Analysing Road Safety for School Routes using STATS19 and Correlation Analysis

Zofia Cywińska, Khushi Sharma, James Cartlidge, Daniel Porges

May 16th, 2025

Executive Summary

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. This report investigates the safety of school routes using the STATS19 road safety dataset, integrating geographic and infrastructural data to identify patterns of risk and propose targeted improvements. The study was commissioned by the South Yorkshire Mayoral Combined Authority (SYMCA) [1]and developed in close collaboration with Ms. Anna Butler [2] whose input was instrumental in shaping its aims and focus.

This project focuses on three primary objectives: (1) ranking school routes by collision and casualty densities, and applying machine learning models to categorise the school routes and identify their commute risk levels; (2) analysing infrastructural data from STATS19 alongside the risk-labelled routes to identify correlations between various factors and commute risk levels; and (3) conducting case studies to assess these factors in depth and generate actionable insights for safer school journeys.

This work done in this project:

- developed several commute risk metrics and applied machine learning analysis to identify abnormally dangerous routes. This analysis provides a scalable framework for future research into road safety,
- identified 10 Critical Risk routes in each borough (Barnsley, Doncaster, Rotherham, Sheffield),
- identified specific lack of pedestrian safety measures in Critical Risk routes, such as high or ignored speed limits and lack of crossings,
- developed a new Python package `googleroutes` to support future research and large-scale route generation.

The report also presents two case studies including an analysis of real-world route usage data from BetterPoints in South Yorkshire, creating the first list of dangerous routes normalised by foot-traffic. This analysis revealed that use of this data provides deeper insight into routes that aren't frequently used by people, so don't have many casualties on them, but are still inherently dangerous, particularly for school children. This first step in a more advanced analysis offers valuable insights for future research to identify and prioritise interventions on traditionally overlooked routes. By focusing on these routes, urban planners and authorities can more precisely and effectively distribute resource to a wider range of urban environments. By building on the methodology of this report, future studies can develop more comprehensive data-driven strategies that protect vulnerable populations, such as school children.

[1]SYMCA's Homepage: https://southyorkshire-ca.gov.uk/

[2]Ms. Anna Butler, Active Travel Manager for SYMCA. LinkedIn: https://uk.linkedin.com/in/anna-butler-33a48880

To support transparency and encourage further exploration, all findings, visualisations, methodology and supplementary resources are publicly available via the project website, PHY346 Road Safety Dashboard[3].

Recommendations Based on our analysis of road safety data in South Yorkshire, we recommend that SYMCA prioritise interventions on school routes identified as *Critical* or *High Risk*, especially in Doncaster and Sheffield. Enhancing pedestrian safety measures - crossings, lights and traffic-calming measures - and speed limit enforcement near schools should be immediate priorities. We recommend incorporating behavioural data from BetterPoints and VivaCity help better understand route usage and highlight high-risk areas that are under analysed. Finally, targeted intervention in areas of high deprivation is needed to ensure equitable protection for all school children.

[3]Project website: https://phy346-road-safety-portfolio.vercel.app

Index

1 Introduction

South Yorkshire's goal to encourage active modes of transport among schoolchildren is exemplified by the Mayor's manifesto, where he pledges to make South Yorkshire the best place in the country for walking and cycling , particularly for young people [1]. This ambition is echoed by the region's Active Travel Commissioner, Ed Clancy OBE, who emphasises that enabling children to walk or cycle not only keeps them active but helps build healthy habits for life in the formative years [2].

However, despite these initiatives, many parents remain hesitant to allow their children to walk or cycle to school, with one of the primary concerns being road safety [3]. This concern is far from unfounded: in 2013, an average of 37 children were seriously injured and one child was killed every week on UK roads [4]. To prevent such tragedies, it is vital to undertake a thorough analysis of the collision and casualty data to uncover actionable insights for planners and policy makers. Despite the urgency, there remains a notable lack of detailed research into schoolchildren casualties and the factors contributing to them, particularly within South Yorkshire.

Although some studies have addressed road safety among school-aged children [5], most have focused on behavioural interventions, such as improving children's road-crossing behaviour, rather than analysing real-world casualty data. Furthermore, several reports related to the STATS19 dataset can be found; however, they are mostly outdated and focus on the whole of the UK or on major cities, such as London [6]. In particular, a study based on the STATS19 dataset for the whole country was conducted to study the effectiveness of additional lighting in reducing casualties, but ultimately concluded a positive relationship - indicating that increased street light frequency, surprisingly, was associated with an increase in casualties [7]. The lack of research focusing particularly on casualties of school-aged children in South Yorkshire is without a doubt a gap in the literature, which needs to be urgently filled.

In response to this gap, this project, commissioned by the South Yorkshire Mayoral Combined Authority (SYMCA) [1]) and developed in collaboration with Ms. Anna Butler [2], Active Travel Manager for SYMCA, aims to analyse road safety patterns using the STATS19 dataset, which is the UK Department of Transport's official record of police-reported road collisions. This project integrates it with geospatial data and generated school routes to explore risk at a granular level. Focusing on child pedestrian and cyclist casualties and their relationship with route characteristics and infrastructural data, our approach combines spatial analysis with machine learning techniques to better understand

[1]SYMCA's Homepage: https://southyorkshire-ca.gov.uk/

[2]Ms. Anna Butler, Active Travel Manager for SYMCA. LinkedIn: https://uk.linkedin.com/in/anna-butler-33a48880

where risk is concentrated, what factors are associated with elevated danger, and how these insights can support targeted improvements in infrastructure and policy.

1.1 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 the following 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 analysing 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, core analytical outcomes are presented, 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 this project. It also includes recommendations, as well as 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 future directions and recommendations for enhancing data availability, methodological approaches, or analysis in similar research are suggested.

1.2 Aims

This project investigates the safety of school routes in South Yorkshire by analysing collision data, generating commute routes, and applying machine learning to uncover risk patterns. The aim is not only to identify which routes are most dangerous, but also to understand what physical or environmental factors are contributing to that risk.

The clients, the Active Travel Team at the South Yorkshire Mayoral Combined Authority (SYMCA), wanted a project which explored road safety with a particular focus on school-age children and their daily commutes. They were interested in identifying the most dangerous school routes across the region and understanding the underlying causes of casualties near schools. In particular, they wanted to know whether specific infrastructural features, such as road type, speed limit, or junction design, could be statistically linked to elevated risk. BetterPoints (described in detail in section 2.1.4) was also a point of interest due to the behavioural data which could be used to contextualise casualty and collision rates by comparing them to actual route usage or popularity.

These interests helped define the scope of the project and guided our focus toward a route-level analysis, combining statistical and spatial methods to produce insights that are both data-driven and actionable. To address these objectives, the report frames its investigation around the following three areas of interest:

Route Risk Identification

1. How can school commute routes be realistically modelled using available mapping data?
2. What is a fair and meaningful way to define commute risk for each route, taking into account length and exposure?
3. Can unsupervised machine learning techniques, such as clustering and anomaly detection, be used to classify school routes into interpretable risk categories?
4. Which routes emerge as the most risky and needing attention, and how are they distributed across different boroughs?

Infrastructure Feature Analysis

1. What road or environmental features are most strongly associated with elevated route-level risk?
2. Can machine learning models reliably predict risk levels based solely on features such as road type, speed limit, junction control, and pedestrian crossing facilities?
3. Which features contribute most to severe outcomes, based on model interpretation methods like SHAP?

Case Studies and Real-World Context

1. Do the modelled risk classifications align with actual infrastructure conditions when routes are inspected visually?
2. How does behavioural data (e.g. BetterPoints route usage) compare with predicted risk levels, and can it help prioritise intervention?
3. What real-world insights can be drawn to support actionable improvements in school commute safety?

2 Data Sources & Tools

2.1 Data Sources

2.1.1 STATS19

The STATS19 dataset is sourced from the UK Department for Transport's official Road Safety Data page, available on data.gov.uk[8]. It contains detailed records of reported road collisions across the entire UK, spanning from 1979 to 2023. For the purposes of this project, the analysis is restricted to more recent years, corresponding to the introduction of updated data collection standards in 2016 and the availability of geospatial coordinates required to integrate collision data with school route information after 1999.

The STATS19 data is divided into three interconnected datasets:

- Collisions dataset: Records 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: Provides information for 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.

2.1.2 City Council Websites

To identify and analyse routes to schools across South Yorkshire, it was first necessary to compile an accurate and up-to-date list of active primary and secondary schools in

the region. This information was obtained from the official websites of the four local authorities: Sheffield, Doncaster, Barnsley, and Rotherham.

Doncaster, Barnsley, and Rotherham each maintain dedicated webpages listing local schools, typically including school names, types, and addresses. In contrast, Sheffield City Council publishes school data in the form of downloadable PDF documents, updated periodically throughout the academic year. For consistency with the available collision data, the school list dated January 2023 was used in this project[9].

Although most council websites do not explicitly state when their lists were last updated, their official status ensures a reasonable level of reliability for spatial analysis and route generation.

2.1.3 School Catchment Zones

To support realistic route generation, geospatial data on school catchment areas was incorporated into the analysis. Among the four South Yorkshire boroughs, Sheffield is the only local authority that publishes official catchment zone boundaries for its schools in a geospatial format.

Catchment zones for both primary and secondary schools in Sheffield were obtained as shapefiles for the 2024–2025 academic year [10][11]. These files, accessible via the Sheffield City Council's online GIS portal, are updated annually and provide reliable spatial boundaries that define expected pupil intake areas.

These catchment shapefiles were used to approximate student origin zones for each school. This enabled the generation of realistic walking routes from residential areas to school entrances, forming a critical input to the route modelling and subsequent risk assessment stages of the project.

2.1.4 BetterPoints Data

BetterPoints is a mobile application designed to incentivise active travel by tracking users' journeys and awarding points for walking, cycling, or using public transport [12]. The app uses GPS data to record the start and end locations of each journey and applies heuristics based on distance and duration to infer the mode of transport.

Through a partnership between Sheffield City Council and BetterPoints, a limited sample of journey data was made available for this project. The dataset consisted of anonymised trips to and from Mundella Primary School in Sheffield. Due to the need for data anonymisation and the scope of the pilot, this was the only school for which data was provided.

While the dataset was too limited to support generalisable behavioural analysis, it served as a valuable case study. In particular, it provided an opportunity to compare the modelled school routes generated via the Google Maps API against actual routes used by families, offering insight into the limitations of algorithmically-generated paths and the value of real-world behaviour data.

2.2 Tools

A range of tools and platforms were used throughout this project to support data acquisition, geospatial processing, route generation, analysis, and visualisation. The primary tools are summarised in the table below, with their respective versions at the time of use noted:

Tool Used	Description	Link
Python 3.12.4	General-purpose programming language used for route generation and data analysis.	-
QGIS	Open-source GIS software used to process and visualise geospatial data.	qgis.org
ArcGIS Online	Cloud-based platform for geospatial visualisation and map sharing.	arcgis.com
OpenRouteService (ORS)	Used to generate school isochrones through a QGIS plugin.	openrouteservice.org
Google Maps Directions API	Used to generate realistic walking routes to schools.	cloud.google.com

Table 2.1: Key software used in the project, a description of the software and a link to their homepage. QGIS citation [13]

In addition, multiple Python packages were used for data analysis, and a new package called `googleroutes` was developed for use with the data in this project. A full list with citations can be found in Appendix B.2.

3 Methodology

This section outlines the processes used to prepare datasets, generate school commute routes, and integrate them for analysis. The objective was to evaluate the safety of school routes in South Yorkshire using STATS19 collision records and modelled walking routes to primary and secondary schools.

The methodology consists of three stages. First, the STATS19 collision, vehicle, and casualty datasets were merged, cleaned, and labelled to ensure consistency and interpretability. Second, walking routes were generated from schools to residential origin points, based on either official catchment zones or estimated isochrones. Third, the generated routes were spatially linked to collision data through buffering and joining techniques, resulting in a unified dataset that connected route geometry with casualty records.

The final dataset enabled route-level analysis of collision and casualty patterns, forming the basis for risk classification and correlation analysis.

3.1 STATS19 Dataset Preparation

This subsection outlines the processing steps applied to the STATS19 collision, vehicle, and casualty datasets to enable route-level risk analysis. The three datasets were first merged, followed by exploratory analysis and data cleaning. Missing values were systematically identified and handled, and categorical codes were converted into human-readable labels using an external keys dataset. The final output was a filtered and cleaned dataset, restricted to South Yorkshire and the years 2016–2023, suitable for integration with route data and further analysis.

3.1.1 Merging the STATS19 Datasets

The STATS19 data is structured across three separate datasets: collisions, vehicles, and casualties. Each contains key information relating to police-reported road incidents, but individually they fail to provide a complete picture. Therefore, the separate datasets must be unified into one, in order to enable a comprehensive analysis. To achieve this, a carefully ordered merging process was followed.

The vehicles dataset was first joined to the collision dataset through the virtue of the shared "accident_index" identifier, which unlike the "accident_reference", also encore

the year. Following the initial merge, the casualties dataset was linked to the combined dataset by utilising the same method. This two-step process was an essential step to ensure the accurate alignment of related records, while avoiding redundant or mismatched entries.

A key challenge in this process is the large combined file size. When merging all available years of STATS19 data (1979–2023), the total size reaches approximately 3.4–3.5 GB. Processing datasets of this size risks exceeding system memory limits and can lead to crashes or extremely slow runtimes. There are two strategies that can address this challenge:

1. **Reducing the temporal scope**: This strategy narrows the data range, such as focusing only on the last five years, significantly lowering the data size. Focused projects may use this approach when historical data is less critical to their specific research questions.
2. **Chunked processing**: Instead of restricting the dataset's temporal scope, this technique relies on processing the full multi-year dataset in smaller segments (chunks). Using pandas' chunk processing capabilities, the data can be loaded, merged, and filtered piece by piece, reducing the memory burden at each step. This approach, implemented in the project and available in the code repository linked on the project website, significantly improved performance and ensured that the merging process could be completed even on devices with limited RAM. It also allowed the full dataset to be retained for a more complete analysis.

The merged dataset was then filtered for South Yorkshire, using the police force code [14] and it was verified that the number of collisions in the dataset matched the numbers reported by the Department for Transport. The following preliminary data cleaning was performed by removing redundant or duplicate columns resultant from the merging process, as well as by reordering the identifier columns (accident_index, vehicle_reference, casualty_reference) for consistency.

3.1.2 Preliminary Exploration and Cleaning

Following the successful merging of the STATS19 dataset, an initial exploratory phase was undertaken to generate preliminary visualisations and descriptive statistics, such as plots of collisions counts by year, distributions of casualty severity, as well as spatial maps of accident locations. While useful for providing early insights, they also highlighted several problems with the raw dataset, which required a detailed evaluation before undertaking further analysis.

Key problems included missing data, incomplete records, and a lack of consistent formatting. For instance, some fields were sparsely populated in early years, leading to artificially low or inconsistent counts. Several columns were only systematically recorded in more recent years, making cross-year comparisons unreliable without assessing data completeness.

To prepare the dataset for statistical analysis and machine learning, it was necessary to identify and minimise missing or invalid values. This helped establish when specific

fields became reliably recorded, highlight any abrupt changes in reporting practices, and determine which years of data could be used with confidence.

Although pandas provides simple tools to check for null values, the STATS19 dataset often uses placeholder values such as -1 to indicate missing or inapplicable entries. These values are not formally marked as missing, so a full missing value analysis was required to detect and convert them to NaN, the standard missing value type in pandas. Care was taken to ensure that such placeholders were genuinely invalid and not legitimate numerical values like zero.

The missing value analysis produced several key outputs, including a ranked list of columns by percentage of missing values and temporal graphs showing how data completeness evolved over time. These results (see Appendix D)are supported by figures and tables illustrating completeness trends. Notably, records prior to 2004 exhibited substantial gaps and inconsistencies. While data quality improved thereafter, the most consistent and complete coverage was observed from 2016 onward. Based on these findings, and to align with the constraints introduced during route generation, the final analytical dataset was restricted to the period from 2016 onward, ensuring higher data quality and minimising bias from missing entries.

This cleaned and temporally filtered dataset provided a robust foundation for the subsequent stages of analysis, including correlation assessment, risk ranking, and machine learning modelling.

3.1.3 Labelling the Data and Keys Dataset

Many columns in the STATS19 dataset use numerical codes to represent categorical information, for example, road type, junction control, weather conditions, or casualty class. While these values are consistent and compact, they are not human-readable and can be easily misinterpreted in both exploratory analysis and machine learning workflows. For instance, models may incorrectly assume ordinal relationships between encoded categories, leading to unintended bias or poor predictive performance.

To make the dataset more interpretable and suitable for classification tasks, a keys dataset was created to provide mappings from numerical codes to their descriptive labels. These mappings were sourced from the official STATS19 variable description file provided by the Department for Transport, distributed in `.ods` (OpenDocument Spreadsheet) format. The relevant entries were extracted, cleaned, and reformatted into a Python dictionary structure, enabling programmatic relabelling of the main dataset.

Once constructed, the dictionary was applied to convert all relevant columns into their labelled equivalents. This transformation allowed categorical data to be correctly interpreted during analysis, improved the clarity of visualisations and summary statistics, and ensured proper handling by encoding pipelines during machine learning.

The final labelled version of the dataset preserved the original structure but significantly improved readability and model compatibility, serving as the basis for all subsequent exploratory and predictive tasks.

3.1.4 Summary

To summarise, first, the scope of our analysis was narrowed to focus on recent years, specifically 2016–2023. This decision was justified by both data quality, as reporting standards changed around 2016, improving consistency, and by the fact that older data was less relevant to the current state of school route safety. Second, the merged dataset was filtered to retain only records corresponding to South Yorkshire, identified using the police force code [14], which also reduced the dataset size to under 150 MB and made it more manageable.

Then null and missing data points were then removed from the filtered dataset, and placed aside for later recording. After, the dataset was converted from numerical values, to human readable strings, using a comprehensive keys dataset, mapping the coded value and its label. This made it possible to be read both by humans and machine learning analysis.

3.2 Routes Generation

To enable route-level analysis of school commute safety, walking routes were generated for every primary and secondary school in South Yorkshire. These routes served as the spatial framework onto which collision data could be mapped and analysed.

While route generation APIs exist, producing thousands of routes programmatically required a consistent, scalable method. This section outlines the full pipeline: identifying school locations, defining realistic origin points around each school, and generating walking routes between these points and school entrances using the Google Maps API. The resulting routes formed the basis for calculating route-specific risk metrics, such as casualty density and exposure-adjusted risk.

The process was designed to ensure geographic realism and broad coverage, balancing available data (e.g., catchment zones or isochrones) with practical limitations such as API access and processing time.

3.2.1 Creation of School Location .csv files

Accurate geolocation of all primary and secondary schools in South Yorkshire was an essential first step in the route generation process. Official city council websites were consulted to obtain comprehensive school lists. These typically included the name, type (primary or secondary), and address of each school but did not provide geographic coordinates.

To address this limitation, each school was manually geolocated using Google Maps. The latitude and longitude for every school were recorded in a shared Google Sheet, with a separate spreadsheet created for each borough. Each entry included the school's name, education level, and coordinates. These spreadsheets were then exported as CSV files, enabling integration into geospatial software such as QGIS. The resulting datasets served as the foundation for generating the walking routes by specifying the endpoint of each

route.

While this manual approach was effective, it was also time-consuming and dependent on the availability of accurate address data. It would be beneficial if local or combined authorities made school location data publicly available in structured, geocoded formats (e.g., CSV or GeoJSON). Making this data accessible on council websites would reduce the overhead for similar projects in the future and support wider data-driven planning efforts.

3.2.2 Selecting Route Origins For Each School

To generate meaningful and geographically accurate walking routes to schools, it was first necessary to define suitable origin areas for each institution. These origins served as starting points for route generation and were designed to represent where students are likely to begin their daily commute. Depending on data availability in each borough, origin zones were based on either official catchment boundaries or generated isochrones. This section outlines the rationale behind each method, the tools used to implement them, and how origin points were placed for later use.

Isochrones vs Catchment Zones

In Sheffield, official catchment zone data was available for both primary and secondary schools and was used wherever possible. These zones define the areas from which schools typically admit students, making them a reliable way to approximate where students live. Using these boundaries helped ground the analysis in actual local authority planning and made the results more directly relevant to school admissions and transport decisions.

However, for Barnsley, Doncaster, and Rotherham, catchment zone data wasn't publicly available, and SYMCA was unable to provide it during the time frame of the project. A Freedom of Information request was considered but ultimately ruled out due to time constraints. In place of catchments, isochrones were used. An isochrone shows the area that can be reached within a certain walking distance or time from a school. While it doesn't represent enrolment boundaries, it does give a practical and geographically grounded estimate of the school's local catchment in terms of walkability. This approach made it possible to apply a consistent method across the whole region, even in areas where official data was lacking.

Isochrone Generation

Isochrones were generated in QGIS using the `ORS Tools` plugin. The `ORS Tools` plugin was chosen due to its ready availability in QGIS, and the reputation of ORS as a reliable API for such tasks. The CSV file containing the coordinates of each school in a region, separated into primary and secondary, was imported as a vector layer, and the `Isochrones from layer` function under `Batch Jobs` was used. A full process for isochrone generation can be found in the `googleroutes` Github repo.

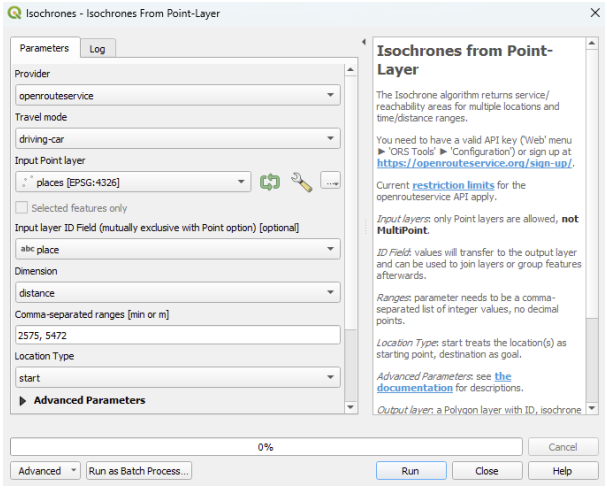


Figure 3.1: Parameters used for isochrone generation in QGIS

The parameters used can be seen in Figure 3.1. The distances of 2575m (∼1.6 miles) and 5472m (∼3.4 miles) were chosen as the average distances walked to school by primary and secondary school students respectively [15]. These distances represent a student's typical walk to school, however, they fail to capture the nuance of locations with smaller catchment areas, where it is uncommon to walk such distances. For journeys longer than the stated values, students are shown to typically use alternate forms of transportation, such as buses or cars [15].

Perimeter Points Generation

Once the catchment areas or isochrones were obtained, evenly spaced origin points were created along their boundaries using the `geometry to points` tool in QGIS. The number of points placed around each zone could be adjusted depending on the desired density of routes to be generated. An example of these perimeter points is shown in QGIS in Figure 3.2). Each set of these generated points was saved in an individual named file corresponding to its school. These files were structured so they could later be matched back to the original school location CSV files, upon which the isochrones were based in order to generate routes.

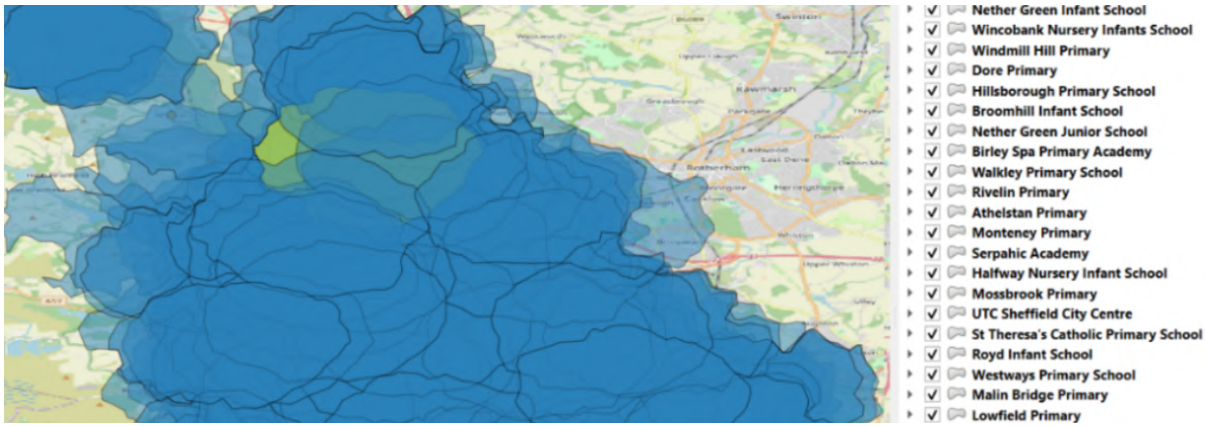


Figure 3.2: Demonstration of isochrones for schools in South Yorkshire.

3.2.3 API selection for route generation

Initially, both the Google Maps Directions API and OpenRouteService (ORS) Directions API were evaluated for generating walking routes. During testing, several issues became apparent with the ORS API. The most significant constraint was its rate limit of 20 requests per minute, which made it impractical for generating over 100,000 routes within a reasonable time-frame. In addition, the routes returned by ORS were considered less representative of how students actually travel to school. Since ORS is not widely used for journey planning, especially among school-aged children, its output was unlikely to reflect real-world commuting behaviour.

Taking these factors into account, the project shifted focus entirely to the Google Maps API. While Google Maps provides more realistic and accurate walking routes, it also has some drawbacks. The main limitation is the financial cost, as the API is part of the paid Google Cloud platform. This was managed by using free credits available when creating new Google Cloud accounts. Over the six-month duration of the project, two accounts were used to cover the required number of requests without incurring additional charges.

Although ORS could theoretically complete 100,000 routes in around three hours due to its higher batch limit, the trade-off in route accuracy and relevance made Google Maps the more appropriate choice for this analysis.

3.2.4 Route generation with the `googleroutes` Python library

To determine the typical walking routes taken to school, a custom Python package named `googleroutes` was developed. This package was built around the Google Maps Directions API and designed to streamline the large-scale generation of realistic pedestrian routes.

As an overview, the `googleroutes` package accepts a dataset containing multiple origin points (such as those placed around the perimeter of a catchment zone) and a destination point. It returns a complete set of walking routes based on user-defined parameters, exporting them in standard geospatial formats such as `.gpkg`, `.shp`, or `.geojson`. These formats allow for straightforward processing and visualisation in GIS software.

`googleroutes` is available publicly at Github, where full documentation and several example `Jupyter` notebooks are also provided. These include step-by-step guides for route generation and integration into geospatial workflows. This section outlines the development process and key considerations involved in building a functional and robust tool.

The initial use case of `googleroutes` was to serve as a simple Python module that would take a KML file containing perimeter points of a school catchment area (or any other set of origin points) and generate routes to a specified destination. The original plan was to support both the Google Maps and ORS APIs as back-end routing services. However, as described in Section3.2.3, the ORS API was eventually dropped, and the tool was built solely around Google Maps.

Using the Google Maps API, a working prototype was created that successfully generated routes from a KML file and destination point. However, this early version had

several limitations:

- It was unclear which style of route the API used, (i.e. fastest, shortest, most economical)
- The code was inefficient, relying on complex string and list manipulations
- Generated routes were incorrectly mapped to places far from South Yorkshire, for example near Seychelles or in the middle of the Atlantic Ocean

Upon reviewing the Google Routes API documentation, it was clarified that the API returns the fastest route by default [16]. Also, through various iterations of the code, the code was made more efficient through use of Python libraries such as `shapely` and `fiona`. The code now robustly handles responses from the API and converts them into usable files.

During initial research, answers were conflicting over whether the API used EPSG:4326 or EPSG:3857 as a coordinate reference system (CRS), and it was decided to use EPSG:3857, which caused the issue of incorrect route mapping. After further investigation, it was determined that the correct CRS was EPSG:4326. With this change, all known issues with the package were resolved. The `googleroutes` package now includes functionality to choose the desired CRS, allowing users to select the appropriate one for their data.

Generic functionality for using any Polygon as a collection of origins was added, although the function `generate_from_isochrones` refers to only isochrones due to its use case in this project.

For further analysis, a length field was added to the routes datasets, the code for this can be found in appendix C.1.

3.3 Combining the Routes with STATS19 Dataset

Integrating the merged STATS19 dataset to the route shapefiles was an essential step to assess the dangerousness of each route in South Yorkshire. In this step, casualties involving primary school-aged children were assigned to primary school routes, while those involving secondary school-aged children were assigned to secondary school routes. What's important to note is the fact that due to the lack of coordinate data in STATS19 prior to 1999, only the data from 1999 onwards was used in this stage.

3.3.1 Data preparation

Given the large volume of individual route shapefiles, the data preparation workflow was automated by creating a series of Python scripts, which are publicly available at GitHub.

Each file resulting from the method described in Section 3.2 contained all the generated routes for each school in South Yorkshire. The STATS19 dataset was imported into QGIS as a Delimited Text Layer, along with route shapefiles imported as Vector Layers. In order to assign the relevant casualties to individual routes, each route shapefile was

split, resulting in a group of individual route files for every catchment zone/isochrone. A unique route ID was assigned to each route to allow for referencing.

The route geometries were extremely narrow and therefore did not intersect most of the casualty points. To address this, spatial buffering of each route was a necessary step. Upon visual inspection, a 14-metre buffer was determined to be the most optimal, as it captured relevant casualty points while minimising overlap with unrelated incidents.

Finally, the Join Features by Location function was utilised to join the buffered routes to each casualty point that they intersect. This entire process was repeated separately for each borough in South Yorkshire.

3.3.2 Routes Dataset

After generating routes for each primary and secondary school in South Yorkshire, the individual route `.gpkg` (GeoPackage) files were converted to CSV and were merged into a unified routes dataset. Each route was assigned a unique route_id identifier to ensure consistent tracking across further analyses.

This combined routes dataset was then merged with the STATS19 dataset by joining on the accident_index key. This step allowed us to link each recorded collision to the specific school routes on which it occurred. The result was a final merged dataset that contained only those STATS19 collision records falling within the defined school route buffers, with each relevant collision assigned a corresponding route_id.

An important feature introduced through the routes dataset was route_length, representing the total length of each school route. This variable was critical for normalization purposes, as it allowed route-level metrics (such as total collisions or casualties) to be adjusted for route length. Without this adjustment, longer routes would be unfairly penalised in subsequent risk rankings, simply due to their size, rather than reflecting true relative danger.

One challenge encountered during this process was that a single collision could be geographically assigned to multiple route_ids. This occurred because the route buffers could overlap, resulting in multiple identical entries for the same collision, each linked to a different school route. While this duplication does not inherently compromise the data, it raises important analytical questions: should such overlapping collisions be counted separately for each route, or should they be de-duplicated to avoid inflating collision counts in overlapping areas? This consideration was carefully noted and its implications evaluated during the subsequent analytical stages.

4 Analysis

This section presents the analytical methods used to assess the safety of school walking routes across South Yorkshire. Building on the dataset prepared in Section 3, the goal was to identify the most dangerous routes, understand the factors contributing to route-level risk, and evaluate the potential for machine learning to assist in prioritising interventions.

The analysis is divided into two main parts. The first involves the development of risk metrics and route-level indicators based on collision and casualty data from STATS19. These metrics were used to classify routes into risk categories using clustering techniques. Special attention was given to correcting for biases introduced by route length and estimated usage, using a popularity-based proxy to approximate exposure.

The second part focuses on exploring the relationship between infrastructural features-such as speed limits, junction types, and lighting conditions-and route-level risk. This was achieved through a combination of correlation analysis and feature importance modelling using explainable machine learning methods. The aim was not only to rank routes but to gain insight into what conditions are most strongly associated with elevated risk.

These methods are later applied to real-world examples in the Results section, where selected case study schools are analysed to evaluate how well the risk metrics and route classifications align with observed patterns of danger and usage. This provides a practical test of their value for future risk monitoring and intervention planning.

4.1 Route Risk Identification

The first goal of the analysis was to identify which school routes posed the highest levels of risk to student pedestrians. This required defining route-level indicators that could be used to evaluate and compare risk across thousands of unique routes in a consistent and interpretable way.

Several candidate metrics were considered, each with different advantages and limitations. Early methods relied on simple counts of collisions or casualties, normalised by route length, to estimate incident density. However, these metrics failed to capture other relevant dimensions of risk, such as exposure or statistical outliers. To address this, additional metrics were developed to account for route usage and expected incident frequency. These formed the basis for a multidimensional clustering model used to classify all routes into categorical risk levels.

The following subsections outline the evolution of this risk metric framework, begin-

ning with baseline measures and progressing toward a composite, exposure-adjusted risk model suitable for classification.

4.1.1 Casualty Density as a Baseline Metric

A natural first step in evaluating route safety is to consider the number of casualties recorded along each school route. When normalised by route length, this provides a simple casualty density metric, defined as the number of casualties per kilometre. This approach is straightforward to calculate and offers an initial sense of which routes have experienced more incidents relative to their size.

However, this method comes with several limitations. Most notably, it does not account for how frequently each route is used. A busy route with multiple incidents might appear worse than it is, while a less popular route with just one serious incident could represent a much higher risk per user but remain hidden in the ranking. It also treats all casualties equally, regardless of severity, and unfairly penalises longer routes, which naturally intersect more potential conflict zones.

In addition to these conceptual drawbacks, simple normalised metrics such as collisions or casualties per kilometre also introduced technical issues:

- Small Denominator Effect: Short routes with only one or two incidents could appear disproportionately dangerous due to inflated per-kilometre rates.
- Data Skewness: Collision and casualty counts were heavily skewed, making it difficult to distinguish statistically significant outliers from random variation.
- Single-Dimension Limitations: Simple metrics failed to capture the multidimensional nature of risk, including incident severity and route exposure.

For these reasons, casualty density was treated only as a baseline indicator. More robust metrics were developed to incorporate route popularity and statistical risk ratios, allowing for more meaningful comparisons across routes of varying lengths and usage. These adjusted metrics are described in the following section.

Despite its limitations, casualty density-along with other simple normalised metrics-was used in the Results section 5.2) to provide an initial risk ranking and borough-level mapping of high-risk routes, serving as a benchmark for evaluating the effectiveness of more advanced models.

4.1.2 Defining the Commute Risk Metric

To address the limitations of casualty density and simple per-kilometre normalisation, a more comprehensive set of route-level risk metrics was developed. These were designed to capture different dimensions of risk-including raw incident frequency, relative exposure, and deviation from statistically expected behaviour-while remaining interpretable and scalable.

A total of six core metrics were selected for the final model, along with route length as a contextual reference:

- Collision Count: The total number of collisions recorded along each route.
- Casualty Count: The total number of individuals injured in those collisions.
- Collisions per km: A normalised measure of collision frequency per unit length.
- Casualties per km: A normalised measure of casualty frequency per unit length.
- Poisson Risk Ratio (Collisions): The ratio of observed to expected collisions based on overall patterns and route length.
- Poisson Risk Ratio (Casualties): The same, applied to casualty counts.

These metrics together allowed the analysis to account for both raw volume and underlying structure in the data. While raw counts tend to favour longer routes, normalised values adjust for scale but can exaggerate risk on very short routes. Poisson-based ratios provided a statistical baseline, helping to highlight routes where incident levels were meaningfully higher than expected.

Several other approaches were explored but ultimately not used in the final model. Weighted composite scores were considered, where collision and casualty metrics could be combined into a single score using manually selected weights. However, this introduced subjectivity and reduced interpretability. Rank-based scoring, where each route was ranked on individual metrics and then aggregated, also provided a single outcome but failed to capture more complex relationships between variables.

Statistical modelling approaches, such as Poisson or Negative Binomial regression, were also explored. These were useful for understanding expected incident rates and potential predictors, but were more appropriate for inferential analysis than for assigning categorical risk levels across thousands of routes.

Casualty severity was also considered as a dimension of risk, with the intention of weighting incidents based on outcome (e.g., slight, serious, or fatal). However, incorporating severity added modelling complexity and introduced further subjectivity in selecting appropriate weights. Due to time constraints and difficulty in validating the results, this component was excluded from the final analysis, but may be worth revisiting in future iterations.

In the end, the six selected metrics were used directly as input features for the clustering model described in the next section. This avoided the need to compress multidimensional risk into a single number and allowed for a more flexible, data-driven classification of route risk levels.

4.1.3 Machine Learning Approach for Risk Classification - Clustering

To identify risk patterns across school routes, an unsupervised clustering approach was used to categorise routes based on multiple commute risk indicators. Seven metrics were

selected: route length, raw collision and casualty counts, normalised values (collisions and casualties per kilometre), and Poisson-based excess risk ratios. Together, these features captured both absolute and relative safety risks, accounting for severity, frequency, and exposure.

All features were standardised using Z-score normalisation to prevent scale imbalances. Although the codebase included `DBSCAN` and `hierarchical` clustering setups, only `KMeans` was used in the final analysis due to time constraints, and its ease of interpretation. The number of clusters was determined using the elbow method (See Figure 4.1, and a value of $k = 5$ was selected based on diminishing returns in inertia.

To interpret the multidimensional clustering results, **Principal Component Analysis (PCA)** was applied. The first two components were retained, with **PC1** strongly influenced by casualty/collision density and Poisson excess risk, and **PC2** driven by raw incident counts and total route length. These components were labelled *"Collision and Casualty Density Composite"* and *"Route Length and Exposure Variability"*, respectively, as shown by their loadings in Figure 4.3.

To validate the PCA interpretation, scatterplots of route length against casualties and collisions per kilometre were plotted (Figures 4.5 and 4.6). The visual similarity between these scatterplots and the PCA distribution confirmed that the principal components successfully captured meaningful structure in the data.

Each K-Means cluster was ranked by average casualties per kilometre and assigned an internal label ranging from "Minimal Risk" to "Very High Risk." However, one cluster contained no valid members and was excluded from final visualisations. To supplement this, an **Isolation Forest** was applied for anomaly detection, identifying statistical outliers not well captured by clustering. These routes were reassigned to a new category: **Critical Risk**, subject to additional thresholds for minimum route length and incident counts.

The raw cluster and anomaly labels were then refined into a final, simplified public-friendly classification:

- **Severe Risk Route**: Routes with high density and excess risk, or those flagged as statistical anomalies.
- **High Risk Route**: Routes with consistently elevated values across multiple metrics.
- **Moderate Risk Route**: Routes slightly above average in incident frequency or density.
- **Baseline Risk Level**: Routes within expected risk levels for their length and estimated exposure.

This categorisation was visualised using Principal Component Analysis (PCA) to project routes into two dimensions for clearer interpretation (See Results Figure 5.14. The first principal component corresponded to a composite of collision and casualty density, while the second related to variation in length and exposure. Separate visualisations highlighted borough-level differences, top-risk routes, and category distributions.

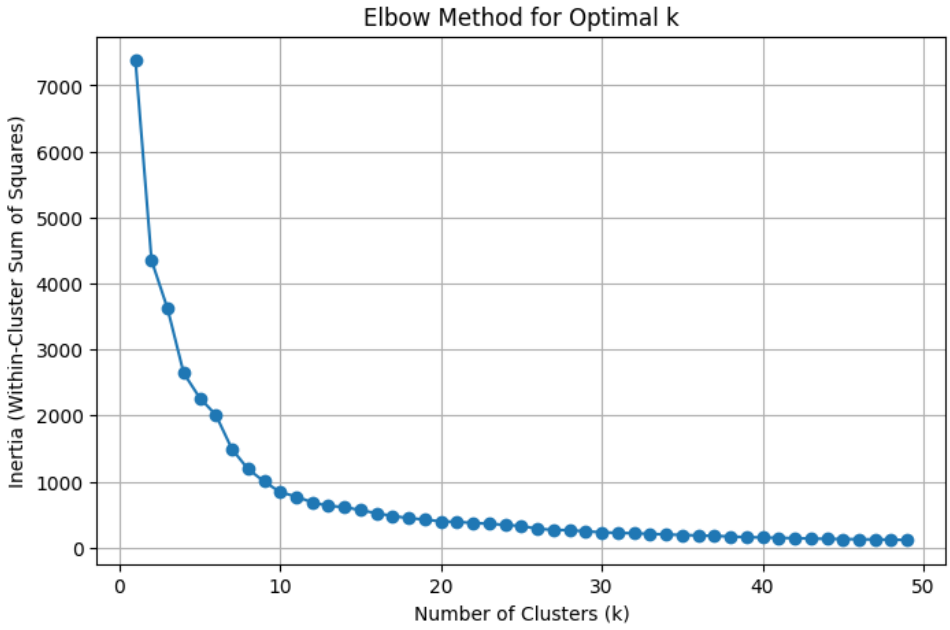


Figure 4.1: Elbow plot showing within-cluster sum of squares (inertia) for increasing values of k in `K-Means` clustering. The point of diminishing returns is observed around $k = 5$ to $k = 8$, however, $k = 5$ was selected as the optimal number of clusters due to easier interpretability of lower number of clusters.

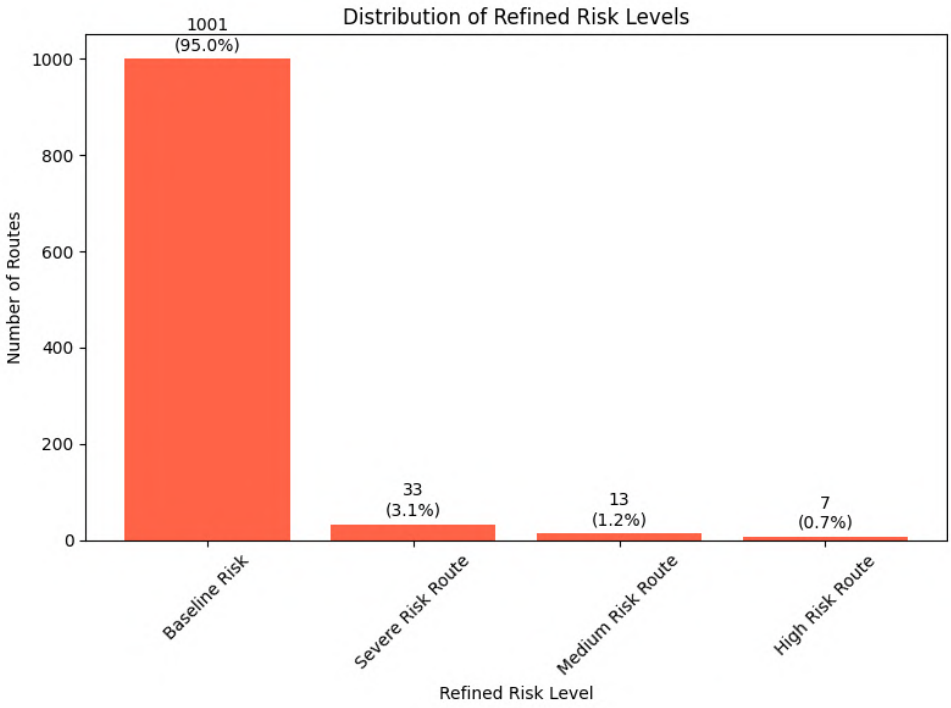


Figure 4.2: Distribution of refined risk categories across all school routes. The vast majority of routes (95%) were classified as "Baseline Risk", with only a small proportion labelled as “Severe”, “High”, or “Moderate” risk based on clustering and anomaly detection. The "Statistical Risk Concern" category did not contain any routes, and thus was not included.

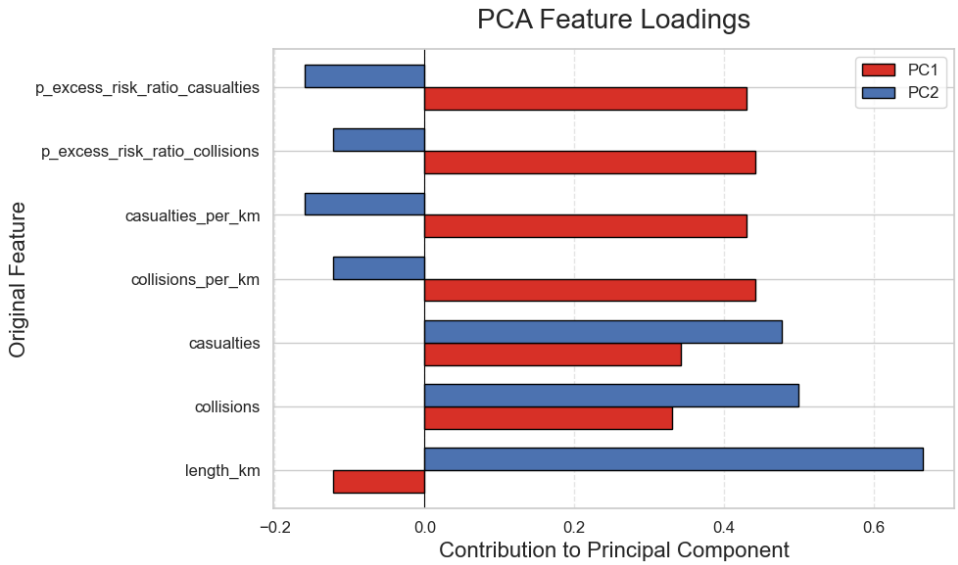


Figure 4.3: Principal Component Analysis (PCA) feature loadings showing the contribution of each risk metric to the first two principal components. PC1, labelled "Collision and Casualty Density Composite", is driven by normalised and excess risk metrics, while PC2, "Route Length and Exposure Variability", is influenced by total incidents and route length.

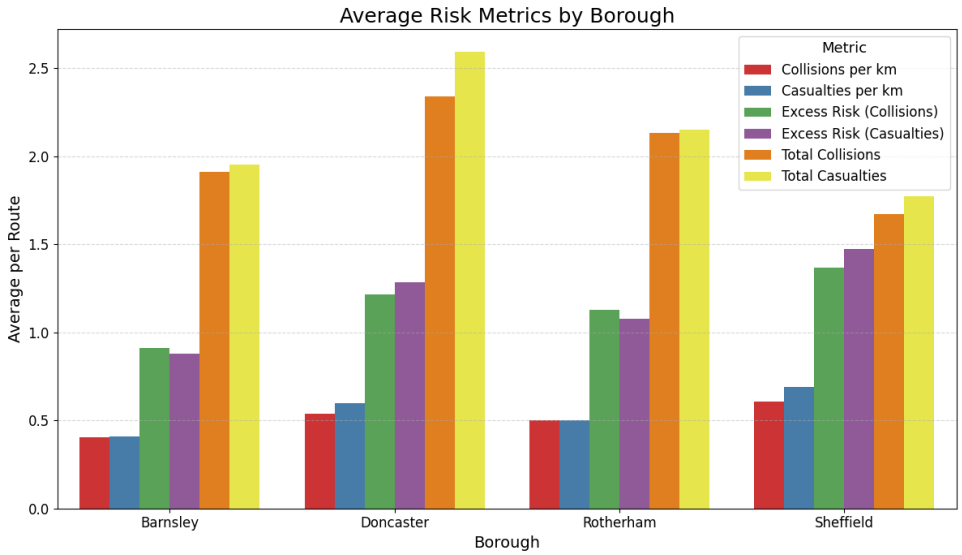


Figure 4.4: Average route-level risk metrics across South Yorkshire boroughs. Doncaster Rotherham, and Barnsley exhibit the high average excess risk-ratio of collisions and casualties per route, while Sheffield generally shows high average values for both normalised and Poisson metrics.

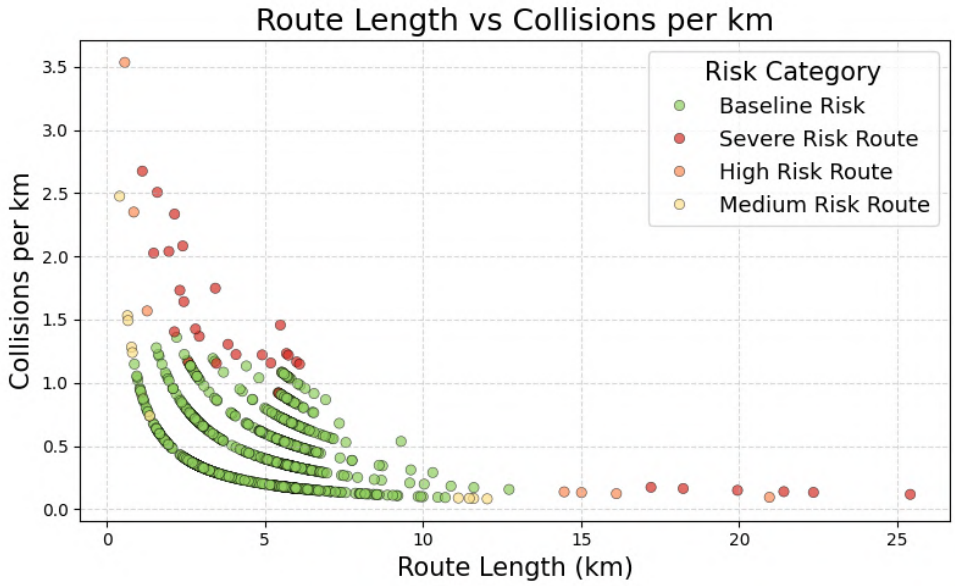


Figure 4.5: Scatterplot of route length versus collisions per kilometre. Patterns resemble those seen in the PCA projection, further validating the principal components: PC1 aligns with incident density, while PC2 reflects variation in length and scale.

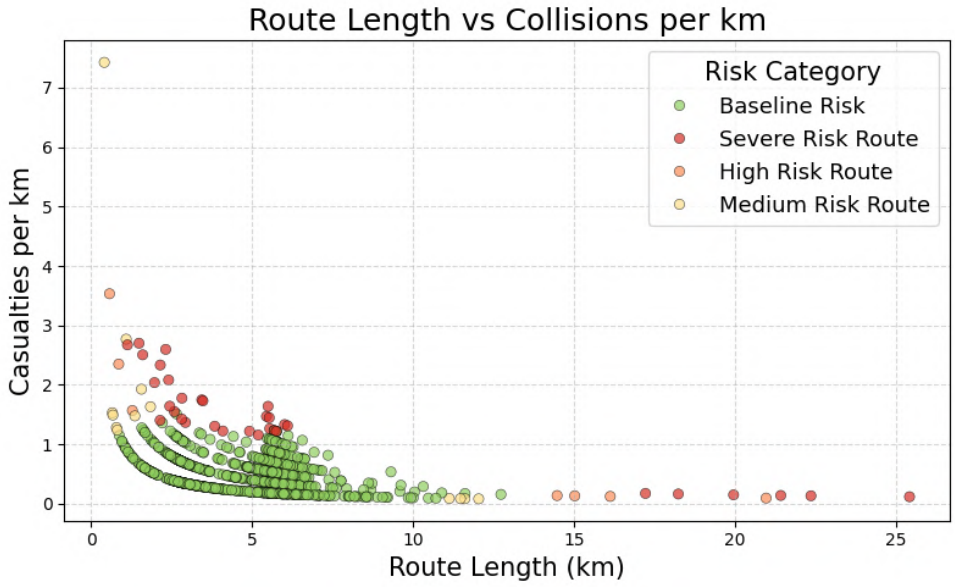


Figure 4.6: Scatterplot of route length versus casualties per kilometre, coloured by risk category. A visible gradient emerges along both axes, similar to the structure captured by the PCA projection. This supports the interpretation of PC1 as a collision/casualty density axis and PC2 as related to exposure and route length.

4.1.4 Accounting for route popularity

Current route risk metrics are useful for representing the net safety of a route, i.e. the number of collisions per km for a given time frame, but while this metric is important in its own right, it does not reveal how inherently dangerous a route is, failing to capture routes that are rarely used because of how dangerous they are. A popularity score, P, is defined to calculate this metric.

First, a buffer zone is defined around each route using the `GeoSeries.buffer` method in `geopandas` [17]. From the `geopandas` docs, "The buffer of a geometry is defined as the Minkowski sum (or difference, for negative distance) of the geometry with a circle with radius equal to the absolute value of the buffer distance.". The chosen buffer distance depends on the use case, and size of the road. To allow for the size of an average road, plus some tolerance due to the 2m resolution of EPSG:4326, a distance of 20m was chosen.

Given a real-world set of data containing journeys travelled by people, the number of intersections N_{int}, and average length of these intersections $\langle L_{int} \rangle$, with the buffers can be calculated using the `GeoSeries.geometry.intersection` method. The method returns a dataframe with the spatial intersection of the journeys with the buffers. The number of intersections is the size of the list (the number of LineStrings in the MultiLineString) and the length can be accessed with `gdf.geometry.length` (divided by N_{int} to get the average length $\langle L_{int} \rangle$.

The popularity score P is defined as the number intersection $\times$ the average proportion of by length the route that the intersection travels through

$$P = \frac{N_{int} \langle L_{int} \rangle}{L_{route}}. \tag{4.1}$$

The casualties per length per person, then, is defined as casualties per km normalised by popularity score

$$C_{lp} = \frac{N_{casualties}}{L_{route} P}. \tag{4.2}$$

Substituting P into equation 4.2 reveals that the metric does not actually depend on the length of the routes, only the distance travelled by each person on the route, but the P score can still be used as an important metric for the popularity of a route.

The problem is, however, that there are very few ways of accounting for the popularity of a route without access to large amounts of data. As a small University project, we would normally be unable to access this data, as it would be locked behind privacy regulations or large paywalls. Fortunately, at the start of the project, BetterPoints reached out and offered access to anonymised data for South Yorkshire. Unfortunately, due to time constraints and problems with anonymising the data, only data for one school was obtained. The analysis of this data and a case study on Mundella Primary School is found in Section 5.5.

4.2 Infrastructure Correlation and Feature Importance Analysis

Following the categorisation of school routes into distinct risk levels via unsupervised clustering, this analysis aimed to investigate the role of infrastructure features in contributing to these risk levels. Understanding which physical and environmental characteristics are most associated with high-risk routes can inform targeted interventions and policy decisions to improve road safety, particularly for schoolchildren.

To perform this analysis, the route risk labels obtained from the clustering analysis were merged into the combined STATS19 and routes dataset using each route's `route_id`. A refined set of nine infrastructure-related features was selected for study, including `speed_limit`, `road_type`, `junction_detail`, `junction_control`, and `pedestrian_crossing_physica` These features are categorical in nature and represent the physical or regulatory road environment present at the time and location of each collision. Records with missing values in any of the selected features were removed to ensure consistency in model input.

A Random Forest Classifier was used to model the relationship between infrastructure characteristics and the assigned route-level risk category. The model was chosen due to its ability to handle categorical inputs, robustness to noise and non-linear interactions, and its provision of built-in feature importance scores. All categorical features were label-encoded prior to model training.

To interpret the classifier and determine which features most influenced the output, SHAP (SHapley Additive Explanations) was applied. SHAP values were used to quantify each feature's marginal contribution to the model's predictions, both at a global level and for individual classes. SHAP was selected due to its theoretical robustness and ability to capture complex interactions among variables. A summary plot was generated to visualise the mean absolute impact of each feature across all predictions (Fig. 5.19), alongside the Random Forest feature importances (Fig. 5.18). Additional class-specific SHAP plots were produced to explore how different values of key features influenced assignment to particular risk levels.

5 Results

5.1 General Insights from the STATS19 dataset

To address the lack of data regarding child casualties in South Yorkshire, the STATS19 dataset was analysed to provide some general insights for this specific region. For all figures in this section, "child casualties" were defined as casualties involving children from 5 to 16 years old. To begin with, the casualty data was separated, in order to visualise how the total number of child casualties classed as pedestrian or cyclist varied in each local authority district between years 2016-2023. The number of casualties was then normalised by the school-aged child (5-16 years old) population in each region, which was obtained from the 2021 UK Census [18]. Although Sheffield has the highest total number of child casualties, Doncaster and Rotherham are noted to have the highest number of child casualties per child, which is a result visualised in Figure 5.1.

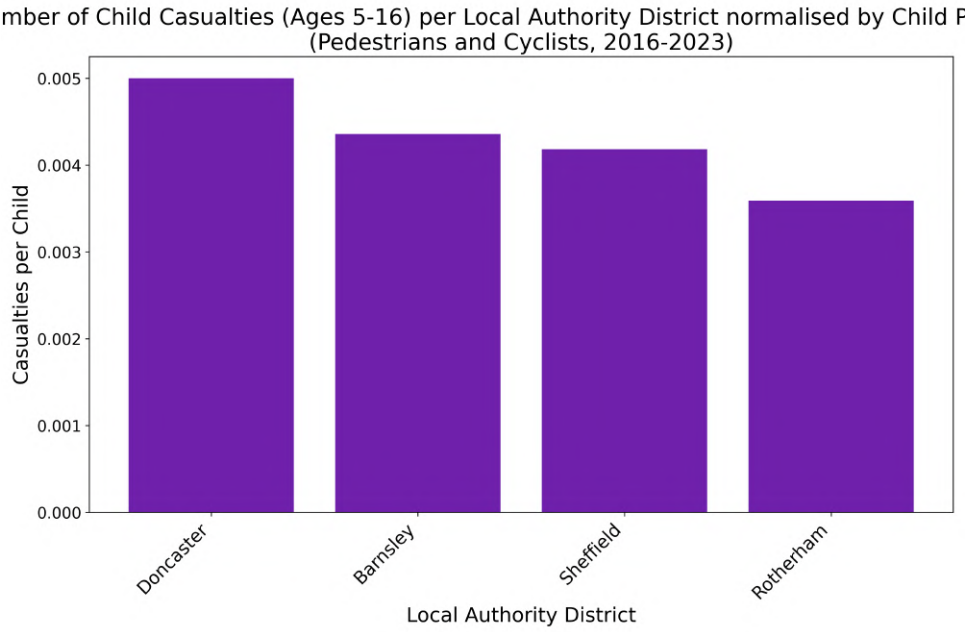


Figure 5.1: A barchart representing the number of child casualties normalised by child population (5-16 years old) in each local authority district.

Additionally, to study the trends in the STATS19 data from the most recent years (2016-2023), the total number of child casualties classed as pedestrian or cyclist in South Yorkshire was plotted for each month. Linear regression analysis was then performed on

the data, as seen in Figure 5.2.

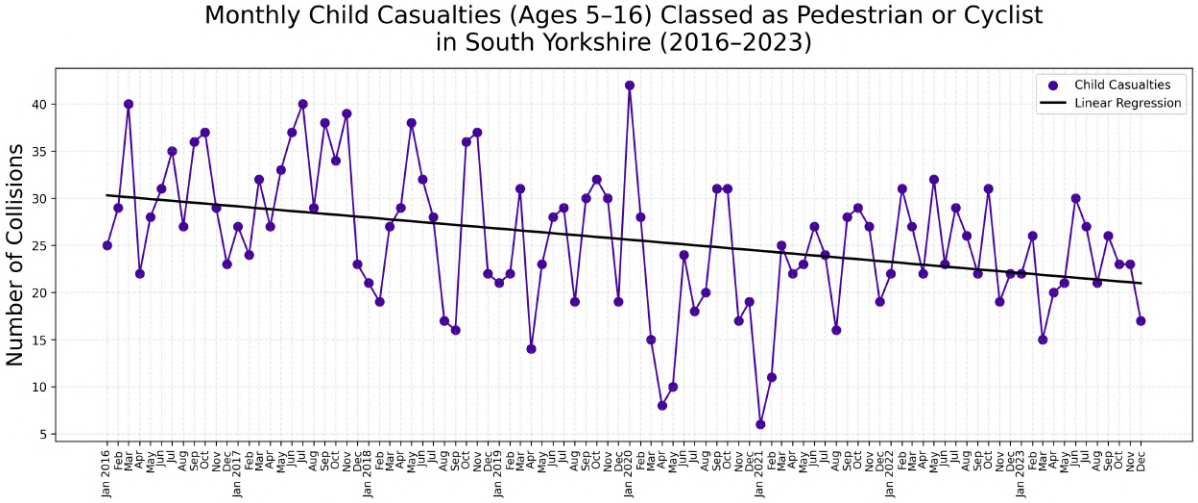


Figure 5.2: Number of child casualties classed as pedestrian or cyclist for every month between January 2016 and December 2023. Linear regression was performed on the data, which can be seen as a straight line in the graph.

From linear regression analysis, the line gradient was found to be $m = -0.0983$, indicating that on average the number of child casualties in South Yorkshire decreased by 0.0983 every month between January 2016 and December 2023. The coefficient of determination was $R^2 = 0.1404$, meaning that only approximately 14% of the variation in the number of child casualties per month can be explained by the linear trend. This is a relatively low value, suggesting that other factors affect casualty rates - not just time. Lastly, the p-value was determined to be $p = 1.69 \times 10^{-4}$, meaning that the general decrease in the number of child casualties per month between January 2016 and December 2023 is unlikely to be due to random chance.

Taking inspiration from the Department for Transport Child Casualty Factsheet from June 2015 [4], further statistics were obtained for South Yorkshire to compile a Child Casualty Factsheet specific to this region (available in Appendix A). These include the number of casualties between different ages (from 5 to 16 years old), across all available years in the STATS19 dataset, normalised by the total population of school-aged children in South Yorkshire (Figure 5.3). Linear regression was also performed on this data (not included in Figure 5.3), yielding the slope value $m = 0.001194$, coefficient of determination $R^2 = 0.7718$, and p-value 1.69×10^{-4}. The slope value shows that there is a strong upward trend in the data, meaning that casualties involving older children are more likely. Additionally, the coefficient of determination indicates that age is a strong predictor of casualty probability, as the relationship between age and the normalised number of child casualties is mostly linear. Lastly, the p-value implies that the relationship between the normalised number of child casualties and age is statistically significant, and unlikely to be a result of random chance.

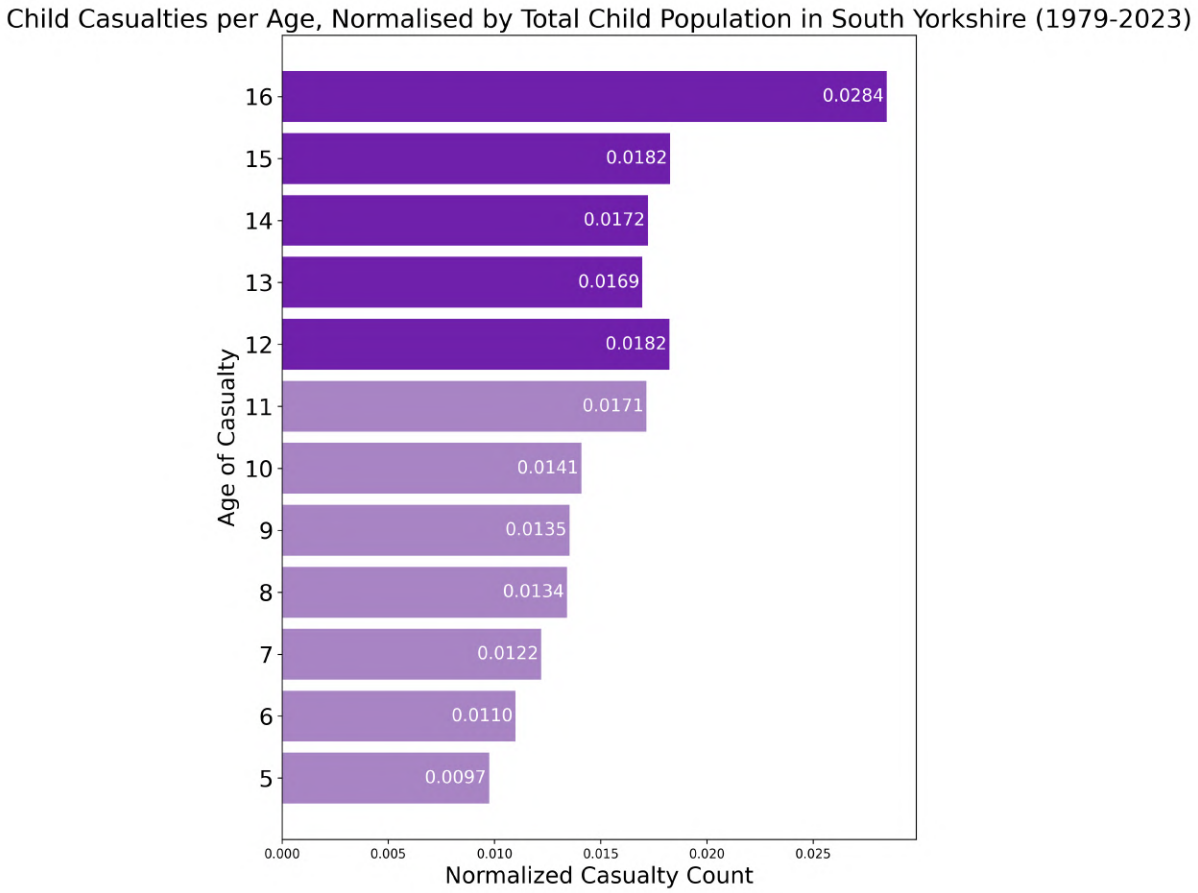


Figure 5.3: A bar chart showing the number of child casualties between ages from 5 to 16 years old, normalised by the total school-aged children population in South Yorkshire (1979-2023). The darker shade of purple represents secondary school-aged children, while the lighter shade represents primary school-aged children.

Furthermore, it was found that between the years 2021-2023, 41.9% of the child casualties were classed as passengers, 38.7% as pedestrians, and 19.5% as cyclists (Figure 5.4).

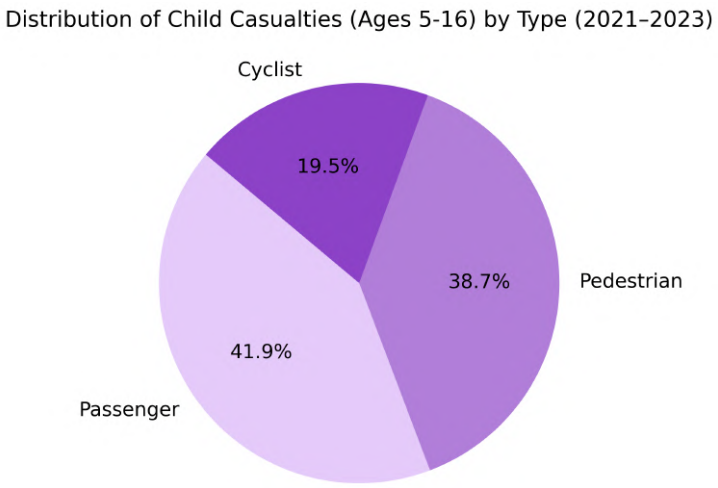


Figure 5.4: A pie chart showing the distribution of different casualty types represented as percentages of the total number of child casualties for the years 2021-2023

The full STATS19 dataset was further studied to investigate what portion of child casualties occurred when travelling to and from school. It was found that:

1. 7.14% of child casualties occurred between 8 and 8:59 am on a weekday
2. 35.85% of child casualties occurred between 3 and 6:59 pm on a weekday
3. 57.01% of child casualties occurred during any other time of the week

This indicates that while most of the child casualties occurred during other times of the week, there is a significant risk associated with commuting home from school.

Additionally, a basic geospatial analysis of the same dataset allowed to conclude that:

1. 6.07% of the child casualties occurred within a radius of 2 km from a school.
2. 23.04% of the child casualties occurred within a radius of 5 km from a school.
3. 36.48% of the child casualties occurred within a radius of 10 km from a school

Finally, it was found that on average, 17 children were slightly injured and 8 children were severely injured every month in 2023.

5.2 Casualty Density of routes to school

The initial approach to ranking school routes in terms of their commute risk was based on a basic metric - casualty density. To determine the casualty density for each route, the total number of collisions along a particular route was divided by the total length of each route in kilometres. For this ranking, the STATS19 entries from 2016-2023 were filtered to include only the child casualties, which were classed as pedestrian or cyclist.

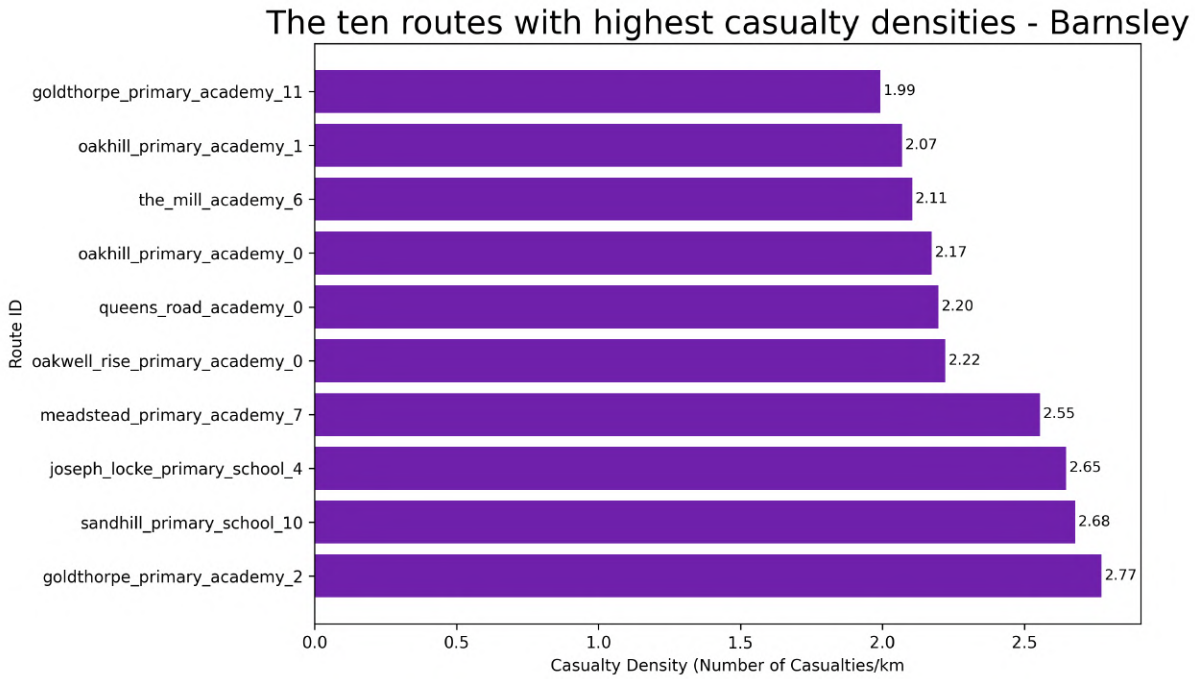


Figure 5.5: 10 routes with the highest casualty densities in Barnsley.

Overall, the mean casualty density in Barnsley was found to be 0.7480 casualties/km, while the standard deviation to be 0.4795 casualties/km. The standard deviation is approximately 64% of the mean.

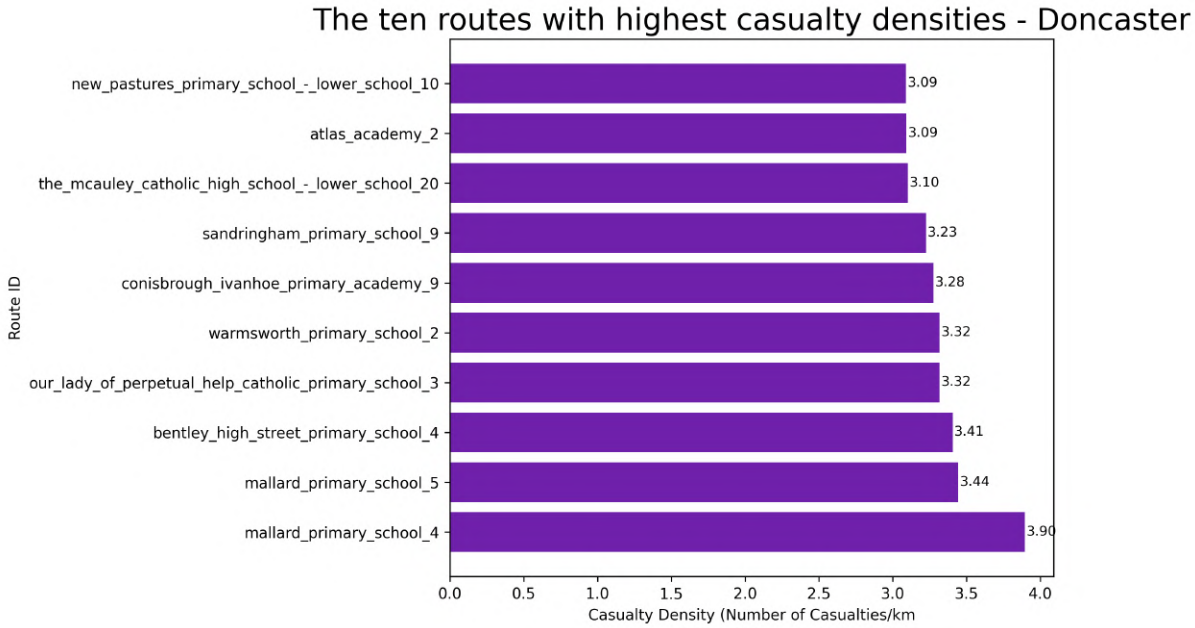


Figure 5.6: 10 routes with the highest casualty densities in Doncaster.

The mean casualty density of routes in Doncaster is 0.8881 casualties/km, with the standard deviation of 0.6722 casualties/km. The standard deviation is roughly 76% of the mean casualty density.

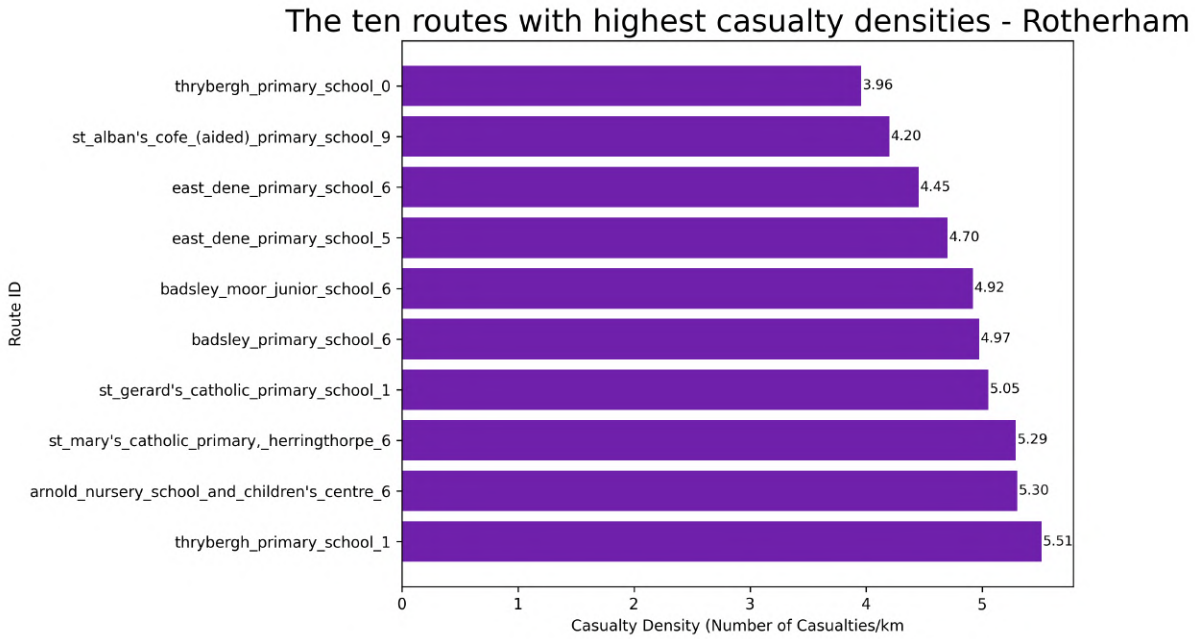


Figure 5.7: 10 routes with the highest casualty densities in Rotherham.

In Rotherham, an average route's casualty density was 0.8531 casualties/km with a standard deviation of 0.7582 casualties/km. The standard deviation is approximately 89% of the average casualty density.

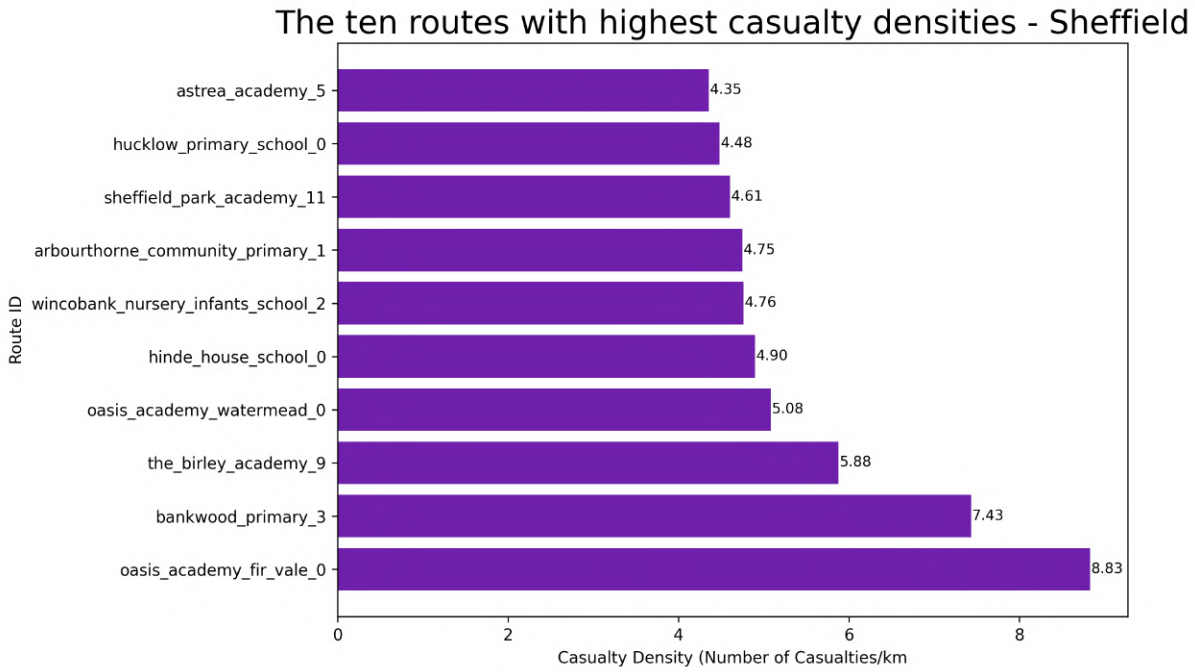


Figure 5.8: 10 routes with the highest casualty densities in Sheffield.

Finally, the routes in Sheffield had a mean casualty density of 1.0660 casualties/km and a standard deviation of 1.0694. These are the highest average and variability values in the whole of South Yorkshire. The standard deviation is 100% of the average casualty density.

Across all boroughs, the standard deviation ranges from 64% to 100% of the average casualty density, indicating a substantial degree of variability in casualty density of routes in each local authority district. From these, Barnsley was found to have the most consistent routes, while Sheffield has the most variable routes in terms of casualty density.

The above results were also mapped to provide an insight into the risk distribution in each borough.

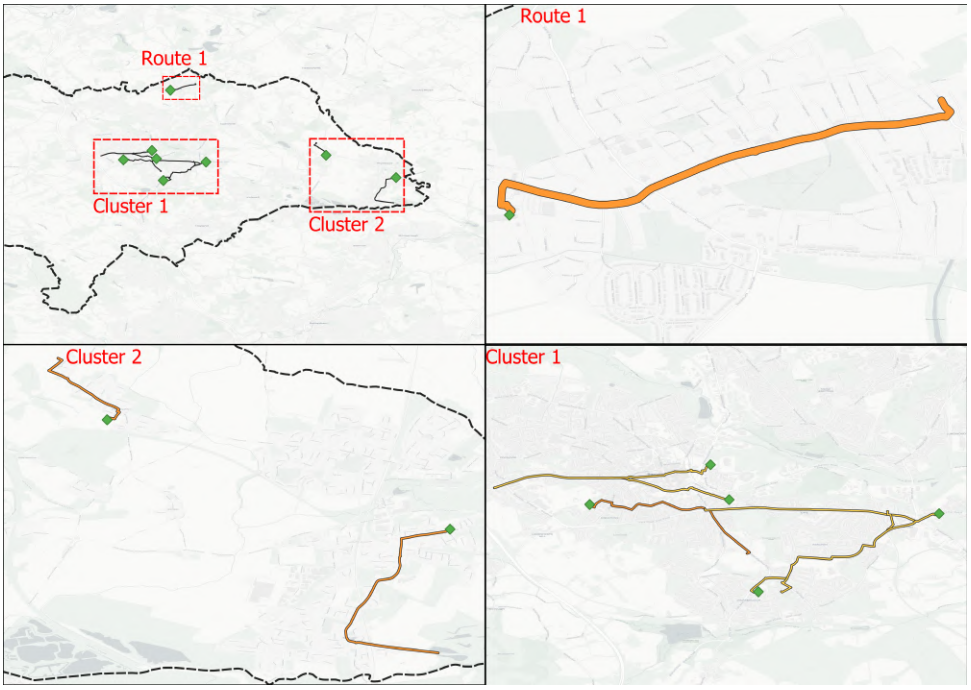


Figure 5.9: The ten routes with the highest casualty densities mapped for Barnsley. The top-left panel highlights clusters of high-ranking routes across the borough, while the remaining panels provide zoomed-in views of each cluster. The green points represent schools that these routes are associated with.

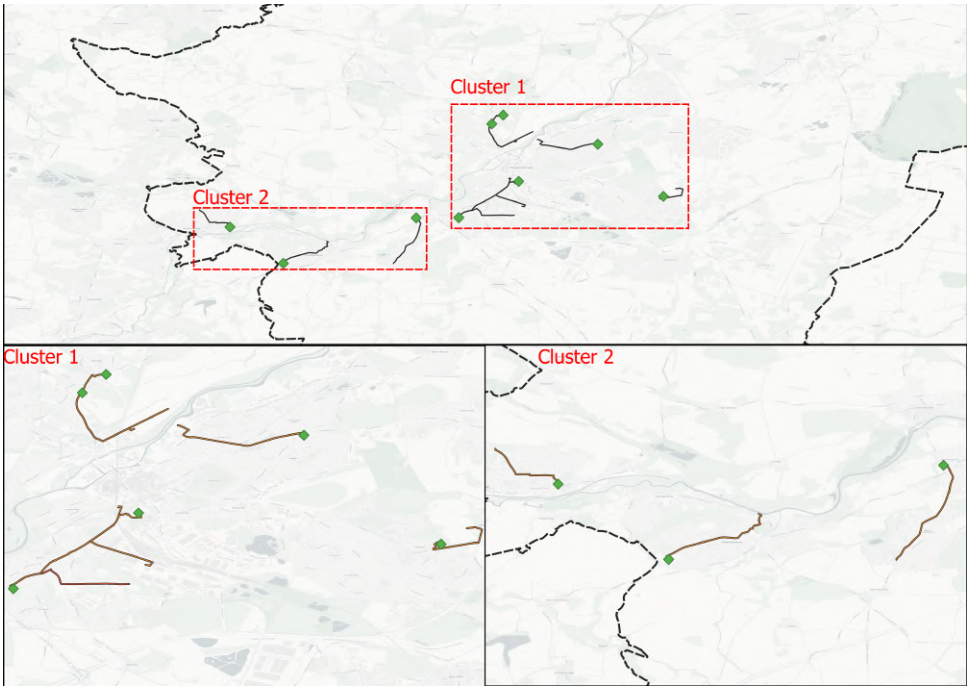


Figure 5.10: The ten routes with the highest casualty densities mapped for Doncaster. The top panel highlights the clusters of high-ranking routes across the borough, while the remaining panels provide zoomed-in views of each cluster. The green points represent schools that these routes are associated with.

Figure 5.11: The ten routes with the highest casualty densities mapped for Rotherham. The top panel highlights the cluster of high-ranking routes across the borough, while the bottom panel provides a zoomed-in view of the cluster. The green points represent schools that these routes are associated with.

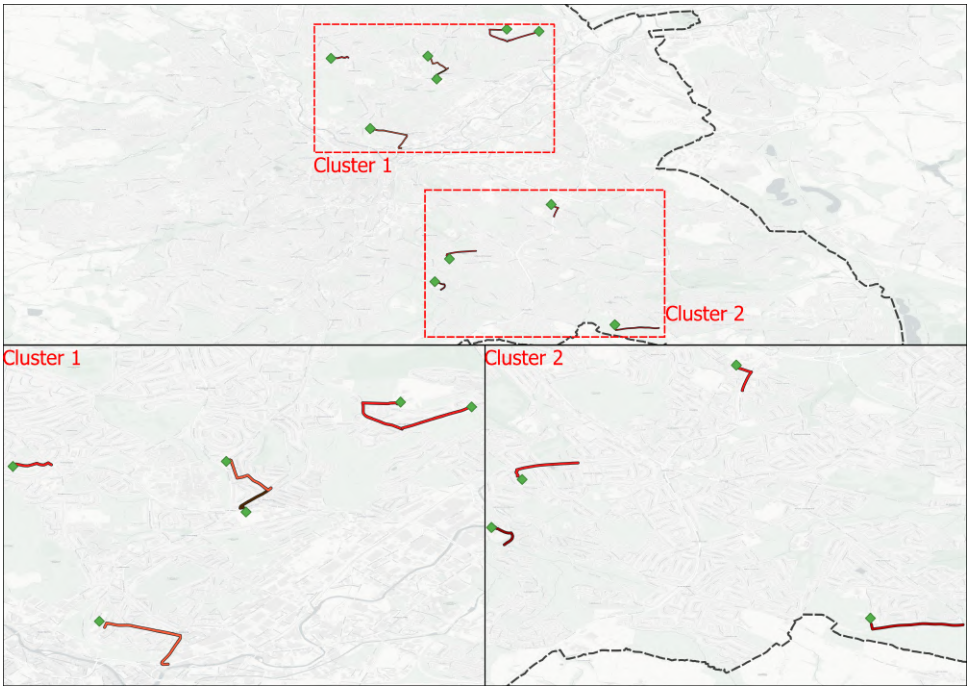


Figure 5.12: The ten routes with the highest casualty densities mapped for Sheffield. The top panel highlights the clusters of high-ranking routes across the borough, while the remaining panels provide zoomed-in views of each cluster. The green points represent schools that these routes are associated with.

Interesting insights were discovered when the maps of routes with highest casualty densities were compared to the 2019 English Deprivation Indices data. It was discovered that for all boroughs, the clusters of routes with the highest casualty densities always lie in more deprived areas. This is the case for boroughs with highly spread out routes, such as Barnsley, as well as for those with routes narrowly clustered, such as Rotherham. A visual representation of the 2019 English Deprivation Indices data can be seen in Figure 5.13. This indicates that casualty densities are closely linked to levels of deprivation.

Figure 5.13: The 2019 English Deprivation Indices data for all local authority districts in South Yorkshire is visualised via the Ministry of Housing, Communities & Local Government IoD2019 Interactive Dashboard [19]. The darker shades of blue represent higher relative deprivation. The top-left panel represents Barnsley, the top-right shows Doncaster, the bottom-left visualises Rotherham, and the bottom-right shows Sheffield.

While casualty density offers a basic indication of collision frequency normalised by route length, it undoubtedly comes with several limitations. Namely, it fails to account for the volume of users on each route. This may skew the results, as heavily trafficked urban routes may appear safer that they are, simply because of their longer lengths diluting the casualty density value, despite posing a higher overall commute risk. Following the same logic, very short routes with only a few casualties will exhibit disproportionately high casualty densities. Additionally, the metric does not distinguish between severities, ultimately treating minor and fatal casualties as equivalent. This can lead to routes with multiple fatal incidents ranking lower than those with a greater number of minor injuries.

Considering these limitations, it becomes clear that a more sophisticated metric must be developed, in order to determine the commute risk of each route more accurately.

5.3 Results from the Machine Learning approach to risk analysis of school routes

This section presents the results of applying machine learning techniques to classify school walking routes across South Yorkshire based on safety risk. The classification draws upon the commute risk metrics described in Section 4.1.2, which were designed to address limitations of traditional measures such as casualty density. These refined metrics, incorporating length-normalised incident rates and statistical risk-ratios, formed the input to a clustering model used to label each route with an interpretable risk category. This section evaluates the outputs of that process and examines the spatial and statistical distribution of risks across the South Yorkshire.

5.3.1 Risk Classification Results

Using the KMeans clustering algorithm, school routes were categorised into the following five risk levels:

1. Severe Risk Route – Routes with significantly elevated collision and casualty rates relative to expected values.
2. High Risk Route – Routes with high but not extreme levels of recorded incidents.
3. Moderate Risk Route – Routes with moderate risk indicators, often slightly above the regional baseline.
4. Statistical Risk Concern – Routes flagged by the clustering model but not exhibiting critical values (not populated in the final dataset).
5. Baseline Risk – The majority of routes, with minimal recorded incidents and typical exposure.s

Although the original clustering framework supported five categories, the "Statistical Risk Concern" category did not appear in the final dataset due to the distribution of cluster assignments of the routes.

5.3.2 Principal Component Analysis of Risk Category Clusters

To better understand the structure of route-level risk in the dataset, Principal Component Analysis (PCA) was applied to the standardised input features. This dimensionality reduction technique allows complex, multivariate data to be projected into two primary axes, PC1 and PC2, capturing the most significant variance.

The first principal component (PC1) predominantly reflects collision and casualty density, combining multiple metrics related to incidents normalised by route length. The second principal component (PC2) captures route length and exposure variability, indicating differences in spatial extent and inferred route popularity.

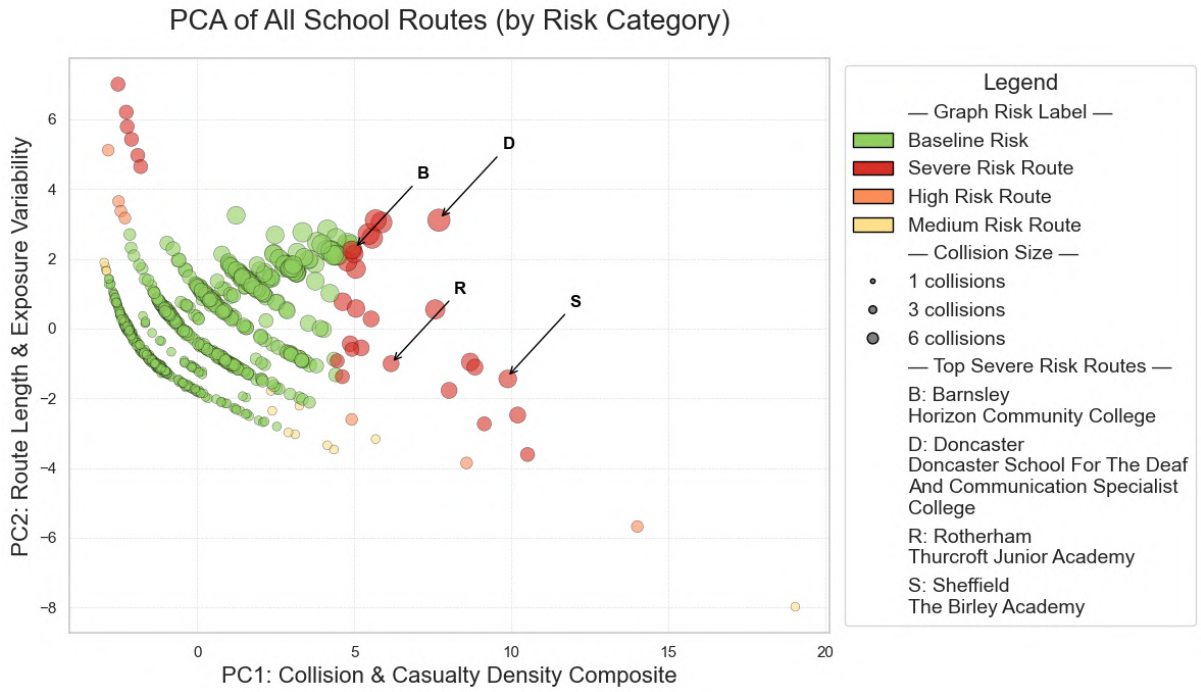


Figure 5.14: **Principal Component Analysis (PCA) of all school routes, colour-coded by risk category.** Letters B, D, R, and S mark the highest-risk routes in each borough, based on collision severity and frequency. The risk level of a school route can be interpreted to increase generally from top-left to bottom-right of the plot. It is observed that the severe risk routes (red) are concentrated in the centre of the plot, the baseline risk routes (green) concentrated all together in the centre-left, while the high (orange) and moderate risk routes (yellow) are more spread-out in the top-left and central areas.

Figure 5.14shows the full PCA projection of all routes, coloured by assigned risk category. Annotated markers indicate the highest-risk routes per borough, which cluster distinctly along PC1. The separation of risk levels along PC1 validates its interpretation as a proxy for composite incident density.

To examine regional differences in higher-risk routes, Figure 5.15 presents a PCA plot limited to routes above Baseline Risk. Borough-level clustering is evident, with Sheffield and Doncaster exhibiting a broader spread across both axes-implying greater diversity in route exposure and severity.

Finally, Figure 5.16 zooms in on the Severe Risk category, revealing how these critical routes are spread across both principal components. The plot further supports PC1 as capturing incident severity, while PC2 aligns with route size and complexity.

5.3.3 Distribution of Risk Categories in Each Borough

The clustering results were disaggregated by borough to examine regional differences in route safety. Figure 5.17 presents the distribution of risk categories within each borough. Table 5.1 also presents the number of routes in each risk category in each borough.

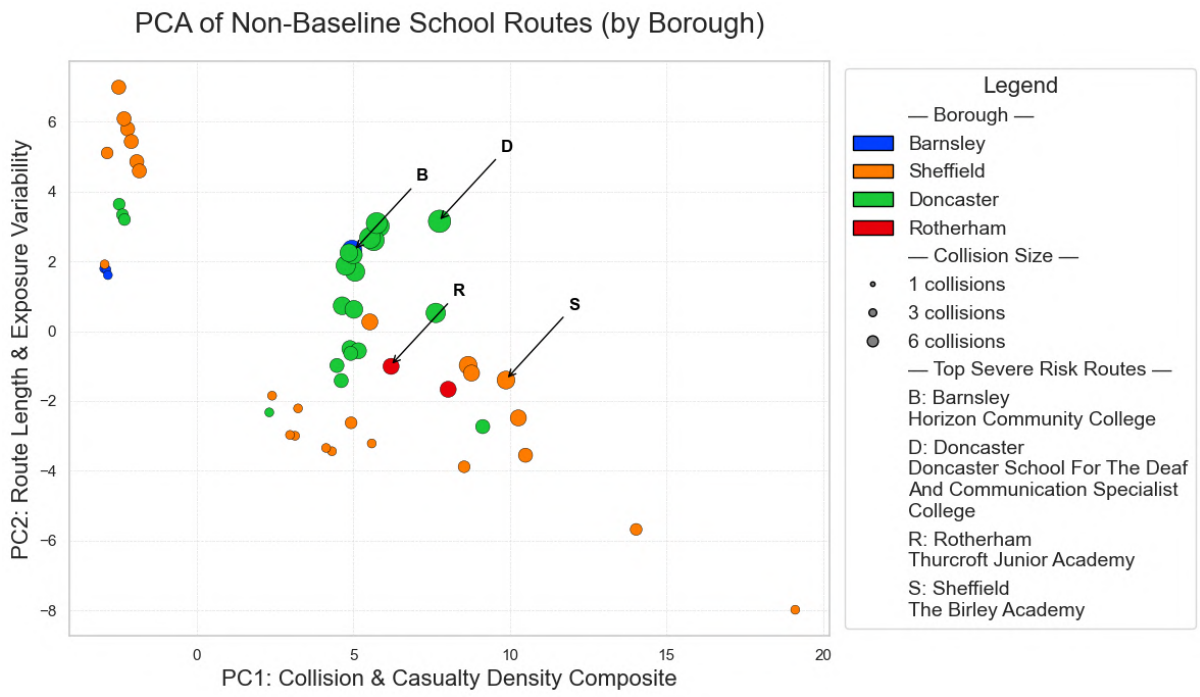


Figure 5.15: PCA projection of all school routes above Baseline Risk. Boroughs vary in the distribution of high-risk routes, and annotated markers highlight the top Severe Risk routes in each region.

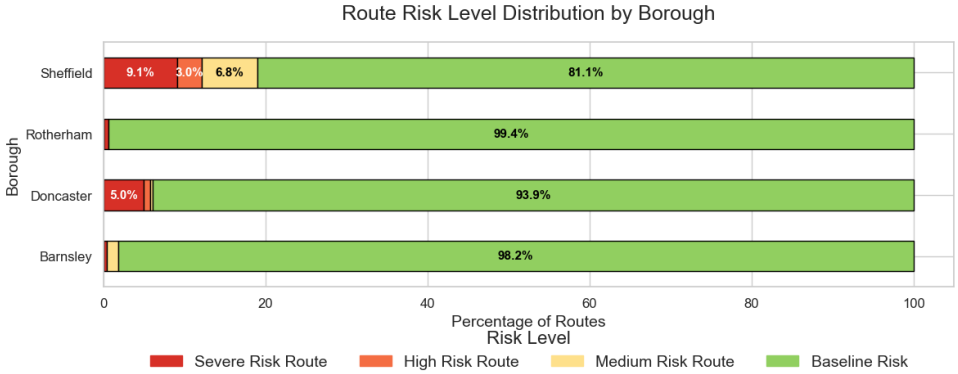


Figure 5.17: Distribution of school routes by risk category within each borough. Most boroughs have over 90% of routes classified as Baseline Risk, with Sheffield showing a noticeably higher proportion of Severe and High Risk routes, followed by Doncaster. In contrast, Rotherham and Barnsley have only a minimal share of non-baseline risk routes. *It should be noted that these figures have not been normalised by borough population. As such, observed differences may partially reflect variation in population size rather than purely infrastructure-related risk. Future work should account for population and student density to enable more accurate cross-borough comparisons.*

The distribution of risk severity reveals that Sheffield has the highest number of routes classified as Critical Risk, High Risk, and Moderate Risk. Doncaster also shows a notably elevated number of Critical Risk routes. In contrast, Rotherham and Barnsley have significantly fewer non-baseline risk routes. The prominence of Sheffield in higher risk categories may be partially attributed to its larger population size compared to other

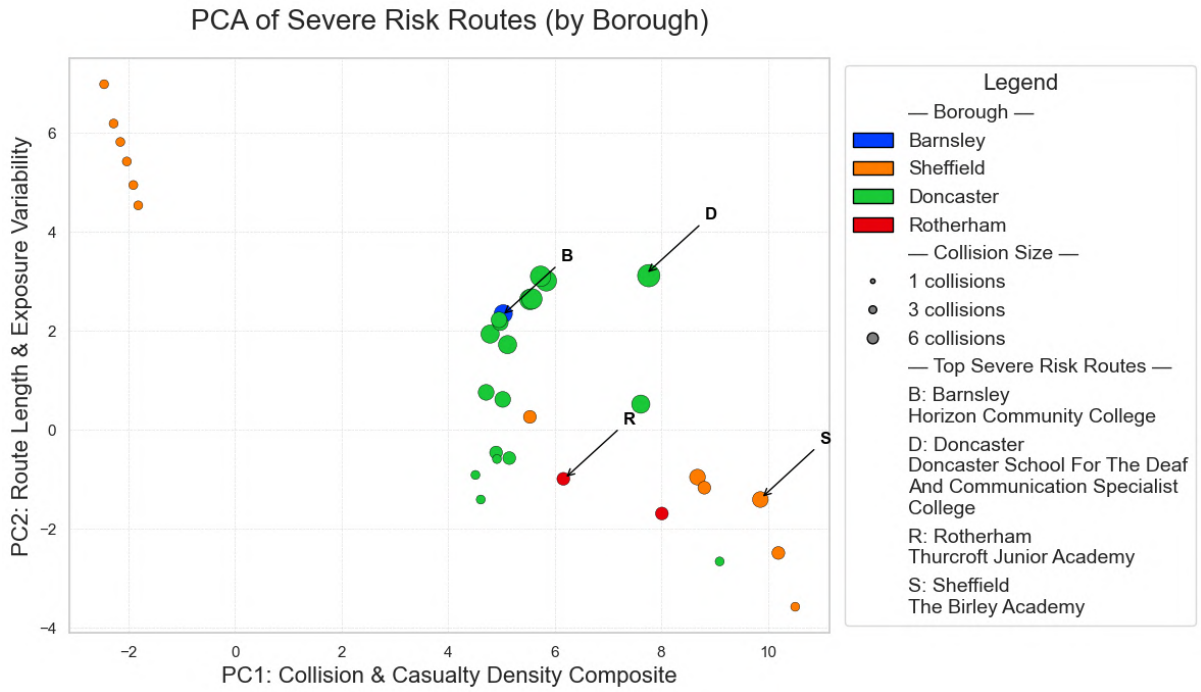


Figure 5.16: Focused PCA plot of Severe Risk Routes only. Variation along PC1 and PC2 shows differences in route severity and exposure across boroughs. Top-ranked Severe Risk routes are annotated.

boroughs [20], as increased pedestrian and vehicle interaction can naturally elevate collision risk. However, Doncaster presents an interesting case: despite having a population size similar to Rotherham [21, 22], it displays a much higher share of severe risk routes. This discrepancy suggests that beyond population, infrastructural differences, such as road layout, crossing design, or traffic management-may play a more decisive role in determining route safety.

The clustering-based risk classification was applied to 1,054 school routes across South Yorkshire, providing a refined understanding of where road safety interventions may be most needed. The large majority, 1,001 routes (94.97%), were classified as Baseline Risk, with the remaining 53 routes (5.03%) split between higher risk categories.

The final breakdown is:

	Risk Category			
Borough	Severe Risk	High Risk	Moderate Risk	Baseline Risk
Barnsley	1	0	3	223
Doncaster	18	3	1	341
Rotherham	2	0	0	330
Sheffield	12	4	9	107

Table 5.1: Percentage of routes in each risk category across South Yorkshire boroughs. Sheffield contains the highest proportion of Critical and High Risk routes, while Rotherham is dominated by Baseline Risk routes.

- Severe Risk Routes: 33 routes (3.13%)
- High Risk Routes: 7 routes (0.66%)
- Moderate Risk Routes: 13 routes (1.23%)
- Baseline Risk Routes: 1001 routes (94.97%)

These proportions are reflected in Figure 5.1, where the distribution of elevated-risk routes is shown to vary significantly across boroughs. Sheffield and Doncaster contained the highest number of non-baseline risk routes, while Barnsley and Rotherham were overwhelmingly dominated by Baseline Risk routes.

- Doncaster stood out with the highest number of Severe Risk Routes (18), followed by Sheffield (12).
- Sheffield also had the highest number of Moderate Routes (9) and High Risk Routes (4), pointing to a broader spread of risk conditions around schools.
- In contrast, Rotherham had only 2 Severe Risk Routes and no other elevated-risk categories.
- Barnsley, despite having a larger total route count than Sheffield, recorded only 1 Severe Risk Route and no High Risk routes.

5.3.4 Key Findings

1. Risk Distribution: Over 94% of school routes are Baseline Risk, but the remaining 5% show significantly elevated safety concerns.
2. Geographic Patterns: Doncaster has the most Severe Risk routes, indicating a high-priority area for interventions.
3. Meanwhile, Sheffield has the highest proportion of Severe Risk routes. It also has the widest variety of risk profiles, with substantial counts in all non-baseline categories.
4. PCA confirmed that routes with high per-kilometre risk and shorter lengths are the most variable and severe, helping to validate the classification scheme and highlight high-risk outliers.

5.3.5 Priority Routes for Intervention

The following list highlights the top 10 most critical school routes identified in the region. These were selected based on a combination of high counts, high per kilometre densities, and statistical risk-ratios of collisions and casualties, and a resulting classification in the Severe Risk category:

1. **Town Field Primary School** (Route 8): 3 collisions, 4 casualties, 2.7 casualties per kilometre.

2. **Meadowhead School Academy** (Route 10): 3 collisions, 3 casualties, 2.67 casualties per kilometre.
3. **Parkwood E-Act Academy** (Route 6): 4 collisions, 6 casualties, 2.6 casualties per kilometre.
4. **The Birley Academy** (Route 3): 4 collisions, 4 casualties, 2.51 per kilometre.
5. **The Birley Academy** (Route 3): 5 collisions, 5 casualties, 2.33 per kilometre.
6. **The Birley Academy** (Route 7): 5 collisions, 5 casualties, 2.08 per kilometre.
7. **Thurcroft Junior Academy** (Route 9): 4 collisions, 4 casualties, 2.04 casualties per kilometre.
8. **St Francis Xavier Catholic Primary School** (Route 7): 3 collisions, 4 casualties, 2.7 casualties per kilometre.
9. **Heatherwood School** (Route 20): 6 collisions, 6 casualties, 1.75 casualties per kilometre.
10. **Parkwood E-Act Academy** (Route 7): 4 collisions, 6 casualties, 1.73 casualties per kilometre.

In addition to the top 10 region-wide, the highest-risk routes are also summarised on a per-borough basis to support localised decision-making. The tables **??** present the top Severe Risk routes from each borough, ranked by number of collisions and casualties:

5.4 Results from Correlation Analysis of Infrastructure-related data in STATS19 using Random Forests

To better understand which physical road features are most strongly associated with elevated school commute risk, a Random Forest model was trained using ten key infrastructure variables extracted from the STATS19 dataset. The aim was to identify which conditions, such as speed limits, junction types, or pedestrian facilities, most influence whether a route is classified as Baseline Risk or flagged for intervention. Two interpretability techniques were used: Random Forest feature importances and SHAP (SHapley Additive Explanations), which allow for both global and class-specific insight into feature impact.

5.4.1 Feature Importances

The Random Forest output (See figure 5.18) highlights `speed_limit` as the most important predictor of risk classification by a large margin. This is consistent with prior research and policy guidance, as vehicle speed is known to be a critical factor in both collision likelihood and severity. Following this, `pedestrian_crossing_physical_facilities`, `junction_detail`, and `junction_control` were the next most influential features. These

#	School Name (Route)	Length (km)	Colli--sions	Coll--sions /km	Excess Collision Risk	Casua--lties	Casu--alty /km	Excess Casualty Risk
Barnsley								
1.	Horizon Community College (Route 8)	5.54	6	1.08	2.44	7	1.26	2.7
Doncaster								
1.	Town Field Primary School (Route 8)	1.48	3	2.03	4.57	4	2.7	5.78
2.	St Francis Xavier Catholic Primary School (Route 7)	2.81	3	1.07	2.4	5	1.78	3.8
3.	Heatherwood School (Route 20)	3.43	6	1.75	3.94	6	1.75	3.74
4.	Doncaster School For The Deaf And Communication Specialist College (Route 2)	5.49	8	1.46	3.29	9	1.64	3.51
5.	Town Field Primary School (Route 9)	2.58	3	1.16	2.62	4	1.55	3.32
6.	Don Valley Academy (Route 15)	5.44	5	0.92	2.07	8	1.47	3.15
7.	Utc (Route 21)	5.53	5	0.9	2.04	8	1.45	3.1
8.	Conisbrough Ivanhoe Primary Academy (Route 8)	2.8	4	1.43	3.22	4	1.43	3.05
9.	Conisbrough Ivanhoe Primary Academy (Route 9)	2.14	3	1.4	3.17	3	1.4	3.0
10.	Sandringham Primary School (Route 0)	2.92	4	1.37	3.09	4	1.37	2.93

Table 5.2: School route-level risk metrics across boroughs (route numbers in parentheses).

#	School Name (Route)	Length (km)	Colli--sions	Coll--sions /km	Excess Collision Risk	Casua--lties	Casu--alty /km	Excess Casualty Risk
Rotherham								
1	Thurcroft Junior Academy (Route 9)	1.96	4	2.04	4.6	4	2.04	4.36
2.	Thurcroft Junior Academy (Route 10)	2.44	4	1.64	3.7	4	1.64	3.51
Sheffield								
1.	Meadowhead School Academy (Route 10)	1.12	3	2.67	6.03	3	2.67	5.72
2.	Parkwood E-Act Academy (Route 6)	2.31	4	1.73	3.91	6	2.6	5.56
3.	The Birley Academy (Route 3)	1.6	4	2.51	5.65	4	2.51	5.36
4.	The Birley Academy (Route 1)	2.14	5	2.33	5.27	5	2.33	4.99
5.	The Birley Academy (Route 2)	2.4	5	2.08	4.7	5	2.08	4.46
6.	Parkwood E-Act Academy (Route 7)	3.47	4	1.15	2.6	6	1.73	3.7
7.	Halfway Nursery Infant School (Route 4)	17.21	3	0.17	0.39	3	0.17	0.37
8.	Halfway Nursery Infant School (Route 3)	18.23	3	0.16	0.37	3	0.16	0.35
9.	Halfway Nursery Infant School (Route 2)	19.95	3	0.15	0.34	3	0.15	0.32
10.	Halfway Nursery Infant School (Route 1)	21.42	3	0.14	0.32	3	0.14	0.3

Table 5.3: School route-level risk metrics across boroughs (route numbers in parentheses).

variables describe how pedestrian access and vehicle movement are managed, particularly at points of intersection or crossing, both of which are critical in determining safety on routes used by schoolchildren.

Less influential features included `road_surface_conditions`, `urban_or_rural_area`, and `road_type`, which contributed modestly to model predictions. Features like `carriageway_hazards`, `pedestrian_crossing_human_control`, and `special_conditions_at_site` had minimal impact, likely due to sparse reporting or weaker correlations with risk.

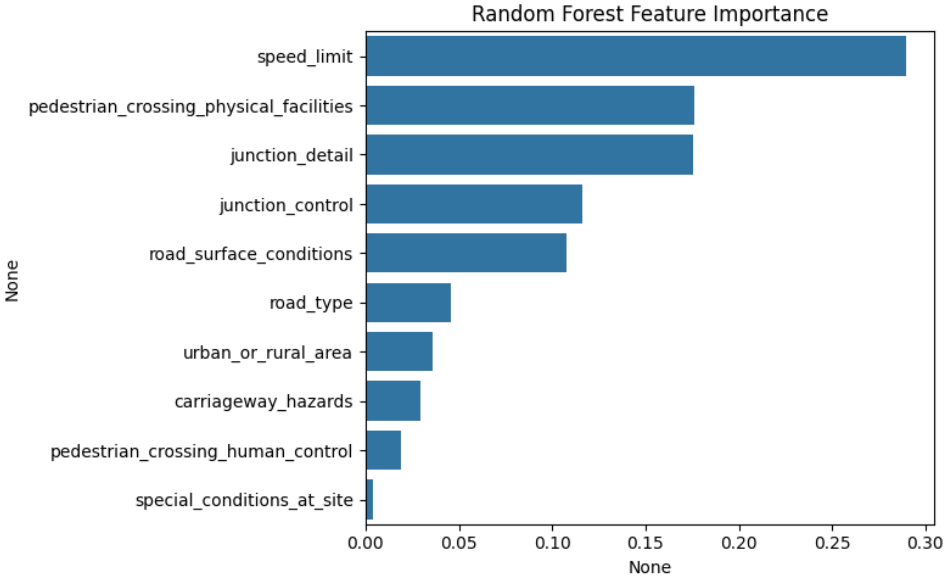


Figure 5.18: Random Forest Feature plot highlighting the most correlated infrastructural features in STATS19 with ranked routes by commute risk. It is observed the features such as speed limit, pedestrian crossing physical facilities, and junction-related features have a high correlation amongst other.s

SHAP Summary Insights

SHAP values offer a more detailed picture of how each feature contributes to predictions across different risk levels. In the SHAP summary plot (figure 5.19), speed_limit again dominates across all classes, with especially strong influence in pushing predictions toward Severe and High Risk Route classifications.

`junction_control` and `junction_detail` are also shown to play important roles in the prediction of higher-risk classes, particularly when intersections lack proper signage or have complex layouts. This supports the hypothesis that poorly managed or ambiguous junctions increase pedestrian vulnerability, especially for school children.

`pedestrian_crossing_physical_facilities` appears as another key variable, contributing to safer classifications when present. Its impact spans all classes but is particularly relevant for routes that fall into the Moderate or High-Risk categories, suggesting that the absence of these facilities may increase risk without necessarily triggering the most severe labels.

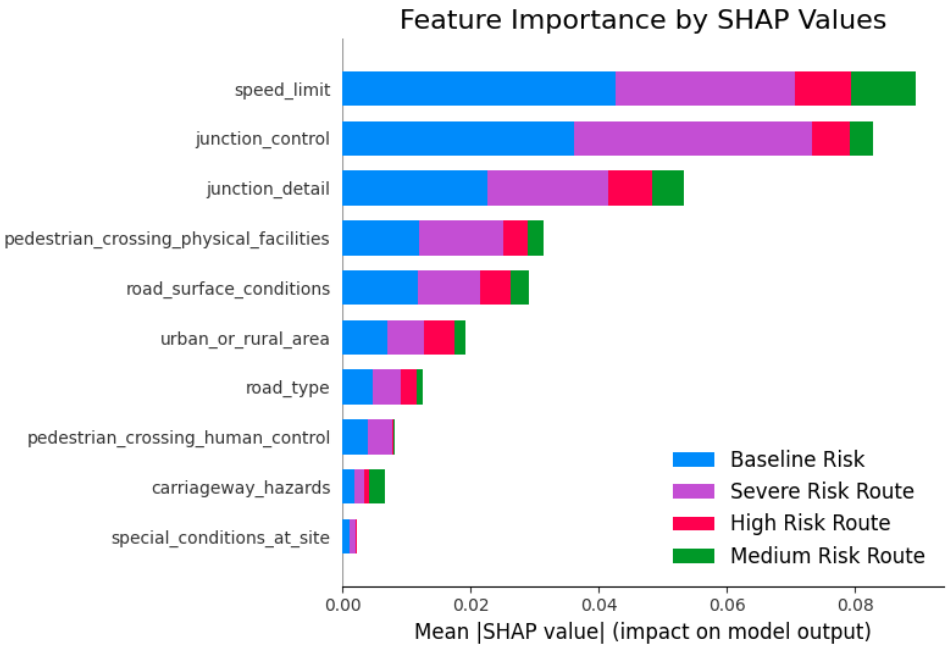


Figure 5.19: A SHAP Summary plot, showing impact of different features on different commute risk categories. It is observed that junction control has the highest correlation with severe risk routes, which speed_limit has a relatively higher correlation with high and moderate risk routes, when compared with junction control.

SHAP Impact Breakdown for Speed Limits

The third figure 5.20 breaks down SHAP impact by specific speed limit categories, showing how different limits influence predictions toward Baseline Risk (Class 0).

- Routes with 20 or 30 mph limits are associated with negative SHAP values, meaning they actively contribute to lower risk predictions. This reinforces the idea that slower speed environments are safer for school travel.
- 50 to 70 mph zones are associated with positive SHAP impacts, pushing routes away from the Baseline category. Among these, 70 mph shows the strongest effect, implying that school routes intersecting high-speed roads are systematically identified as higher risk.
- Interestingly, 40 mph zones have a slightly negative average SHAP value. This suggests that while they are not as protective as 30 mph limits, they may still offer a measurable reduction in perceived risk compared to national-speed-limit routes.

Summary of Findings

- Speed limit was the most influential predictor of route-level risk, with higher speed zones (50–70 mph) consistently associated with Severe and High Risk classifications.
- Junction-related variables, specifically `junction_control` and `junction_detail`.

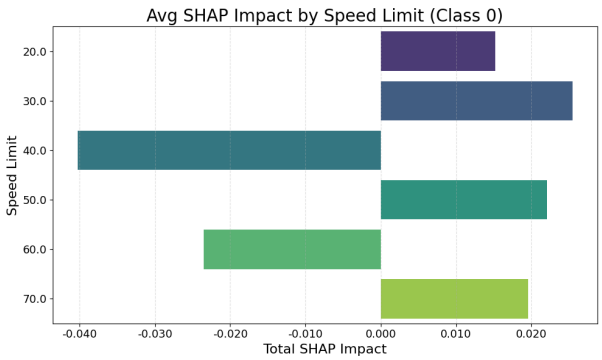


Figure 5.20: Average SHAP impact by speed limit on severe risk routes (Class 0). It is observed that while speed limits of 20, 30, 50 and 70 kmph have a positive correlation with a route being severe risk, speed limits of 40 and 60 kmph have a negative correlation with the same.

showed strong influence on model predictions, suggesting that complex or uncontrolled intersections increase route risk.

- Pedestrian infrastructure, such as the presence of physical crossing facilities, reduced predicted risk and was particularly relevant in differentiating Moderate and High Risk routes from Baseline.
- Lower speed limits (20–30 mph) were consistently associated with safer outcomes, while higher limits substantially increased modelled risk.
- Intermediate speed zones (40 mph) showed a slight protective effect compared to national-speed-limit roads, indicating a potential threshold for effective speed calming.
- Less impactful features included `road_surface_conditions`, `road_type`, and `urban_or_rural_a` which contributed marginally to predictions.
- Special case features, such as `carriageway_hazards`, `pedestrian_crossing_human_control`, and `special_conditions_at_site`, had minimal influence, likely due to sparse representation or weak correlation with risk.

These findings highlight the central role of infrastructure in shaping commute risk and suggest clear priorities for intervention. However, model-based insights must be interpreted with caution. In the following sections, case studies and correlation analyses are used to evaluate the real-world accuracy and generalisability of these findings, and to explore whether high-risk classifications correspond with on-the-ground conditions and behavioural data.

5.5 Case Study: Mundella Primary School with BetterPoints Data

1. Case studies

(a) BetterPoints data - Normalisation by Route Usage

i. If route popularity is incorporated into the danger score, do our results change?

ii. Do our routes accurately reflect routes walked by real people?

One of the aims of this project, Route Risk Identification, involves determine the most dangerous routes in South Yorkshire in 2 ways:

- Casualties per km - provides a metric for the most dangerous routes in terms of the most casualties, regardless of the popularity of the route
- Casualties per km per person - provides a metric for the most inherently dangerous routes, normalised for how busy the route is

The latter is important to capture routes that have fewer casualties because they are not popular routes to use. This could be for a variety of reasons, including the route being so dangerous that no one wants to risk using it. This metric requires a measure of foot traffic on the routes, which is difficult to obtain under normal circumstances. It is unknown whether the government collects this data, and any access would likely be tightly controlled. Therefore, the most viable way to collect this data is through crowd-sourcing. For a project of this size and due to the year-long time frame, it would not be feasible to collect enough data on our own to ensure a reliable analysis.

This previously inaccessible analysis has now been made possible with access to real-world data from BetterPoints. BetterPoints is a fitness and active travel app released in 2013. It allows users to track their journeys and collect points when using active travel methods, such as walking, cycling or taking the bus. These points can then be redeemed for rewards, such as cafe vouchers or gift cards. In 2023, BetterPoints released in Sheffield in collaboration with Sheffield City Council [23] after a trial phase at the University of Sheffield. BetterPoints allowed us access to anonymised data for trips to and from Mundella Primary School in Sheffield. Given more time, it would have been helpful to obtain data for the whole of Sheffield, however the limited data provided was used for a focused analysis of Mundella Primary School.

Data was provided as a list of anonymised LineStrings, with starting locations removed from the data. Data for all modes of transport were provided and

Comparison to Google Routes

Real-world data was helpful in determining the extent to which the routes generated through `googleroutes` matched the actual routes people use. As a crude first look, isochrone routes with a distance of 2575m were used and a visual comparison of the routes for Mundella Primary against the BetterPoints data was done and shown in Figure 5.21.

Blue lines show the generated routes and the purple lines represent the BetterPoints data, which is unfiltered, i.e. includes all modes of transport. As is visually apparent, the routes do not do a good job of representing real-world data, only capturing about $\sim 20\%$ it. Figure 5.22 shows the same routes overlaid on the BetterPoints data filtered

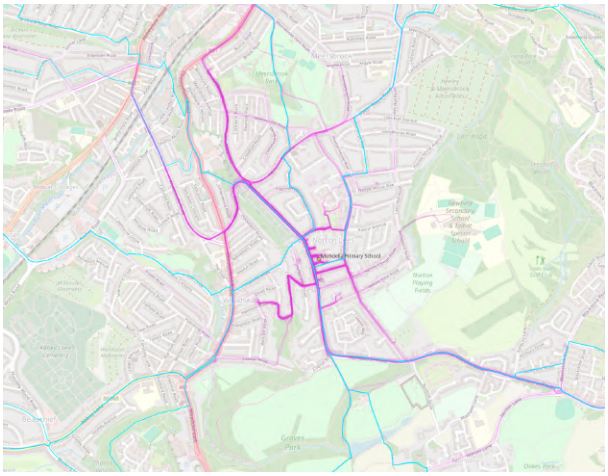


Figure 5.21: Comparison of BetterPoints data with 2575m isochrone routes for Mundella Primary School, generated with Google Maps Routes API. Routes are shown in blue and unfiltered BetterPoints data shown is in purple.

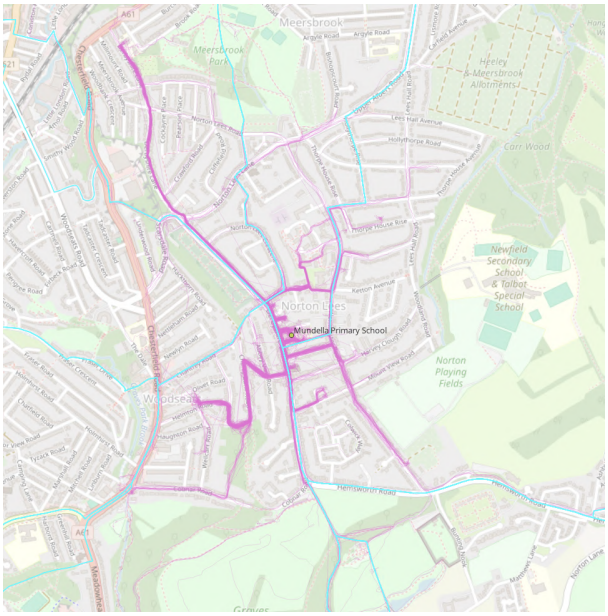


Figure 5.22: Comparison of BetterPoints data with 2575m isochrone routes for Mundella Primary School, generated with Google Maps Routes API. Routes are shown in blue and walking only BetterPoints data is shown in purple.

for walking trips only. It does a better job of capturing the data at $\sim 40\%$ captured. Both of these models fall short of capturing the small side roads around the school as there is a tendency for Google Maps to converge routes onto major roads, especially at further distances. This is a clear limitation of the route generation method, however the impact of this is reduced when using catchment zones instead of isochrones. Figure 5.23 shows the catchment zone for Mundella Primary as the black hatched region, routes generated from the perimeter of the catchment zone are shown in blue and the walking only BetterPoints data is in purple. This is a clear improvement over the isochrone method, capturing $\sim 80\%$ of the BetterPoints data that lies inside the catchment zone. Recall, however, that data for catchment zones is quite difficult to obtain, placing a clear hurdle in the way of this project and other similar projects.

A big limitation of the BetterPoints data is lack of age data. This means the equivalence between the Stats19 casualties and the data isn't perfect. Also, younger children are very unlikely to use BetterPoints, and even younger children wouldn't have a phone anyway. This likely isn't a massive issue as the journeys provided end at Mundella Primary and it is safe too assume an adult BetterPoints user's route wouldn't be massively different from a child's.

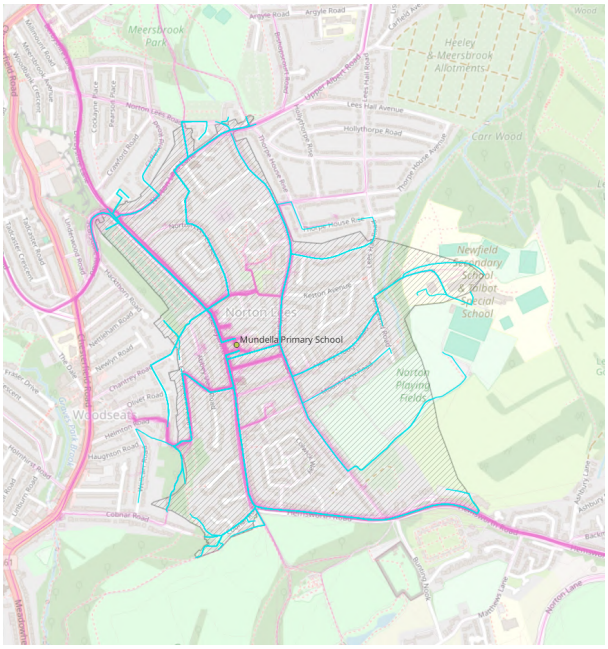


Figure 5.23: Comparison of BetterPoints data with catchment zone routes for Mundella Primary School, generated with Google Maps Routes API. Routes are shown in blue and walking only BetterPoints data is shown in purple. The Mundella Primary School catchment zone is shown as a black hatched region.

As seen in Figure 5.23, the generated routes represent the BP data quite well, especially on Hemsworth Road to the south and Derbyshire Lane to the north, going on to Chesterfield Road. The routes do however fail to capture some of the finer details, such as the smaller side roads particularly close to the school.

Normalised route risk analysis

Access to BetterPoints data allows the second metric of route risk (casualties per km per person) C_{lp} to be calculated. The popularity score and casualty metric are discussed and justified in Section 4.1.4, but, from Equation 4.2, C_{lp} was calculated for each route, filtering the casualties for pedestrian or cyclist children aged 5-16. Figure 5.24 shows the results of this analysis (left), and the same routes with the simple casualties per km metric (right). The C_{lp} analysis meaningfully changes which routes are ranked at the top, with the area to the north-east rising in the ranking significantly. The top 3 routes for each metric are shown in Table 5.4.

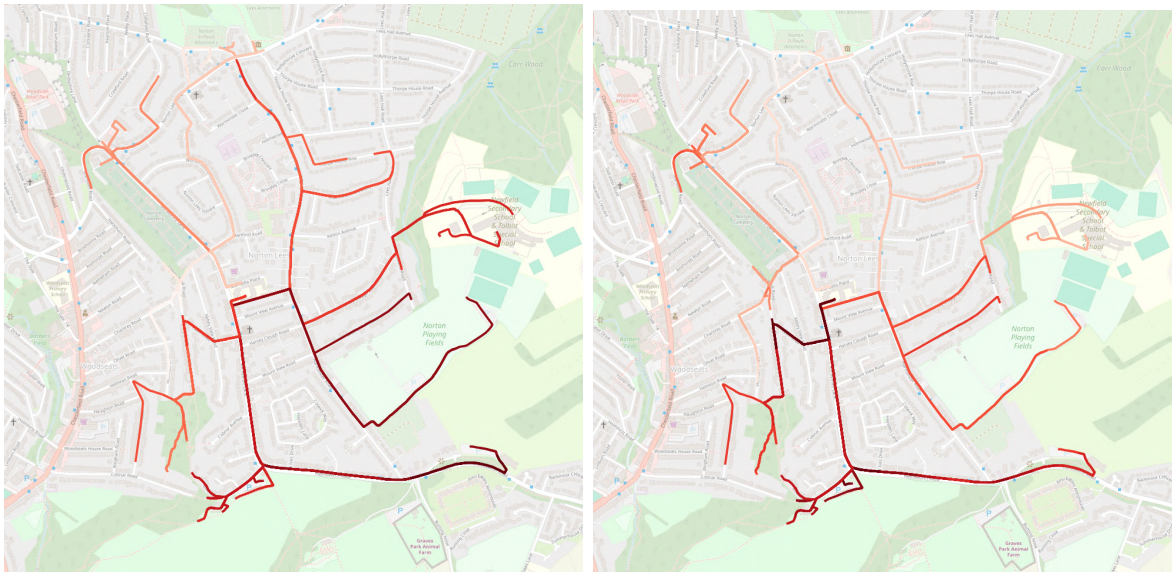
ranking	route id	Casualties per km C_l	Casualties per km per person C_{lp}
Simple metric			
1	122	34.0	(0.1134)
2	119	33.7	(0.1108)
3	118	31.6	(0.1060)
Popularity metric			
1	50	(22.9)	0.2433
2	49	(21.9)	0.2428
3	47	(20.5)	0.2427

Table 5.4: Comparison of the top 3 most dangerous routes according to the simple metric C_l (casualties per length) and the popularity metric C_{lp}. The route id is the internal id for the Mundella Primary School routes. The (bracketed) values are the values for the parameter that does not correspond to the metric the routes have been ranked by.

It is apparent from the figure that routes that spend more time on main roads are the most dangerous. This is perhaps unsurprising, however some key areas stood out when looking at the results that warranted further investigation.

Bochum Parkway

A large cluster of casualties can be seen at the intersection of Bochum Parkway and Dyche Lane, shown in figure 5.25. Although this area is not actually within the catchment area for Mundella Primary School, it stood out as both on a high-risk route and due high frequency of casualties. Bochum Parkway is a dual carriageway and forms part of the Sheffield Outer Ring Road. Despite being a dual carriageway with 2 lanes in each direction, the road has a 40mph limit, which is regularly ignored by drivers, including lorry drivers and even Police, according to some forum users [24] [25]. This is hotly debated on these internet forums, with some users saying that the speed limit should be increased, as the road is not residential and it is a dual carriageway. User WiseOwl182 says "Agreed on the first point, but not the second - what's the good reason? It's a dual carriageway, non-residential and with a central reservation. Virtually anywhere else would be 50mph minimum." [24]. Some users, on the other hand point out the non-



(a) Routes ranked by casualties per km per person, C_{lp}

(b) Routes ranked by casualties per km

Figure 5.24: Route ranking for Mundella Primary School for the popularity adjusted risk metric (left) and the simple risk metric (right). More dangerous routes are represented by darker shades of red. Routes are generated from the perimeter of the catchment zone.

controlled pedestrian crossing, which is on a public bridleway, allowing access to green areas to the south-east of the city. User thorphanger says "There are two very well used public footpaths one of which crosses over the parkway from Cinderhill Lane. There is also a lane which Farm vehicles and cars turn into and the speed vehicles come up behind you possibly not believing you are indicating to turn down this lane.". [24]

The cluster at the intersection with Dyche Lane most likely comes from drivers in excess of the speed limit turning left onto Dyche Lane from Bochum Parkway. Dyche Lane has a 30mph limit, although the sign could be missed, and the graduated turn onto the road encourages higher speeds into the turn. A view from Bochum Parkway onto Dyche Lane can be seen in Figure 5.26. The pedestrian crossing seen in the image was installed at some point between 2008-2011 according to historical images on Google Earth, (Figures 5.27a & 5.27b).

23 casualty points are selected near the intersection from the years 1979 - 2024. Of these, only two have occurred since 2011 (one in 2012 and another in 2022), or a decrease from 6.6 casualties per 10 years to 2 per 10 years. Comparing 2000 - 2011 and 2012 - 2024, there were 9 casualties in the former range and 2 in the later. Safety has clearly improved over time, and since installing a pedestrian-controlled crossing, the number of casualties has dropped significantly.

A change to the speed limit on Bochum Parkway from 40mph to 50mph was proposed in 2010, but it was rejected. Considering the location of the road next to Meadowhead School, and the various controlled and uncontrolled pedestrian crossings, this was almost certainly the correct decision, but drivers still continue to ignore the 40mph limit. A solution is unclear, but could start with stricter enforcing of the speed limits, although that could be said for almost every road in the UK where pedestrians are regularly

Figure 5.25: Cluster of casualties at the intersection of Bochum Parkway and Dyche Lane. The dark red line is a high risk route crossing the intersection. The blue dots are casualties in the Stats19 dateset filtered for pedestrians and cyclists age 5-16. 22 out of 23 of the casualties are 'slight injury', 1 is 'serious injury'.

Figure 5.26: The view from Bochum Parkway onto Dyche Lane. The 30mph speed limit sign is circled in red and the direction of travel onto Dyche Lane is represented by the red arrow. Map data: ©2025 Google

crossing.

5.6 A Further Look into the Highest Ranked School Routes by Commute Risk Metrics

With the previous sections having found the casualty density per route, created a risk classification, and used these data points in a case study with BetterPoints, the respective discoveries can be correlated to represent the highest ranking routes to school in each borough for danger level. This would be achieved by viewing some of the severe risk locations, provided in section 5.3.1 and then looking at each location using Google street view. This further allows the individual routes to be investigated and allows the suggestion of reasons why the risk values are larger in specific locations.

St Mary's and Thryberg, Horizon and Newstead

A possible initial reason for high commute risk values is that the school is near a major road or series of roads. With these roads contained within our isochrone or catchment zone, a route drawn through it would conversely record more crashes. This can be demonstrated with several of the higher ranking routes being placed near each other. The St Mary's and Thryberd primary schools both have a route (St Mary's 6th and Thryberg's 1st) that goes through the roundabout on Fitzwilliam road, and in Barnsley, Horizon community college (route 8) and Newstead academy (route 6) both share a large main road and a roundabout. It seems to be that there is a trend of dangerous routes being ones that are difficult to travel through as a pedestrian due to a larger influx of cars and less places to cross.

Bankwood Community Primary School

There were also specific locations where the road safety dangers were within both close proximity to a school and far away from any main road. Bankwood Community primary school, located in Sheffield, is very close to the nearby residential area, but the main entrance is only accessible by crossing an already thin road that additionally has cars parked on the side to make it even more difficult to fit a single vehicle in. Furthermore, the road is downhill, has little visibility for the turn, and no signs or zebra crossings (as of Google street view of September 2024). All of this serves as evidence that in areas without sufficient planning or interventions, the number of casualties will increase heavily.

Meadstead Primary Academy

Meadstead Primary Academy (Barnsley) lacks any large main roads or intersections nearby, but the 7th route moves through a long winding road on main street, with no zebra crossings, a lack of visibility, and zero speed bumps. For any child taking this route on their way to school, there is a chance a vehicle suddenly appears and a casualty occurs.

Summary

What these examples illustrate is that dangerous routes are caused by a combination of the traffic of the route itself (shown with main roads causing more casualties) and a lack of safety measures (speed bumps, zebra crossings, stop lights, pathways for alternate routes). Individual analysis is required, but this section should help illustrate the utility of using the risk values as a criteria to investigate specific school routes to find where dangers lie.

Additionally notable is the extent to which a route is considered. Analysing the school locations can show where casualties are likely to occur, but there are a multitude of routes which are dangerous because of features far away from the school. Being able to understand the journey of a school child in terms of the routes they can travel rather than the proximity to the school is an important distinction which can be used to improve child safety, but more work is needed to fully understand the link. In particular, additional BetterPoints data is instrumental in defining the link between population and casualty data, as without sufficient data it is difficult to correlate the Google routes with the number of people who are using the routes.

(a) 2008. Map data: ©2008 Google, Infoterra Ltd & Bluesky

(b) 2011. Map data: ©2011 Google

Figure 5.27: Google Earth Historical Satellite picture of the intersection of Bochum Parkway and Dyche Lane on 27 September 2008 (top) and 28 September 2011 (bottom). The pedestrian control has been put in place within this time-frame, but it is unknown when.

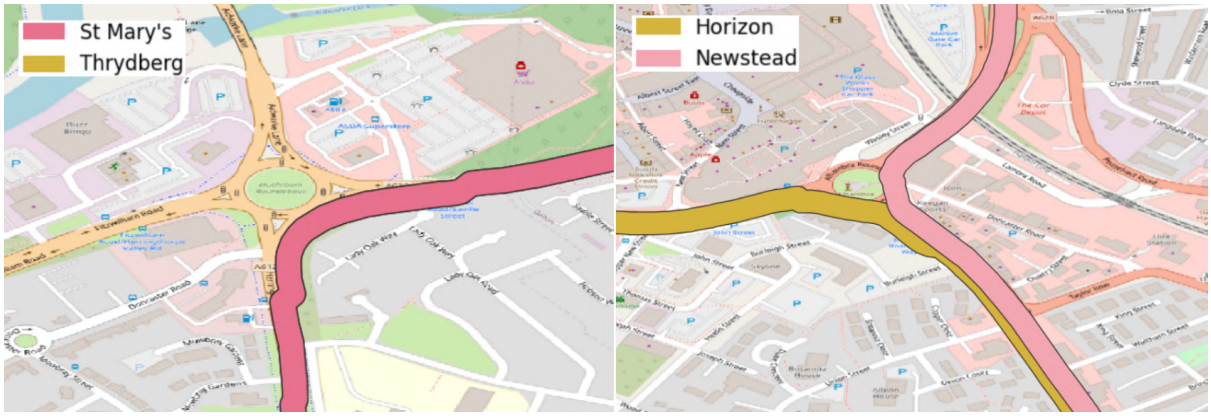


Figure 5.28: A QGIS map of 4 routes coinciding with roundabouts. Left is St Mary's and Thrydberg, right is Horizon and Newstead

Figure 5.29: Google Street View image of Bankwood community school. This road is the only one that leads to the entrance.

Figure 5.30: *A Google streetview image of Main street in Barnsley near Meadstead. There are no crossings at all until the end of this long road, where an intersection is instead.*

6 Discussion

This project successfully met its core aims 1.2 by developing a data-driven approach to identifying and understanding school route risk across South Yorkshire. It combined realistic route modelling, risk classification using clustering, and infrastructure-based machine learning analysis, with focused case studies and real-world behavioural analysis to ground the findings in real-world examples.

Route Risk Identification School routes were modelled using the Google Maps API, with start points chosen from the perimeter of each school's catchment zone. Validation against BetterPoints data showed that these routes more closely reflected actual walking patterns than routes derived from isochrones. This provided a practical way to simulate realistic school journeys in the absence of detailed pedestrian tracking for most schools.

The first approach to assessing risk focused on casualty density (casualties per km). This offered a quick snapshot of where collisions were concentrated, revealing, for instance, that routes in Rotherham 5.11 were more clustered, while Barnsley's 5.9 were more spread out. However, casualty density alone was too simplistic, it didn't consider severity or statistical context. This limitation led to the development of a more flexible, multi-metric model.

Multiple indicators, total casualty and collision counts, collisions per km, casualties per km, and excess risk scores, were standardised and used in an unsupervised clustering model. This allowed routes to be classified into four categories (Baseline, Moderate, High, Severe) based on both frequency and structure in the data. Unlike threshold-based methods, clustering accounted for outliers and uneven route lengths, giving a more robust picture of risk.

Infrastructure Feature Analysis To interpret the influence of individual infrastructure features on the route risk classifications produced by the Random Forest model 5.18, SHAP (SHapley Additive exPlanations) values were used to quantify feature impact. As shown in Figure 5.19, speed limit emerged as the most impactful feature overall, contributing significantly to the model's predictions of higher risk categories. Routes with speed limits above 20 mph, particularly those at 30 mph or higher, were associated with increased likelihood of being classified as High or Severe Risk. This trend was consistent in the SHAP impact on speed limits visualisations5.20, where higher speed limits had both a stronger and more consistent positive SHAP value for Severe Risk predictions. This finding supports established literature on pedestrian injury severity and vehicle speed, while also validating the model's internal logic.

Other features such as road type, junction detail, and presence or absence of pedestrian crossings showed lower but still meaningful SHAP impact values. For instance, A-roads and uncontrolled or roundabout-type junctions contributed to elevated risk classifications, while designated pedestrian crossings had a protective effect when present. However, limitations in STATS19 coverage of infrastructure variables mean that these effects may be underestimated. Despite this, the SHAP analysis provided a robust, interpretable breakdown of how individual environmental features influenced model outputs, offering clear pathways for targeted policy interventions—particularly around traffic speed reduction and pedestrian infrastructure near schools.

Case Studies and Real-World Context Focused case studies provided contextual validation to the analysis of modelled routes. In particular, it was found that risk significantly increased in the absence of pedestrian control measures. This indicates a critical opportunity for further analysis; with access to infrastructure location data a more comprehensive machine learning analysis could be done to assess the success and risk of these measures. A case study on Mundella Primary School using BetterPoints data highlighted high risk areas that would be missed without access to foot traffic data. It was shown that an involved analysis using this real-world data can help prioritise intervention to a wider range of urban environments and lays a foundation for further use of this data across South Yorkshire. A case study into a busy intersection near a school revealed that the introduction of pedestrian control measures decreased risk for children aged 5-16, but highlighted a problem with the enforcement of speed limits, in school areas.

Understanding risk contributory factors beyond infrastructure When visualing the spatial distribution of routes with the highest casualty densities for each local authority district, an interesting pattern became evident - all of the highly ranking routes were located in more deprived areas in each borough, indicating a positive relationship between casualty density and deprivation level. While not a primary aim of this project, it's an important insight, as it possibly highlights infrastructural inequalities present in more deprived areas in South Yorkshire.

6.1 Recommendations

Based on our analysis, we recommend the following actions:

- Annual analysis of the STATS19 data combined with geospatial data - Figure 5.2 demonstrated a statistically significant decrease in the number of child casualties between January 2016 and December 2023. To ensure a continued decrease in the number of child casualties over time, the STATS19 data should be analysed and reviewed on annual basis to inform policy makers. Additionally, geospatial analysis of the STATS19 child casualties may highlight a need for infrastructural changes in areas identified as high-risk.
- Infrastructure improvements targeted towards secondary schools: Figure 5.3 reveals a statistically significant positive relationship between the number of casualties and

age, suggesting that secondary-school aged children are more likely to be involved in casualties. This result can be explained by the increased independence and longer commute times associated with this age group, ultimately calling for more particular focus on improving the infrastructure around secondary schools.

- Time-specific traffic calming measures and restrictions: Over 40% of child casualties occurred during school commute hours, calling for direct measures to be taken to reduce traffic-related risks faced by children during these unavoidable journeys.

- Prioritise infrastructural investments in areas with high deprivation levels: Figures 5.9-5.12, as well as Figure 5.13 collectively highlight the evident correlation between increased child casualty density and the deprivation index of a given area. These results emphasise the need for special focus to be dedicated towards analysing and improving the road infrastructure in more deprived areas.

- Infrastructure assessment : Our results (see Section 5.4) found that features like junction design, crossing types, and speed control strongly influenced route risk. Case studies showed that severe-risk routes often lacked pedestrian crossings, signage, or traffic-calming features. These routes should be audited and prioritised for safety upgrades.

6.2 Limitations

It is important to acknowledge certain limitations that may have impacted the results or reduced the scope of this report. Many of these relate to difficulty in accessing certain data due to complications and delays in the approval process to acquire this data. An understanding of these limitations could be helpful for future research in this area.

In the initial meeting with Anna Butler, SYMCA, and BetterPoints, a number of datasets were offered. These included data from 17 VivaCity cameras around Sheffield and BetterPoints data for Sheffield. Both of these would have provided insight into population and route density across Sheffield, allowing a more comprehensive view of route safety, as in specific locations casualties might be low because of low foot traffic. However, despite initial agreements about access to these datasets, the VivaCity data was never provided and the BetterPoints data was delayed due to the strict anonymisation process. Fortunately, a sample of the BetterPoints data for Mundella Primary School was obtained, and even this small section of the entire dataset allowed a deep and informed analysis of the area around the school (Section 5.5).

Another significant limitation was the challenge in capturing a complete set of routes using mapping APIs. Services like the Google Maps API often ignore smaller side streets while prioritising the fastest route along main roads. As an unfunded project, our choice of API also proved to be significant limitation. A more comprehensive analysis involving a larger number routes would have been prohibitively expensive for a group of undergraduate students using the Google Maps API. The alternative, OpenRouteService (ORS) API was free but the rate limits were too restrictive for our purposes. With more time, a third alternative could be explored, or the cost of the Google Maps API could be paid if given funding.

Further constraints included the absence of data on the location of pedestrian control measures, which were requested but not delivered. The use cases of this data is discussed in the next section 6.3.

This study was also impacted by time constraints. As full-time undergraduate students, we were unable to dedicate our full attention to the project for the whole year. The first half of the year was primarily spent familiarising ourselves with a few of the software tools used that we hadn't used before, such as QGIS and the Google Maps API. A longer time frame would have allowed for more extensive testing and a larger number of focused case studies, especially for Critical Risk Routes.

6.3 Further Improvements

Our road safety analysis has provided valuable insights through the integration of STATS19 collision data, school journey routes generated via the Google Maps API, and sample BetterPoints data from one school. However, several promising opportunities exist to enhance future iterations of this work, particularly in expanding data coverage, refining analytical approaches, and improving implementation strategies.

Strengthening Data Coverage

While we successfully utilised key datasets, future work could benefit from:

- Scaling BetterPoints Integration: With data currently available from only one school, behavioural insights remain limited. Expanding BetterPoints deployment across more schools would enable borough-wide comparisons and deeper understanding of mobility and exposure patterns.
- Maintenance and Intervention Records: Access to road maintenance and intervention logs-such as resurfacing, new crossings, or signage updates-would support temporal analyses of safety before and after improvements.
- Detailed Infrastructure Layers: Augmenting STATS19 with geospatial data on crossings, speed bumps, traffic calming measures, and cycling infrastructure would significantly enhance our correlation models and help isolate high-impact safety features.
- Vivacity Sensors and Usage Data: Deploying pedestrian and cycle sensors at key junctions would provide real-time, non-school-specific usage data to complement BetterPoints and fill behavioural gaps.
- Weather and Seasonal Data: Incorporating meteorological data could reveal how adverse conditions affect safety outcomes and travel behaviours throughout the year.

Expanding Analytical Approaches

Our machine learning models have enabled clustering of route risk and infrastructure correlation analysis. Future directions include:

- School vs. General Route Comparison: Comparing school journey routes to general traffic corridors would help identify whether school travel faces unique risks or reflects broader road safety issues.
- Longitudinal Impact Studies: With time-stamped maintenance and intervention data, we could evaluate the effectiveness of specific measures through before-and-after or time-series analysis.
- Predictive Modelling: Expanding ML applications to forecast high-risk areas could enable proactive interventions before incidents occur.
- Perception vs. Reality Gap Analysis: Combining objective safety data with community perception surveys could help identify mismatches and guide targeted education or infrastructure efforts.
- Equity-Focused Insights: Overlaying IMD (Index of Multiple Deprivation) data with risk maps could reveal whether infrastructure improvements are equitably distributed across communities.

Engagement and Implementation Pathways

To maximise the impact of our findings, we suggest:

- Community Feedback Mechanisms: Enabling residents to report near-misses or safety concerns would supplement formal data and help identify under-reported risks.
- Multi-Agency Data Sharing: Coordinating with emergency services, healthcare providers, and transport agencies would create a richer, multidimensional view of road incidents and outcomes.
- Evidence-Based Intervention Tracking: Systematically documenting interventions and their outcomes would build an evidence base for identifying the most effective road safety measures. By pursuing these improvements , we can evolve this analysis into a powerful decision-support tool for ensuring safer journeys for schoolchildren and the wider community.

7 Conclusion

The analyses presented in this report offer an objective, data-driven framework for identifying and prioritising high-risk school routes for targeted safety interventions. By focusing on routes with the most critical risk characteristics, limited resources can be allocated more strategically to maximise safety outcomes. The framework developed here can be consistently applied across South Yorkshire and adapted for use in other regions.

A range of analytical tools and techniques were utilised throughout the project, including QGIS, Python, and machine learning. Each one played an instrumental role in producing and interpreting informative data. Through continuous refinement, these methods were consistently improved to ensure that raw data can be effectively transformed into meaningful outputs. These methods allowed meaningful insights into casualties across South Yorkshire to be extracted to produce actionable outcomes. It was found that although Sheffield had the highest recorded number of child casualties, Doncaster had highest child casualty rates when normalised by child population. Linear regression analysis of monthly child casualties from January 2016 and December 2023 revealed a statistically significant downward trend over time, which is most likely a results of additional, external factors. Linear regression analysis was also performed on the number of casualties per age, normalised by total child population in South Yorkshire, indicating that secondary school-aged children are more likely to be involved in casualties. Furthermore, between 1979 and 2023, over 40% of child casualties occurred during hours and that a significant proportion of them occurred within a radius of 10 km from a school.

Investigation into casualty densities for routes in each borough revealed significant variations, with the routes in Sheffield demonstrating the highest mean and variability. Mapping the highest casualty density routes in each local authority district revealed that they were consistently located in more deprived areas.

Real-world data from BetterPoints was compared against routes generated by the `googleroutes` library to determine their accuracy. It was found that routes generated from catchment zones were significantly more accurate that those generated from isochrones around the school. These routes tend to favour major roads, often ignoring smaller side roads near the school. It was found that integrating BetterPoints data into the analysis helped to reveal routes that don't have a large number of casualties, but do have a large number in proportion to foot traffic on that route. However, more data is required to create any significant results. The data drew attention to Bochum Parkway, where the 40Mph speed limit is frequently ignored despite the numerous pedestrian crossings on the road. It was found that since putting in a new pedestrian crossing at the intersection of Bochum Parkway and Dyche Lane, casualties of pedestrians and cyclists aged 5-16 have dropped significantly. In a case study on Critical Risk routes, locations

deprived of pedestrian safety measures - crossings, speed control measures, suitable footpaths - were identified with recommendations made to improve safety in these areas.

The discussion section 6 explored the implications of this project, with key recommendations for future planning. For example annual model iteration, investment in infrastructure and targeted interventions in deprived areas were identified as important next steps. The limitations of the study were also acknowledged, including difficulty in generating a representative routes dataset, and API and data restrictions. These limitations offer clear pathways for future improvement, and this project has laid the foundations for this work.

Ultimately, annual application of the methodology developed in this project can be used to monitor changes in risk patterns and evaluate the effectiveness of interventions, contributing to a comprehensive, long-term approach to improving student safety. Further studies to increase the scope of this project to other vulnerable populations and urban environments would contribute to a more comprehensive understanding of road safety, in and outside of school areas. By building on this foundation, decision makers and urban planners can work towards safer and more equitable school journeys for children in South Yorkshire and beyond.

Bibliography

[1]M. o. S. Y. Coppard Oliver 1, *My plan for south yorkshire*, English, 2024.

[2]Y. S. Foundation, *South Yorkshire's Mayor launches 2025 Walk and Wheel Challenge*, en.

[3]National Travel Survey, *National Travel Survey 2022: Travel to and from school*, en.

[4]Department for Transport, *Facts on child casualties*, en.

[5]M. S. Zeedyk et al., "Children and road safety: Increasing knowledge does not improve behaviour", en, British Journal of Educational Psychology **71**, 573–594 (2001).

[6]H. Ward et al., "Reporting of Road Traffic Accidents in London: Matching Police STATS19 Data with Hospital Accident and Emergency Department Data", en, (2002).

[7]P. Marchant, J. D. Hale, and J. P. Sadler, "Does changing to brighter road lighting improve road safety? Multilevel longitudinal analysis of road traffic collision frequency during the relighting of a UK city", English, Journal of epidemiology and community health, Num Pages: 467-472 Publisher: BMJ Publishing Group LTD Section: Original research, 467–472 (2020).

[8]Department for Transport, *Road Safety Data*, en, Jan. 2025.

[9]Sheffield City Council, *Sheffield_ schools_ list_ january_ 2023_ 5.pdf*, en, 2023.

[10]Sheffield City Council, *Primary Catchment Boundaries 2024 - 2025*, en-gb.

[11]Sheffield City Council, *Secondary Catchment Boundaries 2024 - 2025*, en-gb.

[12]Betterpoints, *BetterPoints Ltd – Behaviour change technology.*

[13]QGIS Development Team, *Qgis geographic information system*, QGIS Association (2025).

[14]O. for National Statistics, *Police Force Areas (December 2024) Names and Codes in the UK - ArcGIS Hub Dataset - data.gov.uk*, en.

[15]Department for Transport, *National Travel Survey: 2014*, en.

[16]Google, *Google maps routes api documentation*, 2025.

[17]J. V. den Bossche et al., *Geopandas/geopandas: v1.0.1*, version v1.0.1, July 2024.

[18]nomis, *Nomis - 2021 Census Area Profile - Yorkshire and The Humber Region*, 2025.

[19]Ministry of Housing Communities and Local Government, *Power bi dashboard*, https://app.powerbi.com/view?r=eyJrIjoiOTdjYzIyNTMtMTcxNi00YmQ2LWI1YzgtMTUyYzMxOWQ3NzQ2Iiw Accessed: May 2025.

[20]Sheffield City council, *Census and Population | Sheffield City Council*, 2021.

[21]Office for National Statistics, *How the population changed in Doncaster, Census 2021 - ONS*, en, 2021.

[22]Office for National Statistics, *How the population changed in Rotherham, Census 2021 - ONS*, en, 2021.

[23]Sheffield City Council, *Betterpoints sheffield launches to inspire healthier journeys and support local businesses*, Sept. 2023.

[24]Freebooter and Various, *Sheffieldforum.co.uk: bochum parkway speed limit*, Feb. 2019.

[25]StewySphinx65, *Reddit/r/sheffield: bochum parkway speed limit*, 2021.

[26]S. Gillies, R. Buffat, J. Arnott, et al., *Fiona*, version 1.10.1, 2024.

[27]Filipe et al., *Python-visualization/folium: v0.18.0*, version v0.18.0, Oct. 2024.

[28]T. Clifford, B. Morgan, C. Broadfoot, et al., *Google-maps-services-python*, version 4.10.0, 2024.

[29]C. R. Harris et al., "Array programming with NumPy", Nature **585**, 357–362 (2020).

[30]T. pandas development team, *Pandas-dev/pandas: pandas*, version v2.2.3, Sept. 2024.

[31]J. Lawhead et al., *Pyshp*, version 2.3.1, 2024.

[32]T. scikit-learn developers, *Scikit-learn*, version 1.5.2, Sept. 2024.

[33]S. Gillies et al., *Shapely*, version 2.0.6, Aug. 2024.

[34]D. Porges, *Googleroutes*, version 1.0.0, 2025.

A Supplementary Figures

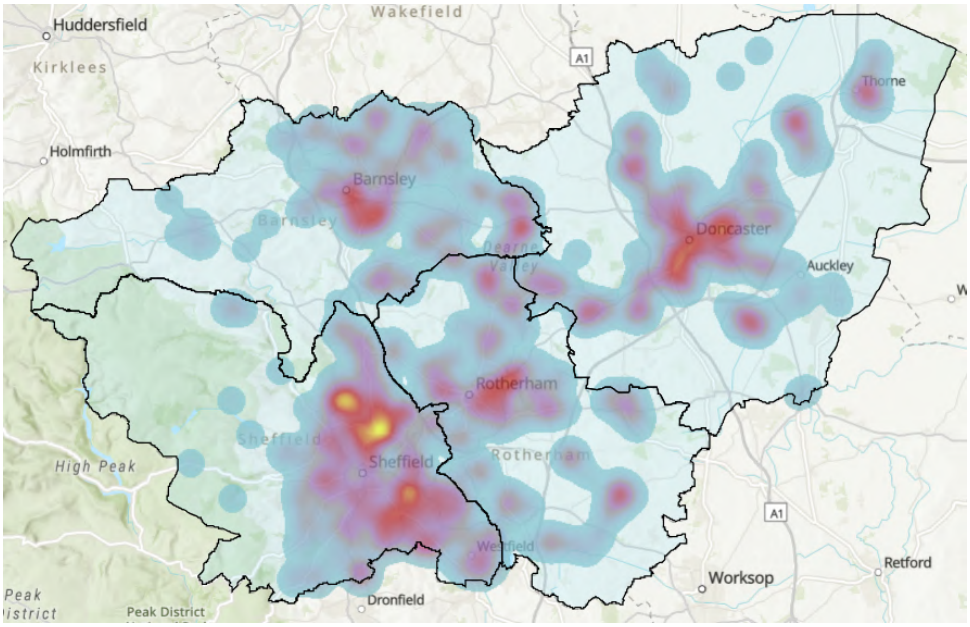
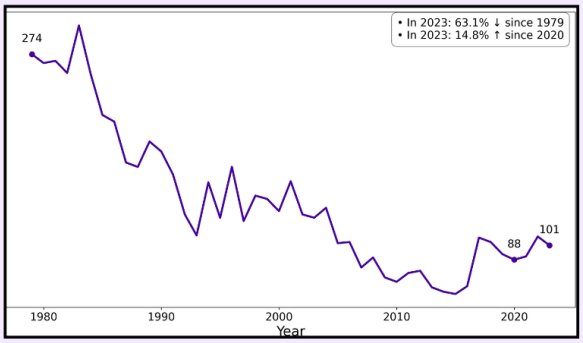


Figure A.1: Heatmap of Child Casualties (Ages 5-16) classed as pedestrian or cyclist in South Yorkshire (2016-2023)

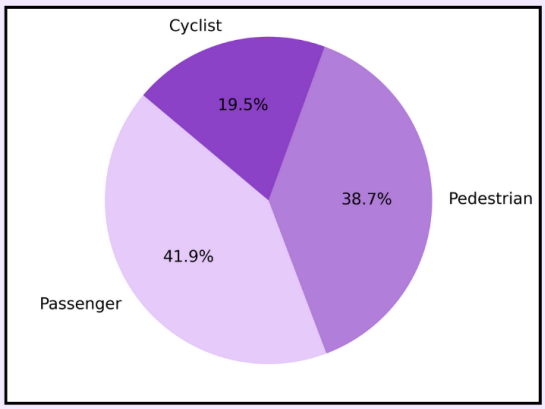
Children Casualties Factsheet May 2025

In 2015, the Department of Transport released a children casualties factsheet, providing a visualisation of the STATS19 data. As part of our 3rd year project, focusing on the safety of school children in South Yorkshire, we are providing a similar-styled factsheet, visualising trends in casualties for school aged children (5-16 years old) in South Yorkshire.

Number of children killed or seriously injured (1979-2023)



Child casualties per road user type (2021-2023)

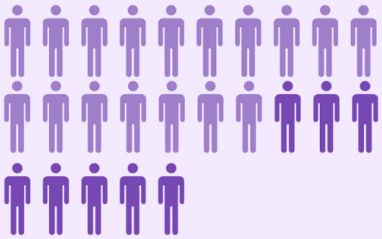


Child casualties in relation to school location (1979-2023)

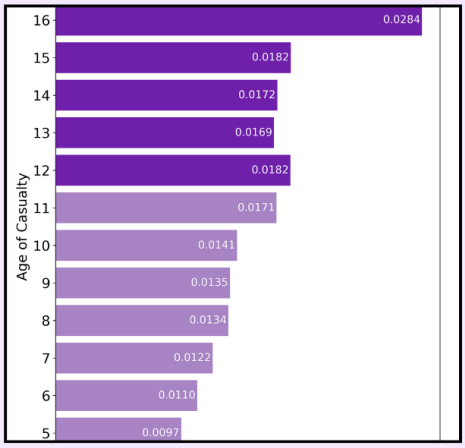
- **6.09%** of child casualties happened within a 2km radius from a school
- **23.04%** of child casualties happened within a 5km radius from a school
- **36.48%** of child casualties happened within a 10km radius from a school

Average number of casualties per month in 2023

- **17** children were slightly injured
- **8** children were seriously injured
- **0** children died



Child casualties per age normalised by population (1979-2023)



Child casualties per time of day (1979-2023)

- **7.14%** of child casualties occur between 8am and 8:59am am on a weekday
- **35.85%** of child casualties occur between 3pm and 6:59pm on a weekday
- **57.01%** of child casualties occur at any other time

Figure A.2: Children Casualties Factsheet for South Yorkshire

Link to download

B Resources, Software & Libraries

B.1 Links

Main project website (correlation plots, codes, resources, and datesets available here): https://phy346-road-safety-portfolio.vercel.app/ Stats19 analysis Github repo: https://github.com/Khu Route generation Github repo: https://github.com/dwporges/PHY-Road-Safety-Route-Generation

B.2 Python libraries

The analysis presented in this report was performed using Python 3.12.4. Several libraries were used to make this analysis possible

- **Fiona (1.10.1)**: For reading and writing of geographic data formats [26]
- **Folium (0.18.0)**: For visualisation of routes and locations [27]
- **GeoPandas (1.0.1)**: For working with geospatial data in dataframe format [17]
- **googlemaps (4.10.0)**: For interfacing with the Google Directions API [28]
- **NumPy (2.1.3)**: For data manipulation and analysis [29]
- **Pandas (2.2.3)**: For data manipulation and analysis [30]
- **PyShp (2.3.1)**: For handling of ESRI Shapefiles [31]
- **Scikit-learn (1.5.2)**: For machine learning algorithms [32]
- **Shapely (2.0.6)**: For manipulation and analysis of geometric objects [33]
- **googleroutes (1.0.0)**: For generating and manipulating routes [34]

N.B. this is not a requirements list for the `googleroutes` package, it is the software used in the analysis and the versions used. An up-to-date requirements.txt file is found in the Github repo

C Code Snippets

C.1 Length field for route generation

The following code was used to get the lengths of each route:

```
import geopandas as gpd

routes_gdf['length_m'] = routes_gdf.geometry.length
routes_gdf['length_km'] = routes_gdf.geometry.length / 1000
```

D Missing Value Analysis of the STATS19 Dataset

The merged STATS19 dataset's biggest issue is undoubtedly the high volume of missing values seen across different column over the years. For some columns, the missing values make up to 92% of the total values (See Figure D.1). The practices involving the recording of casualties in the STATS19 dataset changed several times since 1976. Some fields are sparsely populated in early years, as they were only implemented and systematically recorded in later years (See Figure D.2). By contrast, some fields were recorded in early years, but eventually dropped (See Figure D.2).

Name of column	No.of missing values	Percentage out of 243191 total values
casualty_imd_decile	224932	92.491910
enhanced_severity_collision	216823	89.157493
enhanced_casualty_severity	216812	89.152970
casualty_distance_banding	163013	67.030852
casualty_home_area_type	162835	66.957659
pedestrian_road_maintenance_worker	160638	66.054254
did_police_officer_attend_scene_of_accident	160619	66.046441
second_road_number	94677	38.931128
trunk_road_flag	89358	36.743958
urban_or_rural_area	89237	36.694203
junction_control	88623	36.441727
local_authority_district	8305	3.415011
age_band_of_casualty	3157	1.298157
age_of_casualty	3157	1.298157
second_road_class	238	0.097865
carriageway_hazards	141	0.057979
car_passenger	110	0.045232
road_surface_conditions	101	0.041531
special_conditions_at_site	48	0.019738
bus_or_coach_passenger	45	0.018504
pedestrian_crossing_human_control	41	0.016859
pedestrian_crossing_physical_facilities	37	0.015214
sex_of_casualty	14	0.005757
first_road_number	13	0.005346
light_conditions	1	0.000411

Figure D.1: A table showing the number of missing values from different columns in the STATS19 dataset, as absolute values and as percentages of the total amount of values in each column.

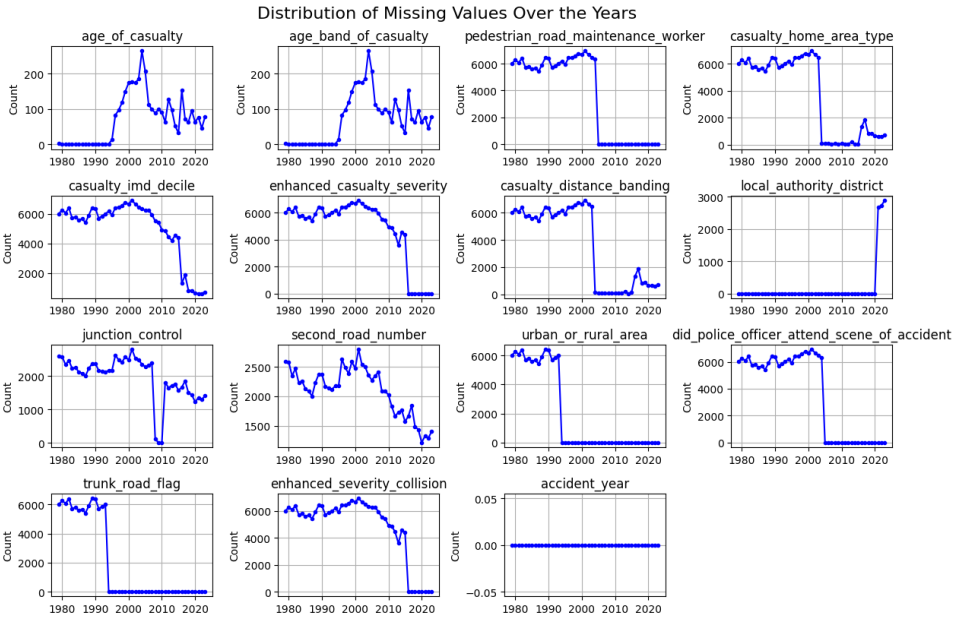


Figure D.2: A series of graph showcasing how the number of missing values varied over the each for each column in the STATS19. Most graphs showcase very abrupt changes in the number of missing values, while others show a constant variation or no variation at all.