Analysing Road Safety for School Routes using STATS19 and Correlation Analysis

Zofia Cywinska, Khushi Sharma, James Cartlidge, Daniel Porges

May 16th, 2025

Abstract

Road safety for school children remains a significant public concern in the United KIngdom, particularly in regions such as South Yorkshire, where rising casualty rates have drawn increasing attention from local authorities and stakeholders.[check] This report investigates the safety of school routes using the STATS19 road safety dataset, integrating geographic, infrastructural, and socio-economic data to identify patterns of risk and propose targeted improvements.

This project focuses on three primary objectives: (1) ranking school routes by collision and casualty risk; (2) applying machine learning models to predict casualty severity and identify key environmental, infrastructural, and demographic factors associated with elevated danger; and (3) further analyse these factors to provide actionable insights for creating safer commuting journeys to schools.

Through a combination of statistical analysis and predictive modelling, this report uncovers spatial trends and feature correlations that can inform practical recommendations. The insights gained aim to support local councils, transport planners, and policymakers in enhancing school route safety, reducing child casualties and promoting safer, more sustainable travel in South Yorkshire and beyond.

Index

1 Introduction

Road safety remains a critical public health concern across the United Kingdom, particularly when it comes to protecting vulnerable groups such as schoolchildren during their daily commutes. Despite numerous efforts to improve infrastructure and raise public awareness, road traffic collisions continue to pose significant risks, with severe outcomes not only for the individuals directly involved but also for the broader communities they affect. In 2023, there were [] road fatalities and [] serious injuries across the UK.

...

This project aims to analyse road safety patterns using the STATS19 dataset, a comprehensive collection of reported road collision data maintained by the UK Department of Transport. By integrating this dataset with geographic and route information, the study seeks to:

1. Identify and rank school routes based on collision and casualty risk.
2. Investigate environmental, infrastructural, and socio-economic factors associated with elevated risk levels.
3. Develop and apply machine learning models to casualty severity and assess feature importance.
4. Generate actionable recommendations for improving road safety on school routes.

Through this approach, the project endeavors to uncover spatial and statistical patterns that can inform policymakers, urban planners, and stakeholders in implementing targeted interventions to enhance the safety of school commutes.

2 Report Structure

This report is organised into several key sections to provide a clear and logical narrative of the project. The **Aims** section outlines the project's main goals and research questions, which guide the focus of our analysis and interpretation of the results. The **Data Sources & Tools** section describes the datasets, external sources, and software environments used, explaining how the data was acquired and processed throughout the project.

The **Methodology** section details the step-by-step process followed for preparing, cleaning, and analyzing the STATS19 data, generating school routes, and merging the two for combined analysis. It also covers the machine learning approaches and correlation analysis methods used in the project. The **Analysis** section presents the exploration of routes and preliminary data insights, including key aspects such as missing value analysis, correlation analysis, and other investigative steps leading into the main findings. In the **Results** section, we present the core analytical outcomes, highlighting key patterns, relationships and predictive insights derived from the dataset.

The **Discussion** section interprets these results, explores their significance, and connects them back to the project's stated aims, while also addressing the broader implications of our work. It also includes a necessary reflection on the limitations and challenges encountered throughout the project, ensuring transparency about the constraints of the data, methods, and interpretations. The **Conclusions** section summarises the key takeaways from the project, followed by the **Further Improvements** section, where we suggest future directions and recommendations for enhancing data availability, methodological approaches, or analysis in similar research.

Finally, the **Acknowledgements** section formally recognises the project contributions, while the **References** section lists all academic, data, and software sources cites or consulted. Any supplementary materials, such as additional figures, tables, analysis code, or supporting details, are provided in the **Appendices**.

3 Aims

The aims of this report are to investigate and understand the factors affecting road safety on school routes using the STATS19 road collision dataset and associated geographic data. Our goals are structured to guide both exploratory analysis and predictive modeling, ultimately aiming to support data-driven recommendations for safer school commutes.

Our clients asked us to investigate road safety in South Yorkshire, focusing on the demographics of school children. Specifically, they had an interest in which features of these most dangerous routes were correlated, and hence how this danger was caused.

To answer these questions, the report seeks to: Identify the most dangerous and safest routes to schools across South Yorkshire. Predict the seriousness of casualties using machine learning models, examining which features (variables) most strongly influence casualty severity. Analyse how these features correlate with the identified dangerous or safe routes, aiming to uncover patterns related to infrastructure, socioeconomic factors (e.g. Index of Multiple Deprivation (IMD)), and environmental conditions. Understand why certain routes stand out as particularly dangerous or safe, by exploring possible correlations between key risk factors and route characteristics.

We aim to deliver a clear, analytical study that follows these guiding research questions:

3.1 Route Risk Identification

This section addresses the identification and ranking of school routes based on risk.

1. What school routes are used? *(edit / make routes)*
2. Can we use machine learning to identify risk levels associated with each school route? *(define the risk level of routes)*
3. Which school routes have the highest and lowest levels of commute safety? *(rank the routes by their risk levels)*

3.2 Correlation Analysis

We explore which factors are most strongly associated with dangerous or safe routes.

1. What factors are most strongly associated with the dangerous or safe routes?
2. Why are these factors correlated with risk levels?

3.3 Further Analysis

Finally, we investigate potential improvements and actionable insights.

1. Can we analyse these factors and the historical data to find improvements or potential actionable insights?

2. How can the insights from this analysis inform recommendations for improving school route safety?

[Placeholder: write a paragraph on our interpretations and solutions for the questions above]

4 Data Sources & Tools

4.1 Data Sources

4.1.1 STATS19

The STATS19 dataset is sourced from the UK Department for Transport's official Road Safety Data page, available on data.gov.uk. This dataset provides detailed records of reported road collisions across the entire UK, covering the period from 1979 to 2023. While the full dataset spans multiple decades, we primary focus on recent subsets where relevant to our analysis.

The STATS19 data is divided into three interconnected datasets:

- Collisions dataset: Contains core information for each police reported collision, including the location, date, time, severity, environmental conditions, and key identifiers (such as 'accident_reference') that link to the other two datasets.
- Vehicles dataset: Provides detailed information on each vehicle involved in a reported collision, including vehicle characteristics and driver details where available.
- Casualties dataset: Includes information on individuals who were injured or killed as a result of the collision, including demographic details and injury severity.

The three datasets are linked through the shared accident_reference identifier, which ensures that vehicle and casualty records can be accurately matched to their corresponding collision events. Additional identifiers, such as casualty_reference and vehicle_reference, are unique only within each collision, reflecting the many-to-one relationship between a single collision event and the multiple vehicles and casualties it may involve.

It is important to note that while the most recent data up to 2023 is included, the 2024 dataset was excluded from this project as it has not yet been verified or finalised for official use.

5 Methodology

5.1 STATS19 Dataset Preparation

intro

5.1.1 Merging the 3 Datasets

6 Analysis

7 Results

8 Discussion

8.1 Limitations

9 Conclusion

10 Further Improvements

11 References

12 Appendices