1. Purpose of Simulation

I created synthetic telematics data to prototype the pipeline without depending on sensitive real-world data. The simulation mimics realistic driving patterns, including location, speed, acceleration, time of day, and road types.

This dataset serves as a foundation for building the trip-level and driver-level aggregations used in risk modeling and premium pricing.

---

2. Simulation Setup

- **Drivers & Trips**
    - Number of drivers: 30 (configurable).
    - Trips per driver: 10.
    - This balance provides enough variation to test aggregation logic while keeping the dataset manageable.
- **Trip Duration**
    - Randomized between 1 minute and 1 hour.
    - Matches realistic variability (short urban trips vs. longer highway journeys).
- **Sampling Interval**
    - 5 seconds between telemetry points.
    - Enough granularity to capture changes in acceleration and speed without inflating dataset size.

---

3. Columns (Raw Telemetry Level)

- **timestamp** → allows temporal analysis and detecting day/night driving.
- **trip_id, driver_id** → relational identifiers for aggregation later.

- **lat, lon** → Midwest bounding box, mimics GPS tracks.
- **speed** → core safety feature; varies by road type.
- **acceleration** → enables harsh braking/acceleration detection.
- **road_type** → city, residential, or highway; affects speed limits and event probabilities.
- **engine_on** → constant in this sim, but placeholder for future engine state tracking.

Reasoning:
I kept only features that:

1. Are realistic outputs of telematics hardware.
2. Tie directly to risk scoring (speeding, harsh events, road environment).
3. Allow richer aggregations later without bloating the dataset.

---

4. Number of Rows

- Each trip duration (1–60 minutes) × sampling interval (5s) → ~12–720 rows per trip.
- With 30 drivers × 10 trips = 300 trips total, dataset ≈ 100k rows.
- This scale ensures:
    - Enough variation for aggregation and ML training.
    - Still lightweight for local development.

---

5. Data Quality Measures

- **Speed realism:** Capped to legal/typical ranges by road type.
- **Harsh events:** Probabilities tied to environment (e.g., more harsh braking in city, more harsh acceleration on highways).

- **Consistency:**
    - Ensured average speed never exceeds max speed.
    - Time increments always uniform (5s).
    - No negative speeds or accelerations outside harsh event windows.
- **Geospatial realism:** Latitude/longitude restricted to Midwest bounding box, keeps location data internally consistent.
- **Temporal realism:** Trips randomized across recent 10-day window, with ~30% probability of being night trips.

---

6. Why This Approach

- Balances **control** (probabilistic harsh events, road type fractions) with **randomness** (driver behavior, trip timing) to simulate both structured and unpredictable elements of real driving.
- Produces data rich enough to extract higher-level features (trip summaries, driver history), which is the real goal of this project.
- Keeps dataset interpretable, so when visualizing or debugging, I can easily validate whether results are “reasonable.”