SC3021 Data Science Fundamentals

Singapore Road Safety Analysis

Lab 01: Problem Formulation, Requirement Analysis, and Source Identification

Presented By: **BALODI SHALOK U2423095K**
MANISH VISHIN KUMAR CHELLANI

Background and Necessity

Road traffic accidents continue to impose social, economic, and healthcare costs in Singapore.

Despite strong road infrastructure and enforcement, accidents resulting in fatalities and serious injuries still occur.

Even when datasets are aggregated (monthly/annual), they allow:

- building severity proxies (e.g., casualty counts),
- testing relationships with rainfall,
- controlling for vehicle population and vehicle types involved.

Problem Formulation

How do weather conditions, vehicle composition, and temporal patterns influence traffic accident severity in Singapore, and what data-driven insights can be used to inform targeted road safety and prevention strategies?

Underlying Problem

While overall accident numbers are monitored, there is **limited integrated analysis** on what combinations of environmental conditions, vehicle composition, and temporal patterns are associated with higher accident severity.

Existing data is fragmented across multiple public datasets and largely aggregated, making it difficult to derive actionable insights that can inform targeted road safety interventions and resource allocation.

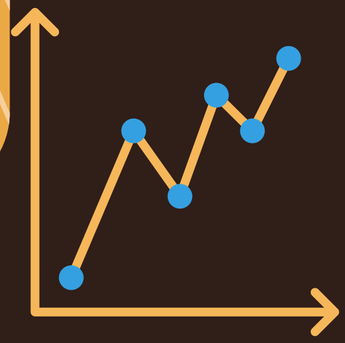
THE PROBLEM:RELEVANCE AND GOALS

This problem is relevant to olice) / public safety policy. Other major stakholders include

- Land Transport Authority (LTA) / transport planning
- Emergency response agencies / demand planning

The primary goal is to identify measurable factors associated with higher traffic accident severity.

A secondary goal is to evaluate whether regression-based models are feasible given the available data



Requirement Analysis

All datasets must be publicly available, reproducible, and accessible as CSV files.

Datasets must include temporal information to enable alignment and trend analysis.

Weather data (rainfall) is required to test hypotheses relating adverse conditions to accident severity. Vehicle-related datasets are required to capture exposure and vehicle mix effects.

Severity is represented using injury severity categories and monthly casualty counts, enabling both continuous and binary severity measures.

The analysis requires a representation of accident severity which is defined as seriousness of outcomes (fatal / injury categories) using casualty counts or injury severity categories.

Exposure variables like vehicle population and composition of vehicles involved are required to normalize severity outcomes and avoid misleading trends.

Data Requirements

- Severity outcome data is required to define continuous and binary severity measures.
- Weather data is needed to assess the impact of adverse conditions on accident severity.

To address the research question, the selected datasets should collectively provide:
Outcome (severity)

- accident severity categories OR casualty counts (monthly/annual)

Explanatory features

- weather: rainfall (monthly)
- vehicle mix: vehicles involved by type (annual) + vehicle population (monthly) to normalize exposure
-

IAll datasets must include a time dimension (monthly or annual) to allow trend analysis and integration.

Dataset 1: Causes of Accidents by Severity

- **Contains** accident causes classified by injury severity

- Features
  1. cause,
  2. severity type,
  3. time period

	year	accident_classification	road_user_group	causes_of_accident	number_of_accidents
0	2012	FATAL	Drivers, Riders or Cyclists	Failing to Keep a Proper Lookout	59
1	2012	FATAL	Drivers, Riders or Cyclists	Failing to Have Proper Control	50
2	2012	FATAL	Drivers, Riders or Cyclists	Failing to Give Way to Traffic with Right of Way	9
3	2012	FATAL	Drivers, Riders or Cyclists	Changing Lane without Due Care	6
4	2012	FATAL	Drivers, Riders or Cyclists	Disobeying Traffic Light Signals Resulting in ...	9

- Advantages: explicit severity information
- Limitations: aggregated data, possible multi-cause counting

Dataset 2: Monthly Accident Casualties

- Description: monthly counts of traffic accident casualties
- Features in the dataset
  1. month
  2. year
  3. casualty count
  4.

	DataSeries	2025Nov	2025Oct	2025Sep	2025Aug	2025Jul	2025Jun	2025May	2025Apr	2025Mar	2025Feb	2025Jan	2024
0	Total Casualties Fatalities	13	13	12	10	11	17	18	11	12	9	12	
1	Pedestrians	4	1	3	2	2	4	3	3	4	1	4	
2	Personal Mobility Device Users	0	0	1	0	0	0	0	0	0	0	0	
3	Cyclists & Pillions	4	0	1	1	0	2	1	0	0	3	2	
4	Motor Cyclists & Pillion Riders	5	10	5	4	8	10	11	7	5	4	6	

5 rows × 204 columns

- Advantages: suitable severity proxy with temporal resolution
- Limitations: not event-level, requires normalization

Dataset 3: Vehicles Involved by Type (Annual)

- Contains vehicles involved in fatal and injury accidents by type

- Features In the dataset
  1. vehicle category
  2. year,
  3. counts

	DataSeries	2023	2022	2021	2020	2019	2018	2017	2016	2015
0	Total	13507	12346	10964	9852	14133	14062	14168	15369	14982
1	Bicycles And Power Assisted Bicycles	598	745	812	581	473	513	605	633	643
2	Motor Cycles & Scooters	4157	4102	3636	3364	4860	4748	4619	4913	4694
3	Motor Cars & Station Wagons	6409	5420	4716	4374	6643	6423	6680	7172	6930
4	Goods Vans & Pick-Ups	689	581	577	499	552	549	539	657	617

- Advantages: captures vehicle mix risk differences

- Limitations: annual granularity

Dataset 4: Monthly Rainfall Total

- Contains monthly rainfall measurements
- Features in the dataset
  1. month
  2. rainfall
  3. amount

	month	total_rainfall
0	1982-01	107.1
1	1982-02	27.8
2	1982-03	160.8
3	1982-04	157.0
4	1982-05	102.2

- Advantages: high-quality meteorological data
- Limitations: rainfall alone may not capture all adverse conditions

Dataset 5: Vehicle Population (Monthly)

- Contains registered motor vehicle population by type

- Features:
  1. vehicle type
  2. Month
  3. population count

	month	vehicle_type	number
0	2012-01	Cars	593555
1	2012-01	Rental Cars	13970
2	2012-01	Taxi	27059
3	2012-01	Buses	17037
4	2012-01	Goods & Other Vehicles	159854

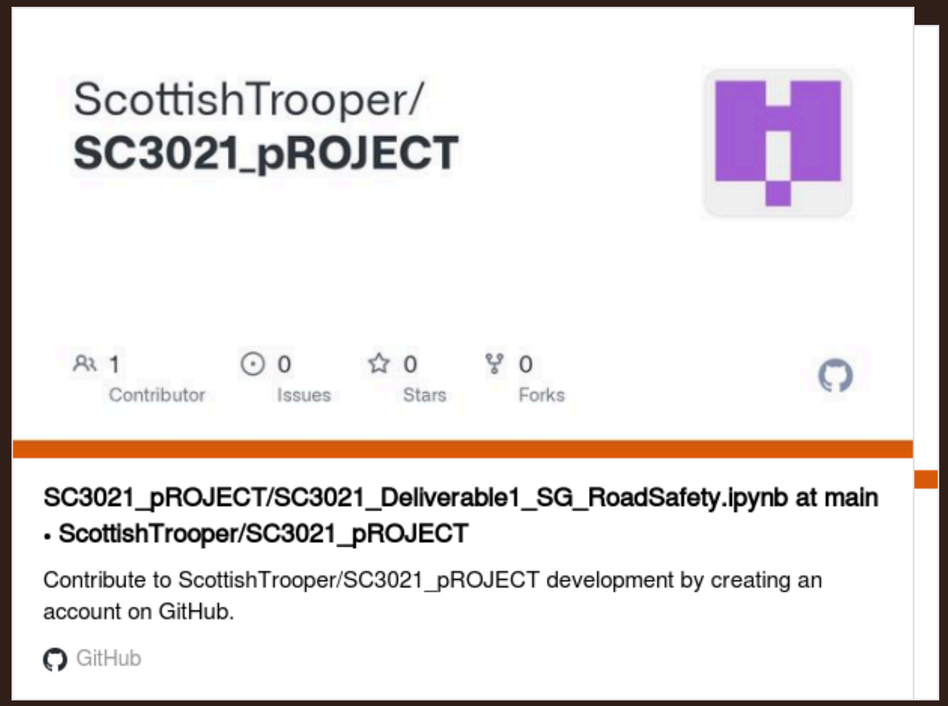
- Advantages: exposure normalization variable

- Limitations: population does not reflect usage intensity

Next Steps

- Clean and align datasets temporally
- Perform detailed data cleaning and exploratory data analysis (EDA)

- Apply linear and logistic regression models
- Interpret results for policy and safety insights



https://github.com/ScottishTrooper/SC3021_pROJECT/blob/main/SC3021_Deliverable1_SG_RoadSafety.ipynb