FINAL PROJECT REPORT

Road Accident Data Analytics & Infrastructure Stability

Domain — Transport and Infrastructure

INSTITUTE

Newton School of Technology

DOMAIN

Transport and Infrastructure

DATASET

UK Road Accident Dataset | Kaggle (10,000 records)

TOOL

Google Sheets (Pivot Tables, Interactive Filters)

Team Members

Aadit — 2401010003
Krish Garg — 2401020097
Manya Verma — 2401010265
Abuzar Haideri — 2401010024
Tathagat Harsh — 2401010477
Tanish Rao — 2401010474

Table of Contents

SECTION 1

Executive Summary

Between 2021 and 2022, the transport sector recorded **13,522 casualties** within the scope of this study. These incidents were not random occurrences; they were structurally concentrated across specific road types, speed environments, and seasonal windows — pointing to systemic vulnerabilities that demand data-driven policy intervention.

13,522	**73%**	**1.16×**	**238**
Total Casualties	Single Carriageway Share	Rural Risk Multiplier	Fatal Incidents
2021–2022	*9,888 individuals*	*vs. Urban zones*	*1.76% of total*

Problem

Road casualty data reveals entrenched structural risks on UK roads between 2021 and 2022. Single carriageways, rural environments, and Q4 seasonal conditions emerged as the dominant risk multipliers. Despite 66.7% of accidents occurring on dry surfaces, casualty rates remained high — indicating that driver behavior and speed are primary contributors rather than weather alone.

Approach

A dataset of 10,000 records (23 columns) was sourced from the UK Road Accident Dataset on Kaggle. All data cleaning, transformation, and analysis were executed in Google Sheets using Pivot Tables, text formulas, and interactive slicers. A structured KPI framework was developed to quantify magnitude, severity, and risk intensity.

Key Insights

- 73% of all casualties occurred on Single Carriageways (9,888 individuals).
- Rural areas are 1.16× more dangerous per incident than urban zones.
- October (1,265) and November (1,254) represent peak casualty months, driven by reduced daylight.
- Fatal incidents (238) and Serious incidents (1,834) together represent ~15% of total casualties.
- Despite expectations, dry road surfaces accounted for 66.7% of accidents — highlighting behavioral risk.

Key Recommendations

- Implement stage-adjusted safety audits for high-speed rural single carriageway zones.
- Develop seasonal risk buffers (October–December) for emergency service resource allocation.
- Build real-time monitoring dashboards to detect early instability signals in infrastructure.

SECTION 2

Sector & Business Context

Sector Overview

The transport and infrastructure sector underpins economic activity across every industry vertical. It is characterised by high transaction volume, strong dependency on environmental and seasonal stability, and rapid growth in vehicle density — particularly in rural and semi-urban corridors where road infrastructure has not kept pace with demand.

Current Challenges

- **Funding volatility** for road repairs and preventive maintenance, creating deferred risk accumulation.
- **Over-expansion during capital surges** without corresponding safety infrastructure investments.
- **High-speed infrastructure risks** on single carriageways, where design speeds conflict with local usage patterns.
- **Seasonal blind spots** in emergency service deployment during Q4 winter months.

Why This Problem Was Chosen

Road accident patterns offer a measurable, data-rich lens into systemic infrastructure instability. Understanding structural casualty patterns enables better workforce planning for emergency services, more targeted infrastructure investment, and sustainable safety growth strategies. This project was selected because it represents a high-impact, analytically tractable problem that bridges data science with real-world public safety outcomes.

SECTION 3

Problem Statement & Objectives

Formal Problem Definition

To systematically analyze road accident trends across 2021–2022 and identify structural patterns of instability across road types, speed limits, lighting conditions, and geographic zones — enabling data-driven policy and resource allocation decisions.

Project Scope

- Measure total casualty magnitude and proportional intensity across severity tiers.
- Identify high-risk road surfaces, lighting conditions, and road type combinations.
- Evaluate the rural vs. urban risk differential using a quantified multiplier.
- Map seasonal trends to identify Q4 systemic risk corridors.
- Develop an executive-level interactive dashboard for stakeholder use.

Success Criteria

A well-defined KPI framework with measurable outputs, a fully functional interactive dashboard in Google Sheets, and a documented set of actionable recommendations validated against the data. Insights must be expressed in decision language appropriate for executive and policy-level audiences.

SECTION 4

Data Description

Dataset Source

Source: **UK Road Accident Dataset** — available on Kaggle. This dataset compiles police-reported road traffic accident records from across the United Kingdom, covering incident-level details including severity, location type, road conditions, lighting, and vehicle involvement.

Data Structure

Column / Feature	Description
Accident_Severity	Categorical: Fatal / Serious / Slight
Road_Type	Type of road: Single Carriageway, Dual, Roundabout, etc.
Urban_or_Rural_Area	Binary classification: Urban or Rural
Number_of_Casualties	Integer count of casualties per incident
Light_Conditions	Lighting at time of incident (Daylight, Dark, etc.)
Road_Surface_Conditions	Surface state: Dry, Wet, Snow, Frost, etc.
Vehicle_Type	Classification of vehicles involved
Date	Date of accident (used to extract Month and Year)
Speed_Limit	Posted speed limit at accident location
Junction_Detail	Type of junction, if applicable
Weather_Conditions	Weather at time of incident

Data Size & Limitations

Records: 10,000 rows | **Columns:** 23 | **Period:** 2021–2022

- Absence of detailed financial/profitability impact data limits economic quantification.
- Potential under-reporting bias in 'Slight' severity incidents due to discretionary reporting.
- Vehicle count and traffic volume data not available — limiting exposure-normalised risk rates.

SECTION 5

Data Cleaning & Preparation

All primary cleaning and transformation steps were executed in **Google Sheets** as per capstone requirements. The following procedures were applied to ensure data integrity:

Missing Values

Blank region fields were replaced with 'Unknown'. Missing road surface entries were flagged and imputed using modal values within matched severity categories. Null entries in lighting conditions were treated as 'Unknown' to preserve record counts.

Duplicate Removal

Duplicate rows identified by composite key (Date + Location + Severity) were removed. A total of 47 duplicates were eliminated, leaving 9,953 clean records.

Date Transformations

Month and Year were extracted from the raw date field using =TEXT() and =YEAR() formulas, enabling time-series aggregation by month and quarter.

Feature Engineering

A Severity_Intensity categorical column was created: Fatal → 'Critical', Serious → 'High', Slight → 'Moderate'. A Rural_Risk_Flag binary column was added to enable rural vs. urban pivot filtering.

Outlier Treatment

Casualty count outliers (values > mean + 3 SD) were reviewed. Multi-vehicle pile-up events with high counts were retained as valid extreme events.

Assumptions

Urban/Rural classification is taken as reported. Speed limit data is assumed accurate to the posted limit at time of incident. All monetary impact figures in Section 11 are directional estimates, not audited values.

SECTION 6

KPI & Metric Framework

KPI	Formula / Definition	Value	Why It Matters
Total Casualties	SUM(Number_of_Casualties)	13,522	Primary magnitude indicator
Casualty Intensity%	Avg casualties per incident x 100	~1.35	Measures severity density
Severity Ratio	Fatal : Serious : Slight count distribution	238 : 1,834 : 11,450	Identifies critical tier volume
Rural Risk Multiplier	Rural avg casualties / Urban avg casualties	1.16x	Quantifies geographic risk gap
Peak Month Index	Month with highest casualty count	Oct (1,265)	Informs seasonal planning
Dry Surface %	% accidents on dry road surface	66.7%	Challenges weather-first assumptions
Single Carriageway %	% casualties on single carriageway	73%	Identifies dominant risk road type

Each KPI was mapped directly to a project objective: magnitude KPIs address the scale of the problem; intensity KPIs reveal structural severity; geographic and surface KPIs isolate root cause variables. Together, they form a layered decision framework for stakeholders at both operational and executive levels.

SECTION 7

Exploratory Data Analysis (EDA)

Trend Analysis — Seasonal Patterns

Casualty data across 2021–2022 reveals a consistent Q4 peak, with October recording the highest monthly count at 1,265 casualties, closely followed by November at 1,254. This pattern aligns with reduced daylight hours, increased commuter traffic, and deteriorating road visibility. Q1 and Q2 months consistently record the lowest casualty volumes, confirming a strong seasonal cycle.

Month	Casualties	Trend (relative to peak month)
Jan	890	70%
Feb	810	64%
Mar	950	75%
Apr	960	76%
May	1,010	80%
Jun	1,020	81%
Jul	1,050	83%
Aug	1,080	85%
Sep	1,100	87%
Oct	**1,265**	100%
Nov	1,254	99%
Dec	1,133	90%

Figure 1: Monthly Casualty Distribution (2021–2022) — Indicative representation

Road Type Impact Analysis

Single carriageways dominate the casualty distribution, contributing **73% of all casualties (9,888 individuals)**. This disproportionate share — relative to their traffic volume — suggests a structural design-speed mismatch. Dual carriageways and roundabouts contribute significantly less, likely due to physical separation of opposing traffic flows.

Road Type	Casualties	Share %
Single Carriageway	**9,888**	**73.1%**
Dual Carriageway	1,756	13.0%
Roundabout	676	5.0%
One-Way Street	542	4.0%
Slip Road	406	3.0%
Other/Unknown	254	1.9%

Figure 2: Casualty Distribution by Road Type

Environmental & Surface Analysis

Counter-intuitively, **66.7% of all accidents occurred on dry road surfaces**. This finding challenges the common assumption that adverse weather is the primary accident driver. The data suggests that driver behavior — including speeding and inattention — under normal conditions represents a larger systemic risk than environmental degradation. Wet surface accidents account for approximately 27% of incidents, with snow, ice, and frost contributing the remainder.

Rural vs. Urban Risk Analysis

Rural areas recorded an average casualty intensity of **1.16× higher per incident** compared to urban zones. This differential is attributed to higher speed limits on rural roads, longer emergency response times, and lower lighting provision. Urban areas, while contributing higher absolute casualty volumes due to traffic density, show lower per-incident severity.

SECTION 8

Dashboard Design

Platform & Objective

The dashboard was built entirely in **Google Sheets** using Pivot Tables, dynamic named ranges, and interactive slicers — as per capstone requirements. The objective was to deliver an executive-level analytical interface enabling non-technical stakeholders to identify risk concentration patterns without requiring data expertise.

View Structure

Panel / View	Content	Chart Type
KPI Overview	Total Casualties, Fatal count, Severity Ratio, Rural Multiplier	KPI Cards
Road Type Risk	Casualties by road type, ranked bar view	Bar Chart
Monthly Trend	Casualty volume by month, 2021 vs 2022	Line Chart
Severity Breakdown	Fatal / Serious / Slight distribution	Donut Chart
Rural vs. Urban	Comparative casualty averages by zone	Grouped Bar
Surface Analysis	Accident distribution by road surface condition	Stacked Bar

Interactive Elements

- **Year Slicer:** Filter all views by 2021, 2022, or combined.
- **Urban/Rural Slicer:** Isolate geographic zone for comparative analysis.
- **Accident Severity Slicer:** Focus on Fatal, Serious, or Slight incidents.
- **Road Type Dropdown:** Drill into single carriageway vs. dual vs. other.

The dashboard was designed using industry-standard visual hierarchy: KPI magnitudes at the top, distribution charts in the middle section, and trend/comparison panels at the bottom. Color coding follows a consistent red (high risk) → amber → green (low risk) severity scale.

SECTION 9

Insights Summary

The following insights were extracted through systematic EDA and dashboard analysis. Each is expressed in decision language for executive and policy-level action.

01 **Road casualties are structural, not random**

73% of all casualties concentrated on a single road type (Single Carriageways) — indicating an infrastructure design problem, not an isolated behavior problem.

02 **High-speed rural roads are the primary danger zone**

Rural areas produce 1.16× higher casualty intensity per incident. Combined with faster speeds and longer response times, rural single carriageways represent the highest-priority intervention zone.

03 **Q4 is a systemic risk window**

October and November alone account for 18.6% of annual casualties. Emergency services must pre-position resources for this predictable seasonal surge.

04 **Dry roads are deceptively dangerous**

66.7% of accidents on dry surfaces disproves the weather-first assumption. Speed, distraction, and driver behavior — not weather — are the dominant risk variables.

05 **Fatal incidents are concentrated and predictable**

238 fatal incidents (1.76% of total) represent the critical tier. Cross-referencing road type and lighting data reveals that >60% of fatals occurred on rural single carriageways at night.

06 **Severity gradient is manageable**

85% of incidents are 'Slight' — meaning targeted intervention at the Fatal/Serious tier (15%) could yield disproportionate safety and cost benefits.

07 **Lighting conditions amplify rural risk**

Dark conditions with no lighting are overrepresented in rural fatal incidents. Infrastructure investment in rural lighting addresses multiple risk factors simultaneously.

08 **2022 shows marginal improvement over 2021**

Year-on-year comparison reveals a ~3% reduction in total casualties in 2022, suggesting existing interventions are marginally effective but insufficient at scale.

SECTION 10

Recommendations

Recommendation 1: Targeted Seasonal Enforcement

Insight: Q4 peak (Oct–Nov accounts for 18.6% of casualties).

Action: Align police enforcement schedules and speed camera operations with peak volatility months. Pre-deploy mobile speed enforcement units on rural single carriageway corridors from October through December.

Impact & Feasibility: High feasibility — leverages existing infrastructure. Estimated 8–12% reduction in Q4 fatals within 2 years.

Recommendation 2: Rural Single Carriageway Safety Audits

Insight: 73% casualty concentration + 1.16× rural risk multiplier.

Action: Commission stage-adjusted safety audits for all rural single carriageways with speed limits ≥50mph. Audit outputs should trigger targeted interventions: rumble strips, speed reduction zones, and improved lane delineation.

Impact & Feasibility: Medium-term investment with long-term payoff. Reduces structural risk concentration.

Recommendation 3: Rural Lighting Infrastructure Program

Insight: Dark conditions overrepresented in rural fatal incidents.

Action: Prioritize solar-powered or smart lighting installation on rural single carriageway sections with historical fatality clusters. A phased 3-year rollout targeting top 200 risk segments would address >60% of lighting-related fatal incidents.

Impact & Feasibility: High impact. Addresses multiple risk factors (speed, visibility, behavior) simultaneously.

Recommendation 4: Predictive Monitoring Dashboard (Real-Time)

Insight: Data-driven patterns are predictable with high confidence.

Action: Build a real-time accident monitoring system connected to police reporting infrastructure. Early warning triggers should activate when monthly casualty counts exceed Q4 seasonal thresholds, enabling pre-emptive resource deployment.

Impact & Feasibility: Medium feasibility — requires API integration with STATS19 reporting system. High strategic value.

SECTION 11

Impact Estimation

The following impact projections are directional estimates derived from applying recommendation logic to observed casualty patterns. They are not audited financial projections but represent reasonable first-order approximations for policy planning purposes.

Impact Area	Current State	Expected Improvement	Mechanism
Reactive Safety Costs	High concentration in Q4 and rural zones	10–15% cost reduction	Disciplined pre-positioning of resources reduces reactive emergency spend
Fatal Incident Rate	238 fatals / 13,522 total (1.76%)	8–12% reduction in fatals	Targeted lighting + enforcement on highest-risk rural segments
Emergency Response Efficiency	Rural response lag amplifies casualty severity	Improved alignment of EMS velocity with peak cycles	Seasonal pre-deployment informed by predictive monitoring
Infrastructure Resilience	Reactive repair cycle post-accident	Proactive audit-driven investment	Stage-adjusted audits reduce deferred risk accumulation
Seasonal Risk Window	Q4 contributes ~28% of annual casualties	Q4 share reducible to ~22–24%	Combined enforcement + monitoring interventions during Oct–Dec

SECTION 12

Limitations

- **Absence of traffic volume data:** Without normalised exposure rates (casualties per vehicle-km), absolute casualty counts may overstate risk on high-traffic roads and understate risk on low-traffic rural segments.
- **Reporting bias in 'Slight' severity:** Minor incidents may be under-reported due to voluntary reporting thresholds, potentially inflating the proportion of Serious/Fatal incidents in the dataset.
- **No causation inference:** The analysis identifies correlational patterns across road type, surface, and lighting variables. Causal claims require controlled experimental or quasi-experimental design.
- **Dataset scope:** 10,000 records represent a sample, not the full UK accident population. Results are directionally indicative but should be validated against the full STATS19 dataset before policy implementation.
- **No financial impact quantification:** Cost-of-accident data was not included, limiting the economic case for recommendations to qualitative arguments.
- **Assumption risks:** Urban/Rural classifications, speed limit accuracy, and weather condition reporting are taken as given from police reports. Recording inconsistencies may introduce noise.

SECTION 13

Future Scope

Additional Analysis

- Time-series forecasting (ARIMA or Prophet) to predict Q4 casualty volumes 90 days in advance.
- Geospatial clustering analysis to identify exact road segments with highest fatality density.
- Multi-variable regression to isolate the independent contribution of speed limit, lighting, and surface condition to fatal incident probability.
- Cohort analysis comparing 2021 vs. 2022 response to existing interventions.

New Data Requirements

- Full STATS19 dataset (all UK police-reported accidents, not a 10k sample).
- Traffic volume data by road segment to enable exposure-normalised risk rates.
- Infrastructure spend data to correlate investment with casualty reduction.
- Real-time weather and lighting condition APIs for predictive monitoring integration.
- Emergency service response time logs to quantify the rural severity amplification effect.

SECTION 14

Conclusion

This project has delivered a structured, evidence-based analysis of UK road accident patterns across 2021–2022, transforming 10,000 raw records into a clear set of actionable insights and policy recommendations.

The central finding — that **73% of casualties concentrate on a single road type** (Single Carriageways), compounded by a **1.16× rural risk multiplier** and a predictable **Q4 seasonal peak** — points conclusively to a structural, not random, safety crisis. The high proportion of accidents on dry surfaces further redirects policy attention from weather mitigation to behavioral and speed-management interventions.

The KPI framework, EDA, and interactive Google Sheets dashboard produced in this project provide stakeholders with a reusable analytical infrastructure — one that can be updated with new data to track the effectiveness of recommended interventions over time.

With targeted implementation of the four recommendations outlined in Section 10, a conservative **10–15% reduction in reactive safety costs** and an **8–12% reduction in fatal incident rates** are achievable within a 2–3 year horizon — representing a meaningful, data-justified return on public safety investment.

SECTION 15

Appendix — Contribution Matrix

The following table documents the contribution of each team member across all project stages. Contribution claims are verifiable through Google Sheets Version History and submitted working files.

Team Member	Dataset & Sourcing	Data Cleaning	KPI & Analysis	Dashboard Development	Report Writing	PPT Preparation	Overall Role
Aadit 2401010003	Dataset coordination & validation	Category standardization	Pivot tables & statistical checks	Industry visuals support	Recommendations drafting	Formatting support	Analysis Lead
Krish Garg 2401020097	Primary dataset sourcing	Cleaning support	KPI framework support	Dashboard structuring support	Major report writing	Major PPT making	PPT & Overall Lead
Manya Verma 2401010265	Dataset review	Cleaning validation check	Funding & trend KPI calculations	Dashboard layout & filter testing	Data section support	Led PPT design & finalization	Dashboard Lead
Abuzar Haideri 2401010024	Dataset validation support	Final review support	Risk % validation	Final dashboard refinement	Problem framing input	PPT structure support	Strategy Lead
Tathagat Harsh 2401010477	Sourcing discussion support	Major cleaning execution	KPI support & risk metric inputs	Layout support	EDA writing	Presentation rehearsal support	Project Lead
Tanish Rao 2401010474	Sourcing support	Category creation & cleaning support	Metric cross-check support	Trend charts support	Advanced analysis inputs	Analysis slides	Data Lead