on a synthetic dataset representing **50 drivers** over a **90-day period** (January 1, 2025, to March 31, 2025). The dataset contains **4,500 records** intended to verify the logic behind driver performance scoring before moving to database setup.

Data Quality and Preparation

- **Structure:** The data includes metrics for delays_minutes, behavioral_problems, violations_count, accidents_count, and the resulting rating.
- **Data Integrity:** A quality check confirmed the dataset is complete with **no missing (NaN) values** across any columns.
- **Preprocessing:** The date column was successfully converted from an object to a datetime data type to facilitate time-series analysis.

Statistical Profile

Descriptive statistics provided a baseline for driver performance:

- **Ratings:** The average driver rating is **3.91**, with a range spanning from a minimum of **1.0** to a maximum of **4.8**.
- **Delays:** On average, drivers faced **15.3 minutes** of delay, with the worst delay reaching **54 minutes**.
- **Incidents:** Accidents and behavioral problems are relatively rare, with mean values near zero (0.01 and 0.05 respectively), indicating they are exception events rather than the norm.

Correlation Analysis (Heatmap Findings)

A correlation matrix was visualized to determine which negative behaviors most heavily penalize a driver's rating:

- **Punctuality is Critical:** The strongest negative correlation exists between **delays and rating (-0.58)**, indicating that lateness is the heaviest penalty in the scoring logic.
- **Accident Severity:** Accidents are the second most impactful factor, showing a strong negative correlation of **-0.49**.
- **Minor Infractions:** Violations (**-0.21**) and behavioral problems (**-0.19**) hurt the rating, but significantly less than delays or accidents.
- **Behavioral Independence:** Interestingly, there is almost no correlation between accidents and delays (**0.06**). This implies that drivers who are late are not necessarily more reckless or prone to accidents; these are distinct behavioral issues.

Conclusion

The analysis concludes that the synthetic data accurately reflects the intended scoring logic, where punctuality and safety are the primary drivers of performance scores.