

Project: Roadway Crash Risk Explorer

Project Overview

Crash frequency on roads is influenced by a variety of geometric and environmental factors — from **lane width** and **curve radius** to **lighting** and **weather conditions**. In this project, you'll explore a synthetic dataset of roadway segments to identify which features are most associated with **crash frequency**. Your goal is to uncover which types of roads are most at risk and propose design strategies or policy interventions to reduce crash occurrences.

Objective

You'll work with a **synthetic dataset** of roadway segments, each with geometric and environmental descriptors and an associated number of crashes per year. Your task is to analyze what combinations of roadway features are associated with higher crash frequencies and to make design recommendations based on your insights.

Major Questions for Your Proposal

1. What Makes a Road Risky?

Which roadway features (e.g., narrow lanes, high speeds, poor lighting) do you think are most strongly associated with crashes? Why?

2. Exploration Strategy

- Will you group by speed, lane width, or curve type? Compare lighting conditions? What relationships are you planning to examine (e.g., speed + weather, curve radius + shoulder type, etc.)?
- Will you use statistics or visualizations to explore the data?
- Will you create a regression or classification model to predict crashes based on input variables?
- How will you handle missing values?

3. Visualization Plan

- What visuals will you use to explore the data? Will you look at correlations between input variables?
- How will you show your findings? Bar charts, scatter plots, grouped box plots?

Project Definitions

Variables

- **road_segment_id**: Unique ID for each road segment
- **avg_daily_traffic**: Average daily vehicle volume (vehicles/day)
- **posted_speed_kph**: Posted speed limit (km/h)
- **lane_width_m**: Width of each lane (m)
- **shoulder_type**: Paved, Unpaved, None

- **curve_radius_m**: Radius of curvature for road segment (m; straight roads will have high values or **None**)
- **lighting**: Good, Poor, None
- **weather_condition**: Dry, Wet, Snowy, Foggy
- **crashes_per_year**: Number of reported crashes per year on the segment

Data Source

You'll be provided with a **synthetic CSV dataset** containing approximately 400 roadway segments. The dataset includes realistic trends and relationships (e.g., higher crashes for sharp curves, poor lighting, wet conditions).

Project Requirements

1. Data Cleaning & Grouping

- Handle missing values (e.g., **None** for curve radius on straight segments)
- Consider grouping speeds, lane widths, or crash frequency levels
- Consider creating subsets of the data based on conditions (e.g., speed + weather)

2. Exploratory Analysis

- Identify roadway features that are associated with high crash rates
- Explore interactions between conditions (e.g., high speed **and** poor lighting)
- Investigate whether traffic volume or design features are stronger predictors

3. Visualization

- Use scatter plots, grouped bar charts, or box plots to illustrate trends
- Consider visualizing how combinations of variables (e.g., speed + weather) affect crash counts

4. Discussion

- Which features had the most consistent effect on crashes?
- Did anything surprise you?
- What design or policy changes might reduce crash frequency?